

Previsão de Mortalidade por COVID-19 sob Efeito de Concept Drift: Uma Abordagem Contínua com Janelas Deslizantes e XGBoost

Mateus Souza Lopes¹, Manoel Limeira de Lima Júnior¹, Rafael Alves Braga¹

¹Centro de Ciências Exatas e Tecnológicas – Universidade Federal do Acre (UFAC)
Rio Branco – AC – Brasil

lopes.mateus@sou.ufac.br, manael.limeira@ufac.br, alves.rafael@sou.ufac.br

Abstract. *Predicting COVID-19 mortality using Machine Learning presents challenges related to severe class imbalance and temporal changes caused by public health interventions, particularly vaccination. This study analyzed 644,628 clinical notifications collected between 2020 and 2025 in the state of Acre, Brazil. To preserve the chronological structure of the data and reduce temporal bias, a sliding window validation strategy was adopted. The XGBoost model, combined with a dynamic class-weight adjustment approach, achieved a recall of 0.7622 for the minority class (death) and an AUC-ROC of 0.9043. Temporal interpretability analysis revealed changes in the epidemiological behavior throughout the pandemic, showing that the model captured the transition of the main mortality predictors associated with vaccination progress. While age was the dominant predictor during the pre-vaccination period, post-vaccination stages were increasingly influenced by attributes related to olfactory disorders and the number of infections.*

Resumo. *A previsão de mortalidade por COVID-19 utilizando Aprendizado de Máquina enfrenta o desafio do desbalanceamento de classes e do concept drift gerado pelas intervenções de saúde, especialmente a vacinação. Este artigo analisou 644.628 notificações clínicas (2020-2025) do estado do Acre. Para mitigar o viés cronológico, utilizou uma estratégia de validação temporal com janelas deslizantes. O modelo XGBoost, aliado ao ajuste dinâmico de pesos entre classes, alcançou um recall de 0,7622 na identificação da classe minoritária (óbito) e AUC-ROC de 0,9043. A análise temporal de interpretabilidade evidenciou a evolução do comportamento epidemiológico ao longo da pandemia, demonstrando que o modelo capturou a transição dos principais fatores preditivos de mortalidade associada ao avanço da vacinação, passando da variável idade, predominante no período pré-vacinação, para atributos relacionados a distúrbios olfativos e ao número de infecções no período pós-vacinação.*

1. Introdução

A pandemia de COVID-19 gerou um volume massivo de dados clínicos, impulsionando a aplicação de Inteligência Artificial e Aprendizado de Máquina no apoio à gestão em saúde pública [Solanki et al. 2021]. No entanto, a predição da evolução clínica da doença, particularmente da mortalidade, apresenta complexidades inerentes à dinâmica dos dados do mundo real. Dois desafios metodológicos principais se destacam: a heterogeneidade nos

níveis de imunidade da população, decorrente de campanhas de vacinação e exposição prévia ao vírus, que gera um forte desvio de conceito (*concept drift*) na base de dados ao longo do tempo [Sahiner et al. 2023]; e o extremo desbalanceamento de amostras entre as classes, em que o desfecho de óbito representa uma fração pequena das notificações. Esses fatores tornam a modelagem preditiva mais complexa, exigindo abordagens capazes de lidar simultaneamente com mudanças temporais nos dados e com a baixa representatividade da classe minoritária.

A maioria das técnicas tradicionais de classificação assume que os dados seguem sempre o mesmo padrão ao longo do tempo [Lu et al. 2018]. Porém, em cenários epidemiológicos, esse comportamento pode mudar devido à evolução da doença e às intervenções de saúde. Assim, ignorar a ordem cronológica dos registros em validações convencionais pode causar vazamento de dados, utilizando informações do futuro para prever eventos do passado [Kaufman et al. 2012].

