Análise do risco de diabetes tipo 2 em mulheres Pima (Akimel O’odham) a partir de dados clínicos históricos

Bruno Cerqueira Gianotti | 10721759

Daniel Fernandes Saraiva | 10381985

Gabrielle Solange Ferreira | 10414956

Reginaldo Rogério de Campos | 10743942

Sumário

Introdução

A diabetes mellitus tipo 2 é uma das principais doenças crônicas não transmissíveis no mundo, associada a altos índices de morbidade, mortalidade e custos em saúde pública. Entre as populações indígenas da América do Norte, destaca-se a comunidade **Pima (Akimel O'odham)**, localizada no estado do Arizona (EUA), que apresenta uma das maiores prevalências de diabetes tipo 2 já registradas globalmente.

Desde 1965, o **National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)** tem conduzido um estudo longitudinal sobre essa população, com o objetivo de compreender os fatores clínicos, genéticos e de estilo de vida que contribuem para o desenvolvimento da doença. Como fruto desse esforço, surgiu o **Pima Indians Diabetes Dataset**, que reúne informações de mulheres com 21 anos ou mais, contemplando variáveis como glicose, pressão arterial, índice de massa corporal, histórico de gravidez, entre outras.

A importância desse conjunto de dados vai além da análise epidemiológica: ele tornou-se um dos **datasets mais utilizados em ciência de dados e aprendizado de máquina**, servindo como referência para o desenvolvimento e validação de modelos preditivos em saúde. Assim, ele contribui não apenas para avanços acadêmicos, mas também para a formulação de estratégias de prevenção, diagnóstico precoce e políticas de saúde voltadas a populações vulneráveis.

Nesse contexto, este trabalho tem como objetivo analisar o Pima Indians Diabetes Database, disponibilizado atualmente em repositórios como o Kaggle, buscando compreender como variáveis clínicas e históricas se relacionam com o diagnóstico de diabetes tipo 2. O estudo também representa uma oportunidade de aplicar conceitos de ciência de dados, explorando padrões, correlações e potenciais fatores de risco.

Todo o código, dataset e versões deste projeto estão disponíveis no repositório GitHub: https://github.com/BrunoCerqueiraGianotti/projeto-diabetes-pima.

Premissas do projeto

Diante dessa relevância, este projeto parte do princípio de que a análise de dados clínicos e sociodemográficos pode fornecer insights valiosos para a predição e compreensão da diabetes tipo 2.

A premissa central consiste em que, por meio de técnicas de ciência de dados, é possível identificar padrões de risco, avaliar os fatores mais influentes no desenvolvimento da doença e construir modelos preditivos capazes de apoiar estratégias de prevenção e diagnóstico precoce.

Objetivo Geral

Desenvolver um projeto aplicado de ciência de dados utilizando o Pima Indians Diabetes Dataset, com o propósito de analisar fatores de risco e construir modelos preditivos capazes de estimar a probabilidade de ocorrência de diabetes tipo 2, contribuindo para a compreensão da doença e para estratégias de prevenção.

Objetivos Específicos

1. Realizar análise exploratória dos dados (EDA) para compreender a distribuição das variáveis, identificar outliers e tratar dados faltantes.
2. Investigar a correlação entre fatores clínicos/demográficos e a presença de diabetes, destacando os mais relevantes.
3. Aplicar técnicas de pré-processamento, como normalização e balanceamento de classes, para preparar os dados para modelagem.
4. Construir e avaliar modelos de aprendizado de máquina (ex.: regressão logística, árvore de decisão, random forest, redes neurais) para prever a ocorrência de diabetes.
5. Comparar métricas de desempenho dos modelos (acurácia, precisão, recall, F1-score, AUC) para selecionar o mais adequado.
6. Interpretar os resultados obtidos e discutir suas implicações para o contexto de saúde pública da população Pima.

Referências de aquisição do dataset

Os dados foram coletados por meio de exames clínicos e entrevistas estruturadas conduzidas pela equipe do NIDDK, como parte de um estudo epidemiológico de longo prazo com foco na diabetes tipo 2. A população estudada consiste exclusivamente em mulheres da comunidade Pima, acima de 21 anos, localizadas no sul do Arizona. O contexto envolve variáveis relacionadas à saúde metabólica, fatores de risco e histórico familiar, com o objetivo de rastrear padrões emergentes e fatores predisponentes.

Período da coleta: desde 1965, em ciclos bienais de exames clínicos e entrevistas.

População: mulheres de ascendência Pima, residentes no Arizona, com idade ≥ 21 anos.

Limitações: presença de valores zero em variáveis que não poderiam assumir esse valor (e.g., glicose, IMC), indicando a necessidade de pré-processamento.

Licenciamento e uso: o dataset é amplamente utilizado em pesquisas e práticas de ensino, servindo como benchmark em algoritmos de classificação e aprendizado supervisionado.

