RELATÓRIO AP1 PROJETO MACHINE LEARNING-BRENO DE SOUZA

Dataset escolhido: Diabetes Dataset EDA

LINK: https://www.kaggle.com/code/chanchal24/diabetes-dataset-eda-prediction-with-7-models/notebook

Sobre o dataset:

Esse dataset é sobre **diabetes** e contém **informações clínicas** de pacientes com o objetivo de **prever se a pessoa tem ou não diabetes**, com base em características fisiológicas e exames médicos.

O conjunto de dados é conhecido como **Pima Indians Diabetes Dataset**, um dos mais usados em tarefas de classificação em Machine Learning na área de saúde.

O conjunto foi originalmente coletado pelo **National Institute of Diabetes and Digestive and Kidney Diseases** e se refere a **mulheres Pima de pelo menos 21 anos de idade**, um grupo indígena dos EUA com alta taxa de diabetes.

Variáveis Presentes:

Pregnancies - Número de gestações

Glucose- Concentração de glicose no sangue (após jejum).

Blood Pressure – Pressão Arterial Diastólica(mm Hg)

SkinThickness - Espessura da dobra cutânea do tríceps (mm)

Insulin - Nível de insulina no soro (mu U/ml)

BMI - Índice de massa corporal (peso em kg / (altura em m)^2)

DiabetesPedigreeFunction - Histórico familiar de diabetes (relação genética ponderada)

Age- Idade do paciente (em anos)

Outcome - 0 = não diabético, **1 = diabético (variável alvo)**

Objetivo do Trabalho:

Criar um **modelo preditivo de classificação** que, com base nas variáveis clínicas, consiga prever o valor da variável **Outcome**, ou seja, **se o paciente tem ou não diabetes**.

Estatísticas descritivas básicas:

Código: summary(diabetes)

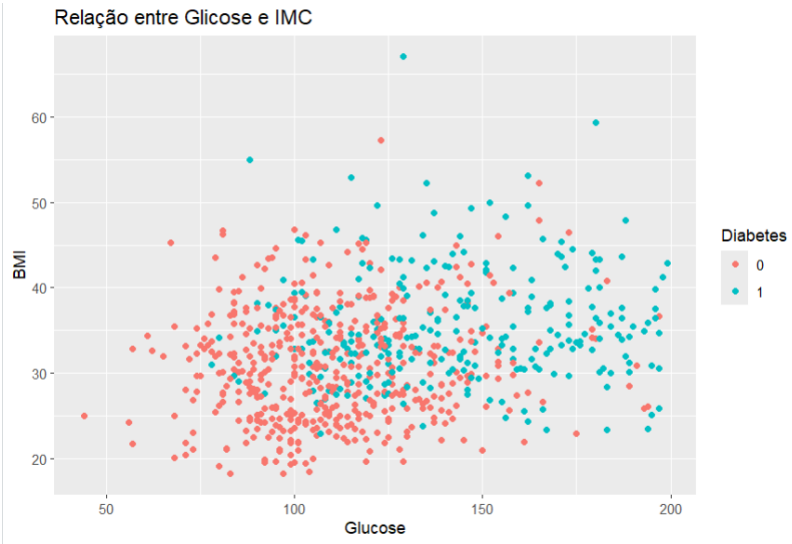
```
> # Estatísticas descritivas básicas
> summary(diabetes)
  Pregnancies        Glucose      BloodPressure    SkinThickness      Insulin
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00   Min.   :  0.0
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00   1st Qu.:  0.0
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00   Median : 30.5
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54   Mean   : 79.8
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00   3rd Qu.:127.2
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.0
      BMI        DiabetesPedigreeFunction      Age          Outcome
 Min.   : 0.00   Min.   :0.0780           Min.   :21.00   Min.   :0.000
 1st Qu.:27.30   1st Qu.:0.2437           1st Qu.:24.00   1st Qu.:0.000
 Median :32.00   Median :0.3725           Median :29.00   Median :0.000
 Mean   :31.99   Mean   :0.4719           Mean   :33.24   Mean   :0.349
 3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00   3rd Qu.:1.000
 Max.   :67.10   Max.   :2.4200           Max.   :81.00   Max.   :1.000
```

Como podemos observar, A média da glicose é **120.9**, com valores variando de **0 a 199**. O índice de massa corporal (BMI) tem média **31.99**, indicando um possível sobrepeso médio entre os pacientes. A idade dos pacientes varia de **21 a 81 anos** e a média de Outcome é **0.349**, indicando que **aproximadamente 35% dos pacientes no dataset têm diabetes**. Após isso eu substitui valores 0 por NA nas colunas específicas e substituir os valores NA pela mediana da respectiva coluna.

