Projeto Final: Análise descritiva em um *dataset* sobre diabetes

Beatriz Magiore e Guilherme Rodrigues

Professor: Prof. Dr. Thomas Nogueira Vilches

Apresentação do Projeto Final
Programa de Pós-Graduação em Biometria
Universidade Estadual Paulista "Júlio de Mesquita Filho"

Instituto de Biociências de Botucatu - Unesp

13 de julho de 2023

Sumário

Sumário

Diabetes

Definição

Diabetes Mellitus é uma síndrome metabólica de origem múltipla, decorrente da falta de insulina e/ou da incapacidade de a insulina exercer adequadamente seus efeitos [Ministério da Saúde, 2023].

Tipos:

- Tipo 1: causada pela destruição das células produtoras de insulina, em decorrência de defeito do sistema imunológico.
- Tipo 2: resulta da resistência à insulina e de deficiência na secreção de insulina.
- Diabetes Gestacional: é a diminuição da tolerância à glicose, diagnosticada pela primeira vez na gestação, podendo ou não persistir após o parto.
- Outros tipos: decorrentes de defeitos genéticos associados com outras doenças ou com o uso de medicamentos.

Estatísticas mundiais e no Brasil

No mundo

Cerca de 422 milhões de pessoas convivem com o diabetes, a maioria vivendo em países de baixa e média renda, e 1,5 milhão de mortes são diretamente atribuídas ao diabetes a cada ano [WHO, 2023].

No Brasil

De acordo com a Sociedade Brasileira de Diabetes, existem atualmente mais de 13 milhões de pessoas vivendo com a doença, representando 6,9% da população nacional [Ministério da Saúde, 2023].

Causas e riscos para a saúde

- O diabetes é uma das principais causas de cegueira, insuficiência renal, ataques cardíacos, derrame e amputação de membros inferiores.
- Uma dieta saudável, atividade física regular, manutenção de um peso corporal normal e evitar o uso de tabaco são formas de prevenir ou retardar o aparecimento do diabetes.

Sobre o conjunto de dados

O conjunto de dados de previsão de diabetes [MUSTAFA, 2023] consiste em uma compilação de informações médicas e demográficas de pacientes, acompanhadas da condição de diabetes (1 para a presença de diabetes e 0 para ausência de diabetes). Os atributos abrangem uma variedade de características, incluindo idade, gênero, índice de massa corporal (IMC), hipertensão, doenças cardíacas, histórico de tabagismo, nível de HbA1c (hemoglobina glicada) e nível de glicose no sangue.

Sumário

Objetivos

Estudar a relação entre o diabetes e as demais variáveis, além de colocar em prática as ferramentas aprendidas no decorrer da disciplina, trabalhando em colaboração por meio da ferramenta de versionamento de texto Git.

Sumário

Metodologia

Todos os resultados foram obtidos utilizando linguagem de programação R e o trabalho foi desenvolvido em colaboração por meio da ferramenta Git e um repositório no GitHub.

Análises

Realizamos três análises:

- Análises descritivas dos dados de modo geral para conhecer o *dataset* e saber os tipos de variáveis presentes.
- Análises descritivas das variáveis categóricas: gênero, hipertensão, doenças cardíacas, histórico de tabagismo, diabetes.
- Análises descritivas considerando as variáveis numéricas: idade, IMC, nível de HbA1c, nível de glicose no sangue.

Sumário

Análises descritivas dos dados de modo geral

```
str(sapply(dados, unique))
```

```
## List of 9
##  $ gender             : chr [1:3] "Female" "Male" "Other"
##  $ age                : num [1:102] 80 54 28 36 76 20 44 79 42 32 ...
##  $ hypertension       : int [1:2] 0 1
##  $ heart_disease      : int [1:2] 1 0
##  $ smoking_history    : chr [1:6] "never" "No Info" "current" "former" ...
##  $ bmi                : num [1:4247] 25.2 27.3 23.4 20.1 19.3 ...
##  $ HbA1c_level        : num [1:18] 6.6 5.7 5 4.8 6.5 6.1 6 5.8 3.5 6.2 ...
##  $ blood_glucose_level: int [1:18] 140 80 158 155 85 200 145 100 130 160 ...
##  $ diabetes           : int [1:2] 0 1
```

Figura 1: Informações das variáveis do *dataframe*.

