LVI SBPO SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL
DE 04 ATÉ 07 NOVEMBRO 2024
FORTALEZA | CEARÁ

OTIMIZAÇÃO E ANÁLISE MULTICRITÉRIO EM MODELOS DE APRENDIZADO DE MÁQUINA BASEADOS EM ARVORES DE DECISÃO

Marcos Hendrio Pompeu da Silva
Pontifícia Universidade Católica do Paraná, PPGEPS
R. Imac. Conceição, 1155 - Prado Velho, Curitiba – PR
marcos.hendrio@pucpr.edu.br

Gilberto Reynoso-Meza
Pontifícia Universidade Católica do Paraná, PPGEPS
R. Imac. Conceição, 1155 - Prado Velho, Curitiba – PR
g.reynosomeza@pucpr.br

RESUMO

Os modelos de aprendizado de máquina emergiram como ferramentas essenciais no panorama das grandes corporações. Estes modelos atuam como instrumentos analíticos robustos, facilitando a tomada de decisões otimizadas e proporcionando uma vantagem competitiva através da antecipação de tendências e estratégias. No presente estudo, introduzimos uma metodologia para a avaliação de modelos de classificação fundamentados em árvores de decisão. Esta avaliação é conduzida através da otimização multiobjetivo, decisão multicritério e uma análise meticulosa dos paradigmas de design. Em nossa abordagem, três objetivos — precisão, recall e quantidade de variáveis —, que à primeira vista podem parecer antagônicos, são analisados de forma simutanea. Utilizaremos algoritmos de aprendizado de máquina com características estruturais diversificadas. O intuito é elucidar as variações comportamentais conforme variáveis adicionais são integradas ao sistema. Argumentamos que, em vez de se concentrar unicamente em métricas convencionais de maneira segmentada, é mais benéfico adotar uma metodologia que sintetize esses objetivos.

PALAVRAS CHAVE. Otimização multiobjetivo, aprendizado de máquina, análise multicritério.

Tópicos do artigo: ADM, EST&MP, MH

ABSTRACT

Machine learning models have emerged as essential tools in the landscape of large corporations. These models act as robust analytical instruments, facilitating optimized decision-making and providing a competitive advantage by anticipating trends and strategies. In this study, we introduce a methodology for evaluating classification models based on decision trees. This evaluation is conducted through multi-objective optimization, multi-criteria decision-making, and a meticulous analysis of design paradigms. In our approach, three objectives - precision, recall and number of variables - which at first glance may seem antagonistic, are analyzed simultaneously. We will use machine learning algorithms with diverse structural characteristics. The aim is to elucidate behavioral variations as additional variables are integrated into the system. We argue that rather than focusing solely on conventional metrics in a segmented way, it is more beneficial to adopt a methodology that synthesizes these objectives.

KEYWORDS. Multi-objective optimization, machine learning, multi-criteria analysis.

Article topics: ADM, EST&MP, MH

DOI: 10.59254/sbpo-2024-193472
https://proceedings.science/p/193472?lang=pt-br

1. Introdução

No meio acadêmico, diversas técnicas estatísticas foram empregadas por um longo período, resultando em avanços substanciais que beneficiam a sociedade. O aprendizado de máquina, em particular, tem sido um pilar na geração de conhecimento validado por pesquisas rigorosas e experimentações. Estas ferramentas são essenciais para a automatização de processos, aprimoramento de produtos e eficiência operacional nas empresas [Barton e Tribunal 2012]. Contudo, somente na última década, com o advento do crescimento exponencial de dados, avanços em ferramentas de processamento e armazenamento, e progressos na ciência de dados, é que as grandes corporações começaram a adotar e implementar tais práticas em escala ampla [Lee e Shin 2020]. Neste novo cenário, modelos preditivos permeiam várias áreas de negócios, tornando a informação uma commodity vital para a sobrevivência e prosperidade das empresas [Mcafee e Brynjolfsson 2012]. Essas estruturas de aprendizagem podem ser simples como os modelos de regressão linear [Fávero e Belfiore 2017] ou possuir uma maior complexidade nos casos aprendizagem profunda [LeCun, Bengio e Hinton 2015]. Independentemente da abordagem utilizada, a transparência e a interpretabilidade desses modelos são propostas de grande importância, pois permitem uma melhor compreensão do problema.

Embora existam diversas técnicas que visam contribuir para a transparência e interpretabilidade dos modelos, como a importância das características presentes nos modelos baseados em árvores [Kazemitabar, Amini, Bloniarz e Talwalkar 2017] e na técnica SHAP derivada da teoria dos jogos [Holzinger, Saranti, Molnar, Biecek e Samek 2022] persistem desafios relacionados à determinação do ponto ideal de complexidade, à seleção de critérios para tomada de decisão, modelos mais intuitivos para os usuários finais e métricas mais justas e aderentes [Valdivia, Monedero e Casillas 2020]. Esses fatores são cruciais, pois a compreensão dos resultados pelos desenvolvedores e stakeholders são essenciais para a adoção e implementação eficaz de soluções baseadas em aprendizado de máquina.

