Diabetes

Predição da doença a partir de dados de saúde

Diabetes

07/12/2023

Geovana Lopes Batista

Isabela Moreira Silva

Mateus Braga Nascimento

Doenças crônicas

- Doenças que não podem ser resolvidas em um curto prazo;
- Não são consideradas emergências médicas;
- Representam por cerca de 7 em cada 10 mortes por doenças no mundo.

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Diabetes

- Doença metabólica cuja característica é o aumento de glicose no sangue;
- Diabetes tipo I: diminuição ou completa ausência da produção de insulina pelo pâncreas;
- Diabetes tipo II: resistência à insulina e/ou diminuição de produção pelo pâncreas.

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Diabetes

- Segundo dados da OMS:
  - 6ª maior causa de morte nas Américas, causando mais de 284 mil mortes em 2019;
  - Aumento de 70% das mortes por diabetes entre 2000 e 2019.

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Diabetes

- Alta incidência relacionada ao sobrepeso e obesidade, alimentação inadequada e falta de atividade física;
- Pode levar a: cegueira, amputação de membros inferiores, doenças renais, cardíacas e câncer;
- Diabetes tipo II, quando descoberta no estágio inicial de pré-diabetes, pode ser revertida.

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Diabetes e a Saúde Pública

- Gastos com doenças crônicas foi de, aproximadamente, 3,45 bilhões de reais no SUS em 2018;
- 30% desse valor (≈ 1 bilhão de reais) foi gasto no tratamento de diabetes;
- Ferramentas que auxiliem no diagnóstico precoce podem melhorar a qualidade de vida de muitas pessoas além de diminuir gastos em saúde pública.

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Alguns dados

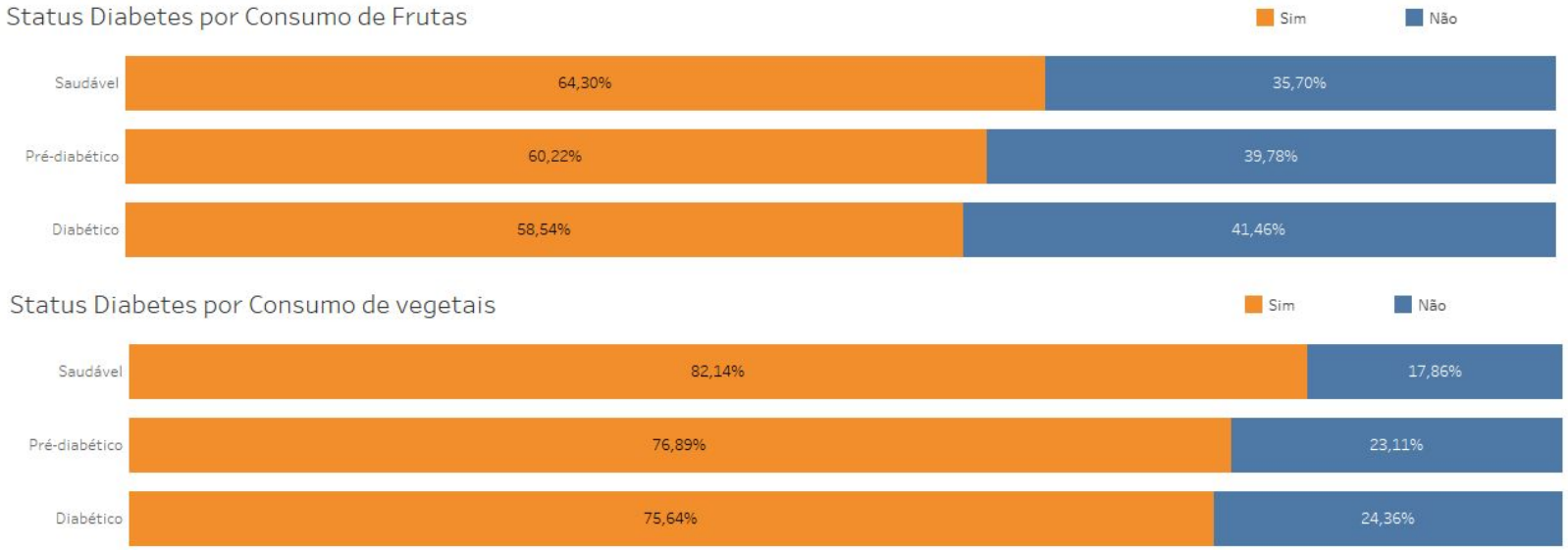
- Dataset obtido a partir do Kaggle, com 253.580 linhas e 21 variáveis.

