Projeto de Estatística

Grupo: Lucas Inojosa
Nathalia Barbosa
Rafael Barros
Sarah Melo

Centro de Informática UFPE
Universidade Federal de Pernambuco

Entendendo a base de dados

CAMPOS DA BASE

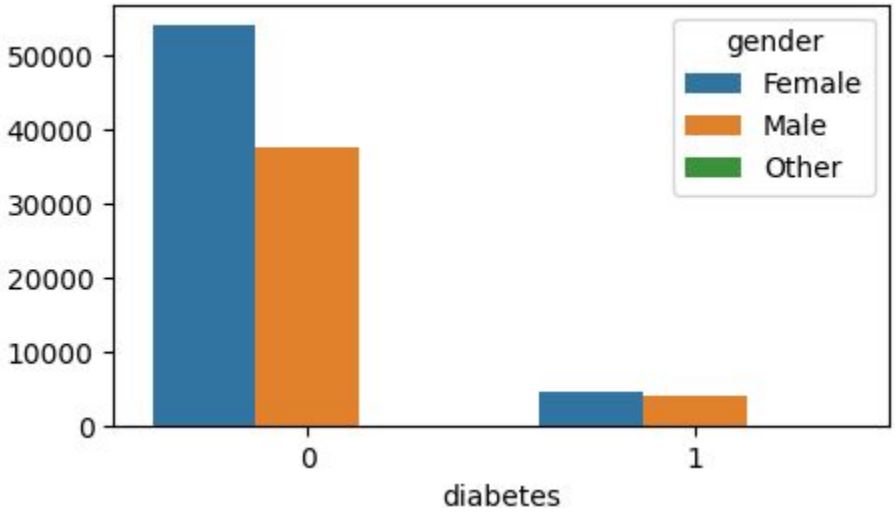
- Gênero
- Idade
- Hipertensão
- Problemas cardíacos
- Histórico de tabagismo
- IMC
- Nível de HbA1c
- Nível de glicose no sangue
- Diabetes

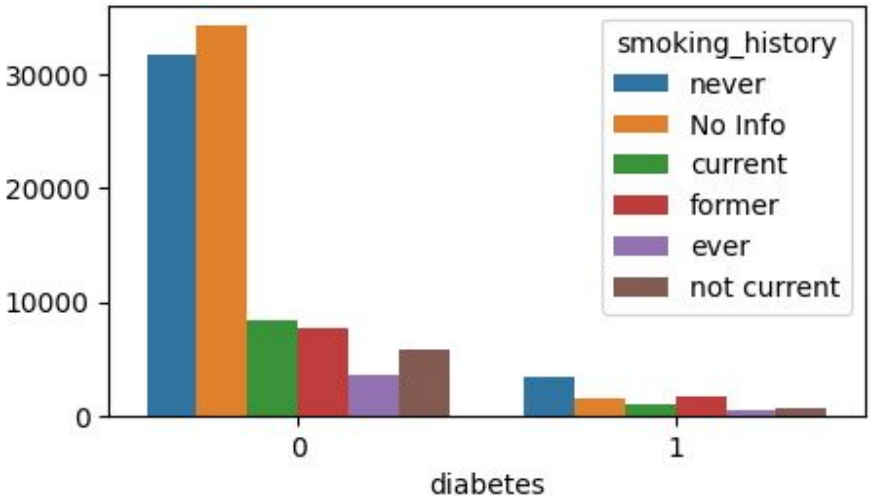
Entendendo a base de dados

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0
...	...	...	...	...	...	...	...	...	...
99995	Female	80.0	0	0	No Info	27.32	6.2	90	0
99996	Female	2.0	0	0	No Info	17.37	6.5	100	0
99997	Male	66.0	0	0	former	27.83	5.7	155	0
99998	Female	24.0	0	0	never	35.42	4.0	100	0
99999	Female	57.0	0	0	current	22.43	6.6	90	0

Entendendo a base de dados

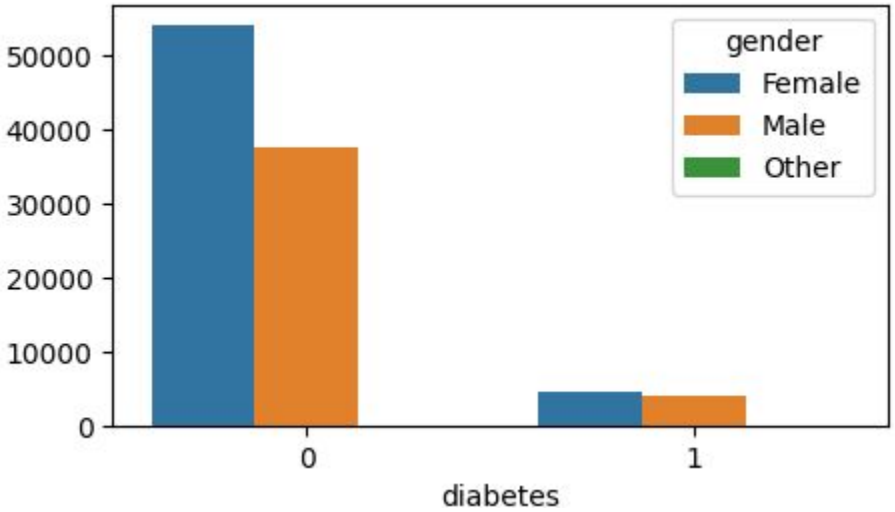
- Dados não-numéricos:

