FACULDADE DE TECNOLOGIA DE BAURU

TECNOLOGIA EM BANCO DE DADOS

Previsão de Doença do Coração Usando Machine Learning

Equipe:
Luis Felipe Paludetto Silva

Bauru/SP
2025

Previsão de Doença do Coração Usando Machine Learning

Equipe:
Luis Felipe Paludetto Silva

Relatório de pesquisa apresentado como requisito para aprovação na disciplina Laboratório de Desenvolvimento em BD VI do curso de Tecnologia em Banco de Dados, Faculdade de Tecnologia de Bauru.

Profa. Dra. Patricia Bellin Ribeiro

Bauru/SP
2025

SUMÁRIO

RESUMO

Por meio de modelos de previsão baseados em Inteligência Artificial, com os dados corretos, tornou-se possível prever as mais diversas informações, desse modo, é possível comparar diversos tipos de modelos de previsão para problemas no coração, verificando sua precisão e agrupando os dados resultantes.

Palavras-chave: Inteligência Artificial. Previsão. Complicações Cardíacas.

ABSTRACT

Through predictive models based on Artificial Intelligence, with the correct data, it has become possible to predict various types of information. In this way, it is possible to compare different types of prediction models for heart problems, verifying their accuracy and grouping the resulting data.

Keywords: Artificial Intelligence. Prediction. Cardiac Complications.

1. INTRODUÇÃO

Com a tecnologia de Inteligência Artificial moderna, a habilidade de analisar dados e padrões se elevou, permitindo que pessoas montem modelos com finalidades muito variadas para tentar prever ou resolver os problemas mais diversos da humanidade.

Um dos problemas mais importantes da humanidade é a saúde, segundo Affonso (2023) em 2022 apenas no Brasil houve cerca de 400 mil mortos por problemas cardiovasculares, mas usando modelos de previsão por Inteligência Artificial pode ser possível identificar esses problemas antes que se tornem fatais.

Usando de conceitos de análise de dados, é possível agrupar as informações fornecidas e criar previsões baseadas em algoritmos criando padrões com alta porcentagem de acurácia, também simplificando todos os dados por meio da geração gráficos.

2. OBJETIVOS

Apresentar e comparar modelos, os quais usam de dados fornecidos, para tentar prever as complicações médicas relacionadas ao coração verificando sua acurácia e agrupando os dados de maneira simples e de entendimento intuitivo.

3. MATERIAIS E MÉTODOS

O projeto foi desenvolvido com o objetivo de construir um modelo preditivo de doença do coração com base em dados clínicos. Utilizou-se a ferramenta gratuita Google Colab com a linguagem Python (versão 3.11.12), junto das bibliotecas numpy, pandas, matplotlib.pyplot, seaborn, além de ferramentas de machine learning como train_test_split, accuracy_score, LogisticRegression, GaussianNB, svm, KNeighborsClassifier, DecisionTreeClassifier, RandomForestClassifier, xgb, e componentes de redes neurais com Sequential e Dense do Keras.

Empregou-se o conjunto de dados Heart Disease UCI Dataset, com 303 amostras e 14 atributos clínicos. Os dados foram organizados, limpos e explorados visualmente. Variáveis categóricas foram codificadas, e variáveis numéricas, padronizadas.

Os dados foram divididos em treino e teste (70% e 30%). Treinaram-se diferentes modelos preditivos supervisionados e redes neurais simples. As performances foram avaliadas por acurácia, matriz de confusão, F1-score e curva ROC. Aplicou-se validação cruzada para maior robustez dos resultados.