Verificando se os valores após isso foram tratados, e sim, foram.

```
rm = TRUE), x))
> # Verificar se os valores foram tratados
> summary(diabetes)
  Pregnancies        Glucose       BloodPressure    SkinThickness      Insulin
 Min.   : 0.000   Min.   : 44.00   Min.   : 24.00   Min.   : 7.00   Min.   : 14.0
 1st Qu.: 1.000   1st Qu.: 99.75   1st Qu.: 64.00   1st Qu.:25.00   1st Qu.:121.5
 Median : 3.000   Median :117.00   Median : 72.00   Median :29.00   Median :125.0
 Mean   : 3.845   Mean   :121.66   Mean   : 72.39   Mean   :29.11   Mean   :140.7
 3rd Qu.: 6.000   3rd Qu.:140.25   3rd Qu.: 80.00   3rd Qu.:32.00   3rd Qu.:127.2
 Max.   :17.000   Max.   :199.00   Max.   :122.00   Max.   :99.00   Max.   :846.0
      BMI        DiabetesPedigreeFunction      Age            Outcome
 Min.   :18.20   Min.   :0.0780           Min.   :21.00   Min.   :0.000
 1st Qu.:27.50   1st Qu.:0.2437           1st Qu.:24.00   1st Qu.:0.000
 Median :32.30   Median :0.3725           Median :29.00   Median :0.000
 Mean   :32.46   Mean   :0.4719           Mean   :33.24   Mean   :0.349
 3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00   3rd Qu.:1.000
 Max.   :67.10   Max.   :2.4200           Max.   :81.00   Max.   :1.000
```

Em seguida, gerei um gráfico de dispersão sobre a relação entre glicose e IMC.

