# Predição da Doença Arterial Coronariana (DAC) Usando Aprendizado de Máquina

*Aplicações de Algoritmos de Inteligência Artificial na Identificação Precoce de Riscos Cardíacos

1st Kauan Divino Pouso Mariano
*Instituto de Informática*
*Universidade Federal de Goiás*
Goiânia, Goiás
kauan@discente.ufg.br

***Abstract*—Este relatório técnico aborda a previsão da doença arterial coronariana (DAC) utilizando algoritmos de aprendizado de máquina. Através de uma análise exploratória, foram identificadas características clínicas significativas do conjunto de dados Z-Alizadeh Sani, que engloba informações médicas de 303 pacientes. Diversos modelos de aprendizado de máquina, incluindo Random Forest, Gradient Boosting, SVM e Redes Neurais, foram testados, revelando potenciais promissores na previsão da DAC. Características como histórico de CVA, idade e indicadores do ECG mostraram-se particularmente relevantes. Apesar da eficácia demonstrada pelos modelos, enfatiza-se a necessidade de uma abordagem complementar ao discernimento clínico, reconhecendo a singularidade e complexidade inerente à predição de patologias médicas.**

***Index Terms*—Doença Arterial Coronariana, Aprendizado de Máquina, Modelagem Preditiva, Algoritmos de Previsão**

## I. Introdução

Doenças cardíacas representam uma das principais causas de óbito globalmente, compreendendo um conjunto de condições que comprometem o sistema cardiovascular, abrangendo primordialmente o coração e os vasos sanguíneos. A importância de um diagnóstico precoce e acurado ressalta a necessidade de métodos eficientes de detecção e prevenção. A evolução na área da tecnologia da informação e da ciência de dados tem possibilitado a utilização de registros médicos como ferramenta no auxílio à detecção e prevenção de patologias cardíacas.

O conjunto de dados denominado Z-Alizadeh Sani, disponibilizado na plataforma Kaggle, apresenta informações médicas e laboratoriais de pacientes, com a finalidade de assegurar a presença ou ausência de doença arterial coronariana (DAC). Este dataset destaca-se pela sua riqueza e diversidade, incorporando características que vão desde dados demográficos até resultados de exames específicos.

O escopo deste relatório técnico engloba uma análise minuciosa do conjunto de dados Z-Alizadeh Sani, com foco em:

- Análise Exploratória de Dados (EDA): Esta seção busca compreender a distribuição das características, identificar padrões e extrair informações relevantes que possam contribuir para a detecção precoce da DAC.
- Modelagem Preditiva: A intenção é conceber e avaliar modelos de aprendizado de máquina capazes de antecipar a presença de DAC com base nas características apresentadas.
- Análise de Importância de Características: Esta parte objetiva discernir as características de maior relevância na previsão da DAC, fornecendo diretrizes para futuras pesquisas médicas e enfatizando determinados exames ou dados clínicos.

Análise de Importância de Características: Esta parte objetiva discernir as características de maior relevância na previsão da DAC, fornecendo diretrizes para futuras pesquisas médicas e enfatizando determinados exames ou dados clínicos. Com esta análise, pretende-se não somente estabelecer modelos preditivos de alta eficácia, mas também aprofundar o entendimento sobre os elementos associados à doença arterial coronariana, visando otimizar estratégias de diagnóstico e terapêuticas.

## II. Metodologia

O conjunto de dados empregado originou-se do Kaggle, uma plataforma de repositório online dedicada a conjuntos de dados públicos destinados à análise e modelagem. Especificamente, o conjunto de dados Z-Alizadeh Sani engloba informações médicas detalhadas referentes a 303 pacientes, com o propósito central de identificar a presença de doença arterial coronariana (DAC).

No âmbito do pré-processamento, o conjunto de dados foi submetido a diversas etapas para assegurar sua qualidade e pertinência. Inicialmente, foi realizada uma verificação para identificar possíveis valores ausentes, garantindo a completude das entradas. Subsequentemente, características categorizadas, como sexo e histórico familiar, foram convertidas em formatos numéricos mediante a técnica de codificação one-hot, preparando-as para serem compatíveis com modelos de aprendizado de máquina. Por fim, o processo de normalização foi aplicado aos dados, utilizando o StandardScaler do pacote scikit-learn, com o intuito de equalizar a importância inicial de todas as características ao serem introduzidas nos algoritmos de aprendizado.