Nesse contexto, este artigo investiga o uso de algoritmos de Aprendizado de Máquina para prever desfechos clínicos, com foco na mortalidade ao longo da evolução da pandemia. As principais contribuições do trabalho são: o uso de uma estratégia de validação temporal com janelas deslizantes para acompanhar as mudanças nos dados ao longo do tempo (*concept drift*); o tratamento do forte desbalanceamento de amostras entre as classes (0,3% de óbitos) por meio de ajustes no modelo XGBoost; e a análise interpretável dos resultados, mostrando como a importância dos atributos clínicos e vacinais mudou durante a pandemia.

2. Trabalhos Relacionados

Diversos estudos utilizaram algoritmos de Aprendizado de Máquina para prever óbito por COVID-19. O trabalho de [Han et al. 2020] utilizou Random Forest e XGBoost sobre sintomas e exames de sangue, mas sua base estava limitada a 375 pacientes. [Adib et al. 2021] aplicou SVM para prever óbito em gestantes, sem incluir dados de vacinação nem discutir a importância dos atributos preditivos. Já [Walia and Jeevaraj 2021] buscou identificar o risco de óbito em estágios iniciais, apontando como limitação a necessidade de bases maiores.

Por sua vez, [Silva and Silva Neto 2023] utilizou dados da Síndrome Respiratória Aguda Grave para prever óbito e cura, revelando a importância das comorbidades. No entanto, não foi possível relacionar vacinação e óbito, já que apenas 37% dos registros eram de indivíduos vacinados no período da coleta. Em comum, essas abordagens não investigam como a importância das variáveis preditivas se altera diante de intervenções de saúde pública, como a vacinação, lacuna que motiva a proposta deste trabalho.

3. Metodologia

Considerando os desafios relacionados à dinâmica temporal dos dados epidemiológicos e ao severo desbalanceamento de amostras entre as classes, a metodologia foi definida voltada à preservação da cronologia dos registros e à adaptação contínua do modelo ao longo da pandemia. Nesse contexto, a pesquisa caracteriza-se como quantitativa e aplicada, com o objetivo de analisar dados referentes à vacinação e à pandemia causada pela COVID-19 no estado do Acre [Da Silva and Menezes 2005].

A pesquisa utilizou uma base da Secretaria de Estado da Saúde do Acre (SESA-CRE), com 644.628 notificações de casos suspeitos de COVID-19 entre março de 2020 e setembro de 2025. O desfecho clínico ‘*evolucao_caso*’ apresentou um desbalanceamento crítico: 99,7% (641.909 instâncias) de cura contra 0,3% (1.889 instâncias) de óbitos. Diante desse contexto, tornou-se necessário construir atributos capazes de representar características temporais relacionadas à evolução imunológica dos pacientes.

Para capturar a evolução do histórico clínico e vacinal dos pacientes, foram derivados novos atributos temporais: (i) ‘*numero_infeccao*’, conforme o protocolo clínico adotado, considerou-se como uma nova infecção a ocorrência de um novo teste positivo com um intervalo mínimo de 90 dias em relação ao evento anterior, seguindo orientações do Ministério da Saúde [COFEN 2020]; (ii) ‘*status_imunidade*’, definido dinamicamente a partir do atributo ‘*intervalo_vacina_infeccao*’, classificando indivíduos em “Não Vacinado”, “Imunidade Ativa” (última dose aplicada até 365 dias) ou “Imunidade Reduzida” (última dose aplicada há mais de 365 dias).

Como o comportamento epidemiológico variou significativamente ao longo da pandemia, tornou-se necessário adotar uma estratégia de validação capaz de preservar a dependência temporal dos dados. Dessa forma, métodos de validação como *Hold-out* e *K-fold cross-validation* não foram utilizados, evitando o uso de informações futuras durante o treinamento do modelo. Nesse contexto, adotou-se a técnica de Janela Deslizante (*Sliding Window*), preservando a cronologia das notificações. A abordagem de Janela Deslizante consiste em dividir os dados em subconjuntos temporais consecutivos, nos quais o intervalo de treinamento é deslocado progressivamente ao longo do tempo. Em cada iteração, o modelo é treinado utilizando apenas observações de períodos anteriores e avaliado em registros posteriores, preservando a ordem cronológica dos dados [Júnior et al. 2018]. O particionamento foi configurado com um bloco de 60 dias consecutivos para treinamento, seguido de 30 dias imediatamente posteriores para teste. A cada iteração, a janela avançava 30 dias de forma contínua até o processamento total da base.

