Tecnologia em Análise e Desenvolvimento de Sistemas - TADS

Reconhecimento de Padrões

TADS – IFRS – 2021/1

Prof. ***Luciano Vargas Gonçalves***

Prof. ***Luís Henrique Gularte Ferreira***

Aplicações com K-means

Aprendizagem de Máquina

- **Aprendizagem de máquina (AM)**
  - O processo de aprendizado **consiste** ***no treinamento de um algoritmo*** **ou modelo para que possa criar regras que relacionam os dados de entrada** (atributos previsores) com os dados de saída (atributo alvo), permitindo a realização de tarefas como classificação, previsão e agrupamento de dados.

  - Dessa forma, o essencial no aprendizado de máquina é o **reconhecimento de padrões**, ou seja, **a busca por semelhanças entre as características de diferentes instâncias de determinado conjunto de dados**.

Aprendizado Supervisionado

- Objetivo AM
  - Diferentes abordagens de aprendizado de máquina foram desenvolvidas ao longo do tempo. **Sejam os algoritmos, orientados para a classificação, regressão ou agrupamento, todos têm por finalidade acumular conhecimento sobre determinado conjunto de dados, utilizando técnicas de treinamento para melhorar o modelo que relacionar os atributos de entrada a um alvo ou grupo de dados**

Aprendizagem de Máquina

- Etapas do processo

Figura 1. Processo geral do aprendizado de máquina.

K-Means Clusterização
Características

- Dados não rotulados
- Aprendizado não supervisionado

Características	Tipos de aprendizado		
	Semissupervisionado		Reforço
	Supervisionado	Não supervisionado	
Conjunto de dados	Valores para atributo previsor e alvo.	Dados não rotulados. Ex. K-Means	Sem atributo-alvo.
Aprimoramento	Treinamento do modelo com base nas instâncias rotuladas.	Análise intrínseca.	Recompensas e punições.
Tarefa	Prever a resposta ou o rótulo correto.	Agrupar instâncias com características similares.	Buscar novas hipóteses no sentido de tentar reduzir as punições e aumentar as recompensas.

Seleção do Dados

- Seleção
  - Primeiramente, os dados devem ser selecionados em quantidade e qualidade suficientes para extrair o conhecimento necessário. Quanto maior a quantidade de exemplos / amostras obtidos, melhor será o aprendizado. Em contrapartida, a qualidade tem relação direta com os atributos escolhidos. Devem ser priorizados os de maior relevância para o modelo.

Seleção do Dados

Seleção

- Aplicação realizada aos dados extraídos de:
  - FEE (Fundação de Economia e Estatísticas do estado do Rio grande do Sul)
    - http://deedados.planejamento.rs.gov.br/feedados/#!pesquisa=0
  - Dados relativos ao consumo de Energia Elétrica nas 35 Microrregiões do RS, no ano de 2009.
    - Consumo Rural, Urbano, Comercial e Industrial

Seleção do Dados

Seleção

Microrregiões	Código	Energia Elétrica			
		Consumo			
		Rural	Residencial	Industrial	Comercial
		2009 (MWh)	2009 (MWh)	2009 (MWh)	2009 (MWh)
Cachoeira do Sul	430422	65.104	71.145	53.723	30.007
Camaquã	430528	51.069	53.691	101.153	25.521
Campanha Central	430630	26.356	98.258	23.862	40.963
Campanha Meridional	430631	43.841	89.725	54.723	33.812
Campanha Ocidental	430629	432.248	208.192	123.380	84.706
Carazinho	430109	33.898	30.575	24.862	16.546
Caxias do Sul	430216	110.497	502.222	1.375.086	336.732
Cerro Largo	430106	47.575	20.054	12.302	10.477
Cruz Alta	430111	79.046	69.500	30.289	38.374
Erechim	430104	63.385	97.791	126.535	55.852
Frederico Westphalen	430103	89.341	49.314	28.894	27.451
Gramado-Canela	430524	35.141	194.799	248.591	114.179
Guaporé	430214	63.961	55.091	180.621	28.864
Ijuí	430108	52.380	18.060	2.835	10.158
Jaguarão	430734	35.010	25.759	5.536	9.622
Lajeado-Estrela	430421	363.694	128.799	283.648	76.305
Litoral Lagunar	430735	90.877	146.150	88.610	93.994
Montenegro	430523	54.695	109.905	277.682	52.199
Não-Me-Toque	430112	37.631	16.189	20.050	10.263
Osório	430527	115.442	314.097	94.709	126.588
Passo Fundo	430110	141.929	166.627	217.116	124.130
Pelotas	430733	82.594	253.295	163.649	123.914
Porto Alegre	430526	64.667	2.658.652	2.343.299	2.049.046
Restinga Seca	430319	25.008	7.504	10.829	3.401
Sananduva	430105	24.619	18.518	10.028	9.768
Santa Cruz do Sul	430420	110.874	158.905	226.043	95.697
Santa Maria	430318	68.991	260.433	66.207	123.052
Santa Rosa	430101	68.633	73.463	80.939	43.920

