Atividade 08

Eduardo Satiro da Cruz
PPGMMC
CEFET-MG
Belo Horizonte, Brasil
eduardo.satiro@gmail.com

***Abstract*—Esse trabalho informa passos que foi feito para utilização do algoritmo *K Neighbors Classifier* no conjuntos de dados *Airlines*.**

***Index Terms*—conjunto de dados, classificação, KMeans, GMM.**

I. Introdução

Na disciplina de Inteligência Computacional, ministrada pelo dr. Alisson Marques da Silva no CEFET-MG, foi solicitado uma atividade que consiste em escolher um conjunto de dados e realizar experimentos computacionais comparando dois ou mais algoritmos de agrupamento. O *dataset* escolhido foi através do Kaggle [1]. Para os procedimentos foi utilizado a linguagem *Python* com as bibliotecas: *pandas, numpy, seaborn, matplotlib, sklearn*.

II. Conjunto de Dados: *Mall Customer Segmentation Data*

O conjunto de dados é o *Mall Customer Segmentation Data* [2] e foi criado apenas para fins de aprendizado dos conceitos de segmentação de clientes, também conhecidos como análise de compras [2]. A descrição do problema é um dono do shopping que deseja entender os clientes para que possa ser dado à equipe de marketing e planejar a estratégia.

Inicialmente foi criado um *DataFrame* da bibioteca *pandas* para a manipulação dos dados. A função utilizada foi *read_csv* que carrega o arquivo baixado com extensão CSV para o *DataFrame* criado.

A Fig. 1 apresenta o *DataFrame*. Existe cinco atributos sendo *CustomerID* é o identificador do cliente, *Gender* é o gênero do cliente, *Age*é a idade, *Annual Income (k$)* é a renda anual do cliente, *Spending Score (1-100)* é a pontuação atribuída pelo shopping com base no comportamento do cliente e na natureza do gasto.

A Fig. 2 apresenta a descrição estatística dos dados, isso foi retorno da função *describe* da biblioteca *pandas*. Observamos a quantidade de atributos, desvio padrão, a média e entre outras informações para analisar os dados.

A Fig. 3 apresenta informações sobre o *DataFrame*, incluindo o tipo do dado e colunas, valores não nulos e uso de memória.

A Fig. 4 demonstra como esta a distribuição dos clientes entre masculino e feminino. Há um equilíbrio entre os dois valores.

```
df_mall
```

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

200 rows × 5 columns

Fig. 1. *DataFrame*

```
df_mall.describe()
```

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

Fig. 2. Análise estatística do conjunto de dados

A Fig. 5 demonstra como esta a distribuição da idade dos clientes. Os valores estão entre 18 a 70 anos e observamos que há mais clientes com idade por volta de 30 anos.

A Fig. 6 demonstra como esta a distribuição de renda dos clientes. Muitos clientes se encontram na faixa de 60 a 90.

A Fig. 7 demonstra como esta a distribuição de pontuação dos clientes.