Status Diabetes por IMC, Saúde Física e Mental

Diabetes status	IMC	Saúde Física	Saúde Mental
Saudável	27,7	4	3
Pré-diabético	30,7	6	5
Diabético	31,9	8	4

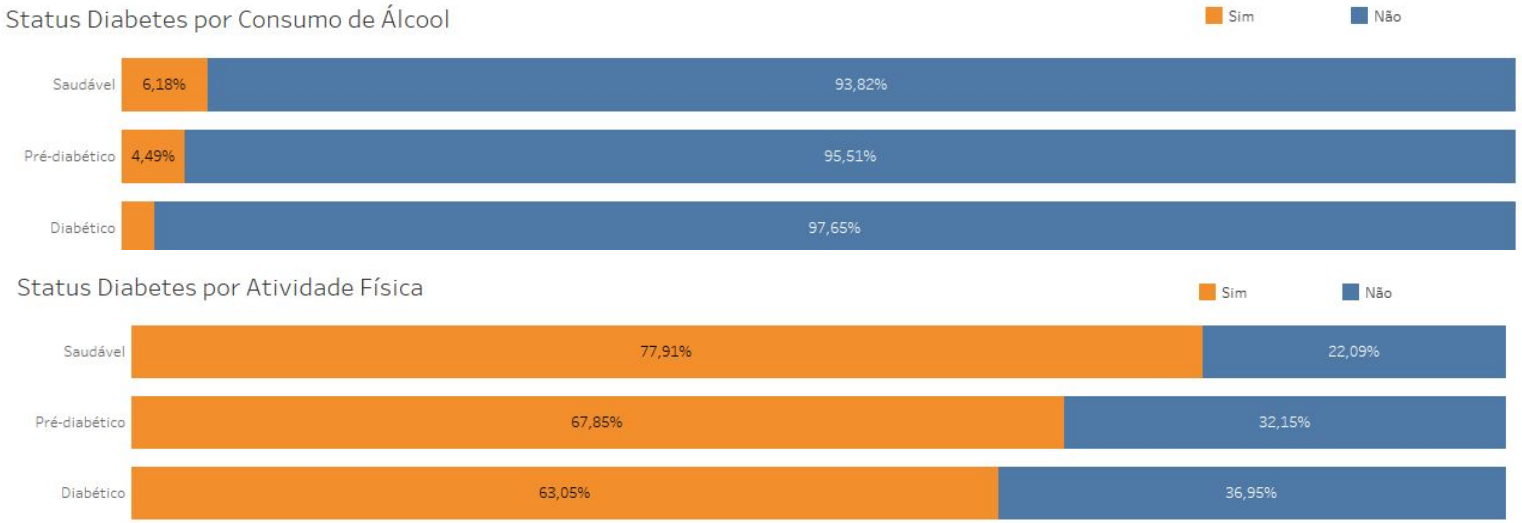
INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Alguns dados



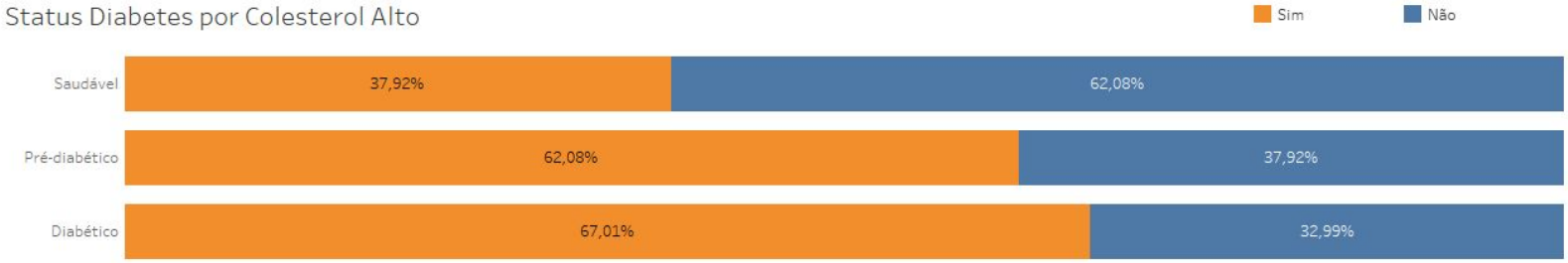
INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Alguns dados



INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Alguns dados



Status diabetes por Faixa Etária

Diabetes status	18-24 anos	25-29 anos	30-34 anos	35-39 anos	40-44 anos	45-49 anos	50-54 anos	55-59 anos	60-64 anos	65-69 anos	70-74 anos	75-79 anos	80+ anos
Saudável	5.601	7.404	10.737	13.055	14.943	17.765	22.808	26.019	26.809	24.939	17.790	12.132	13.701
Pré-diabético	21	54	72	142	163	312	418	550	702	697	602	445	453
Diabético	78	140	314	626	1.051	1.742	3.088	4.263	5.733	6.558	5.141	3.403	3.209

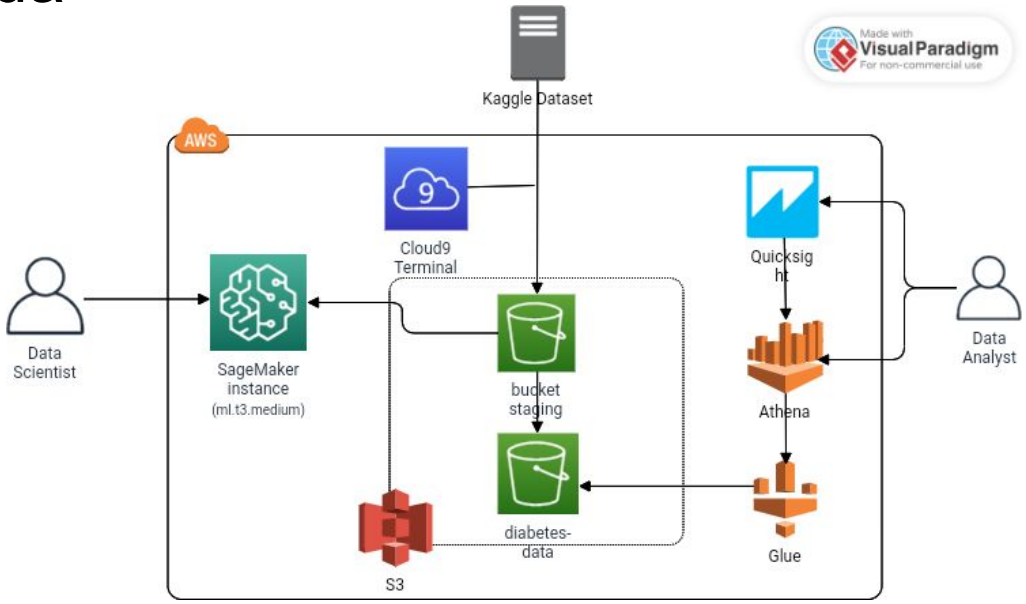
INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Arquitetura AWS Explorada

Utilizamos um bucket S3 para receber o arquivo original do dataset. Usando Cloud9 Terminal, descompactar e copiar para um segundo bucket o csv de interesse.

Através do Athena, criamos um database no Glue que lê o arquivo csv como uma tabela externa. O database foi enriquecido com views que adicionavam descrição às variáveis.

Através do sagemaker, o dado bruto foi utilizado dentro de um notebook jupyter para treinamento e teste dos modelos apresentados.

