AVALIAÇÃO DO DESEMPENHO DE TÉCNICAS DE INTELIGÊNCIA COMPUTACIONAL NA CLASSIFICAÇÃO DE INDIVÍDUOS DIABÉTICOS E NÃO DIABÉTICOS

EQUIPE

Igor Costa Lins de Albuquerque[1], Artur Pereira Neto[1].

[1] Universidade Estadual de Montes Claros (Unimontes).

INTRODUÇÃO

- 537 milhões de pessoas diagnosticadas com diabetes mellitus globalmente (IDF, 2021).
- 10.5% da população brasileira possui diabetes (Costa, 2023).
- 240 milhões pacientes não diagnosticados (IDF, 2021).

- Ferramenta de diagnostico precoce.
- Algoritmos de classificação baseados em aprendizado de máquina.
- Desbalanceamento característico de base de dados da saúde.

- **Objetivo:** avaliação preliminar da performance de modelos classificadores de Machine Learning para diferenciação de pacientes diabéticos e não diabéticos.

MÉTODO

Base de dados:

- Diabetes Health Indicators Dataset (Teboul, 2021).
- 253.680 indivíduos.
- 218.334 indivíduos não diabéticos (86,07%).
- 35.346 indivíduos diabéticos ou pré-diabéticos (13,93%).
- 21 descritores, sendo 14 categóricos binários.

Pré-processamentos:

- Normalização e remoção de dados ausentes e duplicados.
- Índice de correlação.
- GridSearch.
- Oversampling.

MÉTODO

- **Modelos investigados:** Redes Neurais Multicamadas (RNM), Árvores de decisão (AD), Florestas randômicas (FR), CatBoost (CB) e XGBoost (XGB),

- **Métrica:** recall.
- Falsos negativos em relação a base de dados de saúde.

- Divisão 80/20.
- Python com Scikit-learn e TensorFlow.
- Google Colab.

RESULTADOS

- Colunas 'Fruits', 'Veggies', 'AnyHealthcare', 'NoDocbcCost', 'Income', 'Education' e 'CholCheck retiradas.
- Binarização categórica da coluna “GenHlth”.
- Remoção de 24.206 registros duplicados.
- 135.074 registros sintéticos de diabéticos.
- 388.754 registros.

RESULTADOS

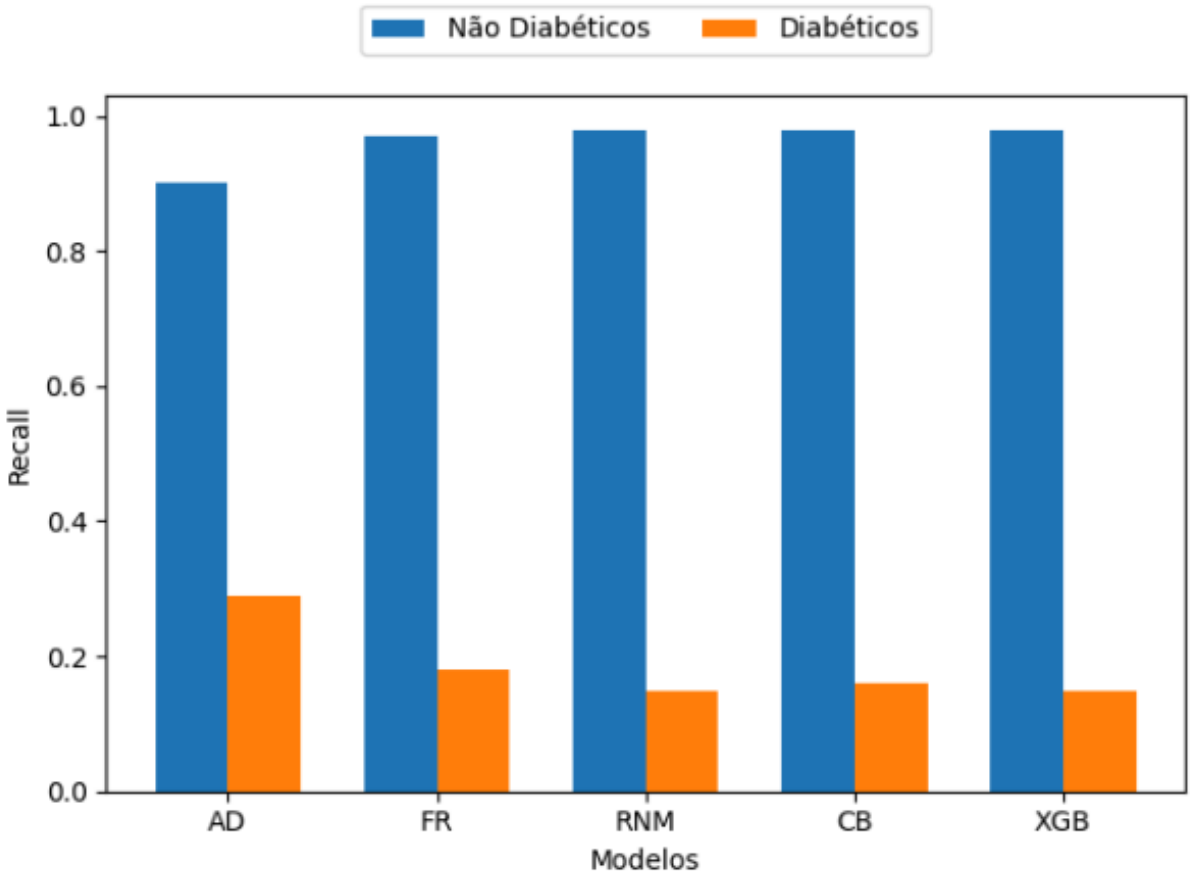

Figura 1. Comparação dos recalls obtidos para cada modelo de classificação sem o uso de oversampling.

