Atividade 3 MO810/MC959

[*] O link contendo o jupyter notebook e outros conjuntos de dados e também o datacar estão no drive:

https://drive.google.com/drive/folders/1XbJe1t7qfeV_G1yKNZamV8UuZNwLuPZP?usp=sharing

1st Betania E R da Silva
IC-Unicamp

2nd Decio Miranda Filho
IC-Unicamp

3rd Felipe Scalabrin Dosso
IMECC-Unicamp

I. Introdução

Este trabalho explora a interpretabilidade de modelos de aprendizado de máquina aplicados ao conjunto de dados *Adult*, que é comumente utilizado para prever a elegibilidade para auxílio social. Compararam-se dois tipos de modelos: um modelo não interpretável, representado por uma *Random Forest,* e um modelo interpretável, a regressão logística. Para investigar a influência de cada característica nas previsões, foram empregados métodos de explicabilidade, como SHAP e *Tree Interpreter* — este último sendo um método que não foi abordado em aula, mas que permite analisar as contribuições individuais de cada variável no modelo baseado em árvores. Além disso, foi utilizado o método contrafactual DiCE, que explora cenários alternativos para ilustrar como mudanças específicas em características individuais poderiam impactar as decisões dos modelos, facilitando a visualização dos fatores determinantes para a classificação.

II. Trabalhos Relacionados

A busca por interpretabilidade em modelos de aprendizado de máquina, especialmente em domínios sensíveis como justiça criminal e assistência social, tem impulsionado o desenvolvimento de métodos que explicam o funcionamento de modelos complexos e avaliam sua equidade. Um exemplo recente é o PolyFIT [1], uma abordagem que constrói modelos polinomiais a partir das interações de características em modelos de caixa-preta, gerando modelos substitutos mais transparentes. Utilizando árvores de interação de características, o PolyFIT identifica e aproveita interações relevantes para transformá-las em representações polinomiais que replicam o desempenho de modelos complexos. Em experimentos com o conjunto de dados *Adult*, o PolyFIT superou modelos lineares e polinomiais tradicionais, obtendo resultados comparáveis ao modelo EBM [2], conhecido por priorizar interpretabilidade e manter alta acurácia.

Jabbari et al. [3] investigam os trade-offs entre interpretabilidade e equidade em modelos de aprendizado de máquina, utilizando dados sintéticos e reais. Os autores identificaram quatro padrões de trade-offs em experimentos com dados sintéticos, que foram posteriormente validados nos conjuntos de dados COMPAS e *Adult*. As curvas geradas com dados reais confirmaram as tendências observadas nos dados sintéticos, demonstrando a relevância da análise em cenários práticos. A quantidade de *features* foi usada como medida de interpretabilidade, permitindo uma avaliação da complexidade do modelo de forma objetiva. O uso de dados sintéticos permitiu uma análise controlada das complexas interações entre interpretabilidade e equidade, fornecendo uma base sólida para explorar esses aspectos em contextos do mundo real.

Quadrianto et al. [4] propõem um método de aprendizado que transforma os dados em uma representação justa e interpretável, preservando seu significado original. Esse método foi aplicado ao conjunto de dados Adult, considerando gênero como uma característica protegida. Os resultados mostraram que a abordagem alcançou igualdade de oportunidade ao ajustar variáveis como status de relacionamento (*wife* e *husband*), mitigando viés de gênero e mantendo alta precisão. A decomposição residual e o critério de independência de Hilbert-Schmidt (HSIC) foram usados para medir a independência estatística entre a nova representação e as características protegidas, permitindo a visualização das modificações implementadas e explicando como a equidade foi alcançada.

Mothilal et al. [5] apresentam o método DiCE para a geração de explicações contrafactuais diversas em modelos de aprendizado de máquina. O método destaca a importância da diversidade nas explicações para facilitar a compreensão das decisões algorítmicas. O framework de otimização introduzido considera tanto a proximidade quanto a diversidade entre contrafactuais, garantindo uma análise mais rica das possíveis variações. A avaliação foi realizada em diversos conjuntos de dados, incluindo o *Adult Dataset*, e demonstrou que as explicações geradas permitem visualizar como alterações em características, como nível educacional e status marital, podem modificar as previsões. Isso promove a interpretabilidade e ajuda a identificar possíveis vieses nos modelos.

