Trabalho Mineração de Dados - Previsão de Diabetes

Alunos: Rafael Martins e Enzo Innecco

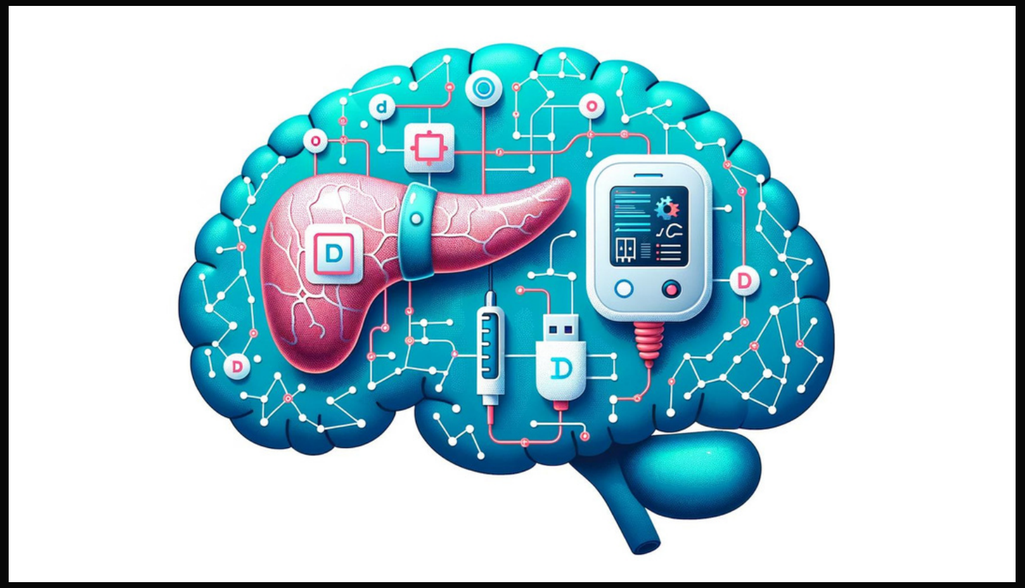
Diabetes

A diabetes é uma doença crônica que ocorre quando o corpo não consegue regular adequadamente os níveis de açúcar no sangue. Isso pode ser devido à produção insuficiente de insulina (um hormônio que regula a glicose no sangue) ou à incapacidade do corpo de usar a insulina efetivamente.

A diabetes é um problema sério de saúde porque pode levar a complicações graves, como danos ao coração, artérias, olhos, rins e nervos. Em casos extremos, pode até levar à morte. Portanto, é crucial entender e gerenciar essa condição para manter uma vida saudável.

Objetivo

Portanto, o objetivo do projeto é criar modelos de predição usando Machine Learning para prever se um paciente, de acordo com suas características, tende a ter diabetes ou não, além de descobrir quais características que mais contribuem e influenciam uma pessoa a ter diabetes. Com isso, os hospitais poderão implementar esses modelos em seus sistemas para identificar precocemente os pacientes em risco de desenvolver diabetes.



Base de Dados

Selecionamos um conjunto que contém uma coleção de 100 mil dados médicos e demográficos de pacientes, juntamente com seu estado de diabetes.

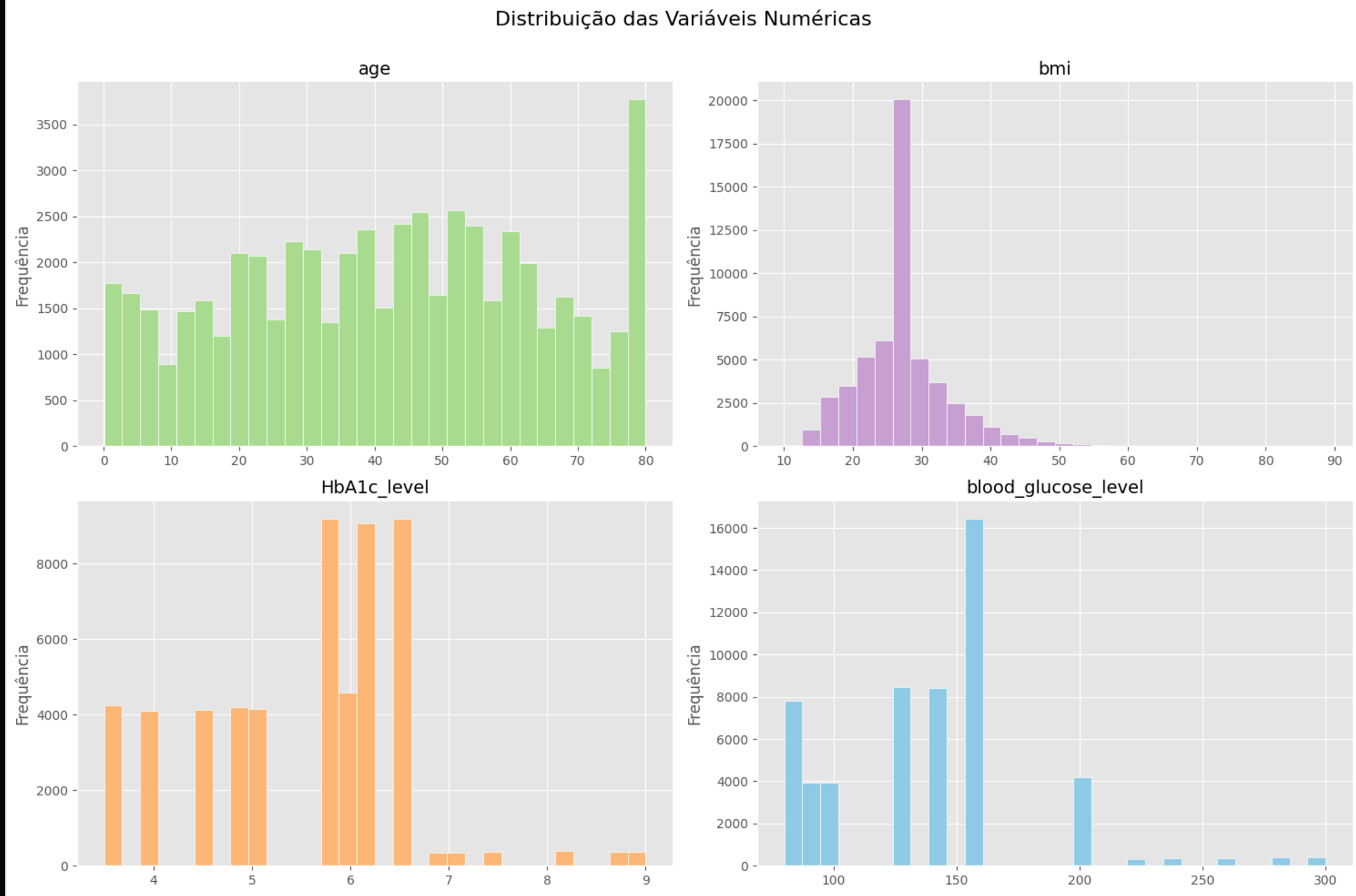
gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Female	80.0	0	1	never	25.19	6.6	140	0
Female	54.0	0	0	No Info	27.32	6.6	80	0
Male	28.0	0	0	never	27.32	5.7	158	0
Female	36.0	0	0	current	23.45	5.0	155	0
Male	76.0	1	1	current	20.14	4.8	155	0
...	...	...	...	...	...	...	...	...
Female	80.0	0	0	No Info	27.32	6.2	90	0
Female	2.0	0	0	No Info	17.37	6.5	100	0
Male	66.0	0	0	former	27.83	5.7	155	0
Female	24.0	0	0	never	35.42	4.0	100	0
Female	57.0	0	0	current	22.43	6.6	90	0

