**MODELO DE ÁRVORE DE DECISÃO BASEADO EM BIOINFORMÁTICA PARA DETECÇÃO DE *DIABETES MELLITUS* TIPO 2**

WESLEY FELIPE CAPUCHO1 e EDERSON CICHACZEWSKI2

1Graduando em Engenharia Biomédica, UNINTER, Cruzeiro-SP, wesleycapucho@outlook.com;

2MSc., Engenheiro de Computação, Professor Orientador, UNINTER, Curitiba - PR, ederson.c@uninter.com

Apresentado no

Congresso Técnico Científico da Engenharia e da Agronomia – CONTECC

8 a 11 de agosto de 2023

**RESUMO**:

O objetivo deste estudo é desenvolver e testar um modelo de Árvore de Decisão para o diagnóstico do Diabetes Mellitus Tipo 2. Essa doença crônica requer uma detecção rápida para prevenir complicações e aumentar a qualidade de vida dos pacientes. Porém, há uma carência de modelos de aprendizado de máquina adequados para esse propósito na literatura, por esse motivo, este estudo contribui para suprir essa necessidade, criando um modelo que utiliza dados demográficos e clínicos. As etapas do estudo incluíram o treinamento do modelo, a separação dos dados em treinamento e teste, a coleta e o pré-processamento dos dados, e a avaliação do desempenho do modelo. Os resultados indicaram que o modelo diagnosticou o Diabetes Mellitus Tipo 2 com cerca de 80% de acurácia. A matriz de confusão mostrou que o modelo classificou corretamente os casos, com altos valores de verdadeiros positivos e verdadeiros negativos. A conclusão é que o modelo de Árvore de Decisão pode auxiliar na detecção precoce e no tratamento adequado de pacientes com Diabetes Mellitus Tipo 2, oferecendo uma forma precisa e eficiente de reconhecer doenças crônicas.

**PALAVRAS-CHAVE:** Inteligência Artificial, Doença Metabólica, Machine Learning, Saúde Pública, Algoritmo

**PROPOSAL OF A BIOINFORMATICS-BASED DECISION TREE MODEL FOR DETECTION OF TYPE 2 DIABETES MELLITUS**

**ABSTRACT**: This study aims to develop and test a Decision Tree model for the diagnosis of Type 2 Diabetes Mellitus. This chronic disease requires quick detection to prevent complications and improve the quality of life of the patients. However, there is a lack of suitable machine learning models for this purpose in the literature. Therefore, this study contributes to filling this gap, creating a model that uses demographic and clinical data. The steps of the study included the training of the model, the splitting of the data into training and testing sets, the collection and pre-processing of the data, and the evaluation of the model’s performance. The results showed that the model diagnosed Type 2 Diabetes Mellitus with about 80% accuracy. The confusion matrix revealed that the model correctly classified the cases, with high values of true positives and true negatives. The conclusion is that the Decision Tree model can assist in the early detection and proper treatment of patients with Type 2 Diabetes Mellitus, providing a precise and efficient way of recognizing chronic diseases.

**KEYWORDS:** Artificial Intelligence, Metabolic Disease, Machine Learning, Public Health, Algorithm.

**INTRODUÇÃO**

Milhões de pessoas em todo o mundo sofre de Diabetes Mellitus Tipo 2 (DMT2), uma doença crônica complexa caracterizada por um desequilíbrio nos níveis de glicose no sangue (Dicow, 2015). A bioinformática tem sido fundamental para a análise de dados biológicos, permitindo a integração e análise de abundantes dados genômicos, transcriptômicos, proteômicos e clínicos por meio de técnicas avançadas (Cassanego, 2020). O aprendizado de máquina é amplamente utilizado em áreas como reconhecimento de imagem, processamento de linguagem natural e medicina (Esteva et al., 2017). No campo médico, o aprendizado de máquina tem se mostrado útil para melhorar o diagnóstico e o tratamento (Rajkomar et al., 2018).

Nesse contexto, o objetivo deste estudo é criar e avaliar um modelo baseado em árvore de decisão para identificar com precisão o Diabetes Mellitus Tipo 2. A técnica envolve a utilização de dados demográficos e clínicos de pacientes que atendam a critérios de seleção específicos. Foram coletados detalhes como idade, sexo, índice de massa corporal, pressão arterial e antecedentes de diabetes de uma amostra representativa da população. Os resultados mostraram que o modelo baseado em árvore de decisão conseguiu detectar o Diabetes Mellitus Tipo 2 com uma precisão de aproximadamente 80%. Esses resultados demonstram que o modelo proposto pode ser uma ferramenta útil para a detecção precoce e o encaminhamento adequado de pacientes com Diabetes Mellitus Tipo 2.