Matriz de dados base extraída do site da FEE.

Preparação do Dados

Preparação

- Preparação
  - Na sequência, os dados devem ser preparados e adequados ao modelo utilizado. Nessa etapa, estão incluídas as transformações de unidade, conversão de escala, normalização, discretização e mudanças de representação dos dados. É importante avaliar o balanceamento dos dados, ou seja, se os dados coletados para diferentes faixas ou classes de previsão estão presentes em quantidades equivalentes.
  - Os dados podem ainda ser separados em dois grupos, um para ser utilizado na etapa de **treinamento e outro para testes.**

Preparação do Dados

Preparação

- Preparação
  - Planilha foi exportada para formado de planilha eletrônica (Excel ou Calc)
  - Formatada conforme as necessidades do programa e dos dados;
  - Padronizada
    - Pelo aplicação da fórmula da variável Z

$$Z = \frac{X - \mu}{\sigma}$$

Z - variável normal padronizada
X - variável normal
μ - média
σ - desvio padrão

Preparação do Dados

Preparação

MDB_2009 (Matriz de Dados Base)

	A	B	C	D	E	F
1	Id	Unidade (Instância)	W	Z	V	T
2	1	Cachoeira do Sul	65103.886	71145.072	53722.653	30006.699
3	2	Camaquã	51068.538	53690.597	101152.884	25521.026
4	3	Campanha Central	26355.83	98258.058	23862.031	40962.698
5	4	Campanha Meridional	43840.784	89724.777	54723.009	33811.568
6	5	Campanha Ocidental	432248.438	208192.252	123380.403	84705.631
7	6	Carazinho	33897.662	30575.477	24861.584	16545.556
8	7	Caxias do Sul	110497.11	502221.559	1375085.758	336732.082
9	8	Cerro Largo	47574.606	20053.877	12301.693	10476.661
10	9	Cruz Alta	79046.232	69500.057	30288.704	38373.853
11	10	Erechim	63384.731	97791.243	126535.095	55852.134
12	11	Frederico Westphalen	89340.864	49313.723	28893.612	27450.59
13	12	Gramado-Canela	35140.788	194798.585	248590.746	114178.646
14	13	Guaporé	63961.228	55091.439	180620.708	28863.639
15	14	Ijuí	52380.02	18060.388	2835.134	10157.619
16	15	Jaguarão	35009.507	25759.175	5536.205	9622.07
17	16	Lajeado-Estrela	363693.825	128799.219	283648.297	76304.829
18	17	Litoral Lagunar	90876.853	146150.35567	88610.412	93993.969
19	18	Montenegro	54694.55	109904.877	277682.24	52198.946
20	19	Não-Me-Toque	37630.626	16188.994	20050.475	10263.048

Valores observados

MDP_2009 (Matriz de Dados Padronizada)