Em relação à Análise Exploratória de Dados (EDA), esta se mostrou fundamental para a compreensão da natureza e distribuição das informações presentes. Durante esta fase, histogramas e gráficos de densidade foram elaborados para visualizar a distribuição de cada característica. Matrizes de correlação foram estabelecidas para discernir as inter-relações entre diferentes características e a influência destas sobre a DAC. Adicionalmente, técnicas como boxplots foram aplicadas visando analisar a distribuição das características em contraste com os resultados (presença ou ausência de DAC).

No contexto da modelagem e avaliação, após a EDA, procedeu-se à etapa de modelagem. Primeiramente, os dados foram segmentados em conjuntos de treinamento e teste, assegurando uma distribuição equitativa de casos positivos e negativos em ambos os segmentos. Diversos modelos de aprendizado de máquina foram testados, incluindo Regressão Logística, Random Forest, Gradient Boosting, Máquinas de Vetores de Suporte (SVM) e Redes Neurais. Posteriormente, cada modelo foi treinado com o conjunto de treinamento e avaliado com o conjunto de teste. Para aqueles modelos que demonstraram resultados promissores, uma busca de grade foi executada visando otimizar hiperparâmetros e potencializar seu desempenho. Finalmente, os modelos foram avaliados com base em métricas diversificadas, como acurácia, matriz de confusão, precisão, recall e pontuação F1.

## III. Resultados

A análise e predição de doenças cardíacas foi realizada utilizando um conjunto de dados específico. A etapa inicial consistiu em uma análise exploratória dos dados, que proporcionou insights valiosos acerca do conjunto em questão, permitindo assim compreender a distribuição das características e identificar correlações potenciais, preparando os dados para subsequente modelagem.

Na fase de análise exploratória, foi observado que características como CVA, Age e St Depression possuíam relevância considerável na detecção de doenças cardíacas. Estas características exibiram uma correlação acentuada com a presença da condição. Adicionalmente, foi constatado que atributos específicos, como Sex e DM, apresentavam distribuições divergentes entre os grupos de pacientes diagnosticados com e sem a condição, ressaltando a importância destas características no contexto diagnóstico.

Na etapa de modelagem, diversos algoritmos de aprendizado de máquina foram testados a fim de determinar o mais adequado para a predição da presença de doenças cardíacas. O modelo Random Forest, baseado em árvores de decisão, atingiu uma acurácia de 86.89%. Este modelo é reconhecido por sua aptidão em processar grandes conjuntos de dados, minimizar overfitting e discernir a relevância das características. Por outro lado, o algoritmo Gradient Boosting, que constrói um modelo preditivo em estágios, alcançou a acurácia mais elevada do estudo, com 90.16%. O algoritmo Support Vector Machines (SVM), focado em separar os dados em classes, obteve uma acurácia de 88.52%. Por fim, as Redes Neurais, com sua estrutura inspirada na biologia neural humana, atingiram uma acurácia de 86.89

A relevância das características foi avaliada para determinar quais variáveis impactavam mais significativamente na predição da doença cardíaca. Utilizando a técnica de importância de características do modelo Random Forest, observou-se que 'CVA' emergiu como a característica mais determinante, sugerindo que históricos de acidentes vasculares cerebrais são indicadores robustos da presença da condição. A 'Idade' posicionou-se como a segunda característica mais influente, reiterando a correlação entre a idade avançada e a probabilidade de desenvolvimento de doenças cardíacas. 'ST Depression', uma métrica derivada do eletrocardiograma, também foi identificada como relevante. Outras características, incluindo 'Sexo', 'DM', 'Pressão Arterial', e 'Q Wave', ainda que de menor relevância relativa, contribuíram para a eficácia do modelo, elucidando a multifacetada natureza das doenças cardíacas e os diversos fatores envolvidos.

## IV. Discussões

A precisão na previsão da doença arterial coronariana (DAC) por meio de algoritmos de aprendizado de máquina sugere uma transformação significativa nas práticas da medicina contemporânea. Este estudo explorou tal potencial, enfatizando como determinadas características clínicas podem indicar a presença de DAC.

