EXPLORING THE APPLICATION OF MACHINE LEARNING MODELS IN THE ANALYSIS OF SUSPECTED DIABETES: A PRELIMINARY INVESTIGATION

EXPLORANDO A APLICAÇÃO DE MODELOS DE APRENDIZAGEM DE MÁQUINA NA ANÁLISE DE SUSPEITA DE DIABETES: UMA INVESTIGAÇÃO PRELIMINAR

SILVA, Jhônatas; HIRATA, Mateus; SANTOS, Flávia Aparecida Oliveira; CARVALHO, Jaqueline Corrêa Silva; CARVALHO, Marcos Alberto; RAMOS, Celso de Ávila; BASTOS, Camila; PEREIRA, Patrícia Carolina de Souza; SILVA, Vinícius Duarte Esteves

Jhônatas Silva, UNIFENAS, Brasil

Mateus Hirata, UNIFENAS, Brasil

Marcos Alberto Carvalho, UNIFENAS, Brasil

Jaqueline Corrêa Silva Carvalho, UNIFENAS, Brasil

Flávia Aparecida Oliveira Santos, UNIFENAS, Brasil

Camila Bastos, UNIFENAS, Brasil

Patrícia Carolina Souza, UNIFENAS, Brasil

Celso de Ávila Ramos, UNIFENAS, Brasil

Vinícius Duarte Esteves Silva, UNIFENAS, Brasil

Revista Científica da UNIFENAS
Universidade Professor Edson Antônio Velano, Brasil
ISSN: 2596-3481
Publicação: Trimestral
vol. 6, nº. 5, 2024
revista@unifenas.br

Recebido: 08/07/2024
Aceito: 28/08/2024
Publicado: 09/09/2024

URL: https://revistas.unifenas.br/index.php/revistaunifenas/issue/view/52

DOI: 10.29327/2385054.6.5-6

ABSTRACT: The conducted study addresses the possibility of employig machine learning models to assist in the early diagnosis of diabetes mellitus, a chronic condition that significantly impacts patients' quality of life and strains healthcare systems. The main objective is to explore the use of machine learning to aid in diabetes risk analysis, complementing traditional medical analysis. The dataset used was processed and normalized, and balancing techniques such as SMOTE and undersampling were employed, preparing the data to train three models: Keras Neural Network, Random Forest, and Gradient Boosting. Results show that the Random Forest model performs best overall, with high accuracy and the ability to minimize false positives, which is crucial given the study's context to prevent incorrect diagnoses or actual disease cases from going unnoticed. The study also highlights that synthetic data generation techniques can enhance the representativeness of imbalanced medical datasets, reinforcing their potential for future applications in medicine.

KEYWORDS: Artificial Intelligence, Machine Learning, Supervised Learning, Diabetes Mellitus, Early Diagnosis

RESUMO: O estudo conduzido aborda a possibilidade de empregar modelos de aprendizado de máquina para auxiliar no diagnóstico precoce do diabetes mellitus, uma condição crônica que impacta significativamente a qualidade de vida dos pacientes e sobrecarrega os sistemas de saúde. O objetivo principal é explorar o uso de aprendizado de máquina para auxiliar na análise de risco para um diagnóstico do diabetes, complementando a análise médica tradicional. Os dados empregados como conjunto de testes foram processados e normalizados, e técnicas de balanceamento como SMOTE e undersampling foram empregadas, preparando os dados para treinar três modelos: Keras Neural Network, Random Forest e Gradient Boosting. Os resultados mostram que o modelo Random Forest tem a melhor performance geral, com alta precisão e capacidade de minimizar falsos positivos,

características especialmente importantes dado o contexto do estudo, evitando diagnósticos incorretos ou casos verdadeiros passando desapercebidos. O estudo também evidencia que técnicas de geração sintética de dados podem melhorar a representatividade de conjuntos de dados médicos desbalanceados, reforçando seu potencial para aplicações futuras em medicina.

PALAVRAS-CHAVE: Flutter, Inteligência Artificial, Aprendizado de Máquina, Aprendizado Supervisionado, Diabetes Mellitus, Diagnóstico Precoce

1 INTRODUÇÃO

O diabetes mellitus é uma doença crônica, caracterizada por índices elevados de açúcar na corrente sanguínea, que traz consigo diversas complicações a curto e longo prazo [1], que diminuem drasticamente a qualidade de vida do paciente [2] e contribuem para a sobrecarga dos órgãos de saúde [3]. Logo, identificar pacientes em situações de risco para o desenvolvimento ou, ainda, pacientes que possuam a doença de forma assintomática é crucial. Tais medidas contribuem de forma geral para a saúde pública ao reduzir custos e auxiliar a preservar a saúde dos pacientes a longo prazo [4].

