Instituição Evangélica de Novo Hamburgo

Previsão de AVC com dataset de classificação binária

João Pedro Barbosa

Novo Hamburgo

2025

**SUMÁRIO**

[**1. INTRODUÇÃO 7**](#_97wcjlzawqnd)

[**2. DESENVOLVIMENTO 7**](#_4ltowocm8orx)

[2.1 O que é o AVC? 7](#_h1h55vfokhha)

[2.2 Tipos de AVC 7](#_ttn7yvku7mcn)

[2.2.1 AVC isquêmico 7](#_p71z1acqr95e)

[2.2.2 AVC hemorrágico 7](#_feq0ekzq9q)

[2.3 Fatores de risco de AVC 8](#_9gtxhyntal41)

[**3. PRÉ PROCESSAMENTO E EXPLORAÇÃO DOS DADOS 8**](#_9h1aef4uym70)

[3.1 Descrição do dataset 8](#_loi5bnigf7dp)

[3.2 Gráfico ocorrência de AVC no dataset 9](#_bmaotnecgqsr)

[3.3 Justificativa para a escolha do dataset 10](#_a4218luk3hjs)

[3.4 Análise exploratória 10](#_3r0yaswi3w6i)

[3.4.1 Casos de AVC em hipertensos 10](#_yuteb7rl80dq)

[3.4.2 Casos de AVC em pessoas que tinham alguma doença do coração 14](#_53qayv3iagc)

[3.4.3 Distribuição de casos de AVC por estado civil 17](#_siiw1e8n76pp)

[3.4.4 Distribuição de AVC conforme o tipo de trabalho 21](#_lg6om9npn5zb)

[3.4.5 Distribuição de AVC conforme zona habitada 25](#_8dydrq3bglb)

[3.4.6 Distribuição de AVC conforme níveis de glicose 29](#_elgdc8ve2w4z)

[3.4.7 Distribuição de AVC conforme IMC 31](#_hcugg3tb5c97)

[3.4.8 Distribuição de AVC por grupo de risco 36](#_wq61qc19up8e)

[3.5 Pré Processamento dos dados 37](#_6oez2j2cxw0r)

[3.5.1 Verificação de valores nulos e análise preliminar 37](#_pibgs2xfofcv)

[3.5.1.1 Verificação do cabeçalho do dataset 37](#_wsmzw5z7wo72)

[3.5.1.2 Verificação de valores nulos dentro do dataset 38](#_iborhiwfyy0d)

[3.5.2 Tratamento de dados 39](#_32alw03uhgay)

[3.5.2.1 Tratamento de pessoas casadas com idade insuficiente 39](#_nji7hukth7bu)

[3.5.2.2 Tratamento de pessoas com idade insuficiente empregadas 39](#_ezfmj4t2fh8h)

[3.5.2.3 Tratamento de valores nulos de IMC 40](#_libe48kb4w83)

[3.5.2.4 Tratamento de conversão para valores numéricos 41](#_przkqlhk2o8e)

[**4. DESENVOLVIMENTO DO MODELO DE DEEP LEARNING 44**](#_hhf6wlviji1p)

[4.1 Modelo utilizado 44](#_9ipflu9cvhpz)

[4.2 Treinamento do modelo 44](#_xsd5gqh4spt4)

[4.2.1 Colunas utilizadas no treinamento 44](#_7v96tqu8kw1e)

[4.2.2 Treinamento sem regularização 45](#_bwjhfyk0vwfp)

[4.2.2.1 Seleção do dado previsor 45](#_nz7wbbbc53xb)

[4.2.2.2 Criação do treino e teste 45](#_wwqv3mv0vgrz)

[4.2.2.3 Camadas e neurônios 46](#_z6l4byqwh49v)

[4.2.2.4 Treino e teste 46](#_r626u5rkqgoo)

[4.2.3 Treinamento com regularização 47](#_oog5m8p2n7be)

[4.2.3.1 Seleção do dado previsor 47](#_oxn50etqqyj)

[4.2.3.2 Criação do treino e teste 47](#_2hcrs6cp0d0o)

[4.2.3.3 Camadas e neurônios 47](#_3luovbaou5lo)

[4.2.3.4 Ajuste de pesos das classes 48](#_ig44doiit0qe)

[4.2.3.5 Treino e teste 48](#_uw21yvo5igwm)

[**5. AVALIAÇÃO DO MODELO 49**](#_u7gf33r8tdqt)

[5.1 Avaliação modelo não regularizado 49](#_lw6kssw6dtr9)

[5.1.1 Matriz de confusão 49](#_jqv69vsy5aal)

[5.1.2 Acurácia 51](#_jt5sumu8xyvf)

[5.1.3 Recall e verdadeiro negativo 51](#_p77rzesf3ucz)

[5.1.4 Resultados, melhorias e ajustes 52](#_5bm36eexq3sd)

[5.2 Avaliação modelo regularizado 52](#_69hbzf4wlblb)

[5.2.1 Matriz de confusão 53](#_nayt6997057m)

[5.2.2 Acurácia 54](#_lmao2cn2p79)

[5.2.3 Recall e verdadeiro negativo 54](#_1wewmcj5zfmf)

[5.2.4 Falsos negativos e positivos 56](#_7a6xbjlllkjn)

[5.2.5 Recall fatores de risco 57](#_etjuoxfi7tm3)

[5.2.6 Recall de pessoas que tiveram AVC e que não tinham fator de risco 59](#_garr9e2r06b2)

[**6. CONCLUSÃO 61**](#_1fatw55s2a2v)

[**REFERÊNCIAS 63**](#_wgantue0wjn)

# 

**LISTA DE FIGURAS**

[Figura 1 - Distribuição de AVC dentro do dataset 10](#_hud0v2e1p3x)

[Figura 2 - Criação do gráfico de quantidade de hipertensos no dataset 11](#_gk9ah0wnoy6b)

[Figura 3 - Gráfico quantidade de hipertensos no dataset 12](#_ww84jvao4aaa)

[Figura 4 - Criação do gráfico de dados de AVC em pessoas com hipertensão 12](#_9nroiji1xd3i)

[Figura 5 - Gráfico de ocorrência de AVC em hipertensos 13](#_n5pen0w3to4v)

[Figura 6 - Criação do gráfico de quantidade de pessoas com doença de coração 14](#_no1j82xrbryn)

[Figura 7 - Gráfico quantidade de pessoas com doença de coração 15](#_f40v082wsr52)

[Figura 8 - Criação do gráfico casos de AVC em pessoas que tinham alguma doença relacionada ao coração 15](#_7k5xm2tz5zh7)

[Figura 9 - Gráfico de casos de AVC em pessoas que tinham alguma doença do coração 16](#_gsyml9pmw2pd)

[Figura 10 - Criação do gráfico do número de pessoas por estado cívil 17](#_fjshdnrsblgb)

[Figura 11 - Gráfico de pessoas por estado cívil 18](#_9o723c3f46sb)

[Figura 12 - Criação do gráfico de casos de AVC por estado cívil 18](#_fy2wze41w37)

[Figura 13 - Criação do gráfico de casos de AVC por estado cívil 19](#_xr2qz5lptlku)

[Figura 14 - Ocorrência de AVC conforme estado civil 20](#_x4xqw0ppyizp)

[Figura 15 - Criação do gráfico distribuição de tipos de trabalho no dataset 21](#_u3yys4rc8y59)

[Figura 16 - Gráfico de pessoas por tipo de trabalho 22](#_w4pkz1ktmibv)

[Figura 17 - Criação do gráfico de casos de AVC em pessoas por tipo de trabalho 22](#_qs4u2xl3gu2z)

[Figura 18 - Criação do gráfico de casos de AVC por tipo de trabalho 23](#_c1dedwc9mmtq)

[Figura 19 - Criação do gráfico de casos de AVC em pessoas por tipo de trabalho 24](#_6r6pywoq5gqw)

[Figura 20 - Criação do gráfico de pessoas conforme zona habitada 25](#_7wyfnjz74rt7)

[Figura 21 - Gráfico de pessoas conforme zona habitada 26](#_c0mjllxjygxt)

[Figura 22 - Criação do gráfico de casos de AVC em pessoas por tipo de zona habitada 26](#_paxfmdfpbvat)

[Figura 23 - Criação do gráfico de casos de AVC em pessoas por tipo de zona habitada 27](#_5unxzij6owa7)

[Figura 24 - Criação do gráfico de casos de AVC em pessoas por tipo de zona habitada 28](#_a0gu7cj2dtkn)

[Figura 25 - Criação do gráfico de distribuição de AVC conforme os níveis de glicose 29](#_2r912g984e7y)

[Figura 26 - Gráfico de distribuição dos níveis de glicose no dataset por ocorrência de AVC 30](#_atiwrwb7gyit)

[Figura 27 - Criação do gráfico de casos de AVC em pessoas conforme IMC 31](#_koxr53hi82og)

[Figura 28 - Gráfico de distribuição do AVC conforme IMC 31](#_wwva0byw6qsn)

[Figura 29 - Criação do gráfico de distribuição de fumantes no dataset](#_7u6xudtb6d4t)32

[Figura 30 - Gráfico de distribuição de fumantes no dataset](#_eafd6rg155f)33

[Figura 31 - Criação do gráfico de casos de AVC em pessoas fumantes 33](#_qqx070nst3dr)

[Figura 32 - Criação do gráfico de casos de AVC em pessoas fumantes 34](#_w02qo1hqyajj)

[Figura 33 - Gráfico de AVC em fumantes 35](#_zew0tneqhavb)

[Figura 34 - Criação do gráfico de fatia de pessoas com fator de risco que tiveram AVC ou não 36](#_w2lbgg8mjygu)

[Figura 35 - Criação do gráfico de fatia de pessoas com fator de risco que tiveram AVC ou não 36](#_2ugkaptik3vc)

[Figura 36 - Gráfico de casos de AVC em pessoas por fatores de risco 37](#_59oehhx0qnpv)

[Figura 37 - Ler os 5 primeiros valores das linhas das colunas 37](#_zfx6cmh79wbq)

[Figura 38 - Valores das primeiras 5 linhas das colunas 38](#_2pbpqpohzacl)

[Figura 39 - Verificação de valores nulos do dataset 38](#_84p2gpk3kzyl)

[Figura 40 - Valores nulos do dataset 38](#_uyt2la3gjo0s)

[Figura 41 - Verifica pessoas com idade insuficiente casadas 39](#_x06idlw7f2xu)

[Figura 42 - Verifica pessoas com idade insuficiente para trabalhar 39](#_fdk43u751e3r)

