Trabalho Prático Aprendizado de Máquina

PNS 2019

Vitor Totaro Fialho
Aprendizado de Máquina
PUC Minas Lourdes
Minas Gerais/Brasil
vitor.t.fialho@gmail.com

1. Descrição do Problema

Análise e Identificação de Fatores Determinantes Associados à Prevalência de Diabetes Mellitus na População Brasileira. A análise foca em uma subpopulação que se estende dos 35 aos 60 anos para os saudáveis e, no caso dos diabéticos, quem recebeu o diagnóstico nessa faixa de idade. Isso com base em Técnicas de Mineração de Dados aplicadas aos microdados da Pesquisa Nacional de Saúde (PNS) 2019.

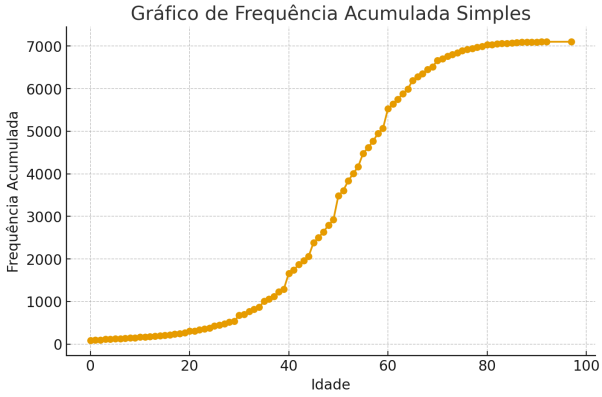

Figura 1: Gráfico de frequência acumulada simples mostrando a idade de diagnóstico da doença

O objetivo é utilizar as diversas variáveis socioeconômicas, demográficas e de estilo de vida da PNS 2019 para construir um modelo de classificação capaz de identificar os principais fatores preditivos que distinguem o grupo de indivíduos diagnosticados com diabetes do grupo de controle (indivíduos sem nenhuma doença crônica, considerados saudáveis).

2. Descrição da Base de Dados

A aplicação eficaz de algoritmos de mineração de dados em saúde requer mais do que apenas o processamento mecânico de grandes volumes de informações; exige uma compreensão profunda do contexto em que esses dados foram gerados. Conforme apontam Zárate *et al.* (2023) ao proporem o método CAPTO, o entendimento prévio do domínio do problema é uma etapa crítica e muitas vezes negligenciada no processo de Descoberta de Conhecimento em Bases de Dados (KDD). Sem essa imersão, corre-se o risco de extrair padrões que, embora estatisticamente válidos, carecem de relevância clínica ou prática.

Neste estudo, adotou-se essa premissa ao estruturar o pipeline de tratamento dos dados da Pesquisa Nacional de Saúde (PNS). Antes da aplicação dos modelos preditivos, foi realizada uma análise conceitual das variáveis disponíveis para garantir que a coluna alvo, referente ao diagnóstico de Diabetes, e seus atributos preditores refletissem adequadamente a realidade epidemiológica capturada pela pesquisa. Essa abordagem visa assegurar que os resultados obtidos não sejam apenas artefatos dos dados, mas sim conhecimentos acionáveis para o contexto da saúde pública brasileira. O mapa e modelo conceitual se utilizados para esse projeto encontram-se ao final do artigo, no módulo 10 (“Modelo conceitual”).

Dito isso, a base de dados em estudo é um subconjunto derivado dos microdados da Pesquisa Nacional de Saúde (PNS) 2019, que originalmente possui 279.382 registros e cerca de 150 atributos. A partir deste universo, foram selecionados cerca de 30 atributos considerados relevantes para o problema da diabetes (incluindo dados demográficos, socioeconômicos, de estilo de vida e medições de saúde).

3. Pipeline de Mineração e Pré-Processamento

Para transformar os dados brutos em um conjunto de dados pronto para a modelagem, o seguinte pipeline de pré-processamento foi executado em ordem cronológica:

3.1. Filtros Iniciais e Tratamento da Subamostra Antropométrica

A primeira etapa consistiu em filtrar a base para o corte de interesse (35 a 60 anos).

Após isso, os registros que não continham as variáveis de peso e altura foram removidos, visto que esses são atributos fundamentais para o problema em questão e que não podem ser imputados de maneira segura e concisa.

Esta filtragem resultou em uma base de dados limpa, porém desbalanceada, contendo 22.663 registros (18.004 indivíduos saudáveis e 4.659 indivíduos com diabetes).

Após isso, foram selecionadas variáveis que se relacionam ao problema em questão, tendo como base um modelo conceitual contruído a partir de literaturas e artigos científicos da área da saúde com foco específico na diabetes.

3.2. Engenharia e Limpeza de Features Numéricas

Com a subamostra definida, as variáveis foram tratadas:

Limpeza de Dados: Dados corrompidos (ex: '71.570.7') e códigos de erro (ex: 999) foram convertidos para NaN (Ausente) usando *pd.to_numeric(errors='coerce').*

Cálculo de Média: Visto que as variáveis que informavam peso e altura possuíam duas respostas (inicial e final), as colunas Peso_Final e Altura_Final_cm foram criadas calculando a média das duas medições *(.mean(skipna=True))*, o que permitiu salvar registros mesmo que uma das duas medições estivesse corrompida (agora como NaN).

Cálculo do IMC: O IMC foi calculado usando a fórmula padrão e arredondado para 2 casas decimais. Peso_Final e Altura_Final_cm também foram arredondados para 2 casas.

Tratamento de vazios (NaN) lógicos: Algumas variáveis possuíam vazios por conta da pergunta não se aplicar a quem está respondendo. Esses registros não estão ausentes, mas são vazios lógicos, por isso foram tratados da forma devida, a maioria substituído por 0 (nos casos de pergunta de frequência por exemplo).

Imputação de dados ausentes: Este processo incluiu a imputação de dados ausentes (utilizando o *MissForest* visto que as variáveis que necessitavam imputação eram todas categóricas).