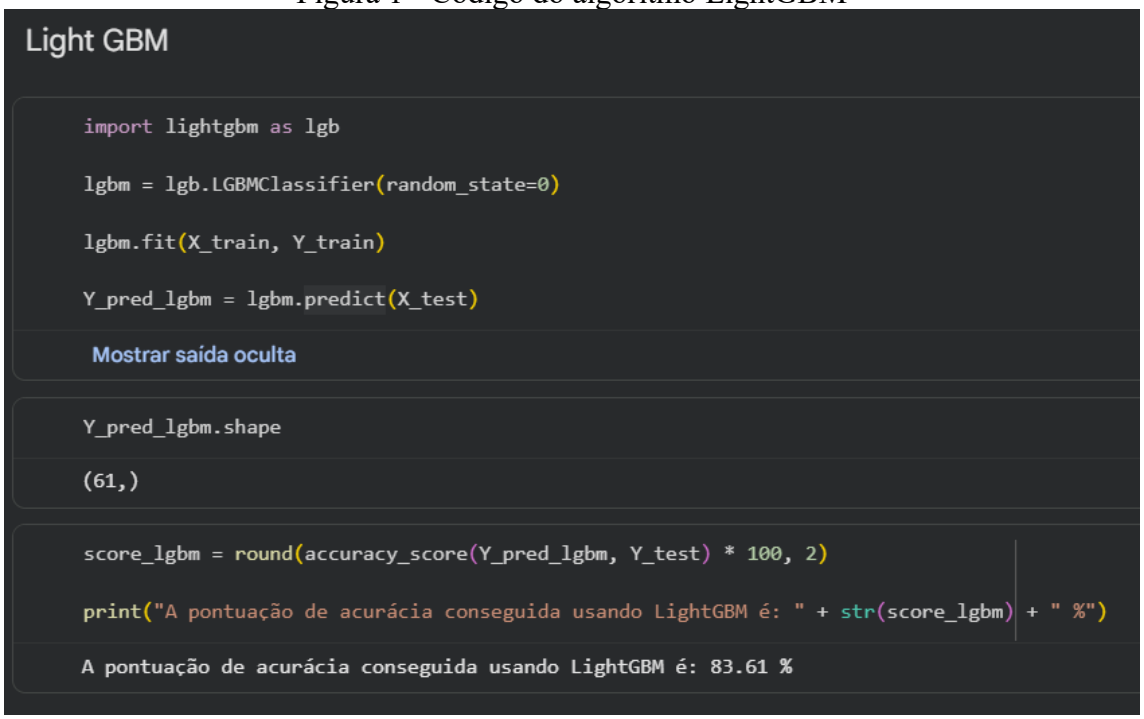
4. RESULTADOS E DISCUSSÕES

A aplicação dos diferentes modelos de aprendizado apresentou resultados variados em relação à acurácia, precisão e sensibilidade. Devido aos dados e configurações fornecidas para cada modelo, o resultado mais promissor foi o modelo de Random Forest, com ‘90,16’ porcento de precisão, e o mais impreciso foi o algoritmo de K-Nearest Neighbors, o qual obteve ‘67,21’ porcento de acurácia, com a possibilidade de ampliação de base de treinamento ou maior quantia de tempo de treino.

É possível notar que modelos mais simples como Logistic Regression e Naive Bayes, quando bem aplicados, podem ser eficazes em tarefas de classificação médica, trazendo o resultado de ‘85,25’ porcento para ambos.

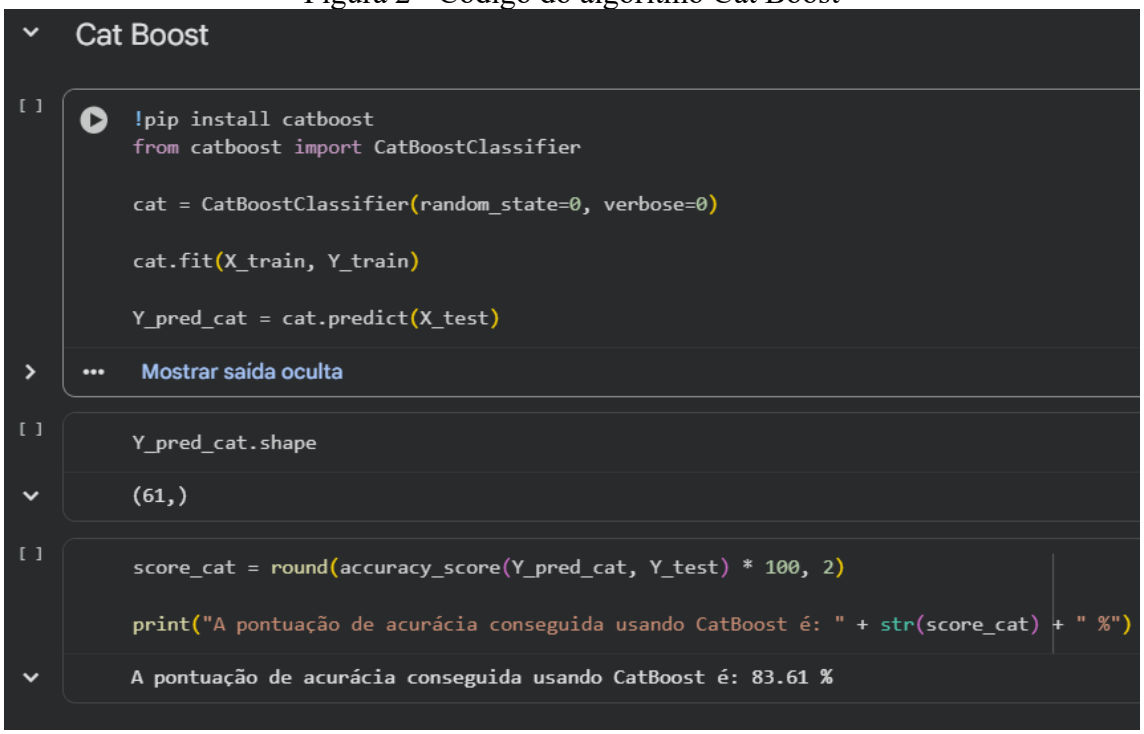
Além disso, foram incluídos no estudo os algoritmos LightGBM e CatBoost, adicionados por iniciativa própria, com o intuito de compará-los aos modelos utilizados no trabalho original. Ambos foram escolhidos por apresentarem fundamentos conceituais semelhantes aos modelos Logistic Regression e Naive Bayes, permitindo uma análise comparativa mais abrangente, assim como é mostrado na Figura 1 e Figura 2.