O crescimento do uso de ferramentas que utilizam dados como fontes de matéria prima para tomada de decisão tem crescido exponencialmente nas empresas que estão na vanguarda de seus seguimentos. Resolver problemas baseados em análises onde os dados são o centro de tudo nunca foi tão frequente no meio corporativo [Kratsch, Manderscheid, Röglinger e Seyfried 2020]. Por esse motivo, o interesse dos pesquisadores sobre o tema tem crescido nos últimos anos e com eles o surgimento de novas técnicas que mesclam áreas de diferentes seguimentos de atuação [Cockburn, Henderson e Stern 2019].

Durante o período de pesquisa para validar a viabilidade do tema, deparamo-nos com diversos artigos que abordam a sinergia entre otimização matemática, aprendizado de máquina e metaheurística [Alipour, Razavi, Derakhshi e Balafar 2018], [Chang 2017], (Cheng e Ying, Pourhejazy e Lin 2021]. Onde é mostrado que essas três técnicas podem trabalhar em harmonia e gerar resultados significativos. Outras abordagens que também podem atuar de forma conjunta são: decisão multicritério, aprendizado de máquina e otimização multiobjetivo [Valdivia, Monedero, Casillas 2020], [Li e Wu 2022], [Wang, Li e Rangaiah 2022] que em conjunto são capazes de gerar soluções simples ainda que esses problemas sejam de grande complexidade. Essa sinergia decorre do amadurecimento dessas técnicas, evidenciando que não existe uma solução única para todos os problemas. Os artigos anteriormente citados oferecem insights concisos sobre como essas metodologias podem atuar de forma integrada. Nos últimos anos, observamos o aumento significativo nas publicações que exploram esses temas, conforme documentado nos artigos [Al-tashi, Abdulkadir, Rais, Mirjalili e Alhussian 2020] e [Karimi-mamaghan, Mohammadi, Meyer e Talbi 2022]. No entanto, as pesquisas estão predominantemente focadas na resolução de problemas de características mono ou biobjetivo, geralmente relacionadas a duas ou no máximo três áreas distintas. Persistindo dessa forma uma lacuna de informações sobre como aplicar a otimização multiobjetivo e decisão multicritério em conjunto para resolver paradigmas tridimensionais ou de maior complexidade no contexto do aprendizado de máquina.

O aprendizado de máquina, em certos contextos, apresenta desafios de natureza multiobjetiva, seja devido à presença de métricas de desempenho conflitantes ou à simplificação

DOI: 10.59254/sbpo-2024-193472
https://proceedings.science/p/193472?lang=pt-br

da complexidade em detrimento da capacidade preditiva dos modelos. Nesse sentido, o objetivo geral deste estudo é: propor uma técnica multiobjetivo para analisar diferentes estruturas de aprendizagem baseadas em árvores de decisão com diferentes variáveis através de decisão multicritério.

Os resultados esperados incluem o desenvolvimento de uma ferramenta capaz não apenas de realizar a seleção de características e de modelos de forma mais aderente a cada tipo de problema, mas também evidenciar através de uma metodologia aspectos importantes para tomada de decisão que são frequentemente ignorados.

Este artigo está estruturado da seguinte forma: marco teórico (seção 2) detalha os procedimentos experimentais e ferramentas empregadas. Os métodos e ferramentas discutem estudos anteriores que fundamentam a relevância do tema. A seção 4 apresenta os resultados, discutindo os casos de uso e comparações. Conclusões com reflexões finais na seção 5.

2. Marco Teórico

2.1 Aprendizado Supervisionado

Aprendizado de máquina refere-se a algoritmos de computador que se aprimoram automaticamente por meio da experiencia com dados [Mitchell 1997]. É de conhecimento comum que existem três grandes áreas de aprendizagem de máquina: aprendizagem por reforço, onde um agente aprende a tomar ações em um ambiente para maximizar recompensas ao longo do tempo, comum em jogos e robótica; não supervisionado, onde os algoritmos encontram padrões em dados não rotulados, útil para clusterização e redução de dimensionalidade; supervisionado, onde os algoritmos aprendem a partir de dados rotulados para fazer previsões. O escopo desse trabalho está concentrado no aprendizado supervisionado, que utiliza como entrada I instancias (observações) com X variáveis (variáveis explicativas) para treinar uma estrutura de aprendizado usando informações confiáveis através da variável alvo T para cada uma das instancias. Para o caso particular da classificação binária o modelo irá treinar usando um conjunto de dados [I x X|T] a fim de ajustar seus parâmetros β na fase de otimização, usando algum critério de avaliação ou função de custo.

Existem diversas técnicas de aprendizagem supervisionada como por exemplo: regressão linear (Fávero e Belfiore 2017] redes neurais artificiais [Lecun, Bengio e Hinton 2015], regressão logística [Fávero e Belfiore 2017], knn [Guo, Wang, Bell, Bi e Greer 2003]. No entanto, essa pesquisa teve como premissa trabalhar com modelo de árvore de decisão que será detalhado a seguir.

2.2 Árvores de Decisão

O algoritmo de arvore de decisão é um dos mais conhecidos e mais usados modelos de aprendizado de máquina para problemas de regressão e classificação [Quinlan 1987]. Um dos grandes fatores que tornam esse modelo popular é a forma como ele atua nas etapas de treinamento, através de uma representação clara dos critérios escolhidos gerando uma visualização amigável em todo processo de aprendizagem. Na prática é possível ver como as principais variáveis se comportam em relação a variável alvo.