Tratamento de *Outliers (Capping):* Para reduzir o impacto de valores extremos sem perder registros, foi aplicada a técnica de *capping*. Valores que excediam os limites superior e inferior (definidos, por exemplo, pelo Percentil 99 e 1) foram substituídos por esses próprios limites.

3.3. Renomeação de Variáveis

Para melhorar a interpretabilidade dos dados e dos modelos, as colunas foram renomeadas de seus códigos originais (ex: C006, P006) para nomes legíveis (ex: SEXO, FEIJAO).

3.4. Conjunto de Dados Final para Modelagem

Divisão Treino-Teste (Train-Test Split)

O que faz: A base de 22.663 registros foi dividida em conjuntos de Treino (70%) e Teste (30%).

Por que faz: O conjunto de teste é "trancado" e jamais usado para treinamento, *encoding* ou balanceamento. Ele representa o "mundo real" e é usado apenas para a avaliação final. O parâmetro *stratify=y* foi usado para garantir que a proporção original de 18.004 saudáveis / 4.659 diabéticos fosse mantida em ambas as divisões.

3.5. Pipeline de Preparação para Modelagem

Código: pré-processamento (Código que pega a base previamente imputada e dividida em train/test, aplica as próximas etapas de pré-processamento e após isso realiza a aplicação dos algoritmos)

Esta é a etapa final de processamento, que segue as melhores práticas para evitar vazamento de dados *(data leakage).*

a) Pré-processamento (Encoding e Scaling)

O pré-processamento foi "aprendido" (fit) no conjunto de treino e apenas "aplicado" (transform) no conjunto de teste.

Label Encoding (Ordinais): Variáveis com ordem (ex: NIVEL_ATIVIDADE_FISICA) foram convertidas em números (ex: 0, 1, 2).

One-Hot Encoding (Nominais): Variáveis sem ordem (ex: SEXO) foram transformadas em colunas binárias (ex: SEXO_M, SEXO_F) para evitar que o modelo aprendesse uma ordem falsa.

StandardScaler (Numéricas): Variáveis numéricas (ex: IMC) foram padronizadas (média 0, desvio 1) para que tivessem a mesma escala de importância para os algoritmos.

b) Balanceamento de Classes (Under, Over e Híbrido)

O que faz: As técnicas de balanceamento foram aplicadas apenas ao conjunto de treino codificado.

Por que faz: A experimentação comparativa entre as estratégias de balanceamento é fundamental porque não existe uma solução universal que garanta o melhor desempenho para todas as distribuições de dados e

algoritmos (princípio do *No Free Lunch*). Enquanto o *undersampling* foca na redução da classe majoritária — melhorando a eficiência computacional mas correndo o risco de descartar informações valiosas —, o *oversampling* preserva os dados originais e sintetiza novos exemplos, podendo, contudo, introduzir ruído ou causar *overfitting*. A abordagem híbrida tenta equilibrar esses extremos, limpando fronteiras de decisão ruidosas enquanto reforça a classe minoritária. Portanto, testar os três cenários é a única maneira empírica de descobrir qual compensação (*trade-off*) entre perda de informação e generalização maximiza as métricas de desempenho especificamente para essa base de dados.

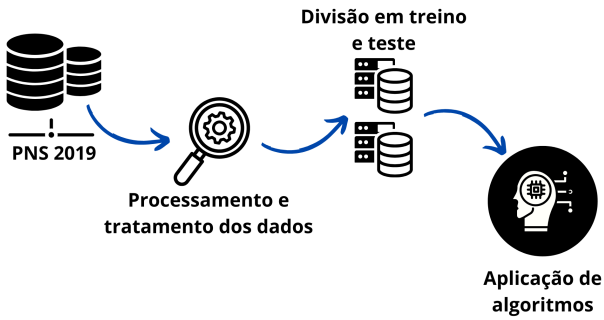

Figura 2: Ilustração do pipeline de processamento

Os atributos estão estruturados da seguinte forma:

Variável	Descrição da variável	Código/valor	Descrição da resposta
SEXO	Sexo	1	Homem
		2	Mulher
FEIJAO	Em quantos dias da semana o(a) Sr(a) costuma comer feijão?	1 a 7	Dias
		0	Nunca ou menos de uma vez por semana
VERDURA_LEGUME	Em quantos dias da semana, o(a) Sr(a) costuma comer pelo menos um tipo de verdura ou legume (sem contar batata, mandioca, cará ou inhame) como alface, tomate, couve, cenoura, chuchu, berinjela, abobrinha?	1 a 7	Dias
		0	Nunca ou menos de uma vez por semana
FREQ_VERDURA_LEGUME	Em geral, o(a) Sr(a) costuma comer esse tipo de verdura ou legume:	1	Uma vez por dia (no almoço ou no jantar).
		2	Duas vezes por dia (no almoço e no jantar).
		3	Três vezes ou mais por dia.
		0	Não consome
CARNE_VERMELHA	Em quantos dias da semana o(a) Sr(a) costuma comer carne vermelha (boi, porco, cabrito, bode, ovelha etc.)?	1 a 7	Dias
		0	Nunca ou menos de uma vez por semana.
FRANGO_GALINHA	Em quantos dias da semana o(a) Sr(a) costuma comer frango/galinha?	1 a 7	Dias
		0	Nunca ou menos de uma vez por semana.
PEIXE	Em quantos dias da semana o(a) Sr(a) costuma comer peixe?	1 a 7	Dias
		0	Nunca ou menos de uma vez por semana.