Com a estratégia de validação temporal definida, a etapa seguinte concentrou-se na seleção de um modelo capaz de lidar com dados desbalanceados e, ao mesmo tempo, fornecer interpretabilidade sobre as variáveis analisadas. Dentre os classificadores experimentados, o XGBoost foi selecionado devido a sua capacidade de interpretação e desempenho. O XGBoost é um algoritmo baseado em árvores de decisão construídas por meio da técnica de *gradient boosting*, o modelo constrói árvores sequencialmente, buscando reduzir os erros das previsões anteriores e melhorar o desempenho preditivo. Para tratar o desbalanceamento, aplicou-se um ajuste na função objetivo do XGBoost via hiperparâmetro *scale_pos_weight*, que permite aumentar o peso associado às instâncias da classe minoritária durante o treinamento do modelo. Para lidar com a flutuação epidemiológica, o peso foi recalculado de forma dinâmica, correspondendo à razão entre o total de instâncias negativas (cura) e positivas (óbito) presentes exclusivamente na janela atual [Chen and Guestrin 2016].

Os códigos utilizados para o pré-processamento dos dados, construção dos gráficos e execução dos experimentos, bem como a base de dados e demais artefatos gerados, estão disponíveis publicamente no repositório Zenodo¹.

¹<https://doi.org/10.5281/zenodo.21051508>

Nesta pesquisa, utilizamos a ferramenta Gemini Pro 3.1 e Writefull Premium para fins de coesão, coerência, revisão ortográfica e gramatical do manuscrito, limitando-nos ao apoio técnico na escrita, sem interferência na concepção intelectual. Os autores revisaram todas as correções e assumem integralmente a responsabilidade pela precisão das informações e pela solidez intelectual do conteúdo apresentado nesta pesquisa.

4. Experimentos e Resultados

Devido à distorção gerada pela acurácia em bases desbalanceadas (onde classificar a cura garante um valor maior que 99%), a seleção do modelo baseou-se na AUC-ROC e no *recall* da classe minoritária (óbito). O processo experimental ocorreu em duas etapas. Inicialmente, avaliou-se diferentes configurações de hiperparâmetros para cada algoritmo: Regressão Logística (regularização inversa $C \in 0,01, 0,1, 1,0, 10,0, 100,0$); Decision Tree (número mínimo de amostras para realizar uma divisão $\in 10, 20, 30, 40, 50$); Random Forest (número de árvores $\in 100, 200, 300, 400, 500$); Support Vector Machine (*kernel* Linear, Polinomial e RBF); e XGBoost (profundidade máxima das árvores $\in 1, 3, 5, 7$, associado ao ajuste dinâmico do *scale_pos_weight*). Em seguida, a melhor configuração de cada algoritmo foi selecionada com base na AUC-ROC e comparada com as demais. Modelos como Decision Tree, Random Forest e Regressão Logística apresentaram falhas na identificação da classe positiva, com *recalls* variando entre 0,01 e 0,03, enquanto o SVM obteve *recall* igual a zero.

O XGBoost (MaxDepth=1) destacou-se com uma AUC-ROC global de 0,9043 e *recall* para óbito de 0,7622. O elevado *recall* obtido demonstra que o ajuste iterativo do *scale_pos_weight* aumentou a sensibilidade do modelo na identificação do padrão associado aos casos de óbito. Em contrapartida, ocorreu um *trade-off*: a precisão para a classe de óbito foi baixa (0,0252), refletindo o alto volume de falsos positivos nesse cenário.