Normalmente uma árvore de decisão DT(β) é definido pelos seus parâmetros β = (n,v) ou seja, um conjunto de nós n e thresholds v. Dado um nó n na árvore de decisão DT(β) que corresponde a uma partição de instancias I_n C I, e a variável x∈X, definimos a partição do nó n em relação a variável x da seguinte forma:

$$I_(n,L)=\{ i\in \llbracket I_n |i\rrbracket _(x) \leq \nu_x\} \quad (1)$$

$$I_(n,R)=\{i\in \llbracket I_n |i\rrbracket _(x) > \nu_x\} \quad (2)$$

Onde, ν_x é o threshold da variável x para o nó n e i_x é o valor para variável x para instancias i. A partição do nó n em relação a variável x divide o conjunto de instancias i_n em dois subconjuntos. I_(n,L) e I_(n,R) que correspondem aos ramos esquerdo e direito do nó da arvore de decisão DT. Cada subconjunto I_(n,L) e I_(n,R) pode ser usado para construir sub árvores descendentes do nó n na árvore de decisão através do particionamento recursivo. Baseados na

DOI: 10.59254/sbpo-2024-193472
https://proceedings.science/p/193472?lang=pt-br

definição anterior, podemos destacar outras estruturas aprendizagem que também utilizam árvores de decisão, porém com características mais complexas conforme descritivo abaixo.

2.2.1 Florestas aleatórias

O modelo de floresta aleatória utiliza uma combinação de árvores, onde cada árvore treinada terá seu resultado avaliado com base nas combinações das amostras e variáveis escolhidas [Breiman 2001]. O fato dessa estrutura possuir diversas árvores torna essa técnica mais robusta, pois teremos mais opções de seleção de características tornado menos suscetível a super ajustes aos dados de treino (overfitting) permitindo uma maior capacidade de generalização. No entanto, é necessário compreender que devido a esse comportamento a técnica será mais custosa computacionalmente e a compreensão um pouco mais difícil.

2.2.2 XGBoost

O terceiro algoritmo é uma combinação entre um modelo baseado em árvore de decisão atuando em conjunto com técnica chamada de aumento de gradiente ou gradiente boosting [Chen e Guestrin 2016]. O aumento de gradiente tem a finalidade de combinar modelos com menor capacidade de predição atuando sequencialmente e conforme novos modelos são treinados os próximos irão se adaptar com base no histórico dos resultados se diferenciando do modelo de floresta aleatória que atua na construção de arvores de forma paralela. Atualmente o XGBoost é dos métodos ensemble mais utilizados pela comunidade devido a sua boa capacidade predição para dados tabulares graças a sua grande eficiência em minimizar as perdas.

2.2.3 LGBM

Assim como o XGBoost o LGBM é um modelo que utiliza aumento de gradiente em sua estrutura, mas apresenta características diferentes [Ke, Meng, Finley, Wang, Chen, Ma, Ye e Liu 2017]. O modelo da Microsoft apresenta maior eficiência nas etapas de construção das árvores reduzindo significativamente o custo computacional, no entanto em alguns casos pode apresentar uma menor capacidade de predição ao compararmos com o XGBoost.

2.3 Otimização multicritério

A seguir, serão descritos problemas relacionados a conceitos multiobjetivos, envolvendo a otimização de múltiplos objetivos, tomada de decisão e análise multicritério, os quais serão aplicados em um contexto específico de problema.

2.3.1 Problema multiobjetivo

Um problema multiobjetivo, ao contrário de um problema de objetivo único, é definido como aquele em que mais de uma variável é otimizada simultaneamente, ou seja, há a busca por um conjunto de soluções que representam o equilíbrio entre os diferentes objetivos, em vez de uma única solução ótima que atenda a todos os critérios de maneira ideal.

Um vetor multi-objetivo pode ser definido como:

$$J(\theta) = \llbracket [J] \rrbracket_1(\theta), \ldots, J_m(\theta) \;] \tag{3}$$

Onde $\theta = [\theta 1, \theta 2, \ldots, \theta n]$ é definido como o vetor de decisão com $\dim(\theta) = n$; $J(\theta)$ como o vetor objetivo. Percebeu-se que não existe uma solução única nos MOPs, pois geralmente não existe uma única solução que é melhor em todos os objetivos. Assim, define-se um conjunto de soluções, o conjunto de Pareto Θ_P. Cada solução no conjunto de Pareto define um vetor objetivo na frente de Pareto J_P. É importante notar que dependemos apenas da frente de Pareto e as aproximações dos conjuntos, $J_P\hat{}*$, $\Theta_P\hat{}*$. Todas as soluções na frente de Pareto são um conjunto de soluções de Pareto ótimas e não dominadas.

2.3.2 Otimização multiobjetivo e técnicas metaheurisiticas

A otimização multiobjetivo é um campo fundamental da pesquisa operacional que lida com a resolução de problemas complexos envolvendo a maximização ou minimização simultânea de duas ou mais funções objetivas conflitantes, enquanto se satisfazem um conjunto de restrições [Branke, Deb, Miettinen e Słowi´nski 2014]. A análise desses problemas leva à identificação da chamada "frente de pareto", que representa um conjunto de soluções não dominadas, ou seja, soluções que não podem ser melhoradas em um critério sem piorar em outro [Branke, Deb,

DOI: 10.59254/sbpo-2024-193472
https://proceedings.science/p/193472?lang=pt-br

Miettinen e Słowi´nski 2014]. A compreensão da frente de Pareto é crucial para a tomada de decisões informadas em ambientes onde múltiplos objetivos devem ser equilibrados.

