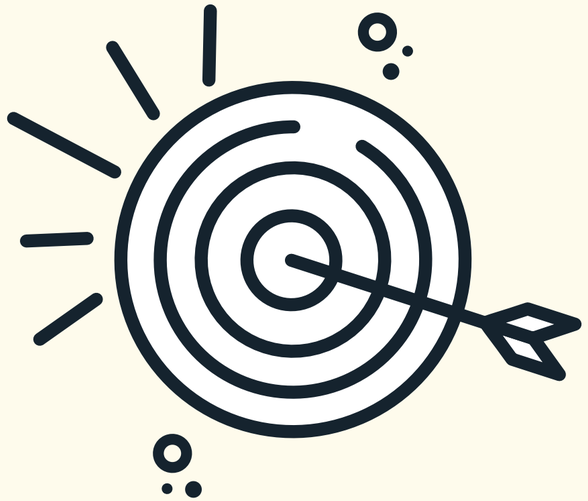
Projeto Paradigmas

João Lucas Felix

Brendo Mendonça

Sumário



Motivação.

Base usada.

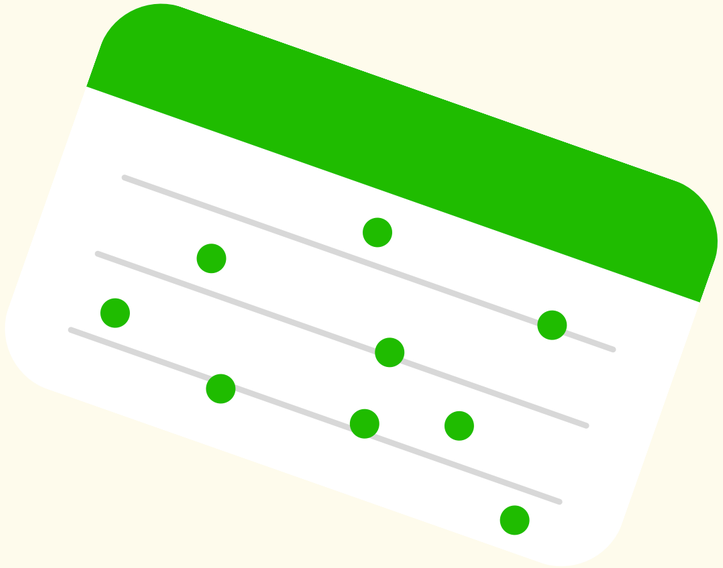
A **Diabetes Health Indicators Dataset** é uma base de dados de saúde pública compilada a partir do Behavioral Risk Factor Surveillance System (BRFSS) de 2015, conduzido pelos Centers for Disease Control and Prevention (CDC) dos EUA.

Características principais:

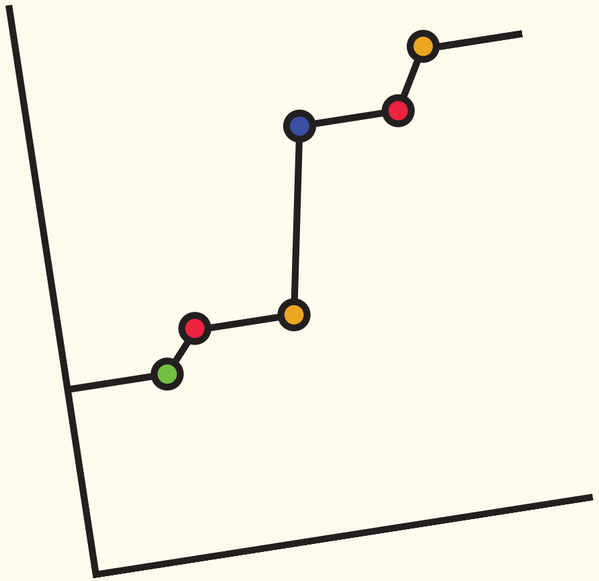
- **Fonte:** CDC BRFSS 2015
- **Amostra:** ≈ 253,680 respostas de pesquisa
- **Período:** Dados de 2015
- **Abrangência:** Nacional (Estados Unidos)

Objetivo

A base foi desenvolvida para **prever** diagnósticos de diabetes com base em indicadores de saúde, comportamentos de risco e fatores demográficos, servindo como ferramenta para triagem inicial de risco de diabetes, identificação de fatores de risco predominantes, e **Desenvolvimento de modelos preditivos em saúde pública**



Variaveis



Variável Alvo.

- Diabetes_binary: Classificação binária (0 = Não diabético, 1 = Diabético)
- Distribuição: 85% não-diabéticos, 15% diabéticos (dataset desbalanceado)

Um dos desafios da base é que a grande maioria dos pacientes não possuem diabetes. Isso dificulta pós as classes são muito desbalanceadas.

Atributos previsores.

1

Condições de Saúde Crônicas, como:

- HighBP: Pressão arterial alta
- HighChol: Colesterol alto
- HeartDiseaseorAttack: Doença cardíaca ou ataque cardíaco
- Stroke: Histórico de AVC

2

Indicadores de Saúde Geral:

- GenHlth: Autoavaliação de saúde (escala 1–5)
- PhysHlth: Dias com saúde física ruim (últimos 30 dias)
- MentHlth: Dias com saúde mental ruim (últimos 30 dias)
- DiffWalk: Dificuldade para caminhar

3

Biometria e Demografia:

- BMI: Índice de Massa Corporal
- Age: Faixa etária (categorizada)
- Sex: Gênero
- Education: Nível de educação
- Income: Renda familiar

Atributos previsores.