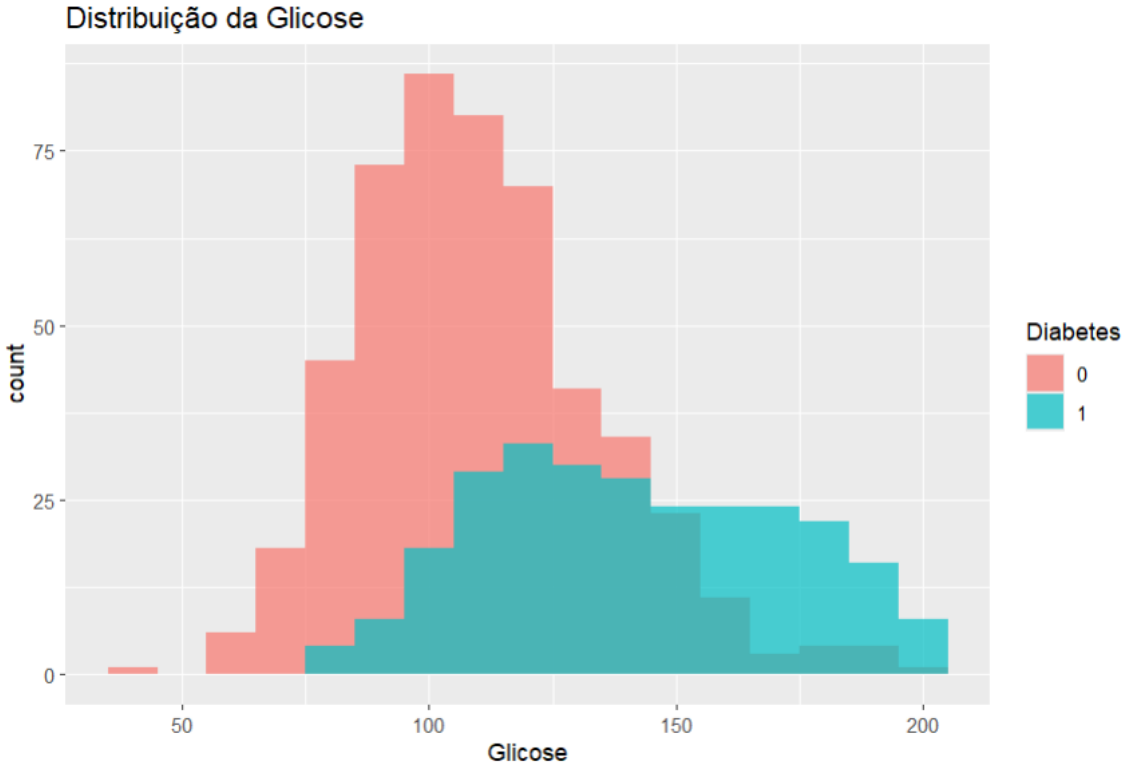


Os pontos **azuis (diabéticos)** aparecem mais concentrados em níveis de glicose **acima de 130**.Os pontos **vermelhos (não diabéticos)** são mais comuns em níveis de glicose **abaixo de 130**. Isso indica que

níveis elevados de glicose estão fortemente associados à presença de diabetes, o que é esperado. Entre pacientes com glicose baixa a moderada (**abaixo de 120**), há uma mistura de diabéticos e não diabéticos, especialmente para valores mais altos de IMC. Isso sugere que **o IMC por si só não é um fator decisivo para diabetes**, mas quando combinado com glicose alta, pode aumentar o risco.

Alguns pacientes com **glicose acima de 180 e IMC elevado (>40)** são diabéticos quase exclusivamente, porém, há alguns casos de **diabéticos com glicose baixa (<100)**, o que pode indicar fatores adicionais influenciando o diagnóstico (exemplo: histórico familiar,etc).

Posteriormente gerei um histograma sobre a distribuição da glicose para confirmar esta tese.



E Em seguida, realizei o teste de Shapiro-Wilk:

```
        Shapiro-Wilk normality test

data:  diabetes$Glucose
W = 0.96962, p-value = 1.523e-11
```

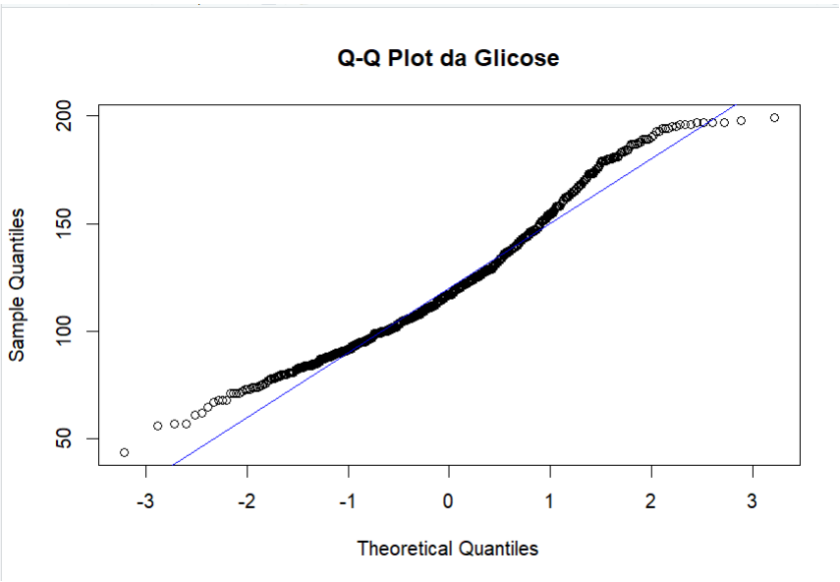
O valor de W deu **0.96962** o que indica aderência a normalidade, pois quanto mais próxima de 1, mais próxima a distribuição está de uma normal. Porém o p-Valor sendo 1.523e-11 (ou seja, 0.00000000001523), é extremamente pequeno, ou seja, **os dados de glicose não seguem uma distribuição normal**, pois o p-valor é muito baixo.

Em suma, o **teste de Shapiro-Wilk** mostrou que a distribuição de **Glicose não é normal (p-valor < 0.05)**. A regressão linear **assume normalidade dos resíduos**, não necessariamente das variáveis individuais.

