ISSN: 2319-0124
josif 2023
15ª JORNADA CIENTÍFICA E TECNOLÓGICA
12º SIMPÓSIO DE PÓS-GRADUAÇÃO
DO IFSULDEMINAS
REALIZAÇÃO INSTITUTO FEDERAL Sul de Minas Gerais
APOIO FAPEMIG CNPq

TESTE DE DIABETES: estimativas de diagnóstico de diabetes usando modelos da Inteligência Artificial

Antônio C. de L. FILHO[1] Vinicius da S. RIBEIRO[2];

RESUMO

A partir do algoritmo de classificação Random Forest, categorizado como um modelo de aprendizado de máquina não supervisionado, desenvolveu-se um código com o objetivo de realizar um pré-diagnóstico de diabetes a partir da Inteligência Artificial (IA). Por meio de critérios biológicos como gênero, idade, hipertensão, doenças cardíacas, histórico de tabagismo, Índice de Massa Corporal (IMC), nível médio de açúcar no sangue (HbA1c) e nível de glicose, determina-se um teste de possibilidade de presença de diabetes no quadro clínico de um indivíduo.

Palavras-chave:
DIAGNÓSTICO; GLICOSE; IA; RANDOM FOREST;

1. INTRODUÇÃO

Nos últimos anos, as Tecnologias de Informação e Comunicação (TICs) têm se tornado um importante aliado no avanço da saúde e da medicina. "Na falta de serviços de saúde formalizados, as TIC (particularmente, usando abordagens sanitárias volantes com telefones celulares & mídias sociais ou mHealth) podem ser usadas para mobilizar os habitantes para ajudar a prevenir a disseminação, cuidar dos doentes, e ter uma compreensão mais profunda da própria doença" (ABBOTT; BARBOSA, 2015, p. 1). Desse modo, buscou-se implementar um exemplo prático da tecnologia como peça fundamental para uma saúde mais precisa, avançada e inclusiva.

A tecnologia pode ser usada para identificar pessoas em risco de desenvolver diabetes, mas não é um substituto para o parecer médico. Um possível paciente ou indivíduo com sintomas e parâmetros característicos do problema de saúde em questão deve sempre consultar um médico para obter um diagnóstico e tratamento adequados.

Tendo isso em vista, criou-se um algoritmo de classificação, baseado no modelo “Random Forest”, o qual realiza um pré-diagnóstico de diabetes ao analisar parâmetros fornecidos pelo usuário. Tal implementação entra como mais um recurso aliado à gestão de saúde entre diversos já criados nesse mesmo intuito. Segundo Lima et al. (2021), "Em conclusão, este estudo mostrou que o algoritmo Random Forest foi capaz de revelar os aspectos mais importante para predição de óbito em pacientes idosos com COVID-19" (p. 445).

[1]Discente de Bacharelado em Ciência da Computação, IFSULDEMINAS – *Campus* Muzambinho. E-mail: antoniolimaaclf@gmail.com
[2]Discente de Bacharelado em Ciência da Computação, IFSULDEMINAS – *Campus* Muzambinho. E-mail: 12201001358@muz.ifsuldeminas.edu.br

2. MATERIAL E MÉTODOS

Na plataforma Kaggle, escolheu-se o dataset “Diabetes prediction dataset” para o modelo de treinamento do algoritmo de IA em questão. Esse conjunto de dados é formado por oito colunas, as quais representam os parâmetros observados em cada pessoa, e 10000 instâncias, sendo elas as observações coletadas, ou seja, cada indivíduo pesquisado. As colunas utilizadas são: gender, age, hypertension, heart_disease, smoking_history, Indice de massa corporal (IMC), HbA1c_level, blood_glucose_level. Outro ponto importante são os parâmetros coletados no dataset, tendo em vista que correspondem a aspectos relevantes no contexto de diagnóstico e avaliação da diabetes, conforme a Tabela 1 apresentada por Gross et al. (2002):

