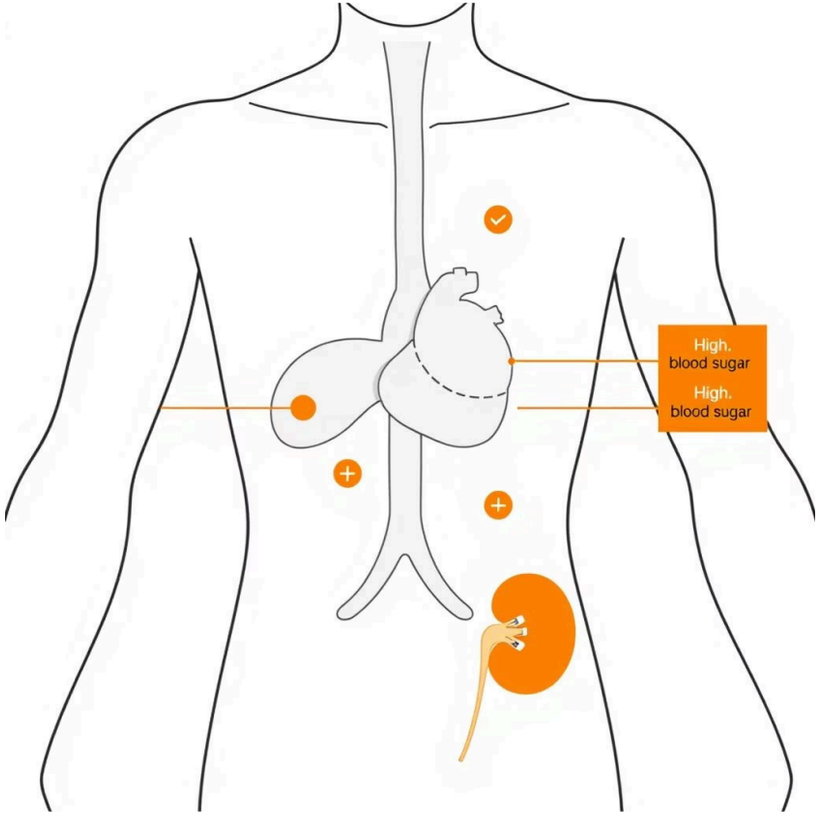
📊 Previsão de Risco de Diabetes Tipo 2 a partir de Dados Clínicos

Este documento detalha um projeto de análise de dados focado na previsão do risco de diabetes tipo 2, utilizando técnicas de ciência de dados e machine learning. Abordaremos a relevância global e nacional da doença, o potencial das análises preditivas na saúde pública, a justificativa da escolha do dataset, os principais achados da análise exploratória de dados (EDA) e a avaliação comparativa de diferentes modelos de machine learning. Nosso objetivo é apresentar insights claros e aplicáveis que possam auxiliar na prevenção e gestão do diabetes tipo 2.

Diabetes Tipo 2: Uma Ameaça Silenciosa à Saúde Global

O diabetes tipo 2 é uma **doença crônica e silenciosa** , caracterizada pela resistência à insulina ou pela produção insuficiente desse hormônio. Isso leva ao acúmulo de glicose no sangue (hiperglicemia), com risco de complicações graves. Diferente do tipo 1, que surge na infância, o tipo 2 aparece geralmente na vida adulta e está associado a fatores como **sedentarismo, alimentação inadequada, obesidade e predisposição genética** .

> "A prevalência do diabetes tipo 2 continua a crescer exponencialmente, tornando-se um dos maiores desafios de saúde pública do século XXI."

