DESTAQUES ACADÊMICOS
UNIVATES | CIÊNCIAS EXATAS E TECNOLÓGICAS
Revista Destaques Acadêmicos, Lajeado, v. 15, n. 4, 2023. ISSN 2176-3070
DOI: http://dx.doi.org/10.22410/issn.2176-3070.v15i4a2023.3635
http://www.univates.br/revistas

ANÁLISE DO DESEMPENHO DE HIPERPARÂMETROS DE APRENDIZAGEM DE MÁQUINA APLICADOS NA PREVISÃO DA TAXA DE ROTATIVIDADE DE CLIENTES

Renan Gnoatto[1], Evandro Franzen[2]

Resumo: O presente artigo apresenta uma análise da performance de diferentes hiperparâmetros de aprendizagem de máquina, aplicados a métodos de aprendizado de máquina, na predição da taxa de rotatividade de clientes de três conjuntos de dados de organizações prestadoras de serviços. Cada um destes conjuntos de dados passou por um processo que consistiu na coleta dos mesmos, no pré-processamento dos dados, na execução dos testes e na avaliação dos resultados. Após a coleta e transformação, foram aplicados os métodos de Floresta Aleatória, Árvore de Decisão e recursos para AutoML (*Auto Machine Learning*). Em cada método foram feitas duas aplicações, uma utilizando os hiperparâmetros padrões e outra com os valores destes parâmetros sendo customizados. Com base nos resultados obtidos, foi possível concluir que o método Floresta Aleatória se mostrou mais eficaz na predição da taxa da rotatividade de clientes, independente do conjunto de dados ou forma de utilização dos hiperparâmetros. Ainda assim, em algumas poucas métricas o melhor resultado foi atingido pelos outros métodos, como para a métrica Recall, classe 0, onde o método MLP atingiu um resultado de 0,96 no primeiro conjunto, utilizando hiperparâmetros padrões. Da mesma forma, para o terceiro conjunto, o método Árvore de Decisão atingiu o valor de 0,85 na métrica Recall, classe 0, utilizando hiperparâmetros personalizados.

Palavras-chave: Aprendizagem de máquina; Hiperparâmetros; Rotatividade de Clientes.

1. INTRODUÇÃO

Devido às mudanças nos mercados e as novas formas de venda, temos ofertas em demasia, o que faz com que o cliente dite os rumos do mercado. Diante deste cenário as organizações acabam tendo que focar nos clientes para entender melhor seus comportamentos e preferências, a fim de gerar maiores lucros e receitas (KHODABANDEHLOU; ZIVARI RAHMAN, 2017).

1 Graduado em Sistemas de Informação - UNIVATES.

2 Doutor em Informática na Educação - UFRGS.

Neste sentido, fidelizar um cliente tornou-se muito mais importante e lucrativo, pois atrair um novo pode custar por volta de 5 vezes mais do que manter um cliente atual satisfeito (KOTLER; KELLER, 2012). O objetivo sempre será a redução da rotatividade dos clientes, visando a mitigação de problemas de insatisfação com o serviço prestado ou com o suporte ao cliente, que concentram o maior número de queixas (LALWANI *et al*., 2021).

O ponto principal para tentar encontrar uma solução para a possível evasão de clientes é a previsão daqueles que possuem alguma possibilidade de abandonar o serviço. Ligado a isso, o conceito da previsão de rotatividade de clientes tem por objetivo encontrar essas soluções para retenção de clientes. Ou seja, ele tenta, de forma precoce, identificar os sinais de um possível abandono, gerando assim tempo para a empresa encontrar soluções para contornar o futuro risco de perda (LALWANI *et al*., 2021).

Diante dos avanços nas áreas de Inteligência Artificial (IA) e aprendizagem de máquina, a possibilidade de prever comportamentos dos clientes aumenta significativamente. Neste cenário, existem duas formas de tentar realizar a predição da rotatividade de clientes. Uma é a forma reativa e a outra é a forma pró-ativa. No caso da reativa a empresa age apenas após o pedido de cancelamento por parte do cliente, tentando o demover da sua decisão. Na forma pró-ativa é feita a predição da rotatividade, para então adiantar-se ao movimento do cliente (LALWANI et al., 2021).

Os métodos para classificação de uma base de dados podem ser entendidos como a simples caracterização do conhecimento adquirido ou a construção da predição de classes, apurado com base nos valores dos atributos de um conjunto de dados. Entre os métodos de classificação temos a Árvore de Decisão, o Floresta Aleatória, o MVS etc. (LEMOS; STEINER; NIEVOLA, 2005).

Levando em consideração que os clientes possuem muitas ofertas de qualquer serviço que tenham interesse, se torna imprescindível entender o comportamento de cada um deles e, se possível, tentar antever qualquer ação que o cliente venha a tomar, a fim de conseguir um diferencial competitivo perante os concorrentes.

Neste cenário, apesar das organizações possuírem bases de dados gigantescas sobre seus clientes, e consequentemente seus gostos, eles acabam não conseguindo extrair as informações valiosas que esses conjuntos de dados possuem, ficando à mercê do mercado, sem saber os motivos que levam seus clientes a trocarem os seus serviços pelos do concorrente. Este artigo visa identificar os métodos de aprendizagem de máquina que apresentam as melhores performances na previsão da taxa de rotatividade de clientes, de empresas prestadoras de serviços.