Além das métricas globais de desempenho, a análise cronológica do modelo permitiu observar como o comportamento discriminativo variou ao longo das diferentes fases da pandemia, conforme apresentado na Figura 1.

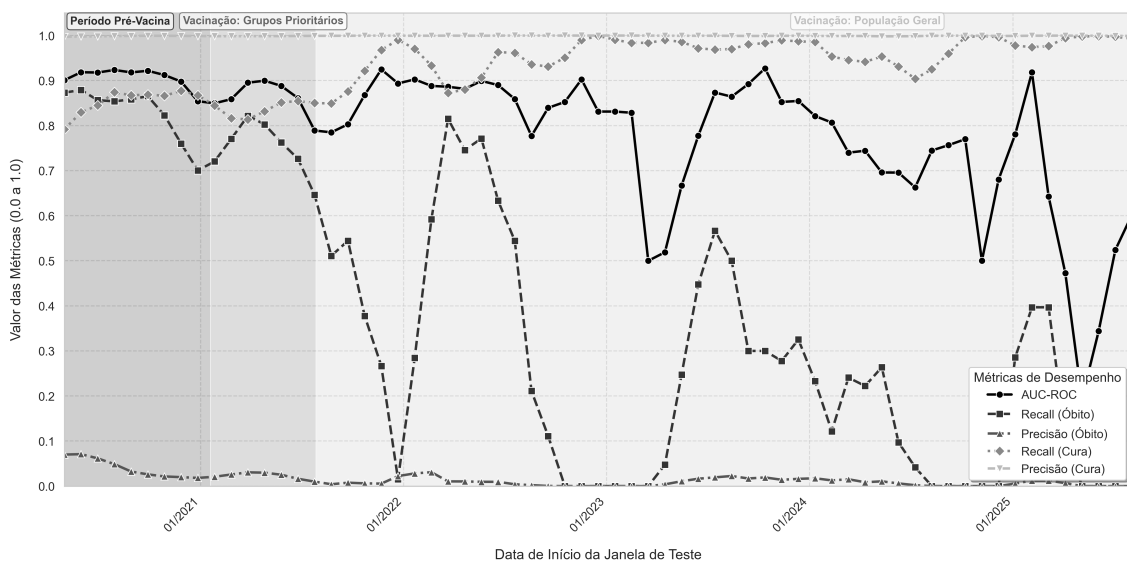


Figura 1. Evolução Temporal do Modelo XGBoost na Previsão de Óbito

Durante o período pré-vacina, o modelo manteve uma AUC-ROC predominantemente entre 0,85 e 0,95, com elevado *recall* para óbitos (acima de 0,8). No entanto, durante a vacinação, o *recall* da classe óbito diminuiu progressivamente, chegando próximo de zero na vacinação da população geral, possivelmente devido à redução da mortalidade e à mudança no perfil epidemiológico. As métricas da classe majoritária permaneceram altas em todo o período.

No período pós-vacinação, a AUC-ROC tornou-se mais instável com oscilações mais acentuadas em algumas janelas temporais. O *recall* para a classe óbito também apresentou elevada instabilidade, variando entre valores próximos de zero e aproximadamente 0,80, possivelmente em função da escassez de instâncias da classe minoritária em determinados períodos. Além disso, a precisão para a classe óbito permaneceu próxima de zero, evidenciando o impacto do desbalanceamento de amostras entre as classes e a dificuldade do modelo em realizar previsões confiáveis para os desfechos de maior relevância clínica.

Além da evolução do desempenho preditivo, também foi analisada a relevância das variáveis ao longo do tempo, permitindo compreender como os fatores associados à previsão de óbito se modificaram durante a pandemia. A relevância dos atributos foi estimada utilizando o mecanismo de importância de atributos do XGBoost, baseado na métrica *gain*, que estima a contribuição média de cada atributo para reduzir o erro do modelo. Com base nisso, a análise foi organizada em três períodos distintos da pandemia.