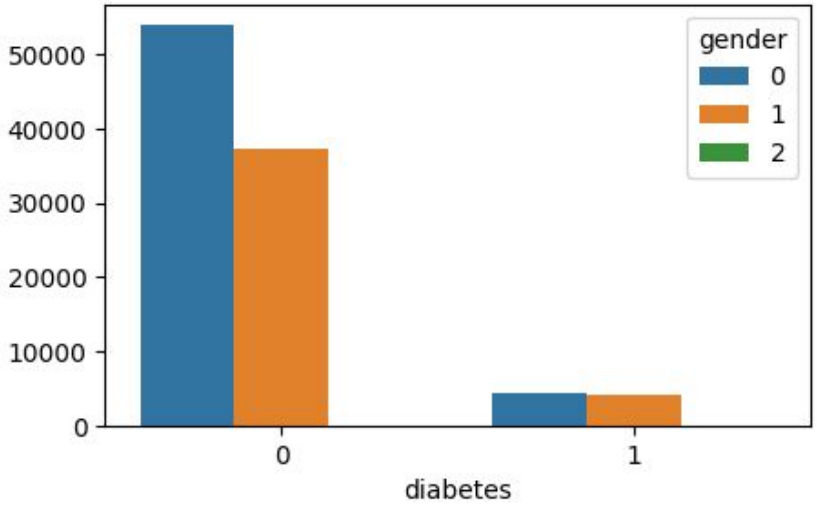




Conversão dos dados não-numéricos

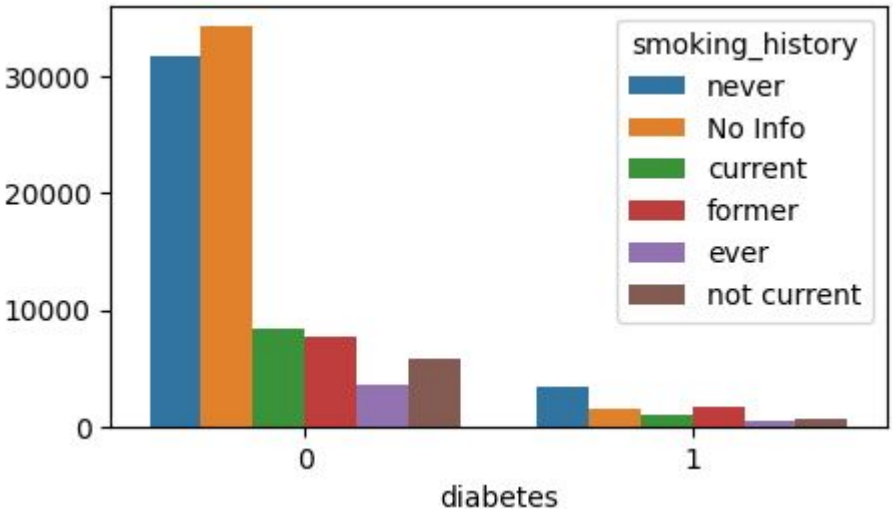
- Gênero

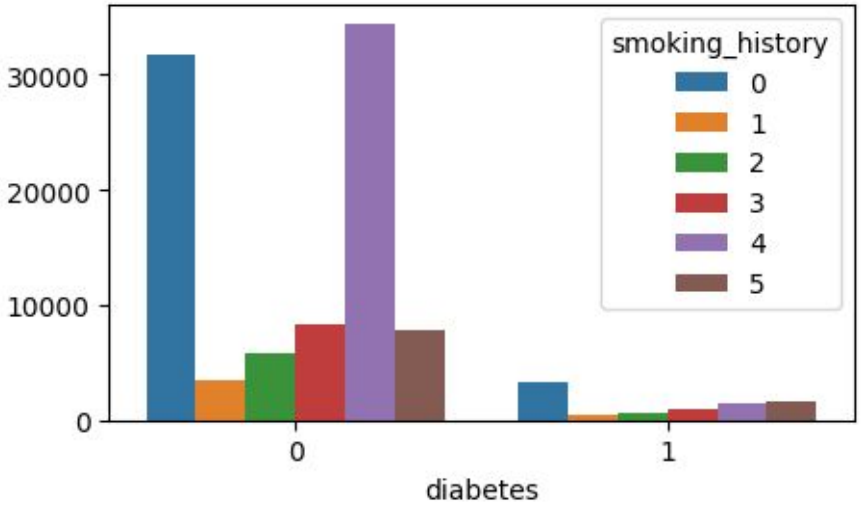




Conversão dos dados não-numéricos

- Histórico de tabagismo

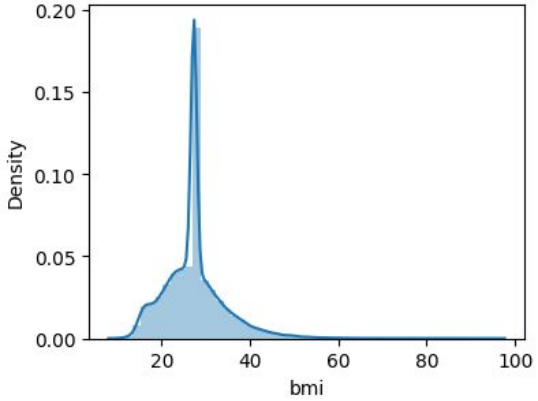


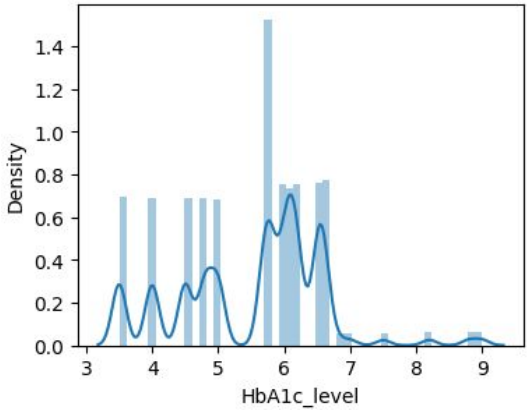


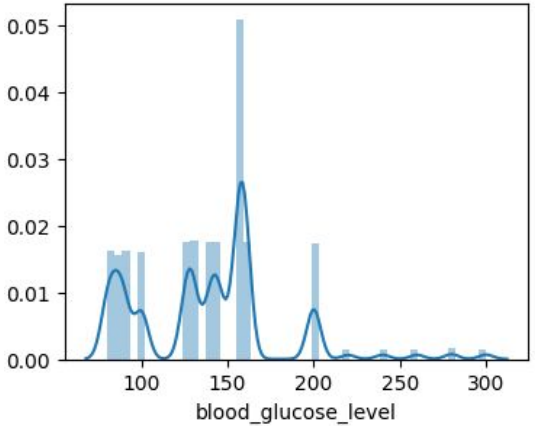
Conversão dos dados não-numéricos

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	1	80.0	0	1	2	25.19	6.6	140	0
1	1	54.0	0	0	3	27.32	6.6	80	0
2	2	28.0	0	0	2	27.32	5.7	158	0