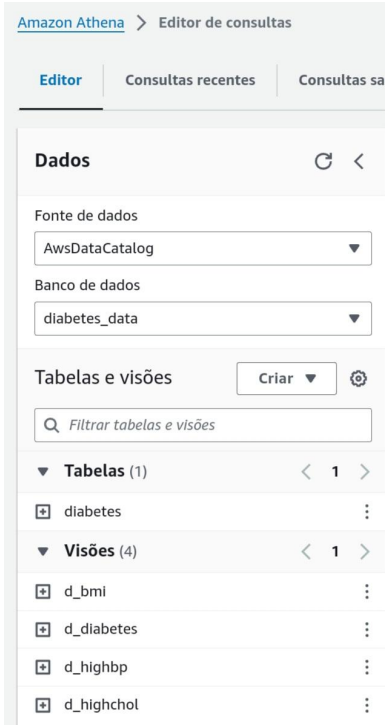


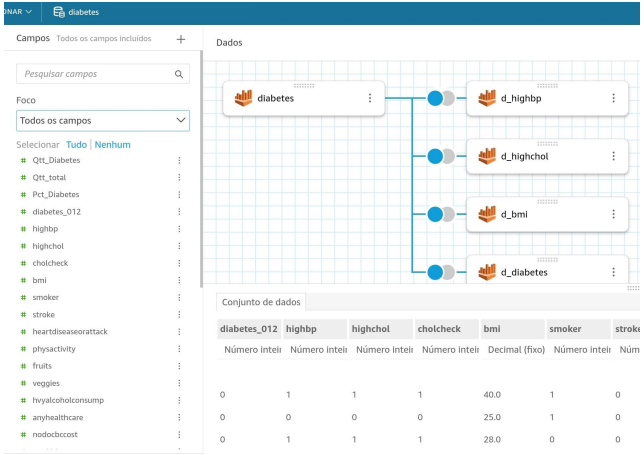
INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Análise de dados com Athena e Quicksight

Athena facilita a análise de dados com SQL nativo, bem como na criação de catálogos de dados serverless no Glue. Uma vez publicado via Glue, os catálogos ficam disponíveis por uma variadade de serviços AWS, dentre eles o Quicksight.

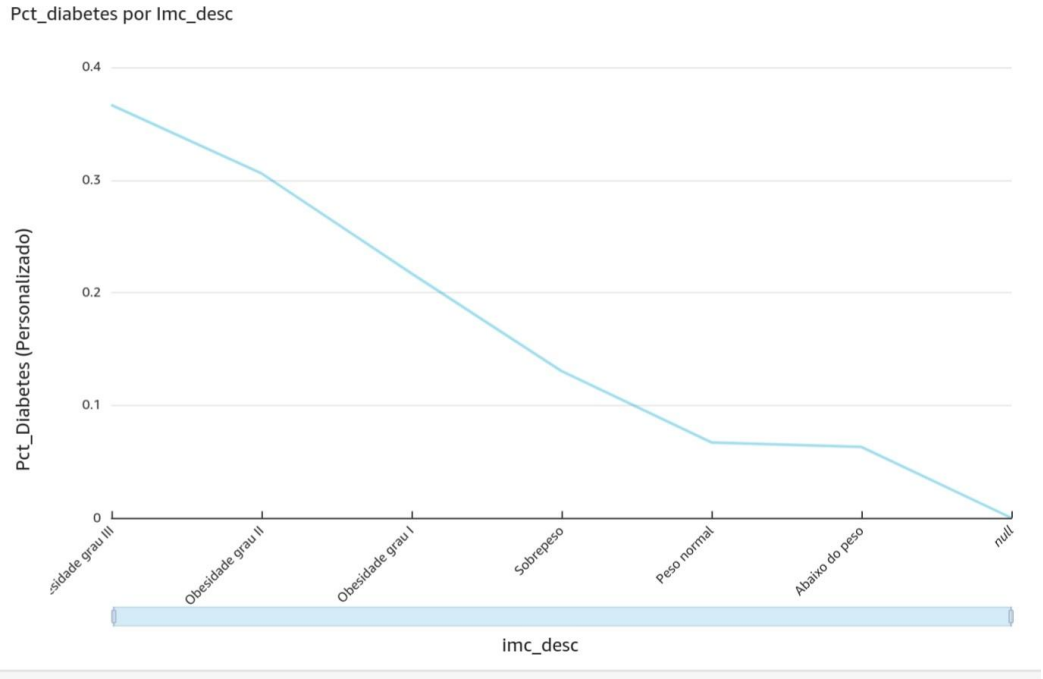
No Quicksight, a definição de modelos de dados analíticos é simples e visual, habilitando a criaçao de relatórios e visualizações de forma rápida.

