

FATEC Rubens Lara
Ciência de Dados
Matemática Básica

Entropia de Dados

Enri Lopes Iwasaki
Leandro Costa Santos

Santos
2023

A base de dados utilizada foi obtida por meio do **IGBE - Instituto Brasileiro de Geografia e Estatística**. A base apresenta dados referentes a **Síntese de Indicadores Sociais**, que analisa a qualidade de vida e os níveis de bem-estar das pessoas, famílias e grupos populacionais, a efetivação de direitos humanos e sociais, bem como o acesso a diferentes serviços, bens e oportunidades, por meio de indicadores que visam contemplar a heterogeneidade da sociedade brasileira sob a perspectiva das desigualdades sociais.

Os dados verificados abordam, mais precisamente, a **Total e respectiva distribuição percentual das pessoas, por classes de rendimento no ano de 2022**. Estando disponível em: https://www.ibge.gov.br/estatisticas/sociais/trabalho/9221sintese_de_indicadores_sociais.html?=&t=resultados.

Tabela 2.3 - Total e respectiva distribuição percentual das pessoas, por classes de rendimento domiciliar per capita, segundo as Grandes Regiões e as Unidades da Federação - Brasil - 2021

Grandes Regiões e Unidades da Federação	Pessoas								
	Total (1 000 pessoas)	Distribuição percentual, por classes de rendimento domiciliar per capita (salário mínimo) (%)							
		Sem rendimento	Mais de zero até ¼	Mais de ¼ até ½	Mais de ½ até 1	Mais de 1 a 2	Mais de 2 a 3	Mais de 3 a 5	Mais de 5
Brasil	212 577	2,0	12,6	19,8	29,1	22,6	6,5	4,2	3,3

População Total do Brasil		
212.577.000		
Rendimento em Salários Recebidos	População	Porcentagem (%)
0	4.251.540	2
0 – 1	130.522.278	61,4
1 – 2	48.042.402	22,6
2 – 3	13.817.505	6,5
3 – 5	8.928.234	4,2
> 5	7.015.041	3,3

$$Classes = \{ '0' , '0 - 1' , '1 - 2' , '2 - 3' , '3 - 5' , '> 5' \}$$

$$| Classes | = 6$$

$$H(X) = - \sum_{x \in classes} P(x) \log_2 (P(x))$$

$$H = - (0,02 \log_2^{0,02} + 0,614 \log_2^{0,614} + 0,226 \log_2^{0,226} + 0,065 \log_2^{0,065} + 0,042 \log_2^{0,042} + 0,033 \log_2^{0,033})$$

$$\log_2^{0,02} = \frac{\log 0,02}{\log 2} = -5,64 \quad \log_2 0,614 = \frac{\log 0,614}{\log 2} = -0,70$$

$$\log_2^{0,226} = \frac{\log 0,226}{\log 2} = -2,15 \quad \log_2^{0,065} = \frac{\log 0,065}{\log 2} = -3,94$$

$$\log_2^{0,042} = \frac{\log 0,042}{\log 2} = -4,57 \quad \log_2^{0,033} = \frac{\log 0,033}{\log 2} = -4,92$$

$$H = -(0,02 \cdot (-5,64) + 0,614 \cdot (-0,70) + 0,226 \cdot (-2,15) + 0,065 \cdot (-3,94) + 0,042 \cdot (-4,57) + 0,033 \cdot (-4,92))$$

$H=1,64$

Entropia Máxima dos Dados

$$\text{Max}_H = \log_2^{|classes|} = \log_2^6 = \frac{\log 6}{\log 2}$$

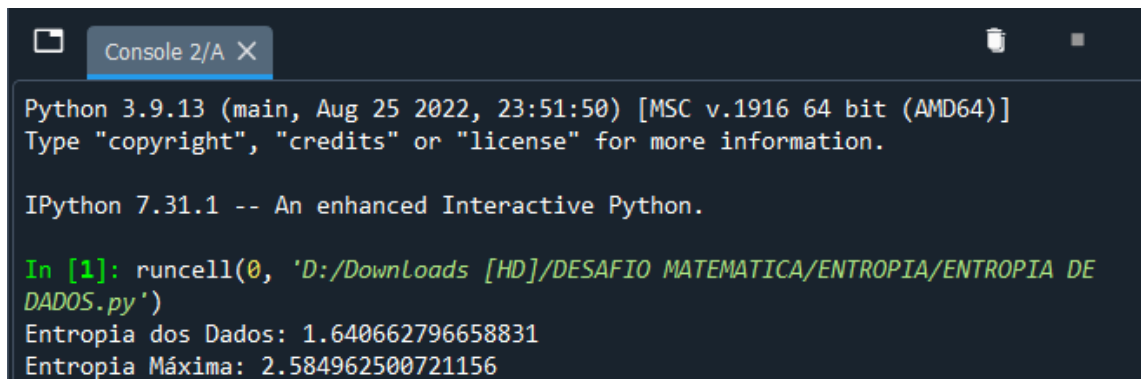
$\text{Max}_H = 2,58$

Conclusão: os dados apresentados não possuem alto grau de aleatoriedade. Visto que o valor da entropia dos dados está distante do valor de entropia máxima, pode-se então concluir que os dados não se distribuem uniformemente.

Programação do método em Python:

```
1  import math
2
3  def calcular_entropia(probabilidades):
4      entropia = 0
5      for probabilidade in probabilidades:
6          if probabilidade > 0:
7              entropia -= probabilidade * math.log2(probabilidade)
8      return entropia
9
10 probabilidades = [0.02, 0.614, 0.226, 0.065, 0.042, 0.033]
11
12 entropia = calcular_entropia(probabilidades)
13 print("Entropia dos Dados:", entropia)
14
15 n = len(probabilidades)
16 entropia_maxima = math.log2(n)
17 print("Entropia Máxima:", entropia_maxima)
```

Demonstração do Console:



The screenshot shows a console window titled "Console 2/A X". The text inside the console is as follows:

```
Python 3.9.13 (main, Aug 25 2022, 23:51:50) [MSC v.1916 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 7.31.1 -- An enhanced Interactive Python.

In [1]: runcell(0, 'D:/Downloads [HD]/DESAFIO MATEMATICA/ENTROPIA/ENTROPIA DE
DADOS.py')
Entropia dos Dados: 1.640662796658831
Entropia Máxima: 2.584962500721156
```