Esses estudos complementam abordagens clássicas de interpretabilidade e justiça, como o LIME [6] de Ribeiro et al. (2016), que fornece explicações locais para predições, e o SHAP [7] de Lundberg e Lee (2017), que utiliza valores de Shapley para quantificar a contribuição de cada característica. Trabalhos como o de Kusner et al. (2017) [8] também expandem o campo ao introduzir a justiça contrafactual, que avalia se decisões permanecem justas entre diferentes grupos sob diferentes cenários.

III. Metodologia

A análise de interpretabilidade dos modelos foi realizada após o treinamento e avaliação no conjunto de dados *Adult*, utilizando o **RandomForestClassifier** para representar um modelo não interpretável, configurado com os seguintes parâmetros: `max_depth=21`, `min_samples_leaf=3`, e `n_estimators=422`. Simultaneamente, utilizou-se o **LogisticRegression** como modelo interpretável, configurado com `C=68.1066`, `class_weight='balanced'`, `fit_intercept=False`, `max_iter=856`, `solver='newton-cg'`, e `tol=0.0001`.

A. Regressão Logística

A regressão logística é um modelo estatístico usado principalmente para tarefas de classificação binária [9]. Ela estima a probabilidade de uma observação pertencer a uma classe específica, aplicando uma transformação logística (sigmoide) sobre uma combinação linear das características de entrada [10]. Esse modelo é considerado interpretável porque seus coeficientes podem ser analisados diretamente para entender o impacto de cada característica na probabilidade prevista [11].

B. SHAP

SHAP (SHapley Additive exPlanations) é uma técnica de explicabilidade baseada nos *valores de Shapley*, originados da teoria dos jogos [12]. Ele distribui o valor da predição entre as características, atribuindo uma "contribuição"de cada característica para a previsão final, garantindo uma distribuição justa e consistente. **SHAP** calcula essas contribuições considerando a diferença na predição ao adicionar ou remover cada característica, em comparação com todas as combinações possíveis das outras características. Essa abordagem torna o *SHAP* uma técnica robusta para explicabilidade tanto local (explicação de uma predição específica) quanto global (explicação do comportamento geral do modelo). Ademais, o *SHAP* tem múltiplas aplicações para amplificar as explicações, abordando questões como comparações entre grupos e diagnóstico de falhas em modelos [13] além de ser utilizado em diversas áreas como farmacologia de identificação de compostos químicos [14], predição de doenças cardíacas [15] e também na área da física [16].

C. Contrafactual

O método contrafactual do DiCE (Diverse Counterfactual Explanations) gera explicações contrafactuais para modelos de machine learning, oferecendo exemplos que mostram como pequenas mudanças nas variáveis de entrada poderiam alterar a previsão [5]. O objetivo é ajudar a entender o comportamento do modelo, propondo alterações mínimas e plausíveis para converter uma previsão original em outra desejada. O algoritmo produz múltiplas explicações diversificadas, promovendo maior interpretabilidade e auxiliando usuários a identificar caminhos variados que levariam a um resultado diferente [5].

D. Tree-Interpreter

O ***Tree Interpreter***, disponível no repositório [17], decompõe a predição de uma instância x de um modelo de árvore de decisão em duas partes: o ***bias***, que representa o valor base ou média das predições do modelo, e as **contribuições das características**, que indicam o impacto individual de cada característica na previsão. Assim, a predição $\hat{y}(x)$ é dada por:

$$\hat{y}(x) = \text{bias} + \sum_{i=1}^{n} \text{contribuição}_i, \tag{1}$$

onde cada **contribuição**$_i$ representa o impacto de uma característica i específica, calculado pela média das alterações na predição ao longo dos caminhos percorridos dentro das árvores. Cada característica, portanto, influencia sobre a predição final com base nas mudanças observadas ao longo desses caminhos [12].

IV. Resultados

A. Explicabilidade

1) Regressão Logística: Para o modelo puramente interpretável, foi escolhida a regressão logística.