**MATERIAL E MÉTODOS**

 Neste estudo, os dados foram pré-processados, corrigindo valores ausentes, removendo discrepâncias e normalizando atributos numéricos. A análise exploratória investigou a relação entre idade e nível de glicose em três grupos: homens, mulheres não grávidas e mulheres grávidas. O modelo de detecção de Diabetes Mellitus Tipo 2 foi construído com o algoritmo de árvore de decisão. Os dados foram divididos em conjuntos de treinamento e teste, e o conjunto de treinamento foi usado para treinar o modelo usando validação cruzada para ajustar os parâmetros do modelo e evitar reajuste.

 A árvore de decisão é uma técnica popular em aprendizado de máquina, conhecida por sua interpretabilidade e facilidade de uso. Ela é amplamente utilizada em uma variedade de aplicações, desde classificação e regressão até detecção de anomalias e tomada de decisão (Witten, 2018). A principal vantagem da árvore de decisão é sua capacidade de aprender e representar relações complexas entre variáveis independentes e dependentes. Em resumo, a árvore de decisão é um poderoso algoritmo que oferece uma abordagem intuitiva e eficiente para tomada de decisões e classificação de dados.

 O banco de dados utilizado neste estudo é o “*Pima Indians Diabetes Database*”, disponível na plataforma Kaggle. Ele contém informações sobre DMT2 sendo composto por variáveis como número de gestações, nível de glicose e índice de massa corporal. É importante para o desenvolvimento e avaliação de modelos preditivos para detecção de diabetes, ao representar uma amostra real de pacientes. Neste estudo, o banco de dados foi pré-processado e filtrado para focar na detecção de DMT2 em diferentes subgrupos.

**Critérios de inclusão de dados**

 Foram considerados pacientes do sexo masculino, feminino não grávidas e feminino grávidas, essa seleção foi realizada para analisar e comparar os dados entre os diferentes grupos, considerando as possíveis diferenças fisiológicas e clínicas associadas ao diabetes.

**Pré-processamento dos dados**

 Verificou-se a presença de valores ausentes e, quando necessário, adotou-se uma estratégia para preenchê-los ou removê-los, garantindo a integridade dos dados.

 Identificou-se a presença de valores discrepantes nos atributos e, quando apropriado, esses valores foram tratados ou excluídos do conjunto de dados.

 Alguns atributos numéricos, como nível de glicose, pressão sanguínea, índice de massa corporal, entre outros, foram normalizados para garantir que todas as variáveis tenham a mesma escala. Essa etapa é importante para evitar que atributos com maiores magnitudes dominem o processo de treinamento.

**Treinamento do modelo**

 Neste estudo, foram selecionados critérios e parâmetros com base em práticas comuns e conhecimentos especializados em bioinformática e aprendizado de máquina para criar um modelo preciso de detecção da Diabetes Mellitus Tipo 2. Utilizou-se uma árvore de decisão com o critério de divisão Gini e uma profundidade máxima definida para evitar *overfitting*. A técnica de validação cruzada foi aplicada durante o treinamento para avaliar o desempenho do modelo de forma mais robusta e reduzir o viés dos dados aleatórios. A seleção desses parâmetros visava obter um modelo de alta precisão, controlando sua complexidade e selecionando características relevantes para a detecção da doença.

 A aplicação da validação cruzada fornece uma avaliação mais confiável do modelo, permitindo uma estimativa precisa de seu desempenho em dados não vistos anteriormente. Em resumo, a seleção cuidadosa e a otimização desses parâmetros contribuem para o desenvolvimento de um modelo robusto e eficaz na detecção da Diabetes Mellitus Tipo 2, oferecendo uma ferramenta promissora para auxiliar na identificação precoce e no tratamento adequado dessa condição médica. A matriz de confusão é uma ferramenta importante na avaliação e análise do desempenho de um modelo de classificação.

**Validação da precisão do modelo**

 A matriz de confusão apresenta uma visão detalhada das classificações corretas e incorretas, permitindo a análise de métricas como precisão, recall e taxa de acerto (Powers, 2011). É composta por quatro elementos principais: verdadeiros positivos (VP), falsos positivos (FP), verdadeiros negativos (VN) e falsos negativos (FN). A precisão é a proporção de observações corretamente classificadas como positivas em relação ao total de observações classificadas como positivas. A sensibilidade é a proporção de observações corretamente classificadas como positivas em relação ao total de observações reais positivas. Já a especificidade é a proporção de observações corretamente classificadas como negativas em relação ao total de observações reais negativas.

**RESULTADOS E DISCUSSÃO**

Os gráficos de dispersão forneceram informações sobre a distribuição e separabilidade das classes. Eles permitem visualizar a relação entre as características dos exemplos e suas respectivas classes, auxiliando na compreensão do desempenho e da capacidade discriminativa do modelo. Cada exemplo é representado por um ponto no espaço, onde a posição no gráfico é determinada pelos valores das duas variáveis. Ao observar os gráficos de dispersão, podemos identificar a tendência dos pontos de diferentes classes a se agruparem ou se sobreporem em determinadas regiões do espaço. Isso fornece informações sobre a separabilidade das classes e indica se o modelo terá facilidade ou dificuldade em classificar corretamente os exemplos. Na Figura 1 ilustra-se o gráfico de dispersão para homens: o gráfico de dispersão mostra a relação entre idade e nível de glicose para homens.

