Aprendizado de Máquina na Predição de Evasão em Academias: Modelos e Estratégias

João Victor Dias Gomes
Departamento de Computação
CEFET-MG, Belo Horizonte, Brasil
jaovdg@gmail.com

Rogerio Martins Gomes
Departamento de Computação
CEFET-MG, Belo Horizonte, Brasil
rogerio@cefetmg.br

Bruno Andre Santos
Departamento de Computação
CEFET-MG, Belo Horizonte, Brasil
bsantos@cefetmg.br

***Resumo*—Este estudo investiga a aplicação de técnicas de aprendizado de máquina na previsão de evasão de clientes em academias. Utilizando um conjunto de dados com 4.000 registros de clientes, foram avaliados seis algoritmos de classificação supervisionada (Decision Tree, Logistic Regression, SVM, Random Forest, XGBoost e Neural Networks), combinados com quatro métodos de seleção de atributos (InfoGain, mRMR, Forward Selection e PCA). A validação cruzada foi aplicada para garantir o rigor experimental e o ajuste de hiperparâmetros via Optuna foi usado visando melhorar as soluções dos modelos. Os resultados mostraram que os atributos mais relevantes para a previsão de evasão foram a frequência de uso, o tempo de associação e a idade do cliente. A Rede Neural apresentou o melhor desempenho geral (acurácia de 95%, revocação de 89%, AUC = 0,931), enquanto o XGBoost destacou-se em precisão (96%) e F1-score (91%). Entre os métodos de seleção, o InfoGain foi o mais eficiente, reduzindo a dimensionalidade sem perda de desempenho. Os resultados evidenciam o potencial de modelos preditivos na antecipação do comportamento de evasão, oferecendo subsídios para estratégias de retenção mais assertivas e orientadas por dados.**

***Palavras-chave*—evasão, aprendizado de máquina, seleção de atributos, academias, predição, redes neurais, XGBoost.**

I. Introdução

A prática regular de atividades físicas desempenha um papel essencial na promoção da saúde física e mental, contribuindo para a prevenção de doenças crônicas como diabetes, hipertensão e obesidade. Em todas as fases da vida, os benefícios da atividade física são amplamente reconhecidos, tornando-se ainda mais críticos com o avanço da idade [1]. Nesse contexto, as academias de ginástica desempenham um papel central ao oferecer um ambiente estruturado para a prática de exercícios físicos. Esses espaços contam com orientação profissional, auxiliando os clientes na busca por seus objetivos de saúde e condicionamento físico [2]. De acordo com dados da Associação Brasileira de Academias (ACAD) [3], aproximadamente 10 milhões de brasileiros frequentaram academias em 2019, consolidando o Brasil como um dos maiores mercados *fitness* do mundo.

No entanto, as academias enfrentam um desafio recorrente: a evasão de clientes. Esse fenômeno pode ser motivado por diversos fatores, incluindo a falta de resultados rápidos, desmotivação, lesões, monotonia dos treinos ou dificuldades de conciliar a rotina com a prática de exercícios [4]. A retenção de clientes, portanto, torna-se um fator estratégico para a sustentabilidade financeira das academias, visto que manter clientes atuais tende a ser mais vantajoso e econômico do que conquistar novos [5] [6].

Diante desse cenário, a aplicação de técnicas avançadas de análise de dados e aprendizado de máquina surge como uma abordagem promissora para compreender padrões de comportamento e prever a evasão de clientes. A capacidade de antecipar quais membros apresentam maior propensão ao abandono pode permitir a implementação de estratégias personalizadas de retenção, promovendo maior engajamento e reduzindo os índices de abandono.

Embora técnicas de aprendizado de máquina já sejam amplamente utilizadas na previsão de evasão em setores como telecomunicações e serviços financeiros, sua aplicação no segmento de academias ainda é escassa [6]. Sendo assim, este estudo visa preencher essa lacuna, explorando a eficácia de diferentes combinações de algoritmos de classificação e métodos de seleção de atributos na predição de evasão. Com isso, busca-se fornecer subsídios práticos para a tomada de decisão gerencial, contribuindo para a implementação de estratégias de retenção baseadas em dados.

II. Trabalhos Relacionados

O uso de algoritmos de aprendizado de máquina na previsão de evasão de clientes é amplamente documentado na literatura, com a maioria dos estudos concentrando-se em setores como telecomunicações e finanças. Tekouabou et al. [6] exploraram o uso de técnicas de aprendizado de máquina para prever esse fenômeno no setor bancário. O estudo destacou o algoritmo Random Forest (RF) por sua robustez e menor suscetibilidade a sobreajuste. Além disso, foram analisadas variáveis específicas, identificando a idade como fortemente associada a evasão, enquanto a posse de cartão de crédito apresentou baixa relevância preditiva.

Na área de telecomunicações, Ullah et al. [7] utilizaram uma abordagem híbrida, combinando modelos preditivos com técnicas de *clustering*, como o *k-means*, para segmentar clientes conforme o risco de evasão. Os resultados apontaram o RF como o modelo com maior acurácia, apesar de seu tempo de treinamento mais elevado, o que o torna especialmente útil para análises direcionadas a grupos de alto risco e para a formulação de estratégias de retenção personalizadas.

Estudos recentes reforçam a eficácia de algoritmos como RF e Gradient Boosting na previsão de evasão, conforme

demonstrado por Wassouf et al. [8], Prabadevi, Shalini e Kavitha [5]. Wassouf et al. [8] concentraram-se no setor de telecomunicações, aplicando técnicas de análise de grandes volumes de dados para identificar padrões comportamentais e históricos dos clientes, possibilitando uma previsão de evasão personalizada para a empresa Syriatel. Em um contexto mais amplo, Prabadevi, Shalini e Kavitha [5] utilizaram uma base de dados genérica do Kaggle, aplicando algoritmos de aprendizado de máquina para prever evasão em múltiplos setores, o que evidenciou a versatilidade dessas técnicas em diferentes contextos organizacionais.