3	1	36.0	0	0	5	23.45	5.0	155	0
4	2	76.0	1	1	5	20.14	4.8	155	0
...	...	...	...	...	...	...	...	...	...
99995	1	80.0	0	0	3	27.32	6.2	90	0
99996	1	2.0	0	0	3	17.37	6.5	100	0
99997	2	66.0	0	0	4	27.83	5.7	155	0
99998	1	24.0	0	0	2	35.42	4.0	100	0
99999	1	57.0	0	0	5	22.43	6.6	90	0

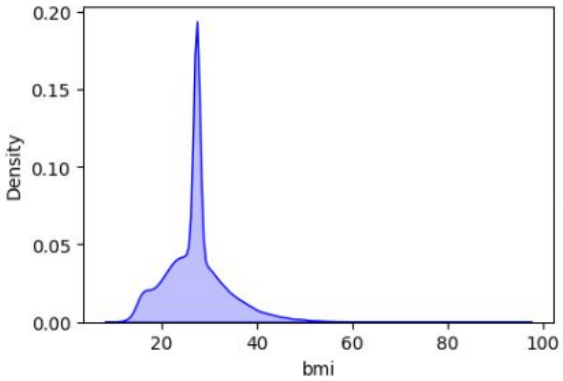
Gráficos

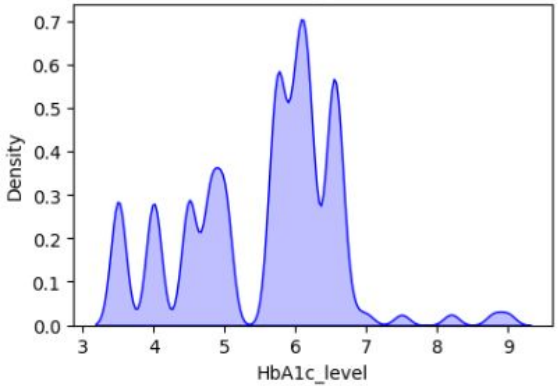

Density
0.20
0.15
0.10
0.05
0.00
20
40
60
80
100
bmi

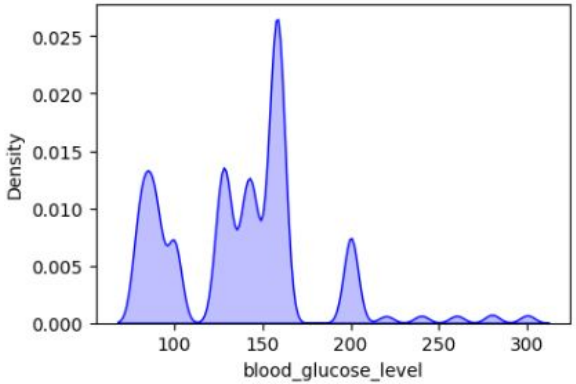

Density
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
3
4
5
6
7
8
9
HbA1c_level