As técnicas metaheurísticas têm emergido como ferramentas essenciais na resolução de problemas de otimização multiobjetivo [Branke, Deb, Miettinen e Słowi´nski 2014], destacando-se pela sua eficácia e versatilidade. Essas estratégias de alto nível fornecem um arcabouço para guiar outros métodos heurísticos na busca por soluções de alta qualidade dentro de um tempo de computação razoável. Entre as metaheurísticas mais amplamente adotadas estão os Algoritmos Genéticos [Jong 1988], otimização por enxame de Partículas [Poli, Kennedy e Blackwell 2007], algoritmos evolutivos [Branke, Deb, Miettinen e słowi´nski 2014]. Uma das principais vantagens dessas técnicas é a sua flexibilidade e adaptabilidade a uma ampla gama de problemas de otimização. Ao contrário de abordagens específicas para um problema, as metaheurísticas podem ser aplicadas de forma genérica, possibilitando sua utilização em diversos contextos e domínios.

2.3.3 Tomada de decisão e análise multicritério

A frente de pareto é uma forma de evidenciar os pontos não dominados, ou seja, as soluções em que pelo menos um dos objetivos apresenta melhor resultado quando comparado com as demais soluções. Esses conjuntos de soluções não dominadas possibilitam selecionar um ponto que mais faça sentido de acordo com os critérios estabelecidos em conjunto com análise multicritério [Reynoso-meza, Blasco, Sanchis e Martıneza 2014]. Existem várias técnicas de visualização para análise da frente pareto, como coordenadas paralelas [Reynoso-meza, Blasco, Sanchis, Herrero, 2013], diagramas de nível [Reynoso-meza, Blasco, Sanchis e Herrero 2013]. Em [Mattson e Messac 2005], algum refinamento é incorporado à noção de frente de pareto para diferentes conceitos de design.

3 Métodos e Ferramentas

3.1 Hardware

Todas as etapas de desenvolvimento desse projeto foram construídas utilizando os seguintes componentes: Notebook dell modelo g15, processador I5 (10500) com 2.5 Ghz, memória RAM 16GB, 1T de ssd, placa de vídeo nvidia 1650, placa mãe 0YJ00M com PCI-express 3.0.

3.2 Software

O trabalho foi desenvolvido usando os seguintes softwares: Sistema operacional Windows 11 versão 22H2, o ambiente de desenvolvimento VScode versão 1.86.2, linguagem python versão 3.9.12 com os respectivos pacotes: pandas [McKinney 2011], numpy [Walt, Colbert, Varoquaux, 2011], matplotlib [Hunter 2007], sklearn (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher e Duchesnay 2011], pymoo [Blank e Deb 2020].

3.3 Métodos e ferramentas

3.3.1 NSGA-III

O algoritmo NSGA-III (Non-dominated Sorting Genetic Algorithm III) é uma técnica de otimização multiobjetivo desenvolvida com base nos princípios dos algoritmos genéticos [Deb 2014]. Ele foi projetado para resolver problemas de otimização multiobjetivo e a sua principal característica está na capacidade de gerar uma população diversificada de soluções que representam uma boa cobertura da frente de Pareto.

3.3.2 Coordenadas paralelas

Coordenadas paralelas é uma técnica muito utilizada que traz uma abordagem bidimensional com sucessivos eixos verticais paralelos e um único eixo horizontal por ponto. Os eixos verticais retratam os objetivos conflitantes que tem como finalidade facilitar a busca pelo ponto ideal. Já o eixo horizontal mostra como as variáveis de decisão estão atuando sobre os objetivos [Reynoso-meza, Blasco, Sanchis e Herrero 2013]. A seguir será mostrado a comparação entre o gráfico de dispersão e o gráfico de Coordenadas paralelas.

Figura 1. Diagrama de dispersão duas dimensões.