2. REFERENCIAL TEÓRICO

Reter clientes à organização é muito importante atualmente pois conseguir angariar novos clientes pode custar entre 5 e 10 vezes mais que manter um atual. A média de perda de clientes, anualmente, varia de 10% a 20%, fazendo com que diminuir essa taxa em apenas 5% pode fazer a empresa lucrar em torno de 25% a 85% a mais (FERREIRA, 2012).

Um dos principais fatores de fidelização de um cliente é a confiança que ele adquire ao ter as suas primeiras experiências com a empresa. Essa confiança é baseada em alguns fatores que estão relacionados à expectativa do cliente em relação à organização e ao produto e o que efetivamente a empresa apresenta. A lealdade do cliente é fundamental, mas também de nada adiantam clientes leais que não são revertidos em lucros para a empresa (FERREIRA, 2012).

Existem alguns bloqueios que acabam contribuindo para baixa rotatividade de clientes, ou seja, que contribuem para que o cliente não troque de fornecedor, mesmo insatisfeito, que são os custos associados à mudança. Por exemplo, no ramo industrial, as barreiras para mudança são mais altas, por isso é mais possível reter um cliente. Outros fatores que podem estar ligados a retenção do cliente são a falta de tempo do mesmo em buscar um novo fornecedor, o seu esforço psicológico para adaptação ao novo serviço e também a falta de certeza na negociação com o novo fornecedor (FERREIRA, 2012).

Desta forma, os clientes mantêm uma rede de fornecedores limitada, onde mantém negócios com o que atender melhor suas necessidades, levando em consideração a relação construída entre as partes. Essa lealdade é construída através do reconhecimento e recompensa, que pressupõe que as duas partes valorizam a relação.

Já Aprendizagem de Máquina tem como objetivo desenvolver a aprendizagem computacional, através de sistemas que capturem o conhecimento automaticamente. Um sistema com esta capacidade possui a característica de realizar escolhas assertivas, baseando-se em um histórico de resoluções satisfatórias anteriores (MONARD; BARANAUSKAS, 2003).

Aprendizagem de Máquina é voltada para concepção de programas que melhoram gradativamente o desempenho dos seus resultados de acordo com a experiência adquirida. É o processo de mutação que os sistemas associados à IA sofrem, executando tarefas como diagnóstico, reconhecimento, planejamento, previsão etc, em um conjunto de dados (NILSSON, 1998). Podemos classificá-la em dois métodos, supervisionada e não supervisionada. No caso da aprendizagem supervisionada o algoritmo aprende com base em um resultado pré-definido, ou seja, ele trabalha já possuindo uma referência do que está certo ou não. Já na aprendizagem não supervisionada o algoritmo trabalha tentando descobrir padrões ou agrupamentos sem ter uma referência de resultado desejado, tentando assim encontrar a melhor solução (BI *et al*., 2019).

Aprendizagem de Máquina, dentre tantas utilidades, mostrou-se ser de grande valia para as empresas em relação à rotatividade de clientes, pois se o algoritmo for concebido de uma forma harmoniosa é capaz de prever informações com base em conjuntos de dados de clientes ativos e não ativos, de forma pró-ativa (LALWANI et al., 2022).

Uma característica comum entre os métodos de Aprendizagem de Máquina e Aprendizagem Profunda é a possibilidade de configurá-los através de conjuntos de hiperparâmetros, que quando definidos da forma mais adequada possível podem resultar na maximização da utilização do método de aprendizagem. Estes hiperparâmetros tem por objetivo configurar inúmeros aspectos do algoritmo e podem fazer o resultado variar muito de acordo com sua utilização (CLAESEN; MOOR, 2015). Os hiperparâmetros sempre são definidos anteriormente ao processo de treinamento de uma base, levando em consideração que todos os métodos de aprendizagem de máquina passam por este processo de treino (PELLICER, 2020).

Para avaliação dos métodos podemos utilizar algumas métricas de desempenho dos algoritmos como acurácia, precisão, recall, F1 score ou matriz de confusão. A acurácia pode ser entendida como o número de instâncias corretamente classificadas em relação a determinado conjunto de dados, e tem sua utilização melhor justificada em uma aprendizagem supervisionada onde o objetivo é a mitigação do erro global (CASTRO; BRAGA, 2011).

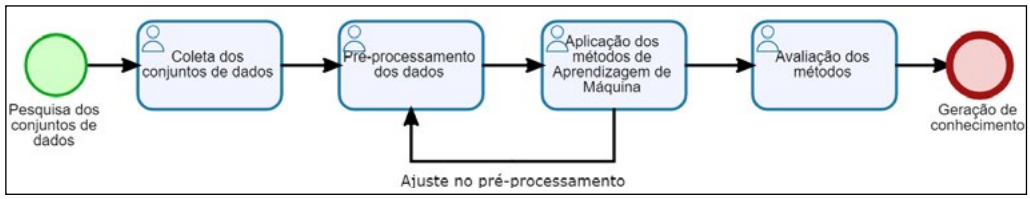
A precisão corresponde à relação entre o número de verdadeiros positivos e total de instâncias classificadas como positivas pelo algoritmo. Neste caso a métrica é dividir o número de verdadeiros positivos pela soma de todos os verdadeiros positivos e falsos negativos. Instâncias classificadas como falsas negativas são positivas classificadas como negativas (CASTRO; BRAGA, 2011; BRAMER, 2013). Recall é a proporção de instâncias que foram corretamente previstas como verdadeiras. Pode ser encontrada através da métrica de verdadeiros positivos divididos pela soma dos verdadeiros positivos e falsos negativos (BRAMER, 2013).