O presente trabalho busca explorar a viabilidade de treinar um modelo de aprendizado de máquina para auxiliar na análise de risco de um paciente para o desenvolvimento ou diagnóstico do diabetes, de forma a facilitar o tratamento futuro e trazer melhorias do nível individual até o nível institucional. A proposta de uso do modelo aconselha que esse seja empregado como ferramenta complementar pelo profissional médico responsável pelo acompanhamento do paciente, auxiliando o profissional a notar situações de risco que, ordinariamente, poderiam ser pouco evidentes.

Essa capacidade dos modelos de aprendizado de máquina em reconhecer padrões em conjuntos de dados que são difíceis de notar para observadores humanos [5], possivelmente emitindo avisos a respeito de situações de risco que, comumente, poderiam passar desapercebidas, é o principal fator que indica a possibilidade de um aumento da eficiência no diagnóstico a partir do emprego da ferramenta. Além disso, a capacidade de processar dados preexistentes de forma rápida permite a esse tipo de modelo efetuar uma análise de risco inicial com base no histórico de saúde já estabelecido de um paciente. Em uma situação em que os sintomas ainda não são evidentes, um alerta prévio acerca do risco pode gerar uma requisição de exames específicos pelo profissional médico antes do que seria usual, possivelmente permitindo opções de tratamento que busquem evitar a progressão do diabetes.

Como objetivo secundário do trabalho, espera-se gerar evidência que reforce ou conteste o uso de técnicas de geração sintética de entradas para normalizar conjuntos anonimizados de dados médicos normalmente pouco representativos, adequando-os para uso como conjuntos de treinamento para modelos de aprendizado de máquina que busquem identificar situações de risco para o diabetes ou outros distúrbios médicos. Embora usualmente se priorize conjuntos de dados reais nesse tipo de treinamento [6], esses podem nem sempre ser representativos o bastante, ou mesmo estar disponíveis para estudo. Logo, é importante saber qual o nível de ruído que esse tipo de técnica gera no sistema, para que se possa consolidar ou não o uso dela como meio de remediar um conjunto pouco representativo de dados [7].

2 METODOLOGIA

Para o desenvolvimento do trabalho proposto, foi utilizado o conjunto de dados “Diabetes prediction dataset”, de autoria de Mohammed Mustafa [8], que traz cerca de cem mil entradas de dados de pacientes com nove colunas de informação, divididas em gênero, idade, indicador de hipertensão, indicador de doenças cardíacas, histórico de fumante, índice de massa corporal (IMC), nível de hemoglobina glicada (HbA1c), nível de glicose sanguínea, e a classe de saída, indicando um diagnóstico positivo ou não de diabetes.

Todo o processo de tratamento de dados e treinamento dos modelos foi feito no ambiente Colab do Google [9], empregando python como linguagem para executar os scripts, junto às bibliotecas pandas, numpy, sklearn, imblearn e keras.

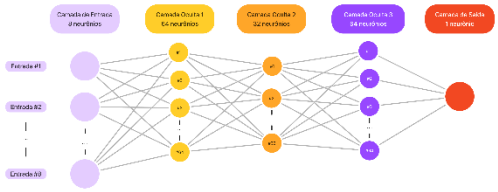
Para a higienização dos dados, a primeira etapa no tratamento dos dados, foram removidas todas as linhas que apresentavam o valor “No Info” na coluna de histórico de fumante, pois acredita-se que inferir um valor numérico para a falta dessa informação poderia interferir significativamente nos resultados do treino. Após essa remoção, o conjunto de dados reteve cerca de 60% (aproximadamente sessenta mil entradas) do seu volume.

Em seguida, os dados foram normalizados empregando a função StandardScaler da biblioteca sklearn, e as colunas categóricas (valores textuais) convertidas em representações numéricas por meio da função LabelEncoder da mesma biblioteca. O primeiro passo é importante porque, em geral, é documentada uma maior eficácia dos modelos quando os dados apresentados a ele estão em uma mesma escala entre si [10]; já o segundo passo é importante para que o modelo possa compreender, de certo modo, o significado dos valores textuais [11].