DOI: 10.59254/sbpo-2024-193472
https://proceedings.science/p/193472?lang=pt-br

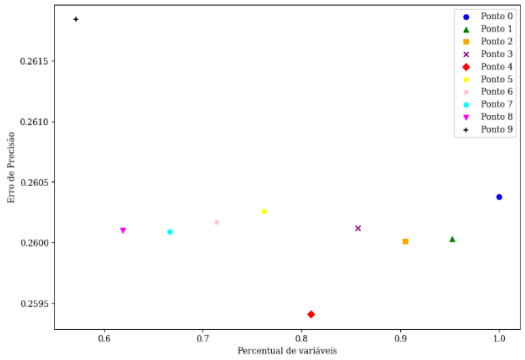
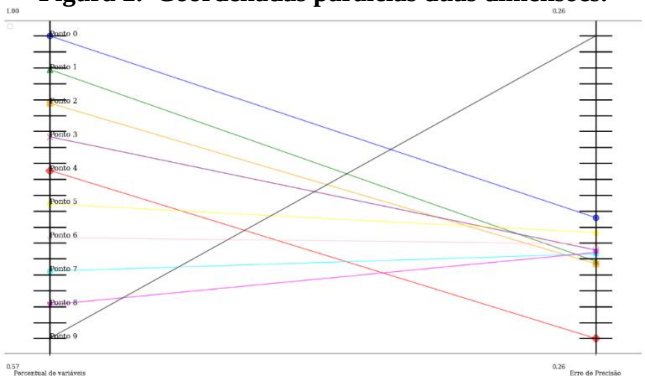
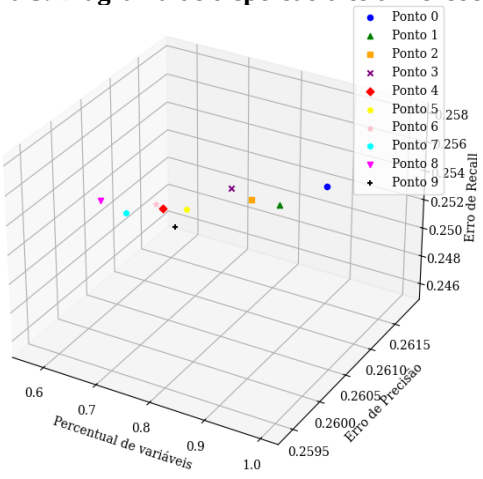


Figura 2. Coordenadas paralelas duas dimensões.



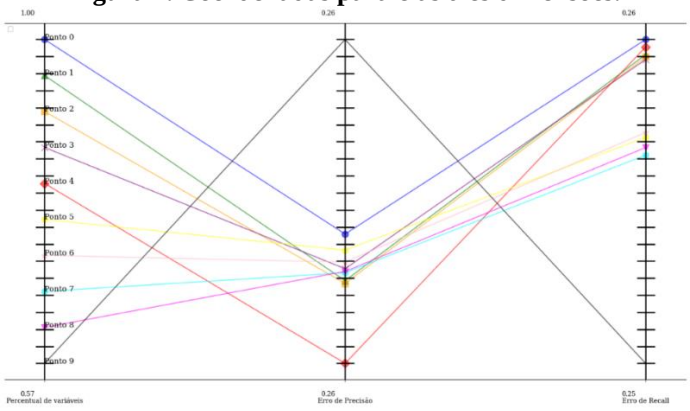
Na figura 1 temos a distribuição de dez pontos que possuem duas dimensões, percentual de variáveis no eixo x e o erro de precisão no eixo y. Cada ponto possui uma cor e forma específica assim como a figura 2. O diferencial está na forma em que os dados são evidenciados, o gráfico de coordenadas paralelas possui dois eixos verticais, um para cada objetivo, e os pontos são mostrados através de linhas horizontais. Para uma análise de comparação é possível notar que o ponto 9 possui o maior erro de precisão, porém utiliza a menor quantidade de variáveis percentualmente. Já ponto 4 possui o menor erro de precisão utilizando um valor percentual de aproximadamente oitenta por cento das variáveis disponíveis. Todas essas características podem ser observadas em ambas as figuras, porém na figura 2 é possível notar de forma mais clara a troca de trade-off dos pontos 4 e 9.

Figura 3. Diagrama de dispersão três dimensões.



DOI: 10.59254/sbpo-2024-193472
https://proceedings.science/p/193472?lang=pt-br

Figura 4. Coordenadas paralelas três dimensões.



Na figura 3 temos um gráfico de dispersão em três dimensões, com o percentual de variáveis no eixo x, o erro de precisão no eixo y e erro de recall no eixo z. E com isso aumentando a complexidade de abstração de como os pontos estão distribuídos. A partir de três dimensões a mente humana começa a ter dificuldades em interpretar e analisar resultados e por isso utilizar abordagens diferentes se torna essencial. A figura 4 mostra uma solução possível para esse problema, pois a utilização de coordenadas paralelas possibilita evidenciar por meio de um plano bidimensional de três ou mais dimensões melhorando a interpretabilidade dos resultados. Além desta técnica existem outras como o diagrama de nível que também utiliza uma abordagem bidimensional só que neste caso teremos um gráfico para cada objetivo e ele será representado no eixo horizontal [Reynoso-meza, Blasco, Sanchis e Herrero 2013].

3.3.3 Hiperparametros

Para garantir a reprodutividade dos resultados iremos utilizar o mesmo random state (parâmetro disponível nos modelos baseados em arvores que visam garantir replicação dos resultados). Além deste, outros hiperparâmetros deverão ser pré-definidos dentre eles os de profundidade da arvore (max_depth), quantidade mínima de instancias para divisão dos nós (min_samples_split), quantidade mínima de instancias para cada nó folha (min_samples_leaf) e para o critério de divisão dos nós, utilizaremos o índice de Gini [Tangirala 2022].

$$GINI(L) = 1 - \sum_{i=1}^{j} p_i^2 \tag{4}$$

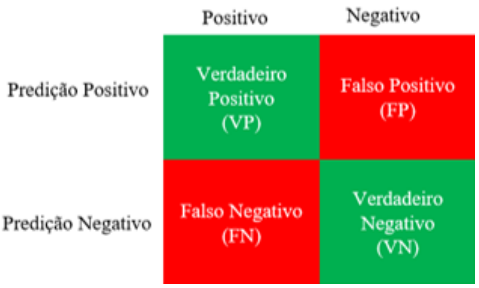
O índice de Gini é soma das probabilidades de as variáveis pertencerem a uma determina classe ponderada pelo número de registros ou instancias. O resultado será subtraído por menos um.