Figura 1: Gráfico de dispersão contendo os casos de DMT2 em homens.



Fonte:Os autores

A Figura 2 demonstra o gráfico de dispersão para mulheres grávidas e não grávidas. Estes gráficos de dispersão mostram a relação entre idade e níveis de glicose. Os pontos vermelhos representam os pontos de dados, o gráfico fornece informações sobre os níveis de glicose mg/L em mulheres não grávidas em diferentes faixas etárias.

Figura 2: Gráfico de dispersão contendo os casos de DMT2 em mulheres grávidas (esq.) e não grávidas (dir.).



Fonte: Os autores.

A Figura 3 (esq.) ilustra que a precisão do modelo foi avaliada com base na matriz de confusão e nos gráficos de dispersão. A matriz de confusão forneceu uma visão detalhada das classificações corretas e incorretas do modelo, enquanto os gráficos de dispersão ajudaram a entender a separação das classes. Com base nessa análise, foram tomadas decisões sobre a eficácia do modelo, ajustes ou melhorias necessárias para aprimorar sua precisão. A matriz de confusão é uma representação tabular que mostra o desempenho do modelo de classificação, comparando suas previsões com as classes reais e inclui métricas como verdadeiro positivo, verdadeiro negativo, falso positivo e falso negativo.

A Figura 3 (dir.) ilustra o resultado da análise de desempenho, contemplando a avaliação da precisão do modelo de árvore de decisão, que é uma métrica utilizada para avaliar o desempenho do modelo na tarefa de prever corretamente a presença ou ausência de diabetes. É calculada comparando as previsões realizadas pelo modelo com as classes reais presentes no conjunto de testes. Neste estudo, a precisão do modelo de árvore de decisão foi calculada que o modelo acertou aproximadamente 75% a 78% das previsões realizadas. Essa faixa de precisão indica que o modelo apresenta um desempenho geralmente correto em suas previsões, embora possa haver variações em relação à concretude em diferentes amostras de teste. A sensibilidade, também conhecida como taxa de verdadeiros positivos, é calculada dividindo-se o número de verdadeiros positivos pelo somatório dos verdadeiros positivos e falsos negativos. Ela representa a capacidade do modelo de identificar corretamente os casos positivos. Neste modelo, a sensibilidade foi de 70%. A especificidade, ou taxa de verdadeiros negativos, é calculada dividindo-se o número de verdadeiros negativos pelo somatório dos verdadeiros negativos e falsos positivos. Ela indica a capacidade do modelo de identificar corretamente os casos negativos. Neste caso, a especificidade foi de 85%. O valor preditivo positivo é calculado dividindo-se o número de verdadeiros positivos pelo somatório dos verdadeiros positivos e falsos positivos. Ele representa a probabilidade de um resultado positivo do modelo ser realmente positivo, no modelo em questão, o valor preditivo positivo foi de 76%. No contexto deste estudo, a precisão do modelo de árvore de decisão indica a sua capacidade de identificar os casos de diabetes baseado no banco de dados utilizado e, fornecendo uma medida da confiabilidade das previsões realizadas.

Figura 3: A matriz de confusão (esq.) e a análise de desempenho (dir.) com os resultados do teste realizado.

 

Fonte: Os autores.

**CONCLUSÃO**

 Com base nos resultados obtidos no modelo de árvore de decisão e na análise exploratória dos dados, pode-se concluir que o trabalho realizado relacionado ao diagnóstico de Diabetes Mellitus Tipo 2 apresentou resultados promissores. Utilizando dados demográficos e clínicos, o modelo desenvolvido conseguiu prever com uma precisão satisfatória a presença ou ausência de diabetes e apesar dos resultados promissores, é importante ressaltar que o modelo desenvolvido é uma ferramenta auxiliar no diagnóstico e não substitui a avaliação médica especializada. No geral, o trabalho demonstrou a viabilidade do uso de técnicas de aprendizado de máquina na detecção de Diabetes Mellitus Tipo 2 com base em informações demográficas e clínicas. Com ajustes e refinamentos adicionais, esse modelo pode se tornar uma ferramenta valiosa para auxiliar médicos e profissionais de saúde no diagnóstico precoce e no tratamento eficaz dessa doença crônica.

**REFERÊNCIAS**

ESTEVA, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature, v. 542, n. 7639, p. 115-118, 2017.

POWERS, D. M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, v. 2, n. 1, p. 37-63, 2011.

RAJKOMAR, A. et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine, v. 1, n. 1, p. 18, 2018.

WITTEN, I. H.; Frank, E.; Hall, M. A. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2018.

Cassanego, Samuel Rech. O que é a bioinformática? Universidade Federal de Santa Maria. Programa de Educação Tutorial (PET) - Sistemas de Informação, 2020. Disponível em: https://www.ufsm.br/pet/sistemas-de-informacao/2020/12/03/o-que-e-a-bioinformatica. Acesso em: 12 jun. 2023.