1	Id	Unidade (Instância)	W	Z	V	T
2	1	Cachoeira do Sul	-0.1621	-0.2438	-0.3272	-0.2457
3	2	Camaquã	-0.3267	-0.2831	-0.2195	-0.2588
4	3	Campanha Central	-0.6165	-0.1826	-0.3949	-0.2137
5	4	Campanha Meridional	-0.4114	-0.2019	-0.3249	-0.2346
6	5	Campanha Ocidental	4.1435	0.0652	-0.1691	-0.0859
7	6	Carazinho	-0.5280	-0.3352	-0.3927	-0.2850
8	7	Caxias do Sul	0.3702	0.7282	2.6712	0.6502
9	8	Cerro Largo	-0.3677	-0.3590	-0.4212	-0.3028
10	9	Cruz Alta	0.0014	-0.2475	-0.3803	-0.2213
11	10	Erechim	-0.1822	-0.1837	-0.1620	-0.1702
12	11	Frederico Westphalen	0.1221	-0.2930	-0.3835	-0.2532
13	12	Gramado-Canela	-0.5135	0.0350	0.1150	0.0001
14	13	Guaporé	-0.1755	-0.2800	-0.0392	-0.2490
15	14	Ijuí	-0.3113	-0.3635	-0.4426	-0.3037
16	15	Jaguarão	-0.5150	-0.3461	-0.4365	-0.3052
17	16	Lajeado-Estrela	3.3395	-0.1138	0.1946	-0.1105
18	17	Litoral Lagunar	0.1402	-0.0747	-0.2480	-0.0588
19	18	Montenegro	-0.2842	-0.1564	0.1810	-0.1809
20	19	Não-Me-Toque	-0.4843	-0.3677	-0.4036	-0.3034

Valores padronizados

Preparação do Dados

Preparação

Dados da planilha foram exportados para CSV e enviados para plataforma do Google - Drive

Abrir — MDB_2009.csv — ~/Documentos/IFRS-2021/RP/Testes/E

```
Id,Unidade (Instância),W,Z,V,T
1,Cachoeira do Sul,65104,71145,53723,30007
2,Camaquã,51069,53691,101153,25521
3,Campanha Central,26356,98258,23862,40963
4,Campanha Meridional,43841,89725,54723,33812
5,Campanha Ocidental,432248,208192,123380,84706
6,Carazinho,33898,30575,24862,16546
7,Caxias do Sul,110497,502222,1375086,336732
8,Cerro Largo,47575,20054,12302,10477
9,Cruz Alta,79046,69500,30289,38374
10,Erechim,63385,97791,126535,55852
11,Frederico Westphalen,89341,49314,28894,27451
12,Gramado-Canela,35141,194799,248591,114179
13,Guaporé,63961,55091,180621,28864
14,Ijuí,52380,18060,2835,10158
15,Jaguarão,35010,25759,5536,9622
16,Lajeado-Estrela,363694,128799,283648,76305
17,Litoral Lagunar,90877,146150,88610,93994
18,Montenegro,54695,109905,277682,52199
19,Não-Me-Toque,37631,16189,20050,10263
20,Osório,115442,314097,94709,126588
21,Passo Fundo,141929,166627,217116,124130
22,Pelotas,82594,253295,163649,123914
23,Porto Alegre,64667,2658652,2343299,2049046
24,Restinga Seca,25008,7504,10829,3401
25,Sananduva,24619,18518,10028,9768
26,Santa Cruz do Sul,110874,158905,226043,95697
27,Santa Maria,68991,260433,66207,123052
28,Santa Rosa,68633,73463,80939,43920
29,Santiago,35698,50195,10820,21173
30,Santo Ângelo,64628,81542,64486,43100
31,São Jerônimo,35745,63899,326840,22878
32,Serras de Sudeste,23155,43545,140411,16182
33,Soledade,17678,16606,3652,9255
34,Três Passos,59499,50619,50864,27660
35,Vacaria,47485,71018,54477,58773
```

