Classificador de Bayes Aplicado em Dados de Doenças Cardíacas

Gabriel Ferreira
Gabriel Soares
José Janailson
Pedro Rodrigues

Apresentação

- **A base de dados escolhida foi a ‘Heart Disease Data Set’ disponível na plataforma UCI**
- **Ele era composto de vários atributos dos quais 13 eram usados para prever o atributo ‘num’ que expunha o diagnostico de uma anomalia no coração.**
- **A base consistia em 4 bancos de dados coletados de forma separada por instituições diferentes**

UCI Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Check out the beta version of the new UCI Machine Learning Repository we are currently testi
Click here to try out the new site.

Heart Disease Data Set

Download: Data Folder, Data Set Description

Abstract: 4 databases: Cleveland, Hungary, Switzerland, and the VA Long Beach

Data Set Characteristics:	Multivariate	**Number of Instances:**	303	**Area:**	Life
Attribute Characteristics:	Categorical, Integer, Real	**Number of Attributes:**	75	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	2177261

Apresentação

- **Os quatro bancos de dados foram nomeados:**
  - 1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
  - 2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
  - 3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
  - 4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Dados

- **Apesar da base consistir em 76 campos apenas 14 deles eram mais consistentes e recomendados pelos autores para serem usados. Alguns deles eram:**
  - Age: idade
  - Cp: tipo de dor no peito
  - Exang : angina induzida por exercício
  - oldpeak : depressão induzida por exercício relativa ao descanso
  - thalach: máxima taxa cardíaca atingida
  - Num: valores de 0 a 4 que identifica a presença de doença cardíaca. Num=0 significa um coração saudável

Dados

- **Analisamos os dados presentes apenas na ‘hungarian.data’ e ‘switzerland.data’ e comparamos a resposta do classificador a essas duas BD**
- **Para a base dados ‘hungarian.data’ e ‘switzerland.data’ usamos os mesmos métodos de tratamento de dados e treinamento do classificador.**
- **A seguir descreveremos apenas o ‘hungarian.data’**