4

Comportamentos de Saúde:

- Smoker: Histórico de tabagismo
- PhysActivity: Atividade física regular
- Fruits: Consumo frequente de frutas
- Veggies: Consumo frequente de vegetais
- HvyAlcoholConsump: Consumo pesado de álcool

5

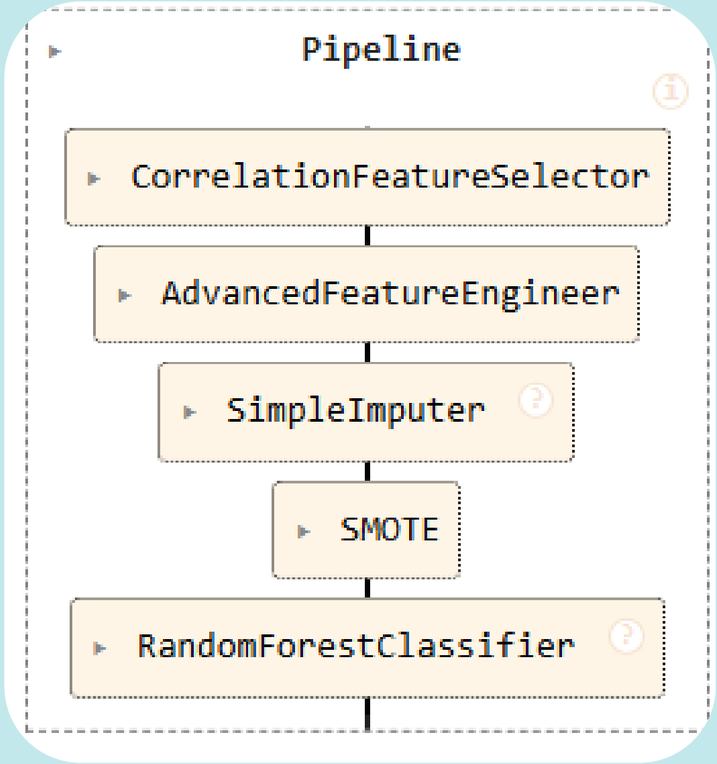
Acesso à Saúde:

- AnyHealthcare: Cobertura de plano de saúde
- NoDocbcCost: Não procurou médico por custo
- CholCheck: Verificação de colesterol nos últimos 5 anos

Metodologia

Pipeline Proposto

Para o nosso problema nos propormos o seguinte pipeline, Que passam por duas classes criada por nÓs que fazem um pre proessamento nos dados. Um SimpleImputer para preencher dados faltantes pela media caso existam, e um Synthetic Minority over-sampling Technique (SMoTE) para tentar lidar com o desbalanceamento dos dados. Seguido por um modelo de classficação.



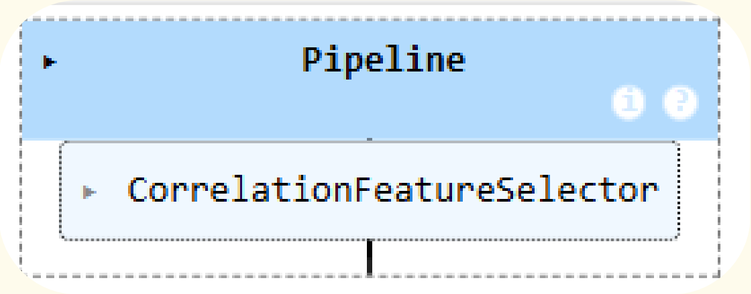
CorrelationFeatureSelection

Na nossa analise sobre os dados encontramos as seguintes correlações:

Correlações com Diabetes:

Diabetes_binary	1.000000
GenHlth	0.293569
HighBP	0.263129
DiffWalk	0.218344
BMI	0.216843
HighChol	0.200276
Age	0.177442
HeartDiseaseorAttack	0.177282
PhysHlth	0.171337
Stroke	0.105816
MentHlth	0.069315
CholCheck	0.064761
Smoker	0.060789
NoDocbcCost	0.031433
Sex	0.031430
AnyHealthcare	0.016255
Fruits	−0.040779
Veggies	−0.056584
HvyAlcoholConsump	−0.057056
PhysActivity	−0.118133
Education	−0.124456
Income	−0.163919

Algumas features tem correlações muito pouco significantes com o atributo alvo, então pouca informação util ela adicionaria com os modelos. Para isso criamos uma classe que faz um corte no valor de 0.1 de correlção descartando qualquer atributo que não tenha correlação igual ou superior a 10% com nosso atributo alvo.



CorrelationFeatureSelection

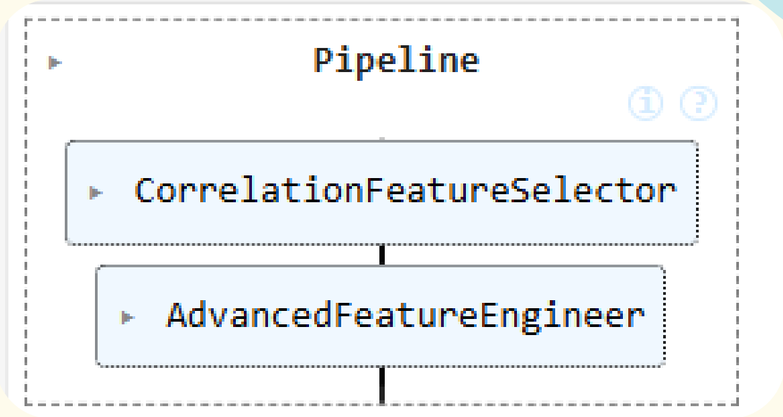
Esta classe cria novas variáveis inteligentes que ajudam o modelo a entender melhor os dados de saúde.

Ela combina informações de forma esperta:

- Junta pressão alta com peso para ver o risco conjunto
- Combina idade com saúde geral para identificar perfis de risco
- Agrupa idades em categorias mais significativas
- Cria razões entre saúde física e mental
- Identifica pacientes com pressão alta E colesterol alto

Por que não é redundante: Em vez de só somar informações que já existem, ela cria novos conceitos que o modelo sozinho teria dificuldade de descobrir, revelando padrões escondidos nos dados.

Resultado: o modelo consegue fazer previsões mais precisas porque enxerga relações complexas que passariam despercebidas.



Classes para lidar com o desbalanceamento.

Esta classe cria novas variáveis inteligentes que ajudam o modelo a entender melhor os dados de saúde.