[Figura 43 - Verificação de Pessoas com idade insuficiente para trabalhar 40](#_ad0py5hasijx)

[Figura 44 - Classifica pessoas com 13 anos ou menos como “criança” 40](#_n68r4j7i4phh)

[Figura 45 - Verifica se o tratamento ocorreu corretamente 40](#_vs2dgrwry9c2)

[Figura 46 - Tratamento dos valores nulos IMC 40](#_4y2f567xrk84)

[Figura 47 - Verifica se há algum valor nulo pendente 41](#_nrsq4fqb2kpt)

[Figura 48 - Verificar valores nulos 41](#_mc201fjvbpm9)

[Figura 49 - Converte Homem e Mulher 41](#_w5ix4zjpypnh)

[Figura 50 - Converte já foi casado Sim = 0 e Não = 1 42](#_hxqj09bkop99)

[Figura 51 - Converte os valores em números inteiros 42](#_6h01p6kxla2x)

[Figura 52 - Converte os valores em números inteiros 42](#_yizwv2v48vop)

[Figura 53 - Converte os valores em números inteiros 42](#_gc1fxs8luijx)

[Figura 54 - Verifica todos os tipos de trabalhos do dataset 42](#_222f3cv7fvm3)

[Figura 55 - Converte os trabalhos para valores numéricos 43](#_ty8w1rciq6eh)

[Figura 56 - Verifica se os valores foram devidamente tratados 43](#_jf1nnzko1sze)

[Figura 57 - Converte os tipos de área de residência para valores numéricos 43](#_2w7oxukdnl24)

[Figura 58 - Verifica os valores da coluna de status de fumante 43](#_2nusegimsxzb)

[Figura 59 - Conversão dos valores de fumante para número 43](#_dt1sxwlqowz9)

[Figura 60 - Verifica se os dados foram devidamente tratados 43](#_rxw8qieosqlw)

[Figura 61 - Verificação do cabeçalho 44](#_8ikku7g4wdjw)

[Figura 62 - Seleção das colunas para treinamento e previsão 45](#_1vsfmx1i14im)

[Figura 63 - Seleção das colunas para treinamento e previsão 45](#_euu7muc7qmca)

[Figura 64 - Seleciona o dado previsor 45](#_1u3bl3a032bg)

[Figura 65 - Teste e Treino 45](#_y4um2kpd2k54)

[Figura 66 - Camadas e neurônios 46](#_jeodkssot6x7)

[Figura 67 - Configuração do treino e teste do algoritmo 46](#_gbatmw116jl1)

[Figura 68 - Saída das 10 últimas épocas 46](#_hq497hhmgoux)

[Figura 69 - Seleção do dado previsor 47](#_t73s60gbk2cq)

[Figura 70 - Teste e Treino 47](#_knrauojvklw3)

[Figura 71 - Padronizando os valores de X 47](#_s8qsyxga35hj)

[Figura 72 - Neurônios e camadas 48](#_7wxr7ivva85u)

[Figura 73 - Ajuste automático do peso das classes 48](#_2rsb9y7z1om)

[Figura 74 - Treino do modelo 48](#_a8gkypd9il78)

[Figura 75 - 5 primeiras épocas 49](#_lhbcxqmhepb)

[Figura 76 - 5 últimas épocas 49](#_39dhflzexocn)

[Figura 77 - Criação da Matriz de Confusão 50](#_szvipenwv3nw)

[Figura 78 - Plot Matriz de Confusão 50](#_ysm5gwnlmcq9)

[Figura 79 - Matriz de confusão 50](#_bgr1yc838g10)

[Figura 80 - Acurácia do modelo 51](#_fvwqfljxc5el)

[Figura 81 - Recall e verdadeiro negativo 51](#_x1budmd37wci)

[Figura 82 - Recall e verdadeiro negativo 51](#_xrgno0qb2w9n)

[Figura 83 - Plot recall e verdadeiro negativo 51](#_6limwkz7lkut)

[Figura 84 - Recall e verdadeiro negativo 52](#_rcctfl7xbhmd)

[Figura 85 - Criação da matriz de confusão 53](#_37qh8nv91rv7)

[Figura 86 - Plot matriz de confusão 53](#_ipr9v77cm7p7)

[Figura 87 - Matriz de confusão 54](#_hxyyere56oe1)

[Figura 88 - Acurácia total do modelo 54](#_gexvtz8w6bv5)

[Figura 89 - Criação do verdadeiro negativo e recall 54](#_u5x53ui50li6)

[Figura 90 - Criação do verdadeiro negativo e recall 55](#_p1pzxx42r7d2)

[Figura 91 - Criação do plot 55](#_6vugawhrarg9)

[Figura 92 - Recall e verdadeiro negativo 55](#_s5tlq8jggj01)

[Figura 93 - Criação das classes 56](#_ohlx7x37njsx)

[Figura 94 - Criação das classes 56](#_mxoin4ryxahm)

[Figura 95 - Criação do plot 56](#_66pwnzeue800)

[Figura 96 - Falsos negativos e positivos 57](#_cvwtujurcbpb)

[Figura 97 - Definição dos fatores de risco 57](#_tnrt59c7mlpg)

[Figura 98 - Criação do plot 58](#_89evrx3xt5j8)

[Figura 99 - Criação do plot 58](#_8owzhndouvk0)

[Figura 100 - Previsão de pessoas com fatores de risco 59](#_tu9n7n8x7c0)

[Figura 101 - Definição dos fatores de risco 59](#_hsxkeyprp6bl)

[Figura 102 - Criação do plot 60](#_eh216ndd9knl)

[Figura 103 - Criação do plot 60](#_vhqr0eo18euc)

[Figura 104 - Previsão de AVC em pessoas que não tinham nenhum fator de risco 61](#_kuskm2fzcww9)

Lista de siglas

AVC Acidente Vascular Cerebral

IMC Índice de Massa Corporal

# 1. INTRODUÇÃO

A pesquisa e o desenvolvimento do algoritmo de Deep Learning tem como o objetivo aprender e classificar corretamente casos de AVC utilizando um dataset, levando em consideração fatores de risco presentes, através de um algoritmo de classificação binária, também realizando uma breve pesquisa para uma melhor compreensão e abordagem sobre o AVC.

# 2. DESENVOLVIMENTO

## 2.1 O que é o AVC?

De acordo com o Ministério da Saúde, o AVC acontece quando os vasos que levam sangue ao cérebro se entopem ou se rompem, provocando paralisia na região do cérebro que ficou sem circulação sanguínea, causando a morte das células nervosas das regiões cerebrais atingidas. É uma doença que é mais comum em homens e é uma das principais causas de morte, incapacitação e internações em todo o mundo, sendo apontado como a segunda maior causa de óbitos no mundo.

Em 2019, de acordo com a Sociedade Brasileira de AVC, que realiza estudos e compila dados mundiais de AVC e diversas outras doenças, foram contabilizados 12.2 milhões de casos de AVC, resultando em 6.55 milhões de mortes.

## 2.2 Tipos de AVC

### 2.2.1 AVC isquêmico

De acordo com o Ministério da Saúde, é responsável por grande parte dos casos de AVC, sendo o causador de 80% das ocorrências de AVC, ocorre quando os vasos sanguíneos se entopem, podendo ocorrer devido a uma trombose, que é a formação de placas em uma artéria principal do cérebro ou uma embolia, quando um trombo ou placa de gordura de outra parte do corpo se solta pela corrente sanguínea e chega até os vasos cerebrais.

### 2.2.2 AVC hemorrágico

De acordo com o Ministério da Saúde, é menos comum que o AVC isquêmico, porém mais letal, e ocorre devido ao rompimento dos vasos sanguíneos no interior do cérebro, causando uma hemorragia intracerebral, também podendo ocorrer uma hemorragia subaracnóide, que é o sangramento entre o cérebro e a membrana aracnóide, que compõe a meninge.

Por consequência dessa hemorragia subaracnóide, há um aumento da pressão intracraniana, que pode resultar em maior dificuldade na circulação sanguínea, agravando a ocorrência e dificultando a chegada de sangue em áreas não afetadas.

## 2.3 Fatores de risco de AVC

De acordo com o Ministério da Saúde, os principais fatores de risco associados a maior probabilidade de ter um AVC, são:

• Hipertensão

• Diabetes tipo 2

• Colesterol Alto

• Estresse

• Ser do sexo masculino

• Obesidade

• Tabagismo

• Consumo excessivo de álcool

• Idade

• Sedentarismo

# 3. PRÉ PROCESSAMENTO E EXPLORAÇÃO DOS DADOS

## 3.1 Descrição do dataset

O dataset utilizado para o treino do algoritmo é o Stroke Prediction Dataset, disponível na plataforma Kaggle, é um dataset para um algoritmo de classificação binária, e ele contém os seguintes dados que foram utilizados para treinar o algoritmo de deep learning:

• ID

• Gênero

• Idade

• Estado Civil

• Tipo de trabalho

• Nível médio de glicose

• IMC

• Tabagismo

• Hipertensão

• Doenças do coração

• Tipo de residência

• AVC (stroke)

O dataset possui no total 5110 linhas, com 12 colunas, sendo 1 das colunas o alvo da previsão (stroke), e uma coluna contendo o ID, que não foi utilizado para a realização do treinamento, e o restante utilizado para o treinamento do algoritmo de classificação binária.

O dataset é desbalanceado, possuindo no total 249 casos positivos de AVC (1) dentro do dataset, sendo os restantes 4861 casos negativos (0).

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

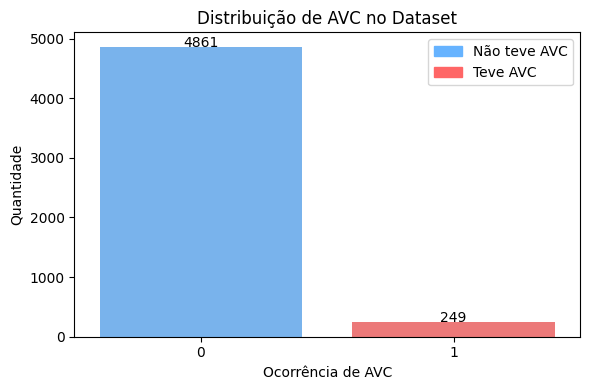
# 

# 

## 3.2 Gráfico ocorrência de AVC no dataset

O gráfico demonstra a distribuição dos casos de AVC dentro do dataset utilizado para o treinamento (Gráfico 1).