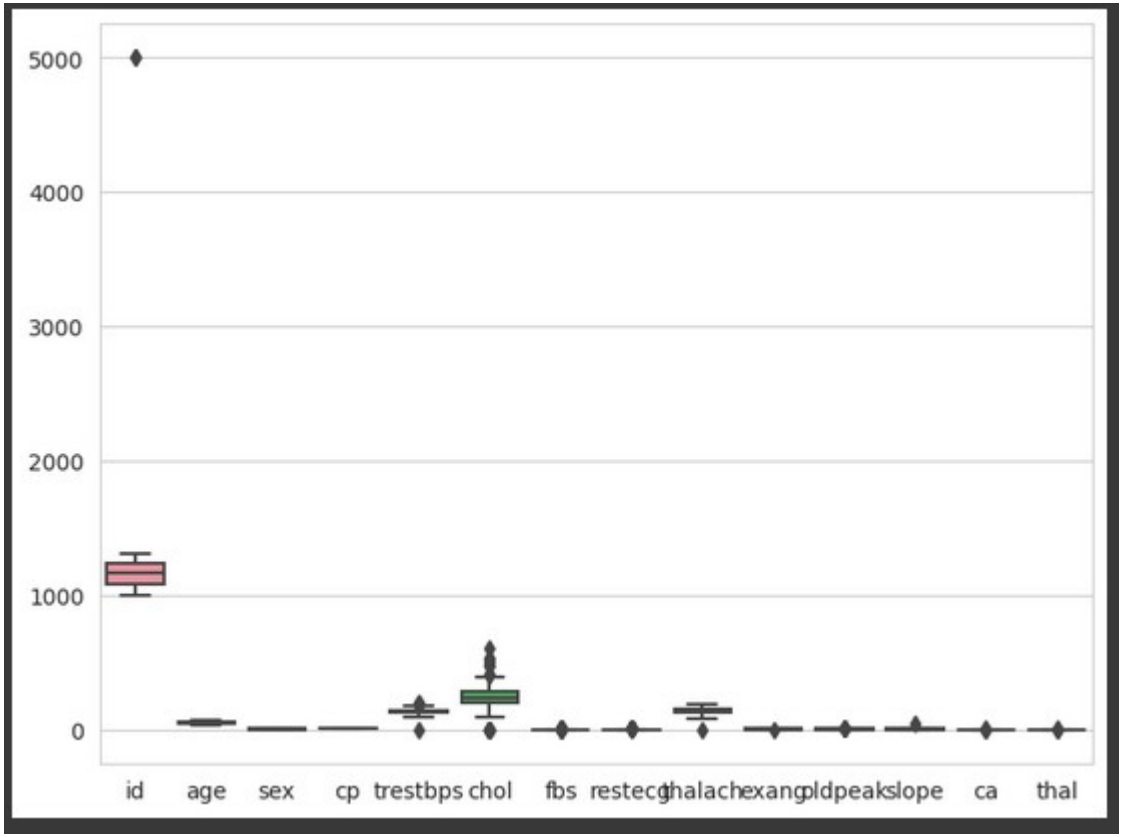
Dados antes de serem processados

	id	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	1254	40	1	2	140	289	0	0	172	0	0.0	41	-9	-9	0
1	1255	49	0	3	160	180	0	0	156	0	1.0	-9	-9	-9	1
2	1256	37	1	2	130	283	0	1	98	0	0.0	2	-9	-9	0
3	1257	48	0	4	138	214	0	0	108	1	1.5	-9	-9	-9	3
4	1258	54	1	3	150	-9	0	0	122	0	0.0	2	-9	-9	0
5	1259	39	1	3	120	339	0	0	170	0	0.0	-9	-9	-9	0
6	1260	45	0	2	130	237	0	0	170	0	0.0	-9	-9	-9	0
7	1261	54	1	2	110	208	0	0	142	0	0.0	-9	-9	-9	0
8	1262	37	1	4	140	207	0	0	130	1	1.5	-9	-9	-9	1
9	1263	48	0	2	120	284	0	0	120	0	0.0	2	-9	-9	0
10	1264	37	0	3	130	211	0	0	142	0	0.0	-9	-9	-9	0

Dados antes de serem processados

	id	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
count	294.000000	294.000000	294.000000	294.000000	294.000000	294.000000	294.000000	294.000000	294.00000	294.000000	294.000000	294.000000	294.000000	294.000000	294.000000
mean	1195.853741	47.826531	0.724490	2.982993	132.102041	230.520408	-0.176871	0.187075	138.62585	0.272109	0.586054	-4.976190	-8.846939	-8.846939	0.792517
std	397.340367	7.811812	0.447533	0.965117	19.437564	95.414336	1.499491	0.707616	25.08408	0.711273	0.908648	5.869698	1.382623	1.382623	1.237006
min	1001.000000	28.000000	0.000000	1.000000	-9.000000	-9.000000	-9.000000	-9.000000	-9.00000	-9.000000	0.000000	-9.000000	-9.000000	-9.000000	0.000000
25%	1080.250000	42.000000	0.000000	2.000000	120.000000	198.000000	0.000000	0.000000	122.00000	0.000000	0.000000	-9.000000	-9.000000	-9.000000	0.000000
50%	1158.500000	49.000000	1.000000	3.000000	130.000000	237.000000	0.000000	0.000000	140.00000	0.000000	0.000000	-9.000000	-9.000000	-9.000000	0.000000
75%	1235.750000	54.000000	1.000000	4.000000	140.000000	277.000000	0.000000	0.000000	155.00000	1.000000	1.000000	2.000000	-9.000000	-9.000000	1.000000
max	5002.000000	66.000000	1.000000	4.000000	200.000000	603.000000	1.000000	2.000000	190.00000	1.000000	5.000000	41.000000	9.000000	9.000000	4.000000