Ela combina informações de forma esperta:

- Junta pressão alta com peso para ver o risco conjunto
- Combina idade com saúde geral para identificar perfis de risco
- Agrupa idades em categorias mais significativas
- Cria razões entre saúde física e mental
- Identifica pacientes com pressão alta E colesterol alto

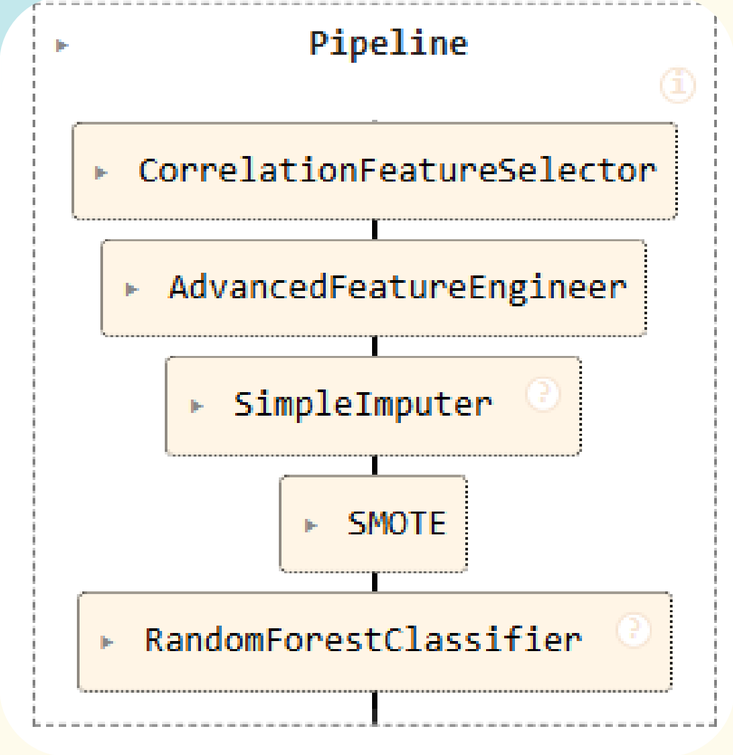
Por que não é redundante: Em vez de só somar informações que já existem, ela cria novos conceitos que o modelo sozinho teria dificuldade de descobrir, revelando padrões escondidos nos dados.

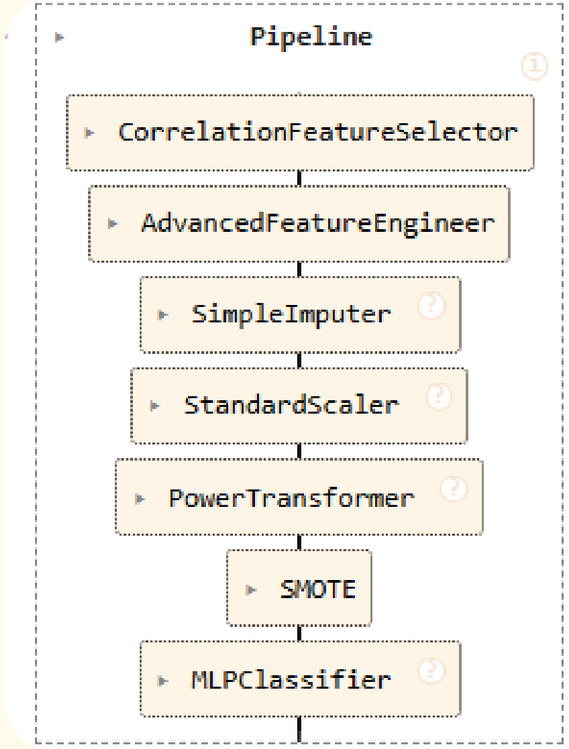
Resultado: o modelo consegue fazer previsões mais precisas porque enxerga relações complexas que passariam despercebidas.

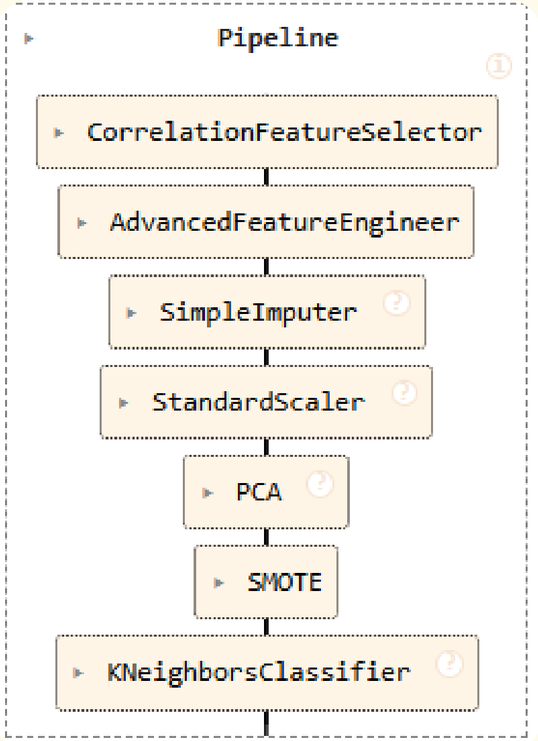
Adicionais praticos para os pipelines de classificação.

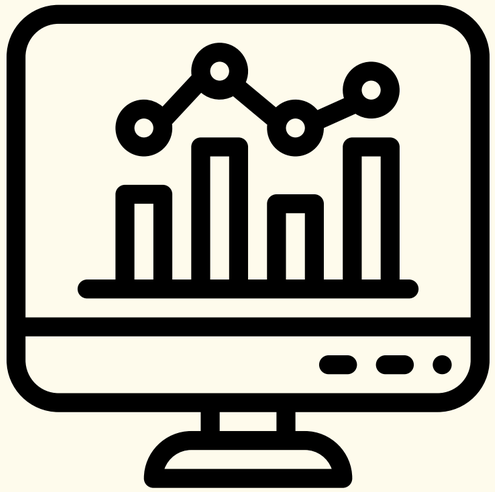
Ainda para o contexto de classificação vamos usar alguns adicionais no pipeline como StandarScaler, ou PCA para ajudar no desenpenho dos modelos. Vamos testar em 3 modelos o RF (Random Forest) KNN e MLP.