0.05
0.04
0.03
0.02
0.01
0.00
100
150
200
250
300
blood_glucose_level

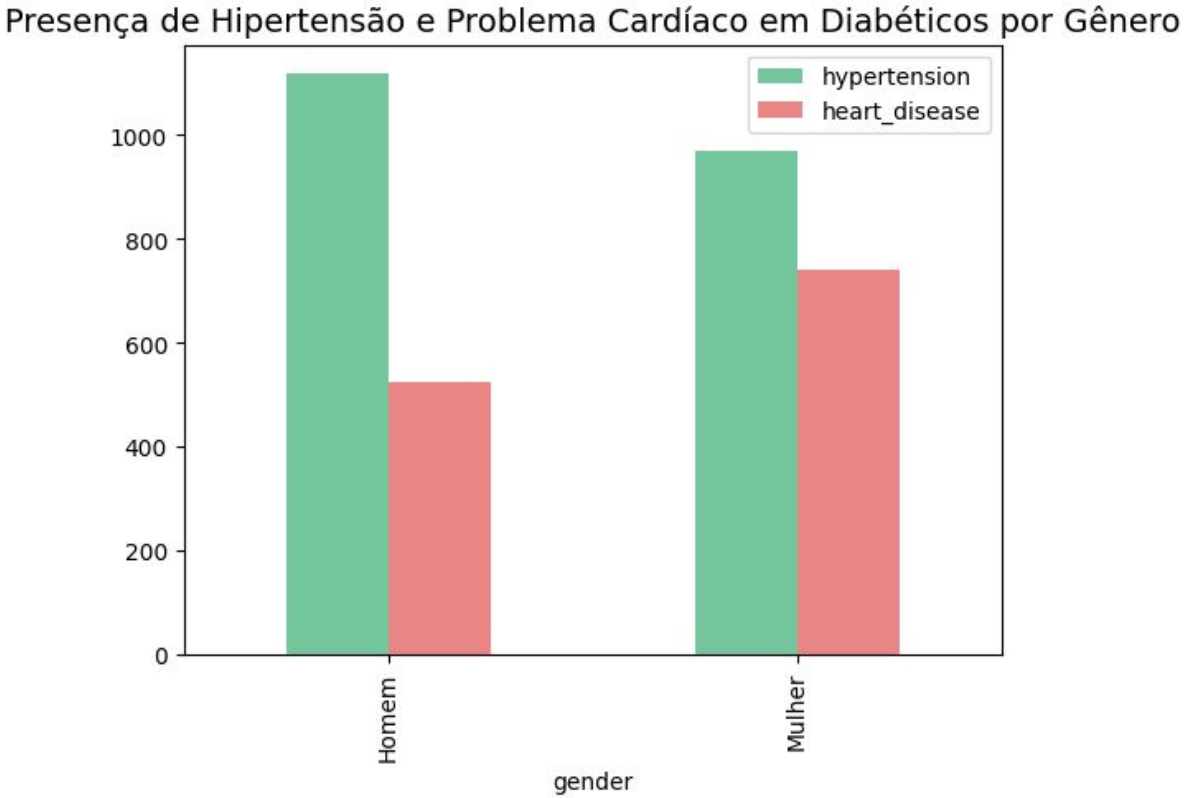
Gráficos


Density
0.20
0.15
0.10
0.05
0.00
20
40
60
80
100
bmi

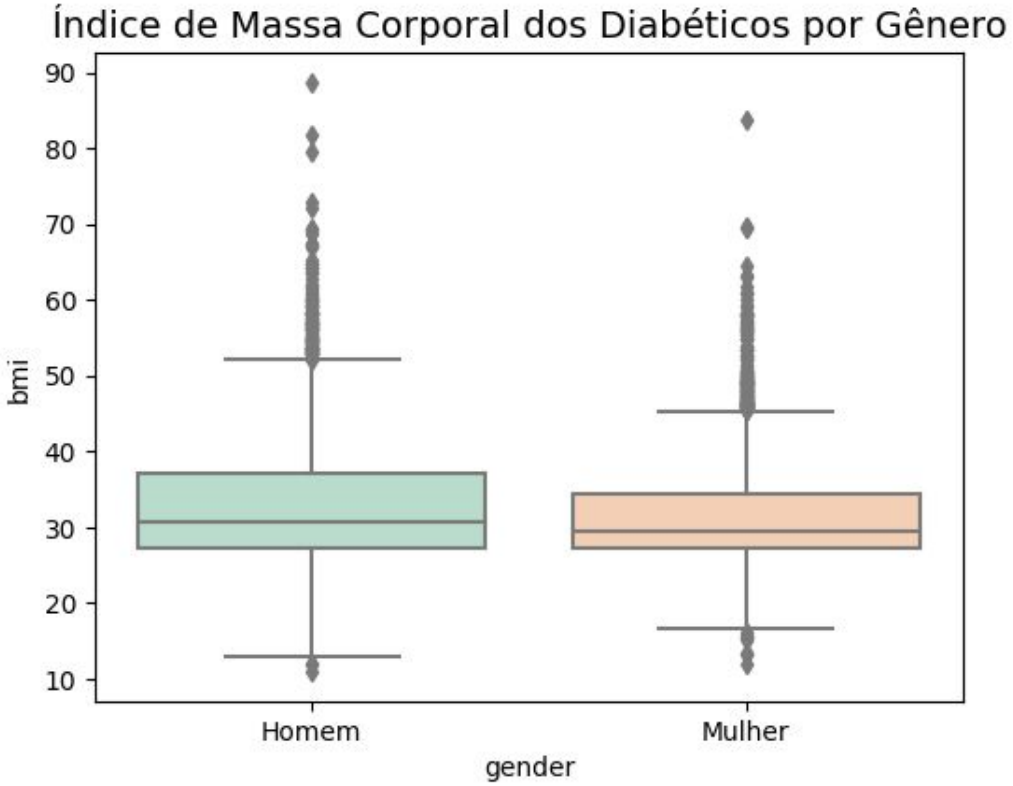

Density
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
3
4
5
6
7
8
9
HbA1c_level


Density
0.025
0.020
0.015
0.010
0.005
0.000
100
150
200
250
300
blood_glucose_level

Gráficos


Presença de Hipertensão e Problema Cardíaco em Diabéticos por Gênero
hypertension
heart_disease
1000
800
600
400
200
0
Homem
Mulher
gender