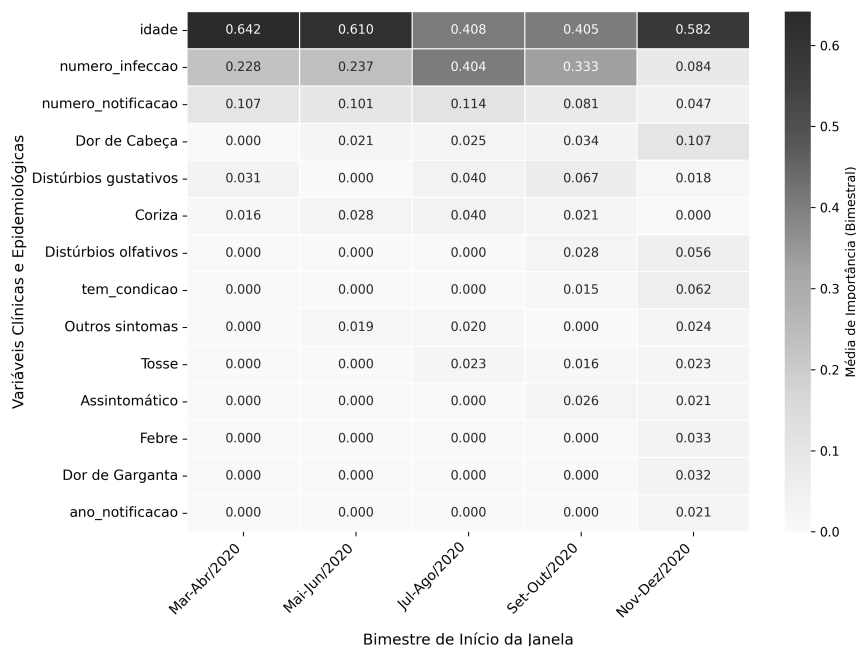


Figura 2. Importância Temporal dos Atributos no Período Pré-Vacina

No período pré-vacina (Mar-Dez de 2020), apresentado na Figura 2, o atributo *idade* dominou amplamente a predição do XGBoost, mantendo alta importância ao longo do período, com índices de importância variando entre 0,405 e 0,642, consolidando-se como o principal fator de risco nesse período. O atributo '*numero_infeccao*' também apresentou relevância significativa, especialmente entre Jul-Ago e Set-Out de 2020, possivelmente associado ao registro da primeira infecção por COVID-19. Contudo, em Nov-

Dez de 2020, sua importância diminuiu consideravelmente, enquanto a idade voltou a predominar no modelo.