Figura 1 - Código do algoritmo LightGBM



Fonte: Elaborado pelo autor (2025)

Figura 2 - Código do algoritmo Cat Boost



Fonte: Elaborado pelo autor (2025)

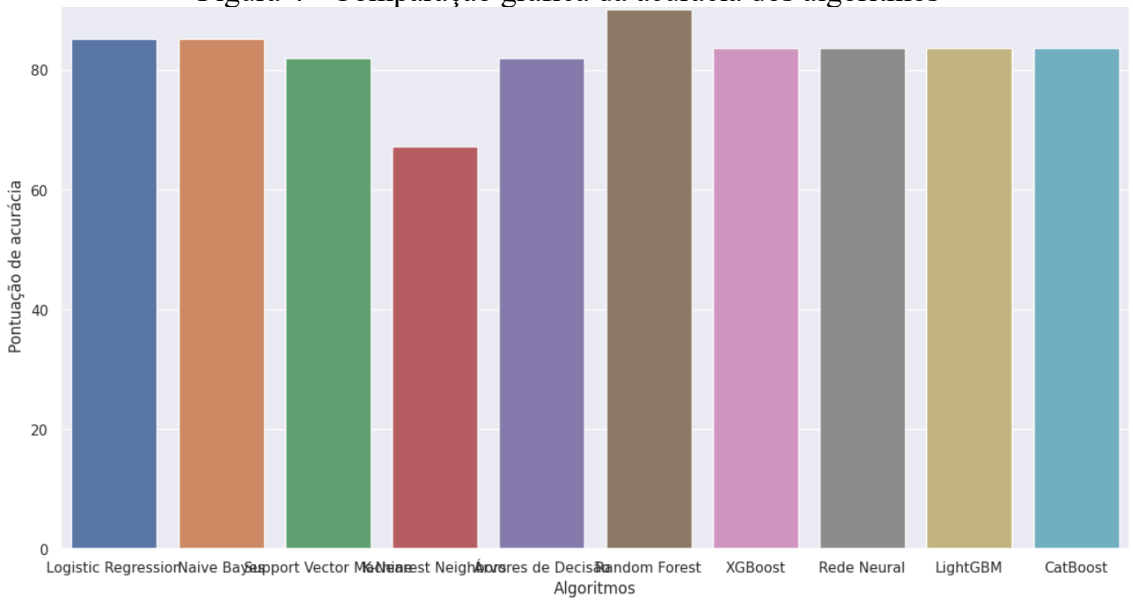
Porém, como o esperado, modelos mais complexos obtiveram bons resultados, com o XGBoost alcançando uma acurácia de ‘83,61%’, também se destacando. Já os modelos SVM, Árvore de Decisão e Rede Neural obtiveram acurácias semelhantes, com ‘81,97%’, indicando desempenho consistente, embora um pouco abaixo dos modelos citados anteriormente, melhor representado na comparação da Figura 3 e Figura 4.