Pipelines finais



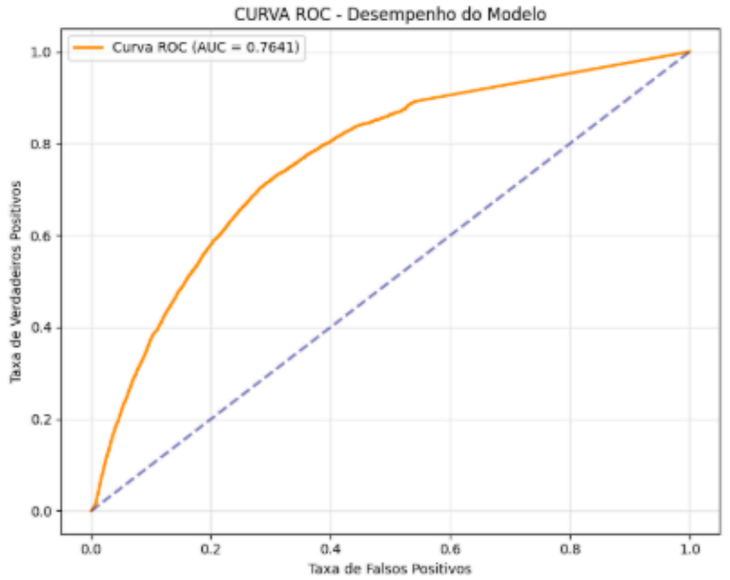


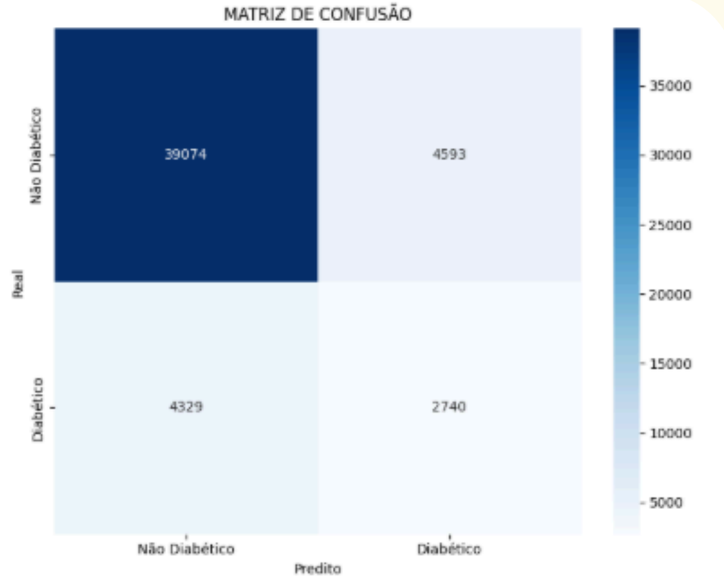


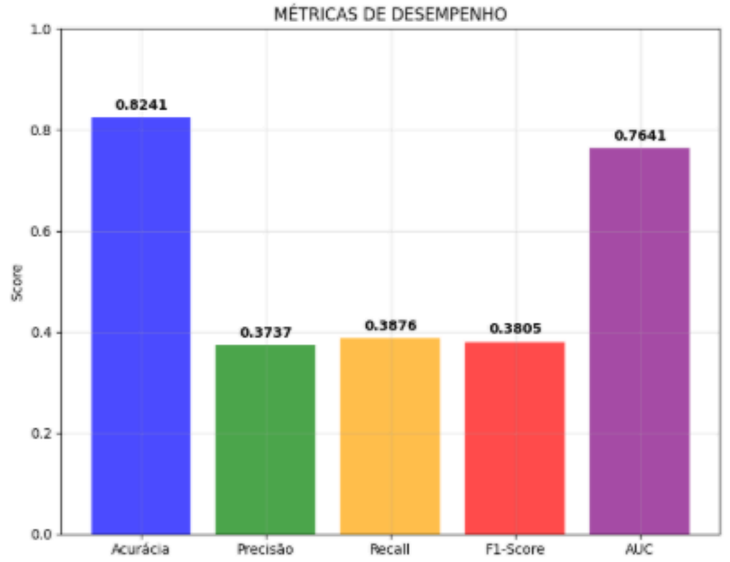


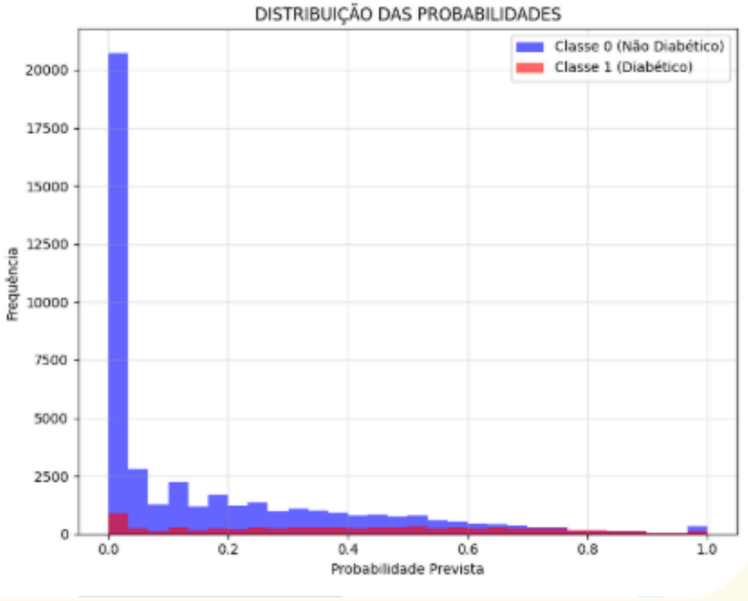
Resultados

KNN


CURVA ROC - Desempenho do Modelo
Curva ROC (AUC = 0.7641)
Taxa de Verdadeiros Positivos
Taxa de Falsos Positivos
0.0
0.2
0.4
0.6
0.8
1.0


MATRIZ DE CONFUSÃO
Não Diabético
Diabético
Real
39074
4593
4329
2740
Não Diabético
Diabético
Predito
35000
30000
25000
20000
15000
10000
5000