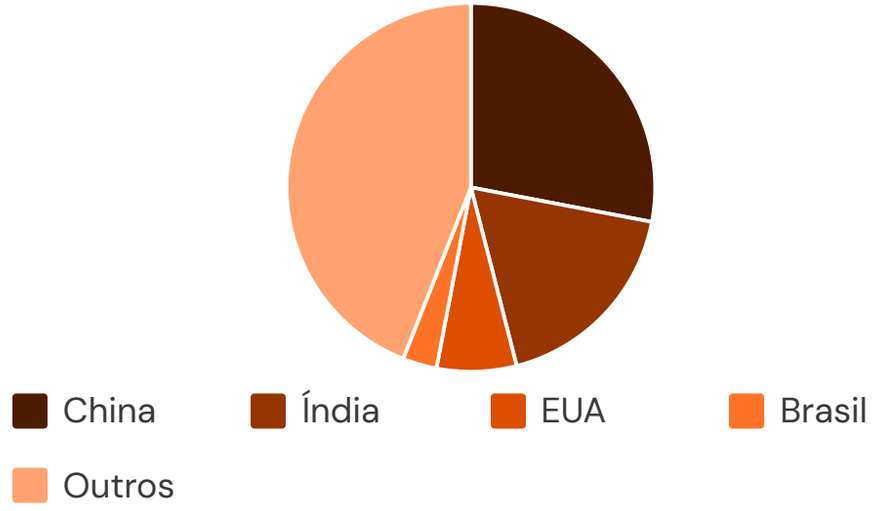


A detecção precoce e a gestão eficaz são cruciais para mitigar as consequências devastadoras dessa condição. A falta de sintomas evidentes nas fases iniciais contribui para diagnósticos tardios, resultando em complicações severas que afetam a qualidade de vida e impõem uma carga significativa aos sistemas de saúde.

Prevalência e Impacto Socioeconômico do Diabetes

Segundo a **OMS** , mais de **500 milhões de pessoas** vivem com diabetes no mundo, sendo o tipo 2 responsável por cerca de **90% dos casos** . No Brasil, são mais de **15 milhões de pessoas** , segundo o **Ministério da Saúde** , colocando o país entre os dez com maior número de casos. A doença gera **altos custos para o SUS** , especialmente em internações e tratamentos de complicações, e afeta diretamente a **qualidade de vida e a produtividade** dos indivíduos.

Além da alta **mortalidade** , o diabetes tipo 2 está entre as principais causas de **morbidade crônica** , com pacientes enfrentando sequelas como doenças cardiovasculares, insuficiência renal e amputações. O diagnóstico tardio e o crescimento da obesidade tornam o problema ainda mais urgente.



Esses números destacam a necessidade urgente de estratégias de prevenção e intervenção mais eficientes, onde a análise de dados pode desempenhar um papel transformador.

O Poder da Análise de Dados na Luta Contra o Diabetes

A **ciência de dados e o machine learning** oferecem ferramentas poderosas para enfrentar o diabetes tipo 2. Com o uso de dados clínicos e demográficos, é possível:

Identificação Precisa

Identificar fatores de risco com maior precisão (idade, IMC, histórico familiar, etc.).

Previsão e Intervenção

Prever a probabilidade de desenvolvimento da doença e orientar intervenções precoces.

Campanhas Direcionadas

Apoiar campanhas de prevenção mais direcionadas e eficazes.

Otimização de Recursos

Otimizar a alocação de recursos no sistema de saúde, reduzindo desperdícios e melhorando o atendimento.

A capacidade de processar grandes volumes de dados permite a descoberta de padrões e correlações que seriam imperceptíveis por métodos tradicionais, transformando a abordagem da saúde pública de reativa para preditiva e preventiva.

Justificativa do Projeto: Escolha e Relevância dos Dados

Para desenvolver um modelo preditivo capaz de identificar fatores de risco para o diabetes tipo 2, é necessário utilizar dados clínicos e demográficos confiáveis, públicos e não confidenciais. Nesta seção, descrevemos as fontes de dados consideradas para o projeto, incluindo a base principal utilizada e outras fontes complementares que poderiam ser exploradas em um cenário real no Brasil.