Durante a parte final de tratamento dos dados, para tratar o desbalanceamento do conjunto de dados (casos negativos em quantidade vastamente superior aos casos positivos), empregou-se uma técnica conhecida como Synthetic Minority Oversampling Technique (SMOTE). De forma simples, essa técnica aumenta o número de casos para as classes sub-representadas, objetivando balancear a distribuição no conjunto de dados [12]. Em seguida, aplicou-se o UnderSampling, para com o objetivo de diminuir o volume do conjunto de dados a ser processado e otimizar a performance [13].

Foram selecionados três modelos para treinar usando esse conjunto de dados processados, em busca daquele que oferecesse a melhor performance: um baseado em redes neurais (Keras), e dois modelos que usam os métodos de ensemble (Random Forest e Gradient Boosting). O modelo Keras emprega múltiplas camadas de neurônios para processar a entrada [14], o Random Forest combina várias árvores de decisão e obtém uma classificação final com base na votação entre elas [15], e o modelo Gradient Boosting cria árvores de decisão sequencialmente, objetivando corrigir os erros do modelo anterior [16].

Figura 1. Modelo Keras



O modelo Keras, ilustrado na figura 1, foi configurado com uma arquitetura sequencial composta por três camadas densamente conectadas. A camada de entrada tem 64 unidades com ativação 'relu' e regularização L2 (0.001), seguida por um dropout de 50% para evitar overfitting. A camada oculta intermediária tem 32 unidades, também com ativação 'relu' e regularização L2 (0.001), seguida por outro dropout de 50%. A terceira camada é similar em configuração à primeira. Por fim, a camada de saída é uma única unidade com ativação 'sigmoid' para classificação binária (indicação de diabetes positiva ou negativa). O modelo é compilado usando o otimizador Adam, com a função de perda binária cross-entropy e métrica de acurácia. O treinamento é realizado ao longo de 150 épocas com um tamanho de lote de 32 e um limiar de decisão de 0.4.

O modelo Random Forest foi configurado com 100 árvores de decisão (estimadores) e um estado aleatório fixo (random state) de 0 para garantir reprodutibilidade. O modelo Gradient Boosting foi configurado com 100 estimadores e um estado aleatório fixo (random state) de 0 para garantir reprodutibilidade, de modo similar ao modelo anterior.

3 RESULTADOS E DISCUSSÃO

Depois de treinados os modelos, todos eles foram testados no conjunto de dados de treino, que consiste em 30% do conjunto de dados higienizado, mas não alterado pelo SMOTE, de modo a reter a distribuição dos dados e garantir uma representação fiel da performance dos modelos desenvolvidos. As métricas de avaliação dos modelos treinados estão dispostas na Tabela 1.

Tabela 1. Métricas de avaiação dos modelos treinados

Legenda: Verdadeiros Negativos (VN), Falsos Negativos (FN), Falsos Positivos (FP), Verdadeiros Positivos (VP)

Com base nas métricas de avaliação apresentadas na tabela anterior, é importante destacar a maior relevância da métrica F1 no contexto do estudo. Isso porque essa é a média harmônica entre a precisão e o recall, oferecendo um modo de visualizar o balanceamento entre as duas métricas. A importância disso é especialmente notável em contextos contendo conjuntos de dados desbalanceados, nos quais a análise isolada de uma das duas métricas anteriores pode levar a conclusões enganadoras.

Além disso, devido ao objetivo do trabalho, que envolve a predição de uma condição de saúde, a célula dos falsos negativos na matriz de confusão deve receber atenção especial. Falsos negativos representam casos em que a condição de saúde não é detectada pelo modelo, o que pode levar a consequências negativas para os pacientes. Portanto, minimizar os falsos negativos é essencial para garantir que a condição de saúde seja identificada e tratada adequadamente.

Felizmente, o modelo Random Forest apresentou a melhor performance geral, mesmo não apresentando o melhor resultado absoluto no caso dos falsos negativos. A diferença entre o modelo Keras (com o menor número de falsos negativos) e o modelo Random Forest foi de apenas dois casos, que é compensada pela precisão quase dez vezes maior do segundo modelo para evitar falsos positivos, uma vez que esses também acarretariam custos em um ambiente de produção.

Apesar dos resultados positivos apresentados, algumas limitações do estudo devem ser consideradas. Em um cenário ideal, todas as entradas do conjunto de dados empregado poderiam fazer parte do treinamento, o que possibilitaria que as técnicas de balancemento de dados gerassem cenários ainda mais próximos da realidade, contribuindo para aumentar a confiabilidade dos modelos. Esse cenário não foi verificado no presente trabalho, mas representa situações passíveis de serem encontradas em trabalhos futuros, que possuam um conjunto de dados com maior aproveitamento de entradas do que o empregado nesse estudo e que, consequentemente, apresentem uma representação mais precisa do mundo real.