Análises descritivas dos dados de modo geral

```
##     gender           age        hypertension heart_disease    smoking_history
##  Female:58552   Min.   : 0.08   0:92515      0:96058       current     : 9286
##  Male  :41430   1st Qu.:24.00   1: 7485      1: 3942       ever        : 4004
##  Other :   18   Median :43.00                              former      : 9352
##                 Mean   :41.89                              never       :35095
##                 3rd Qu.:60.00                              No Info     :35816
##                 Max.   :80.00                              not current: 6447
##       bmi         HbA1c_level    blood_glucose_level diabetes
##  Min.   :10.01   Min.   :3.500   Min.   : 80.0       0:91500
##  1st Qu.:23.63   1st Qu.:4.800   1st Qu.:100.0       1: 8500
##  Median :27.32   Median :5.800   Median :140.0
##  Mean   :27.32   Mean   :5.528   Mean   :138.1
##  3rd Qu.:29.58   3rd Qu.:6.200   3rd Qu.:159.0
##  Max.   :95.69   Max.   :9.000   Max.   :300.0
```

Figura 2: Resumo estatístico dos dados.

Análises descritivas dos dados de modo geral

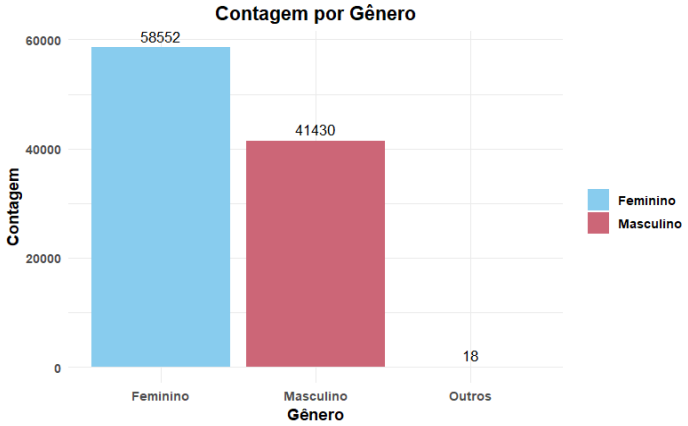


Figura 3: Contagem de pacientes por gênero.

Análises descritivas dos dados de modo geral

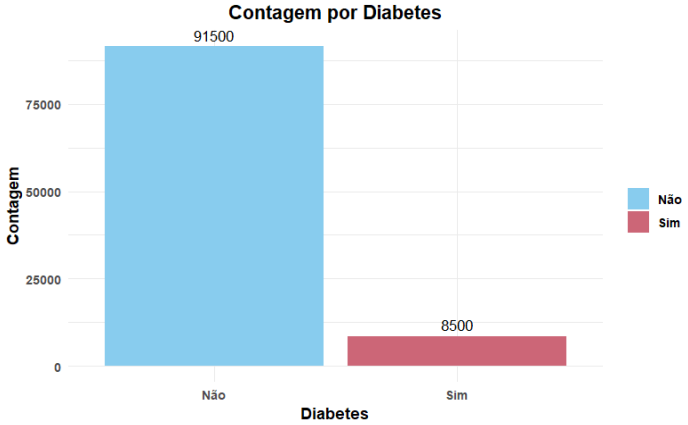
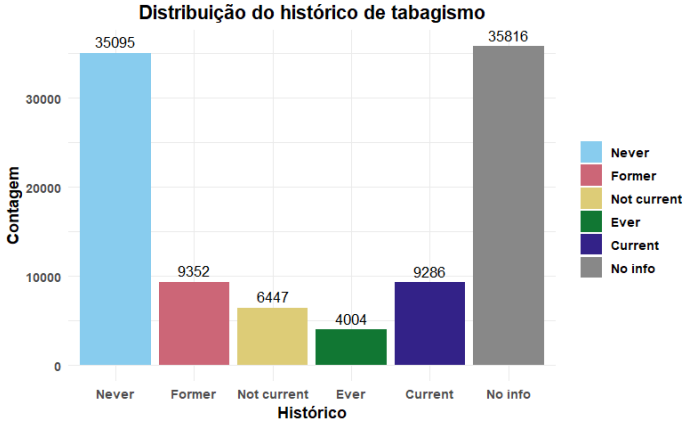