3.4 KPI - métricas

Para avaliar a performasse dos algoritmos de aprendizado de máquina foram escolhidos três métricas. A precisão avalia as afirmações do que realmente são verdadeiras e que foram apontadas pelo modelo, sobre o total dos valores verdadeiros. Já o recall avalia a capacidade do modelo de identificar os valores verdadeiros. Quanto mais valores positivos o modelo identifica, maior será o valor do recall. Além dessas duas métricas já conhecidas no universo do aprendizado de máquina, proporemos uma terceira forma de avaliação chamada de percentual de variáveis utilizadas PVU que será quantidade de variáveis usadas VU pelas estruturas de aprendizado sobre a quantidade total de variáveis disponíveis TV.

Figura 5. Matriz de confusão.

DOI: 10.59254/sbpo-2024-193472
https://proceedings.science/p/193472?lang=pt-br



$$PVU = \frac{VU}{TV} \quad (5)$$

$$Precisão = \frac{VP}{VP+FP} \quad (6)$$

$$Recall = \frac{VP}{TV+FN} \quad (7)$$

O intuito é trabalhar na minimização dos objetivos (métricas) e para isso será necessário realizar alguns ajustes na formulação em relação à precisão e ao recall. Após essa etapa será possível evidenciar os três objetivos com seus valores normalizados entre zero e um através de um plano tridimensional.

$$\text{Erro de Precisão(EP)=Precisão - 1} \quad (8)$$

$$\text{Erro de Recall(ER)=Recall - 1} \quad (9)$$

Portanto, o vector de objetivos a considerar é:

$$J(\theta)=[PVU\ (\theta),EP(\theta),ER(\theta)] \quad (10)$$

Onde θ é um vetor de zeros e uns [0,1] indicando se a característica vai ser utilizada (1) ou não (0) pela estrutura de aprendizagem.

4. Resultados

4.1 Estudo de caso - bases de dados

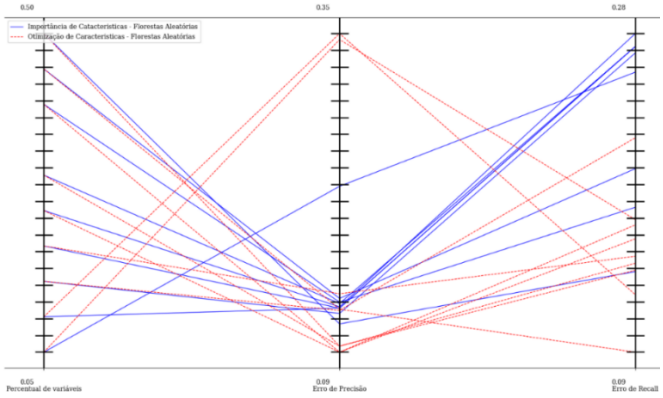
O estudo de caso foi elaborado através da utilização de fontes de dados públicas disponibilizadas pelo Kaggle, uma das maiores comunidades de ciência de dados do mundo. A base de dados selecionada possui uma variável alvo do tipo categórica com duas classes, indicando que estamos lidando com um problema de classificação binária, onde 1 representa a classe positiva e 0 a classe negativa. Além disso, é importante ressaltar que o conjunto de dados será tratado para balanceamento das classes e remoção de dados faltantes.

4.2 Comparação entre a Importância das características e Otimização Multiobjetivo

Os modelos baseados em árvores de decisão possuem uma medida que avalia a importância das variáveis (características), revelando quais têm maior influência na capacidade do modelo de fazer previsões. Essa avaliação é realizada observando como cada característica contribui para a redução da impureza nos nós da árvore durante as divisões. Embora essa medida seja relevante e escalável, ela pode ser menos precisa do que técnicas capazes de testar todas as combinações possíveis. No entanto, a busca exaustiva por todas as combinações entre as variáveis pode ser inviável devido ao custo computacional. Portanto, é essencial empregar métodos que otimizem o tempo de processamento. Nesse sentido, uma abordagem evolucionária é uma escolha promissora, utilizando conceitos de seleção, cruzamento e mutação para evoluir uma população (conjunto de variáveis) de soluções candidatas ao longo de várias gerações. O algoritmo escolhido para esta tarefa é o NSGA-III, conhecido por sua capacidade de lidar eficazmente com muitos objetivos. A seguir será mostrado a comparação entre as duas abordagens.

Figura 6. Comparação entre importância e otimização de características com modelo de florestas aleatórias e base de dados de câncer de mama.