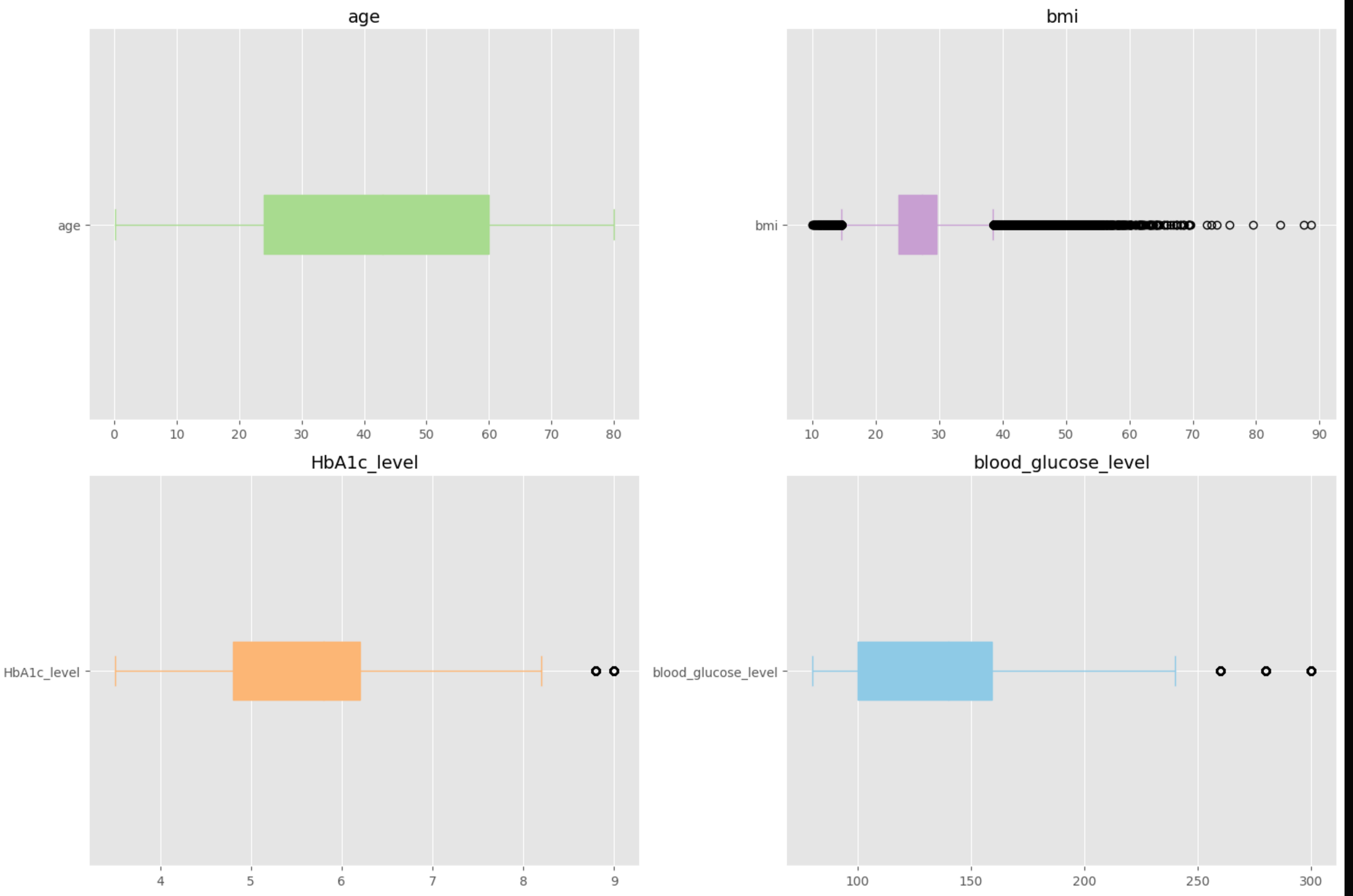
- Gênero
- Idade
- Hipertensão
- Doenças Cardíacas
- Histórico de Tabagismo
- IMC
- Nível de HbA1c
- Nível de Glicose
- Diabetes