Figura 4: Contagem de pacientes com e sem diabetes.

Análises descritivas dos dados de modo geral



Análises descritivas dos dados de modo geral

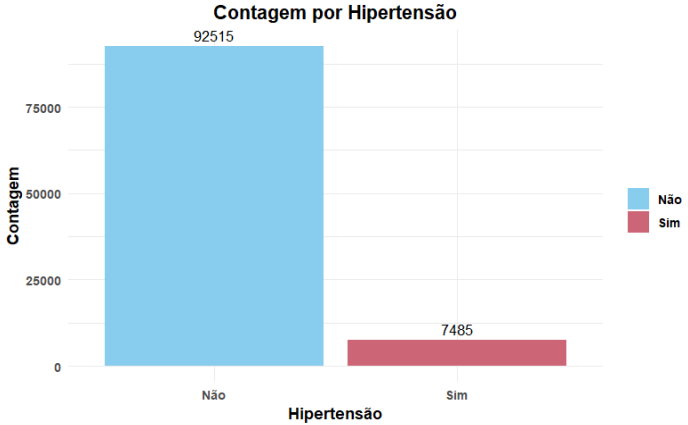


Figura 5: Contagem de pacientes com e sem hipertensão.

Análises descritivas dos dados de modo geral

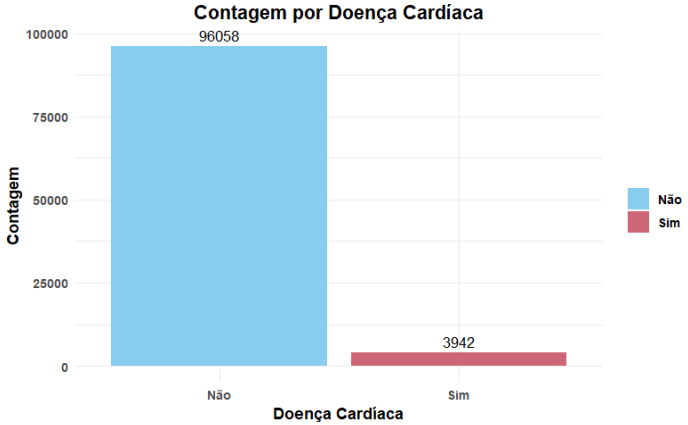


Figura 6: Contagem de pacientes com e sem doença cardíaca.

Análise dos indivíduos de gênero *others*

```
(dados_dicotomicos %>% filter(gender == "Other"))
```

```
##    gender hypertension heart_disease diabetes
## 1   Other            0             0        0
## 2   Other            0             0        0
## 3   Other            0             0        0
## 4   Other            0             0        0
## 5   Other            0             0        0
## 6   Other            0             0        0
## 7   Other            0             0        0
## 8   Other            0             0        0
## 9   Other            0             0        0
## 10  Other            0             0        0
## 11  Other            0             0        0
## 12  Other            0             0        0
## 13  Other            0             0        0
## 14  Other            0             0        0
## 15  Other            0             0        0
## 16  Other            0             0        0
## 17  Other            0             0        0
## 18  Other            0             0        0
```

Figura 7: Descrição dos indivíduos de outro gênero.