O gráfico dos coeficientes (Figura 1) demonstra o impacto das probabilidades das variáveis sobre a necessidade de auxílio social (lembrando que a variável target $y = 1$ equivale a essa necessidade). Destaca-se a variável *education level*, onde níveis mais baixos de educação sugerem maior necessidade de ajuda, enquanto níveis mais altos sugerem menor necessidade/probabilidade. Nota-se também a variável *Ocupacional_Status*, onde diversos tipos de profissão podem ter correlação com a necessidade de auxílio social. Observa-se que *work_class_Without pay* e *occupation_farming_fishing* estão associados a níveis mais altos de renda mais baixa.

Outra variável sugestiva na figura é *capital_gain*, com um coeficiente negativo alto, sugerindo que, à medida que o valor de capital_gain aumenta, as chances de uma pessoa pertencer ao grupo que necessita de auxílio social diminuem drasticamente.

Já a Figura 2 apresenta o resultado do log-odds, permitindo uma outra forma de interpretar os resultados da predição da regressão logística. Essa análise corrobora que *education_level* e *capital_ gain* são extremamente relevantes, indicando que pessoas com maior capital e escolaridade têm menor propensão a necessitar de auxílio. Outra variável de destaque é o *marital_status*, onde relações estáveis sugerem menor necessidade de auxílio, embora com uma sensibilidade inferior às duas variáveis anteriores.

Portanto, o *education_level* é uma variável fortemente preditiva da necessidade de auxílio social, com níveis mais baixos associados a uma maior necessidade. Condições de trabalho e tipo de ocupação também influenciam significativamente, sendo que ocupações de baixa remuneração ou instabilidade estão associadas a uma maior necessidade de auxílio. O *capital_gain* é um importante indicador de riqueza ou segurança financeira, fortemente associado à ausência de necessidade de auxílio.

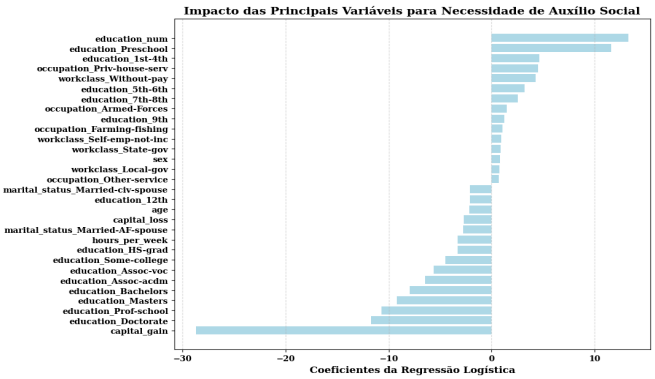

Figura 1. Impacto das Principais Variáveis nas Odds de Precisar de Auxílio Social

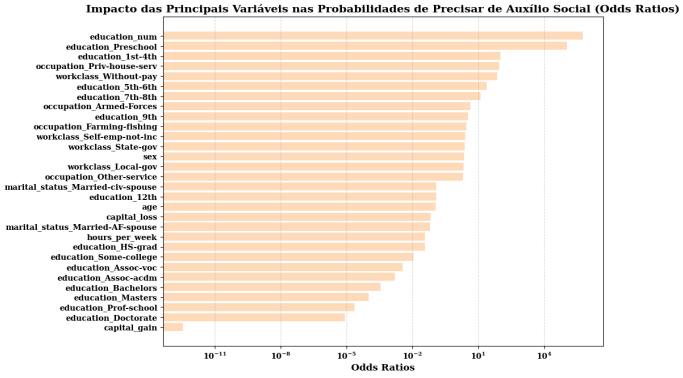

Figura 2. Impacto das Principais Variáveis nas Probabilidades de Precisar de Auxílio Social (Odds Ratios)

2) SHAP: Para realizar a explicabilidade do modelo não interpretável do `Random Forest`, utilizou-se o ***Tree-SHAP***, uma variante do SHAP específica para modelos baseados em árvores, que aproveita a estrutura hierárquica decisória dessas árvores e possui diversas aplicações acadêmicas [18].

Inicialmente, para a primeira amostra do conjunto avaliado, apresentaremos uma explicação local da importância das variáveis.

Com o *force-plot* (Figura 3), percebe-se que as variáveis que mais contribuem para a classificação de indivíduos necessitados de auxílio social incluem: relationship_Not-in-family = 1, indicando que a pessoa não é membro de uma família, correlacionando-se com rendas menores; capital_gain = 0, que significa que estar desempregado ou desalentado mostra extrema necessidade de receber ajuda; *marital_status_Never-married = 1*, indicando que a pessoa nunca foi casada, o que pode representar menor estabilidade financeira; *marital_status_Married-civ-spouse* = 0, indicando que não é casada civilmente; e age = 0.1507, que tem uma leve contribuição para a classe 1.