SUCO_INDUSTRIALIZADO	Em quantos dias da semana o(a) Sr(a) costuma tomar suco de caixinha/lata ou refresco em pó ?	1 a 7	Dias
		0	Nunca ou menos de uma vez por semana.
TIPO_SUCO_INDUSTIALIZADO	Que tipo de suco de caixinha/lata ou refresco em pó o(a) Sr(a) costuma tomar? (Ler as opções de resposta)	1	Diet/Light/Zero
		2	Normal
		3	Ambos.
		0	Não consome
SUCO_NATURAL	Em quantos dias da semana o(a) Sr(a) costuma tomar suco de fruta natural (incluída a polpa de fruta congelada)?	1 a 7	Dias
		0	Nunca ou menos de uma vez por semana.
FRUTA_SEMANA	Em quantos dias da semana o(a) Sr(a) costuma comer frutas?	1 a 7	Dias
		0	Nunca ou menos de uma vez por semana.
FREQ_FRUTA_DIA	Em geral, quantas vezes por dia o(a) Sr(a) come frutas?	1	Uma vez por dia
		2	Duas vezes por dia
		3	Três vezes ou mais por dia.
		0	Não consome
REFRIGERANTE_SEMANA	Em quantos dias da semana o(a) Sr(a) costuma tomar refrigerante?	1 a 7	Dias
		0	Nunca ou menos de uma vez por semana.
TIPO_REFRIGERANTE	Que tipo de refrigerante o(a) Sr(a) costuma tomar?	1	Diet/Light/Zero
		2	Normal
		3	Ambos
		0	Não consome
LEITE_SEMANA	Em quantos dias da semana o(a) Sr(a) costuma tomar leite? (de origem animal: vaca, cabra, búfala etc.)	1 a 7	Dias
		0	Nunca ou menos de uma vez por semana.
TIPO_LEITE	Que tipo de leite o(a) Sr(a) costuma tomar?	1	Desnatado ou semidesnatado.
		2	Integral
		3	Os dois tipos
		0	Não consome
DOCES_SEMANA	Em quantos dias da semana o(a) Sr(a) costuma comer alimentos doces como biscoito/bolacha recheado, chocolate, gelatina, balas e outros?	1 a 7	Dias
		0	Nunca ou menos de uma vez por semana
SUBSTITUIR_REFEICAO_DOCE_ SEMANA	Em quantos dias da semana o(a) Sr(a) costuma substituir a refeição do almoço por lanches rápidos como sanduíches, salgados, pizza, cachorro quente, etc?	1 a 7	Dias
		0	Nunca ou menos de uma vez por semana
Peso_Final	Peso - Média entre 1ª e 2ª pesagem (em kg) (3 inteiros e 1 casa decimal)	20 a 200	Quilogramas
Altura_Final_cm	Altura - Média entre 1ª e 2ª medição (em cm) (3 inteiros e 1 casa decimal)	110 a 210	Centímetros

IMC	IMC calculado utilizando peso e altura (2 inteiros e duas casas decimais)	0 a 50	IMC (decimal)
FAIXA_RENDA_SM	Faixa de renda(em salários mínimos da época)	1	Até 1 SM
		2	Mais de 1 até 2 SM
		3	Mais de 2 até 5 SM7
		4	Mais de 5 até 10 SM
		5	Mais de 10 SM
NIVEL_CONSUMO_ALCOOL	Nível do consumo de álcool	0	Não consome
		1	Consumo Ocasional Leve
		2	Consumo Semanal Leve
		3	Consumo Ocasional Pesado (Binge)
		4	Consumo Semanal Pesado
NIVEL_ATIVIDADE_FISICA	Nível da prática de atividade física	0	Sedentário
		1	Ativo (Nível Baixo)
		2	Ativo (Nível Alto)

Complexidade do problema

Utilizei o algoritmo t-SNE da biblioteca sklearn para ver o quão parecidas eram as duas classes (saudáveis e diabéticos) no conjunto de treino **desbalanceado.**

Resultado:

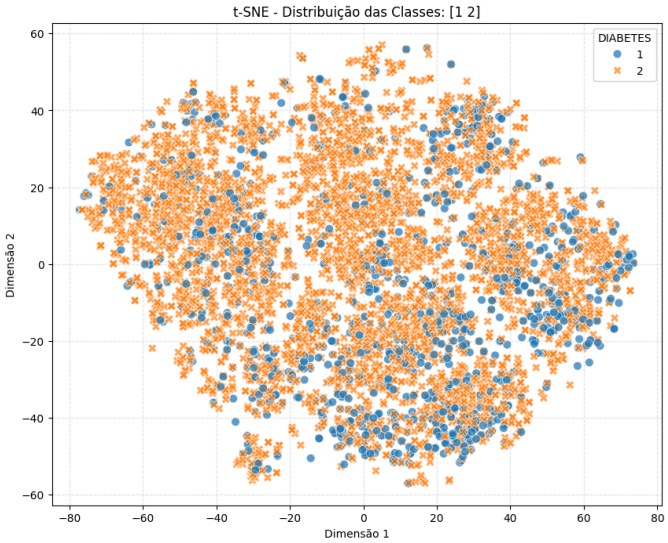

Figura 3: Gráfico t-SNE

5. Modelagem e Otimização de Hiperparâmetros

Código:
https://github.com/VitorTotaro/Machine-Learning/blob/main/processamento_final%20%281%29.ipynb

Nesta etapa, procedeu-se à aplicação e avaliação de algoritmos de Machine Learning para o problema de classificação (identificação de diabetes). Os dados utilizados foram os conjuntos de treino e teste gerados na etapa anterior.

Foram selecionados três algoritmos distintos para comparação: ***K-Nearest Neighbors (KNN)***, **Árvore de Decisão *(Decision Tree)*** e ***Random Forest***.