##### Figura 1 - Distribuição de AVC dentro do dataset



Fonte: Própria (2025)

# 

## 3.3 Justificativa para a escolha do dataset

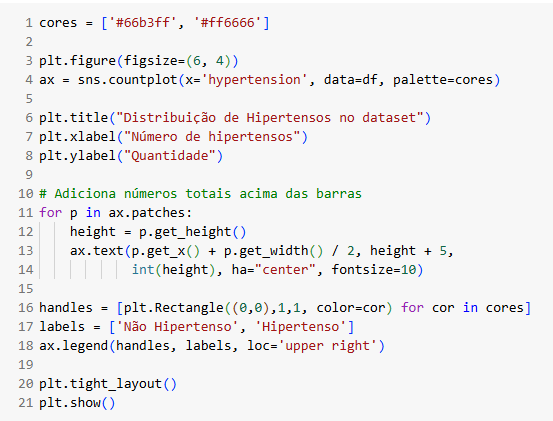
O dataset engloba diversas características presentes em pessoas que fazem parte do grupo de risco do AVC, possui um tamanho adequado para realizar os treinamentos e os testes (5110 linhas e 12 colunas) e possui poucos dados não tratados.

## 3.4 Análise exploratória

Foram gerados diversos gráficos para analisar a distribuição dos casos positivos de AVC dentro do dataset, comparando pessoas que tinham ou não determinado fator de risco que tiveram um AVC.

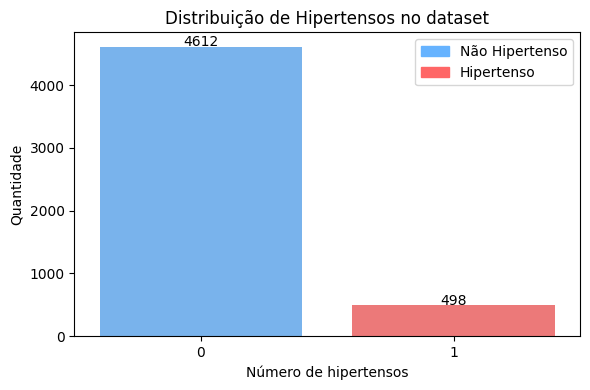
### 3.4.1 Casos de AVC em hipertensos

##### Figura 2 - Criação do gráfico de quantidade de hipertensos no dataset



Fonte: Própria (2025)

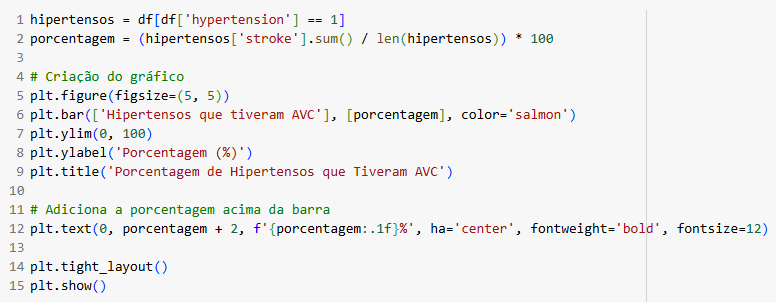
##### Figura 3 - Gráfico quantidade de hipertensos no dataset



Fonte: Própria (2025)

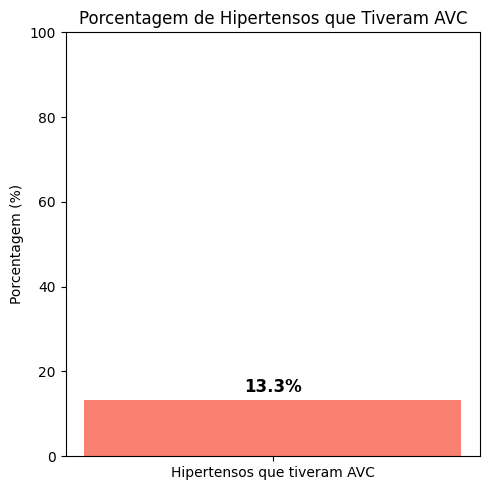
##### 

##### Figura 4 - Criação do gráfico de dados de AVC em pessoas com hipertensão



Fonte: Própria (2025)

##### Figura 5 - Gráfico de ocorrência de AVC em hipertensos

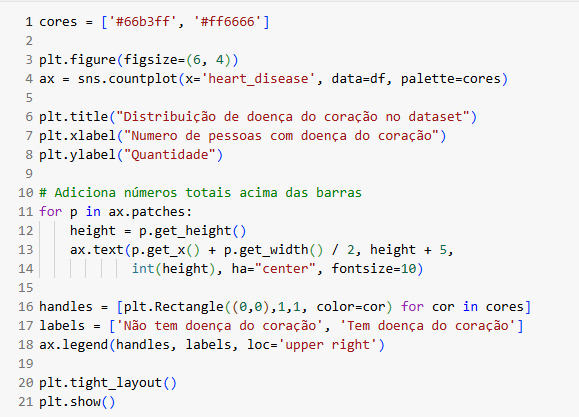


Fonte: Própria (2025)

### 3.4.2 Casos de AVC em pessoas que tinham alguma doença do coração

##### 

##### Figura 6 - Criação do gráfico de quantidade de pessoas com doença de coração



Fonte: Própria (2025)

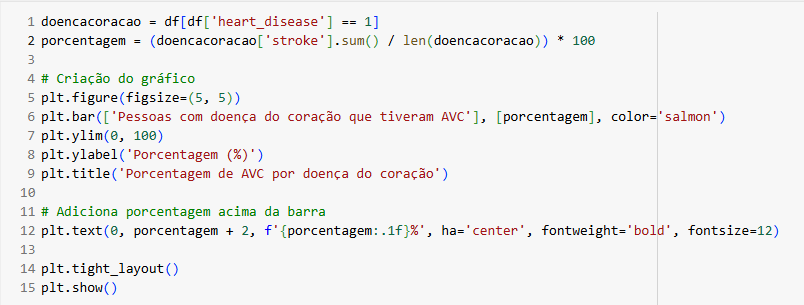
##### Figura 7 - Gráfico quantidade de pessoas com doença de coração

##### 

Fonte: Própria (2025)

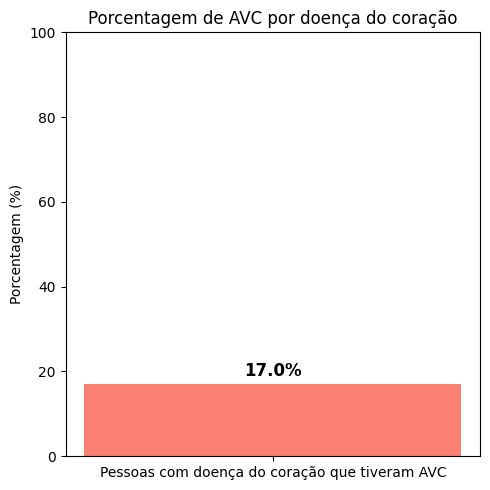
##### 

##### Figura 8 - Criação do gráfico casos de AVC em pessoas que tinham alguma doença relacionada ao coração



Fonte: Própria (2025)

##### Figura 9 - Gráfico de casos de AVC em pessoas que tinham alguma doença do coração



Fonte: Própria (2025)

## 

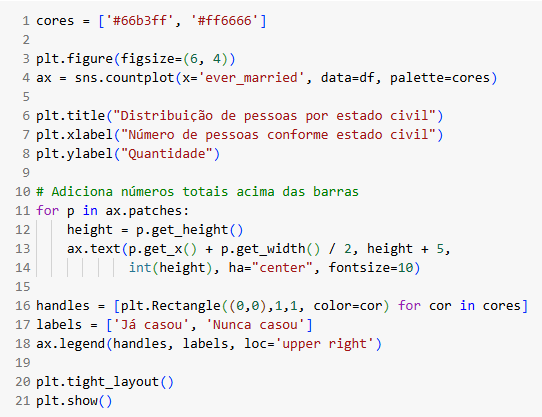
## 

## 

## 

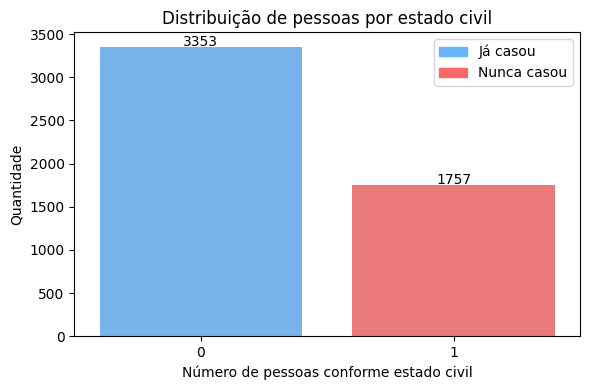
### 3.4.3 Distribuição de casos de AVC por estado civil

##### Figura 10 - Criação do gráfico do número de pessoas por estado cívil



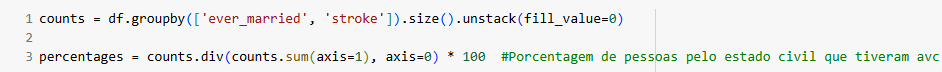
Fonte: Própria (2025)

##### Figura 11 - Gráfico de pessoas por estado cívil

****

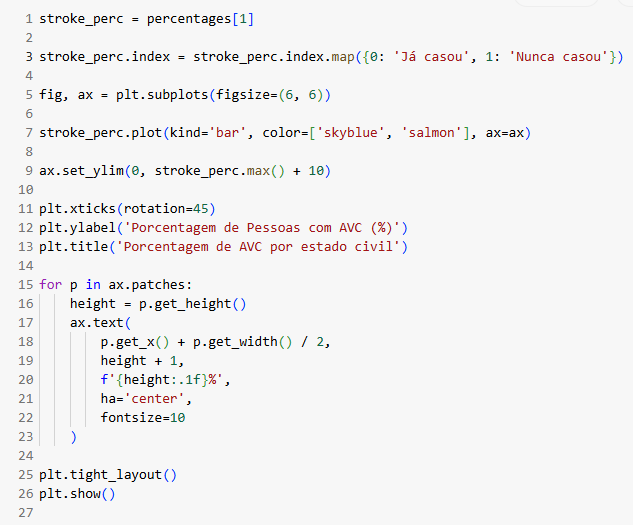
Fonte: Própria (2025)