Ainda no setor de telecomunicações, A.Idris, M.Rizwan A.Khan [9] exploraram a combinação de técnicas de subamostragem e seleção de atributos aplicadas aos classificadores RF e K-Nearest Neighbors (KNN), avaliando o impacto dessas abordagens por meio de métricas como sensibilidade, especificidade e AUC (do inglês, Area Under the Curve). O estudo destacou que a estratégia Chr-PmRF, que integra Otimização por Enxame de Partículas (PSO), Minimum Redundancy Maximum Relevance (mRMR) e RF, como a mais eficaz, consolidando-se como uma alternativa promissora para lidar com desafios computacionais e grandes volumes de dados na previsão de evasão.

No setor de academias, a pesquisa de Jas Semrl e Alexandru Matei [10] comparou o desempenho das plataformas AzureML e BigML para previsão de evasão, com o intuito de avaliar se essas ferramentas, operadas por profissionais sem especialização em ciência de dados, poderiam produzir soluções práticas. Esse estudo propôs duas tarefas de previsão: uma classificação binária para prever o cancelamento do contrato com base na frequência de uso da academia no primeiro mês; e uma classificação de múltiplas classes para estimar o nível de utilização na quinta semana. Os resultados indicaram que o BigML simplifica o treinamento e facilita a comparação entre modelos, com o algoritmo Decision Forest alcançando um AUC de 0,748, uma taxa de falsos positivos de 38,27% e uma acurácia de 66,8% na primeira tarefa. Na segunda tarefa, a acurácia foi de 76%.

Por sua vez, o AzureML destacou-se pela flexibilidade, oferecendo uma variedade de algoritmos, como Regressão Logística, Decision Jungle, Redes Neurais e Decision Forest. A Regressão Logística apresentou o melhor desempenho para a tarefa binária, com AUC de 0,734, enquanto a Rede Neural obteve a maior acurácia na tarefa de múltiplas classes, atingindo 74,6%. A variedade de opções oferecida pelo AzureML revela seu potencial para aplicações que demandam maior personalização dos modelos preditivos.

De maneira similar, Min Shan [11] também usou a plataforma AzureML para prever evasão em academias suecas, adotando uma metodologia ligeiramente distinta. O autor analisou os dados de comportamento dos clientes nos três primeiros meses de uso da academia para prever a evasão nos dois meses subsequentes. Dentre os modelos testados, o Boosted Decision Tree foi o mais eficaz, com AUC de 0,867, superando o desempenho obtido por Jas Semrl e Alexandru Matei [10]. Embora os resultados tenham sido promissores, o estudo ressaltou a importância do aprendizado contínuo e da atualização periódica dos modelos para preservar sua acurácia em ambientes reais e dinâmicos.

Finalmente, o estudo de Sobreiro et al. [12], realizado em academias de Lisboa, utilizou a plataforma Anaconda e a biblioteca Scikit-learn para implementar e avaliar modelos de aprendizado de máquina, incluindo Regressão Logística, Árvore de Decisão, RF e Gradient Boosting Classifier (GBC). Entre os algoritmos testados, o GBC apresentou o melhor desempenho geral, com acurácia de 95,5% e AUC de 0,873, enquanto o RF obteve o maior valor de AUC (0,890), sugerindo alta capacidade de discriminação entre clientes propensos à permanência e à evasão.

Apesar do crescente interesse em modelos de predição de evasão, os estudos existentes focam majoritariamente em setores como telecomunicações e serviços bancários, com aplicação limitada no contexto de academias. Além disso, muitos trabalhos utilizam um único algoritmo ou negligenciam etapas críticas como seleção de atributos e ajuste fino de hiperparâmetros, o que pode comprometer a generalização dos modelos. Poucos estudos avaliam de forma sistemática a influência de diferentes técnicas de seleção de atributos (como InfoGain, mRMR, PCA e Forward Selection) combinadas com múltiplos classificadores. Diante dessa lacuna, este trabalho propõe uma análise comparativa abrangente entre seis algoritmos supervisionados e quatro métodos de seleção de atributos, utilizando validação cruzada e otimização via Optuna, com o objetivo de identificar configurações eficientes e generalizáveis para a predição de evasão em academias.

III. Metodologia

Este estudo adotou uma abordagem estruturada em cinco etapas principais, conforme ilustrado na Figura 1. Inicialmente, foi realizada a obtenção de uma base de dados adequada, contendo informações relevantes sobre a evasão de clientes. Em seguida, foram explorados diferentes métodos de seleção de atributos, visando identificar as variáveis mais significativas para a modelagem preditiva. Na terceira etapa, foram analisados diversos modelos de aprendizado de máquina para determinar aqueles com maior potencial para prever a evasão com precisão. Posteriormente, o desenvolvimento do projeto foi detalhado, abordando as técnicas utilizadas para implementação dos modelos. Por fim, os modelos foram avaliados com base em métricas como acurácia, precisão, sensibilidade e especificidade, a fim de identificar a solução mais eficiente e robusta para prever a evasão de clientes em academias.