5.1. Estratégia de Otimização: *BayesSearchCV*

Para encontrar a melhor combinação de hiperparâmetros para cada modelo, optou-se por não utilizar uma busca exaustiva (*GridSearchCV),* que possui um custo computacional muito elevado. Em vez disso, foi empregada a **Otimização Bayesiana** através da biblioteca *scikit-optimize (skopt)* e sua função *BayesSearchCV.*

Esta abordagem trata a busca de hiperparâmetros como um problema de otimização, onde ela "aprende" com as execuções anteriores para escolher de forma mais inteligente as próximas combinações a serem testadas. A métrica alvo para a otimização foi o ***'f1-score'***, uma vez que é uma métrica robusta para lidar com a avaliação em conjuntos desbalanceados (como o nosso conjunto de teste). Foi utilizada uma validação cruzada de 5 *folds* (cv=5) e um total de 15 iterações de busca *(n_iter=15)* para cada modelo, cada modelo sendo utilizado em todas as abordagens (Desbalanceada, UnderSampling, OverSampling e Híbrido)

5.2. Espaços de Busca Definidos

Os intervalos e parâmetros testados para cada modelo foram:

- **KNN:**
  - *n_neighbors*: (Inteiro) de 3 a 30
  - *weights: (Categórico) 'uniform' ou 'distance'*
  - *metric: (Categórico) 'euclidean' ou 'manhattan'.*
- **Árvore de Decisão:**
  - *criterion: (Categórico) 'gini' ou 'entropy'*
  - *max_depth*: (Inteiro) [3, 5, 10, 15]
  - *min_samples_split*: (Inteiro) [2, 5, 10, 20]
  - *min_samples_leaf:* (Inteiro) [1, 2, 5, 10]
- **Random Forest:**
  - *n_estimators:* (Inteiro) [20, 30, 40, 60, 70, 90, 120, 150]
  - *criterion: (Categórico) 'gini' ou 'entropy'*
  - *max_depth:* (Inteiro) [3, 5, 10, 15]
  - *min_samples_split*: [2, 5, 10, 20]
  - *min_samples_leaf*: (Inteiro) [1, 2, 5, 10]

---

6. Avaliação e Análise de Resultados

Após a conclusão do BayesSearchCV, os melhores estimadores (.best_estimator_) de cada algoritmo e abordagem foram utilizados para fazer previsões no conjunto de teste (teste.csv), que o modelo nunca havia visto. Tabelas com hiperparâmetros encontrados:

(Exemplo de hiperparâmetros encontrados levando em conta a abordagem que utiliza a base desbalanceada. Cada abordagem encontrou um conjunto ideal de hiperparâmetros)

Hiperparâmetro	KNN
Nº de Vizinhos (n_neighbors)	3
Métrica de Distância (metric)	manhattan
Peso dos Vizinhos (weights)	uniform

Hiperparâmetro	Decision Tree	Random Forest
Critério de Divisão (criterion)	Gini	Gini
Profundidade Máxima (max_depth)	15	10
Mín. Amostras (Divisão) (min_samples_split)	2	10
Mín. Amostras (Folha) (min_samples_leaf)	10	10
Qtd. de Árvores (n_estimators)	N/A	60
Max_features	N/A	none

6.1. Métricas de Avaliação

Em um problema de diagnóstico de saúde, a acurácia por si só não é uma métrica suficiente. Dei foco especial às seguintes métricas para a classe positiva (Diabetes):

- **Precisão (Precision):** De todas as previsões "Diabetes" feitas pelo modelo, quantas estavam corretas? (Mede a confiabilidade do diagnóstico positivo, minimizando **Falsos Positivos**).
- **Recall (Sensibilidade):** De todas as pessoas que *realmente* tinham diabetes, quantas o modelo conseguiu identificar? (Mede a capacidade de encontrar a doença, minimizando **Falsos Negativos**).
- **F1-Score:** A média harmônica entre Precisão e Recall, fornecendo uma pontuação única que equilibra ambas as métricas.

6.2. Métricas encontradas

A seguir estão tabelas que evidenciam as métricas de cada modelo em cada uma das abordagens:

Abordagem: Desbalanceado (Base)

Algoritmo: KNN | Acurácia Global: 0.7764

	Precision	Recall	F1-Score
Classe 0 (Não Diabetes)	0.8332	0.8965	0.8637
Classe 1 (Diabetes)	0.4538	0.3240	0.3781

Algoritmo: Decision Tree | Acurácia Global: 0.7773

	Precision	Recall	F1-Score
Classe 0 (Não Diabetes)	0.8315	0.9008	0.8647
Classe 1 (Diabetes)	0.4550	0.3121	0.3702

Algoritmo: Random Forest | Acurácia Global: 0.7966

	Precision	Recall	F1-Score
Classe 0 (Não Diabetes)	0.8292	0.9352	0.8790
Classe 1 (Diabetes)	0.5291	0.2742	0.3612

Abordagem: Híbrido (SMOTEENN)

Algoritmo: KNN | Acurácia Global: 0.5451

	Precision	Recall	F1-Score
Classe 0 (Não Diabetes)	0.8857	0.4873	0.6286
Classe 1 (Diabetes)	0.2831	0.7630	0.4130

Algoritmo: Decision Tree | Acurácia Global: 0.6507

	Precision	Recall	F1-Score
Classe 0 (Não Diabetes)	0.8814	0.6447	0.7447
Classe 1 (Diabetes)	0.3346	0.6732	0.4470

Algoritmo: Random Forest | Acurácia Global: 0.7092

	Precision	Recall	F1-Score
Classe 0 (Não Diabetes)	0.8934	0.7177	0.7960
Classe 1 (Diabetes)	0.3890	0.6774	0.4942

Abordagem: Undersampling

Algoritmo: KNN | Acurácia Global: 0.6652

	Precision	Recall	F1-Score
Classe 0 (Não Diabetes)	0.8943	0.6536	0.7553
Classe 1 (Diabetes)	0.3520	0.7090	0.4705

Algoritmo: Decision Tree | Acurácia Global: 0.5810

	Precision	Recall	F1-Score
Classe 0 (Não Diabetes)	0.9004	0.5282	0.6658
Classe 1 (Diabetes)	0.3049	0.7798	0.4384