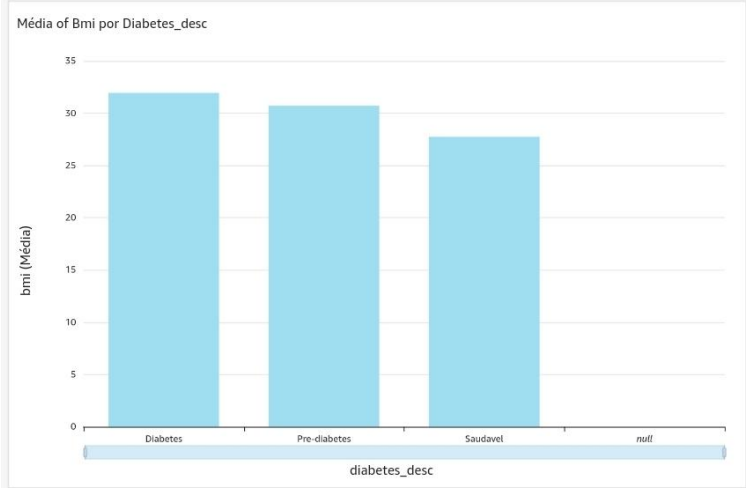




INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Visualizações com o Quicksight





INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Modelos de Classificação

- KNN
- Regressão Logística
- Árvores de Decisão

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Variáveis

Diabetes: 0 (sem diabetes), 1 (pré-diabetes) e 2 (diabetes)
HighBP: pressão alta (0 - não, 1 - sim)
HighChol: colesterol alto (0 - não, 1 - sim)Diabetes: 0 (sem diabetes), 1 (pré-diabetes) e 2 (diabetes)
BMI: Índice de Massa Corpórea - IMC:
abaixo de 18,49: abaixo do peso
entre 18,5 e 24,99: peso normal
entre 25 e 29,99: sobrepeso
entre 30 e 34,99: obesidade grau I
entre 35 e 39,99: obesidade grau II
acima de 40: obesidade grau III
HeartDiseaseorAttack: doença coronariana ou infarto do miocárdio (0 - não, 1 - sim)

PhysActivity: atividade física nos últimos 30 dias (0 - não, 1 - sim)
Fruits: consome fruta uma ou mais vezes por dia (0 - não, 1 - sim)
Veggies: consome vegetais uma ou mais vezes por dia (0 - não, 1 - sim)
HvyAlcoholConsump: alto consumo de álcool - homens adultos mais de 14 doses por semana, mulheres adultas mais de 7 doses por semana (0 - não, 1 - sim)