Análise descritiva considerando as variáveis categóricas

```
## # A tibble: 2 × 11
##   gender media…¹ media…² media…³ dp_hy…⁴ dp_he…⁵ dp_di…⁶ CV_hy…⁷ CV_he…⁸ CV_di…⁹
##   <chr>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 Female  0.0717  0.0267  0.0762   0.258   0.161   0.265    3.60    6.04    3.48
## 2 Male    0.0794  0.0574  0.0975   0.270   0.233   0.297    3.41    4.05    3.04
## # … with 1 more variable: N <int>, and abbreviated variable names
## #   ¹media_hypertension, ²media_heart_desease, ³media_diabetes,
## #   ⁴dp_hypertension, ⁵dp_heart_desease, ⁶dp_diabetes, ⁷CV_hypertension,
## #   ⁸CV_heart_desease, ⁹CV_diabetes
```

Figura 8: Média, desvio padrão e coeficiente de variação por gênero.

Análise descritiva considerando as variáveis categóricas

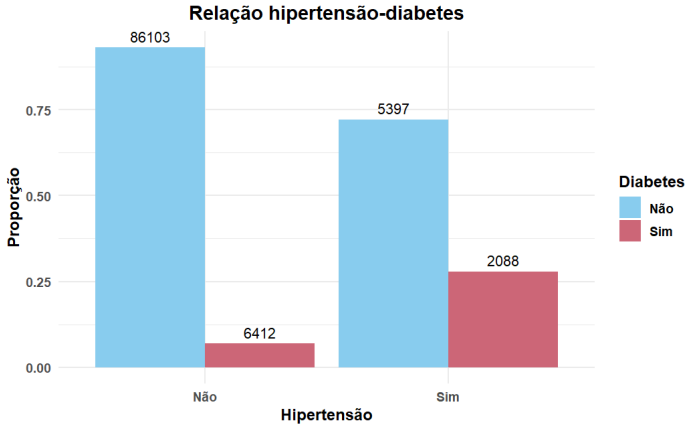


Figura 9: Relação entre hipertensão e diabetes.

Análise descritiva considerando as variáveis categóricas

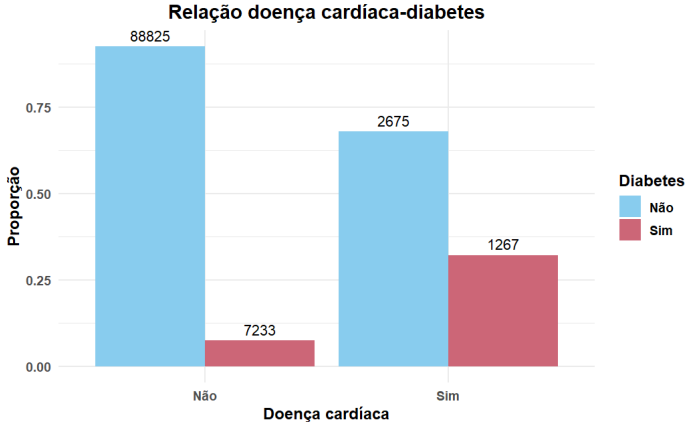


Figura 10: Relação entre doença cardíaca e diabetes.

Análise descritiva considerando as variáveis categóricas

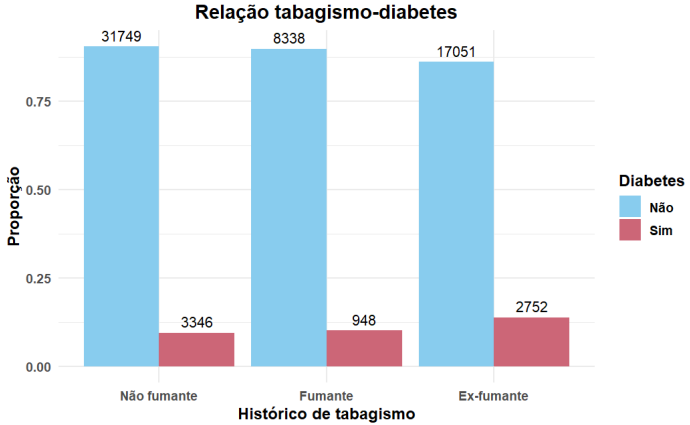
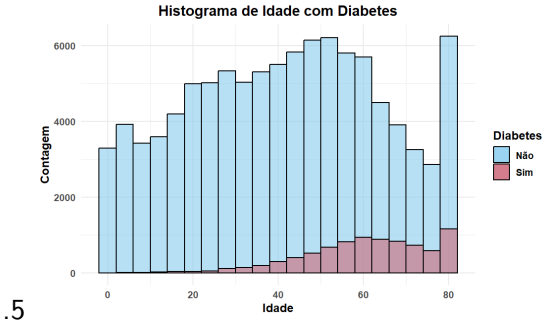