A. Coleta da Base de dados

A base de dados utilizada neste estudo foi obtida na plataforma Kaggle [13] e contém 4.000 registros relacionados à fidelidade e evasão de clientes em academias nos Estados Unidos (EUA). Essa base de dados foi escolhida por sua diversidade de informações relevantes para a previsão de evasão, incluindo 13 atributos que podem influenciar a decisão dos clientes de permanecer ou desistir da academia. Dentre essas

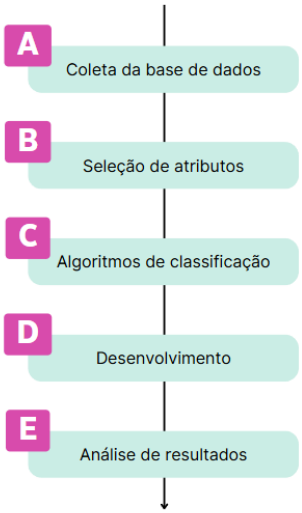

Figura 1. Metodologia aplicada

variáveis, destacam-se a frequência média de uso, o tempo de associação e a idade dos clientes, como fatores essenciais para a construção de modelos preditivos. A Tabela I apresenta uma descrição detalhada dessas variáveis, fornecendo uma visão abrangente dos dados utilizados no estudo.

Tabela I. Descrição dos atributos

Atributo	**Descrição**
Gender	Gênero
Near_Location	Se o usuário mora ou trabalha no bairro onde está localizada a academia
Partner	Se o utilizador trabalha numa empresa associada (a academia tem empresas associadas cujos funcionários recebem descontos; nesses casos, a academia armazena informação sobre os empregadores dos clientes)
Promo_friends	Se o usuário se inscreveu originalmente por meio de uma oferta "traga um amigo"(usou o código promocional de um amigo quando pagou a primeira assinatura)
Phone	Se o usuário forneceu seu número de telefone.
Contract_period	Período do contrato em meses
Group_visits	Visitas em grupo
Age	Idade
Avg_additional_charges _total	Total médio de cobranças adicionais
Month_to_end_contract	O mês em que o contrato acaba
Lifetime	Quantos meses está na academia
Avg_class_frequency_total	Total de frequência média das aulas
Avg_class_frequency _current_month	Frequência média das aulas no mês atual

B. Seleção de atributos

A seleção de atributos desempenha um papel essencial na construção de modelos de aprendizado de máquina, influenciando diretamente sua eficiência e precisão. A escolha de um subconjunto otimizado de variáveis permite que os algoritmos melhorem seu desempenho ao reduzir a complexidade do modelo, evitando sobreajuste e tornando a utilização de recursos computacionais mais eficiente. Além disso, a seleção adequada de atributos pode aumentar a interpretabilidade do modelo, facilitando a compreensão dos fatores que mais impactam a predição.

Este estudo utilizou quatro técnicas distintas de seleção de atributos, que adotam estratégias complementares para identificar os subconjuntos de variáveis com maior poder preditivo:

1) Principal Component Analysis (PCA): Reduz a dimensionalidade do conjunto de dados ao transformá-lo em componentes principais [9].
2) Seleção Progressiva (Forward Selection): Seleciona iterativamente os atributos que mais contribuem para o desempenho do modelo [8] [11].
3) Maximum Relevance and Minimum Redundancy (mRMR): Busca maximizar a relevância e minimizar redundâncias entre as variáveis selecionadas [9].
4) Infogain: Calcula o ganho de informação de cada atributo em relação à variável-alvo [14].

C. Algoritmos de classificação

O estudo empregou seis algoritmos de classificação supervisionada — árvore de decisão, SVM, regressão logística, rede neural, RF e XGBoost — selecionados por representarem diferentes abordagens de aprendizado. Cada modelo teve seus principais hiperparâmetros ajustados para melhora do desempenho:

1) Árvore de Decisão: Organiza a classificação dos dados de forma hierárquica, realizando divisões sucessivas com base nos atributos [15]. Neste estudo, foi usado o critério de Gini e ajustes como profundidade máxima e amostras mínimas por nó/folha.
2) SVM: Define uma fronteira de decisão ideal para separar classes, maximizando a margem entre elas [16] [15]. O trabalho otimizou o fator de regularização, o tipo de kernel e o parâmetro de influência das amostras.
3) Regressão Logística: Utiliza a função sigmoide para calcular probabilidades de pertencimento a uma classe [15]. Neste estudo, foi avaliado o fator de regularização, o tipo de penalidade e o otimizador (com máximo de 100 iterações).
4) Rede Neural: Inspirados no funcionamento do cérebro humano, esse modelo ajusta os pesos das conexões entre neurônios para minimizar erros e melhorar a capacidade de aprendizado [15]. Nesta pesquisa, foram analisados o número de camadas, função de ativação e regularização; utilizou o otimizador ADAM com taxa de aprendizado adaptativa.
5) O Random Forest: Combina múltiplas árvores de decisão para melhorar a precisão [17]. Foram adotados os mesmos critérios da árvore simples. O número de árvores (estimadores) utilizadas foi de 100.
6) XGBoost: Assim como o RF utiliza múltiplas árvores de decisão, porém aprimora o conceito com aprendizado sequencial e técnicas de regularização [18]. Também foram avaliados no ajuste a profundidade máxima da

árvore e o número mínimo de amostras necessárias. O número de estimadores foi mantido em 100.

D. Desenvolvimento computacional

Os experimentos foram conduzidos em ambiente Python com as bibliotecas Scikit-learn, Pandas, NumPy, Matplotlib e Optuna. A biblioteca Scikit-learn foi fundamental para a implementação dos algoritmos de aprendizado de máquina, enquanto as demais bibliotecas auxiliaram no tratamento e visualização dos dados. A infraestrutura foi composta por um computador com processador *multicore* e, suporte a GPU, para acelerar o treinamento das redes neurais.

O conjunto de dados foi dividido em 80% para treinamento e 20% para teste. No conjunto de treinamento, aplicou-se validação cruzada com 5 *folds* (divisão em partes iguais dos dados) para avaliar os modelos e ajustar os hiperparâmetros por meio do Optuna, uma ferramenta de otimização automatizada que utiliza técnicas de busca eficientes, para encontrar as configurações ideais de forma automatizada. Foram testadas diferentes combinações de atributos para cada técnica de seleção, e os melhores modelos foram posteriormente reavaliados no conjunto de teste.