##### Figura 12 - Criação do gráfico de casos de AVC por estado cívil



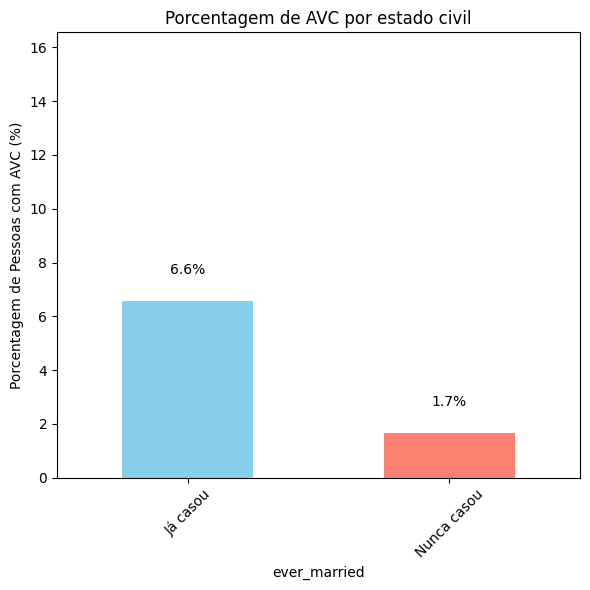
Fonte: Própria (2025)

##### Figura 13 - Criação do gráfico de casos de AVC por estado cívil



Fonte: Própria (2025)

##### Figura 14 - Ocorrência de AVC conforme estado civil



Fonte: Própria (2025)

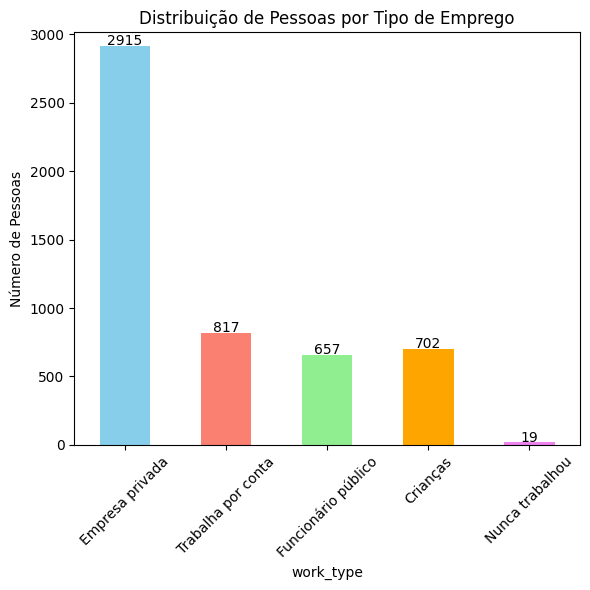
### 3.4.4 Distribuição de AVC conforme o tipo de trabalho

##### Figura 15 - Criação do gráfico distribuição de tipos de trabalho no dataset



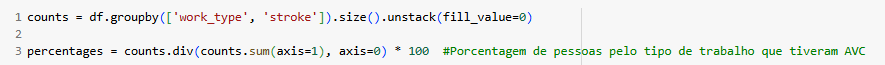
Fonte: Própria (2025)

##### Figura 16 - Gráfico de pessoas por tipo de trabalho



Fonte: Própria (2025)

##### Figura 17 - Criação do gráfico de casos de AVC em pessoas por tipo de trabalho



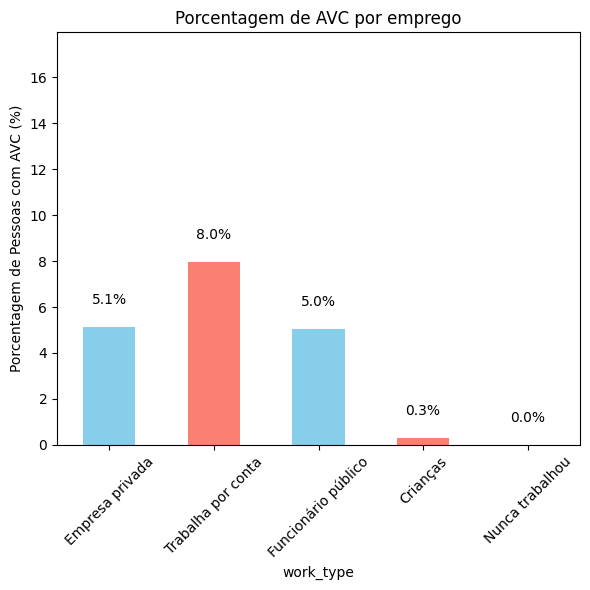
Fonte: Própria (2025)

##### Figura 18 - Criação do gráfico de casos de AVC por tipo de trabalho



Fonte: Própria (2025)

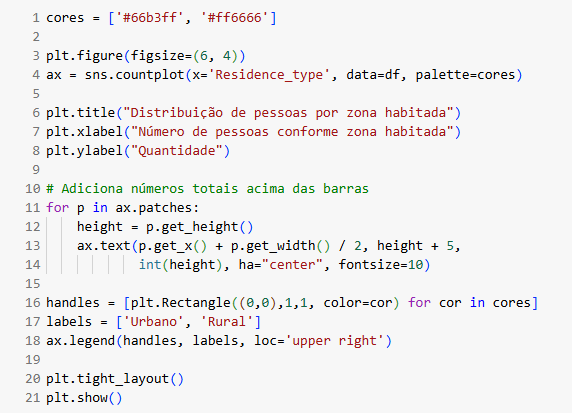
##### Figura 19 - Criação do gráfico de casos de AVC em pessoas por tipo de trabalho



Fonte: Própria (2025)

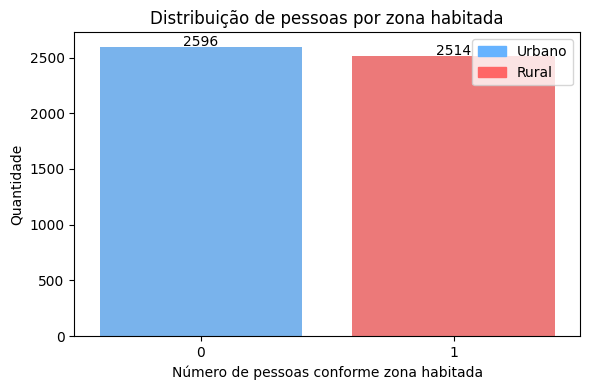
### 3.4.5 Distribuição de AVC conforme zona habitada

##### Figura 20 - Criação do gráfico de pessoas conforme zona habitada



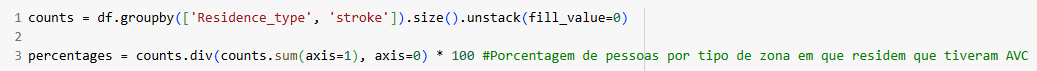
Fonte: Própria (2025)

##### Figura 21 - Gráfico de pessoas conforme zona habitada



Fonte: Própria (2025)

##### Figura 22 - Criação do gráfico de casos de AVC em pessoas por tipo de zona habitada



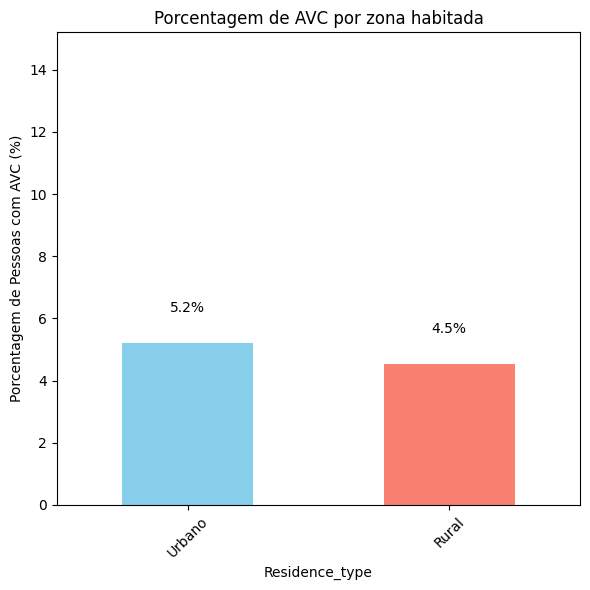
Fonte: Própria (2025)

##### Figura 23 - Criação do gráfico de casos de AVC em pessoas por tipo de zona habitada



Fonte: Própria (2025)

##### Figura 24 - Criação do gráfico de casos de AVC em pessoas por tipo de zona habitada



Fonte: Própria (2025)

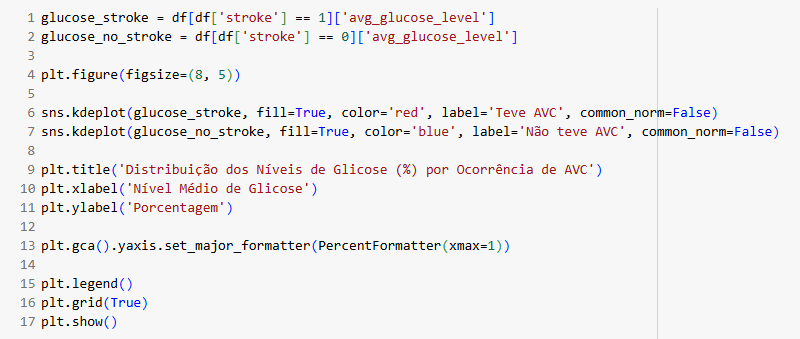
### 

### 

### 

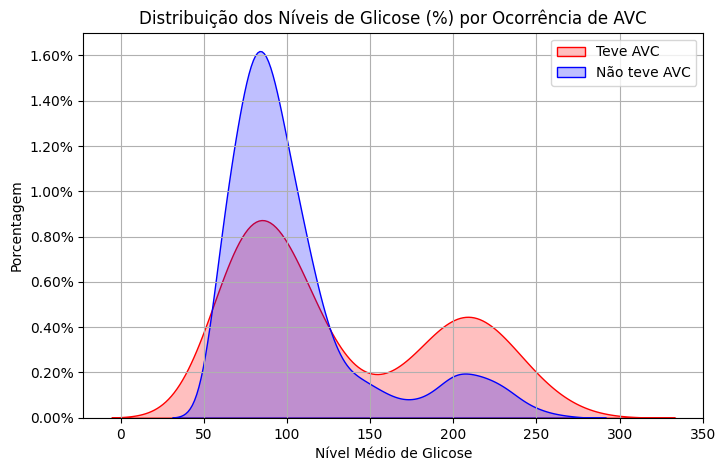
### 3.4.6 Distribuição de AVC conforme níveis de glicose

##### Figura 25 - Criação do gráfico de distribuição de AVC conforme os níveis de glicose



Fonte: Própria (2025)