Por outro lado, as variáveis que contribuem para a classificação contra a necessidade de auxílio são *education_num = 0.9333*, que informa um nível mais alto de escolaridade e, portanto, maior renda; *education_Prof-school = 1*, indicando escolaridade profissional especializada; *hours_per_week = 0.551*, sugerindo que trabalhar mais horas está associado a rendas mais altas; e occupation_Prof-specialty = 1, representando uma ocupação profissional especializada com potencial de maior rendimento.

Em resumo, a previsão final de 0.06 é o resultado de uma interação entre fatores que aumentam a probabilidade de necessidade de auxílio e fatores que a reduzem, indicando que características como ausência de ganho de capital e estado civil influenciam a previsão de renda baixa, enquanto nível educacional e ocupação profissional indicam maior renda.

Na análise global da explicabilidade do modelo, as características mais relevantes, conforme a importância das variáveis, corroboram outras análises anteriores, como mostra a Figura 4.

A variável *marital_status_Married-civ-spouse* tem um forte impacto negativo, onde valores mais altos (pessoas casadas) reduzem a probabilidade de precisar de assistência. Já *education_num* indica que níveis mais altos de escolaridade reduzem a probabilidade de precisar de assistência, conforme refletido pelos valores SHAP negativos na figura.

A variável mais emblemática — que também surgiu nas análises anteriores — é o *capital_gain*, demonstrando um impacto positivo significativo quando o ganho de capital é pequeno ou ausente, influenciando a necessidade de requisitar auxílio social. Outras características influentes incluem a variável age, que apresenta impactos mistos, com certas faixas etárias contribuindo tanto para o aumento quanto para a diminuição da necessidade de assistência prevista; *hours_per_week* mostra que indivíduos que trabalham mais horas por semana têm menor probabilidade de precisar de assistência; *occupation_Exec-managerial* e *occupation_Prof-specialty* indicam que certas funções profissionais estão associadas a menores riscos de precisar de assistência.

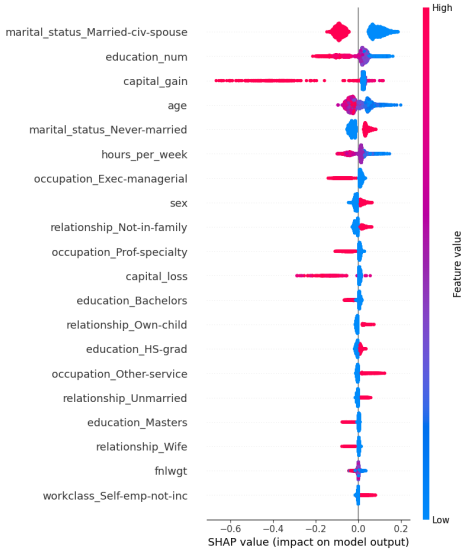

Figura 4. Explicação Global Summary Plot para Tree-SHAP

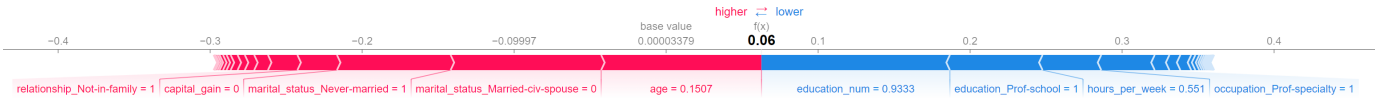

Figura 3. Explicação Local com Force Plot para o Tree-SHAP

Em suma, as características mais influentes do modelo incluem estado civil, nível de educação e ganhos de capital, indicando que ser casado, ter maior escolaridade e possuir ganhos de capital geralmente reduzem a probabilidade de precisar de assistência. Por outro lado, baixos ganhos de capital podem aumentar essa necessidade.

3) Tree-Interpreter: Os resultados do Tree Interpreter corroboram as análises anteriores com as ***feature importances***.