Figura 3 - Comparação entre algoritmos

```
VI. Pontuação final de saída

scores = [score_lr,score_nb,score_svm,score_knn,score_dt,score_rf,score_xgb,score_nn,score_lgbm,score_cat]
algorithms = ["Logistic Regression","Naive Bayes","Support Vector Machine","K-Nearest Neighbors","Árvores de Decisão","Random Forest","XGBoost","Rede Neural","LightGBM","CatBoost"]

for i in range(len(algorithms)):
    print("A pontuação de acurácia conseguida usando "+algorithms[i]+" é: "+str(scores[i])+" %")

A pontuação de acurácia conseguida usando Logistic Regression é: 85.25 %
A pontuação de acurácia conseguida usando Naive Bayes é: 85.25 %
A pontuação de acurácia conseguida usando Support Vector Machine é: 81.97 %
A pontuação de acurácia conseguida usando K-Nearest Neighbors é: 67.21 %
A pontuação de acurácia conseguida usando Árvores de Decisão é: 81.97 %
A pontuação de acurácia conseguida usando Random Forest é: 90.16 %
A pontuação de acurácia conseguida usando XGBoost é: 83.61 %
A pontuação de acurácia conseguida usando Rede Neural é: 83.61 %
A pontuação de acurácia conseguida usando LightGBM é: 83.61 %
A pontuação de acurácia conseguida usando CatBoost é: 83.61 %
```

Fonte: Elaborado pelo autor (2025)

Figura 4 - Comparação gráfica da acurácia dos algoritmos



Fonte: Elaborado pelo autor (2025)

Mas é fundamental considerar não apenas métricas como acurácia, mas também índices como F1-score e sensibilidade, sobretudo em contextos em que falsos negativos podem ter consequências graves. A escolha do modelo ideal depende do equilíbrio entre desempenho técnico, interpretabilidade clínica e recursos computacionais disponíveis.

De forma geral, os resultados mostram que modelos de machine learning, especialmente Random Forest e algoritmos probabilísticos como Naive Bayes ou Regressão Logística, LightGBM ou CatBoost, podem ser bastante úteis na detecção de doenças cardíacas. Apesar disso, é necessário lembrar que esses sistemas devem ser utilizados apenas como apoio, e não usados para substituir uma análise clínica feita por profissionais da saúde.

5. CONCLUSÕES

Os resultados obtidos reforçam o potencial da Inteligência Artificial como ferramenta de apoio à medicina preventiva, especialmente na detecção precoce de doenças cardíacas. Entre os algoritmos testados, os modelos Random Forest, Logistic Regression e Naive Bayes, demonstraram maior acurácia e estabilidade. Mostrando que mesmo algoritmos simples, quando bem configurados, podem apresentar bons resultados.

A replicação desse projeto com bases de dados maiores, mais diversificadas e atualizadas pode contribuir para maior generalização dos resultados. Além disso, o uso de técnicas de seleção de atributos, ajuste fino de hiperparâmetros (como Grid Search) e

aplicação de modelos explicáveis (como SHAP ou LIME) são caminhos promissores para trabalhos futuros.

A integração entre ciência de dados, machine learning e saúde pública mostra-se cada vez mais necessária para antecipar riscos, reduzir custos com internações e melhorar a qualidade de vida da população.

6. REFERÊNCIAS

AFFONSO, Alexandre. Cerca de 400 mil pessoas morreram em 2022 no Brasil por problemas cardiovasculares. Biblioteca Virtual em Saúde, 2023. Disponível em: https://bvsms.saude.gov.br/cerca-de-400-mil-pessoas-morreram-em-2022-no-brasil-por-problemas-cardiovasculares/. Acesso em: 02 abr. 2025.

AHMAD, Ahmad Ayid. Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm. MPDI, 2023. Disponível em: https://www.mdpi.com/2075-4418/13/14/2392. Acesso em: 02 abr. 2025.

Documentação do Colab Enterprise. Disponível em: <https://cloud.google.com/colab/docs?hl=pt-br>. Acesso em: 9 maio. 2025

GOSAVI, Shreekant. Heart Disease Prediction using Machine Learning. Heart-Disease-Prediction-using-Machine-Learning, 2019. Disponível em: https://github.com/g-shreekant/Heart-Disease-Prediction-using-Machine-Learning/tree/master. Acesso em: 02 abr. 2025.

Heart_disease_prediction_original.Ipynb. Disponível em: <https://drive.google.com/file/d/1yls1VBZFT3POZX6j72tJ62tmNvfLN_P2/view?usp=sharing>. Acesso em: 27 jun. 2025.

NASHIF, Shamad. Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System. Scientific Research, 2018. Disponível em: https://www.scirp.org/journal/paperinformation?paperid=88650. Acesso em: 02 abr. 2025.