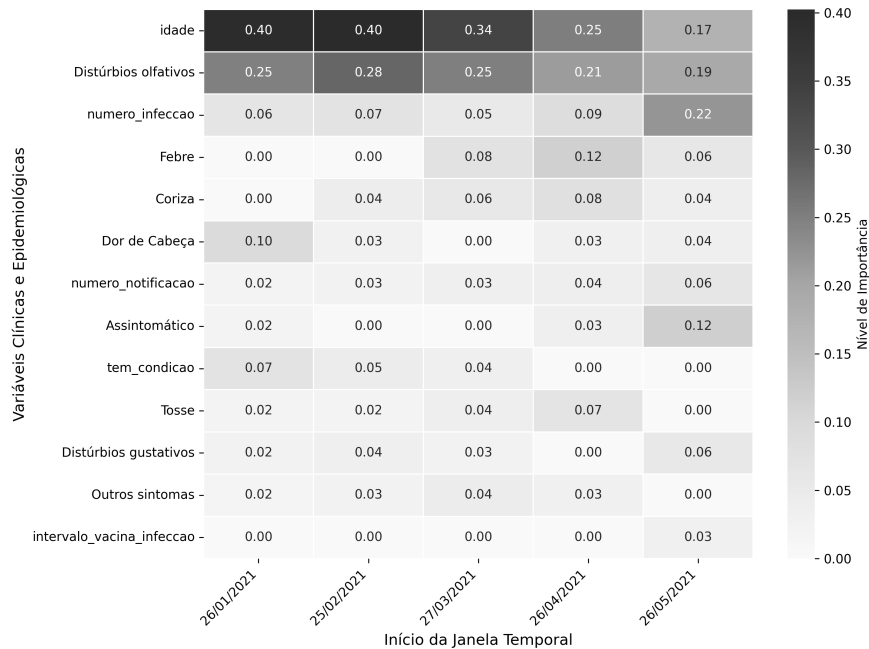


Figura 3. Importância Temporal dos Atributos na Vacinação Prioritária

Entretanto, esse comportamento começou a se modificar com o avanço da vacinação dos grupos prioritários (Jan-Jul de 2021), conforme a Figura 3. Esse intervalo coincide com o início e o escalonamento da campanha de vacinação contra a COVID-19 no Acre, que foi iniciada prioritariamente pelos grupos de maior risco, como idosos e profissionais de saúde [Agência de Notícias do Acre 2021]. Nesse contexto, a importância da variável idade apresentou um rápido declínio (de 0,40 para 0,17), indicando redução de seu poder discriminativo para a previsão do desfecho clínico. Esse comportamento está alinhado ao aumento da cobertura vacinal entre os idosos, reduzindo gradualmente o impacto da faixa etária como fator preditivo. De forma semelhante, atributos relacionados à presença de condições preexistentes também perderam relevância ao longo do período. Em contrapartida, outras *features* como ‘*Distúrbios olfativos*’, mantiveram importância relativamente estável, enquanto ‘*numero_infeccao*’ apresentou crescimento progressivo. Essa redistribuição da importância dos atributos sugere que o modelo foi capaz de captar mudanças no perfil epidemiológico das notificações associadas ao avanço da vacinação.

No período de vacinação ampla (25 de julho de 2021 até setembro de 2025), apresentado na Figura 4, observa-se uma mudança consolidada no perfil preditivo. A idade passa a apresentar importância secundária, reduzindo seu peso para 0,093 em 2025, o que sugere que fatores etários deixaram de atuar como determinantes isolados do desfecho clínico. Em contrapartida, atributos relacionados ao histórico de infecção e ao estado imunológico assumem mais relevância: a variável ‘*numero_infeccao*’ alcançou peso de 0,246, enquanto ‘*intervalo_vacina_infeccao*’ passou a integrar de forma consistente a estrutura de decisão do modelo. Esses resultados indicam que a recorrência de infecções e o declínio da proteção vacinal tornaram-se fatores importantes para a previsão de morte no período pós-vacinação prioritária.



Figura 4. Importância Temporal dos Atributos na Vacinação Ampla (Após 25 de Julho)

5. Conclusão

Este trabalho evidenciou que a previsão de morte em saúde pública deve considerar a dinâmica temporal dos dados epidemiológicos. A combinação da estratégia de Janelas Deslizantes com o classificador XGBoost, utilizando o hiperparâmetro *scale_pos_weight* ajustado em cada janela, mostrou-se eficaz no tratamento do desbalanceamento da base, alcançando AUC-ROC superior a 0,90 e elevado *recall* na identificação de casos de óbito.