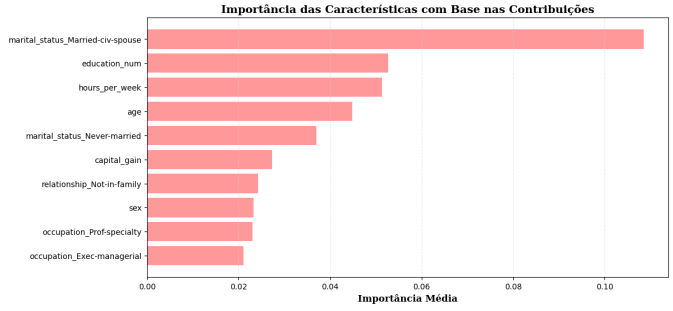

Figura 5. Feature Importance Tree Interpreter

Note na figura 5 que *education_num*, *hours_per_week*,*age* e *capital_gain* são as variáveis mais relevantes para explicar as features - o que ratifica a análise feita anteriormente com a regressão logística e o SHAP. O capital_gain sugere de fato que ter renda/patrimônio altos são grandes indicativos de que dificilmente um indivíduo irão precisar de auxílio, da mesma forma as variáveis age e *hours_per_week* sugere que pessoas com emprego estável e com bastante experiência também não incorrem na necesidade de auxílio social.

B. DiCE

Para a análise contrafactual, foi utilizado o modelo `rf_tun`. O método utilizado foi o **DiCE** por meio de contrafactuais, ou seja, amostras alteradas da instância original com o objetivo de mudar a previsão do modelo de `target=1` (baixa renda) para `target=0` (alta renda).

O processo envolveu a instância original apresentada ao DiCE possuía `target=1`. O DiCE então gerou cinco contrafactuais, ajustando os valores de algumas variáveis para mudar a previsão do modelo para `target=0`. No processo, o DiCE procura fazer o mínimo de alterações possível para manter os contrafactuais próximos da instância original.

Entre as variáveis analisadas, apenas a variável `capital_gain` foi alterada nos contrafactuais gerados. O valor original de `capital_gain` era 0.0, e para que o modelo previsse `target=0`, foram sugeridos pequenos aumentos, como 0.1, 0.2, e 0.5. Esse aumento foi suficiente para que o modelo alterasse sua previsão, enquanto as demais variáveis permaneceram constantes.

Por conseguinte, esses resultados indicam que, para o modelo `rf_tun`, a variável `capital_gain` tem uma influência significativa na diferenciação entre as classes `target=1` e `target=0`. Pequenos aumentos em `capital_gain` foram suficientes para mudar a previsão do modelo, sugerindo que ganhos de capital são um fator chave para o modelo associar uma instância com a classe `target=0`. Outras variáveis, como `age`, `hours_per_week`, e `relationship_Not-in-family`, não precisaram de ajustes, indicando uma influência menor no processo decisório do modelo.

V. Discussão

Nesse módulo da disciplina tratamos quatro modelos de explicabilidade: **Tree-SHAP, Regressão Logísitca, DiCE** além de incluir um modelo não apresentado em sala, o **Tree-Interpreter)**. Dessa forma, múltiplas abordagens de explicação de algoritmos com funcionamentos distintos foram utilizadas. De modo geral, muitas das explicações indicaram que características como *capital_gain*, *education_level* e *estado civil* são preditores importantes na avaliação da necessidade de auxílio social. As análises apontaram que maiores valores de renda e níveis educacionais reduzem a necessidade de auxílio, enquanto características como ocupações instáveis e baixa escolaridade aumentam essa necessidade. Essa diversidade de métodos de explicação revelou nuances importantes nas contribuições das variáveis, demonstrando que abordagens variadas oferecem uma compreensão mais completa das decisões dos modelos.

Referências

[1] J. Jang, M. Kim, C. Bui, and W.-S. Li, *Toward Interpretable Machine Learning: Constructing Polynomial Models Based on Feature Interaction Trees*, 05 2023, pp. 159–170.

[2] H. Nori, S. Jenkins, P. Koch, and R. Caruana, "Interpretml: A unified framework for machine learning interpretability," *arXiv preprint arXiv:1909.09223*, 2019.

[3] S. Jabbari, H.-C. Ou, H. Lakkaraju, and M. Tambe, "An empirical study of the trade-offs between interpretability and fairness," in *ICML 2020 Workshop on Human Interpretability in Machine Learning, preliminary version*, 2020.