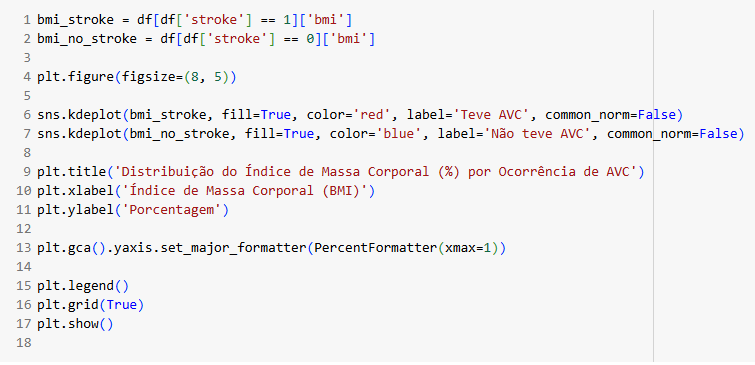
##### Figura 26 - Gráfico de distribuição dos níveis de glicose no dataset por ocorrência de AVC



Fonte: Própria (2025)

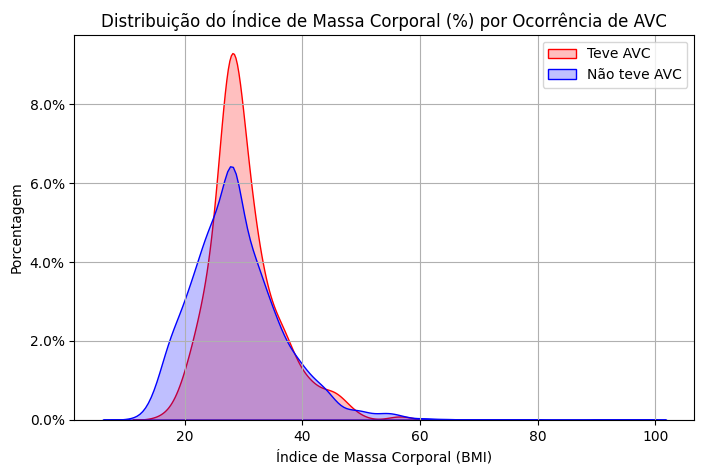
### 3.4.7 Distribuição de AVC conforme IMC

##### Figura 27 - Criação do gráfico de casos de AVC em pessoas conforme IMC



Fonte: Própria (2025)

##### Figura 28 - Gráfico de distribuição do AVC conforme IMC



Fonte: Própria (2025)

**3.4.8 Casos de AVC em fumantes**

No dataset os fumantes são classificados da seguinte forma:

• Já fumou

• Nunca fumou

• Fuma

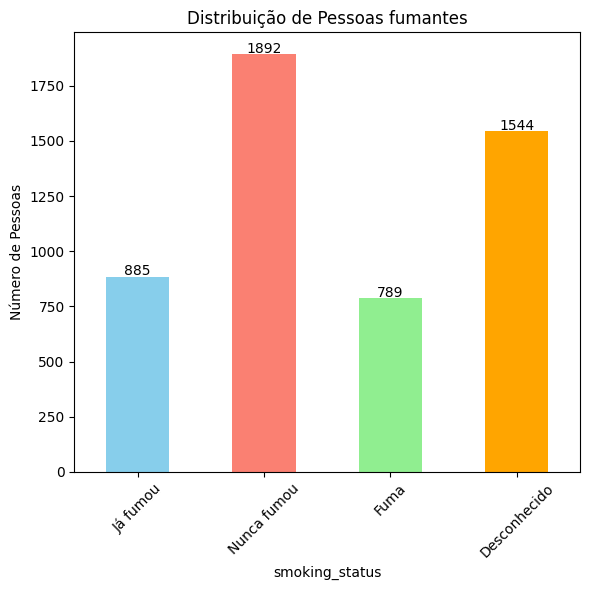
• Desconhecido

##### Figura 29 - Criação do gráfico de distribuição de fumantes no dataset



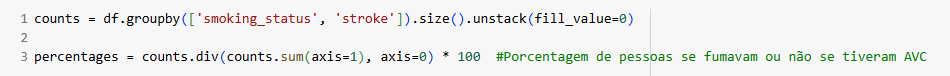
Fonte: Própria (2025)

##### Figura 30 - Gráfico de distribuição de fumantes no dataset



Fonte: Própria (2025)

##### Figura 31 - Criação do gráfico de casos de AVC em pessoas fumantes



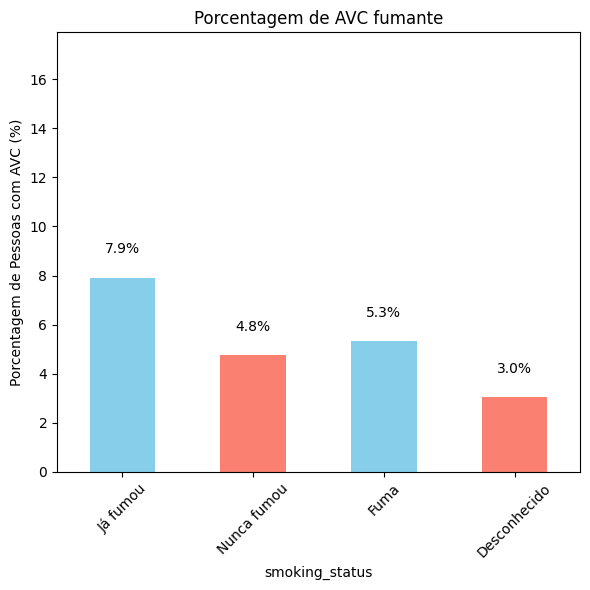
Fonte: Própria (2025)

##### Figura 32 - Criação do gráfico de casos de AVC em pessoas fumantes



Fonte: Própria (2025)

##### Figura 33 - Gráfico de AVC em fumantes



Fonte: Própria (2025)

## 

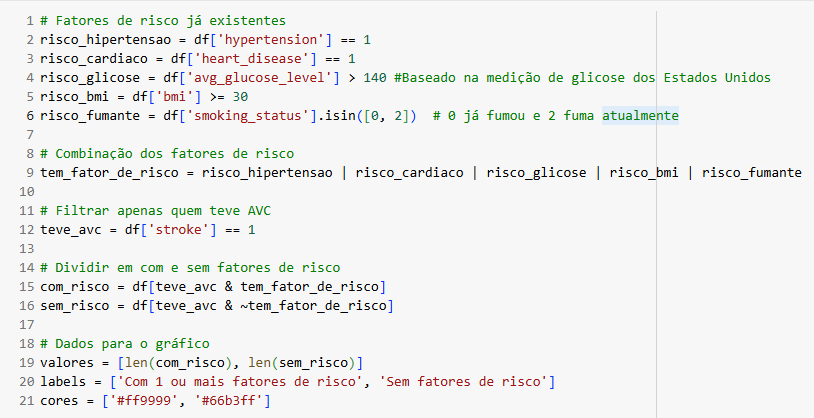
## 

## 

## 

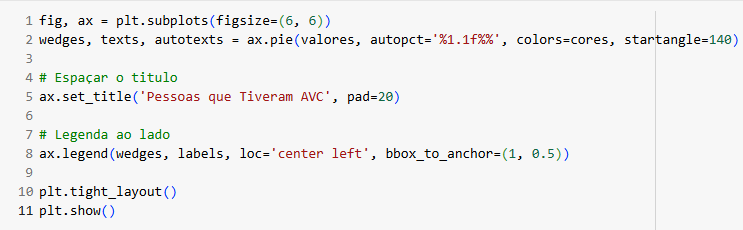
### 3.4.8 Distribuição de AVC por grupo de risco

##### Figura 34 - Criação do gráfico de fatia de pessoas com fator de risco que tiveram AVC ou não



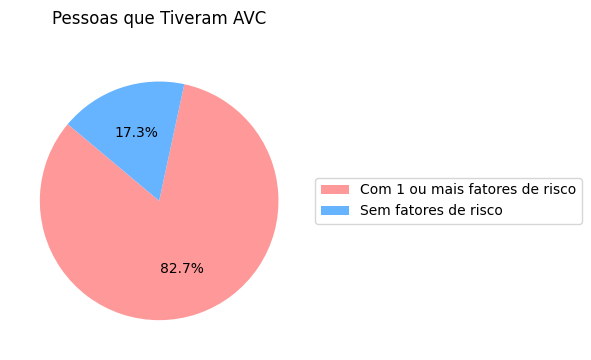
Fonte: Própria (2025)

##### Figura 35 - Criação do gráfico de fatia de pessoas com fator de risco que tiveram AVC ou não



Fonte: Própria (2025)

##### Figura 36 - Gráfico de casos de AVC em pessoas por fatores de risco



Fonte: Própria (2025)

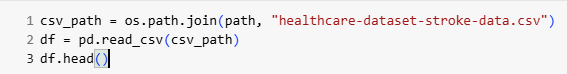
## 3.5 Pré Processamento dos dados

### 3.5.1 Verificação de valores nulos e análise preliminar

Foram verificados valores nulos, dados textuais e inconsistências nos dados do dataset.

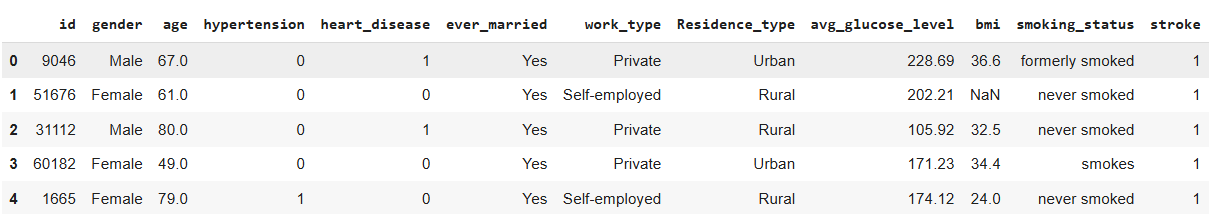
#### 3.5.1.1 Verificação do cabeçalho do dataset

##### Figura 37 - Ler os 5 primeiros valores das linhas das colunas



Fonte: Própria (2025)

##### Figura 38 - Valores das primeiras 5 linhas das colunas



Fonte: Própria (2025)

#### 3.5.1.2 Verificação de valores nulos dentro do dataset

##### Figura 39 - Verificação de valores nulos do dataset



Fonte: Própria (2025)

##### Figura 40 - Valores nulos do dataset



Fonte: Própria (2025)

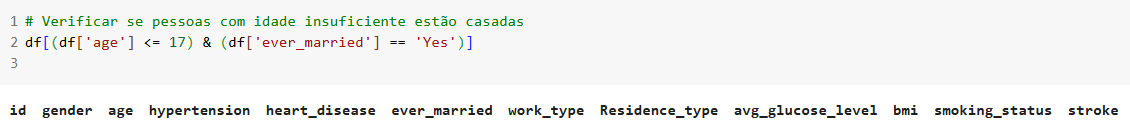
### 3.5.2 Tratamento de dados

Foi realizado um tratamento de dados para corrigir valores incompatíveis, substituir valores nulos por medianas e fazer a conversão de valores não numéricos em valores numéricos para a realização do treinamento.