1	O rastreamento deve ser realizado em todos os indivíduos com sobrepeso (IMC ≥ 25 kg/m²*) e com fatores de risco adicionais: • Sedentarismo • Familiar em primeiro grau com diabetes melito • Grupos étnicos de maior risco (afro-americanos, latinos, índios, asiáticos, moradores das ilhas do Pacífico) • Mulheres com gestação prévia com feto com ≥ 4 kg ou com diagnóstico de DM gestacional • Hipertensão arterial sistêmica (≥ 140/90 mmHg ou uso de anti-hipertensivo) • Colesterol HDL ≤ 35 mg/dL e/ou triglicerídeos ≥ 250 mg/dL • Mulheres com síndrome dos ovários policísticos • HbA1c ≥ 5,7%, TDG ou GJA em exame prévio • Outras condições clínicas associadas à resistência insulínica (ex.: obesidade mórbida, acantose nigricante) • História de doença cardiovascular
2	Na ausência dos critérios acima, o rastreamento do DM2 deve iniciar a partir dos 45 anos
3	Se os resultados forem normais, o rastreamento deve ser repetido a cada três anos, considerando maior frequência dependendo dos fatores de risco iniciais

IMC: índice de massa corporal; TDG: tolerância diminuída à glicose; GJA: glicemia de jejum alterada; HbA1c: hemoglobina glicada.
* O IMC de risco pode ser menor em alguns grupos étnicos.

Zubizarreta et al. (2017) descobriram que o tabagismo agrava os efeitos da diabetes. Para tornar o código mais eficiente, os dados foram convertidos de palavras para números. Isso permitiu que a influência do tabagismo no diagnóstico de diabetes fosse quantificada. As observações "sem informações" ou "nunca" receberam um peso de zero, enquanto "já fumou" e "não fuma atualmente" recebem um peso de um.

Além disso, escolheu-se uma proporção de 30% para teste do algoritmo e a instalação do modelo com nove árvores de classificação. A taxa definida auxiliou a otimizar o desempenho do código, tendo em vista que por se tratar de um conjunto de dados com parâmetros diversos e com muitas observações (dez mil), treinar o modelo com 70% das informações obtidas garantiu maior acurácia.

A acurácia é uma métrica fundamental para avaliar o desempenho de um modelo de Random Forest, a qual mede a taxa de previsões corretas feitas pelo modelo em relação ao número total de amostras de teste. Tal parâmetro elevado indica que o modelo de Random Forest está

fazendo previsões precisas e confiáveis. Isso é crucial em muitos cenários, como detecção de fraudes, diagnóstico médico e classificação de texto. Com uma alta acurácia, é possível tomar decisões mais informadas com base nas previsões do modelo, aumentando a confiança na sua aplicabilidade prática.

3. RESULTADOS E DISCUSSÃO

Quanto aos resultados, obteve-se uma acurácia estimada em cerca de 96,9%, o que denota certa confiabilidade no modelo para predizer sobre uma possível diabetes. Nesse sentido, é válido ressaltar a versatilidade e tolerância com diferentes tipos de variáveis, mas também as condições negativas, como a dificuldade de operações com outliers, os quais se fazem presentes em conjuntos de dados sobre diabetes. Segundo Maniruzzaman et al. (2018), "These classifiers cannot correctly classify diabetic patients when the data contains missing values or has outliers, and therefore, when the machine learning-based classifiers are used for risk stratification[...]." Diante disso, é pertinente citar o fato da classificação ser na forma binária e não probabilística, o que reduz a estratificação de risco, compreendida como processo de categorizar indivíduos ou grupos em diferentes níveis de risco, supracitada.

Os níveis de glicose, tabagismo, idade e doenças cardíacas são fatores importantes para prever o diabetes. Esses fatores podem ser usados para criar modelos precisos e confiáveis usando o algoritmo Random Forest. O modelo pode ser usado para identificar pessoas em risco de diabetes, fornecendo insights valiosos para o diagnóstico precoce, monitoramento contínuo e intervenção personalizada. A adiposidade abdominal aumentada em indivíduos diabéticos também está associada à mortalidade por doenças cardiovasculares e pior controle metabólico da doença.