DOI: 10.59254/sbpo-2024-193472
https://proceedings.science/p/193472?lang=pt-br



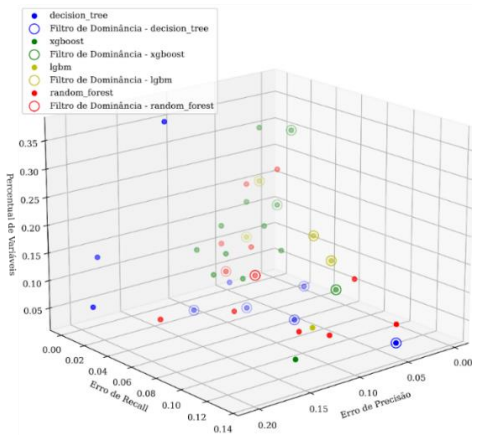
Ao analisar a relação entre o percentual de variáveis e o erro de precisão/recall utilizando a abordagem de importância de características, podemos observar padrões consistentes. À medida que reduzimos o número de variáveis, o erro de precisão geralmente permanece estável ou apresenta uma pequena redução na maioria dos casos. Isso sugere que as variáveis selecionadas possuem um impacto pouco significativo na capacidade do modelo de classificar corretamente os casos. No entanto, o erro de recall tende a aumentar à medida que diminuímos o percentual de variáveis. Isso indica que, embora o modelo possa manter uma boa precisão com menos variáveis, ele pode perder a capacidade de identificar corretamente os casos positivos.

Ao contrário da abordagem de importância de características, a otimização de variáveis revela uma variação mais expressiva no erro de precisão e recall para diferentes percentuais de variáveis. Ou seja, o modelo é mais responsivo a novas variáveis. Isso sugere que a performance do modelo não depende apenas da quantidade de variáveis, mas também da técnica de otimização empregada para combinar as melhores características. Os resultados revelam que, mesmo quando os percentuais das variáveis são idênticos, os modelos otimizados podem apresentar desempenhos superiores nas outras métricas. Essa distinção torna-se evidente ao examinarmos a Figura 6, onde a maioria das linhas vermelhas se situa abaixo das linhas azuis em ambos os objetivos. Quando essa tendência não se manifesta, um objetivo é frequentemente compensado pelo outro.

4.3 Soluções Não Dominadas

Após submeter as estruturas de aprendizagem a uma série de treinamentos, é possível obter os pontos de não dominação para cada modelo. Em outras palavras, cada estrutura de aprendizagem é treinada de maneira individual avaliando quais são as melhores soluções. No entanto, é crucial avaliar esses pontos de não dominação de forma conjunta, ou seja, considerando todas as estruturas de aprendizado simultaneamente. Para isso, faremos uso de um filtro de dominância.

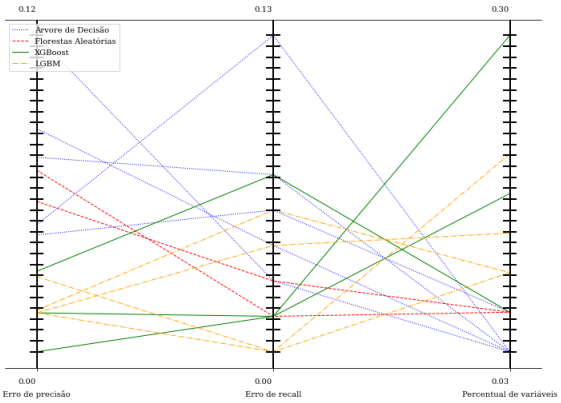
Figura 7. Soluções não dominadas, base de dados câncer de mama.



DOI: 10.59254/sbpo-2024-193472
https://proceedings.science/p/193472?lang=pt-br

Na Figura 7, todas as soluções não dominadas com filtro de dominância estão marcadas com um círculo para indicar sua relação de domínio sobre os pontos das demais estruturas de aprendizagem. Devido à limitação do gráfico de dispersão em evidenciar resultados com mais de duas dimensões, não é possível realizar afirmações conclusivas.

Figura 8. Soluções não dominadas com filtro de dominância, base de dados câncer de mama.



O gráfico de coordenadas paralelas proporciona uma visualização eficaz das estruturas de aprendizado e como elas se comportam em relação as variáveis. Analisando as métricas de desempenho (erro de precisão e erro de recall) juntamente com o percentual de variáveis utilizadas em diferentes modelos, podemos identificar padrões e tendências. Observando as linhas correspondentes aos modelos de árvore de decisão, notamos que, em geral, eles tendem a ter um erro de precisão mais elevado, variando entre 0.015 e 0.122, com um erro de recall oscilando entre valores altos e intermediários. O percentual de variáveis usadas nessas árvores de decisão geralmente é baixo. Em contraste, os modelos xgboost e lgbm possuem outro tipo de comportamento. Ambos tendem a apresentar erros de precisão e de recall mais baixos em comparação com as árvores de decisão, com alguns casos de erro de precisão atingindo 0. A distribuição do percentual de variáveis usadas nestes modelos varia, desde um número moderado até um número substancialmente maior de variáveis. Os modelos de florestas aleatórias exibem apenas duas soluções não dominadas, com erros de precisão e de recall oscilando, ainda que utilizem o mesmo número de variáveis.

5. Conclusões

Até o momento, foram realizados testes em quatro modelos fundamentados em árvores de decisão, aliados à otimização das características. Cada estrutura apresentou comportamentos distintos em relação aos dados de entrada, resultando em critérios de seleção variados. Além disso, foi fornecida uma representação visual alternativa para facilitar a compreensão do desempenho dos modelos de aprendizado de máquina durante as fases de teste. A comparação entre os modelos de forma conjunta, foi realizada com o intuito de simplificar a interpretação dos resultados. O resultado dessa abordagem combinada é um método que busca reduzir a complexidade e aprimorar a eficiência, identificando as variáveis mais relevantes e os modelos mais adequados. Simultaneamente, esse método oferece insights importantes sobre o comportamento dos modelos durante o treinamento e os trade-offs mais adequados para resolver o problema em questão. Isso permite que a subjetividade na escolha da métrica ou valor ideal seja considerada entre as soluções não dominantes.