Algoritmo: Random Forest | Acurácia Global: 0.7067

	Precision	Recall	F1-Score
Classe 0 (Não Diabetes)	0.8982	0.7093	0.7926
Classe 1 (Diabetes)	0.3889	0.6971	0.4992

Abordagem: Oversampling (SMOTE)

Algoritmo: KNN | Acurácia Global: 0.6980

	Precision	Recall	F1-Score
Classe 0 (Não Diabetes)	0.8508	0.7493	0.7968
Classe 1 (Diabetes)	0.3483	0.5049	0.4123

Algoritmo: Decision Tree | Acurácia Global: 0.7279

	Precision	Recall	F1-Score
Classe 0 (Não Diabetes)	0.8430	0.8057	0.8239
Classe 1 (Diabetes)	0.3726	0.4348	0.4013

Algoritmo: Random Forest | Acurácia Global: 0.7929

	Precision	Recall	F1-Score
Classe 0 (Não Diabetes)	0.8534	0.8909	0.8718
Classe 1 (Diabetes)	0.5076	0.4236	0.4618

6.3. Análise dos Relatórios de Classificação

Os resultados consolidados no conjunto de teste (6.799 amostras, sendo 1.426 da classe 'Diabetes' e 5.373 da 'Saudável') revelaram que as técnicas de balanceamento foram essenciais para tornar os modelos úteis para o nosso propósito clínico.

- **Random Forest (com Undersampling):** Foi o modelo que apresentou o **melhor equilíbrio geral (F1-Score: 0.4992)**. Com o balanceamento via Undersampling, ele atingiu um **Recall de aproximadamente 0.70**, o que significa que ele identificou 70% dos casos reais de diabetes, mantendo uma Acurácia razoável de 70%. Ele superou a versão desbalanceada, que só identificava 27% dos doentes.
- **Decision Tree (com Undersampling):** Destacou-se por ter o **maior Recall entre os modelos baseados em árvore (0.78)**. Este modelo foi extremamente agressivo na detecção da doença, errando menos falsos negativos. No entanto, sua Precisão caiu drasticamente (0.30), indicando que para encontrar esses doentes, ele classificou muitos saudáveis incorretamente, resultando em um F1-Score inferior ao do Random Forest.
- **KNN (com SMOTEENN):** Apresentou uma sensibilidade muito alta (**Recall de 0.76**), superando a maioria dos outros cenários. Contudo, sofreu do mesmo problema da árvore de decisão: a baixa precisão (0.28) derrubou a Acurácia global para cerca de 54%. Isso mostra que o KNN, neste dataset da PNS, tende a gerar muito "ruído" quando forçamos o aprendizado da classe minoritária.

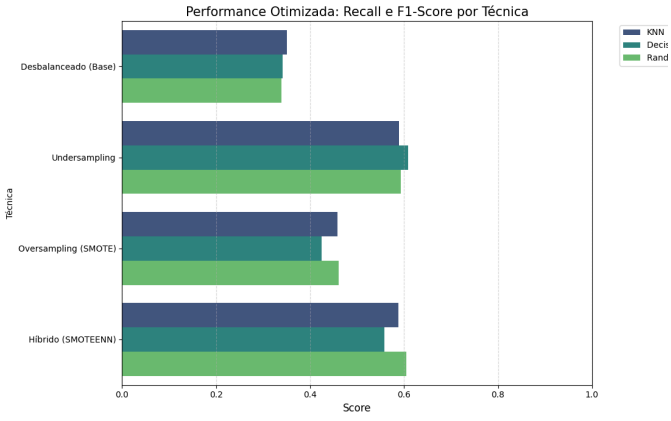

Figura 4: Gráfico comparativo das métricas de cada algoritmo por abordagem

6.4. Análise Visual

Para visualizar os tipos de erros, foram geradas imagens a respeito das métricas:

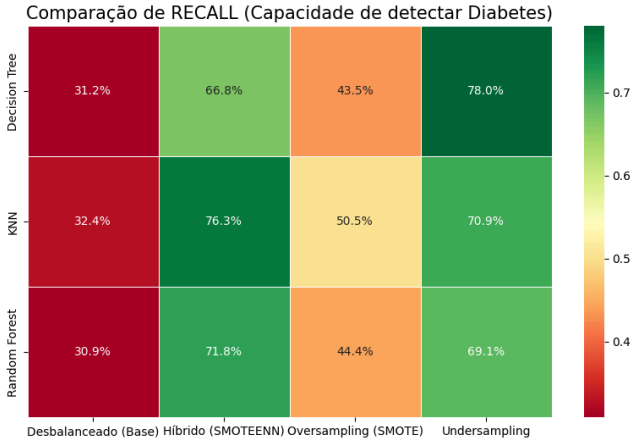

Figura 5: Comparação de recall por algoritmo e abordagem

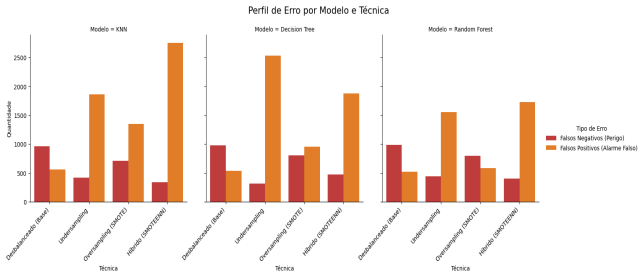

Figura 6: Perfil de erro por modelo e abordagem

A análise das matrizes de confusão e dos gráficos de barras gerados no notebook corroboram os números acima:

- As matrizes dos modelos com **técnicas de balanceamento (Undersampling e SMOTEENN)** mostram claramente uma migração de erros: diminuímos os Falsos Negativos (o que é ótimo para medicina) ao custo de aumentar os Falsos Positivos.
- A matriz do **Baseline (Desbalanceado)** visualmente parece "limpa" na classe majoritária (quase não erra saudáveis), mas é desastrosa na classe alvo, falhando em detectar a grande maioria dos diabéticos.
- O gráfico de barras comparativo evidencia que, enquanto a acurácia cai ligeiramente ao aplicar Undersampling, a barra de **Recall (Sensibilidade)** praticamente dobra de tamanho em comparação ao modelo base, validando a eficácia do pré-processamento focado na classe Diabetes.