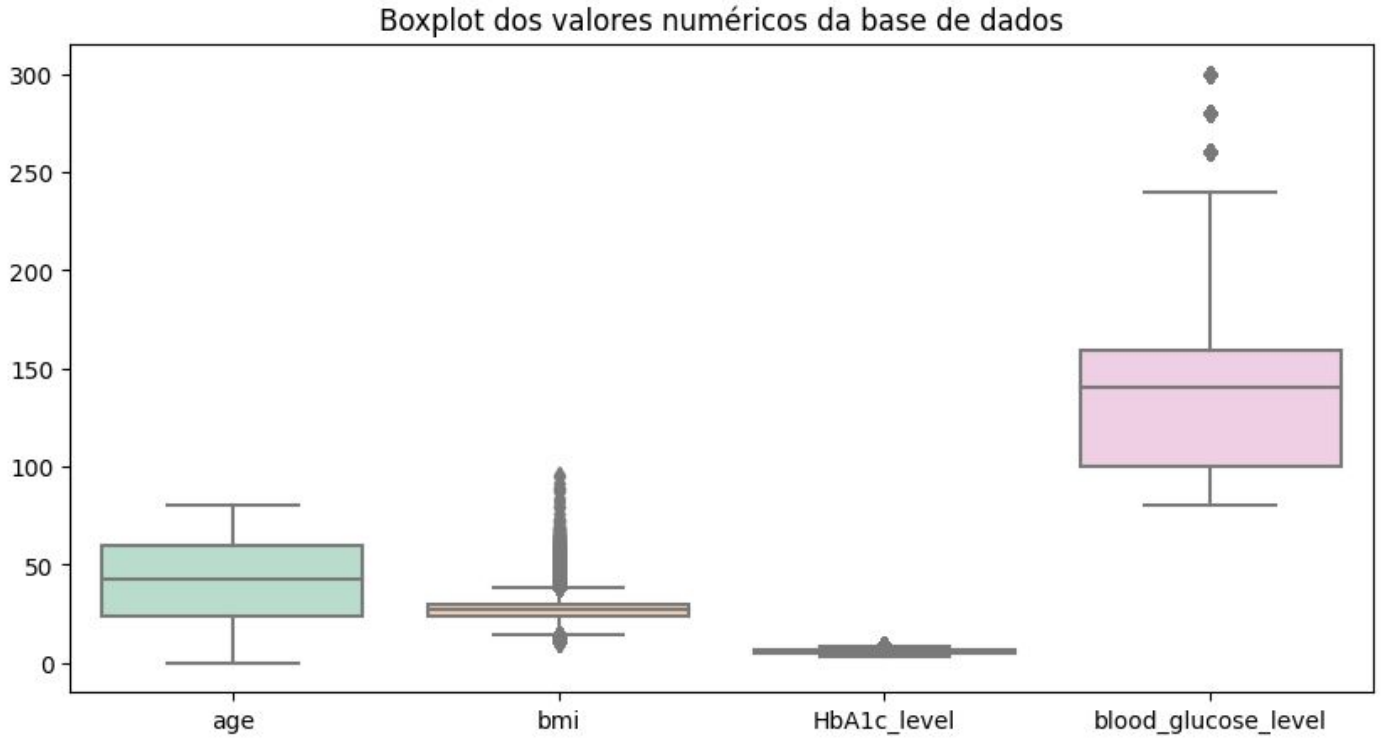
Gráficos


Índice de Massa Corporal dos Diabéticos por Gênero
bmi
90
80
70
60
50
40
30
20
10
Homem
Mulher
gender

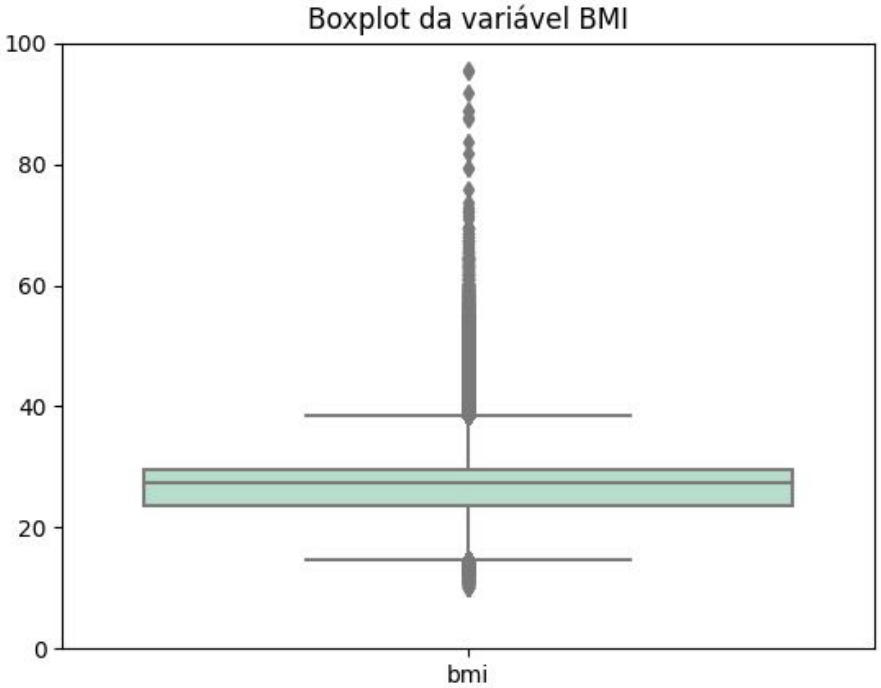
Identificando outliers

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
count	100000.000000	100000.000000	100000.00000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000
mean	0.828780	41.885856	0.07485	0.039420	2.306260	27.320767	5.527507	138.058060	0.085000
std	0.985146	22.516840	0.26315	0.194593	1.314215	6.636783	1.070672	40.708136	0.278883
min	0.000000	0.080000	0.00000	0.000000	0.000000	10.010000	3.500000	80.000000	0.000000
25%	0.000000	24.000000	0.00000	0.000000	1.000000	23.630000	4.800000	100.000000	0.000000
50%	0.000000	43.000000	0.00000	0.000000	3.000000	27.320000	5.800000	140.000000	0.000000
75%	2.000000	60.000000	0.00000	0.000000	3.000000	29.580000	6.200000	159.000000	0.000000
max	2.000000	80.000000	1.00000	1.000000	5.000000	95.690000	9.000000	300.000000	1.000000