Dados antes de serem processados

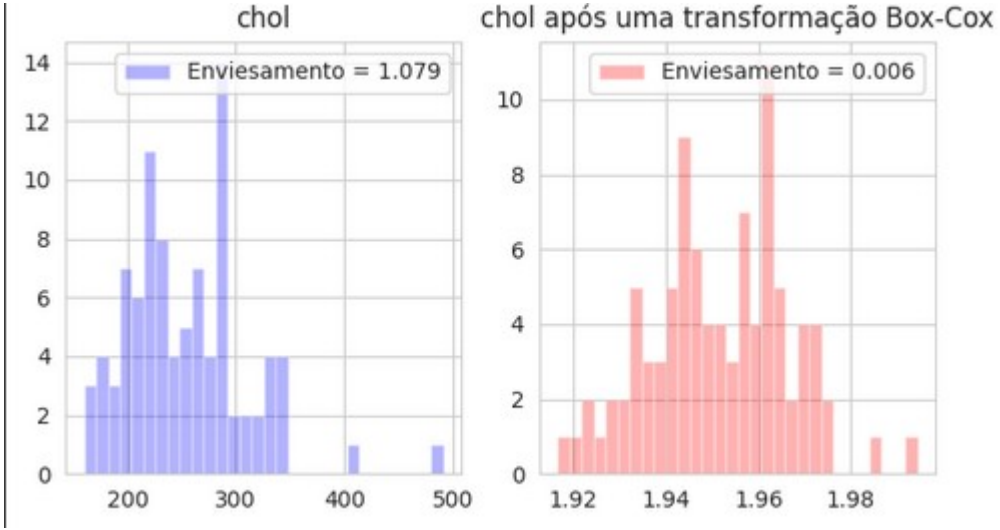


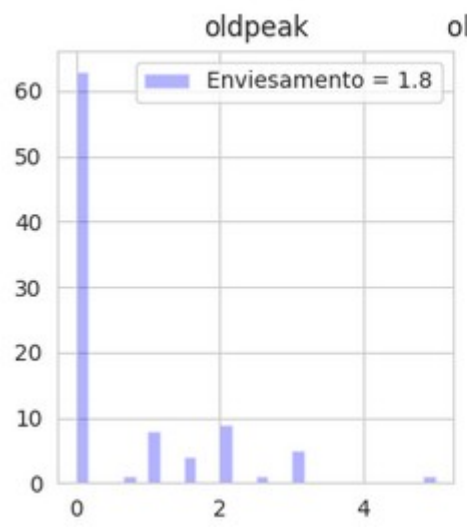

id 1.00 -0.01 -0.03 -0.00 -0.04 -0.01 0.02 -0.02 0.02 -0.03 -0.03 -0.01 -0.01 -0.01 -0.07
age -0.01 1.00 0.01 0.15 0.22 0.04 0.05 -0.01 -0.43 0.15 0.18 -0.00 0.04 0.04 0.21
sex -0.03 0.01 1.00 0.25 0.12 0.03 0.09 -0.10 -0.07 0.17 0.12 0.01 -0.03 -0.03 0.22
cp -0.00 0.15 0.25 1.00 0.10 0.12 0.00 0.08 -0.33 0.36 0.35 -0.06 0.07 0.07 0.43
trestbps -0.04 0.22 0.12 0.10 1.00 0.02 0.06 -0.04 -0.01 0.45 0.15 0.03 0.06 0.06 0.21
chol -0.01 0.04 0.03 0.12 0.02 1.00 -0.11 0.03 -0.09 0.07 0.13 0.03 -0.07 -0.07 0.22
fbs 0.02 0.05 0.09 0.00 0.06 -0.11 1.00 -0.01 -0.00 0.02 0.05 -0.01 0.01 0.01 0.10
restecg -0.02 -0.01 -0.10 0.08 -0.04 0.03 -0.01 1.00 -0.01 -0.01 0.06 -0.01 -0.03 -0.03 -0.05
thalach 0.02 -0.43 -0.07 -0.33 -0.01 -0.09 -0.00 -0.01 1.00 0.02 -0.31 0.01 -0.03 -0.03 -0.33
exang -0.03 0.15 0.17 0.36 0.45 0.07 0.02 -0.01 0.02 1.00 0.34 0.03 0.05 0.05 0.40
oldpeak -0.03 0.18 0.12 0.35 0.15 0.13 0.05 0.06 -0.31 0.34 1.00 -0.01 0.01 0.01 0.58
slope -0.01 -0.00 0.01 -0.06 0.03 0.03 -0.01 -0.01 0.01 0.03 -0.01 1.00 -0.08 -0.08 0.02
ca -0.01 0.04 -0.03 0.07 0.06 -0.07 0.01 -0.03 -0.03 0.05 0.01 -0.08 1.00 1.00 -0.05
thal -0.01 0.04 -0.03 0.07 0.06 -0.07 0.01 -0.03 -0.03 0.05 0.01 -0.08 1.00 1.00 -0.05
-0.07 0.21 0.22 0.43 0.21 0.22 0.10 -0.05 -0.33 0.40 0.58 0.02 -0.05 -0.05 1.00
id age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal
0.8
0.6
0.4
0.2
0.0
−0.2