7. Conclusão da Etapa de Modelagem

Considerando o objetivo do projeto de mineração de dados da PNS, onde a prioridade é a detecção de uma condição crônica (Diabetes):

1. **O Vencedor Técnico:** O **Random Forest com Undersampling** é a escolha mais robusta. Ele oferece o melhor compromisso, detectando cerca de 70% dos casos (Recall ~0.70) sem destruir completamente a precisão do diagnóstico, mantendo o maior F1-Score do experimento (0.499).
2. **Para Triagem Agressiva:** Se o objetivo fosse puramente uma triagem inicial onde "nenhum doente pode ficar para trás", a **Decision Tree com Undersampling** (Recall 0.78) poderia ser considerada, mas o custo operacional de verificar tantos falsos positivos seria alto.
3. **Descarte do Baseline:** Os modelos sem balanceamento (Baseline), apesar de terem acurácia de quase 80%, são inúteis para o nosso problema, pois funcionam praticamente como classificadores da classe majoritária (Saudáveis), ignorando a doença.

Portanto, o modelo selecionado para a fase de interpretação de variáveis e *deploy* será o **Random Forest treinado com dados balanceados via Undersampling**. A próxima etapa focará em entender quais variáveis do questionário da PNS (como IMC, Idade, Hipertensão) foram decisivas para as predições deste modelo.

8. Ordem de Importância de Atributos e Explicabilidade

Após a seleção dos modelos, avançamos da análise de performance pura para a **interpretabilidade clínica**. Entender quais fatores o algoritmo considera decisivos é fundamental para validar a confiança médica na ferramenta. Abaixo, exploramos como o **Random Forest** (nosso modelo mais robusto) e a **Árvore de Decisão** (nosso modelo mais sensível) tomam suas decisões.

8.1. Fatores de Risco no Random Forest

O gráfico abaixo ilustra as variáveis (features) que o modelo **Random Forest (treinado com Undersampling)** considerou mais relevantes para distinguir entre um paciente saudável e um diabético. O eixo X representa o grau de importância (baseado na redução de impureza de Gini acumulada).

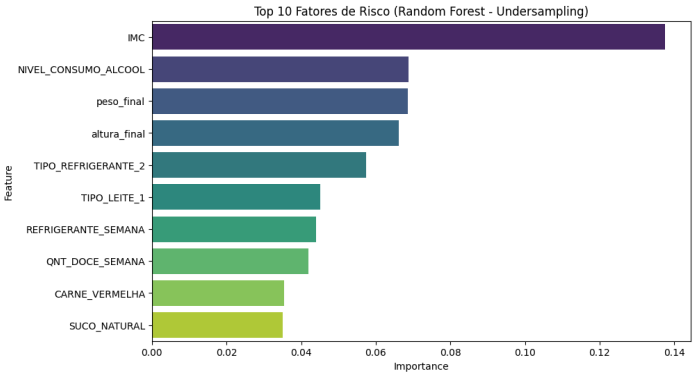

Figura 7: Fatores de risco encontrados pelo Random Forest

Análise: Observe que o modelo não "chuta" aleatoriamente; ele prioriza variáveis que fazem sentido biológico e epidemiológico.

- **Atributos no Topo:** Variáveis que aparecem no topo do gráfico são os divisores de águas mais fortes. O modelo aprendeu que alterações nestes indicadores aumentam drasticamente a probabilidade de diabetes.
- **Cauda Longa:** Variáveis com menor importância ainda contribuem para o ajuste fino da probabilidade, ajudando a desempatar casos limítrofes (borderline).

8.2. As Regras de Decisão (Árvore de Decisão)

Enquanto o Random Forest nos dá a importância geral, a **Árvore de Decisão treinada com SMOTEEN** (sendo a que teve o maior f1-score) nos fornece um fluxograma explícito de triagem. A imagem abaixo mostra a "raiz" da árvore e os primeiros níveis de decisão (profundidade limitada a 3 para melhor visualização).

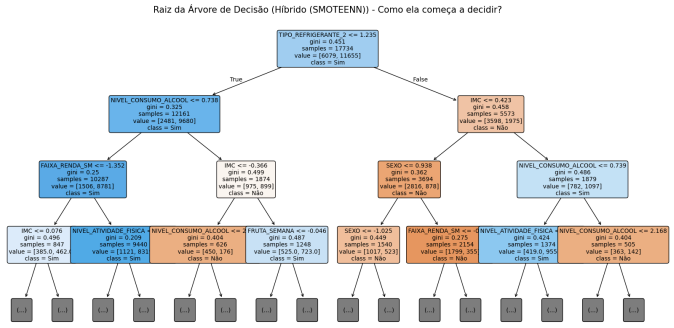

Figura 8: Primeiros níveis da Árvore de decisão SMOTEEN

Interpretação das Regras:

- **Nó Raiz (O topo):** Esta é a pergunta mais discriminante de todo o conjunto de dados. O modelo divide a população inteira baseada nesta única variável. Vemos que é a variável "TIPO_REFRIGERANTE_2", que representa, como visto no dicionário de dados, o refrigerante com açúcar.
- **Ramificações:** Seguindo os ramos para a esquerda ou direita, vemos como o modelo refina o diagnóstico combinando múltiplos fatores.
- **Cor dos Nós:** A intensidade da cor geralmente indica a pureza da classificação. Nós que levam à classe "Diabetes" (Classe 1) com alta certeza demonstram perfis de risco acumulados (ex: Alta Idade + Hipertensão + IMC Elevado).