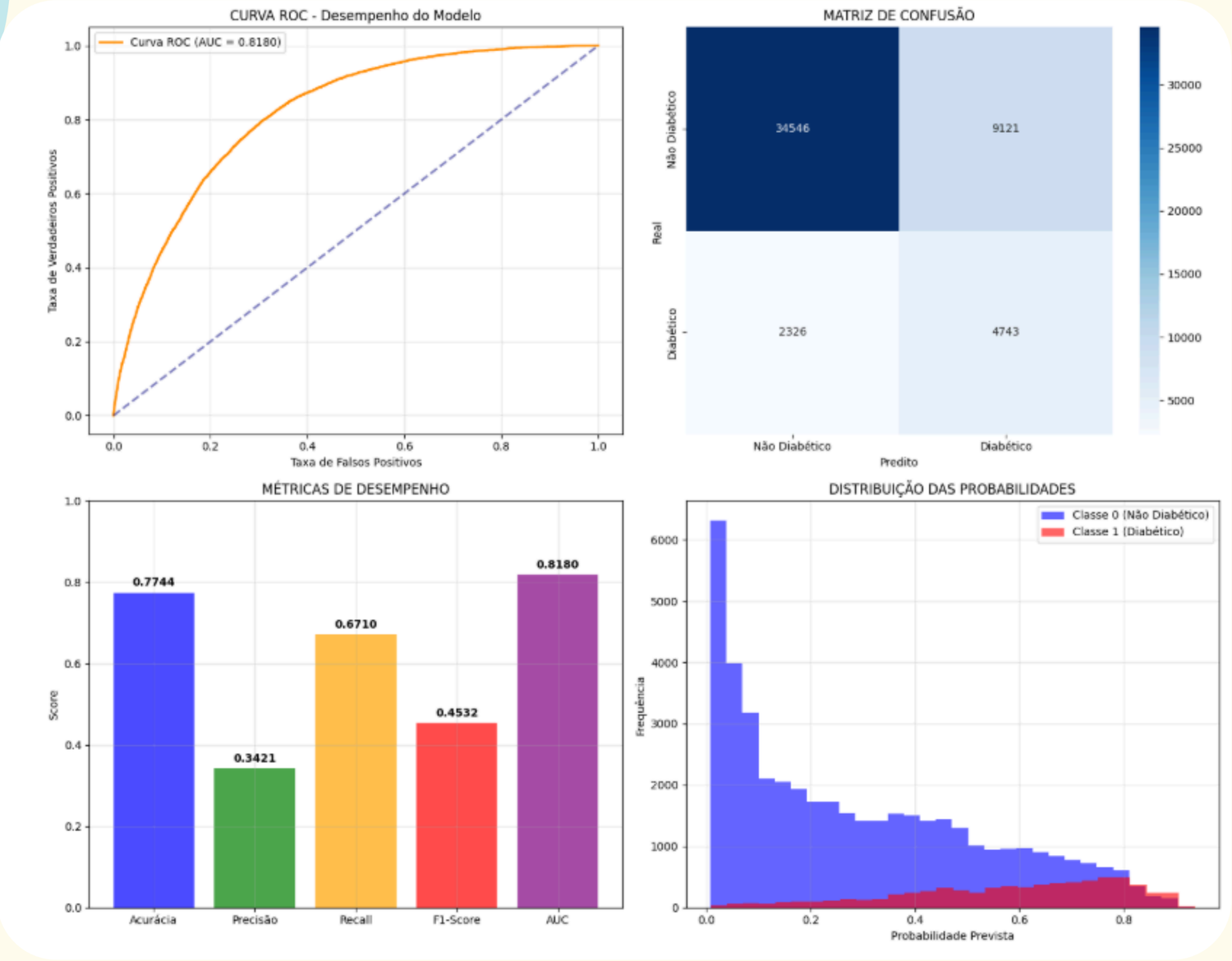

MÉTRICAS DE DESEMPENHO
Score
0.8241
0.3737
0.3876
0.3805
0.7641
Acurácia
Precisão
Recall
F1-Score
AUC


DISTRIBUIÇÃO DAS PROBABILIDADES
Classe 0 (Não Diabético)
Classe 1 (Diabético)
Frequência
Probabilidade Prevista
0.0
0.2
0.4
0.6
0.8
1.0

KNN

o KNN apresentou boa acurácia (82%) e manteve um equilíbrio razoável entre as classes, mas sua capacidade de identificar os pacientes diabéticos foi baixa: precisão de apenas 0.37 e recall de 0.39. Isso significa que o modelo acerta quando diz que um paciente é diabético em cerca de 37% dos casos, e consegue capturar apenas 39% dos verdadeiros diabéticos. Na prática, ele se sai bem em prever os não diabéticos (acima de 89% de acerto), mas deixa passar muitos casos positivos de diabetes. ou seja, é um modelo conservador, mais focado em minimizar falsos positivos do que em capturar casos reais da minoria. o que não é nosso objetivo nesse tipo de problema.