A principal contribuição deste trabalho reside na demonstração da mudança do perfil epidemiológico ao longo da pandemia. Os resultados mostraram que o modelo de aprendizado de máquina foi capaz de identificar a transição das variáveis mais relevantes para a predição de óbito: fatores etários perderam o protagonismo observado no início da pandemia, enquanto atributos relacionados ao acompanhamento imunológico, como histórico vacinal e número de infecções, passaram a exercer maior influência sobre o desfecho clínico. Além disso, o modelo XGBoost destacou-se entre os algoritmos avaliados por apresentar melhor capacidade de identificação da classe minoritária, alcançando *recall* de 0,7622 para a classe óbito, embora a precisão para essa classe tenha permanecido baixa em razão do desbalanceamento dos dados. O modelo apresentou elevada capacidade discriminativa nas fases iniciais da pandemia, com valores de AUC-ROC predominantemente entre 0,85 e 0,95. De forma geral, os achados reforçam o potencial do Aprendizado de Máquina como ferramenta de apoio à tomada de decisão em saúde pública.

Algumas limitações devem ser consideradas. Destaca-se o *trade-off* com a métrica de precisão; a maximização do *recall* resultou em um volume significativo de falsos positivos. Além disso, as análises realizadas dependem da granularidade e dos ruídos inerentes ao sistema de notificação. A partir dessas limitações, trabalhos futuros podem investigar a aplicação de técnicas de superamostragem, como o SMOTE, que podem melhorar o desempenho dos algoritmos, além do uso de estratégias de validação por janelas expansivas para investigar o comportamento do modelo quando utilizada a totalidade dos dados disponíveis, e da experimentação com redes neurais profundas com o objetivo de comparar diferentes tipos de aprendizado.

Referências

- Adib, Q. A. R., Tasmi, S. T., Bhuiyan, S. I., Raihan, M. S., and Shams, A. B. (2021). Prediction model for mortality analysis of pregnant women affected with covid-19. In *2021 24th International Conference on Computer and Information Technology (IC-CIT)*, pages 1–6. IEEE.
- Agência de Notícias do Acre (2021). Acre inicia programa de imunização contra a COVID-19 e apresenta seus primeiros vacinados. Acessado em: 17 jan. 2026.
- Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, page 785–794, New York, NY, USA. Association for Computing Machinery.
- COFEN (2020). Ministério da saúde orienta conduta para casos suspeitos de reinfecção por COVID-19.
- Da Silva, E. L. and Menezes, E. M. (2005). Metodologia da pesquisa e elaboração de dissertação. *UFSC, Florianópolis, 4a. edição*, 123(4):138.
- Han, R., Liu, Z., Philip Chen, C., Xu, L., and Peng, G. (2020). Mortality prediction for covid-19 patients via broad learning system. In *2020 7th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS)*, pages 837–842. IEEE.
- Júnior, M. L. d. L., Soares, D. M., Plastino, A., and Murta, L. (2018). Automatic assignment of integrators to pull requests: The importance of selecting appropriate attributes. *J. Syst. Softw.*, 144(C):181–196.
- Kaufman, S., Rosset, S., Perlich, C., and Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data*, 6(4).
- Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., and Zhang, G. (2018). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, page 1–1.
- Sahiner, B., Chen, W., Samala, R. K., and Petrick, N. (2023). Data drift in medical machine learning: implications and potential remedies. *The British Journal of Radiology*, 96(1150):20220878.
- Silva, R. and Silva Neto, D. R. d. (2023). Inteligência artificial e previsão de óbito por Covid-19 no Brasil: uma análise comparativa entre os algoritmos Logistic Regression, Decision Tree e Random Forest. *Saúde em Debate*, 46:118–129. Publisher: Centro Brasileiro de Estudos de Saúde.
- Solanki, V., Solanki, U., Baliyan, A., Kukreja, V., Lamba, V., and Sahoo, B. K. (2021). Importance of artificial intelligence and machine learning in fighting with covid-19 epidemic. In *2021 5th International Conference on Information Systems and Computer Networks (ISCON)*, pages 1–8. IEEE.
- Walia, H. and Jeevaraj, S. (2021). Early mortality risk prediction in covid-19 patients using an ensemble of machine learning models. In *2021 International Conference on Computational Performance Evaluation (ComPE)*, pages 965–970. IEEE.