3. METODOLOGIA

O processo de desenvolvimento foi separado em etapas, e cada uma foi baseada no processo de geração de conhecimento chamado Knowledge Discovery in Databases (KDD). Este processo é caracterizado como a detecção de padrões, em uma determinada quantidade de dados, que possuem grande possibilidade de serem úteis (MIKUT; REISCHL, 2011). Ele é composto de etapas sequenciais que podem ser seguidas da seguinte maneira: entendimento do campo a ser analisado, entendimento dos dados coletados, que possuem relação com o campo em questão, tratamento de dados irrelevantes e redundantes, busca e recolhimento de padrões de dados e apresentação dos

resultados (AFSHAR et al., 2015). Na Figura 1 há a representação das etapas abrangidas neste trabalho.

Figura 1 - Representação das etapas do desenvolvimento



Fonte: Do autor (2022) adaptado de Fayyad *et al* (1996c)

3.1 Coleta dos conjuntos de dados

O primeiro passo do desenvolvimento da proposta de projeto foi a seleção dos conjuntos de dados que foram utilizados. Neste trabalho foram utilizados os três conjuntos de dados listados abaixo, com seu respectivo link de acesso, que foram captados do site www.kaggle.com. A plataforma Kaggle é uma das mais relevantes na área de ciência de dados, uma vez que possui diversos recursos, como bases de dados públicas. Cada um dos conjuntos de dados possui informações e tamanhos distintos, e são de diferentes ramos de prestação de serviços.

- Churn Telco Europa:

  https://www.kaggle.com/datasets/raumonsa11/churn-telco-europa
- Bank Customer Churn Prediction:

  https://www.kaggle.com/datasets/shantanudhakadd/bank-customer-churn-prediction
- Internet Service Provider Customer Churn:

  https://www.kaggle.com/datasets/mehmetsabrikunt/internet-service-churn

3.2 Pré-processamento dos dados

No segundo passo, cada conjunto de dados foi importado na plataforma do Google Colaboratory e passou por um pré-processamento, que consiste, inicialmente, na análise dos dados brutos, a fim de estabelecer as informações mais relevantes para a predição em questão. Posteriormente a isso, foi efetuada a limpeza do conjunto de dados, removendo assim registros nulos, duplicados e não relevantes, tendo como objetivo a diminuição do tamanho do conjunto. Esta abordagem de pré-processamento dos dados já é conhecida e usualmente testada em vários trabalhos, como o desenvolvido por Lalwani *et al*. (2021),

onde a etapa de pré-processamento dos dados foi separada em três fases, que consistem na localização dos dados mais relevantes, limpeza e filtragem dos dados e seleção de características.

3.3 Aplicação de métodos de aprendizagem de máquina

Antes da aplicação dos métodos de aprendizagem cada conjunto de dados foi separado em conjuntos de treinamento e teste, onde foi utilizada a abordagem 80:20, que visa alocar 80% do total de registros para o treinamento e 20% para o teste (FIGURA 2).

Figura 2 - Separação do conjunto em treinamento e teste

```
numatributos = len(clients.columns) - 1
atributos = list(clients.columns[0:numatributos])

x = clients[atributos]
y = clients['CHURN']

treinamento_x, teste_x, treinamento_y, teste_y=train_test_split(x,
                                                                y,
                                                                test_size=0.20)
```

Fonte: Do autor (2023).

Para cada conjunto de dados foram aplicados os métodos de aprendizagem de máquina de duas formas distintas. Na primeira, utilizando o classificador padrão de cada um deles, e depois personalizando alguns. Os métodos de aprendizagem de máquina utilizados foram Floresta Aleatória, Árvore de Decisão, MVS e Multilayer Perceptron (MLP), da biblioteca Scikit-learn. Além destes, também foi utilizado um modelo de Automated Machine Learning (AutoML), que tem por objetivo afinar automaticamente as configurações dos hiperparâmetros dos métodos de Aprendizagem de Máquina, diminuindo o tempo de configuração dos mesmos. Para este caso foi feito uso da biblioteca PyCaret, onde foram realizadas predições dos métodos Floresta Aleatória, Árvore de Decisão e MVS.

Em relação ao teste com personalização dos hiperparâmetros, foram selecionados alguns de forma aleatória, levando em consideração sua descrição e objetivos no método. Com isso, para cada método de aprendizagem foram selecionados variados hiperparâmetros, que tiveram um valor informado de forma manual. Alguns, inclusive, continuaram com seus valores padrões, porém, em conjunto com outros personalizados. Para todos os conjuntos de dados foram informados os mesmos hiperparâmetros personalizados, conforme abaixo:

- Floresta Aleatória: bootstrap=True, max_depth=80, max_features=30, min_samples_leaf=5, min_samples_split=10, n_estimators=100
- Árvore de Decisão: criterion="gini", max_depth=80, min_samples_split=10, min_samples_leaf=5, max_features=30, max_leaf_nodes=100
- MVS: kernel='linear', degree=8
- MLP: solver='adam', activation='logistic', alpha=1e-5, hidden_layer_sizes=(200,), random_state=1

3.4 Avaliação dos métodos

Após os testes, os desempenhos de cada conjunto de dados foram avaliados, conforme o método de aprendizagem de máquina utilizado, bem como seus hiperparâmetros aplicados. Neste caso, as métricas avaliadas foram Acurácia, Precisão, Recall, F1 score e Area Under Curve (AUC).