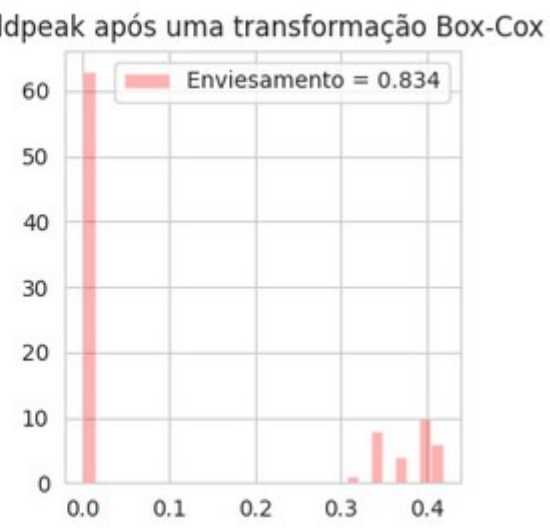
Tratamento dos dados

- **Devido a inconsistência dos dados tivemos de fazer um forte tratamento nos dados:**
- **1) Identificar e eliminar outliers**
- **2) Diminuir o enviesamento dos dados**
- **3) Eliminar dados nulos ou ausentes**
- **Apos esse tratamento nossa base diminuiu consideravelmente de 294 observações para apenas 94.**
- **Além disso duas colunas,thal e cal, foram abandonadas por conterem em sua maioria dados ausentes**