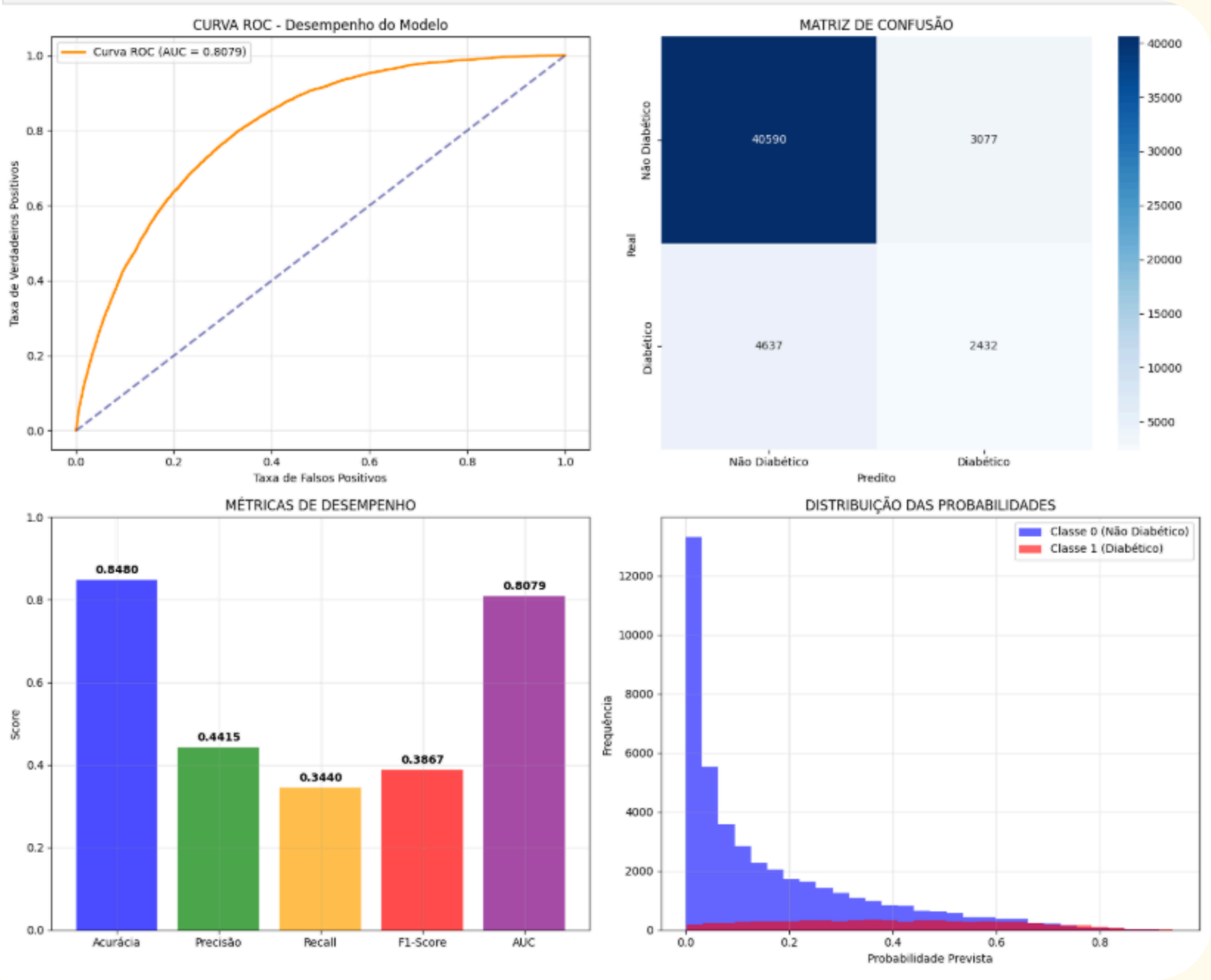
Random Forest



Random Forest

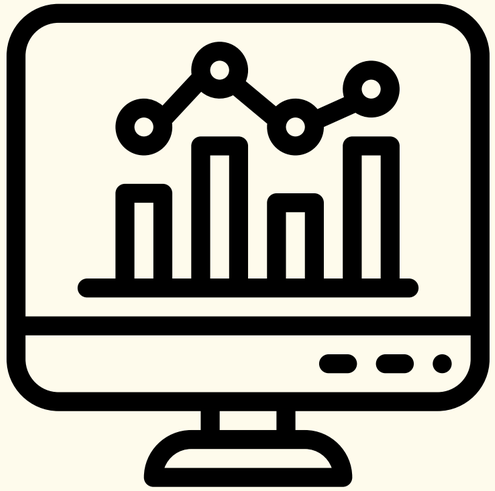
o Random Forest teve a menor acurácia geral (77%), mas foi o modelo que melhor conseguiu capturar os diabéticos (recall 67%), identificando quase 2/3 dos casos verdadeiros, ou seja foi que melhor compriu nosso objetivo. o preço disso foi uma queda na precisão (0.34), indicando que muitos dos casos classificados como diabéticos eram falsos positivos. Isso é comum em datasets desbalanceados quando o algoritmo tenta compensar o peso da classe minoritária. Ainda assim, o AUC foi o maior (0.818), sugerindo que o modelo tem boa capacidade discriminativa. Esse comportamento mostra que ele é mais sensível que os outros, útil em cenários onde não perder casos de diabéticos é mais importante do que evitar alarmes falsos.

MLP


CURVA ROC - Desempenho do Modelo
Curva ROC (AUC = 0.8079)
Taxa de Verdadeiros Positivos
Taxa de Falsos Positivos
MATRIZ DE CONFUSÃO
Não Diabético
Diabético
Real
Predito
40590
3077
4637
2432
MÉTRICAS DE DESEMPENHO
Score
0.8480
0.4415
0.3440
0.3867
0.8079
Acurácia
Precisão
Recall
F1-Score
AUC
DISTRIBUIÇÃO DAS PROBABILIDADES
Classe 0 (Não Diabético)
Classe 1 (Diabético)
Frequência
Probabilidade Prevista

MLP

o MLP alcançou a maior acurácia (85%), mas com desempenho parecido ao KNN em termos de recall da classe positiva, conseguindo capturar apenas 34% dos diabéticos. Apesar disso, teve a melhor precisão (0.44) entre os três, ou seja, quando prevê um caso de diabetes, há maior chance de ser realmente positivo. A rede neural favoreceu muito a classe majoritária, acertando com alta confiança os não diabéticos (93% de recall), mas falhou em identificar a minoria. Isso indica que o MLP está muito enviesado para a classe negativa, funcionando bem para confirmar quem não é diabético, mas pouco confiável para rastreamento da condição. ou seja novamente não foi um bom modelo para prever os casos de diabetes como queriamos.