Matriz de dados Base

Abrir — MDB_PC_2009. — ~/Documentos/IFRS-2021/RP

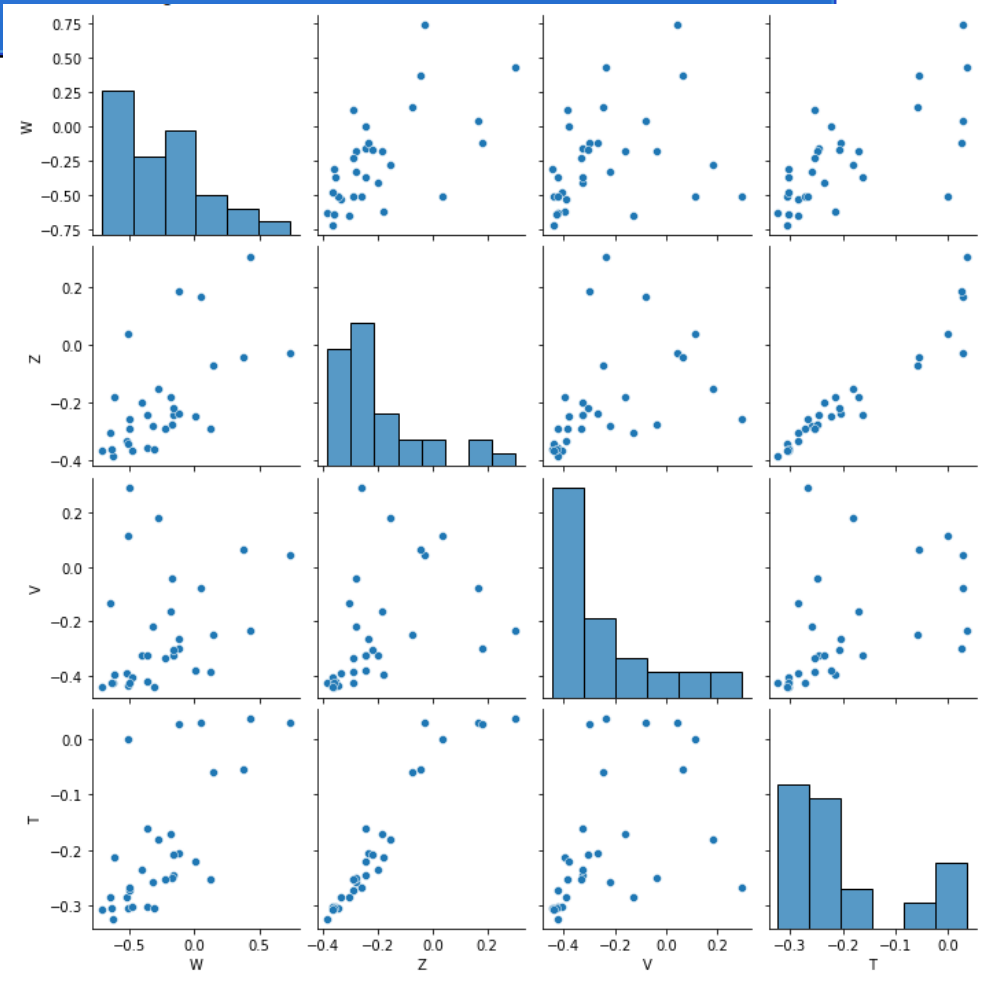
```
Id,Unidade (Instância),W,Z,V,T
1,Cachoeira do Sul,-0.1621,-0.2438,-0.3272,-0.2457
2,Camaquã,-0.3267,-0.2831,-0.2195,-0.2588
3,Campanha Central,-0.6165,-0.1826,-0.3949,-0.2137
4,Campanha Meridional,-0.4114,-0.2019,-0.3249,-0.2346
6,Carazinho,-0.5280,-0.3352,-0.3927,-0.2850
8,Cerro Largo,-0.3677,-0.3590,-0.4212,-0.3028
9,Cruz Alta,0.0014,-0.2475,-0.3803,-0.2213
10,Erechim,-0.1822,-0.1837,-0.1620,-0.1702
11,Frederico Westphalen,0.1221,-0.2930,-0.3835,-0.253
12,Gramado-Canela,-0.5135,0.0350,0.1150,0.0001
13,Guaporé,-0.1755,-0.2800,-0.0392,-0.2490
14,Ijuí,-0.3113,-0.3635,-0.4426,-0.3037
15,Jaguarão,-0.5150,-0.3461,-0.4365,-0.3052
16,Lajeado-Estrela,3.3395,-0.1138,0.1946,-0.1105
17,Litoral Lagunar,0.1402,-0.0747,-0.2480,-0.0588
18,Montenegro,-0.2842,-0.1564,0.1810,-0.1809
19,Não-Me-Toque,-0.4843,-0.3677,-0.4036,-0.3034
20,Osório,0.4282,0.3040,-0.2342,0.0364
21,Passo Fundo,0.7389,-0.0285,0.0436,0.0292
22,Pelotas,0.0430,0.1669,-0.0777,0.0286
24,Restinga Seca,-0.6323,-0.3873,-0.4245,-0.3234
25,Sananduva,-0.6369,-0.3624,-0.4263,-0.3048
26,Santa Cruz do Sul,0.3747,-0.0459,0.0638,-0.0538
27,Santa Maria,-0.1165,0.1830,-0.2988,0.0261
28,Santa Rosa,-0.1207,-0.2385,-0.2654,-0.2051
29,Santiago,-0.5069,-0.2910,-0.4245,-0.2715
30,Santo Ângelo,-0.1677,-0.2203,-0.3027,-0.2075
31,São Jerônimo,-0.5064,-0.2601,0.2926,-0.2665
32,Serras de Sudeste,-0.6540,-0.3060,-0.1305,-0.2861
33,Soledade,-0.7183,-0.3667,-0.4408,-0.3063
34,Três Passos,-0.2278,-0.2901,-0.3337,-0.2526
35,Vacaria,-0.3687,-0.2441,-0.3255,-0.1617
```