A regressão linear pode ser aplicada se houver uma relação aproximadamente linear entre as variáveis. No entanto, o **teste de Shapiro-Wilk** rejeitou a normalidade da **Glicose (p-valor < 0.05)**, e o histograma indica assimetria. Isso pode afetar a normalidade dos resíduos, um pressuposto da regressão.

Se os resíduos não forem normais ou houver heterocedasticidade, creio que a **regressão logística seja o ideal .**

Após isso gerei o gráfico Q-Q plot:



O **Q-Q Plot** mostra que os quantis da glicose desviam da linha teórica nos extremos, sugerindo que a distribuição **não segue uma normal**. Isso reforça o resultado do **teste de Shapiro-Wilk**.

Em seguida realizei a **Correlação de Pearson:**

```
        Pearson's product-moment correlation

data:  diabetes$Glucose and diabetes$BMI
t = 6.5725, df = 766, p-value = 9.144e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1629680 0.2969392
sample estimates:
      cor
0.2310486
```

O **teste de correlação de Pearson** indica uma **correlação positiva fraca (r = 0.231)** entre **glicose** e **IMC (BMI)**.

O **p-valor (< 0.0001)** sugere que essa correlação é estatisticamente significativa, ou seja, não ocorre por acaso. No entanto, a relação é **fraca**, o que **indica que um aumento no IMC não necessariamente leva a um aumento proporcional nos níveis de glicose.**

Após isso, gerei o modelo de regressão linear:

```
Call:
lm(formula = BMI ~ Glucose, data = diabetes)

Residuals:
    Min      1Q  Median      3Q     Max
-13.238  -4.881  -0.440   4.139  34.262

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.10627    0.99572  26.218  < 2e-16 ***
Glucose      0.05219    0.00794   6.572 9.14e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.694 on 766 degrees of freedom
Multiple R-squared:  0.05338,   Adjusted R-squared:  0.05215
F-statistic:  43.2 on 1 and 766 DF,  p-value: 9.144e-11
```

O **modelo de regressão linear** indica que a **glicose** tem um efeito positivo, mas fraco, sobre o **IMC (BMI)**.

Coeficiente de glicose (0.05219): A cada aumento de 1 unidade na glicose, o IMC aumenta, em média, **0.052** unidades.

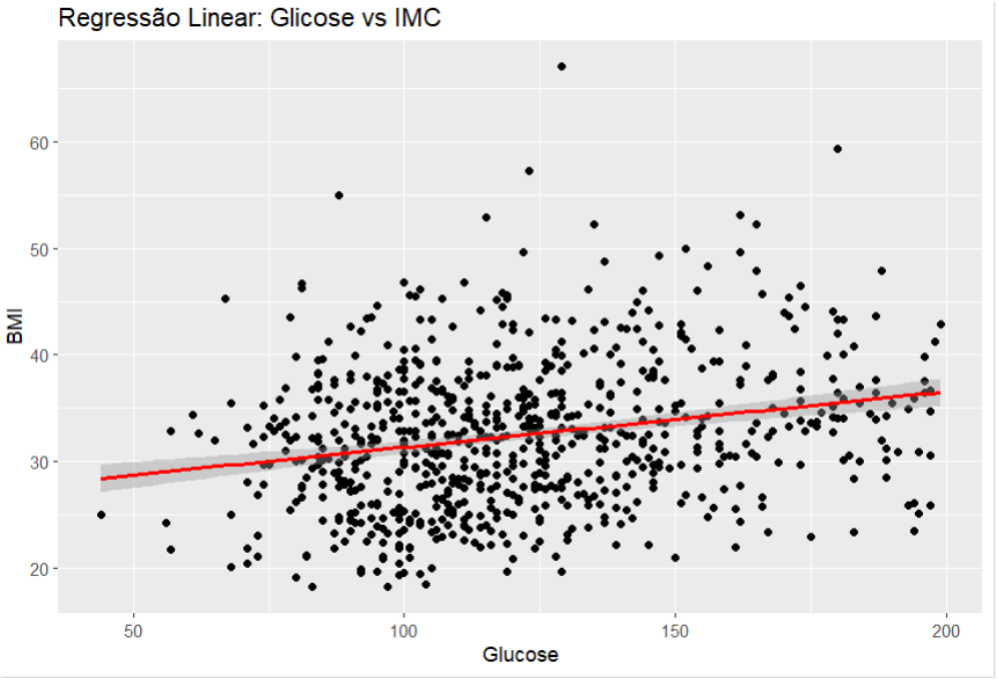
p-valor (< 0.0001): O efeito é estatisticamente significativo