Análise de Clusterização

Encontrando Perfis de Risco de Diabetes de Forma Não Supervisionada

Metodologia

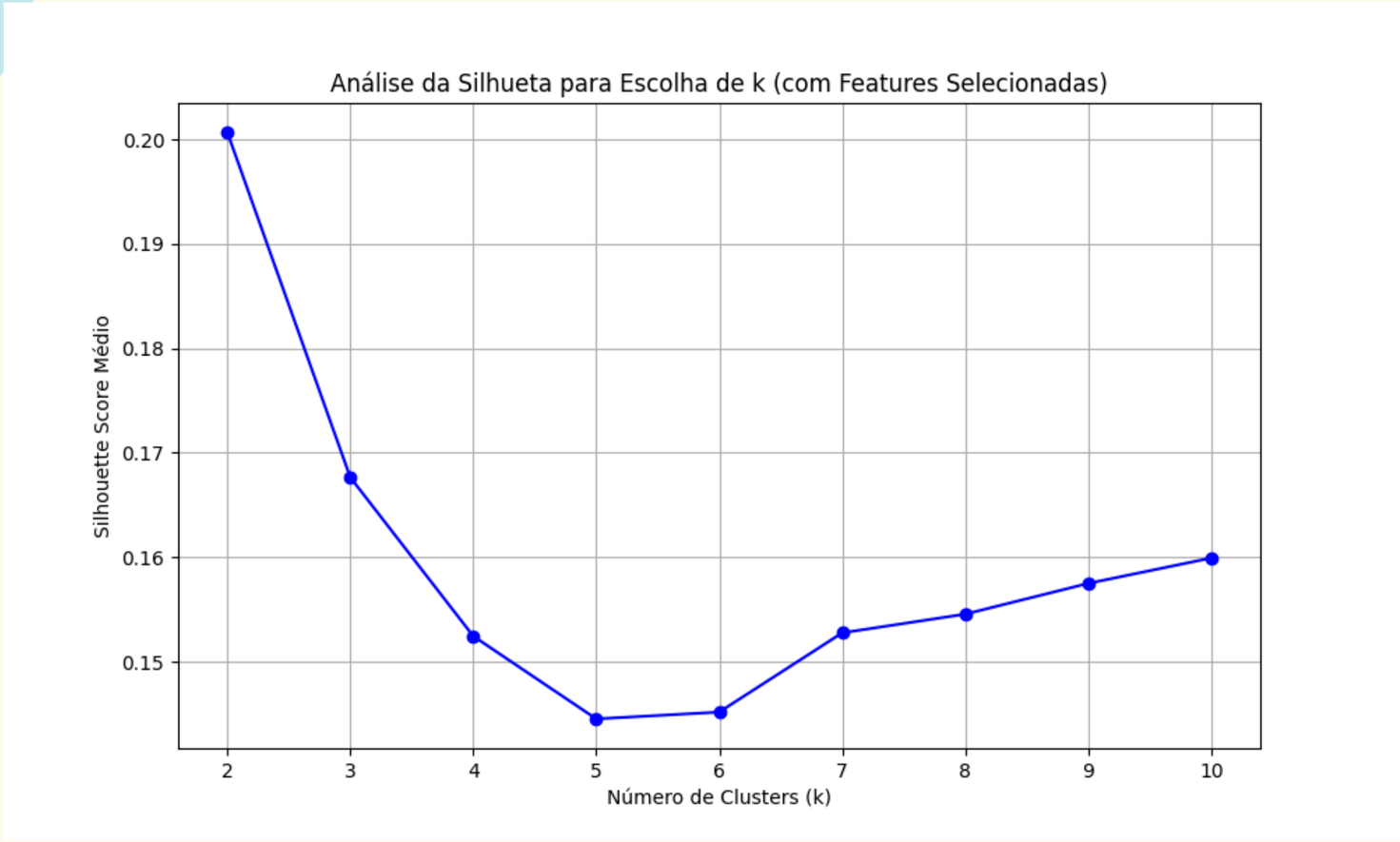
Para a análise não supervisionada, nosso objetivo foi encontrar grupos naturais nos dados. Adotamos uma metodologia em etapas para garantir resultados robustos e significativos.

Padronização dos Dados

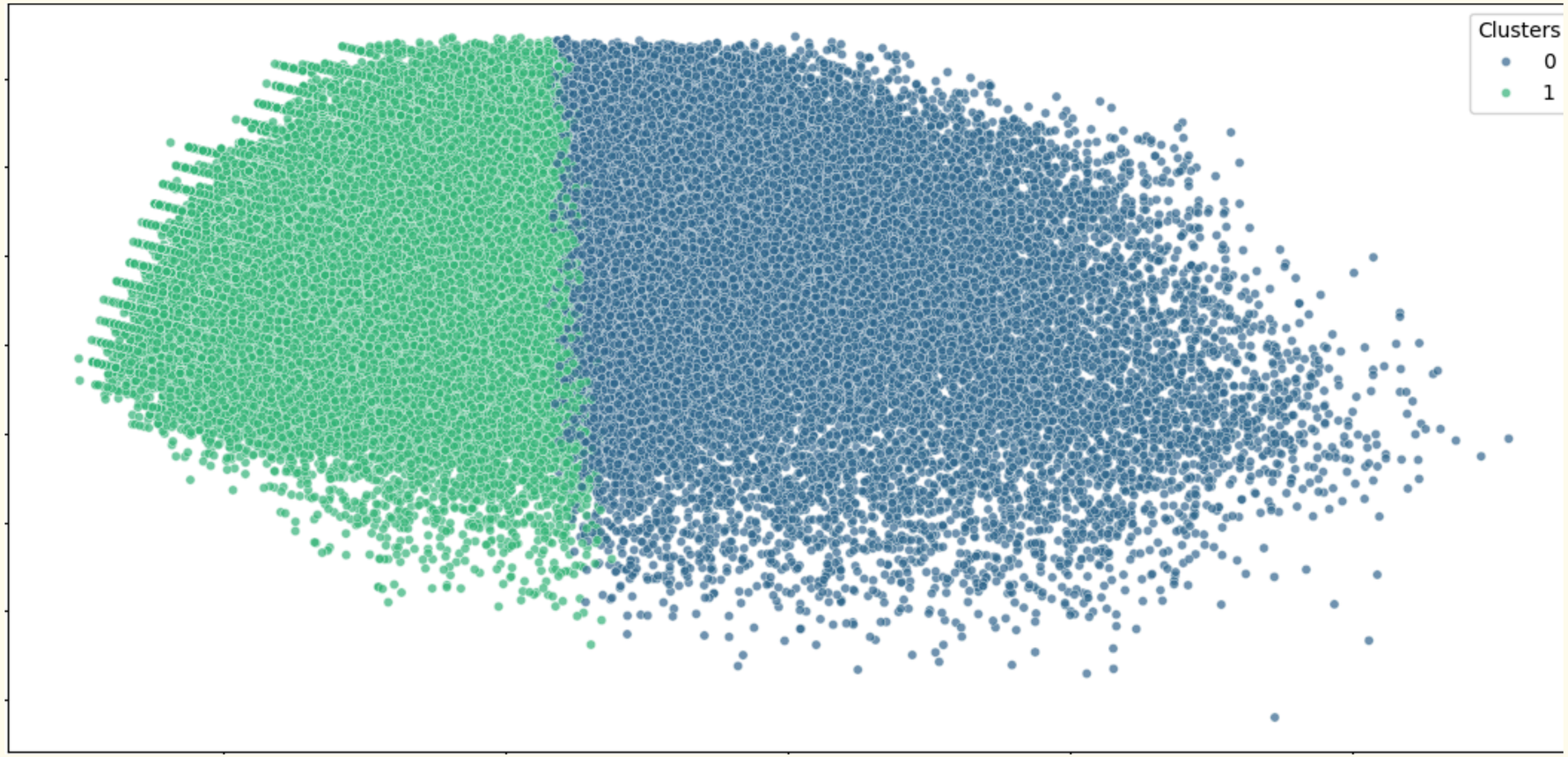
Primeiro, utilizamos o modelo Random Forest para selecionar as features mais preditivas. Com objetivo foi remover o ruído e focar a análise nos indicadores mais relevantes. As features selecionadas foram padronizadas com StandardScaler