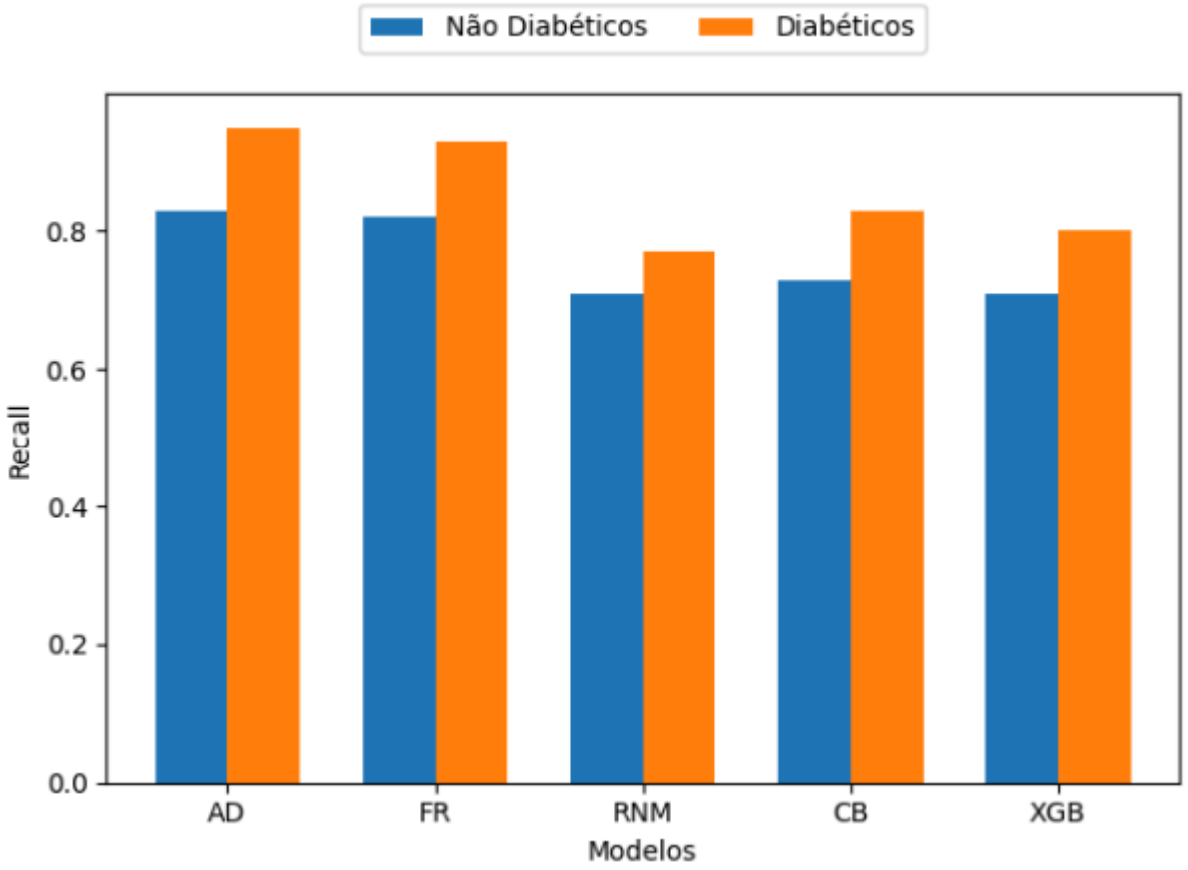

Figura 2. Comparação dos recalls obtidos para cada modelo de classificação com o uso de oversampling

DISCUSSÃO

Base de dados:

- Classificação tendenciosa sem oversampling.
- Aumento de 450% de recall com uso do oversampling.

Classificadores:

- Maiores métricas na Árvore de Decisão e Floresta Randômica.
- Baixa capacidade de generalização da base pela RNM.
- Resultados intermediários do CatBoost e XGBoost.

CONSIDERAÇÕES FINAIS

- Aumento de 450% de recall com uso de oversampling.
- Árvore de Decisão e Floresta Randômica com melhor performance para a classificação de diabéticos nesta base dentre os classificadores analisados.

Sugestão para trabalhos futuros:

- Análise de hiperparâmetros dos modelos propostos.
- Estender a investigação para modelos multiclasses incluindo a condição de pré-diabético.

AGRADECIMENTOS

- Gostaria de agradecer ao PPGMCS e ao LICA por todo auxílio durante a produção deste resumo expandido.

REFERÊNCIAS

- TEBOUL, A. Diabetes Health Indicators Dataset. 2021. Disponível em: https://11nq.com/UYwh5. Acesso em: 01 jul. 2024.
- COSTA, L. F.; SAMPAIO, T. L.; MOURA, L.; ROSA, R. S.; ISER, B. P. M. Time trend and costs of hospitalizations with diabetes mellitus as main diagnosis in the Brazilian National Health System, 2011 to 2019. Epidemiologia e Serviços de Saúde, v. 32, n. 4, p. 1-11, 2023.
- FEDERAÇÃO INTERNACIONAL DE DIABETES - IDF. Atlas de Diabetes da IDF. 9. ed. Bruxelas, Bélgica: IDF, 2021. Disponível em: https://diabetesatlas.org/regional-factsheets/. Acesso em: 10 jul. 2024.

Obrigado.

E-mail: igorcostalins2005@gmail.com.