Descrição da origem dos dados

A origem dos dados é um estudo epidemiológico de longo prazo conduzido pelo NIDDK. O foco principal foi monitorar a alta incidência de diabetes tipo 2 entre mulheres Pima, relacionando-a a fatores metabólicos, clínicos e familiares.

O levantamento foi realizado com:

Exames laboratoriais: medições de glicose, insulina e IMC.

Entrevistas estruturadas: histórico de saúde e condições familiares.

Variáveis antropométricas: idade, número de gestações, espessura da pele, pressão arterial.

Esse contexto justifica a relevância científica do dataset, que não só contribui para pesquisas médicas, como também é amplamente utilizado como referência em projetos de ciência de dados e aprendizado de máquina.

Descrição do dataset

O dataset contém **768 registros** de mulheres Pima, divididas entre **268 (35%) diagnosticadas com diabetes** e **500 (65%) não diagnosticadas**.

Estrutura de atributos

São 8 variáveis de entrada (numéricas) e 1 variável alvo:

Pregnancies: número de gestações.
Glucose: concentração de glicose plasmática.
BloodPressure: pressão arterial diastólica (mmHg).
SkinThickness: espessura da dobra cutânea tricipital (mm).
Insulin: concentração de insulina sérica (mu U/ml).
BMI: índice de massa corporal (kg/m²).
DiabetesPedigreeFunction: risco calculado com base em histórico familiar.
Age: idade em anos.
Outcome: variável de saída (0 = não diabético; 1 = diabético).

Observações sobre os dados

Algumas variáveis contêm valores zero que não são plausíveis (ex.: pressão arterial ou IMC iguais a zero). Isso indica registros incompletos ou dados mascarados, exigindo tratamento antes de análises preditivas.

O dataset apresenta **desbalanceamento de classes**, o que deve ser considerado no uso de modelos de aprendizado de máquina.

Pensamento Computacional

Decomposição

Problema geral: entender por que há alta incidência de diabetes tipo 2 entre os Pima.

Problemas menores:

Começaremos entendendo:

- **Qualidade dos dados** (valores inválidos).
- **Variáveis clínicas** (glicose, pressão arterial, IMC).
- **Fatores familiares** (Diabetes Pedigree Function).
- **Aspectos sociodemográficos** (idade, gestações).
- **Resultado final** (diagnóstico: diabético/não diabético).

Reconhecimento de Padrões

Os passos para identificação de semelhanças e regularidades:

- **Identificação de semelhanças:** verificação de grupos nas métricas das variáveis clínicas, dos fatores familiares, aspectos sociodemográficos para procurar características comuns para os possíveis problemas ou criação de conjuntos de dados.
- **Simplificação de problemas:** após o reconhecimento de um padrão, colocar os problemas em partes menores para ser melhor gerenciado, para comprovar algumas hipóteses que a diabete tipo 2 pode ser mais presente em grupos específicos (por exemplo: que pode ser por idade, por gestação, por estilo de vida).
- **Reutilização de soluções:** durante o estudo, podemos encontrar algumas soluções que podem ser aplicadas tanto para aspectos sociodemográficos quanto às variáveis clínicas, agilizando o processo.
- **Criação de regras:** com a identificação dos motivos que levam ao grupo a ter diabete tipo 2, é possível criar regras para alimentação, peso, estilo de vida, etc, para prevenção e cuidados da doença.

Abstração

- **Problema amplo:** a diabetes tipo 2 envolve fatores clínicos, genéticos, ambientais e culturais.
- **Elementos mantidos:** hábitos alimentares, glicose, pressão arterial, IMC, idade, número de gestações, função de histórico familiar.
- **Elementos abstraídos:** contexto cultural, políticas públicas e fatores ambientais não registrados.
- **Benefício da abstração:** possibilita a aplicação de algoritmos de machine learning para identificar padrões e construir modelos preditivos de risco de alta confiabilidade e eficiência.