Figura 11: Relação entre tabagismo e diabetes.

Análise descritiva considerando as variáveis numéricas



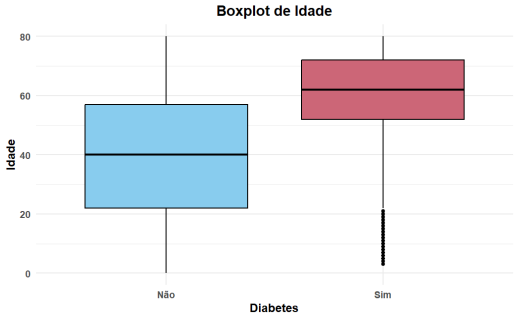


Figura 12: Histograma e boxplot de idade

Análise descritiva considerando as variáveis numéricas

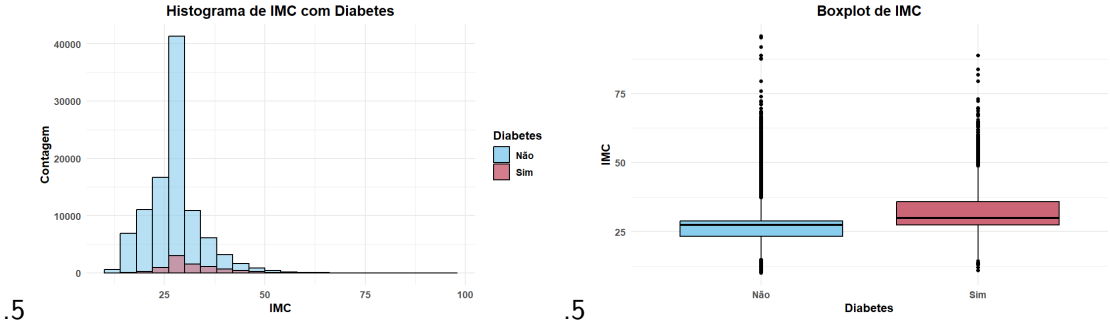


Figura 13: Histograma e boxplot de IMC

Análise descritiva considerando as variáveis numéricas

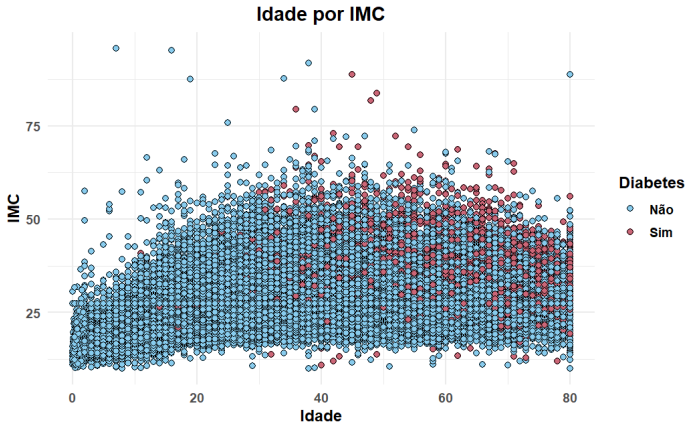


Figura 14: Scatterplot de idade por IMC.