Além disso, devido a limitações de tempo e de poder de processamento, o estudo se limitou a analisar de modo relativamente superficial os resultados gerados por apenas três modelos de aprendizado de máquina, assim como deixou de contemplar a otimização dos hiperparâmentros das redes neurais. Desse modo, esses fatores constituem múltiplos pontos a serem melhorados em iterações futuras de pesquisa, de modo a beneficiar o desenvolvimento de ferramentas mais precisas e eficazes na predição do risco de desenvolvimento de diabetes.

CONCLUSÃO

Com base nos resultados obtidos, o modelo treinado

demonstrou boa precisão, ao mesmo tempo em que o custo financeiro associado ao treinamento se mostrou relativamente baixo, uma vez que o ambiente Colab do Google foi suficiente, sem a necessidade de um plano pago. O presente estudo, embora de escopo limitado, demonstra que um modelo propriamente treinado possui potencial para fazer predições úteis ao campo da medicina. Sugere-se que um próximo passo consista em integrar esse modelo em uma ferramenta prática, que possa ser empregada efetivamente no meio médico, de modo a facilitar tratamentos futuros e promover melhorias abrangendo do âmbito individual até o institucional.

Além disso, o emprego de técnicas de geração sintéticas de dados para viabilizar conjuntos de dados poucos representativos apresentou resultados promissores. Combinando as técnicas de SMOTE e undersampling para balancear o conjunto de dados de treinamento, os modelos mantiveram alta precisão mesmo em conjuntos de dados originalmente desbalanceados, onde os casos positivos eram severamente subrepresentados. Isso fortalece a evidência de que tais abordagens podem ser eficazes na preparação de dados para modelos de aprendizado de máquina mesmo em contextos médicos, onde a representatividade dos dados não é sempre garantida, embora seja essencial para o treinamento de modelos precisos.

REFERÊNCIAS

[1] American Diabetes Association. Standards of Medical Care in Diabetes—2022. Diabetes Care. 2022;45(Suppl 1).

[2] International Diabetes Federation. IDF diabetes atlas. Brussels International Diabetes Federation; 2019.

[3] Bommer C, Heesemann E, Sagalova V, Manne-Goehler J, Atun R, Bärnighausen T, et al. The global economic burden of diabetes in adults aged 20-79 years: a cost-of illness study. The lancet Diabetes & endocrinology [Internet]. 2017;5(6):423–30. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28456416

[4] Diabetes Prevention Program Research Group. Long-term Effects of Lifestyle Intervention or Metformin on Diabetes Development and Microvascular Complications over 15-year follow-up: the Diabetes Prevention Program Outcomes Study. The Lancet Diabetes & Endocrinology [Internet]. 2015 Nov;3(11):866–75. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4623946/

[5] Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine [Internet]. 2018 May 8;1(1). Available from: https://www.nature.com/articles/s41746-018-0029-1

[6] Beaulieu-Jones BK, Greene CS. Semi-supervised learning of the electronic health record for phenotype stratification. Journal of Biomedical Informatics. 2016 Dec;64:168–78.

[7] Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association [Internet]. 2016 Aug 13;24(2):ocw112. Available from: https://academic.oup.com/jamia/article/24/2/361/2631499

[8] Diabetes prediction dataset [Internet]. www.kaggle.com. Available from: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

[9] Google. Google Colaboratory [Internet]. Google.com. 2019. Available from: https://colab.research.google.com/

[10] Han J, Kamber M, Pei J. Data mining : concepts and techniques. Burlington, Ma: Elsevier; 2022.

[11] Géron A. Hands-on machine learning with Scikit-Learn and TensorFlow concepts, tools, and techniques to build intelligent systems. 2nd ed. O'Reilly Media, Inc.; 2019.

[12] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002 Jun 1;16(16):321–57.

[13] He H, Garcia EA. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering. 2009 Sep;21(9):1263–84.

[14] CholletF. Deep learning with Python. Shelter Island, NY: Manning Publications; 2018.

[15] Breiman L. Random Forests. Machine Learning [Internet]. 2001;45(1):5–32. Available from: https://link.springer.com/article/10.1023/A:1010933404324

[16] Friedman JH. Greedy function approximation: A gradient boosting machine. The Annals of Statistics. 2001 Oct;29(5):1189–232.