Identificando outliers



Identificando outliers



	bmi
count	100000.000000
mean	27.320767
std	6.636783
min	10.010000
25%	23.630000
50%	27.320000
75%	29.580000
max	95.690000

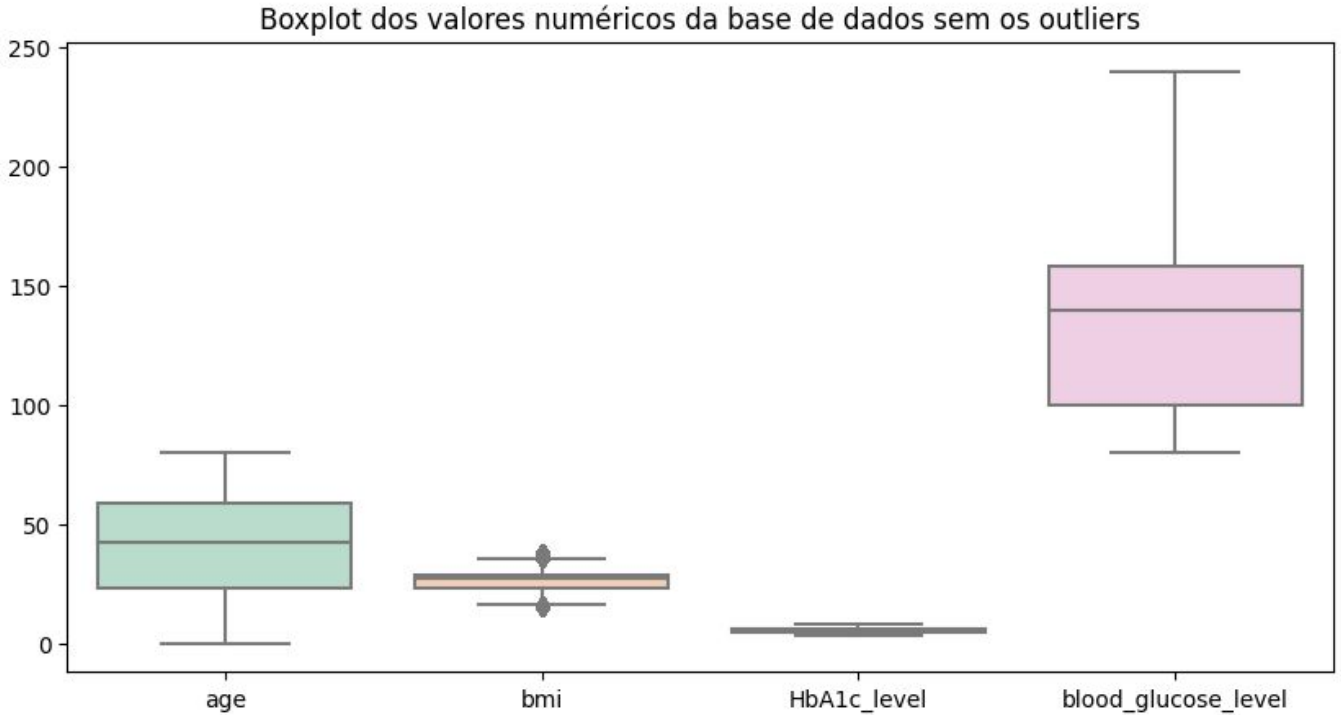
Removendo outliers

```
[ ]  def remove_outliers_1_5_iqr(df, columns):
         for column in columns:
             q1 = df[column].quantile(0.25)
             q3 = df[column].quantile(0.75)
             iqr = q3 - q1
             lower_limit = q1 - 1.5 * iqr
             upper_limit = q3 + 1.5 * iqr
             df = df[(df[column] >= lower_limit) & (df[column] <= upper_limit)]
         return df
```

Removendo outliers

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
count	90387.000000	90387.000000	90387.000000	90387.000000	90387.000000	90387.000000	90387.000000	90387.000000	90387.000000
mean	0.837344	41.407823	0.065651	0.035625	2.317037	26.317686	5.456326	134.572571	0.049731
std	0.986593	22.558809	0.247672	0.185353	1.309170	4.878880	0.994868	35.197882	0.217389
min	0.000000	0.080000	0.000000	0.000000	0.000000	14.710000	3.500000	80.000000	0.000000
25%	0.000000	23.000000	0.000000	0.000000	1.000000	23.370000	4.800000	100.000000	0.000000
50%	0.000000	42.000000	0.000000	0.000000	3.000000	27.320000	5.800000	140.000000	0.000000
75%	2.000000	59.000000	0.000000	0.000000	3.000000	28.280000	6.200000	158.000000	0.000000
max	2.000000	80.000000	1.000000	1.000000	5.000000	38.500000	8.200000	240.000000	1.000000

Removendo outliers



Gaussian Bayes

- Conjunto de dados

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	0	80.0	0	1	1	25.19	6.6	140	0
1	0	54.0	0	0	3	27.32	6.6	80	0
2	2	28.0	0	0	1	27.32	5.7	158	0
3	0	36.0	0	0	4	23.45	5.0	155	0
4	2	76.0	1	1	4	20.14	4.8	155	0

Separando os Parâmetros

```
[ ]  X = bd.iloc[:,0:8].values
     X
```

```
array([[  0.  ,  80.  ,   0.  , ...,  25.19,   6.6 , 140.  ],
       [  0.  ,  54.  ,   0.  , ...,  27.32,   6.6 ,  80.  ],
       [  2.  ,  28.  ,   0.  , ...,  27.32,   5.7 , 158.  ],
       ...,
       [  2.  ,  66.  ,   0.  , ...,  27.83,   5.7 , 155.  ],
       [  0.  ,  24.  ,   0.  , ...,  35.42,   4.  , 100.  ],
       [  0.  ,  57.  ,   0.  , ...,  22.43,   6.6 ,  90.  ]])
```