Análise descritiva considerando as variáveis numéricas

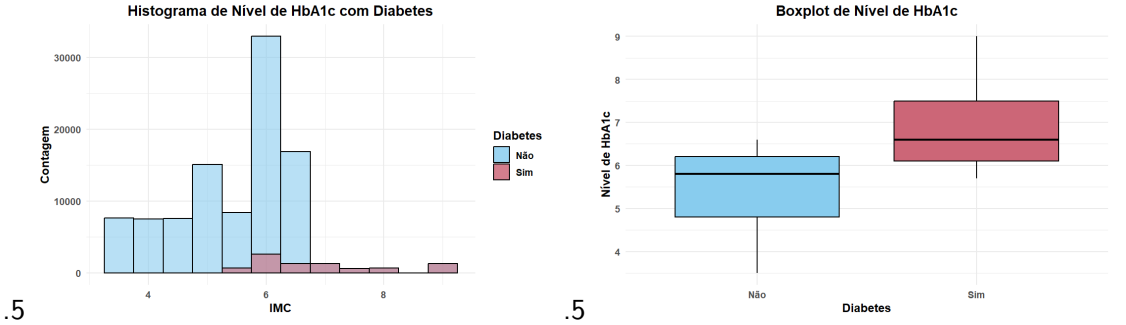


Figura 15: Histograma e boxplot de nível de HbA1c

Análise descritiva considerando as variáveis numéricas

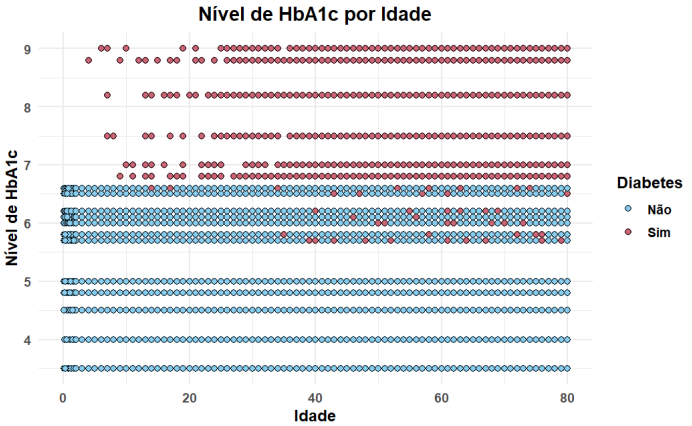


Figura 16: Scatterplot de idade por nível de HbA1c.

Análise descritiva considerando as variáveis numéricas

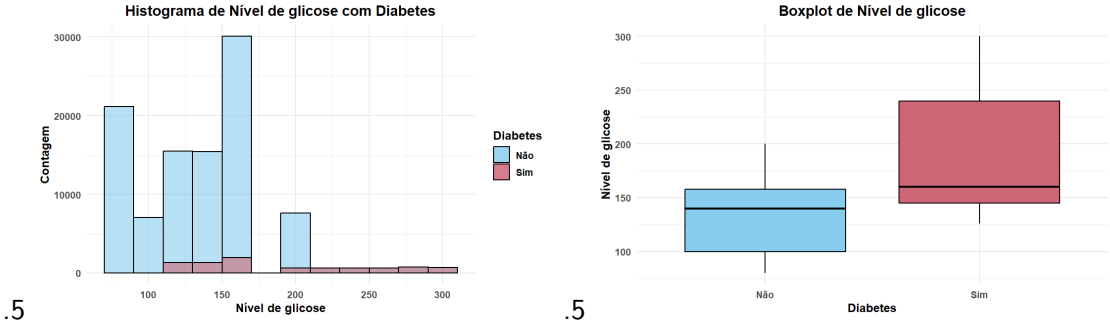


Figura 17: Histograma e boxplot de nível de glicose no sangue.