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

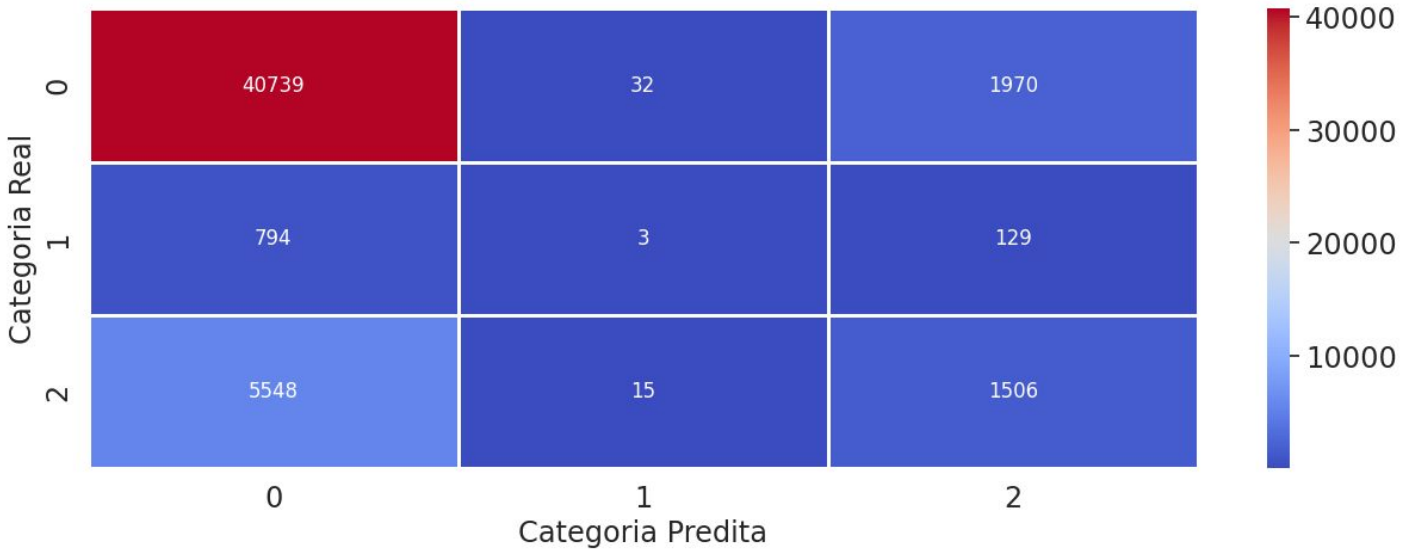
Variáveis

GenHlth: saúde em geral (1 = excelente, 2 = muito boa, 3 = boa, 4 = razoável, 5 = ruim)
MentHlth: por quantos dias, nos últimos 30 dias, teve algum problema de saúde mental (estresse, depressão e problemas com emoções)

PhysHlth: por quantos dias, nos últimos 30 dias, ficou doente ou teve algum ferimento
DiffWalk: dificuldade para caminhar ou subir escadas (0 - não, 1 - sim)
Sex: sexo (0 - feminino, 1 - masculino)
Age: idade (1 = 18-24 anos, 2 = 25-29 anos, 3 = 30-34 anos, 4 = 35-39 anos, 5 = 40-44anos, 6 = 45-49 anos, 7 = 50-54 anos, 8 = 55-59 anos, 9 = 60-64 anos, 10 = 65-69 anos, 11 = 70-74 anos, 12 = 75-79 anos, 13 = 80+ anos)

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

KNN



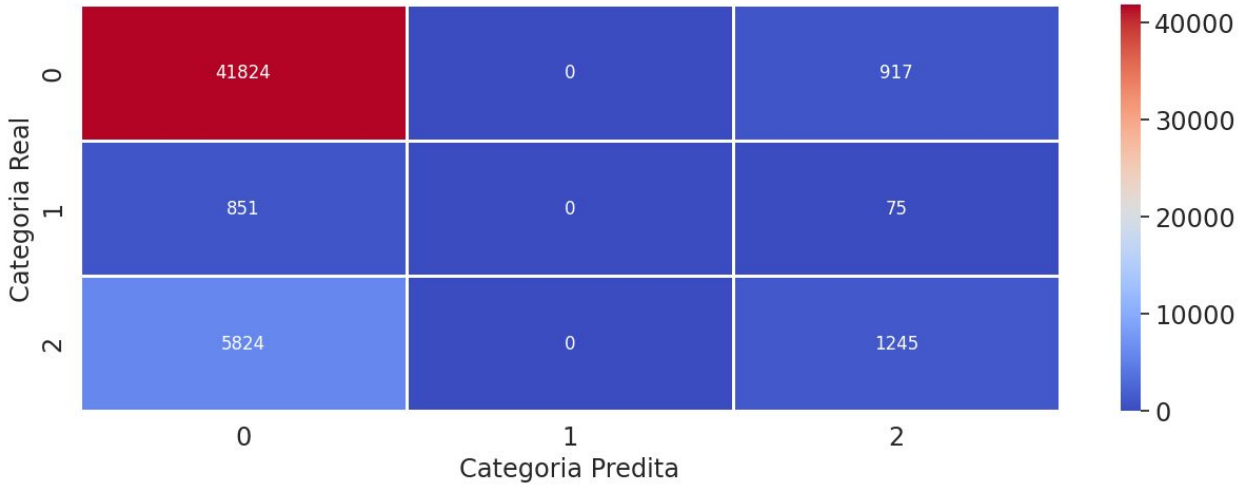
INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