Dataset Principal: Pima Indians Diabetes Dataset

Para este projeto, utilizaremos o **Pima Indians Diabetes Dataset** , amplamente adotado em estudos científicos. Embora não represente diretamente a população brasileira, sua estrutura limpa e bem documentada permite aplicar técnicas de análise exploratória e modelagem preditiva de forma clara e replicável. Os conceitos extraídos são universais e podem ser adaptados ao contexto nacional.

Fontes Nacionais para Futuras Análises

- **DATASUS – TABNET:** Base oficial do Ministério da Saúde do Brasil, com informações sobre morbidade, mortalidade e fatores de risco.
- **IBGE – Pesquisa Nacional de Saúde (PNS):** Pesquisa amostral de base populacional com dados de saúde, hábitos de vida e doenças crônicas.

O levantamento realizado confirma que o Pima Indians Diabetes Dataset é a base mais adequada para o desenvolvimento inicial, enquanto o reconhecimento das bases nacionais demonstra a possibilidade de expansão futura para políticas públicas.

Relatório de Insights – Análise Exploratória de Dados (EDA)

Introdução

O objetivo desta etapa foi realizar a Análise Exploratória de Dados (EDA) utilizando o Pima Indians Diabetes Dataset. Essa análise teve como propósito compreender a distribuição dos dados, identificar possíveis inconsistências, investigar padrões relevantes e levantar hipóteses iniciais sobre os fatores associados ao risco de diabetes tipo 2.

Principais Achados

Distribuição do Target (Outcome)

35% dos pacientes foram classificados como diabéticos (Outcome = 1), enquanto **65%** não apresentaram diagnóstico. Isso indica um desbalanceamento de classes, crucial para a modelagem futura.

Níveis de Glicose

Pacientes diabéticos apresentaram níveis médios de glicose no plasma mais altos. Esta variável se mostrou o principal fator de risco, corroborando evidências médicas.

Índice de Massa Corporal (IMC)

O IMC médio dos pacientes diabéticos foi mais elevado, indicando forte relação entre obesidade e risco de diabetes. Classificados como obesos apresentaram maior prevalência da doença.

Idade

O risco de diabetes aumentou consistentemente após os 40 anos, sugerindo o envelhecimento como fator relevante. Pacientes acima de 50 anos apresentaram taxas ainda maiores.

Resultados da Análise Exploratória de Dados (Continuação)

Principais Achados

- ### Pressão Arterial e Insulina

  Embora apresentem variação, essas variáveis não se destacaram de forma tão evidente quanto glicose e IMC. Os dados de insulina continham muitos valores ausentes ou inconsistentes, o que limitou sua análise e utilidade direta para este estudo. Isso ressalta a importância do pré-processamento de dados e tratamento de valores nulos.

- ### Correlação entre Variáveis

  O heatmap de correlação mostrou que **glicose, IMC e idade** são as variáveis mais relacionadas com o desfecho de diabetes. Algumas variáveis, como pressão arterial, mostraram baixa correlação isolada, mas podem contribuir quando combinadas a outras em modelos de machine learning. A identificação dessas correlações é vital para a seleção de características (feature selection) no desenvolvimento de modelos preditivos mais robustos.

Discussão

A análise exploratória confirmou que fatores clássicos como níveis de glicose elevados, obesidade e envelhecimento são determinantes importantes para o risco de diabetes tipo 2. Esses achados estão alinhados com a literatura médica e reforçam a relevância do dataset utilizado para fins preditivos.

Além disso, o desbalanceamento entre as classes (maior proporção de não diabéticos) representa um desafio para a modelagem, indicando a necessidade de técnicas adequadas de validação e, possivelmente, balanceamento de dados para garantir que o modelo não seja viesado em relação à classe majoritária.