[4] N. Quadrianto, V. Sharmanska, and O. Thomas, "Discovering fair representations in the data domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[5] R. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," 05 2019.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144.
[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
[8] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf
[9] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. Wiley, 2013.
[10] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
[11] J. S. Long and S. A. Mustillo, "Using predictions and marginal effects to compare groups in regression models for binary outcomes," *Sociological Methods Research*, vol. 50, pp. 1284 – 1320, 2018.
[12] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lulu. com, 2019.
[13] D. Bowen and L. Ungar, "Generalized shap: Generating multiple types of explanations in machine learning," *ArXiv*, vol. abs/2006.07155, 2020.
[14] R. Rodríguez-Pérez and J. Bajorath, "Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions," *Journal of Computer-Aided Molecular Design*, vol. 34, no. 10, pp. 1013–1026, October 2020.
[15] T. A. Assegie, "Evaluation of the shapley additive explanation technique for ensemble learning methods," *Proceedings of Engineering and Technology Innovation*, 2022.
[16] R. Pezoa, L. Salinas, and C. Torres, "Explainability of high energy physics events classification using shap," *Journal of Physics: Conference Series*, vol. 2438, 2023.
[17] Andosa, "Treeinterpreter: Package for interpreting scikit-learn's decision tree and random forest predictions," 2015, accessed: 2024-11-11. [Online]. Available: https://github.com/andosa/treeinterpreter
[18] O. Bifarin, "Interpretable machine learning with tree-based shapley additive explanations: Application to metabolomics datasets for binary classification," *PLOS ONE*, vol. 18, 2022.
[19] K. R. Varshney, *Trustworthy Machine Learning*. Chappaqua, NY, USA: Independently Published, 2022.
[20] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, 2019.
[21] Asniar, N. Maulidevi, and K. Surendro, "Smote-lof for noise identification in imbalanced data classification," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, pp. 3413–3423, 2021.
[22] B. Zhou, Y. Ying, and S. Skiena, "Online auc optimization for sparse high-dimensional datasets," *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 881–890, 2020.
[23] Y. Liu, Y. Li, and D. Xie, "Implications of imbalanced datasets for empirical roc-auc estimation in binary classification tasks," *Journal of Statistical Computation and Simulation*, vol. 94, pp. 183 – 203, 2023.
[24] Y.-C. Wang and C.-H. Cheng, "A multiple combined method for rebalancing medical data with class imbalances," *Computers in biology and medicine*, vol. 134, p. 104527, 2021.

APÊNDICE

Apenas o modelo de *Random Forest* havia sido treinado e avaliado no trabalho anterior. Dessa forma, o treinamento da regressão logística seguiu os mesmos passos dos modelos anteriormente treinados, com a mesma proporção de divisão dos conjuntos de treino e validação e otimização do threshold pelo método bayesiano, conforme discutido em sala de aula [19].

A otimização da regressão logística foi realizada com o auxílio do framework de otimização ***Optuna***, proposto por Akiba et al. (2019) [20], com uma busca dinâmica de hiperparâmetros. Os hiperparâmetros a serem otimizados com este método estão listados na Tabela I e com **50 trials** e assim como o Random Forest, otimizado visando à maximização do roc AUC score.

Tabela I
MALHA DE HIPERPARÂMETROS

Modelo	Hiperparâmetro	Intervalo ou Valor
LogisticRegression	C	(0.01, 100)
	penalty	['l2', None]
	solver	['newton-cg', 'lbfgs']
	max_iter	(100, 1000)
	tol	(1e-4, 1e-2)
	fit_intercept	[True, False]
	class_weight	['balanced', None]

A motivação por escolher o **Random Forest** e **Logistic Regression** foram o bom desempenho dos classificadores comparados a todos os outros. Dessa forma, os 2 modelos requisitados a serem explicados obtiveram:

- **LogisticRegression** `accuracy = 0.84663, brier score =0.126678 ,roc AUC score = 0.907616`;
- **Random Forest Classifier** `accuracy = 0.8340, brier score=0.1002,roc AUC score = 0.9118`;

O AUC foi escolhido como métrica principal por se tratar de um conjunto desbalanceado e apresentar a área sobre a curva ROC, o qual é recomendado por diversos trabalhos nesse caso [21] [22] [23] [24].