Posteriormente gerei os seguintes valores:

```
>
> cat("R²:", r2, "\n")
R²: 0.05338343
> cat("MAE:", mae, "\n")
MAE: 5.28348
> cat("RMSE:", rmse, "\n")
RMSE: 6.684793
```

O modelo tem baixo poder preditivo e erros relativamente altos, indicando que a **glicose não é um bom preditor do IMC**.

Após verificar os valores gerei um gráfico de dispersão com linha de regressão para confirmar.



Pontos pretos: Representam as observações individuais da base de dados.

Linha vermelha: Linha de regressão linear ajustada, que representa a relação média entre glicose e IMC.

Inclinação positiva: Indica que, à medida que os níveis de glicose aumentam, há uma leve tendência de aumento do IMC.

A inclinação da linha é **pequena**, o que confirma que o impacto da glicose sobre o IMC **não é significativo**.

O espalhamento dos pontos ao redor da linha indica **alta variabilidade** e pouca previsibilidade.

O coeficiente de determinação $\mathbf{R^2 \approx 0.05}$ confirma que a glicose explica apenas **5,3% da variação no IMC**, sugerindo que outros fatores influenciam mais o IMC.

Em seguida, fiz o modelo de **regressão logística**:

```
Call:
glm(formula = Outcome ~ Glucose, family = binomial, data = diabetes)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.701627   0.436938  -13.05   <2e-16 ***
Glucose      0.040565   0.003377   12.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 793.64  on 766  degrees of freedom
AIC: 797.64

Number of Fisher Scoring iterations: 4
```

Intercept (-5.7016): Representa o log-odds (logaritmo da razão de chances) do Outcome = 1 quando **Glucose = 0**.

Glucose (0.0406): Indica que um aumento de **1 unidade** na glicose **aumenta os log-odds** de desenvolver diabetes em **0.0406**.

Para interpretar de forma mais intuitiva, podemos calcular o **odds ratio**:

e0.0406≈1.041e^{0.0406} \approx 1.041e0.0406≈1.041

Isso significa que **um aumento de 1 unidade na glicose aumenta a chance de um diagnóstico positivo em aproximadamente 4,1%**.

Significância Estatística

> O p-valor de **Glucose** é **< 2e-16**, indicando que a variável é estatisticamente significativa para prever o Outcome.

- **A glicose é um preditor significativo** do desfecho (diabetes ou não).
- O aumento da glicose **eleva a chance de desenvolver diabetes**.
- O modelo é estatisticamente válido, mas pode precisar de mais variáveis para melhorar sua precisão.

Em seguida, gerei a predição e avaliação do modelo de regressão logística:

```
Call:
glm(formula = Outcome ~ Glucose, family = binomial, data = diabetes)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.701627   0.436938  -13.05   <2e-16 ***
Glucose      0.040565   0.003377   12.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 793.64  on 766  degrees of freedom
AIC: 797.64

Number of Fisher Scoring iterations: 4
```

[(predict(log_model, type = "response")]: Gera **probabilidades previstas** de Outcome = 1 (diabetes).

[(ifelse(... > 0.5, 1, 0)] : Converte essas probabilidades em **classes binárias preditas**:

- Se a probabilidade for maior que 0.5 → prevê 1 (diabetes).
- Caso contrário → prevê 0 (sem diabetes).

Após isso gerei a matriz confusão

```
          Actual
Predicted   0   1
        0 440 136
        1  60 132
```

TP (Verdadeiro Positivo) = 132
→ O modelo previu corretamente que 132 pessoas **têm diabetes**.

TN (Verdadeiro Negativo) = 440
→ O modelo previu corretamente que 440 pessoas **não têm diabetes**.

FP (Falso Positivo) = 60
→ O modelo previu diabetes para 60 pessoas que **não têm**.