#### 3.5.2.1 Tratamento de pessoas casadas com idade insuficiente

Verifica se pessoas que têm 17 anos ou menos estão classificadas como casadas.

##### Figura 41 - Verifica pessoas com idade insuficiente casadas



Fonte: Própria (2025)

#### 3.5.2.2 Tratamento de pessoas com idade insuficiente empregadas

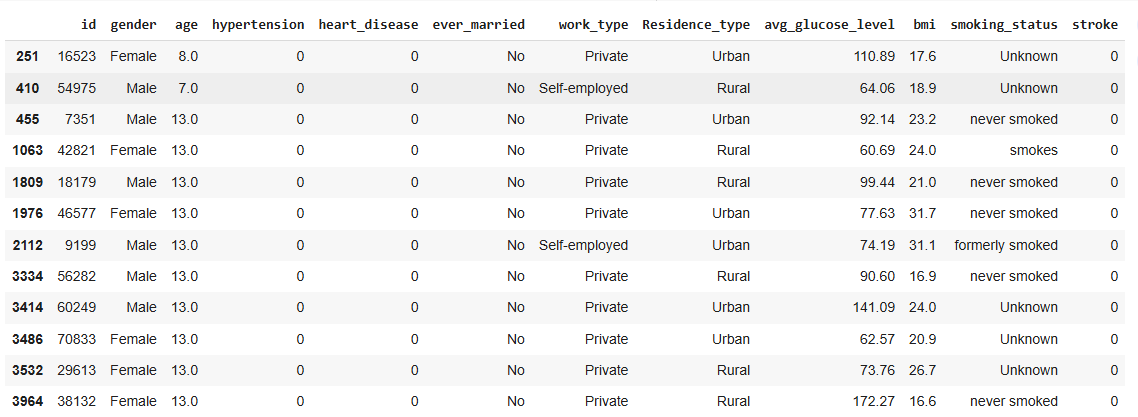
Busca por crianças da coluna work type que estão classificadas como se estivessem trabalhando e que tem 13 anos ou menos.

##### Figura 42 - Verifica pessoas com idade insuficiente para trabalhar



12 Pessoas com idade insuficiente para trabalhar estavam classificadas como empregadas.

##### Figura 43 - Verificação de Pessoas com idade insuficiente para trabalhar



Fonte: Própria (2025)

##### Figura 44 - Classifica pessoas com 13 anos ou menos como “criança”



Fonte: Própria(2025)

##### Figura 45 - Verifica se o tratamento ocorreu corretamente

Fonte: Própria (2025)

#### 3.5.2.3 Tratamento de valores nulos de IMC

Os valores nulos dos valores de IMC foram substituídos por uma mediana dos demais valores.

##### Figura 46 - Tratamento dos valores nulos IMC



Fonte: Própria (2025)

##### Figura 47 - Verifica se há algum valor nulo pendente



Fonte: Própria (2025)

##### Figura 48 - Verificar valores nulos



Fonte: Própria (2025)

#### 3.5.2.4 Tratamento de conversão para valores numéricos

Na coluna gender, o valor Male foi alterado para 0, e o valor Female alterado para 1.

##### Figura 49 - Converte Homem e Mulher



Fonte: Própria (2025)

Na coluna ever married, o valor de Yes foi alterado para 0, e No foi alterado para 1.

##### Figura 50 - Converte já foi casado Sim = 0 e Não = 1



Fonte: Própria (2025)

Converte o valor de todas as colunas que são inteiros em números em inteiros

##### Figura 51 - Converte os valores em números inteiros



Fonte: Própria (2025)

##### Figura 52 - Converte os valores em números inteiros



Fonte: Própria (2025)

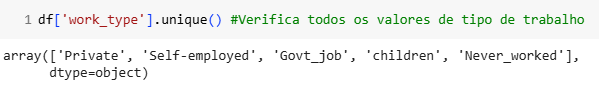
##### Figura 53 - Converte os valores em números inteiros



Fonte: Própria (2025)

Na coluna work type, Private foi alterado para 0, self-employed para 1, govt\_job para 2, children para 3 e never\_worked para 4

##### Figura 54 - Verifica todos os tipos de trabalhos do dataset



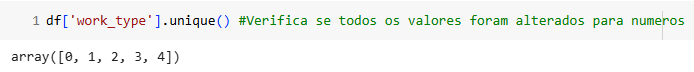
Fonte: Própria (2025)

##### Figura 55 - Converte os trabalhos para valores numéricos



Fonte: Própria (2025)

##### Figura 56 - Verifica se os valores foram devidamente tratados



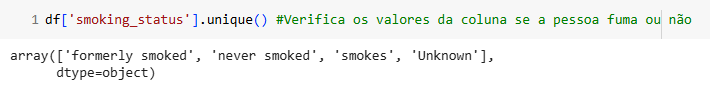
Fonte: Própria (2025)

##### Figura 57 - Converte os tipos de área de residência para valores numéricos



Fonte: Própria (2025)

##### Figura 58 - Verifica os valores da coluna de status de fumante



Fonte: Própria (2025)

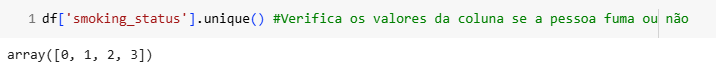
Na coluna smoking\_status, formerly smoked foi alterado para 0, never smoked para 1, smokes para 2 e Unknown para 3.

##### Figura 59 - Conversão dos valores de fumante para número



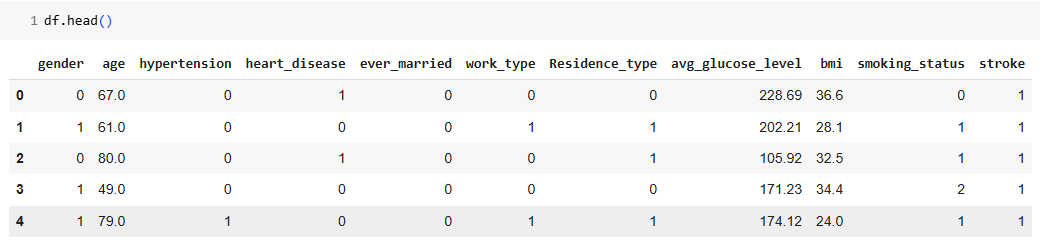
Fonte: Própria (2025)

##### Figura 60 - Verifica se os dados foram devidamente tratados



Fonte: Própria (2025)

##### Figura 61 - Verificação do cabeçalho



Fonte: Própria (2025)

# 4. DESENVOLVIMENTO DO MODELO DE DEEP LEARNING

## 4.1 Modelo utilizado

O modelo utilizado do algoritmo é de classificação binária, onde o algoritmo de Deep Learning deve classificar corretamente se a pessoa teve um AVC (1) ou não (0)

## 4.2 Treinamento do modelo

### 4.2.1 Colunas utilizadas no treinamento

Foram utilizadas as seguintes colunas para a realização do treinamento:

• gender

• age

• hypertension

• heart\_disease

• ever\_married

• work\_type

• Residence\_type

• avg\_glucose\_level

• bmi

• smoking\_status

Coluna alvo da previsão:

• stroke

A coluna ID foi desconsiderada para a realização do treinamento por ser irrelevante para o modelo.

##### Figura 62 - Seleção das colunas para treinamento e previsão



Fonte: Própria (2025)

##### Figura 63 - Seleção das colunas para treinamento e previsão



Fonte: Própria (2025)

### 4.2.2 Treinamento sem regularização

Foi realizado um treinamento com o modelo sem nenhum tipo de regularização para realizar um comparativo entre o desempenho de um modelo com regularização.

#### 4.2.2.1 Seleção do dado previsor

O dado previsor é stroke.

##### Figura 64 - Seleciona o dado previsor



Fonte: Própria (2025)

#### 4.2.2.2 Criação do treino e teste

Foi utilizado 70% dos dados para treino, 30% para teste e o random\_state foi definido para 42.

##### Figura 65 - Teste e Treino



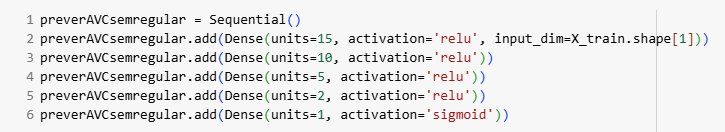
Fonte: Própria (2025)

#### 4.2.2.3 Camadas e neurônios

Foi utilizado uma camada com 15 neurônios com método de ativação relu,

uma camada com 10 neurônios relu, uma camada intermediária com 5 neurônios relu, uma camada com 2 neurônios relu e uma camada de saída com um neurônio de ativação sigmoide.

##### Figura 66 - Camadas e neurônios



Fonte: Própria (2025)

#### 4.2.2.4 Treino e teste

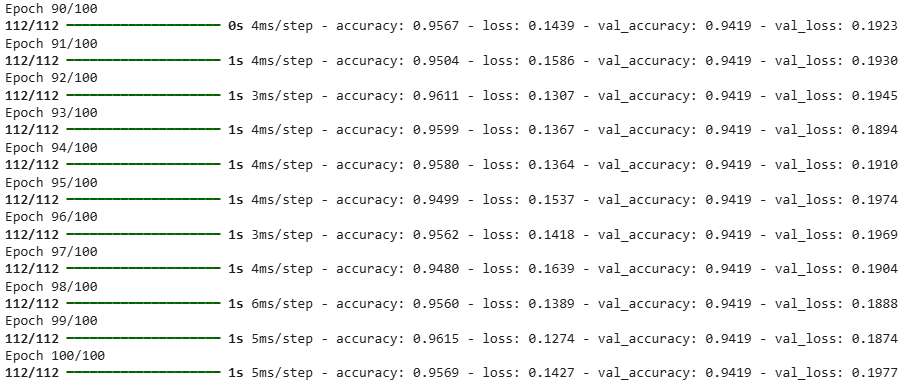
Foi realizado um treino e teste com 100 épocas, utilizando adam como otimizador, binary\_crossentropy como função de perda e acurácia como métrica.