Dados tratados



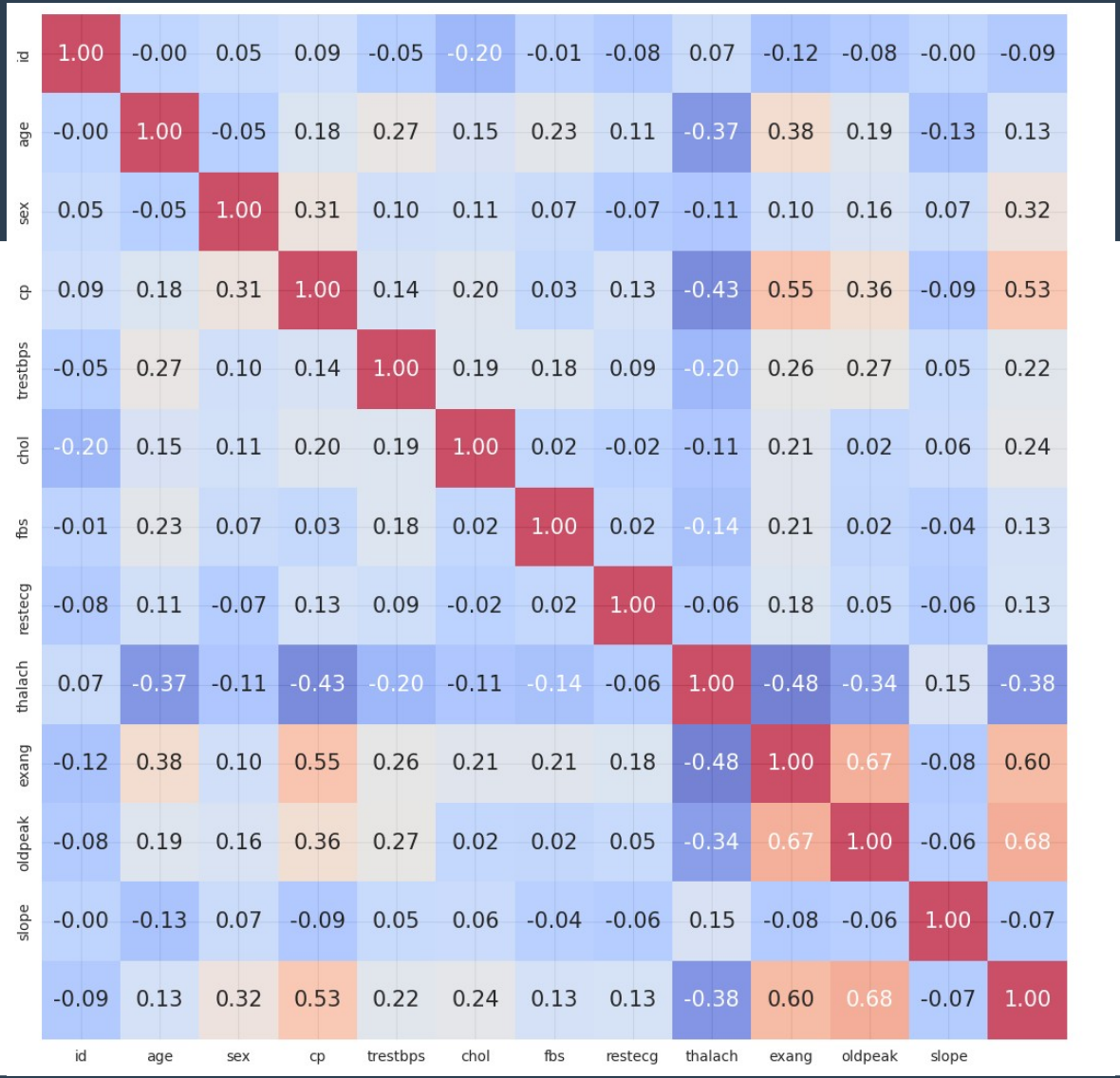




	id	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	num
0	1254	40	1	2	140	289	0	0	172	0	0.0	41	0
1	1256	37	1	2	130	283	0	1	98	0	0.0	2	0
2	1263	48	0	2	120	284	0	0	120	0	0.0	2	0
3	1266	39	1	2	120	204	0	0	145	0	0.0	2	0
4	1268	42	0	3	115	211	0	1	137	0	0.0	2	0
5	1270	38	1	4	110	196	0	0	166	0	0.0	2	1
6	1273	36	1	2	120	267	0	0	160	0	3.0	2	1
7	1274	43	0	1	100	223	0	0	142	0	0.0	2	0
8	1276	49	0	2	124	201	0	0	164	0	0.0	2	0
9	1278	40	1	3	130	215	0	0	138	0	0.0	2	0
10	1281	52	1	2	120	284	0	0	118	0	0.0	2	0
11	1287	41	1	4	130	172	0	1	130	0	2.0	2	3
12	1288	43	0	2	150	186	0	0	154	0	0.0	2	0
13	1291	41	0	2	110	250	0	1	142	0	0.0	2	0
14	1294	54	0	2	150	230	0	0	130	0	0.0	2	0

Matriz de correlações

- **Após o tratamento dos dados verificamos que eles eram relativamente independentes entre si o que é um requisito do classificador de Bayes**
- **Porem como complemento observamos quais atributos tinham uma maior correlação com o ‘num’, variável que estávamos interessados**


id 1.00 -0.00 0.05 0.09 -0.05 -0.20 -0.01 -0.08 0.07 -0.12 -0.08 -0.00 -0.09
age -0.00 1.00 -0.05 0.18 0.27 0.15 0.23 0.11 -0.37 0.38 0.19 -0.13 0.13
sex 0.05 -0.05 1.00 0.31 0.10 0.11 0.07 -0.07 -0.11 0.10 0.16 0.07 0.32
cp 0.09 0.18 0.31 1.00 0.14 0.20 0.03 0.13 -0.43 0.55 0.36 -0.09 0.53
trestbps -0.05 0.27 0.10 0.14 1.00 0.19 0.18 0.09 -0.20 0.26 0.27 0.05 0.22
chol -0.20 0.15 0.11 0.20 0.19 1.00 0.02 -0.02 -0.11 0.21 0.02 0.06 0.24
fbs -0.01 0.23 0.07 0.03 0.18 0.02 1.00 0.02 -0.14 0.21 0.02 -0.04 0.13
restecg -0.08 0.11 -0.07 0.13 0.09 -0.02 0.02 1.00 -0.06 0.18 0.05 -0.06 0.13
thalach 0.07 -0.37 -0.11 -0.43 -0.20 -0.11 -0.14 -0.06 1.00 -0.48 -0.34 0.15 -0.38
exang -0.12 0.38 0.10 0.55 0.26 0.21 0.21 0.18 -0.48 1.00 0.67 -0.08 0.60
oldpeak -0.08 0.19 0.16 0.36 0.27 0.02 0.02 0.05 -0.34 0.67 1.00 -0.06 0.68
slope -0.00 -0.13 0.07 -0.09 0.05 0.06 -0.04 -0.06 0.15 -0.08 -0.06 1.00 -0.07
-0.09 0.13 0.32 0.53 0.22 0.24 0.13 0.13 -0.38 0.60 0.68 -0.07 1.00
id age sex cp trestbps chol fbs restecg thalach exang oldpeak slope

Correlação entre atributos

- **Atributos com maior correlação com diagnostico positivo:**
  - Cp: tipo de dor no peito
  - Exang : angina induzida por exercício
  - oldpeak : depressão induzida por exercício relativa ao descanso
- **Atributos com maior anti correlação com diagnostico positivo:**
  - thalach: máxima taxa cardíaca atingida