A Fig. 8 demonstra a relação entre o atributo idade e

```
df_mall.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual Income (k$)      200 non-null    int64
 4   Spending Score (1-100)  200 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

Fig. 3. Informações sobre o *DataFrame*

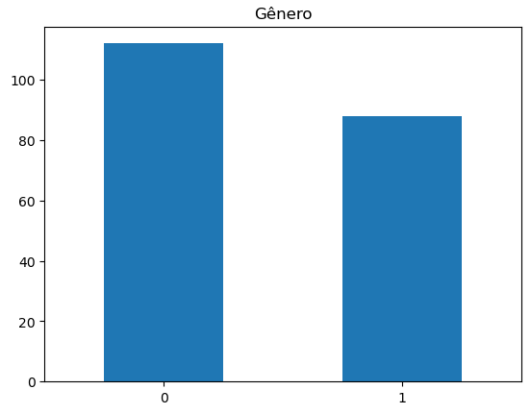

Fig. 4. Gráfico do atributo Gênero

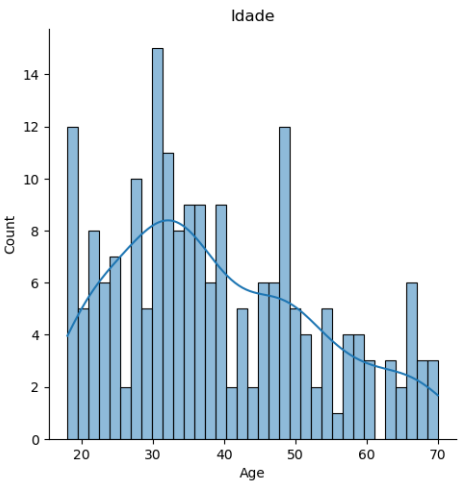

Fig. 5. Gráfico do atributo Idade

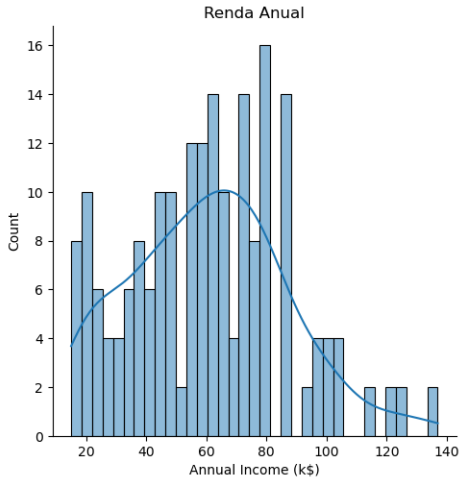

Fig. 6. Gráfico do atributo Renda anual

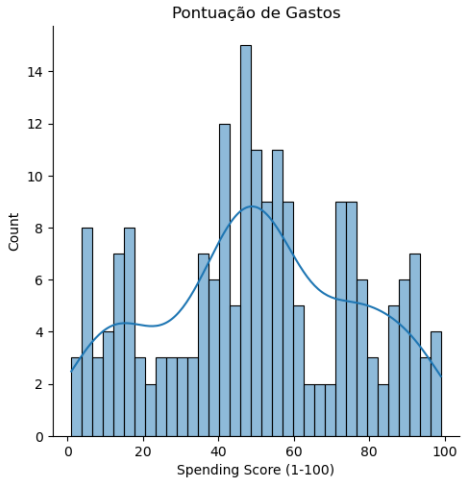

Fig. 7. Gráfico do atributo Pontuação

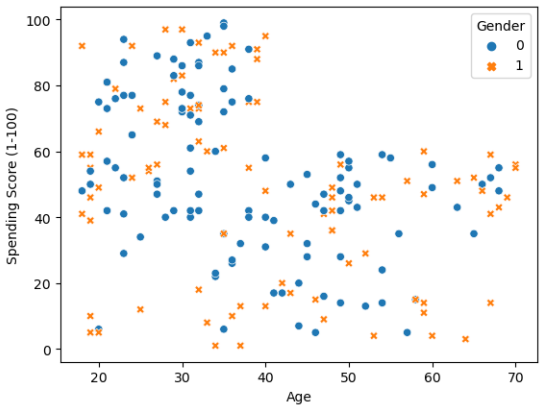

Fig. 8. Gráfico Idade x Pontuação

pontuação. Há poucos clientes com alta pontuação e idade avançada.

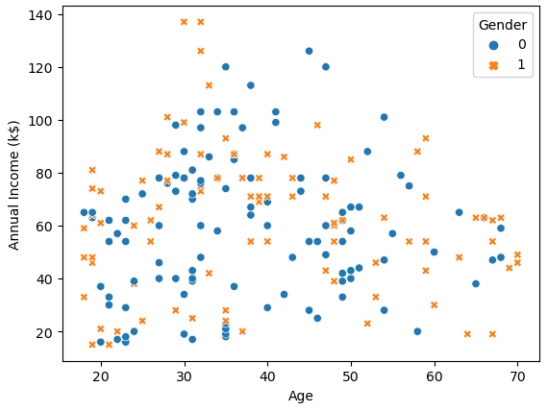

Fig. 9. Gráfico Renda x Idade

A Fig. 9 demonstra a relação entre o atributo idade e renda. Há mais clientes na parte inferior do gráfico.

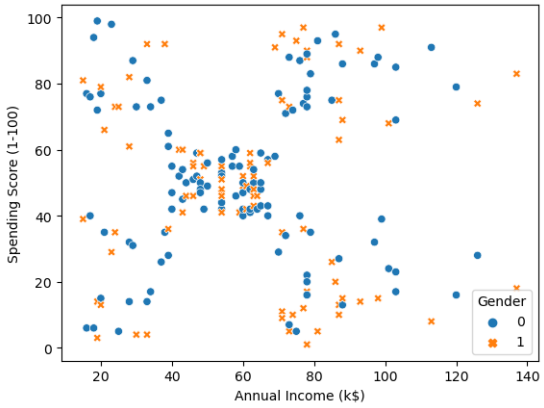

Fig. 10. Gráfico Renda x Pontuação

A Fig. 10 demonstra a relação entre o atributo pontuação e renda. Podemos identificar 5 grupos nesse gráfico.

III. Metodologia

A. Airlines

Para esse experimento, foi utilizado inicialmente o método do cotovelo. O método experimenta vários cluester e cria um gráfico com a variância dos dados em relação ao número de clusters.