##### Figura 67 - Configuração do treino e teste do algoritmo



Fonte: Própria (2025)

##### Figura 68 - Saída das 10 últimas épocas



Fonte: Própria (2025)

### 4.2.3 Treinamento com regularização

Foi realizado um treinamento com regularização, com a utilização de Standard Scaler, Class Weights (para ajustar automaticamente o peso das classes para lidar com o desbalanceamento de valores positivos do previsor), dropout para desligar aleatoriamente neurônios para otimizar o treinamento, e earlystopping para evitar overfitting no modelo.

#### 4.2.3.1 Seleção do dado previsor

Foi utilizado o mesmo dado previsor do primeiro treinamento

##### Figura 69 - Seleção do dado previsor



Fonte: Própria (2025)

#### 4.2.3.2 Criação do treino e teste

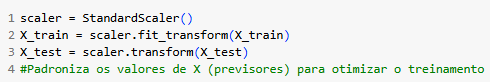
Foi utilizado 70% dos dados para treino, 30% para teste e o random\_state foi definido para 42, com a utilização de StandardScaler para padronizar os previsores para otimizar o treinamento.

##### Figura 70 - Teste e Treino



Fonte: Própria (2025)

##### Figura 71 - Padronizando os valores de X

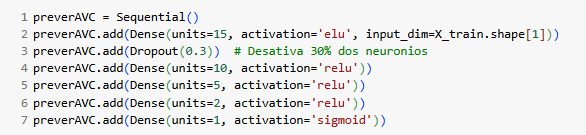


Fonte: Própria (2025)

#### 4.2.3.3 Camadas e neurônios

Foi utilizado uma camada com 15 neurônios com método de ativação elu e dropout de 30%, uma camada com 10 neurônios relu, uma camada intermediária com 5 neurônios relu, uma camada com 2 neurônios relu e uma camada de saída com um neurônio de ativação sigmoide.

##### Figura 72 - Neurônios e camadas

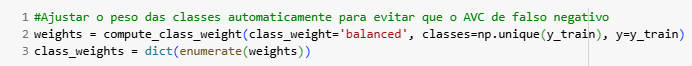


Fonte: Própria (2025)

#### 4.2.3.4 Ajuste de pesos das classes

Foi realizado um balanceamento de pesos das classes utilizando weights, que faz um balanceamento automático das classes para evitar o overfitting, fazendo com que o modelo leve mais em consideração os valores positivos (1) que estão em menor quantidade no dataset.

##### Figura 73 - Ajuste automático do peso das classes



Fonte: Própria (2025)

#### 4.2.3.5 Treino e teste

No treino e teste, foram utilizados, o otimizador adam, a função de perda binary\_crossentropy, a metrica acurácia, ajuste automatico de peso das classes EarlyStopping com o monitoramento de val\_loss do modelo com o limite de até 10 épocas com a restauração dos melhores pesos e 100 épocas.

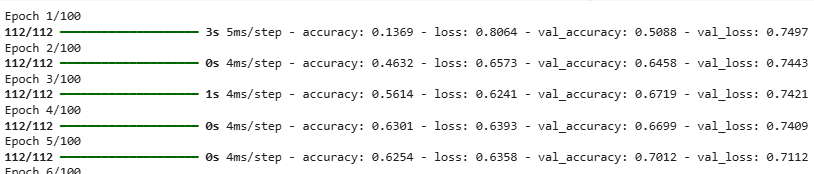
##### Figura 74 - Treino do modelo





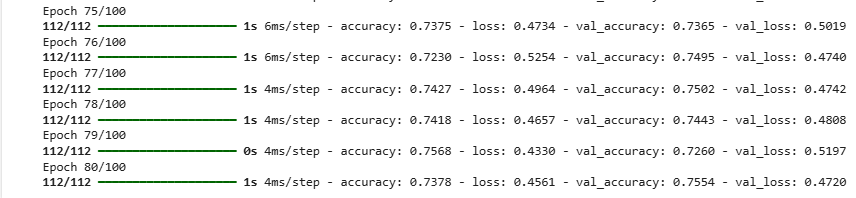
Fonte: Própria (2025)

##### Figura 75 - 5 primeiras épocas



Fonte: Própria (2025)

##### Figura 76 - 5 últimas épocas



Fonte: Própria (2025)

# 5. AVALIAÇÃO DO MODELO

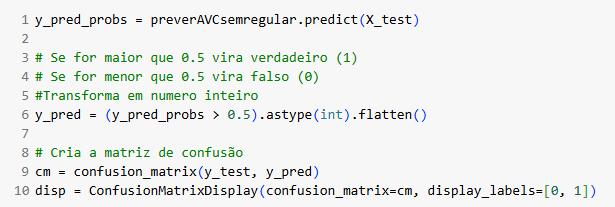
Foram utilizadas métricas de acurácia total, matriz de confusão, verdadeiro negativo e recall(Todos os valores positivos que o modelo previu corretamente).

## 5.1 Avaliação modelo não regularizado

Foi avaliado o desempenho do modelo não regularizado para comparar com o modelo regularizado.

### 5.1.1 Matriz de confusão

##### Figura 77 - Criação da Matriz de Confusão



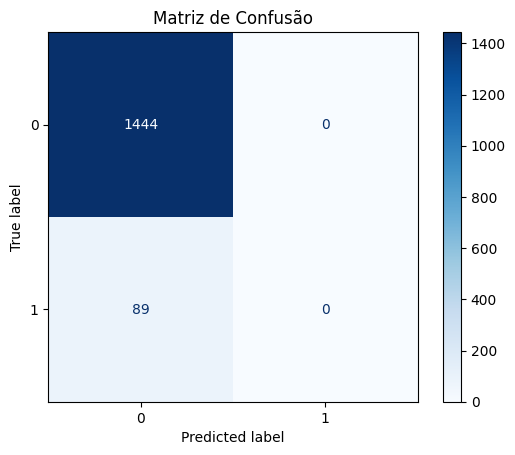
Fonte: Própria (2025)

##### Figura 78 - Plot Matriz de Confusão



Fonte: Própria (2025)

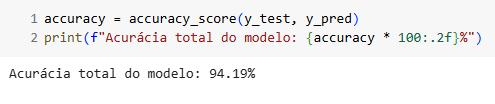
##### Figura 79 - Matriz de confusão



Fonte: Própria (2025)

### 5.1.2 Acurácia

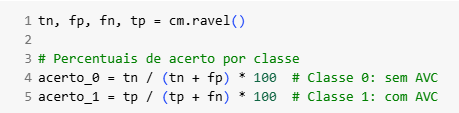
##### Figura 80 - Acurácia do modelo



Fonte: Própria (2025)

### 5.1.3 Recall e verdadeiro negativo

##### Figura 81 - Recall e verdadeiro negativo



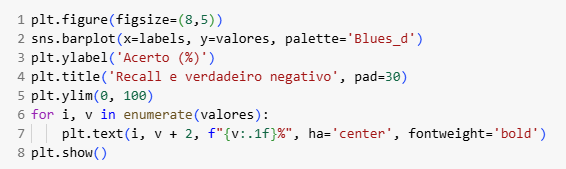
Fonte: Própria (2025)

##### Figura 82 - Recall e verdadeiro negativo



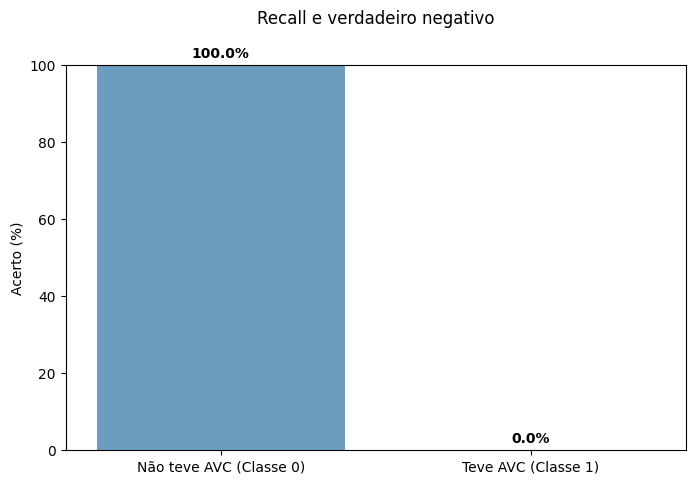
Fonte: Própria (2025)

##### Figura 83 - Plot recall e verdadeiro negativo



Fonte: Própria (2025)

##### Figura 84 - Recall e verdadeiro negativo



Fonte: Própria (2025)

### 5.1.4 Resultados, melhorias e ajustes

O recall do modelo foi de 0%, enquanto os acertos de verdadeiros negativos foram 100%, o modelo teve um overfitting, onde ele previu com precisão todos os valores negativos, e errou todos os valores positivos, a acurácia de 94,17% é decorrente de overfitting.

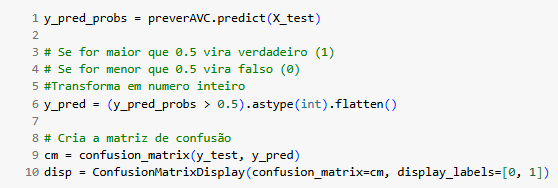
Foi realizado um novo treinamento utilizando regularização para melhorar o aprendizado do algoritmo.

## 5.2 Avaliação modelo regularizado

Foram utilizadas as mesmas métricas para avaliar o desempenho do modelo, acurácia, matriz de confusão, verdadeiro negativo, recall(Todos os valores positivos que o modelo previu corretamente) e os acertos que o modelo teve de pessoas que tinham ou não fator de risco, para avaliar se o modelo está associando corretamente os fatores de risco a uma maior probabilidade de ter um AVC.