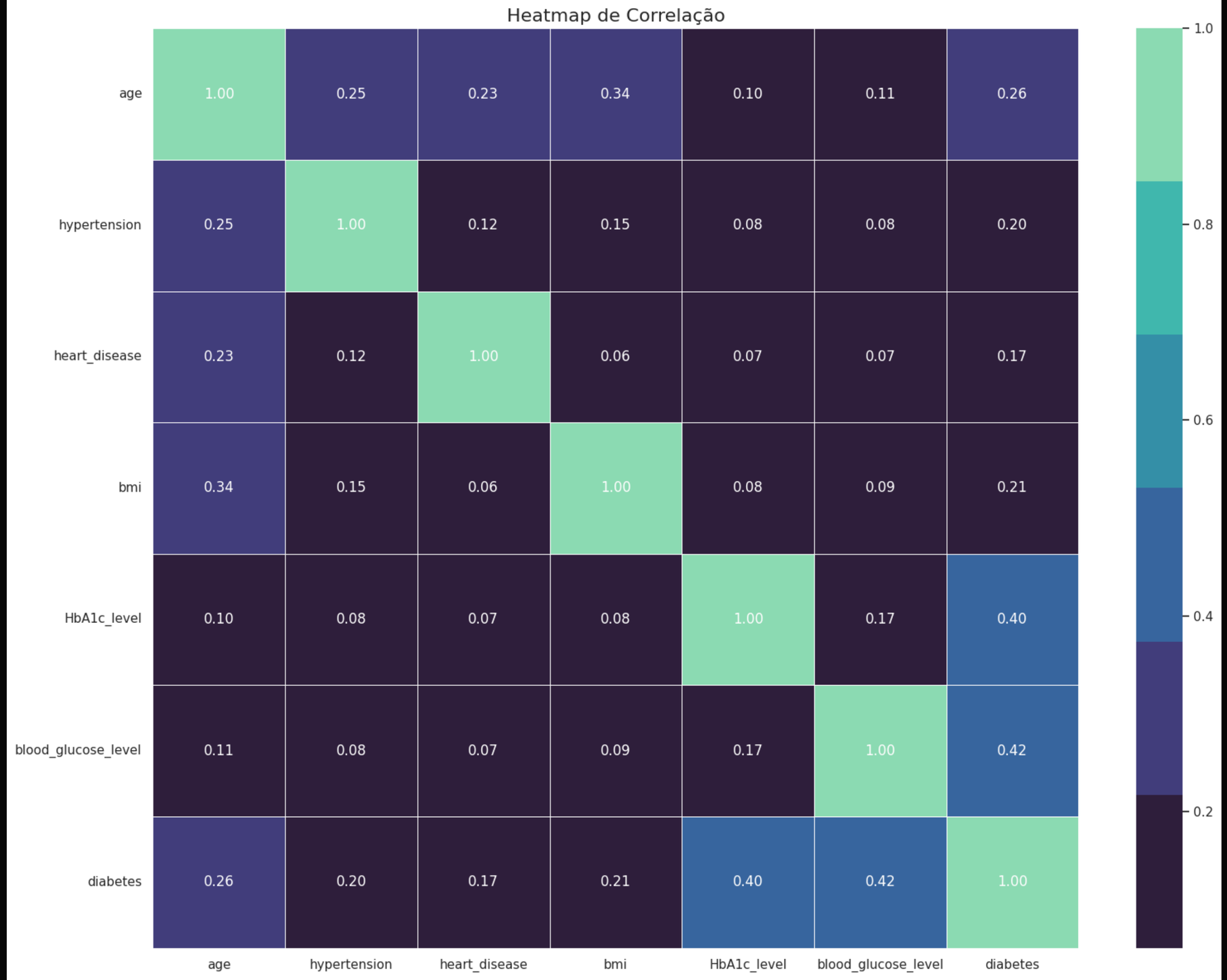
Pré-Processamento

- A base não possui valores NA
- Aplicamos o One-Hot Encoding nas variáveis categóricas

smoking_history_current	smoking_history_ever	smoking_history_former	smoking_history_never	smoking_history_not current
0	0	0	1	0
0	0	0	0	0
0	0	0	1	0
1	0	0	0	0
1	0	0	0	0


Distribuição das Variáveis Numéricas
age
Frequência
0
500
1000
1500
2000
2500
3000
3500
0
10
20
30
40
50
60
70
80
bmi
Frequência
0
2500
5000
7500
10000
12500
15000
17500
20000
10
20
30
40
50
60
70
80
90
HbA1c_level
Frequência
0
2000
4000
6000
8000
4
5
6
7
8
9
blood_glucose_level
Frequência
0
2000
4000
6000
8000
10000
12000
14000
16000
100
150
200
250
300


age
age
0
10
20
30
40
50
60
70
80
bmi
bmi
10
20
30
40
50
60
70
80
90
HbA1c_level
HbA1c_level
4
5
6
7
8
9
blood_glucose_level
blood_glucose_level
100
150
200
250
300


Heatmap de Correlação
age
hypertension
heart_disease
bmi
HbA1c_level
blood_glucose_level
diabetes
1.00 0.25 0.23 0.34 0.10 0.11 0.26
0.25 1.00 0.12 0.15 0.08 0.08 0.20
0.23 0.12 1.00 0.06 0.07 0.07 0.17
0.34 0.15 0.06 1.00 0.08 0.09 0.21
0.10 0.08 0.07 0.08 1.00 0.17 0.40
0.11 0.08 0.07 0.09 0.17 1.00 0.42
0.26 0.20 0.17 0.21 0.40 0.42 1.00
1.0
0.8
0.6
0.4
0.2

Normalização

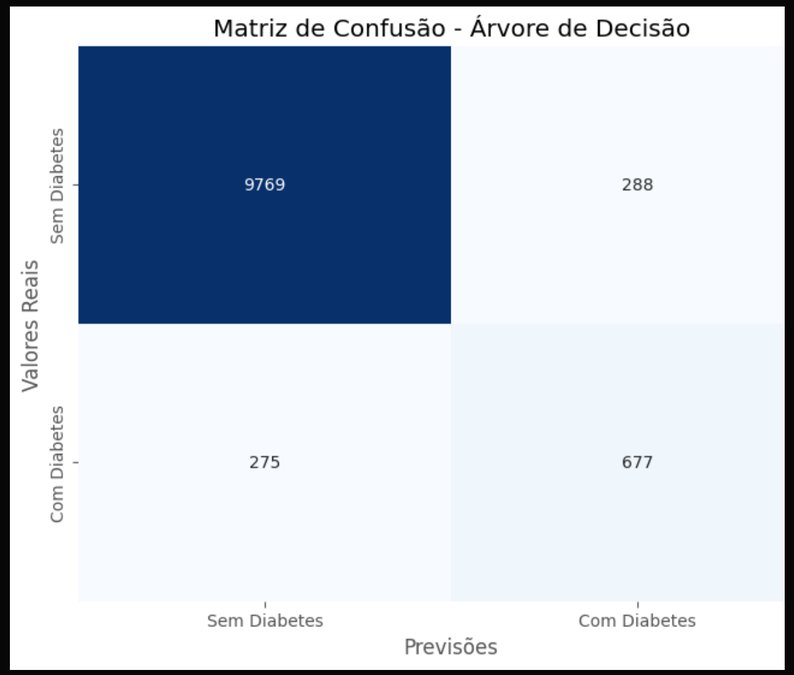
Por conta do o intervalo e a escala dessas variáveis, normalizamos os dados para garantir que todas as características tenham o mesmo peso nos algoritmos de aprendizado de máquina. Utilizamos a normalização Min-Max, que transforma os dados para terem um valor mínimo de 0 e um máximo de 1.