A Fig. 11 apresenta o gráfico do método do cotovelo, nele obtemos o número de cluster. O cluster ideal é 5.

Para os experimentos foi utilizado dois métodos, o *K-Means* e o *Gaussian Mixture*. Para os experimentos foi utilizado os parâmetros *default* de cada método.

O *K-Means* é um algoritmo agrupa os dados tentando separar as amostras em n grupos de variância igual, minimizando um critério conhecido como inércia ou soma dos quadrados dentro do *cluster*. Este algoritmo requer que o

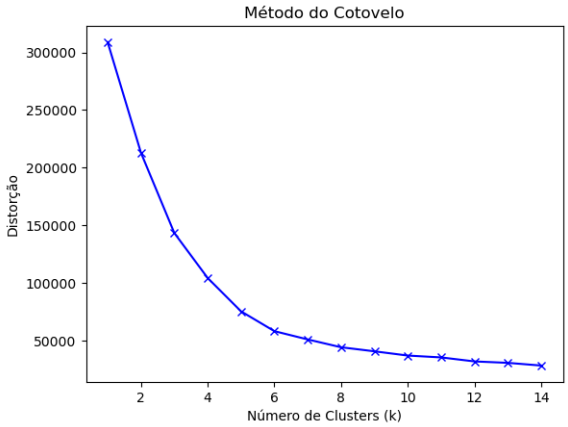

Fig. 11. Método Cotovelo

número de *clusters* seja especificado. Ele se adapta bem a um grande número de amostras e tem sido usado em uma grande variedade de áreas de aplicação em muitos campos diferentes. [5].

A *Gaussian Mixture* é um modelo de mistura gaussiana é um modelo probabilístico que assume que todos os pontos de dados são gerados a partir de uma mistura de um número finito de distribuições gaussianas com parâmetros desconhecidos. Pode-se pensar em modelos de mistura como generalizando o agrupamento *k-means* para incorporar informações sobre a estrutura de covariância dos dados, bem como os centros dos gaussianos latentes [4].

IV. Experimentos

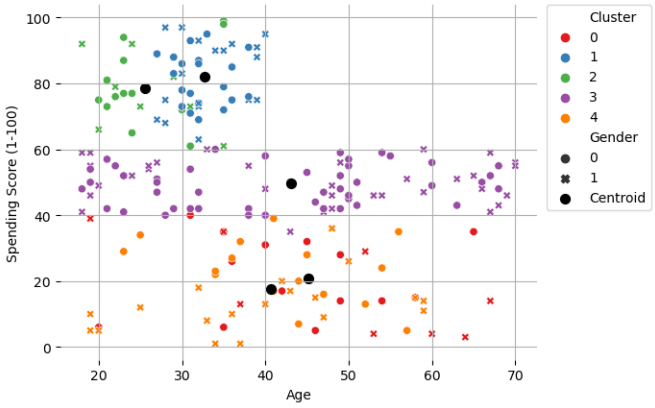

Fig. 12. Gráfico Idade x Pontuação do modelo K-Means

A Fig. 12 apresenta a distribuição entre idade e pontuação.
A Fig. 13 apresenta a distribuição entre idade e renda.