Determinação de Clusters

Para encontrar o número ideal de clusters, aplicamos a Análise da Silhueta, que busca o k que maximiza a separação entre os grupos

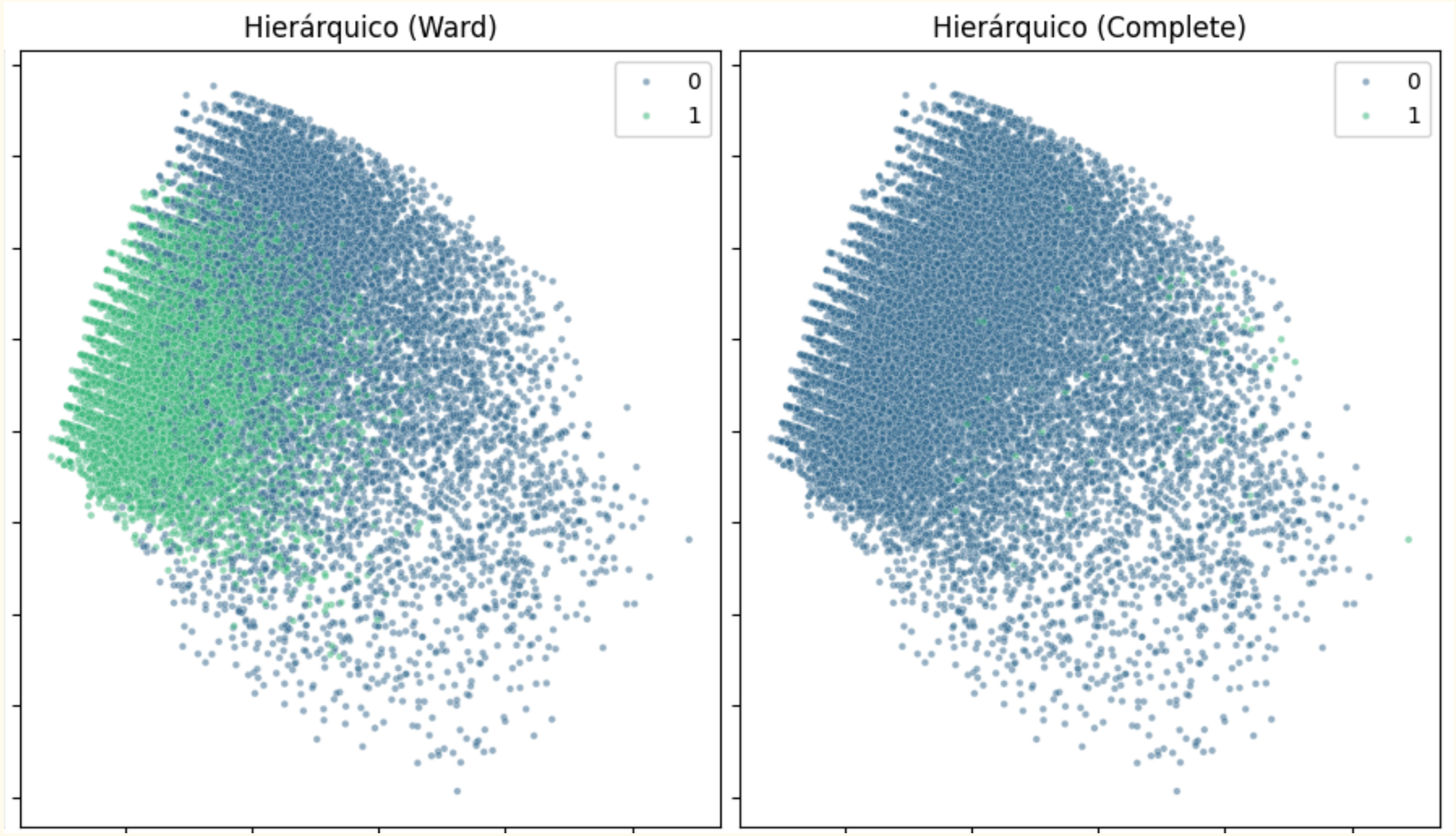


Visualização dos Clusters




```
--- Análise Cruzada: Rótulos Verdadeiros vs. Clusters do K-Means ---
Rótulo Verdadeiro  Diabético  Não Diabético
Cluster K-Means
0                        9999          134425
1                       20670           54906
```

Limitações



Possíveis Melhorias

Engenharia de Features

Modelos Alternativos: XGBoost, LightGBM...

Obrigado Pela Atenção