Matriz de dados Padronizados

Cidades com atributos acima de +-2 desvio padrão foram retiradas:

- Porto Alegre
- Caxias
- Campanha Ocidental

Preparação do Dados



Seleção do Modelo

- Seleção do modelo
  - Algoritmos e implantação
  - Os modelos podem incluir regressões lineares, regressões logísticas, classificação, agrupamento, aprendizado profundo, entre outros.
  - K-Means será o modelo utilizados no exemplo.

Seleção do Modelo

- Modelo Aplicado
  - K-Means
    - Realiza a clusterização (agrupamento) de objetos não rotulados, através do cálculo da menor distância entre os centroides;
  - No exemplo foi utilizada a plataforma Colab (google), na linguagem Python.
    - Uso do Drive como repositório

Seleção do Modelo

Modelo

```
[2] from google.colab import drive

    drive.mount('/content/drive')

    Mounted at /content/drive
```

```
[3] import pandas as pd

    df = pd.read_csv('/content/drive/My Drive/Colab Notebooks/Bases/MDP_PC_2009.csv')
    df.head()
```

```
[ ] import numpy as np
    import pandas as pd
    import matplotlib.pyplot as pl
    import seaborn as sb
```

```
[ ] sb.pairplot(df)
```

```
[ ] X = np.array(df.drop('Unidade (Instância)',axis = 1))
    X
```

```
[ ] from sklearn.cluster import KMeans

    kmeans = KMeans(n_clusters=6, random_state=50)

    kmeans.fit(X)
```

Kmeans

Plataforma Colab

Treinamento

Treino

- Treinamento
  - É a intenção é aprimorar o modelo a cada nova amostra de treinamento avaliada.
  - K-means:
    - Processo iterativo.
    - De forma a maximizar a distância Entre os Grupos (Inter-Cluster) e minimizar a distância Dentro dos Grupos (Intra-cluster).

Treinamento (fit)

Treino

```
[ ]  from sklearn.cluster import KMeans

     kmeans = KMeans(n_clusters=6, random_state=50)

     kmeans.fit(X)
```

Preparação Algoritmo

Treinamento

```
[ ]  kmeans.labels_
```

```
[10] kmeans.labels_

     array([1, 1, 3, 1, 3, 3, 0, 1, 0, 4, 1, 1, 3, 0, 4, 3, 2, 5, 2, 3, 3, 5,
            2, 1, 3, 1, 4, 3, 3, 1, 1], dtype=int32)
```

Classificação

Agrupamento das 32 amostras (microrregiões)

Avaliação do Modelo

- Avaliação
  - Em seguida, a etapa de avaliação utiliza os dados separados inicialmente para efetuar os testes e determinar se o modelo obtido após o treinamento tem a precisão esperada na predição do alvo a partir de conjuntos de dados até então desconhecidos.
  - Concluída a etapa de avaliação, podem ser definidos novos hiperparâmetros (que controlam o processo de aprendizado em si) que incluem:
    - quantidade de repetições de todo o processo sobre o conjunto de dados de treinamento;
    - taxa de aprendizado, fator para as mudanças de parâmetros do modelo;
    - entre outros.