Análise descritiva considerando as variáveis numéricas

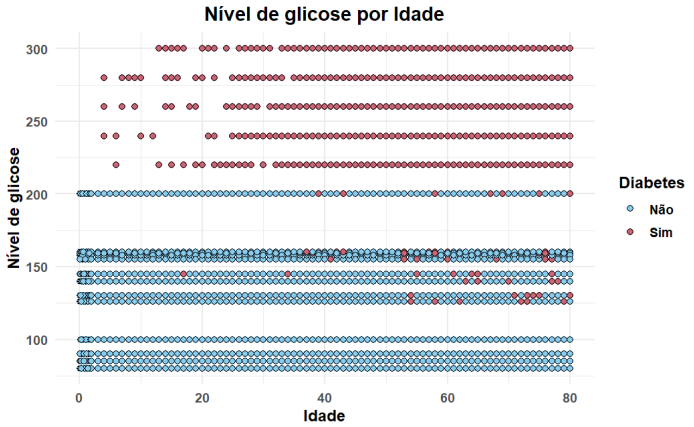


Figura 18: Scatterplot de idade por nível de glicose.

Correlação das variáveis com a diabetes

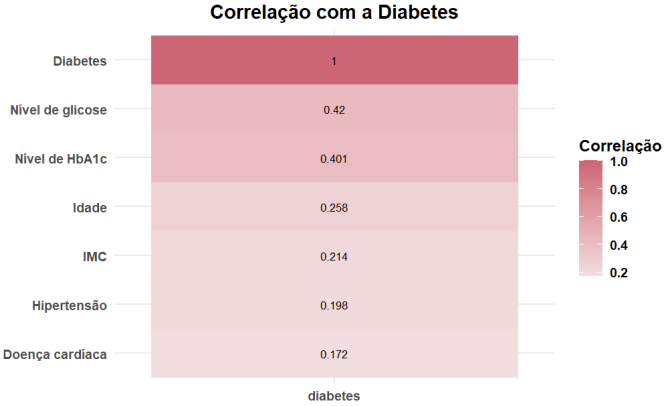


Figura 19: Correlação com a diabetes.

Sumário

Conclusões

- Observamos que a prevalência de diabetes é maior em homens em comparação com mulheres. Além disso, pacientes com hipertensão e doenças cardíacas têm maior propensão a ter diabetes.
- Verificamos que o histórico de tabagismo apresenta uma relação sutil com o diagnóstico de diabetes, sem uma associação clara.
- Em relação às variáveis numéricas, observamos que a idade e o IMC têm uma associação positiva com o diagnóstico de diabetes. Indivíduos com idade acima de 40 anos e IMC acima de 30 apresentam maior incidência da doença.

Conclusões

- Os níveis de HbA1c e glicose no sangue apresentaram uma relação mais clara com o diagnóstico de diabetes. Valores elevados dessas variáveis estão associados a um maior risco de desenvolver a doença.
- Ao considerar a correlação da diabetes com as variáveis, confirmamos que as maiores correlações estão com os níveis de glicose e HbA1c, reforçando sua importância como indicadores da doença.
- Com base nessas análises, é possível concluir que fatores como idade, gênero, hipertensão, doenças cardíacas, níveis de HbA1c e glicose no sangue desempenham um papel significativo na incidência de diabetes. Esses resultados podem auxiliar na identificação de indivíduos em risco e no desenvolvimento de estratégias de prevenção e tratamento personalizados.

Bibliografia

Ministério da Saúde. Plano de reorganização da atenção à hipertensão arterial e ao diabetes mellitus: hipertensão arterial e diabetes mellitus. Online, 2023. URL https://bvsms.saude.gov.br/diabetes/. Acesso em 11/07/23.

M. MUSTAFA. Diabetes prediction dataset: A comprehensive dataset for predicting diabetes with medical & demographic data (kaggle). Online, 2023. URL https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset?select=diabetes_prediction_dataset.csv. Acesso em 13/06/23.

WHO. Diabetes. Online, 2023. URL https://www.who.int/news-room/fact-sheets/detail/diabetes. Acesso em 11/07/23.