	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes
0	1.000000	0.0	1.0	0.192860	0.563636	0.272727	0.0
1	0.674675	0.0	0.0	0.219921	0.563636	0.000000	0.0
2	0.349349	0.0	0.0	0.219921	0.400000	0.354545	0.0
3	0.449449	0.0	0.0	0.170753	0.272727	0.340909	0.0
4	0.949950	1.0	1.0	0.128700	0.236364	0.340909	0.0

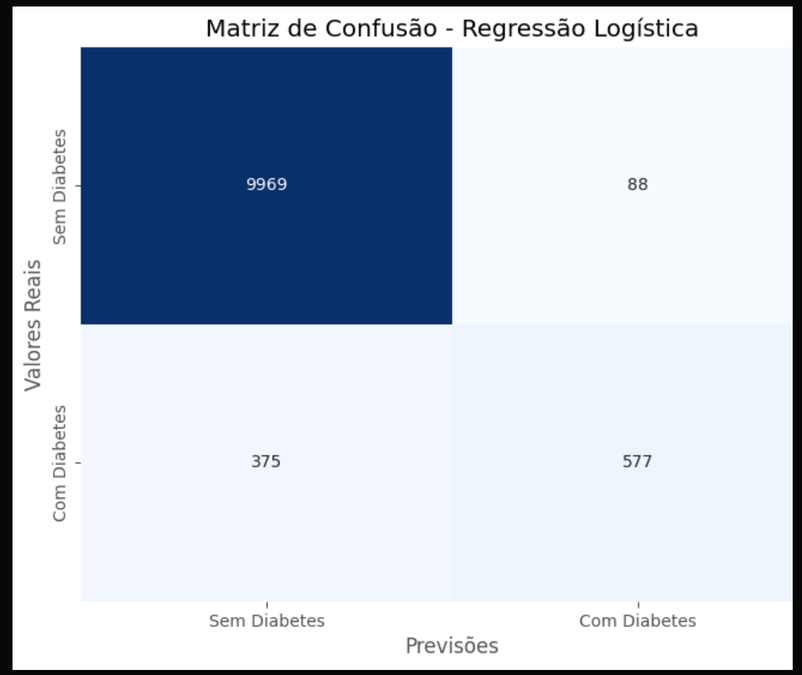
Modelos usados

Foram usados 3 modelos de Machine Learning: Árvore de Decisão, Regressão Logística e KNN