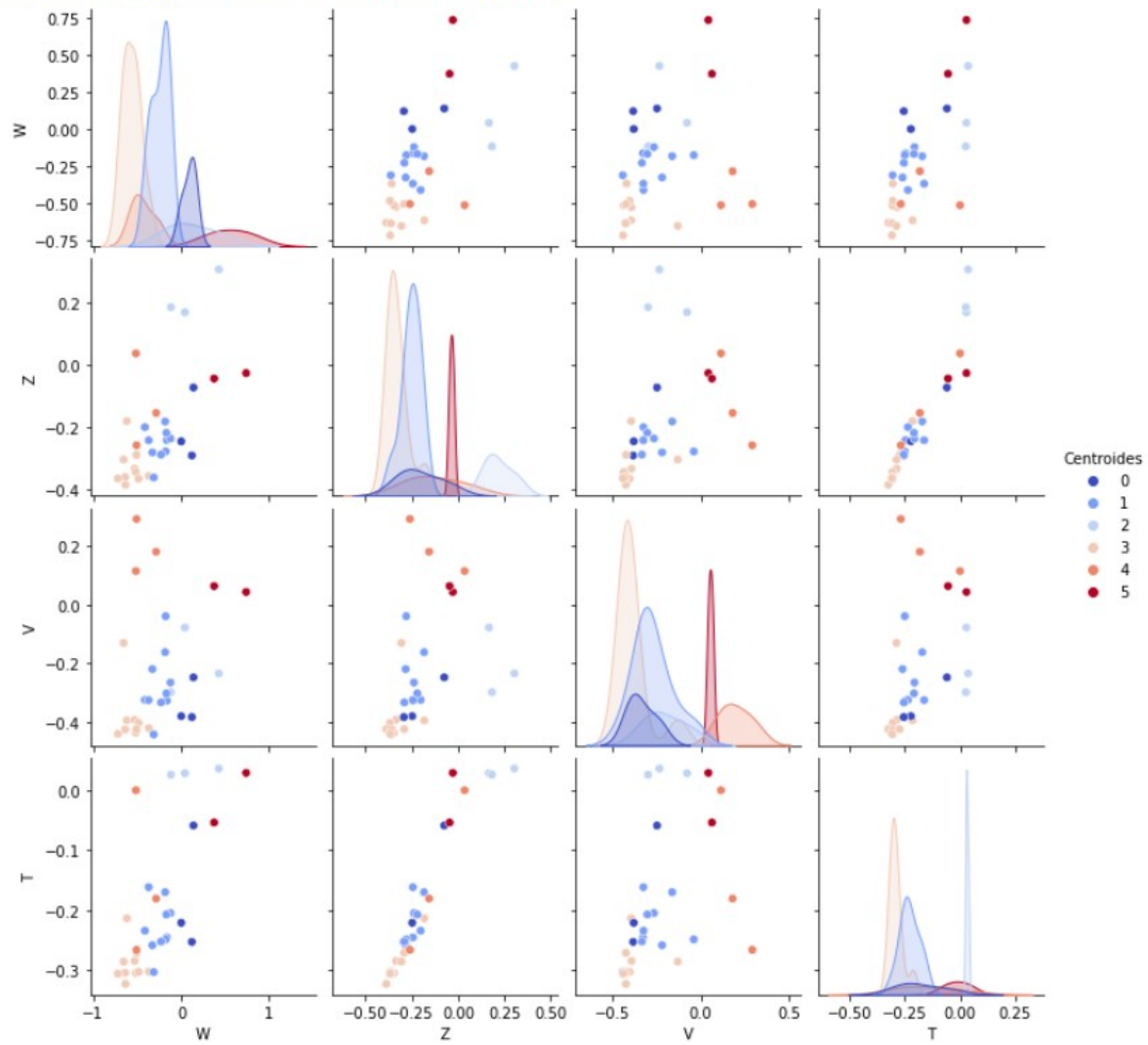
Avaliação do

- Avaliação da Dispersão

Avaliação

```
sb.pairplot(df,hue='Centroides',palette='coolwarm')
```

```
<seaborn.axisgrid.PairGrid at 0x7f3d5afc4710>
```



Avaliação do Modelo

Avaliação

- Avaliação dos Centroides:

```
kmeans.cluster_centers_
```

```
array([[ 0.0879    , -0.20506667, -0.33726667, -0.17776667],
       [-0.24541   , -0.2549    , -0.27427   , -0.22889   ],
       [ 0.11823333,  0.21796667, -0.20356667,  0.03036667],
       [-0.56599   , -0.3304    , -0.38955   , -0.29022   ],
       [-0.4347    , -0.12716667,  0.1962    , -0.1491    ],
       [ 0.5568    , -0.0372    ,  0.0537    , -0.0123    ]])
```