FN (Falso Negativo) = 136
→ O modelo previu que 136 pessoas **não têm diabetes**, mas na verdade **têm**.

Posteriormente, calculei a acurácia:

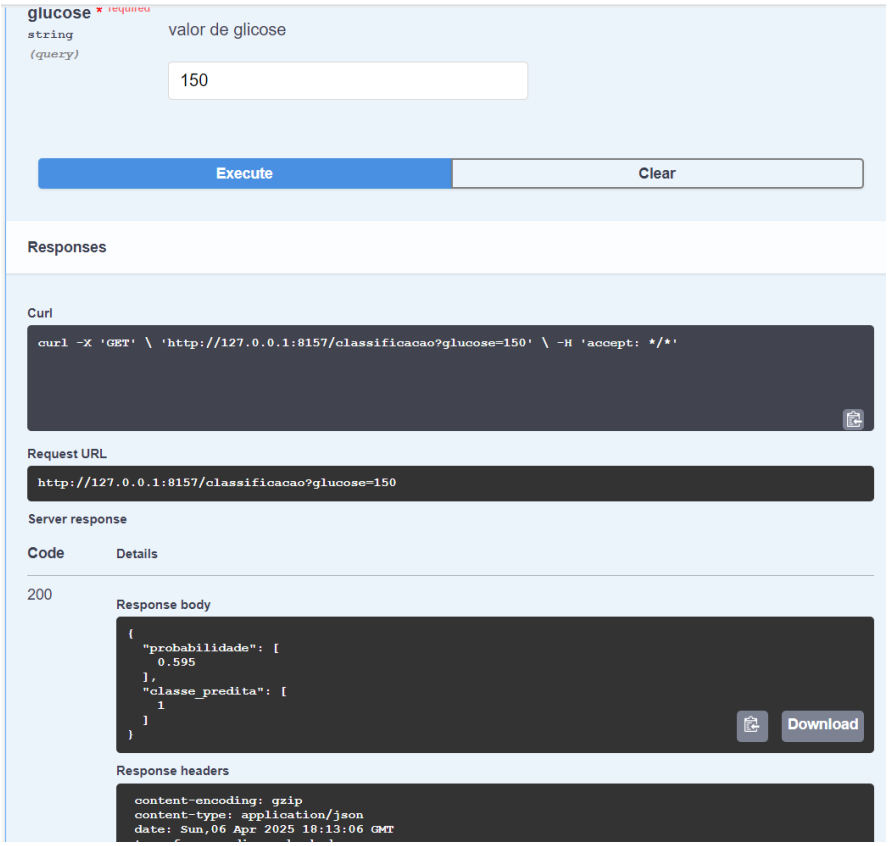
```
> # Acurácia
> accuracy <- sum(diabetes$pred_class == diabetes$Outcome) / nrow(diabetes)
> cat("Acurácia:", accuracy)
Acurácia: 0.7447917
>
```