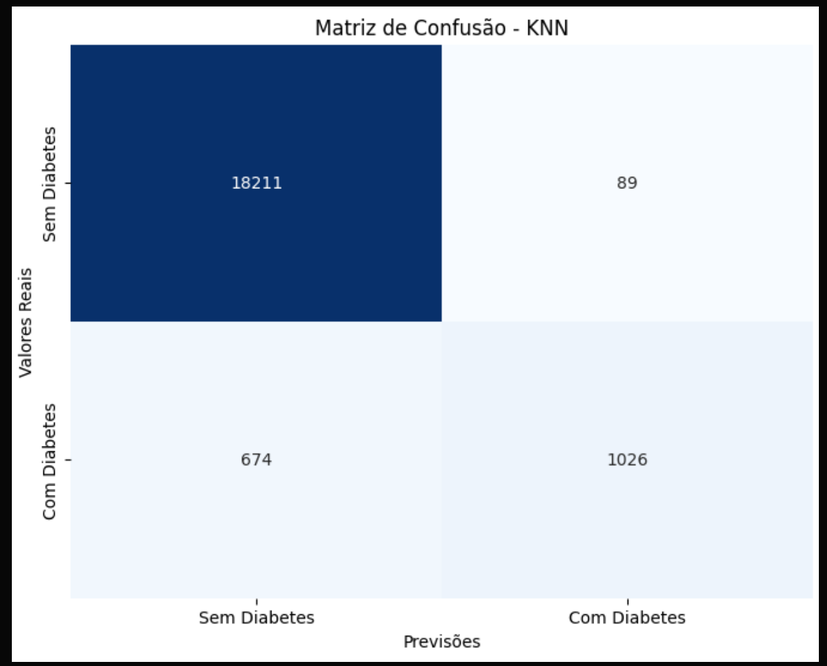
Acurácia: 94.89%

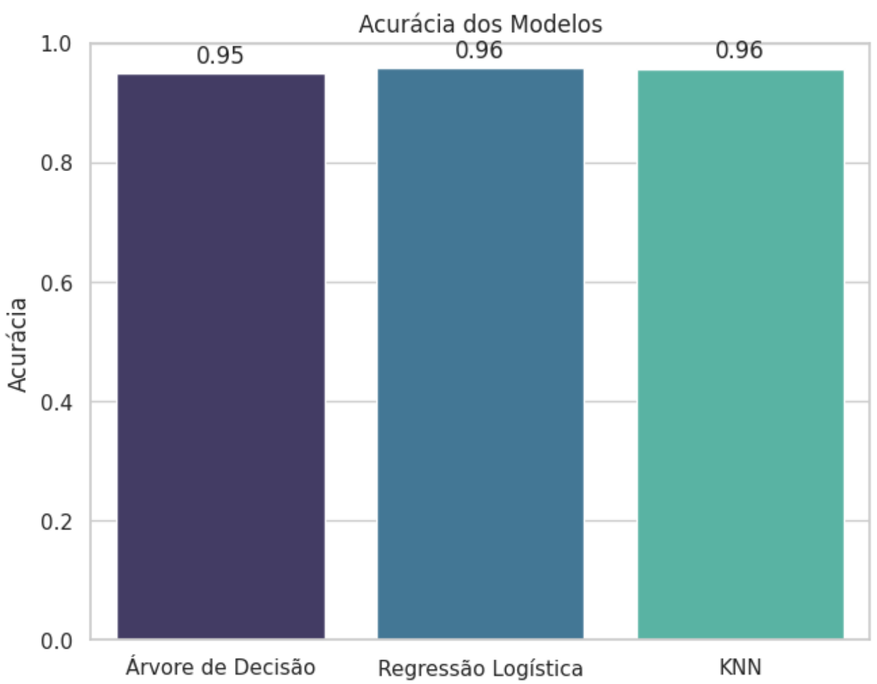


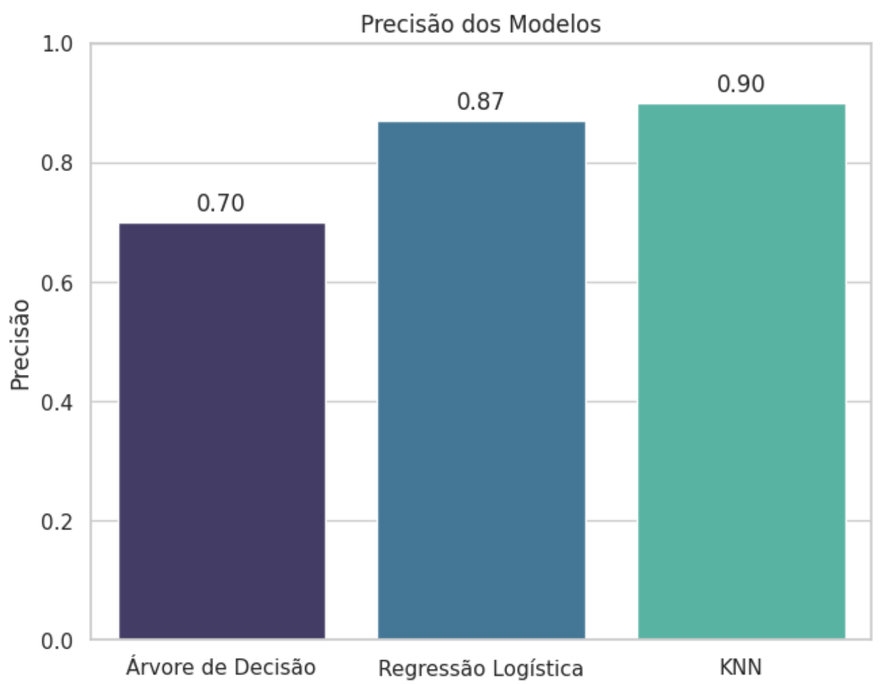
Acurácia: 95.80%

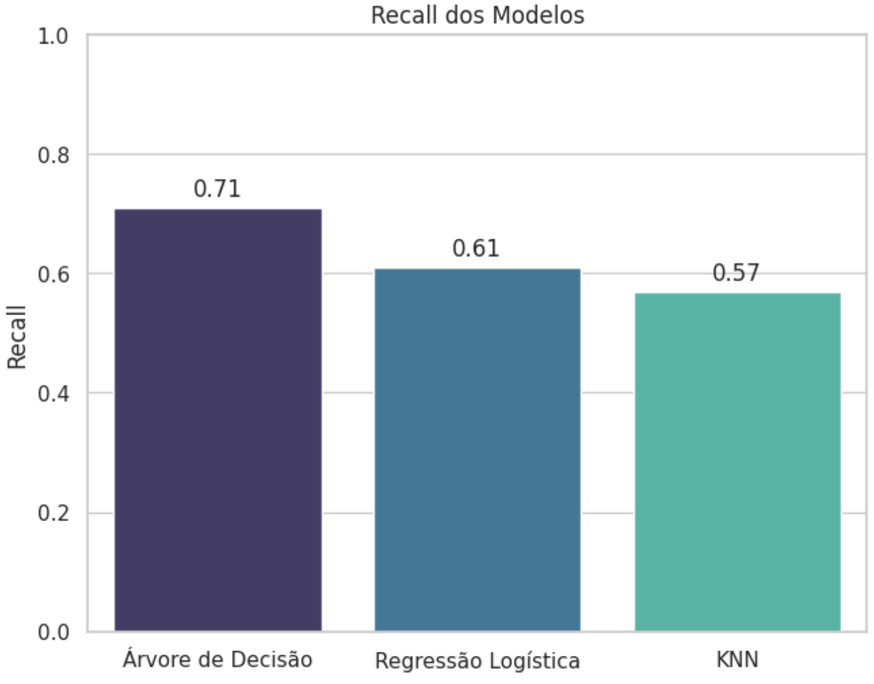


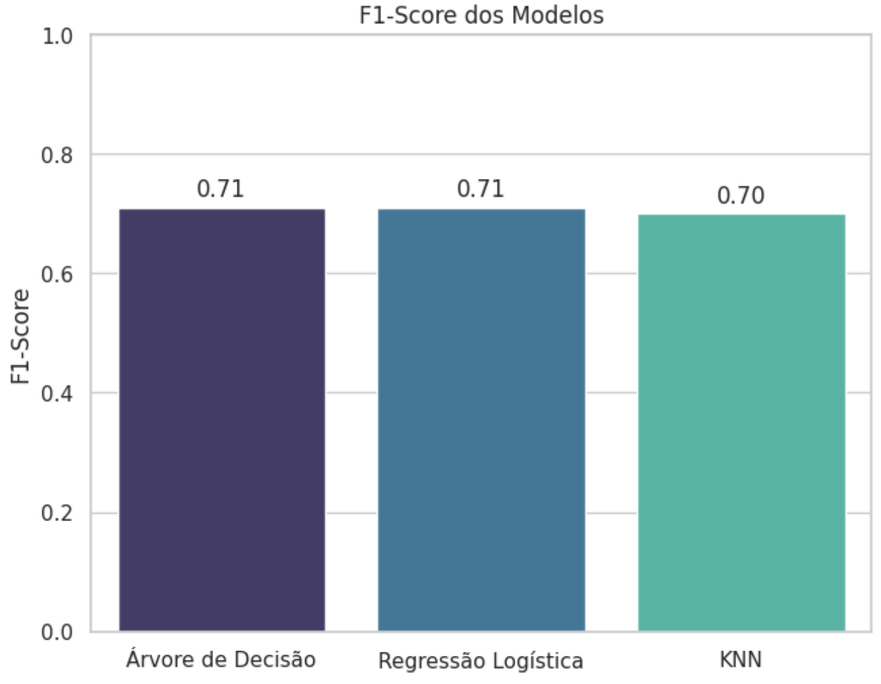
Acurácia: 95.71%




Acurácia dos Modelos
Acurácia
1.0
0.8
0.6
0.4
0.2
0.0
0.95
0.96
0.96
Árvore de Decisão
Regressão Logística
KNN