8.3. Validação Clínica

A coerência entre os atributos mais importantes do Random Forest e as regras de corte da Árvore de Decisão sugere que os modelos conseguiram capturar padrões reais da fisiopatologia do Diabetes, e não apenas ruídos estatísticos da base de dados. Isso aumenta a segurança para um eventual *deploy* da solução como ferramenta de apoio à decisão médica.

9. Conclusão e Trabalhos Futuros

O presente estudo alcançou seu objetivo principal de desenvolver e avaliar modelos de *Machine Learning* capazes de identificar fatores determinantes associados à prevalência de Diabetes Mellitus na população brasileira, utilizando dados da PNS 2019. A análise focou no desafio crítico de lidar com classes desbalanceadas em um cenário de saúde pública.

As principais conclusões extraídas do experimento são:

1. **Impacto do Balanceamento de Dados:** Ficou evidente que a aplicação de algoritmos em dados brutos (desbalanceados) tende a enviesar o resultado para a classe majoritária (saudáveis). Embora a acurácia fosse alta (~79%), a capacidade de detectar doentes (Recall) era inaceitável para triagem médica (~30%). As técnicas de reamostragem (*Undersampling*, *SMOTE* e *SMOTEENN*) foram cruciais para corrigir essa distorção, elevando a sensibilidade dos modelos para patamares acima de 70%.
2. **Melhor Modelo:** O algoritmo **Random Forest**, combinado com a técnica de **Undersampling**, demonstrou ser a solução mais equilibrada e robusta. Ao atingir um *F1-Score* de aproximadamente 0.50 e um *Recall* de 0.70, ele ofereceu o melhor compromisso entre identificar corretamente os pacientes diabéticos e minimizar o número de falsos positivos, superando o KNN (que gerou muitos alarmes falsos) e a Árvore de Decisão (que apresentou maior instabilidade).
3. **Relevância Clínica:** A análise de importância de atributos (*feature importance*) validou o modelo sob a ótica médica. Variáveis como **IMC (Índice de Massa Corporal)** e **Consumo de Álcool** foram identificadas como os preditores mais fortes. Isso confirma que o modelo não apenas encontrou padrões estatísticos, mas capturou a fisiopatologia real da doença, alinhando-se à literatura médica existente sobre fatores de risco para Diabetes Tipo 2.

Trabalhos Futuros

Como próximos passos para a evolução deste projeto, sugere-se:

- **Deploy do Modelo:** Implementar o modelo em uma aplicação web (ex: Streamlit) para simular uma ferramenta de triagem em unidades básicas de saúde.
- **Análise de Causalidade:** Investigar mais a fundo a relação de variáveis socioeconômicas (como Renda e Escolaridade) para entender se atuam como fatores de risco diretos ou variáveis de confusão.
- **Testar Novos Algoritmos:** Avaliar o desempenho de modelos baseados em *Gradient Boosting* (como XGBoost ou LightGBM), que costumam lidar bem com dados tabulares complexos.

REFERÊNCIAS

[1] C Whiteley, F Benton, L Matwiejczyk e N Luscombe-Marsh (2023). Determining Dietary Patterns to Recommend for Type 2 Diabetes: An Umbrella Review. *Nutrients*, 15(4), 861.

[2] Y Liu, et al. (2025). Effects of 12 nutritional interventions on type 2 diabetes: a systematic review with network meta-analysis of randomized trials. *Nutrition & Metabolism*, 22(94).

[3] RR Mir, NU Haq, K Ishaq, N Safie e AB Dogar (2025). Impact of machine learning on dietary and exercise behaviors in type 2 diabetes self-management: a systematic literature review. *PeerJ Computer Science*, 11:e2568.

[4] AD Sarma e M Devi (2025). Artificial intelligence in diabetes management: transformative potential, challenges, and opportunities in healthcare. *Hormones (Athens)*.

[5] E Fix e JL Hodges (1951). Discriminatory Analysis. Nonparametric E. Fix e JL Hodges (1951). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *USAF School of Aviation Medicine. Project No. 21-49-004, Report No. 4.*
[6] L Breiman, J Friedman, R Olshen e C Stone (1984). *Classification and Regression Trees*. Chapman & Hall/CRC.
[7] L Breiman (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
[8]Zárate, L. E., Petrocchi, B., Maia, C. D., Felix, C., & Gomes, M. P. (2023). CAPTO - A method for understanding problem domains for data science projects. *Concilium*, *23*(15), 922-9

10. Modelo conceitual

Mapa:

Descrição do Mapa Conceitual – Domínio de problema: Diabetes Tipo 2		
Dimensão: Hábitos de Saúde		
Aspectos (conhecimento explícito e estudo científico vinculado)	**Atributos associados ao aspecto**	**Atributos vinculados com as fontes de dados**
Uso dos serviços de saúde: O acompanhamento do paciente por profissionais de saúde, e a frequência de exames clínicos, se mostraram relacionados com o uso de medicamentos, e com a necessidade de busca de emergência/ internação, por indivíduos com diabetes (Freitas et al 2018).	- Motivo de saúde que requereu o uso dos serviços de saúde - Diagnóstico médico - Última consulta - Uso de medicamentos	Módulo J - Utilização dos serviços de saúde e Módulo Q – Doenças Crônicas: J4a, J7, J11a, J14, J15a, Q32a, Q33b, Q34c, Q38a3 e Q39a; Fonte: Base de dados PNS (BD-PNS)
Consumo de Drogas: O uso de álcool (Sarit Polsky e Halis K. Akturk), fumo e drogas influenciam negativamente a imunidade dos indivíduos (Pastor et al 2020).	-Quais drogas são consumidas - Qual a frequência de consumo de drogas - Qual a quantidade de drogas consumidas	Módulo P – Estilos de Vida: P27, P28a, P29, P50, P54, P56, P67 e P67a; Fonte: BD-PNS