Após a escolha dos melhores atributos e configurações de um modelo, foi usado o conjunto de dados de teste para avaliar os modelos por meio das métricas: acurácia, precisão, revocação, F1-score e AUC. Essas métricas permitem comparar o desempenho geral dos classificadores e identificar a melhor abordagem preditiva para o problema de evasão em academias.

IV. Análise dos Resultados

Os resultados deste estudo foram discutidos a partir de quatro eixos principais: (i) avaliação do desempenho dos métodos de seleção de atributos, considerando sua capacidade de reduzir a dimensionalidade sem comprometer a acurácia preditiva; (ii) análise dos melhores hiperparâmetros encontrados pelo Optuna; (iii) identificação das variáveis mais relevantes para a previsão de evasão; e (iv) comparação entre os modelos de aprendizado de máquina quanto à performance geral.

A. Desempenho dos métodos de seleção de atributos

A Figura 2 apresenta os resultados obtidos com o método de classificação Árvore de Decisão, considerando diferentes métodos de seleção de atributos. Observa-se que o Forward Selection e o mRMR apresentaram um desempenho similar até a inclusão de cinco atributos. A partir desse ponto, o mRMR teve uma ligeira queda de desempenho entre cinco e oito variáveis, mas posteriormente recupera-se, atingindo novamente o nível do Forward Selection. Ambos estabilizaram seu desempenho a partir da oitava variável, sugerindo um platô na capacidade de ganho preditivo com a adição de novos atributos.

O InfoGain destacou-se por apresentar uma rápida ascensão no desempenho com apenas três variáveis, atingindo estabilidade precoce. Esse comportamento evidencia sua eficiência em identificar atributos altamente relevantes com um conjunto mínimo de variáveis, o que é vantajoso em cenários com restrições computacionais ou necessidade de interpretabilidade.

Por outro lado, o PCA apresentou desempenho inferior na maior parte das configurações, especialmente com poucos componentes. Isso se deve à sua natureza de técnica não supervisionada, que não considera a variável alvo na transformação dos dados, podendo resultar em perda de informação discriminativa relevante. Contudo, observou-se uma melhora significativa no desempenho do PCA a partir de 12 componentes principais — ou seja, quando quase todas as variáveis originais foram incluídas na representação.

Esse comportamento é coerente com a lógica do PCA, que redistribui a variância total dos dados de forma linear entre os componentes. Em modelos baseados em particionamento, como as Árvores de Decisão, os primeiros componentes nem sempre são os mais informativos para a tarefa de classificação, o que pode explicar o desempenho inferior com poucos componentes. O aumento expressivo da acurácia com 12 componentes sugere que a informação preditiva está dispersa em múltiplas direções no espaço dos dados, exigindo a retenção de praticamente toda a variância original.

Por fim, vale ressaltar que, ao utilizar um número de componentes igual ao total de atributos, o PCA realiza apenas uma rotação da base vetorial no espaço original, sem perda de informação. No entanto, o fato de o melhor desempenho ter sido alcançado com 12 (e não 13) componentes pode indicar que o uso da última componente — associada a uma variância residual muito baixa — introduz ruído ou padrões espúrios, prejudicando ligeiramente a capacidade de generalização do modelo, especialmente em contextos de validação cruzada. Esse efeito é mais pronunciado em classificadores sensíveis à variação estatística, como é o caso das Árvores de Decisão.

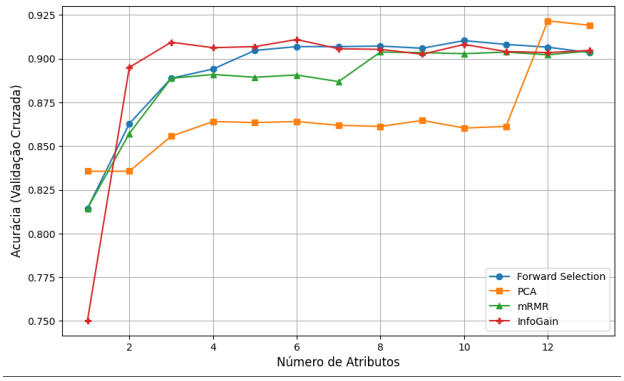

Figura 2. Desempenho do treino do classificador árvore decisão utilizando diferentes métodos de seletores de atributos

A Figura 3 apresenta os resultados obtidos com o classificador SVM. Observa-se novamente o desempenho superior do InfoGain, que alcança uma acurácia acima de 92% com apenas três atributos evidenciando sua capacidade de selecionar características discriminativas de forma eficiente. Os métodos Forward Selection e mRMR mostraram trajetória de desempenho semelhante até a inclusão da quarta variável,

com leve vantagem para o primeiro. Contudo, o mRMR foi o único a atingir a melhor acurácia global com oito atributos, demonstrando sua efetividade em capturar variáveis relevantes mesmo em conjuntos maiores.

Por outro lado, o PCA apresentou desempenho consistentemente inferior em comparação aos demais métodos. Seu crescimento linear e gradual indica baixa eficácia da transformação neste contexto. Isso se deve ao fato de que, por ser uma técnica de redução de dimensionalidade não supervisionada, o PCA não considera a variável alvo, o que pode resultar em componentes principais com baixa relevância preditiva para o modelo SVM.