Separando os Parâmetros

```
[ ]  Y = bd.iloc[:, 8].values
     Y

     array([0, 0, 0, ..., 0, 0, 0])
```

Previsões do Modelo

- mulher (0),
- 55 anos,
- sem hipertensão (0),
- sem doença no coração (0),
- sem histórico de tabagismo (0),
- IMC = 25,
- nível de HbA1c = 7,
- nível de glicose no sangue = 150.

```
# Dados dos novos indivíduos
novos_individuos = [
    [0, 55, 0, 0, 0, 25, 7, 150],
    [2, 55, 0, 0, 0, 25, 7, 150], # mulher -> homem
    [0, 25, 0, 0, 0, 25, 7, 150], # 55 anos -> 25
    [0, 55, 1, 0, 0, 25, 7, 150], # sem hipertensão -> com
    [0, 55, 0, 1, 0, 25, 7, 150], # sem doença no coração -> com
    [0, 55, 0, 0, 2, 25, 7, 150], # sem histórico de tabagismo -> com
    [0, 55, 0, 0, 0, 38, 7, 150], # IMC 25 -> 38
    [0, 55, 0, 0, 0, 25, 5, 150], # nível de HbA1c 7 -> 5
    [0, 55, 0, 0, 0, 25, 7, 80]   # nível de glicose no sangue 150 -> 80
```

Previsões do Modelo

- Considerando outliers:

```
Novo indivíduo 1: [0, 55, 0, 0, 0, 25, 7, 150]
Probabilidade de ser diabético: 2.82%
O novo indivíduo provavelmente não é diabético.
--------------------------------------------------
Novo indivíduo 2: [2, 55, 0, 0, 0, 25, 7, 150]
Probabilidade de ser diabético: 1.46%
O novo indivíduo provavelmente não é diabético.
--------------------------------------------------
Novo indivíduo 3: [0, 25, 0, 0, 0, 25, 7, 150]
Probabilidade de ser diabético: 0.15%
O novo indivíduo provavelmente não é diabético.
--------------------------------------------------
Novo indivíduo 4: [0, 55, 1, 0, 0, 25, 7, 150]
Probabilidade de ser diabético: 95.40%
O novo indivíduo provavelmente é diabético.
--------------------------------------------------
Novo indivíduo 5: [0, 55, 0, 1, 0, 25, 7, 150]
Probabilidade de ser diabético: 100.00%
O novo indivíduo provavelmente é diabético.
```

```
Novo indivíduo 6: [0, 55, 0, 0, 2, 25, 7, 150]
Probabilidade de ser diabético: 5.43%
O novo indivíduo provavelmente não é diabético.
--------------------------------------------------
Novo indivíduo 7: [0, 55, 0, 0, 0, 38, 7, 150]
Probabilidade de ser diabético: 12.43%
O novo indivíduo provavelmente não é diabético.
--------------------------------------------------
Novo indivíduo 8: [0, 55, 0, 0, 0, 25, 5, 150]
Probabilidade de ser diabético: 0.16%
O novo indivíduo provavelmente não é diabético.
--------------------------------------------------
Novo indivíduo 9: [0, 55, 0, 0, 0, 25, 7, 80]
Probabilidade de ser diabético: 1.66%
O novo indivíduo provavelmente não é diabético.
```

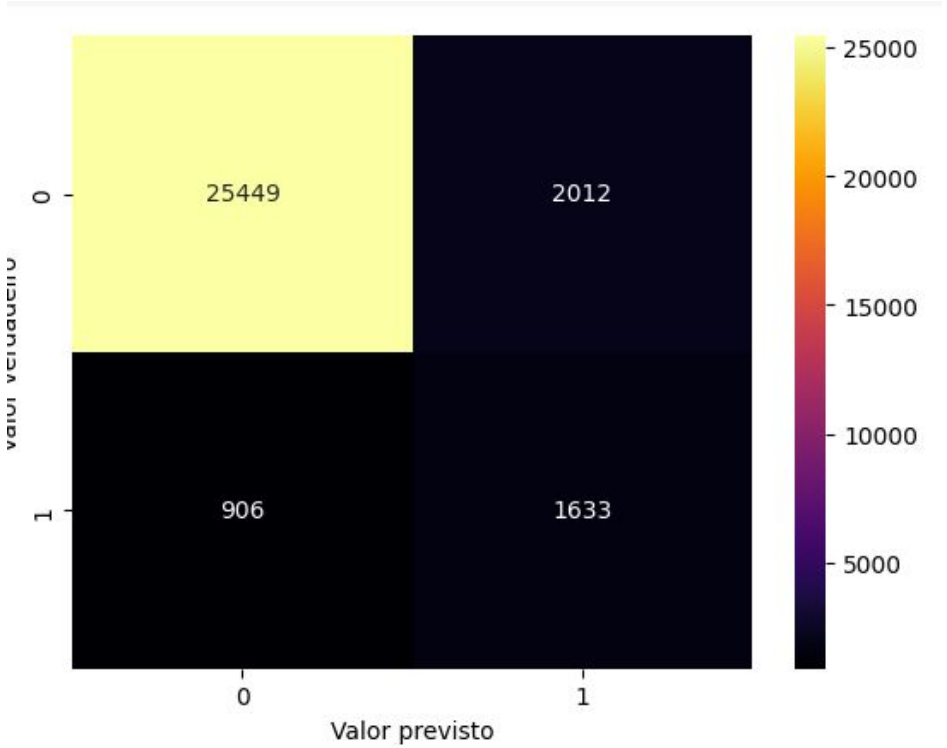
Treinando o Modelo e Obtendo Estatísticas

```
[ ]  naive_diabetes = GaussianNB()
     naive_diabetes.fit(X,Y)
```

```
[ ]  acuracia = naive_diabetes.score(X,Y)
     print("Acurácia do modelo:", acuracia)
```

```
Acurácia do modelo: 0.90352
```