Centroides

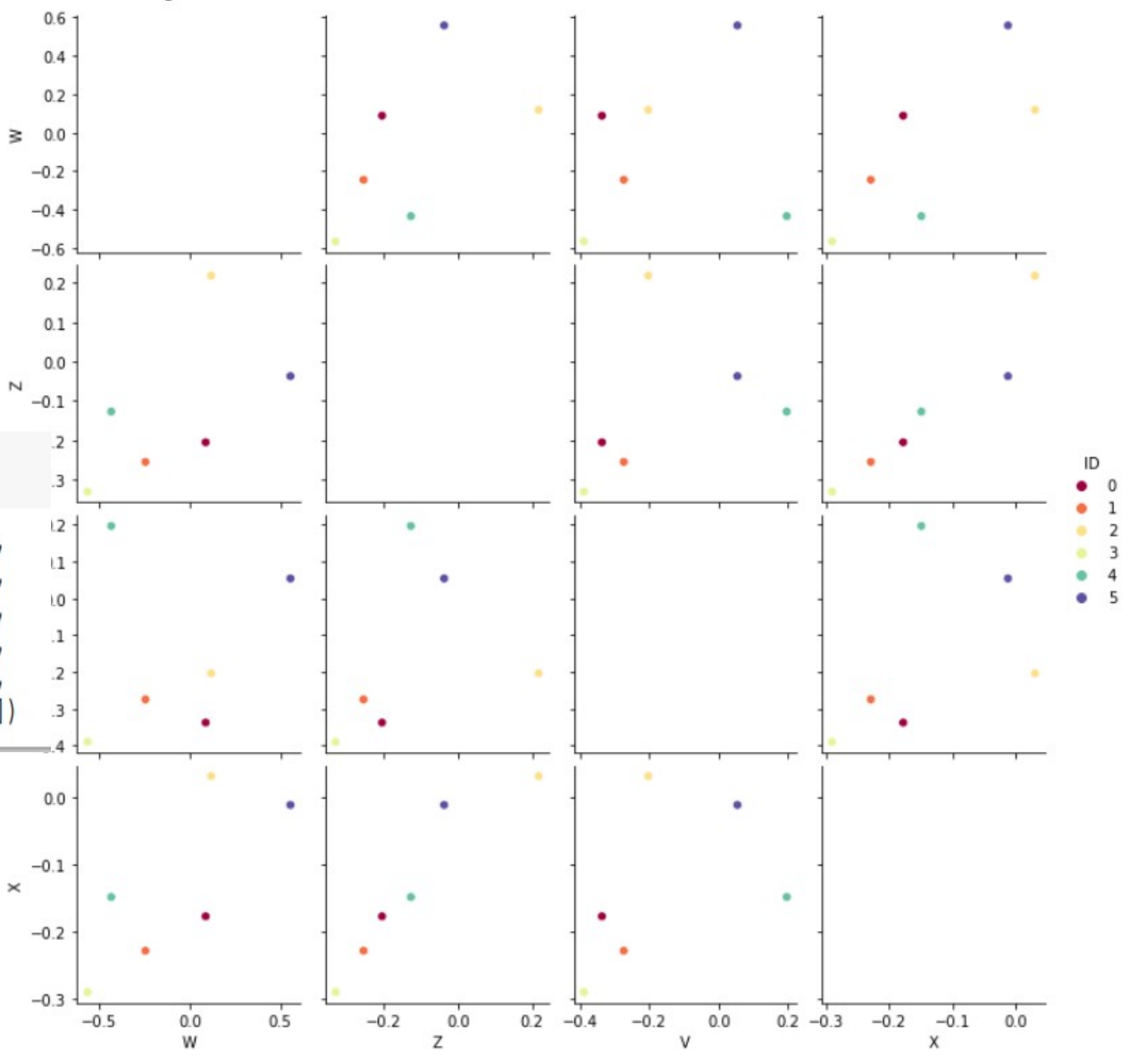
0
1
2
3
4
5

Avaliação do

- Centroides
  - 0 à 5

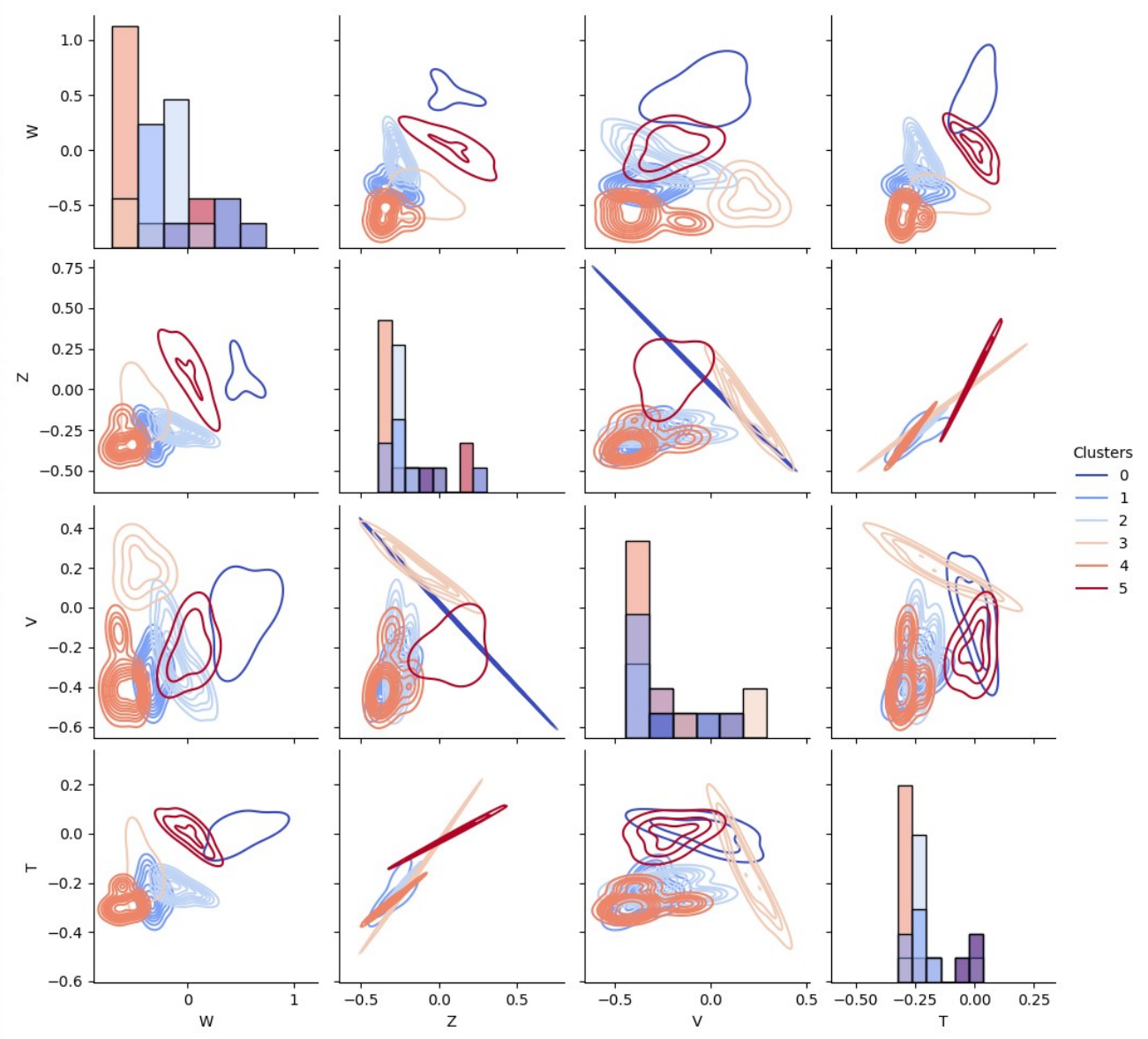
```
means.cluster_centers_
```

```
rray([[ 0.0879    , -0.20506667, -0.33726667, -0.17776667],
      [-0.24541   , -0.2549    , -0.27427   , -0.22889   ],
      [ 0.11823333,  0.21796667, -0.20356667,  0.03036667],
      [-0.56599   , -0.3304    , -0.38955   , -0.29022   ],
      [-0.4347    , -0.12716667,  0.1962    , -0.1491    ],
      [ 0.5568    , -0.0372    ,  0.0537    , -0.0123    ]])
```



Avaliação do

- Centroides
  - 0 à 5



Aplicação

Avaliação

- Aplicação
  - Por fim, a etapa de aplicação diz respeito ao uso do modelo para a realização de previsões a partir da máquina já treinada

Seleção do Dados

Dúvidas??