A Fig. 14 apresenta a distribuição entre pontuação e renda.
A Fig. 12 apresenta a distribuição entre idade e pontuação.
A Fig. 13 apresenta a distribuição entre idade e renda.
A Fig. 14 apresenta a distribuição entre pontuação e renda.

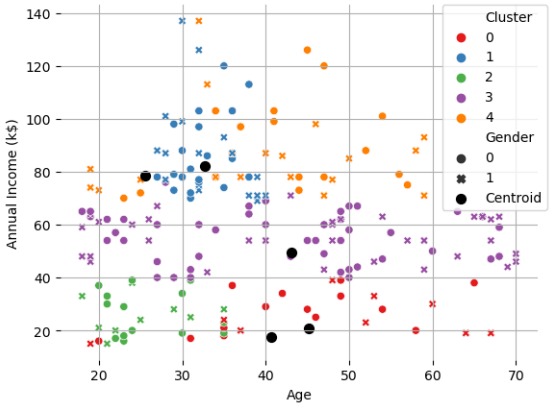

Fig. 13. Gráfico Idade x Renda do modelo K-Means

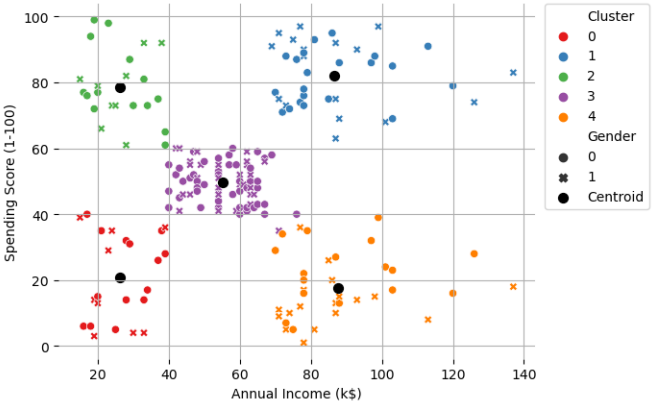

Fig. 14. Gráfico Renda x Pontuação do modelo K-Means

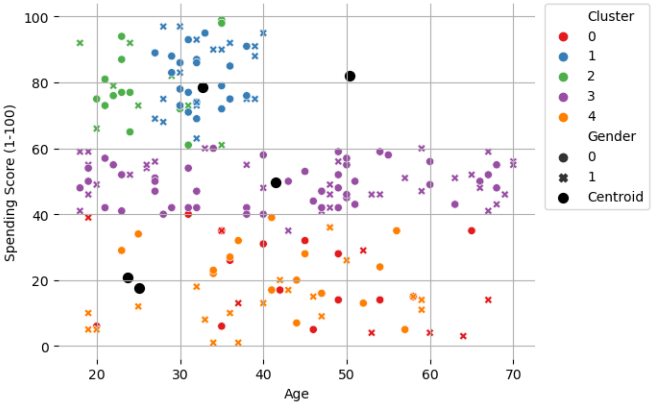

Fig. 15. Gráfico Idade x Pontuação do modelo Gaussian Mixture

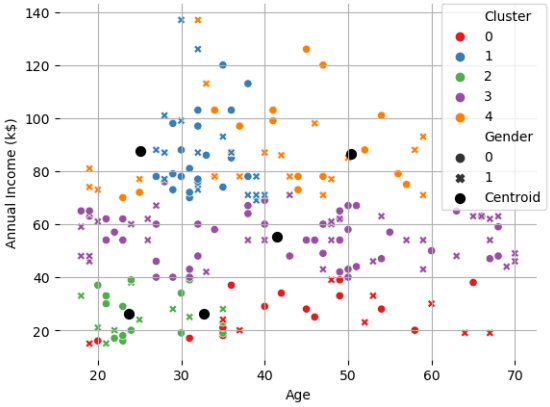

Fig. 16. Gráfico Idade x Renda do modelo Gaussian Mixture

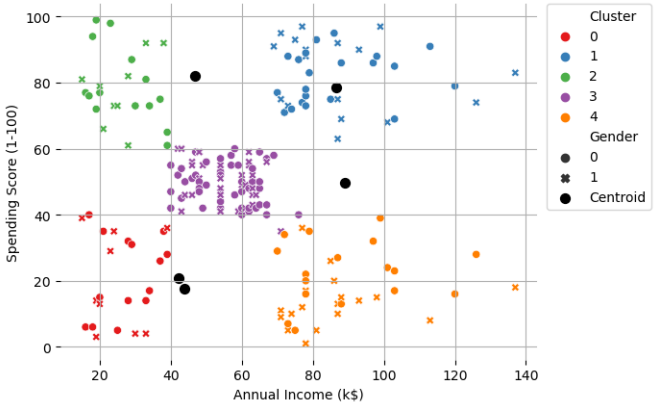

Fig. 17. Gráfico Renda x Pontuação do modelo Gaussian Mixture

REFERENCES

[1] ”Kaggle” https://www.kaggle.com (accessed Jun. 03, 2023)

[2] Choudhary, V. (2018, August). Mall Customer Segmentation Data, Version 1. Retrieved June 3, 2023 from https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python.
VIJAY CHOUDHARY · ATUALIZADO HÁ 5 ANOS

[3] Choi, M. (2018, February). Medical Cost Personal Datasets, Version 1. Retrieved March 21, 2023 from https://www.kaggle.com/datasets/mirichoi0218/insurance.

[4] ”Scikit Learn” https://scikit-learn.org/stable/modules/generated/sklearn.mixture.Gaussia
(accessed Jun. 3, 2023)

[5] ”Scikit Learn” https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.
(accessed Jun. 3, 2023)