3.4.1 Acurácia

A acurácia pode ser entendida como o número de instâncias corretamente classificadas em relação a determinado conjunto de dados, e tem sua utilização melhor justificada em uma aprendizagem supervisionada onde o objetivo é a mitigação do erro global (CASTRO; BRAGA, 2011). O valor da acurácia pode ser conhecido através da soma dos valores classificados como verdadeiros positivos e verdadeiros negativos, divididos pelo número total da amostra. Valores verdadeiros positivos são totais de instâncias positivas classificadas como positivas e os verdadeiros negativos são o total de instâncias negativas classificadas como negativas (BRAMER, 2013).

3.4.2 Precisão

Corresponde a relação entre o número de verdadeiros positivos e total de instâncias classificadas como positivas pelo algoritmo. Neste caso a métrica é dividir o número de verdadeiros positivos pela soma de todos os verdadeiros positivos e falsos negativos. Instâncias classificadas como falsas negativas são positivas classificadas como negativas (CASTRO; BRAGA, 2011; BRAMER, 2013).

3.4.3 Recall

Recall é a proporção de instâncias que foram corretamente previstas como verdadeiras. Pode ser encontrada através da métrica de verdadeiros positivos divididos pela soma dos verdadeiros positivos e falsos negativos (BRAMER, 2013). O Recall pode ser considerado como o principal indicador para determinar se o modelo foi capaz de classificar corretamente os registros de uma determinada classe.

3.4.4 F1 score

F1 score realiza a combinação dos métodos de avaliação Precisão e Recall, sendo considerado a média harmônica das duas métricas. Esta avaliação pode ser obtida pela fórmula: (2 x precisão x recall) / precisão + recall (BRAMER, 2013). Em situações nas quais a precisão e o recall apresentam valores distintos, este indicador é usado para analisar o equilibro do modelo, busca-se, portanto, um modelo capaz de exibir bons indicadores tanto para recall, quanto para precision.

3.4.5 AUC

AUC pode ser entendida como um resumo da exatidão discriminatória de um teste ou como um resumo da própria curva ROC. A escala de AUC vai de 0 a 1, sendo 1, ou mais perto possível disso, sinônimo de alta precisão no teste proposto (GONÇALVES; SUBTIL; OLIVEIRA, 2014).

4. RESULTADOS E DISCUSSÃO

Neste capítulo serão apresentados os resultados da aplicação dos métodos de aprendizado de máquina nos diferentes conjuntos de dados e com diferentes configurações de parâmetros. Após a realização dos testes foi possível analisar os resultados obtidos para cada conjunto de dados e estabelecer conclusões sobre quais métodos apresentaram os melhores resultados, bom como quais combinações de parâmetros são adequadas para resolução do problema de retenção de clientes.

O primeiro conjunto de dados a ser analisado foi o *Churn Telco Europa*, que apresentou valores elevados de Performance ao utilizar os Métodos de Aprendizagem de Máquina Floresta Aleatória e Árvore de Decisão, tanto com hiperparâmetros padrões quanto personalizados. Inclusive, a personalização dos hiperparâmetros não surtiu um efeito muito grande nestes dois métodos mencionados, pois muitas métricas apresentaram o mesmo valor nas duas predições. Inclusive, em algumas, houve uma pequena queda de performance com hiperparâmetros personalizados. Já para os métodos MVS e MLP, que apresentaram uma performance menor, a personalização dos hiperparâmetros foi um pouco mais eficaz, com destaque para o MLP que apresentou uma melhora significativa em oito dos dez valores de métrica avaliados. A Tabela 1 mostra os valores obtidos em cada métrica e destaca com cores vermelho e verde se o valor com hiperparâmetros personalizados foi menor ou maior que com hiperparâmetros padrões, respectivamente.

Tabela 1 - Todos os resultados do conjunto Churn Telco Europa

	Hiperparâmetros	Acurácia		Precisão		Recall		F1		AUC	
		0	1	0	1	0	1	0	1	0	1
Floresta Aleatória	Padrões	0,95	0,95	0,98	0,92	0,93	0,98	0,96	0,95	0,96	0,96
Floresta Aleatória	Personalizados	0,95	0,95	0,98	0,93	0,94	0,97	0,96	0,95	0,95	0,95
Árvore de Decisão	Padrões	0,93	0,93	0,95	0,91	0,92	0,94	0,94	0,92	0,93	0,93
Árvore de Decisão	Personalizados	0,93	0,93	0,96	0,90	0,91	0,96	0,93	0,92	0,93	0,93
MVS	Padrões	0,73	0,73	0,72	0,77	0,85	0,60	0,78	0,67	0,72	0,72
MVS	Personalizados	0,71	0,71	0,78	0,65	0,65	0,79	0,71	0,71	0,72	0,72
MLP	Padrões	0,67	0,67	0,61	0,90	0,96	0,39	0,75	0,55	0,67	0,67
MLP	Personalizados	0,78	0,78	0,76	0,80	0,81	0,74	0,79	0,77	0,78	0,78

Fonte: Do autor (2023).