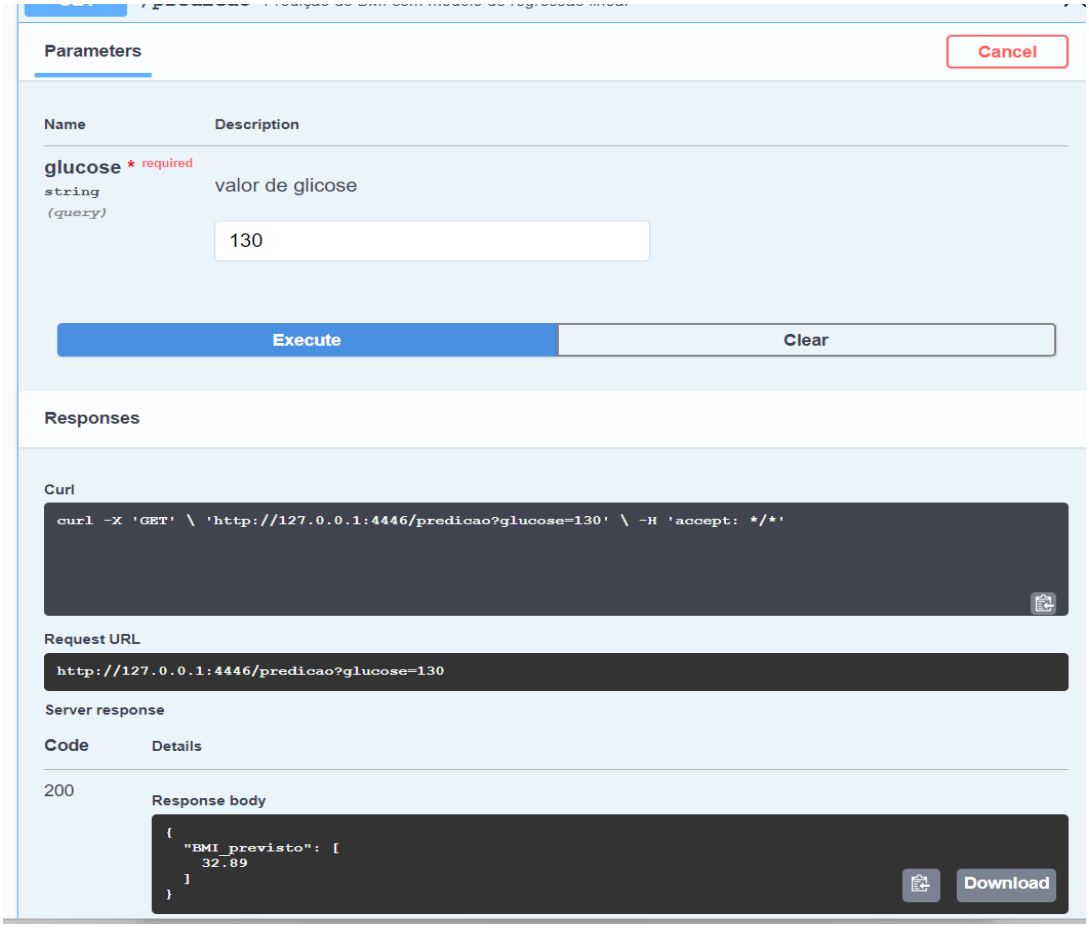
O modelo tem **boa acurácia (74%)**, mas:

- **Alta taxa de falsos negativos** (FN = 136): pessoas com diabetes que o modelo não detectou.
- Isso é **crítico em aplicações médicas**, pois deixar de identificar uma condição como diabetes pode ser perigoso.

A **precisão está boa**, mas a **sensibilidade (recall)** é baixa.

E para finalizar, conclui no Swagger que você insere o valor da glicose e retorna a probabilidade da pessoa possuir diabetes e a conclusão nos valores de **0(não possui diabetes) e 1(possui diabetes).**





Conclusão

Para concluir, neste projeto, desenvolvi dois modelos preditivos a partir do conjunto de dados diabetes: um modelo de **regressão logística** para classificar o risco de diabetes com base nos níveis de **glicose**, e um modelo de **regressão linear** para prever o **Índice de Massa Corporal (IMC)** com base na mesma variável.

O modelo de regressão logística demonstrou uma **boa acurácia (74,4%)**, sendo eficaz em identificar pacientes não diabéticos (**especificidade de 88%**), embora apresente uma **sensibilidade limitada (49,3%)**, indicando que parte dos casos de diabetes não são detectados. Ainda assim, a variável **glicose** mostrou-se altamente significativa (**p < 0.001**) na previsão do desfecho, reforçando sua importância como fator de risco para diabetes. A **precisão geral** do

modelo foi de **0,66**, indicando um desempenho equilibrado, especialmente na correta classificação dos casos positivos e negativos.

No modelo de **regressão linear**, observou-se uma **relação positiva entre os níveis de glicose e o IMC**, sugerindo que, à medida que os níveis de glicose aumentam, também há tendência de aumento no índice de massa corporal. Embora a correlação não seja extremamente forte, o modelo apresentou significância estatística e mostrou-se útil como uma estimativa inicial. Essa relação pode indicar que pacientes com níveis elevados de glicose tendem a apresentar maiores valores de IMC, o que reforça o papel do sobrepeso e da obesidade como fatores associados ao risco de desenvolver diabetes.

Trabalho Elaborado por **Breno de Souza**

Turma: Projeto de Machine Learning- 2025.1