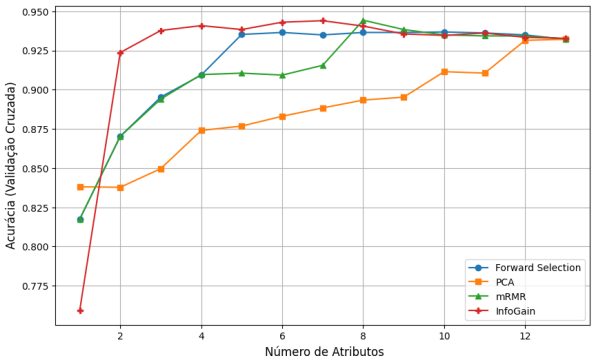

Figura 3. Desempenho do treino do classificador SVM utilizando diferentes métodos de seletores de atributos

O treinamento utilizando o modelo de Regressão Logística, conforme pode ser visto na Figura 4, revela comportamento semelhante ao observado no SVM. O InfoGain mais uma vez atingiu rapidamente a estabilização do desempenho, sugerindo que poucas variáveis concentram a maior parte da informação preditiva. Os métodos mRMR e Forward Selection apresentaram desempenho comparável, embora tenham selecionado subconjuntos distintos entre cinco e sete atributos. O PCA, novamente, mostrou-se menos eficiente neste cenário, reforçando sua limitação em modelos lineares.

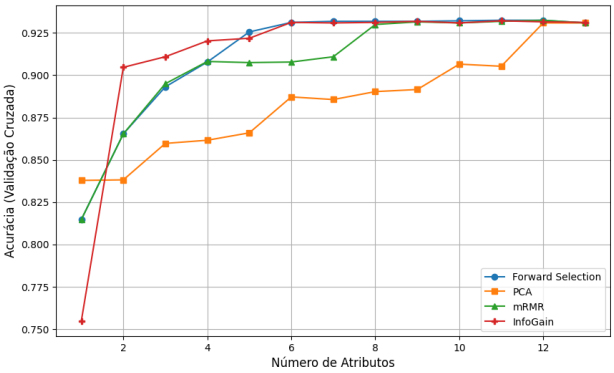

Figura 4. Desempenho do treino do classificador regressão logística utilizando diferentes métodos de seletores de atributos

No caso do RF (Figura 5), o comportamento foi distinto. O PCA apresentou desempenho competitivo, superando os demais métodos em determinadas configurações. Esse resultado, como já dito anteriormente, sugere que modelos baseados em árvores são mais tolerantes à transformação dos dados, sendo capazes de extrair padrões relevantes mesmo quando os atributos são linearmente combinados, como ocorre no PCA. Isso ocorre porque algoritmos como RF baseiam-se em partições sucessivas dos dados e são menos sensíveis à correlação entre variáveis.

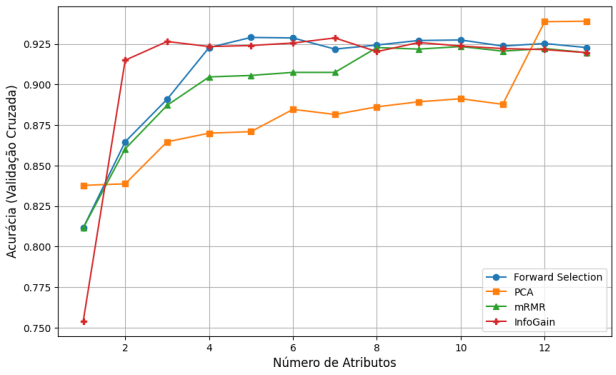

Figura 5. Desempenho do treino do classificador RF utilizando diferentes métodos de seletores de atributos

A Figura 6 apresenta os resultados obtidos com o treinamento da Rede Neural, os quais os padrões anteriores se mantiveram. Embora o PCA tenha demonstrado uma leve melhoria ao final do processo, não foi suficiente para superar os demais métodos. O melhor desempenho foi registrado com o Forward Selection, utilizando seis variáveis, atingindo uma acurácia de 96,09

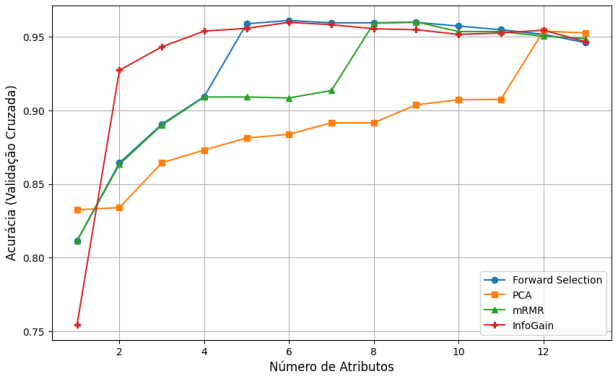

Figura 6. Desempenho do treino do classificador rede neural utilizando diferentes métodos de seletores de atributos

Por fim, na Figura 7, o XGBoost confirmou a consistência dos padrões observados nos modelos baseados em árvores. O PCA obteve seu melhor desempenho com doze componentes, alcançando acurácia de 95,03%, indicando novamente que classificadores do tipo *ensemble* toleram bem a transformação linear dos dados quando a variância é suficientemente preservada.

De maneira geral, o InfoGain destacou-se como um método altamente eficiente, atingindo excelente desempenho com um número reduzido de variáveis. Essa característica o torna ideal

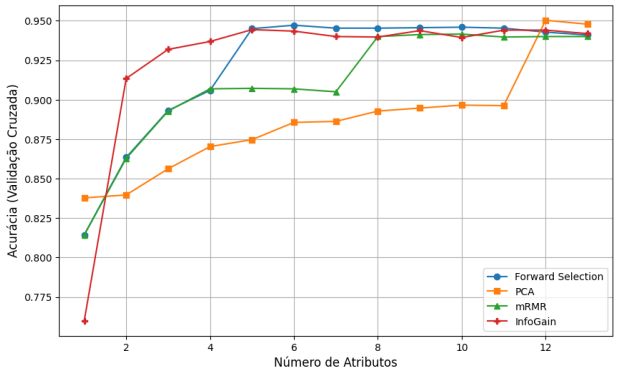

Figura 7. Desempenho do treino do classificador XGBoost utilizando diferentes métodos de seletores de atributos

para cenários em que simplicidade, rapidez de processamento e interpretabilidade são prioridades. O mRMR e o Forward Selection também se mostraram eficazes na construção de subconjuntos de atributos relevantes, contribuindo de forma significativa para a performance dos modelos, embora com maior custo computacional.