Precisão dos Modelos
Precisão
1.0
0.8
0.6
0.4
0.2
0.0
0.70
0.87
0.90
Árvore de Decisão
Regressão Logística
KNN


Recall dos Modelos
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.71
0.61
0.57
Árvore de Decisão
Regressão Logística
KNN

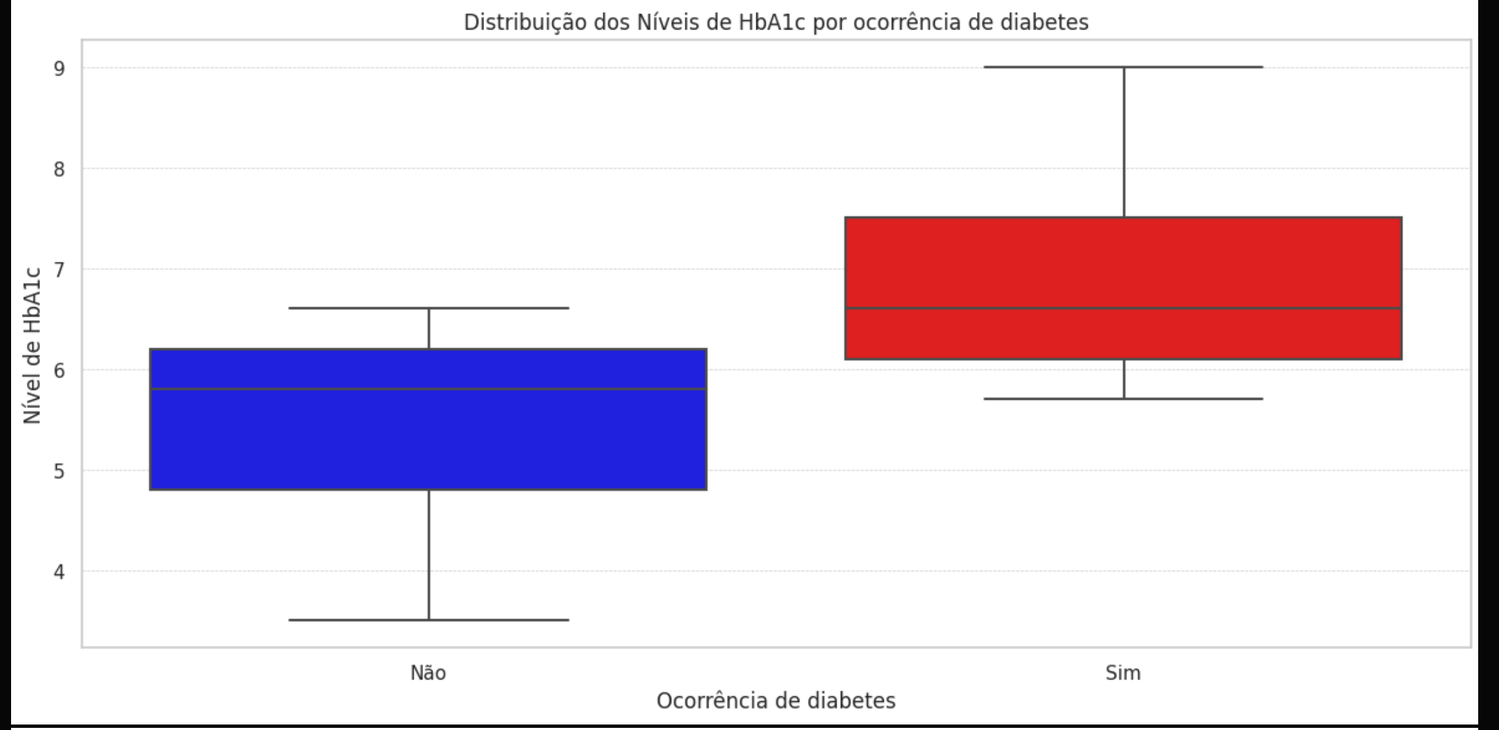

F1-Score dos Modelos
F1-Score
1.0
0.8
0.6
0.4
0.2
0.0
0.71
0.71
0.70
Árvore de Decisão
Regressão Logística
KNN