Ao analisar apenas os resultados da predição com hiperparâmetros padrões é possível notar que o método Floresta Aleatória foi o melhor em basicamente todos as performances, com exceção para o Recall da classe 0, onde o que melhor performou foi o método MLP (TABELA 2).

Tabela 2 - Resultados com hiperparâmetros padrões para o conjunto Churn Telco Europa

	Hiperparâmetros Padrões									
	Acurácia		Precisão		Recall		F1		AUC	
	0	1	0	1	0	1	0	1	0	1
Floresta Aleatória	0,95	0,95	0,98	0,92	0,93	0,98	0,96	0,95	0,96	0,96
Árvore de Decisão	0,93	0,93	0,95	0,91	0,92	0,94	0,94	0,92	0,93	0,93
MVS	0,73	0,73	0,72	0,77	0,85	0,60	0,78	0,67	0,72	0,72
MLP	0,67	0,67	0,61	0,90	0,96	0,39	0,75	0,55	0,67	0,67

Fonte: Do autor (2023).

Já para a predição com hiperparâmetros personalizados o método Floresta Aleatória foi o melhor em todas as métricas, incluindo o Recall da classe 0 (TABELA 3).

Tabela 3 - Resultados com hiperparâmetros personalizados para o conjunto Churn Telco Europa

	Hiperparâmetros personalizados									
	Acurácia		Precisão		Recall		F1		AUC	
	0	1	0	1	0	1	0	1	0	1
Floresta Aleatória	0,95	0,95	0,98	0,93	0,94	0,97	0,96	0,95	0,95	0,95
Árvore de Decisão	0,93	0,93	0,96	0,90	0,91	0,96	0,93	0,92	0,93	0,93
MVS	0,71	0,71	0,78	0,65	0,65	0,79	0,71	0,71	0,72	0,72
MLP	0,78	0,78	0,76	0,80	0,81	0,74	0,79	0,77	0,78	0,78

Fonte: Do autor (2023).

Em relação à predição utilizando o AutoML os resultados obtidos para o conjunto de dados *Churn Telco Europa* seguiram o mesmo padrão já apresentado, onde o método de aprendizagem Floresta Aleatória apresentou as melhores performances para praticamente todas as métricas, com exceção do valor de Recall do método MVS, que atingiu o valor máximo, 1,00. Vale ressaltar que o método Árvore de Decisão mostrou-se bastante eficiente, pois apresentou valores parecidos com o método Floresta Aleatória, ficando 0,01 abaixo em três das cinco métricas avaliadas, conforme Tabela 4.

Tabela 4 - Todos os resultados do AutoML para o conjunto Churn Telco Europa

	Acurácia	Precisão	Recall	F1	AUC
Floresta Aleatória	0,95	0,95	0,99	0,97	0,94
Árvore de Decisão	0,94	0,94	0,99	0,96	0,90
MVS	0,88	0,87	1,00	0,93	0,61

Fonte: Do autor (2023).

Já para o conjunto de dados *Bank Customer Churn Prediction* a predição utilizando os hiperparâmetros personalizados se mostrou mais positiva, pois dentre os 40 resultados de Performance das métricas avaliadas, utilizando os métodos da biblioteca Scikit-learn, 30 apresentaram aumento no valor, enquanto em 5 o valor foi menor e em 5 a performance foi igual, conforme pode ser visto na Tabela 5.

Tabela 5- Todos os resultados do conjunto Bank Customer Churn Prediction

	Hiperparâmetros	Acurácia		Precisão		Recall		F1		AUC	
		0	1	0	1	0	1	0	1	0	1
Floresta Aleatória	Padrões	0,82	0,82	0,80	0,84	0,81	0,83	0,81	0,84	0,82	0,82
Floresta Aleatória	Personalizados	0,82	0,82	0,82	0,83	0,77	0,87	0,79	0,85	0,82	0,82
Árvore de Decisão	Padrões	0,73	0,73	0,67	0,80	0,77	0,70	0,72	0,75	0,74	0,74
Árvore de Decisão	Personalizados	0,76	0,76	0,74	0,77	0,74	0,78	0,74	0,78	0,76	0,76
MVS	Padrões	0,63	0,63	0,68	0,61	0,43	0,82	0,53	0,70	0,63	0,63
MVS	Personalizados	0,74	0,74	0,83	0,70	0,51	0,92	0,63	0,80	0,71	0,71
MLP	Padrões	0,63	0,63	0,63	0,63	0,39	0,82	0,48	0,71	0,60	0,60
MLP	Personalizados	0,66	0,66	0,76	0,63	0,38	0,90	0,51	0,74	0,64	0,64

Fonte: Do autor (2023).

Ao analisar cada predição de forma isolada, podemos notar que, novamente para ambas as predições, a melhor performance foi encontrada com o método Floresta Aleatória. Ao utilizar os hiperparâmetros padrões, em todas as métricas analisadas a melhor performance foi do método Floresta Aleatória, apresentando uma diferença considerável se comparado com o segundo método de melhor performance, que foi o Árvore de Decisão (TABELA 6).