O PCA, por sua vez, apresentou desempenho inferior na maioria dos modelos, mas mostrou-se vantajoso em classificadores baseados em árvores, como o RF e o XGBoost. Esses algoritmos são capazes de lidar com atributos correlacionados ou transformados, o que potencializa a utilidade do PCA nesses casos.

Portanto, a escolha do método de seleção de atributos deve considerar o equilíbrio entre desempenho preditivo e custo computacional, bem como as características do algoritmo de classificação empregado. O mRMR e o Forward Selection tendem a oferecer bons resultados, especialmente quando ajustados ao modelo de aprendizado de máquina utilizado. O PCA, por sua vez, pode ser vantajoso em cenários com modelos baseados em árvores. No entanto, o InfoGain foi o único método que apresentou um desempenho robusto e consistente em diferentes modelos e configurações, consolidando-se como uma ferramenta versátil e confiável para seleção de atributos em tarefas de predição de evasão.

B. Ajuste dos hiperparâmetros

Durante o processo de treinamento, foram avaliadas diferentes combinações de hiperparâmetros com o auxílio da biblioteca Optuna. A Tabela II resume os melhores conjuntos de hiperparâmetros identificados para cada modelo, incluindo o método de seleção de atributos correspondente, aplicados à fase final de testes.

Para o classificador de Árvore de Decisão, o melhor desempenho foi obtido com atributos transformados via PCA, utilizando 12 componentes principais. A configuração ideal incluiu profundidade máxima da árvore igual a 7, no mínimo 11 amostras por divisão e pelo menos 9 amostras por folha terminal.

No caso do SVM, o método de seleção mais eficaz foi o mRMR, com oito variáveis selecionadas. O modelo alcançou melhor desempenho com penalidade $C = 91{,}86$, função de núcleo (kernel) polinomial e parâmetro Γ ajustado automaticamente (scale), indicando preferência por uma separação não linear entre as classes.

A Regressão Logística obteve sua melhor performance com o método Forward Selection, selecionando 11 atributos. A configuração ideal envolveu penalidade do tipo L2, valor de C=3,55, e o solucionador lbfgs, amplamente utilizado para otimização de funções convexas.

Para o modelo RF, o PCA com 13 componentes foi a melhor estratégia de pré-processamento. Os hiperparâmetros ideais incluíram profundidade máxima de 18 níveis, divisão mínima de 8 amostras e mínimo de 1 amostra por folha, o que indica uma configuração profunda com alto poder de ajuste.

A Rede Neural apresentou seu melhor desempenho com variáveis selecionadas via mRMR (9 atributos). A arquitetura otimizada foi composta por três camadas ocultas com 11, 38 e 20 neurônios, respectivamente, utilizando a função de ativação tanh e fator de regularização $\alpha = 0{,}0001$, indicando uma rede moderadamente complexa com controle eficiente de sobreajuste.

Por fim, o XGBoost também teve desempenho otimizado quando combinado com PCA (12 componentes). A melhor configuração incluiu profundidade máxima de 2, peso mínimo por folha igual a 2 e taxa de aprendizado de 0,21, refletindo um modelo mais raso, porém com atualizações mais graduais e estáveis.

Tabela II. Hiperparâmetros ajustados por modelo e método de seleção de atributos

Modelo	**Seletor de Atributos**	**Hiperparâmetros ajustados**
Árvore de Decisão	PCA (12 componentes)	Profundidade máxima = 7; mínimo de amostras para divisão = 11; mínimo de amostras por folha = 9
SVM	mRMR (8 atributos)	C = 91,86; kernel = polinomial; gamma = *scale*
Regressão Logística	Forward Selection (11 atributos)	C = 3,55; penalidade = L2; solucionador = *lbfgs*
Random Forest	PCA (13 componentes)	Profundidade máxima = 18; mínimo de amostras para divisão = 8; mínimo de amostras por folha = 1
Rede Neural	mRMR (9 atributos)	Arquitetura: (11, 38, 20); função de ativação = *tanh*; regularização $\alpha = 0{,}0001$
XGBoost	PCA (12 componentes)	Profundidade máxima = 2; peso mínimo por folha = 2; taxa de aprendizado = 0,21

Esses resultados evidenciam o impacto substancial da combinação entre o método de seleção de atributos e o ajuste de hiperparâmetros sobre o desempenho preditivo dos modelos. O ajuste fino dos parâmetros permitiu adaptações específicas às características de cada algoritmo, maximizando sua capacidade de generalização e tornando os modelos mais robustos para a tarefa de previsão de evasão.

C. Análise das características mais relevantes

Os métodos de seleção de atributos apresentaram forte convergência quanto às variáveis mais relevantes, indepen-

dentemente do modelo de aprendizado empregado. Essa consistência indica que as informações preditivas estão concentradas em um subconjunto específico de atributos.

Os métodos mRMR e Forward Selection atribuíram maior importância ao tempo de associação (*lifetime*), à idade e à duração do contrato. O InfoGain, por sua vez, priorizou variáveis relacionadas à frequência de uso, como a média de aulas frequentadas no mês corrente e a frequência média total. Embora esses atributos também tenham sido selecionados posteriormente pelos demais métodos, suas posições na hierarquia de importância diferiram, refletindo diferentes critérios de avaliação entre os seletores.