Atividade Física: Atividades físicas estão intrinsecamente ligadas uma vez que tornam seu corpo mais sensível à insulina (o hormônio que permite que as células do seu corpo usem o açúcar do sangue como energia), o que ajuda a prevenir e controlar a diabetes, (LaMonte et al. 2005).	-Quais atividades físicas são realizadas - Com qual frequência são realizadas atividades	Módulo P – Estilos de Vida: P34, P35, P37 e P36; Fonte: BD-PNS
Dimensão: Hábitos alimentares		
Dieta - Ingestão de alimentos: A alimentação está ligada à diabetes como citado por Reni Aparecida Barsaglini e Ana Maria Canesqui (Barsaglini e Canesqui, 2010).	Quais alimentos são consumidos - Qual a frequência de consumo de alimentos - Qual a quantidade de alimentos consumidos	Módulo P – Estilos de Vida: P6a até P26a; Fonte: BD-PNS
Dieta - Ingestão de bebidas: A alta ingestão de álcool aumenta o risco de diabetes (Polsky e Akturk, 2017)	- Quais bebidas são consumidas - Qual a frequência de consumo de bebidas - Qual a quantidade de bebidas consumidas	Módulo P – Estilos de Vida: P6b até P24a; Fonte: BD-PNS
Dimensão: Condições Físicas e Mentais		
Deficiências: As deficiências podem ter impacto direto nas atividades físicas que uma pessoa realiza e atividades físicas tem impacto direto na diabetes (LaMonte el al, 2005).	- Informações sobre deficiências que possa possuir	Módulo G – Pessoas com Deficiências; Fonte: BD-PNS
Gravidez: Grávidas têm maior suscetibilidade a diabetes (McCance 2011).	-Status (condição S/N) - Influência da diabetes na gravidez	Módulo P – Estilos de Vida e Módulo Q – Doenças Crônicas: P5 e Q30b; Fonte: BD-PNS
Saúde Mental: Doenças mentais como a depressão são fatores de risco para a diabetes (Fráguas et al 2009).	- Possui alguma doença mental - Qual o efeito na vida cotidiana caso possua uma doença mental	Módulo J - Utilização dos serviços de saúde e Módulo Q – Doenças Crônicas: J7, Q92, Q110a e Q115; Fonte: BD-PNS
Percepção: Por meio da percepção individual, as pessoas relatam dores ou sintomas, que podem estar ligados a diabetes (Lee et al 2020).	-autoavaliação - Percepção de sintomas de Diabetes	Módulo N – Percepção do estado de saúde: N1, N1a e N14; Fonte: BD-PNS
Doenças crônicas: Algumas doenças crônicas, como doenças renais, estão intimamente ligadas a diabetes (Koye et al 2018):	- Possui alguma doença crônica (especialmente a própria diabetes) - Efeitos da diabetes	Módulo Q – Doenças Crônicas: Q30a e Q55a; Fonte: BD-PNS
Dados laboratoriais: Dados laboratoriais podem prever diabetes previamente (Gagliardino et al 2017)	- Quais exames foram requisitados a respeito de diabetes	Módulo Q – Doenças Crônicas: Q29a, Q47a e Q51a; Fonte: BD PNS
Dimensão: Condições Sócio-econômicas		
Renda: A renda tem relação com a diabetes diretamente, uma vez que ela reflete ela é fator importante para diversos pontos, como acesso a saúde, dieta, entre outros (Bird 2015).	- Rendimento de trabalhos - Rendimento de outras fontes	Módulo E – Características de trabalho das pessoas 14 anos ou mais de idade e Módulo F – Rendimentos de outras fontes: E16, E18 e F1a até F14a; Fonte: BD-PNS

Acesso a serviços de saúde: Acesso a serviços de saúde é importante para diagnósticos (Zhang et al 2012).	- Acesso a farmácias e profissionais de saúde - Dificuldades de acesso a farmácias e profissionais de saúde	Módulo Q – Doenças Crônicas: Q33a, Q34d, Q37a, Q38a4, Q38a6, Q40a, Q43 e Q50; Fonte: BD-PNS
Plano de saúde: Planos de saúde refletem nos acessos à saúde do indivíduo (IMCCU 2002).	- Informações sobre o plano de saúde caso possua	Módulo I – Cobertura de Plano de Saúde; Fonte: BD-PNS
Trabalho: O trabalho é relacionado com atividades físicas que é intimamente ligado a ocorrência de diabetes (Hu et al 2003).	- Informações sobre o trabalho caso possua - Consequências do trabalho sobre a saúde - Interferências de doenças crônicas no trabalho	Módulo E – Características de trabalho das pessoas 14 anos ou mais de idade e Módulo M – Características do trabalho e apoio social: E12, E14a, E17, E19, M5d e M6; Fonte: BD PNS
Dimensão: Características do indivíduo		
Sexo: Sexo esta ligado a diabetes, que afeta mais os homens em nosso país (Gale e Gillespie 2001).	- Sexo	Módulo C – Características gerais dos moradores: C6; Fonte: BD-PNS
Idade: A idade afeta principalmente a população mais velha (Laakso e Pyörälä,1985	- Idade	Módulo C – Características gerais dos moradores: C7 e C8;Fonte: BD-PNS
Dimensão: Genética		
Predisposição individual: Os fatores de risco para desenvolver o diabetes são: sedentarismo, história familiar de diabetes em parentes de 1º grau (Ali 2013).	- Recomendações para reduzir os efeitos da diabetes	Módulo Q – Doenças Crônicas: Q46a; Fonte: BD-PNS
Dimensão: Antropometria		
Peso e altura: Utilizados para calcular o IMC, fazendo referência a características como a obesidade, que é um fator causador da diabetes (Leong e Wilding, 1999).	- Peso (Múltiplas medições) - Altura (Múltiplas medições)	Módulo P – Estilos de Vida e Módulo W – Antropometria: P1a, P4a, W00201, W00202, W00101 e W00102; Fonte: BD PNS