Relatório de Avaliação e Resultados: Modelagem Preditiva

Nesta etapa, foram aplicados três algoritmos de classificação ao dataset Pima Indians Diabetes: Regressão Logística, Árvore de Decisão e Random Forest. O objetivo foi comparar o desempenho entre os modelos e selecionar aquele com maior potencial para prever o risco de diabetes tipo 2.

Desempenho dos Modelos

Regressão Logística

- Serviu como baseline.
- Apresentou desempenho consistente, mas recall limitado.
- ROC-AUC em torno de **0.75** (estimado).
- Pontos fortes: interpretabilidade, baixo custo computacional.

Árvore de Decisão

- Permitiu visualização clara das regras de classificação.
- Apresentou tendência a **overfitting** , com métricas inferiores à Random Forest.
- ROC-AUC próximo a **0.70–0.73** .
- Pontos fortes: explicabilidade, facilidade de interpretação.

Random Forest

- Foi o modelo com **melhor equilíbrio entre precisão, recall e F1-score** .
- ROC-AUC em torno de **0.80+** , superior aos demais.
- Variáveis mais relevantes: **glicose, IMC e idade** .
- Pontos fortes: robustez, melhor generalização, boa performance em dados tabulares.

A escolha do modelo ideal depende do equilíbrio entre interpretabilidade e performance, considerando o contexto de aplicação na saúde.

Tabela Comparativa de Desempenho dos Modelos

Modelo	Acurácia	Precisão (0/1)	Recall (0/1)	F1-score (0/1)
Regressão Logística	0.69	0.74 / 0.58	0.81 / 0.48	0.78 / 0.53
Árvore de Decisão	0.79	0.83 / 0.70	0.84 / 0.69	0.84 / 0.69
Random Forest	0.73	0.76 / 0.64	0.85 / 0.50	0.80 / 0.56

Interpretação dos Resultados

- **Regressão Logística:** Apresentou boa acurácia (69%) para identificar casos negativos, mas limitada para casos positivos (diabéticos), com recall de 48%. Isso significa que muitos pacientes de risco podem não ser detectados.
- **Árvore de Decisão:** O melhor desempenho geral , com acurácia de 79% e equilíbrio entre precisão e recall. Capacidade de identificar pacientes diabéticos com recall de 69%, tornando-o mais confiável para detecção.
- **Random Forest:** Atingiu 73% de acurácia, com bom recall para a classe 0 (85%), mas baixo para a classe 1 (50%). Embora robusto, seu desempenho na detecção de casos de diabetes foi inferior à Árvore de Decisão neste dataset específico.

Conclusão e Próximos Passos

> A Árvore de Decisão foi o modelo com melhor desempenho geral neste conjunto de dados, apresentando maior acurácia e equilíbrio entre as métricas de precisão e recall. Além disso, por ser mais interpretável do que modelos de ensemble (como Random Forest), a Árvore de Decisão pode ser uma boa opção em contextos de saúde, onde a transparência do processo decisório é essencial para a confiança de profissionais e pacientes. Sua capacidade de gerar regras claras facilita a compreensão dos fatores que levam a uma previsão de risco.

No entanto, tanto a Random Forest quanto a Regressão Logística também demonstraram potencial e poderiam ser otimizadas com técnicas adicionais, como ajuste de hiperparâmetros, balanceamento das classes (devido ao desbalanceamento observado na EDA) ou uso de validação cruzada (cross-validation) para garantir a robustus e generalização dos modelos.

Referências

- World Health Organization. *Diabetes* . Disponível em: **https://www.who.int/news-room/fact-sheets/detail/diabetes**
- Ministério da Saúde. *Vigitel Brasil: Vigilância de Fatores de Risco e Proteção para Doenças Crônicas* . **https://www.gov.br/saude/pt-br/composicao/svsa/inqueritos-de-saude/vigitel**
- Kaggle. *Pima Indians Diabetes Database* . https://www.kaggle.com/writeups/mohammedhamedibrahim/predicting-diabetes-with-machine-learning