Tabela 6 - Resultados com hiperparâmetros padrões para o conjunto Bank Customer Churn Prediction

	Hiperparâmetros Padrões									
	Acurácia		Precisão		Recall		F1		AUC	
	0	1	0	1	0	1	0	1	0	1
Floresta Aleatória	0,82	0,82	0,80	0,84	0,81	0,83	0,81	0,84	0,82	0,82
Árvore de Decisão	0,73	0,73	0,67	0,80	0,77	0,70	0,72	0,75	0,74	0,74
MVS	0,63	0,63	0,68	0,61	0,43	0,82	0,53	0,70	0,63	0,63
MLP	0,63	0,63	0,63	0,63	0,39	0,82	0,48	0,71	0,60	0,60

Fonte: Do autor (2023).

Ao comparar as performances utilizando os hiperparâmetros personalizados, o método Floresta Aleatória continuou à frente dos demais, porém, neste caso, ficando abaixo do MVS em duas métricas, que foram Precisão da classe 0 e Recall da classe 1 (TABELA 7).

Tabela 7 - Resultados com hiperparâmetros personalizados para o conjunto Bank Customer Churn Prediction

	Hiperparâmetros Personalizados									
	Acurácia		Precisão		Recall		F1		AUC	
	0	1	0	1	0	1	0	1	0	1
Floresta Aleatória	0,82	0,82	0,82	0,83	0,77	0,87	0,79	0,85	0,82	0,82
Árvore de Decisão	0,76	0,76	0,74	0,77	0,74	0,78	0,74	0,78	0,76	0,76
MVS	0,74	0,74	0,83	0,70	0,51	0,92	0,63	0,80	0,71	0,71
MLP	0,66	0,66	0,76	0,63	0,38	0,90	0,51	0,74	0,64	0,64

Fonte: Do autor (2023).

Ao utilizar o AutoML, o método Floresta Aleatória ainda continuou apresentando a melhor performance na maioria das métricas, porém neste caso o método Árvore de Decisão igualou em Acurácia e até foi melhor em Precisão, levando em consideração neste caso o fato de ter sido gerada uma performance apenas para ambas as classes preditas (TABELA 8).

Tabela 8 - Todos os resultados do AutoML para o conjunto Bank Customer Churn Prediction

	Acurácia	Precisão	Recall	F1	AUC
Floresta Aleatória	0,84	0,72	0,50	0,59	0,85
Árvore de Decisão	0,84	0,75	0,43	0,55	0,83
MVS	0,69	0,25	0,20	0,22	0,51

Fonte: Do autor (2023).

Por fim, ao analisar o conjunto *Internet Service Provider Customer Churn* foi possível notar que ele seguiu o mesmo padrão dos demais, apresentando a melhor performance ao utilizar o método Floresta Aleatória. Neste caso, ao comparar os hiperparâmetros padrões com os personalizados, temos o número de 23 performances apresentando valor maior ao utilizar os personalizados, enquanto 13 apresentaram diminuição na comparação com a predição padrão. Outras 4 performances mantiveram o mesmo valor, conforme Tabela 9.

Tabela 9- Todos os resultados do conjunto Internet Service Provider Customer Churn

	Hiperparâmetros	Acurácia		Precisão		Recall		F1		AUC	
		0	1	0	1	0	1	0	1	0	1
Floresta Aleatória	Padrões	0,81	0,81	0,78	0,85	0,85	0,78	0,81	0,81	0,81	0,81
Floresta Aleatória	Personalizados	0,82	0,82	0,79	0,85	0,84	0,80	0,82	0,82	0,82	0,82
Árvore de Decisão	Padrões	0,76	0,76	0,75	0,75	0,74	0,74	0,74	0,74	0,75	0,75
Árvore de Decisão	Personalizados	0,81	0,81	0,78	0,84	0,85	0,77	0,81	0,81	0,81	0,81
MVS	Padrões	0,74	0,74	0,70	0,78	0,80	0,67	0,75	0,72	0,74	0,74
MVS	Personalizados	0,72	0,72	0,72	0,73	0,69	0,76	0,71	0,74	0,72	0,72
MLP	Padrões	0,77	0,77	0,73	0,83	0,83	0,72	0,78	0,77	0,78	0,78
MLP	Personalizados	0,77	0,77	0,74	0,81	0,81	0,73	0,77	0,77	0,77	0,77

Fonte: Do autor (2023).

Na utilização dos hiperparâmetros padrões, o método Floresta Aleatória foi o melhor para todas as 10 performances, das 5 métricas analisadas, conforme Tabela 10. Já utilizando os hiperparâmetros personalizados o método Floresta Aleatória só não apresentou a melhor performance para o Recall da classe 0, que foi onde o método Árvore de Decisão atingiu o maior valor (TABELA 11).

Tabela 10 - Resultados com hiperparâmetros personalizados para o conjunto Internet Service Provider Customer Churn

	Hiperparâmetros Padrões									
	Acurácia		Precisão		Recall		F1		AUC	
	0	1	0	1	0	1	0	1	0	1
Floresta Aleatória	0,81	0,81	0,78	0,85	0,85	0,78	0,81	0,81	0,81	0,81
Árvore de Decisão	0,76	0,76	0,75	0,75	0,74	0,74	0,74	0,74	0,75	0,75
MVS	0,74	0,74	0,70	0,78	0,80	0,67	0,75	0,72	0,74	0,74
MLP	0,77	0,77	0,73	0,83	0,83	0,72	0,78	0,77	0,78	0,78

Fonte: Do autor (2023).