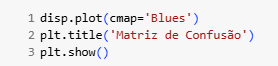
### 5.2.1 Matriz de confusão

##### Figura 85 - Criação da matriz de confusão



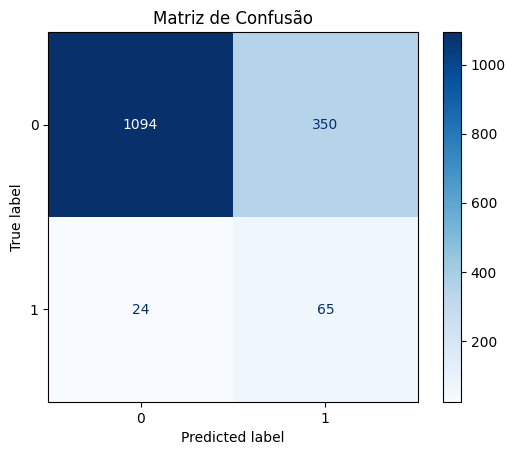
Fonte: Própria (2025)

##### Figura 86 - Plot matriz de confusão



Fonte: Própria (2025)

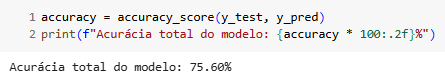
##### Figura 87 - Matriz de confusão



Fonte: Própria (2025)

### 5.2.2 Acurácia

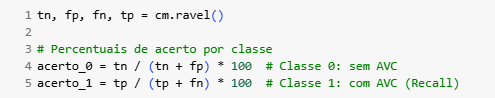
##### Figura 88 - Acurácia total do modelo



Fonte: Própria (2025)

### 5.2.3 Recall e verdadeiro negativo

##### Figura 89 - Criação do verdadeiro negativo e recall



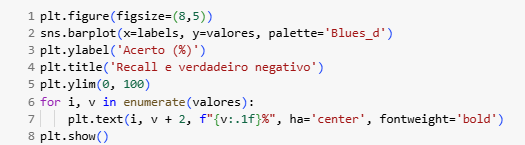
Fonte: Própria (2025)

##### Figura 90 - Criação do verdadeiro negativo e recall



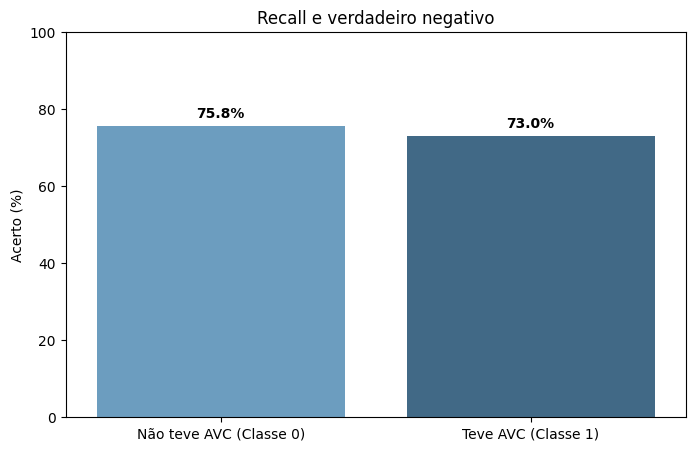
Fonte: Própria (2025)

##### Figura 91 - Criação do plot



Fonte: Própria (2025)

##### Figura 92 - Recall e verdadeiro negativo



Fonte: Própria (2025)

### 5.2.4 Falsos negativos e positivos

##### Figura 93 - Criação das classes



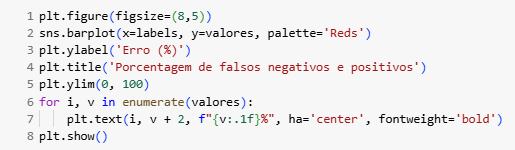
Fonte: Própria (2025)

##### Figura 94 - Criação das classes



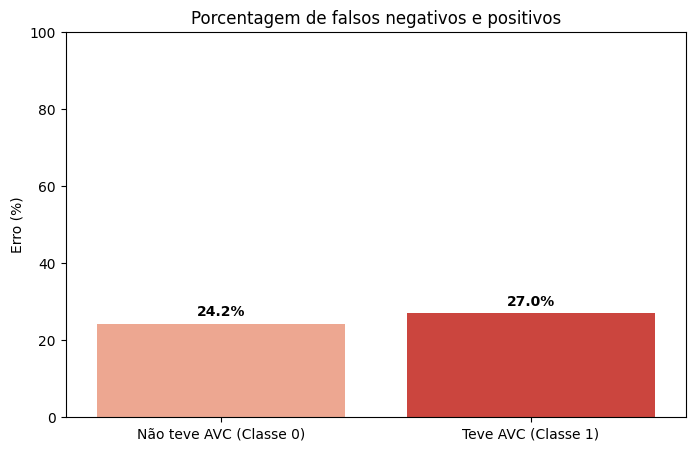
Fonte: Própria (2025)

##### Figura 95 - Criação do plot



Fonte: Própria (2025)

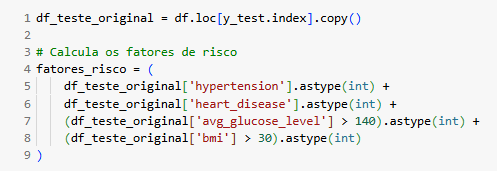
##### Figura 96 - Falsos negativos e positivos



Fonte: Própria (2025)

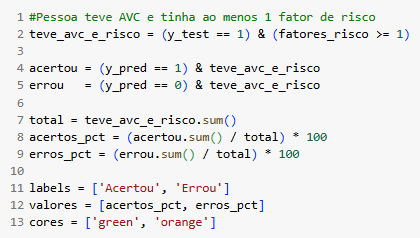
### 5.2.5 Recall fatores de risco

##### Figura 97 - Definição dos fatores de risco



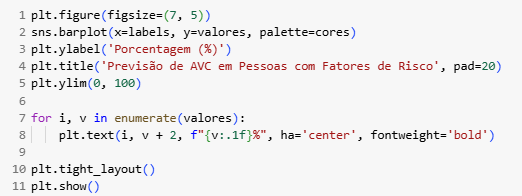
Fonte: Própria (2025)

##### Figura 98 - Criação do plot



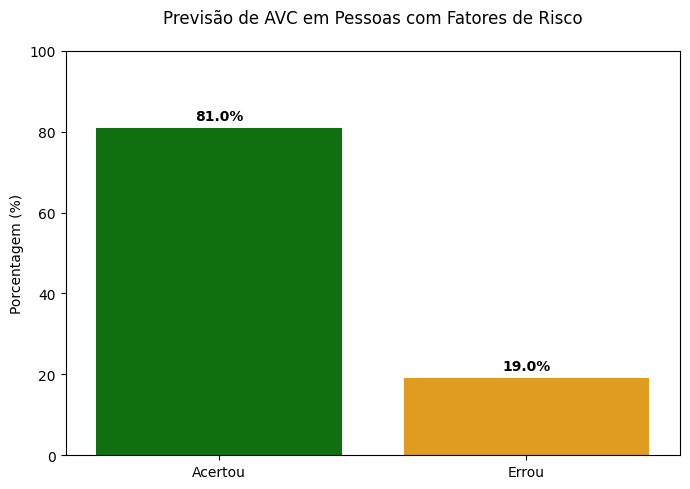
Fonte: Própria (2025)

##### Figura 99 - Criação do plot



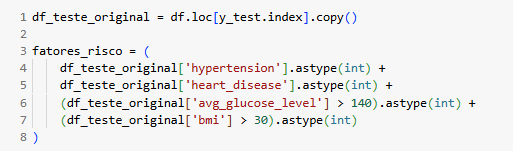
Fonte: Própria (2025)

##### Figura 100 - Previsão de pessoas com fatores de risco



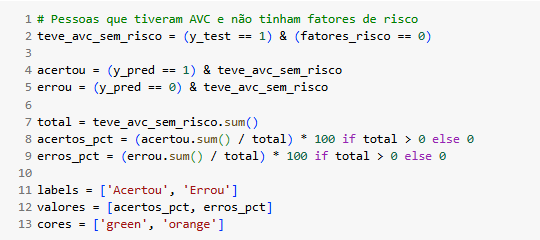
### 5.2.6 Recall de pessoas que tiveram AVC e que não tinham fator de risco

##### Figura 101 - Definição dos fatores de risco



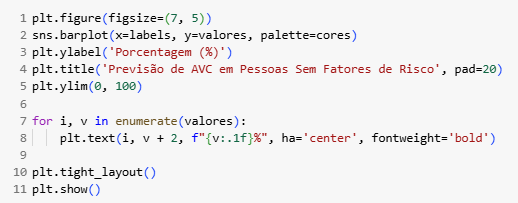
Fonte: Própria (2025)

##### Figura 102 - Criação do plot



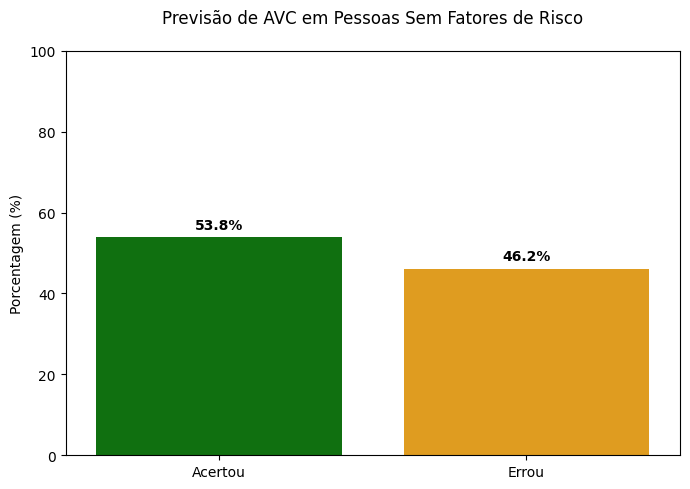
Fonte: Própria (2025)

##### Figura 103 - Criação do plot



Fonte: Própria (2025)

##### Figura 104 - Previsão de AVC em pessoas que não tinham nenhum fator de risco



Fonte: Própria (2025)

# 6. CONCLUSÃO

Foi possível realizar uma pesquisa prévia sobre o AVC e criar um modelo de Deep Learning de classificação binária com a capacidade de identificar fatores de risco e associar a ocorrência de AVC dentro do dataset, corrigir o overfitting do primeiro modelo e aprimorar a previsão do segundo modelo com regularização.

Houve um recall de 73% no modelo, onde desses 73% o algoritmo classificou corretamente 81% das pessoas que tiveram AVC e tinham ao menos um fator de risco.

# REFERÊNCIAS

BRASIL. Governo Federal. Ministério da Saúde. **Acidente Vascular Cerebral**. Disponível em: https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/a/avc. Acesso em: 14 jun. 2025.

MARAMÉLIA MIRANDA (Brasil). Sociedade Brasileira de Avc (ed.). **Números do AVC**. 2025. Disponível em: https://avc.org.br/numeros-do-avc/. Acesso em: 14 jun. 2025.

BRASIL. Governo Federal. Ministério da Saúde (ed.). **Acidente vascular cerebral (AVC)**. 2015. Disponível em: https://bvsms.saude.gov.br/avc-acidente-vascular-cerebral/. Acesso em: 14 jun. 2025.

PALACIOS, Federico Soriano. **Stroke Prediction Dataset**. 2021. Disponível em: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data. Acesso em: 30 maio 2025.