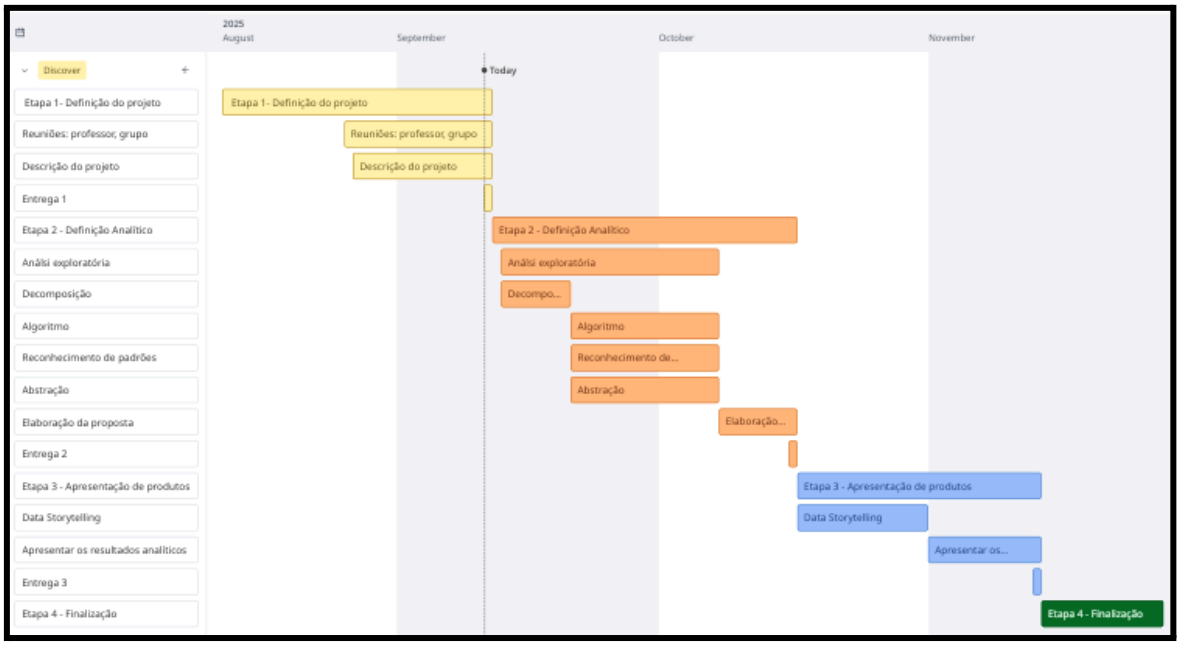
Algoritmo

Aplicação no dataset:

- **Regressão logística:** calcula a probabilidade de um indivíduo ser diabético com base em suas características.
- **Árvore de decisão:** cria regras de decisão, como “se glicose > 125 e IMC > 30, então risco alto”.
- **Random forest:** combina várias árvores de decisão para melhorar a acurácia.
- **Redes neurais:** detectam padrões complexos não lineares entre todas as variáveis.

Cronograma

Para o projeto, o primeiro modelo de cronograma está abaixo:



Referências Bibliográficas

NATIONAL INSTITUTE OF DIABETES AND DIGESTIVE AND KIDNEY DISEASES. KAGGLE. *Pima Indians Diabetes Dataset*. Disponível em: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. Acesso em: 8 set. 2025.

AOBESIDADE. Obesidade mórbida: o paradoxo de Pima. Disponível em: https://aobesidade.com.br/obesidade-morbida/#:~:text=O%20paradoxo%20de%20Pima&text=Mas%20veja%20a%20seguir%20um,os%20%C3%ADndios%20Pima%20do%20Arizona. Acesso em: 9 set. 2025.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. Diabetes in the Pima Indians. In: **The Genetic Basis of Common Diseases**. Bethesda: NCBI Bookshelf, 1992. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK233089/. Acesso em: 9 set. 2025.

NATIONAL INSTITUTES OF HEALTH. NIH's work with Native communities drives diabetes research. In: **NIH Intramural Research Program Catalyst**, v. 29, n. 6, 2021. Disponível em: https://irp.nih.gov/catalyst/29/6/nihs-work-with-native-communities-drives-diabetes-research. Acesso em: 9 set. 2025.

PUBMED. KNOWLER, W. C.; BENNETT, P. H.; HANSON, R. L.; et al. Diabetes incidence and prevalence in Pima Indians: influence of obesity and family history. *Diabetes Care*, v. 16, n. 1, p. 120–126, 1993. DOI: 10.2337/diacare.16.1.120. Disponível em: https://pubmed.ncbi.nlm.nih.gov/8422779/. Acesso em: 9 set. 2025.

PUBMED. ESCOBEDO, J. et al. Prevalence of diabetes and obesity in the Pima Indians of Mexico. *Diabetes Care*, v. 29, n. 8, p. 1852–1856, 2006. DOI: 10.2337/dc06-0040. Disponível em: https://pubmed.ncbi.nlm.nih.gov/16873794/. Acesso em: 9 set. 2025.

RESEARCHGATE. Pima Indian Diabetes dataset attributes. In: *Article: A comparative study of classification algorithms for diabetes prediction*. Disponível em: https://www.researchgate.net/figure/Pima-Indian-Diabetes-dataset-attributes_tbl1_325653625. Acesso em: 9 set. 2025.

SCIENTIFIC AMERICAN / SAGE. Pima Indians and Obesity. In: *Encyclopedia of Obesity*. Thousand Oaks: Sage Publications, 2008. Disponível em: https://sk.sagepub.com/ency/edvol/embed/obesity/chpt/pima-indians. Acesso em: 9 set. 2025.

SCIENTIFIC DIRECT. KUMAR, A. et al. An interpretable machine learning model for diabetes prediction using the Pima Indians dataset. *Heliyon*, v. 10, n. 1, 2024. DOI: 10.1016/j.heliyon.2024.e23957. Disponível em: https://www.sciencedirect.com/science/article/pii/S240584402400567X. Acesso em: 9 set. 2025.