Treinando o Modelo e Obtendo Estatísticas



Acurácia 90,8%
de certeza

Sem Diabetes
93% de certeza

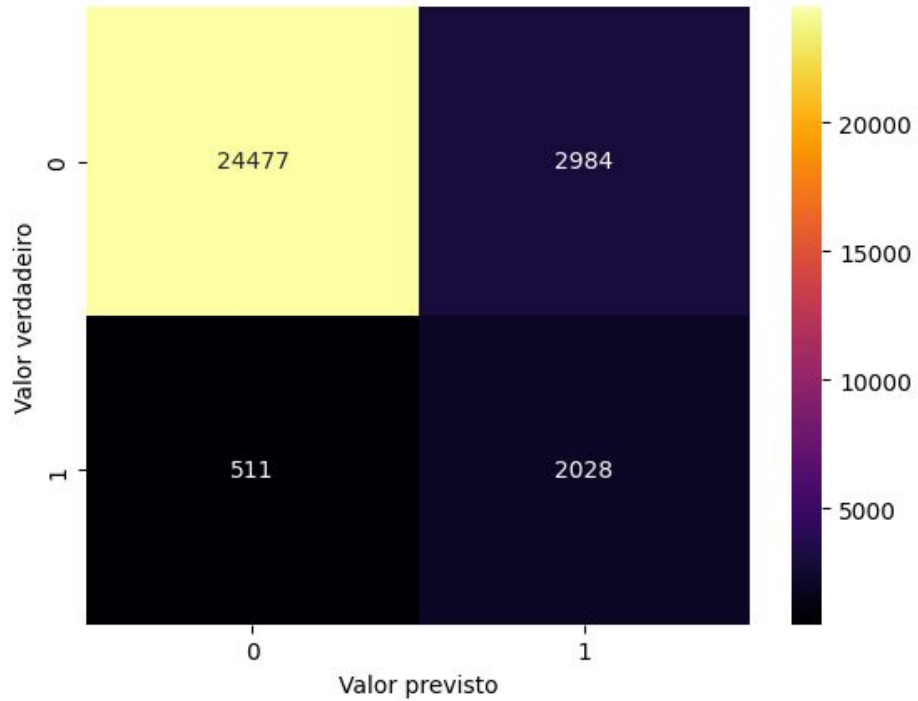
Com Diabetes
64% de certeza

Problema

Número de Falsos Negativos muito Grande

Apenas 8% do nosso banco de dados possui Diabetes

Aplicando o Método Smote para os dados com outliers

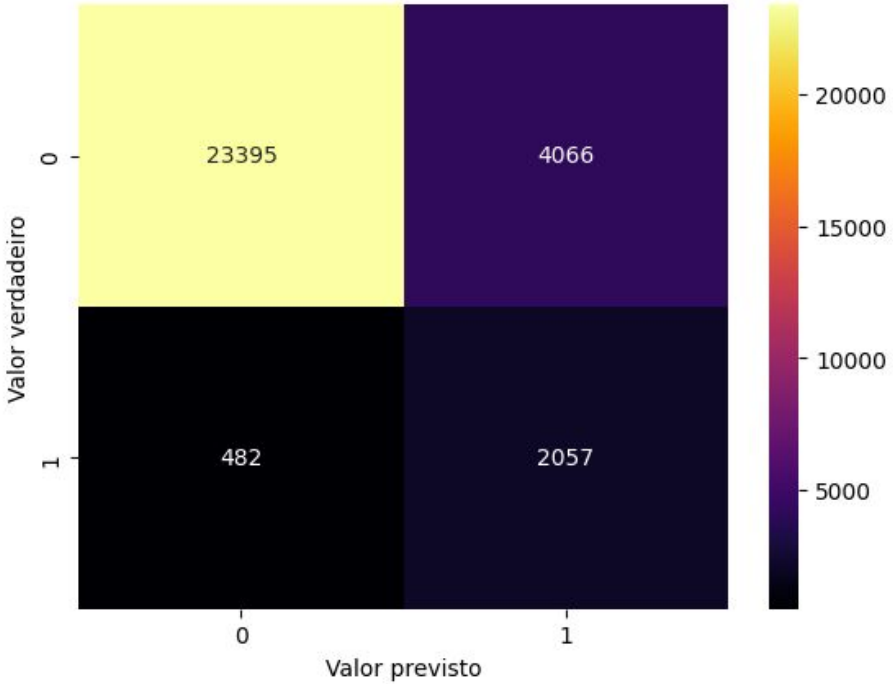


Acurácia 85,2%
de certeza

Sem Diabetes
89% de certeza

Com Diabetes
79% de certeza

Aplicando o Método Smote para os dados sem outliers



Acurácia 88,8% de certeza

Sem Diabetes 85% de certeza

Com Diabetes 81% de certeza

Melhorias

Muitas previsões de pessoas diabéticas que não estão diabetes (falsos positivos)

Geração de dados sintético não aumenta a qualidade dos dados

Obter Mais dados reais de pessoas com diabetes para melhorar o modelo