Treinamento do classificador

- **Treinamos com o modelo gaussiano do classificador de bayes que assume que a afinidade entre os atributos segue uma distribuição gaussiana**
- **Originalmente usamos 20% dos dados para testes e 80% para treinamento.**
- **Depois usamos 95% de para treinamento e 5% para testes para obter uma maior acurácia**
- **Por fim trocamos a base de dados de hungarian.data para switzerland.data**

resultados

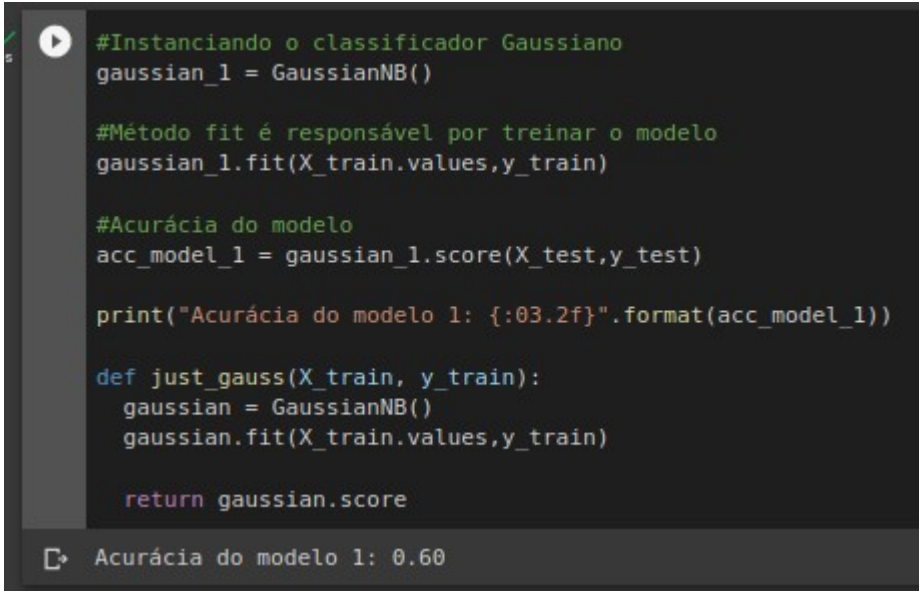
- **Primeiro usamos 20% para testes e 80% para treinamento.**
- **Obtemos para esta base de dados uma acurácia de 0.42**

```
#Instanciando o classificador Gaussiano
gaussian_1 = GaussianNB()

#Método fit é responsável por treinar o modelo
gaussian_1.fit(X_train.values,y_train)

#Acurácia do modelo
acc_model_1 = gaussian_1.score(X_test,y_test)

print("Acurácia do modelo 1: {:03.2f}".format(acc_model_1))

def just_gauss(X_train, y_train):
  gaussian = GaussianNB()
  gaussian.fit(X_train.values,y_train)

  return gaussian.score
```

```
Acurácia do modelo 1: 0.42
```

resultados

- **Depois aumentamos o tamanho dos dados de treinamento para 95% e diminuímos os de testes para 5% e obtemos uma acurácia de 0.60**



```
#Instanciando o classificador Gaussiano
gaussian_1 = GaussianNB()

#Método fit é responsável por treinar o modelo
gaussian_1.fit(X_train.values,y_train)

#Acurácia do modelo
acc_model_1 = gaussian_1.score(X_test,y_test)

print("Acurácia do modelo 1: {:03.2f}".format(acc_model_1))

def just_gauss(X_train, y_train):
  gaussian = GaussianNB()
  gaussian.fit(X_train.values,y_train)

  return gaussian.score
```

```
Acurácia do modelo 1: 0.60
```

resultados

- **Depois trocamos a base de dados para switzerland.data que passou pelos mesmos métodos e obtivemos uma acurácia de apenas 0.20**
- **Isso se deve por ela ter mais atributos ausentes, menos dados e de ter mais uma coluna com valores nulos.**

```
#Instanciando o classificador Gaussiano
gaussian_1 = GaussianNB()

#Método fit é responsável por treinar o modelo
gaussian_1.fit(X_train.values,y_train)

#Acurácia do modelo
acc_model_1 = gaussian_1.score(X_test,y_test)

print("Acurácia do modelo 1: {:03.2f}".format(acc_model_1))

def just_gauss(X_train, y_train):
  gaussian = GaussianNB()
  gaussian.fit(X_train.values,y_train)

  return gaussian.score
```

```
Acurácia do modelo 1: 0.20
```