O desempenho superior do InfoGain, mesmo com um número reduzido de atributos, sugere que a frequência de participação é o fator mais determinante para prever a evasão, seguida pela idade e tempo de associação. Esse resultado reforça a hipótese de que a assiduidade reflete diretamente o nível de engajamento do cliente com o serviço, sendo, portanto, um indicador-chave de fidelização.

Por outro lado, atributos como gênero, cadastro telefônico, vínculo com empresas conveniadas e inscrição por meio de promoções apresentaram baixa relevância preditiva. A variável de localização também teve influência limitada, possivelmente devido à sua representação binária ("reside ou não no bairro da academia"), o que restringe sua expressividade. Para análises futuras, recomenda-se utilizar representações mais granulares, como a distância em quilômetros até a academia, a fim de reavaliar sua relevância.

Como estratégia complementar de interpretação, foi empregada a biblioteca SHapley Additive Explanations (SHAP), que permite mensurar o impacto individual de cada variável na predição. A Figura 8 exibe um gráfico de importância das variáveis, no qual o eixo X representa a contribuição de cada atributo na saída do modelo. Valores mais à direita indicam aumento na probabilidade de evasão, enquanto os valores à esquerda sinalizam maior chance de permanência do cliente. A coloração do gráfico indica o valor da variável (vermelho para valores altos, azul para baixos). Por exemplo, a variável Avg class frequency current month, quando apresenta valores elevados, está associada à redução da probabilidade de evasão, corroborando a relevância da frequência de uso como preditor.

A Figura 9 mostra um gráfico tipo *waterfall*, que detalha a decomposição da predição de uma instância individual (um cliente). A previsão parte de uma média base de 0,22, sendo ajustada incrementalmente pelas contribuições de cada variável até atingir um valor final próximo de zero, o que indica baixa probabilidade de evasão para o caso analisado.

Essas análises evidenciam que o modelo é fortemente influenciado por variáveis comportamentais, como hábitos de frequência, tempo de relacionamento e tipo de contrato — características também destacadas pelos métodos de seleção de atributos. Essas variáveis devem ser priorizadas em estratégias de retenção e personalização de serviços. Em contrapartida, atributos com baixa relevância, como gênero ou localização aproximada, podem ser descartados ou simplificados conforme os objetivos de implementação prática

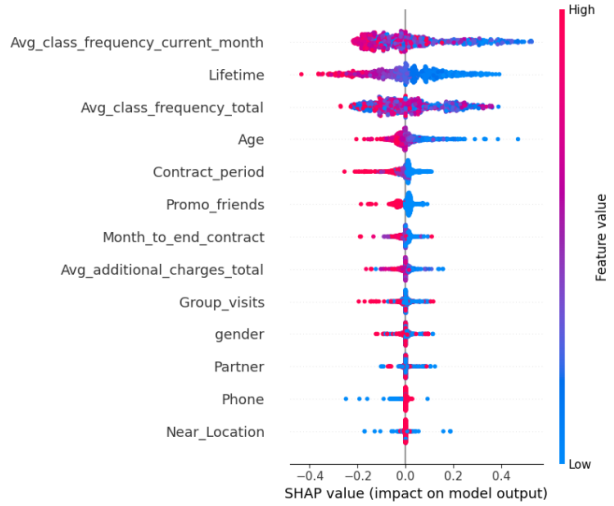

Figura 8. Importância das características na previsão de evasão

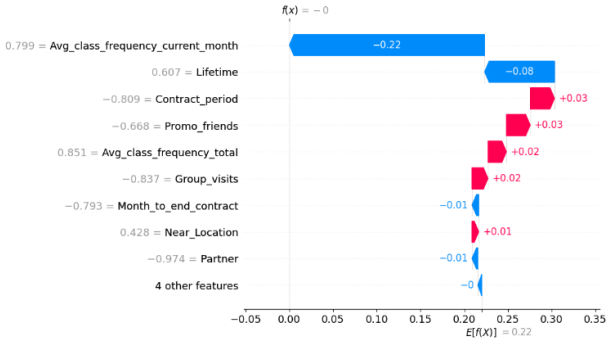

Figura 9. Impacto das características na previsão de evasão de um único cliente

D. Desempenho dos modelos de aprendizado

A Tabela III resume os principais resultados obtidos pelos modelos avaliados. Os classificadores baseados em Redes Neurais, o RF e o XGBoost apresentaram os melhores desempenhos, todos com acurácia de 95%. O XGBoost destacou-se nas métricas de precisão (96%) e F1-Score (91%), empatado com a Rede Neural, que, por sua vez, obteve o maior revocação (89%) e a maior área sob a curva ROC (AUC = 0,931), evidenciando sua superioridade na detecção de clientes com alta probabilidade de evasão.

O RF apresentou desempenho intermediário, com métricas equilibradas (F1-Score de 90% e AUC de 0,919), sendo uma alternativa robusta e de menor complexidade computacional em relação à rede neural.

Os modelos SVM e Árvore de Decisão também apresentaram desempenhos competitivos, ambos com acurácia de 93% e F1-score de 87%. No entanto, a Regressão Logística foi o modelo com o pior desempenho geral, com acurácia de 92% e os menores valores em todas as métricas analisadas, refletindo limitações em capturar padrões não lineares.

Tabela III. Resultados dos Modelos

Modelo	Acurácia	Precisão	*revocação*	F1-Score	AUC
Rede Neural	**95%**	93%	**89%**	**91%**	**0.931**
XGBoost	**95%**	**96%**	86%	**91%**	0.925
Random Forest	**95%**	94	86%	90%	0.919
Árvore de Decisão	93%	88%	86%	87%	0,911
SVM	93%	91%	83%	87%	0.902
Regressão Logística	92%	88%	81%	84%	0.884

V. Conclusão

Os métodos de seleção de atributos mostraram-se eficazes para identificar variáveis importantes na evasão de clientes em academias, com destaque para o InfoGain, que obteve alta acurácia usando poucos atributos, sendo ideal para contextos que exigem eficiência computacional e interpretabilidade.