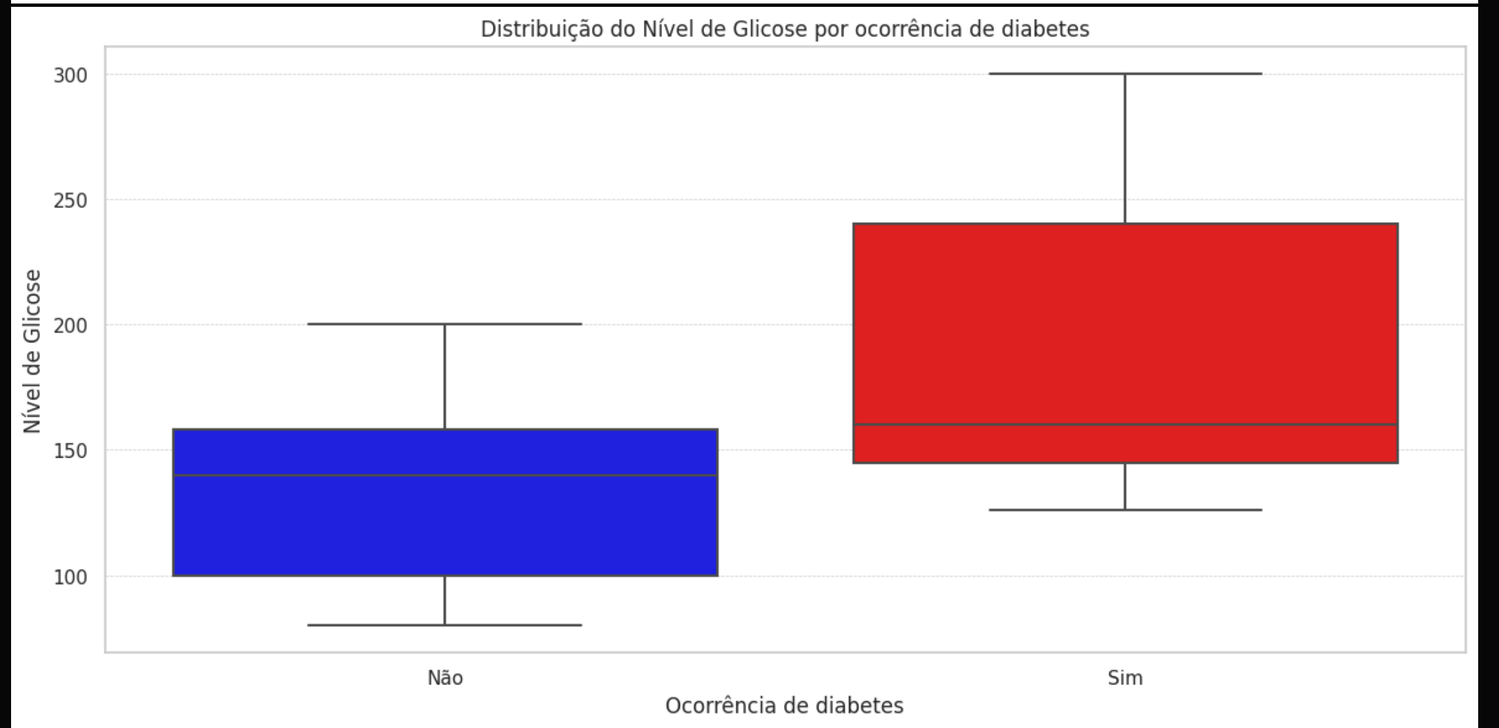
Teste dos Modelos

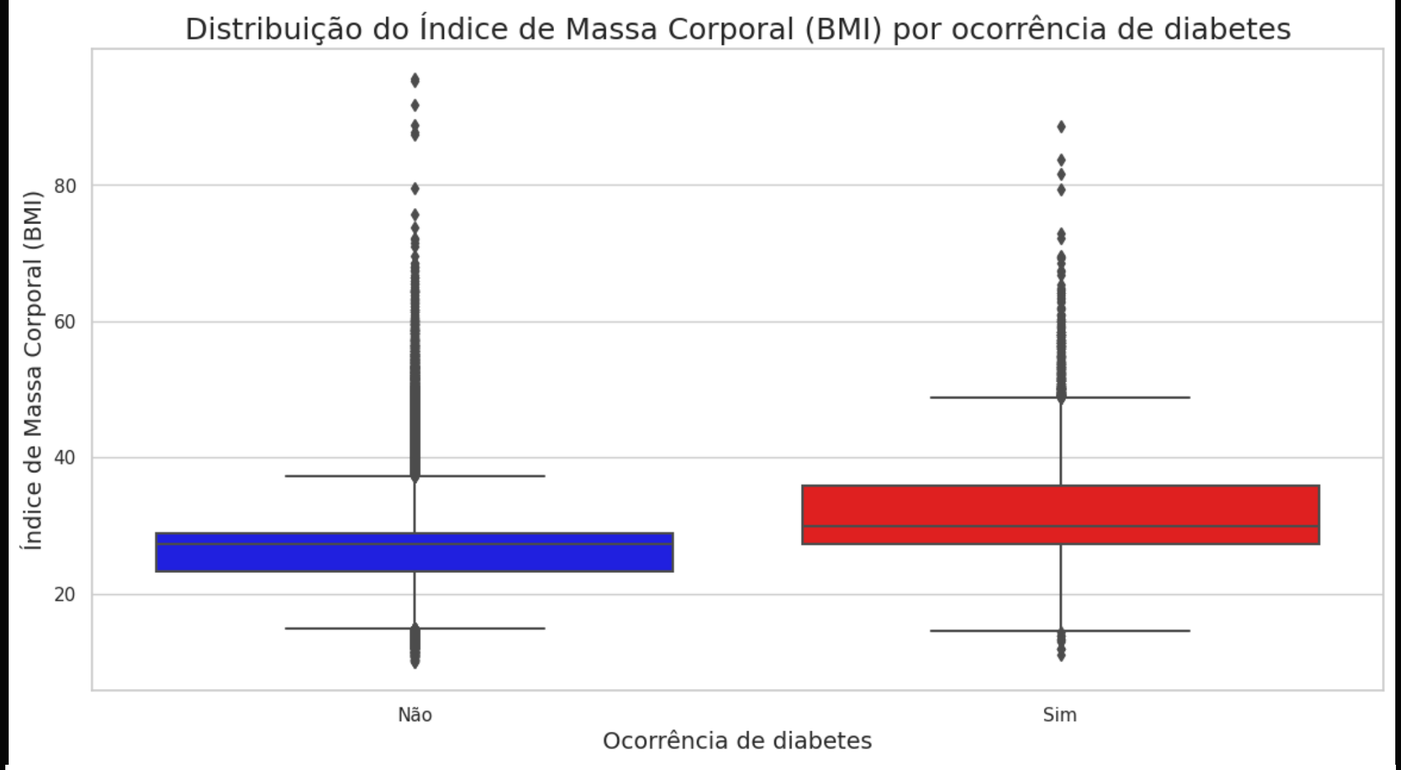
Para testar os modelos, criamos uma base de dados igual a original com dados fictícios de 15 pacientes para servirem de teste.