KNN

	precision	recall	f1-score	support
0.0	0.87	0.95	0.91	42741
1.0	0.06	0.00	0.01	926
2.0	0.42	0.21	0.28	7069
accuracy			0.83	50736
macro avg	0.45	0.39	0.40	50736
weighted avg	0.79	0.83	0.80	50736

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Regressão Logística



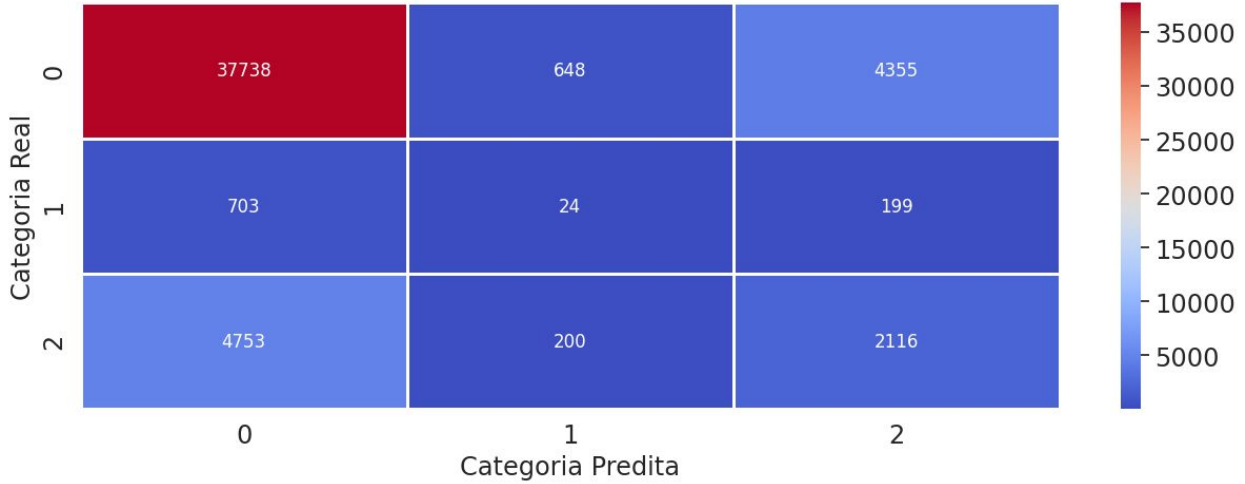
INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Regressão Logística

	precision	recall	f1-score	support
0.0	0.86	0.98	0.92	42741
1.0	0.00	0.00	0.00	926
2.0	0.56	0.18	0.27	7069
accuracy			0.85	50736
macro avg	0.47	0.38	0.39	50736
weighted avg	0.80	0.85	0.81	50736

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Árvores de Decisão



INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Árvores de Decisão

	precision	recall	f1-score	support
0.0	0.87	0.88	0.88	42741
1.0	0.03	0.03	0.03	926
2.0	0.32	0.30	0.31	7069
accuracy			0.79	50736
macro avg	0.41	0.40	0.40	50736
weighted avg	0.78	0.79	0.78	50736

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Conclusões

Embora os modelos tenham alcançado alta acurácia e precisão, a matriz de confusão revela um viés significativo para a categoria de pessoas sem diabetes, devido ao desbalanceamento da base de dados.

No estudo em questão, é preferível o erro do tipo I (falso positivo) para incentivar a busca por ajuda médica.

A Árvore de Decisão se aproximou mais das necessidades, mas ainda requer aprimoramento para reduzir o erro do tipo II (falso negativo).

Futuramente, planeja-se explorar novos modelos, como redes neurais, e equilibrar a distribuição dos dados. Após o aperfeiçoamento, o objetivo é expandir o modelo para outras doenças crônicas.

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

MUITO OBRIGADA!