O diabetes é uma doença complexa causada por uma variedade de fatores, incluindo genética, estilo de vida e ambiente. Compreender os mecanismos subjacentes à doença e os diferentes tipos de diabetes é essencial para desenvolver modelos de previsão eficazes. Esses modelos podem ajudar a identificar pessoas em risco de diabetes, fornecendo insights valiosos para o diagnóstico precoce, monitoramento contínuo e intervenção personalizada.

4. CONCLUSÃO

Em resumo, este estudo explorou o uso do algoritmo Random Forest para prever a diabetes com base em parâmetros clínicos relevantes, como glicose, tabagismo, idade e doenças cardíacas. Os resultados obtidos mostraram que o Random Forest é uma abordagem promissora para a previsão da diabetes, apresentando uma acurácia satisfatória e fornecendo insights valiosos sobre os fatores de risco associados à doença. A inclusão desses parâmetros no modelo de previsão permitiu

uma estratificação de risco mais precisa e personalizada, auxiliando na identificação de indivíduos em maior risco de desenvolver a doença. No entanto, é importante ressaltar que este estudo apresenta algumas limitações. Por exemplo, a generalização dos resultados pode ser restrita ao contexto do conjunto de dados utilizado e à população estudada. Além disso, outros fatores relevantes para a diabetes podem não ter sido considerados neste estudo, e a inclusão de dados adicionais, como marcadores genéticos ou informações sobre estilo de vida, poderia fornecer uma visão mais abrangente e precisa.

O uso do algoritmo Random Forest mostrou-se promissor na previsão da diabetes. O aprimoramento contínuo dos modelos de previsão e a incorporação de mais informações clínicas podem ter um impacto significativo na prevenção, diagnóstico e tratamento da diabetes. Antunes (2009) define tecnologia como "a procura dos resultados úteis da investigação científica e as consequências que lhes estão associadas". O uso do algoritmo Random Forest é um exemplo de como a tecnologia pode ser usada para melhorar a saúde humana.

REFERÊNCIAS

ABBOTT, Patricia A.; BARBOSA, Sayonara FF. Usando tecnologia da informação e mobilização social para combater doenças. **Acta Paulista de Enfermagem**, v. 28, p. 1-1, 2015.

ANTUNES, João Lobo. Medicina e tecnologia. **JANUS 2009: Aliança de civilizações: um caminho possível?**, 2009.

FERREIRA, Celma Lúcia Rocha Alves; FERREIRA, Márcia Gonçalves. Características epidemiológicas de pacientes diabéticos da rede pública de saúde: análise a partir do sistema HiperDia. **Arquivos Brasileiros de Endocrinologia & Metabologia**, v. 53, p. 80-86, 2009.

GROSS, Jorge L. et al. Diabetes melito: diagnóstico, classificação e avaliação do controle glicêmico. **Arquivos Brasileiros de Endocrinologia & Metabologia**, v. 46, p. 16-26, 2002.

LIMA, Tiago Pessoa Ferreira et al. Previsão de óbito e importância de características clínicas em idosos com COVID-19 utilizando o Algoritmo Random Forest. **Revista Brasileira de Saúde Materno Infantil**, v. 21, p. 445-451, 2021.

MANIRUZZAMAN, Md et al. Accurate diabetes risk stratification using machine learning: role of missing value and outliers. **Journal of medical systems**, v. 42, p. 1-17, 2018.

ZUBIZARRETA, Marco López et al. Tabaco y diabetes: relevancia clínica y abordaje de la deshabituación tabáquica en pacientes con diabetes. **Endocrinologia, Diabetes y Nutrición**, v. 64, n. 4, p. 221-231, 2017.