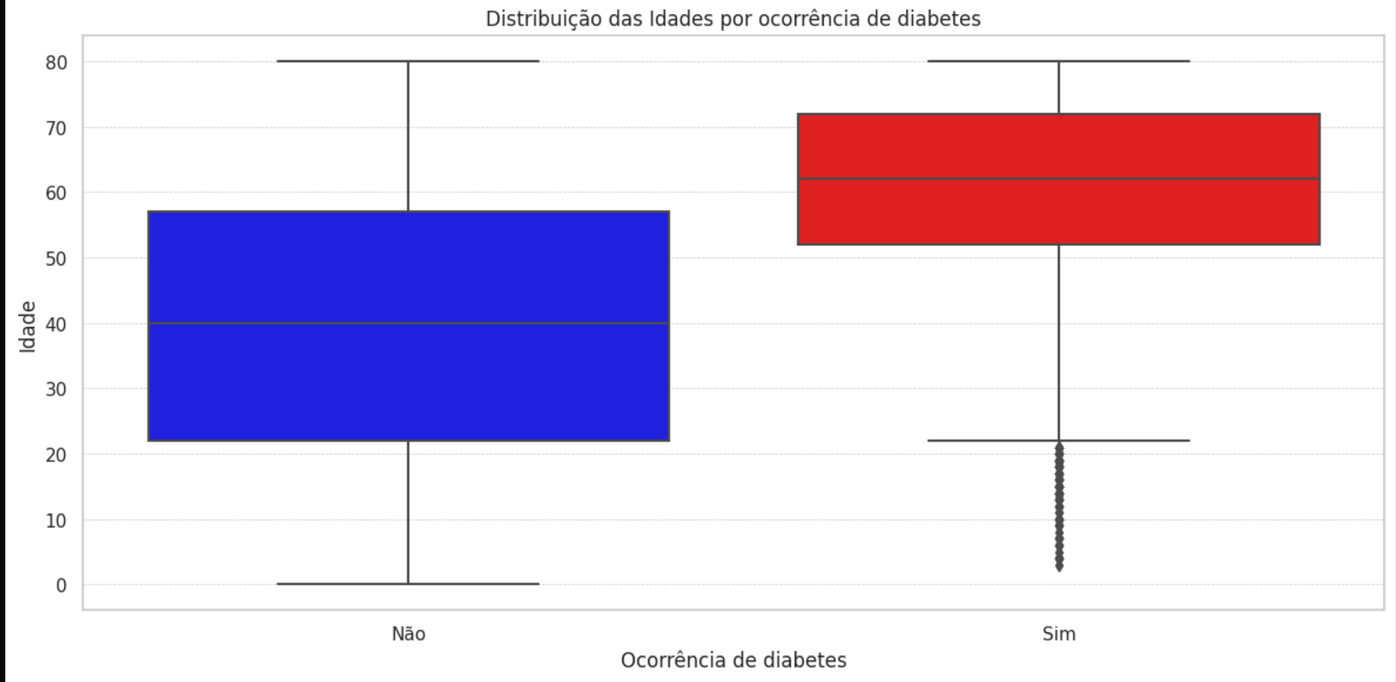
age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	gender	smoking_history
50	1	1	39.13	6.82	175.33	Male	never
55	1	0	31.75	6.50	168.88	Male	former
57	0	0	33.08	6.89	167.24	Female	current
62	1	1	34.93	6.61	172.52	Male	former
58	0	1	32.70	6.68	170.44	Male	never
26	0	0	19.84	4.71	72.93	Female	former
36	0	0	21.14	4.94	79.32	Female	never
36	0	0	22.55	4.28	80.06	Male	never
32	0	0	20.46	4.75	94.26	Male	never
35	0	0	21.08	4.43	86.87	Female	current
52	0	0	20.95	4.68	94.96	Male	current
30	1	1	26.69	4.08	168.22	Male	never
57	1	0	34.01	4.79	175.92	Male	never
29	0	0	28.66	5.12	115.17	Male	never
47	1	1	25.95	5.77	119.13	Male	former

```
Previsões do modelo Árvore de Decisão:
[1. 1. 1. 1. 1. 0. 0. 0. 0. 0. 0. 1. 1. 0. 1.]
Previsões do modelo Regressão Logística:
[1. 1. 1. 1. 1. 0. 0. 0. 0. 0. 0. 0. 1. 0. 1.]
Previsões do modelo KNN:
[1. 1. 1. 1. 1. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0.]
```


Distribuição dos Níveis de HbA1c por ocorrência de diabetes
Nível de HbA1c
9
8
7
6
5
4
Não
Sim
Ocorrência de diabetes


Distribuição do Nível de Glicose por ocorrência de diabetes
Nível de Glicose
300
250
200
150
100
Não
Sim
Ocorrência de diabetes


Distribuição do Índice de Massa Corporal (BMI) por ocorrência de diabetes
Índice de Massa Corporal (BMI)
80
60
40
20
Não
Sim
Ocorrência de diabetes


Distribuição das Idades por ocorrência de diabetes
Idade
80
70
60
50
40
30
20
10
0
Não
Sim
Ocorrência de diabetes

Conclusão

A vasta maioria dos pacientes (91,5%) não apresenta diabetes. A média de idade é de aproximadamente 42 anos, e o Índice de Massa Corporal (IMC) médio de 27,32 indica uma tendência ao sobrepeso na população estudada. O nível médio de glicose no sangue, situando-se em 138,06 mg/dL, se aproxima do limite superior do considerado normal, sugerindo um monitoramento mais atento em relação ao risco de diabetes. Além disso, a análise destacou a importância de considerar múltiplos fatores de risco ao avaliar a predisposição ao diabetes. Investigações futuras devem se aprofundar na relação entre características como hipertensão, doenças cardíacas e histórico de tabagismo com a incidência da doença.

Conclusão

Desempenho dos Modelos:

- Regressão Logística e KNN: Acurácias altas (~95.7%).
- Árvore de Decisão: Acurácia ligeiramente menor (94.9%), mas melhor equilíbrio e interpretabilidade.

Recall Importante:

- Regressão Logística e KNN: Precisos, mas com mais falsos negativos.
- Árvore de Decisão: Melhor em capturar casos reais de diabetes.

Portanto, com base nas análises dos modelos, para a previsão de diabetes, a árvore de decisão poderia ser a melhor opção, apesar da acurácia um pouco menor. A capacidade deste modelo de fornecer resultados mais equilibrados, sua interpretabilidade e seu maior recall para casos de diabetes (essencial para minimizar falsos negativos) são vantagens significativas no contexto de previsões médicas.