Referências

Barton, D., Tribunal D. (2012). Making Advanced Analytics Work For You. Harvard Business Review.

DOI: 10.59254/sbpo-2024-193472
https://proceedings.science/p/193472?lang=pt-br

Lee, I., Shin Y. J. (2020). Machine Learning for enterprises: applications, algorithm selection, and challenges. Elsevier.

McAfee, A., Brynjolfsson, E. (2012). Big Data: The Management Revolution. Harvard Business Review.

Mingoti, S. (2023). Análise de dados através de métodos de análise multivariada. Editora ufmg.

Fávero, L. P., Belfiore, P. (2017). Manual de análise de dados. Elsevier.

LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature.

Kazemitabar, J., Amini, A., Bloniarz, A., Talwalkar, A. S. (2017). Variable Importance Using Decision Trees. NEURIPS.

Holzinger, A., Saranti, A., Molnar, C., Biecek, P., Samek, W. (2022). Explainable AI Methods - A Brief Overview. Springer.

Kratsch, W., Manderscheid, J., Röglinger, M., Seyfried, J. (2020). Machine Learning in Business Process Monitoring: A Comparison of Deep Learning and Classical Approaches Used for Outcome Prediction. Springer.

Alipour, M. M., Razavi, S. N., Derakhshi, M. R. F., Balafar, M. A. (2018). A hybrid algorithm using a genetic algorithm and multiagent reinforcement learning heuristic to solve the traveling salesman problem. Springer.

Cockburn, I. M., Henderson, R., Stern, S. (2019). The Impact of Artificial Intelligence on Innovation An Exploratory Analysis. Nber.

Chang, Y. (2017). Using k-means clustering to improve the efficiency of ant colony optimization for the traveling salesman problem. IEEE.

Cheng, C., Pourhejazy, P., Ying, K., Lin. Unsupervised Learning-based Artificial Bee Colony for minimizing non-value-adding operations. Elsevier, 2021.

Valdivia, A., Sánchez-Monedero, J., Casillas J. (2020). How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness. Wiley.

Li, W., Wu, Z (2020). A methodology for dam parameter identification combining machine learning, multi-objective optimization and multiple decision criteria. Elsevier.

Wang, Z., Li, J., Rangaiah, G. P. (2022). Machine learning aided multi-objective optimization and multi-criteria decision making: Framework and two applications in chemical engineering. Elsevier.

Al-tashi, Q., Abdulkadir, S. J., rais, H. m., mirjalili, S., Alhussian H. (2020). Approaches to Multi-Objective Feature Selection: A Systematic Literature Review. IEEE.

Karimi-Mamaghan, M., Mohammadi, M., Meyer, P., Karimi-Mamaghan, A. M., Talbi, E. (2022). Machine learning at the service of meta-heuristics for solving combinatorial optimization problems: A state-of-the-art. Elsevier.

Mitchell, T. (1997). Machine Learning.

DOI: 10.59254/sbpo-2024-193472
https://proceedings.science/p/193472?lang=pt-br

Quinlan, J.R. (1987). Simplifying decision trees. International Journal of Man Machine Studies. Elsevier.

Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K. (2003). KNN Model-Based Approach in Classification. Springer.

Breiman, L. (2001). Random Forests. Springer.

Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. ACM digital library. Ke1, G., Meng, Q., Finley, T. Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Neurips.

Reynoso-Meza, G., Blasco, X., Sanchis, J., Herrero, J. M. (2013). Comparison of Design Concepts in Multi-Criteria Decision-Making Using Level Diagrams. Polytechnic University of Valencia.

Reynoso-Meza, G., Blasco, X., Sanchisa, J., Martıneza, M. (2014). Controller tuning using evolutionary multi-objective optimisation: current trends and applications. Polytechnic University of Valencia. Elsevier.

Mattson, C. A., Messac, A. (2005). Pareto Frontier Based Concept Selection Under Uncertainty, with Visualization. Springer.

Jürgen Branke, Kalyanmoy Deb, Kaisa Miettinen, Roman Słowi´nski. Multi-Objective Optimization. Springer, 2014.

Poli, R., Kennedy, J., Blackwell T. (2007). Particle swarm optimization. Springer.

Jong, K. (1988). Learning with Genetic Algorithms: An Overview. Springer.

Deb, K. (2014). An Evolutionary Many-Objective Optimization Algorithm Using Reference-point Based Non-dominated Sorting Approach, Part I: Solving Problems with Box Constraints. IEEE.

McKinney, W. (2011). Pandas: a Foundational Python Library for Data Analysis and Statistics.

Walt, S. V., Colbert, S. C., Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. IEEE.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. IEEE.

Pedregosa, F., Varoquaux, G., Gramfort, A. Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot M., Duchesnay E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.

Blank, J., Deb, K. (2022). Pymoo: Multi-Objective Optimization in Python. IEEE.

Tangirala, S. (2020). Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm. International Journal of Advanced Computer Science and Applications.

DOI: 10.59254/sbpo-2024-193472
https://proceedings.science/p/193472?lang=pt-br