Tabela 11 - Resultados com hiperparâmetros personalizados para o conjunto Internet Service Provider Customer Churn

	Hiperparâmetros Pesonalizados									
	Acurácia		Precisão		Recall		F1		AUC	
	0	1	0	1	0	1	0	1	0	1
Floresta Aleatória	0,82	0,82	0,79	0,85	0,84	0,80	0,82	0,82	0,82	0,82
Árvore de Decisão	0,81	0,81	0,78	0,84	0,85	0,77	0,81	0,81	0,81	0,81
MVS	0,72	0,72	0,72	0,73	0,69	0,76	0,71	0,74	0,72	0,72
MLP	0,77	0,77	0,74	0,81	0,81	0,73	0,77	0,77	0,77	0,77

Fonte: Do autor (2023).

Em relação à predição utilizando o AutoML seguiu-se o mesmo padrão apresentado para as demais predições, que foi o método Floresta Aleatória apresentando a melhor performance em todas as métricas, conforme Tabela 12.

Tabela 12 - Todos os resultados do AutoML para o conjunto Internet Service Provider Customer Churn

	Acurácia	Precisão	Recall	F1	AUC
Floresta Aleatória	0,81	0,84	0,78	0,81	0,88
Árvore de Decisão	0,75	0,78	0,72	0,75	0,82
MVS	0,74	0,74	0,77	0,75	0,74

Fonte: Do autor (2023).

Os resultados encontrados neste trabalho são semelhantes a outros voltados para a mesma área, como por exemplo o trabalho de Ahmad, Jafar e Aljoumaa (2019), intitulado "Previsão da rotatividade de clientes em telecomunicações utilizando a aprendizagem de máquina em grandes plataformas de dados", que aplicou quatro métodos de aprendizagem de máquina para previsão da taxa de rotatividade de clientes. Nele, podemos ver que o método Floresta Aleatória obteve melhor performance que Árvore de Decisão, se comparamos a métrica utilizada no trabalho, que foi AUC.

5. CONCLUSÕES

Reter um cliente sempre foi um desafio que as empresas enfrentam a muitos anos, e com o avanço da tecnologia e o grande número de informações que as empresas geram de cada cliente, essa tarefa ganha poderosas armas que podem auxiliar de forma muito positiva.

Neste sentido, este trabalho objetivou, inicialmente, levantar um número suficiente de trabalhos relacionados ao problema de pesquisa definido, a fim de criar uma base de conhecimento para o desenvolvimento do trabalho. Baseando-se no problema de pesquisa e no referencial teórico produzido, foram definidas as etapas de desenvolvimento do trabalho, na qual se baseiam no processo de geração de conhecimento KDD. Foram definidas as seguintes etapas para o trabalho: seleção dos conjuntos de dados, seleção dos dados relevantes, pré-processamento dos dados, execução do treinamento e teste dos métodos, avaliação dos métodos e geração de conhecimento.

Ao final do processo de desenvolvimento foi possível chegar a resultados concretos, para cada conjunto de dados e método de aprendizagem de máquina. Analisando os dados obtidos para todos os conjuntos de dados, foi possível constatar que que a predição com hiperparâmetros personalizados pode ser considerada satisfatória para os métodos da biblioteca Scikit-learn, pois apresentou uma melhora em 68 das 120 performances analisadas, se

comparado com a predição com hiperparâmetros padrões. Em 29 a performance foi menor. Em 23 o valor foi o mesmo para ambas predições.

Para efeito de entendimento desta análise, cada classe predita foi considerada uma performance, então, como para cada métrica (Acurácia, Precisão, Recall, F1 score e AUC) analisada existiam 2 classes preditas, para cada método de aprendizagem foram consideradas 10 performances, totalizando 120 para todos os conjuntos de dados, nos métodos da biblioteca Scikit-learn.

Considerando os conjuntos de dados, o que apresentou as melhores performances foi o Churn Telco Europa, que é o que possui o maior número de dados (188754 registros). Os valores de performance deste conjunto foram os mais altos tanto na predição com hiperparâmetros personalizados quanto com os padrões e também com AutoML. Mesmo tendo sido rebalanceado, ficou claro que quanto mais registros o conjunto possuir, melhor trabalham os métodos de aprendizagem de máquina, pois possuem uma variedade muito maior de informações.

O conjunto de dados Bank Customer Churn Prediction foi o que respondeu melhor à personalização de hiperparâmetros, pois em 30 das 40 performances obtidas, dos métodos da biblioteca Scikit-learn, ele obteve uma melhora, se comparada com a predição padrão. Este conjunto, inclusive, é o que possui o menor número de registros, sendo possível concluir que a personalização de hiperparâmetros respondeu melhor ao baixo número de registros e conseguiu achar alternativas para entregar uma predição mais assertiva.

Apenas dois métodos de Aprendizagem de Máquina apresentaram performance maior, em todas as métricas analisadas, com personalização de hiperparâmetros, em relação aos padrões. Neste caso foram o método MVS, do conjunto Bank Customer Churn, e Árvore de Decisão, do conjunto Internet Service Provider Customer Churn.

De modo geral o método de aprendizagem de máquina que apresentou a melhor performance foi o Floresta Aleatória, pois ao comparar as predições com hiperparâmetros padrões e personalizados, em apenas 1 das 10 performances este método não apresentou o valor mais alto. Da mesma forma, na predição com AutoML o método Floresta Aleatória foi o melhor em 4 das 5 métricas, tendo em vista que para estas predições foi gerado apenas um valor de performance para cada métrica, diferentemente dos métodos da biblioteca Scikit-learn, que geraram uma performance para cada classe, sendo duas por métrica.