Os métodos Forward Selection e mRMR também apresentaram desempenho competitivo, sendo eficazes na identificação de subconjuntos com alta capacidade discriminativa. Contudo, ambos exigem maior tempo de processamento devido ao seu caráter iterativo ou à necessidade de cálculos envolvendo múltiplas interdependências entre atributos. O PCA, por sua vez, obteve desempenho inferior em modelos lineares, como SVM e Regressão Logística, mas demonstrou ser uma alternativa viável quando aplicado a algoritmos baseados em árvores de decisão, como o RF e o XGBoost — modelos menos sensíveis à transformação linear dos dados.

As análises identificaram como principais preditores da evasão a frequência de uso, tempo de associação, idade e duração do contrato, enquanto variáveis como gênero, uso de promoções, vínculo com empresas e localização tiveram baixa relevância, podendo ser descartadas para simplificar os modelos sem perda de desempenho.

Entre os algoritmos de aprendizado de máquina testados, três modelos — Rede Neural, XGBoost e RF — atingiram a mesma acurácia de 95%. No entanto, a Rede Neural obteve os melhores resultados em métricas críticas: maior revocação (89%), maior área sob a curva ROC (AUC = 0,931) e F1-score elevado (91%). Isso evidencia sua superioridade na detecção de casos positivos de evasão, o que a torna a opção mais robusta e eficaz entre os modelos avaliados.

O estudo evidencia que técnicas de aprendizado de máquina e seleção de atributos são ferramentas úteis na prevenção da evasão em academias, permitindo o uso de modelos preditivos para apoiar estratégias de retenção mais eficazes e baseadas em dados.

Como direções para pesquisas futuras, propõe-se: i) a utilização de variáveis geográficas mais expressivas, como a distância em quilômetros entre o cliente e a academia, visando aprimorar a representação espacial e sua influência no comportamento de evasão; ii) a generalização da abordagem para outros setores baseados em prestação contínua de serviços, como instituições de ensino presenciais e plataformas de educação *online*; iii) a validação externa dos modelos em bases de dados maiores, heterogêneas e provenientes de diferentes regiões ou realidades operacionais, a fim de ampliar a robustez e a aplicabilidade dos resultados obtidos.

Agradecimentos

Os autores agradecem ao Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) pelo apoio institucional e pelas condições oferecidas para o desenvolvimento desta pesquisa.

Referências

[1] S. M. M. Matsudo, "Envelhecimento, atividade física e saúde," *Boletim do Instituto de Saúde-BIS*, no. 47, pp. 76–79, 2009.

[2] J. J. de Oliveira Toscano, "Academia de ginástica: um serviço de saúde latente," *Revista brasileira de ciência e Movimento*, vol. 9, no. 1, pp. 40–42, 2001.

[3] ACAD Brasil, "Revista ACAD Brasil - Edição 82," 2019, 16 out. 2024. [Online]. Available: https://www.acadbrasil.com.br/wp-content/uploads/2019/03/edicao-82.pdf

[4] C. M. d. Liz and A. Andrade, "Análise qualitativa dos motivos de adesão e desistência da musculação em academias," *Revista Brasileira de Ciências do Esporte*, vol. 38, pp. 267–274, 2016.

[5] B. Prabadevi, R. Shalini, and B. R. Kavitha, "Customer churning analysis using machine learning algorithms," *International Journal of Intelligent Networks*, vol. 4, pp. 145–154, 2023.

[6] S. C. Tékouabou, S. C. Gherghina, H. Toulni, P. N. Mata, and J. M. Martins, "Towards explainable machine learning for bank churn prediction using data balancing and ensemble-based methods," *Mathematics*, vol. 10, no. 14, p. 2379, 2022.

[7] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector," *IEEE access*, vol. 7, pp. 60 134–60 149, 2019.

[8] W. N. Wassouf, R. Alkhatib, K. Salloum, and S. Balloul, "Predictive analytics using big data for increased customer loyalty: Syriatel telecom company case study," *Journal of Big Data*, vol. 7, no. 1, p. 29, 2020.

[9] A. Idris, M. Rizwan, and A. Khan, "Churn prediction in telecom using random forest and pso based data balancing in combination with various feature selection strategies," *Computers & Electrical Engineering*, vol. 38, no. 6, pp. 1808–1819, 2012.

[10] J. Semrl and A. Matei, "Churn prediction model for effective gym customer retention," in *Proc. Int. Conf. on Behavioral, Economic, Socio-Cultural Computing (BESC)*, Kraków, Poland, 2017, pp. 1–3.

[11] M. Shan, "Building customer churn prediction models in fitness industry with machine learning methods," 2017.

[12] P. Sobreiro, P. Guedes-Carvalho, A. Santos, P. Pinheiro, and C. Gonçalves, "Predicting fitness centre dropout," *International journal of environmental research and public health*, vol. 18, no. 19, p. 10465, 2021.

[13] Adrian Vinueza, "Gym customers features and churn," 2024, 20 out. 2024. [Online]. Available: https://www.kaggle.com/datasets/adrianvinueza/gym-customers-features-and-churn

[14] T. Pehlivanova and V. Nedeva, "Attributes selection using machine learning for analysing students' dropping out of university: a case study," in *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 1031, no. 1, 2021, p. 012055.

[15] M. Caetano, "Modelos de classificação: aplicações no setor bancário," Ph.D. dissertation, Ph.D. dissertation, Instituto de Computação, Universidade Estadual de Campinas, Campinas, Brasil, 2015.

[16] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2019.

[17] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[18] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, San Francisco, CA, USA, 2016, pp. 785–794.