A relevância atribuída ao CVA (Acidente Vascular Cerebral) na previsão da doença é digna de nota. A correlação entre um histórico de acidente vascular cerebral e a presença de DAC não se restringe a um mero indicativo estatístico, mas traz consigo implicações clínicas profundas. Tal relação pode apontar para uma possível conexão previamente subestimada entre ambas as condições, indicando que indivíduos com antecedentes de CVA poderiam beneficiar-se de uma avaliação clínica mais rigorosa quanto à DAC.

Conforme antecipado, a idade emergiu como uma característica de relevância determinante. Tal observação corrobora a compreensão médica de que o risco associado a diversas patologias, incluindo a DAC, se amplifica com o avanço da idade. A quantificação dessa relação pelo algoritmo oferece uma perspectiva objetiva da importância relativa da idade, especialmente quando contrastada com outras características.

A métrica "Depressão ST", derivada do eletrocardiograma, foi igualmente identificada como uma característica essencial. Essa constatação alinha-se à literatura médica existente, uma vez que determinadas anormalidades no eletrocardiograma, como a Depressão ST, são associadas a complicações cardíacas. A habilidade do modelo em salientar tal aspecto reforça a autenticidade e pertinência de suas previsões.

No entanto, ao avaliar a aplicabilidade dos modelos propostos em contextos práticos, torna-se imperativo adotar um enfoque equilibrado. A despeito da precisão evidenciada pelos modelos, é imperativo que estes atuem como complementos, e não substitutos, ao discernimento clínico. Um modelo pode analisar variáveis e formular previsões baseadas em dados, mas a interpretação e aplicação destas em contextos clínicos

reais exigem a expertise inerente ao profissional de saúde. Adicionalmente, é fundamental considerar a singularidade de cada paciente, reconhecendo que variáveis externas ao modelo podem influenciar a detecção da DAC.

Em síntese, enquanto os modelos demonstram ser ferramentas valiosas no auxílio à detecção da DAC, sua aplicação clínica necessita ser abordada com prudência e discernimento.

## V. Conclusão

O estudo em foco abordou de maneira aprofundada a desafiadora tarefa de prever a doença arterial coronariana (DAC) por meio de algoritmos de aprendizado de máquina. A análise exploratória do conjunto de dados revelou a vastidão de informações nele contidas, englobando desde características demográficas até métricas elaboradas derivadas de exames laboratoriais e eletrocardiogramas.

A implementação de diversos modelos de aprendizado de máquina evidenciou que, mediante calibração apropriada, é factível atingir níveis de precisão encorajadores na predição da DAC. A ênfase na relevância do histórico de CVA, idade e indicadores do ECG, como a Depressão ST, não somente corroborou a eficiência dos modelos propostos, mas também trouxe à tona insights clínicos de grande valor.

A investigação também enfatizou a complexidade subjacente à predição de patologias médicas. O organismo humano, como sistema complexo, e a DAC, caracterizada por sua natureza multifatorial, resistem à simplificação e à predição com base em indicadores isolados ou em grupos restritos de variáveis. Assim, a estratégia de modelagem empregada buscou uma perspectiva holística, contemplando um espectro amplo de características a fim de construir um retrato mais abrangente da condição.

A pesquisa reafirmou o potencial significativo do aprendizado de máquina no âmbito da medicina contemporânea, mas também destacou a necessidade de uma perspectiva equilibrada. Enquanto os modelos representam ferramentas potentes, estes devem ser vistos como complementares ao discernimento clínico e não como substitutos.

Antevendo desenvolvimentos futuros, e considerando a evolução contínua das tecnologias de aprendizado de máquina e o incremento na disponibilidade de dados, percebe-se um vasto horizonte para aprimoramentos e refinamentos. A incorporação e assimilação dessas ferramentas nas rotinas clínicas pode catalisar transformações no modo como a DAC é diagnosticada e tratada, culminando em melhores desfechos para os pacientes e um sistema de saúde otimizado e mais efetivo.

## References


[1] Classification of Coronary Artery Disease. Disponível em: https://www.kaggle.com/datasets/saeedeheydarian/classification-of-coronary-artery-disease. Acesso em: 21 set 2023.

[2] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. doi: 10.1023/A:1010933404324

[3] Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). The Annals of Statistics, 28(2), 337-407.

[4] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE international conference on computer vision, 1026-1034.

[6] Chollet, F. (2015). Keras. GitHub repository. Disponível em: https://github.com/fchollet/keras.