Por conta deste motivo não é possível realizar a comparação das performances das predições dos métodos da biblioteca Scikit-learn com as predições do AutoML, porém o único método que chegou ao valor máximo de predição foi o MVS, utilizando AutoML, que atingiu o valor de 1,00 para a métrica Recall. Em relação a todas as ferramentas utilizadas, o Google Collaboratory

mostrou-se ser muito eficaz para a criação do modelo de predição, tendo em vista que nele é possível organizar o código de forma sucinta e dinâmica. Outra ferramenta que foi satisfatória foi a biblioteca Scikit-learn, que disponibilizou os métodos de aprendizagem de máquina utilizados no decorrer do trabalho, e que se mostraram eficientes no problema de pesquisa.

Por fim, levando em consideração o conhecimento gerado na fase de análise dos resultados deste trabalho, conclui-se que o método de aprendizagem de máquina Floresta Aleatória, dentre os métodos utilizados no trabalho, é o mais recomendado para realizar a predição da taxa da rotatividade de clientes, independentemente de haver personalização ou não dos hiperparâmetros, pois a performance foi satisfatória para ambas predições. Inclusive, se a predição for feita através do AutoML, a performance apresentada pode ser ainda maior.

REFERÊNCIAS

AFSHAR, Hadi Lotfnezhad et al. Prediction of Breast Cancer Survival Through Knowledge Discovery in Databases. Global Journal of Health Science, [s. l.], v. 7, n. 4, p. 392–398, 2015. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4802184/. Acesso em: 16 ago. 2022.

AHMAD, Abdelrahim Kasem; JAFAR, Assef; ALJOUMAA, Kadan. Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, [s. l.], v. 6, n. 1, p. 28, 2019. Disponível em: https://doi.org/10.1186/s40537-019-0191-6. Acesso em: 30 jul. 2022.

BI, Qifang et al. What is Machine Learning? A Primer for the Epidemiologist. American Journal of Epidemiology, [s. l.], p. kwz189, 2019. Disponível em: https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwz189/5567515. Acesso em: 14 ago. 2022.

BRAMER, Max. Principles of Data Mining. 2ª ed. London: Springer, 2013. (Undergraduate Topics in Computer Science). E-book. Disponível em: https://doc.lagout.org/Others/Data%20Mining/Principles%20of%20Data%20Mining%20%282nd%20ed.%29%20%5BBramer%202013-02-21%5D.pdf. Acesso em: 21 ago. 2022.

CASTRO, Cristiano Leite de; BRAGA, Antônio Pádua. Aprendizado supervisionado com conjuntos de dados desbalanceados. Sba: Controle & Automação Sociedade Brasileira de Automatica, [s. l.], v. 22, n. 5, p. 441–466, 2011. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-17592011000500002&lng=pt&tlng=pt. Acesso em: 23 ago. 2022.

FAYYAD, U.M., PIATETSKY-SHAPIRO, G., SMYTH, P. From data mining to knowledge discovery in databases. Artificial Intelligence Magazine, v. 17, n. 3, p. 37-54, 1996c.

FERREIRA, Célia Marina Costa. Um estudo sobre fidelização e retenção de clientes na área do fitness. Dissertação de Mestrado, [s. l.], 2012. Disponível em: https://repositorio.ipcb.pt/handle/10400.11/1701. Acesso em: 7 set. 2022.

GONÇALVES, Luzian et al. ROC Curve Estimation. REVSTAT-Statistical Journal, p. 1-20 Pages, 2014. Disponível em: <https://revstat.ine.pt/index.php/REVSTAT/article/view/141>. Acesso em: 8 maio 2023.

KHODABANDEHLOU, Samira; ZIVARI RAHMAN, Mahmoud. Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. Journal of Systems and Information Technology, [s. l.], v. 19, n. 1/2, p. 65–93, 2017. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/JSIT-10-2016-0061/full/html. Acesso em: 8 out. 2022.

KOTLER, Philip; KELLER, Kevin L. Administração de Marketing. 14ª ed. São Paulo: Pearson Education do Brasil, 2012.

LALWANI, Praveen et al. Customer churn prediction system: a machine learning approach. Computing, [s. l.], v. 104, n. 2, p. 271–294, 2022. Disponível em: https://link.springer.com/10.1007/s00607-021-00908-y. Acesso em: 30 jul. 2022.

MIKUT, Ralf; REISCHL, Markus. Data mining tools: Data mining tools. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, [s. l.], v. 1, n. 5, p. 431–443, 2011. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/widm.24. Acesso em: 16 ago. 2022.

MITCHELL, Tom M. Machine Learning. New York: McGraw-Hill, 1997.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos Sobre Aprendizado de Máquina. Sistemas Inteligentes Fundamentos e Aplicações. 1ª ed. Barueri, São Paulo: Manole Ltda, 2003.

NILSSON, Nils J. Introduction to Machine Learning: an early draft of a proposed textbook. Department of Computer Science, Stanford University, [s. l.], 1998. Disponível em: https://ai.stanford.edu/~nilsson/MLBOOK.pdf. Acesso em: 14 ago. 2022.