Estudo de Caso

Email: vitortbarboza@hotmail.com

Telefone: (91)98611-4022

Objetivo do estudo de caso

Uma grande multinacional varejista do ramo de supermercados deseja ingressar no mercado brasileiro.

É necessário fazer uma análise que servirá de base para a estratégia (tomada de decisão) e escolha das cidades de entrada dessa empresa no Brasil.

Tarefas a serem realizadas:

- Realizar uma classificação dos municípios brasileiros com base nas informações disponíveis e faça uma caracterização dos municípios em grupos.
- Elaborar uma forma de classificar um novo município entre os grupos.
- Responder quais grupos de municípios deveriam ser a porta de entrada para empresa no país e porque.

A resolução do Case será feito em Python

Análise Exploratória dos Dados

Primeiramente, realizou-se uma breve Análise nos dados com a finalidade de obter um melhor entendimento a respeito dos mesmos.

```
# Análise Exploratória dos Dados - EDA
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 5507 entries, Abadia de Goiás (GO) to Zortéa (SC)
Data columns (total 23 columns):
 #   Column                                                                Non-Null Count  Dtype
---  ------                                                                --------------  -----
 0   Área (km²)                                                            5507 non-null   float64
 1   Densidade demográfica, 2000                                           5507 non-null   float64
 2   Distância à capital (km)                                              5507 non-null   float64
 3   Esperança de vida ao nascer, 2000                                     5507 non-null   float64
 4   Mortalidade até um ano de idade, 2000                                 5507 non-null   float64
 5   Taxa de fecundidade total, 2000                                       5507 non-null   float64
 6   Percentual de pessoas de 25 anos ou mais analfabetas, 2000            5507 non-null   float64
 7   Renda per Capita, 2000                                                5507 non-null   float64
 8   Índice de Gini, 2000                                                  5507 non-null   float64
 9   Intensidade da indigência, 2000                                       5507 non-null   float64
 10  Intensidade da pobreza, 2000                                          5507 non-null   float64
 11  Índice de Desenvolvimento Humano Municipal, 2000                      5507 non-null   float64
 12  Taxa bruta de freqüência à escola, 2000                               5507 non-null   float64
 13  Taxa de alfabetização, 2000                                           5507 non-null   float64
 14  Média de anos de estudo das pessoas de 25 anos ou mais de idade, 2000 5507 non-null   float64
 15  População de 25 anos ou mais de idade, 1991                           5507 non-null   int64
 16  População de 25 anos ou mais de idade, 2000                           5507 non-null   int64
 17  População de 65 anos ou mais de idade, 1991                           5507 non-null   int64
 18  População de 65 anos ou mais de idade, 2000                           5507 non-null   int64
 19  População total, 1991                                                 5507 non-null   int64
 20  População total, 2000                                                 5507 non-null   int64
 21  População urbana, 2000                                                5507 non-null   int64
 22  População rural, 2000                                                 5507 non-null   int64
dtypes: float64(15), int64(8)
memory usage: 1.0+ MB
```

A base de dados utilizada no presente trabalho possui 5507 observações e 22 variáveis.

Análise Exploratória dos Dados

- Em seguida, foi utilizada a função *describe* para se obter as estatísticas descritivas na qual incluem aquelas que resumem a tendência central, dispersão e forma de distribuição de um conjunto de dados.

```
df.describe()
```

	Área (km²)	Densidade demográfica, 2000	Distância à capital (km)	Esperança de vida ao nascer, 2000	Mortalidade até um ano de idade, 2000	Taxa de fecundidade total, 2000	Percentual de pessoas de 25 anos ou mais analfabetas, 2000	Renda per Capita, 2000	Índice de Gini, 2000	Intensidade da indigência, 2000	...	Taxa de alfabetização 200(
count	5507.000000	5507.000000	5507.000000	5507.000000	5507.000000	5507.000000	5507.000000	5507.000000	5507.000000	5507.000000	...	5507.00000(
mean	1549.211476	96.731869	253.212620	67.748925	34.083376	2.864845	26.668411	170.814160	0.560734	49.786210	...	78.23062(
std	5738.392465	524.006185	163.210532	4.860915	18.470551	0.744454	15.164462	96.425347	0.058663	10.571483	...	12.46025'
min	2.900000	0.100000	0.000000	54.350000	5.380000	1.560000	2.020000	28.380000	0.360000	0.020000	...	39.34000(
25%	205.700000	11.300000	121.858906	64.530000	18.640000	2.320000	13.970000	86.495000	0.520000	42.830000	...	67.93000(
50%	417.200000	23.600000	228.262939	68.240000	29.510000	2.670000	22.600000	159.100000	0.560000	49.510000	...	82.04000(
75%	1031.450000	48.000000	358.072044	71.440000	46.150000	3.230000	39.780000	232.695000	0.600000	56.455000	...	88.34000(
max	161445.900000	12881.400000	1474.314590	78.180000	109.670000	7.790000	70.260000	954.650000	0.820000	88.350000	...	99.09000(

8 rows × 23 columns

Análise Exploratória dos Dados

- Posteriormente, verificou-se a existência de *missing values* nos dados. De acordo com a figura abaixo, não há *missing values* presente nos dados

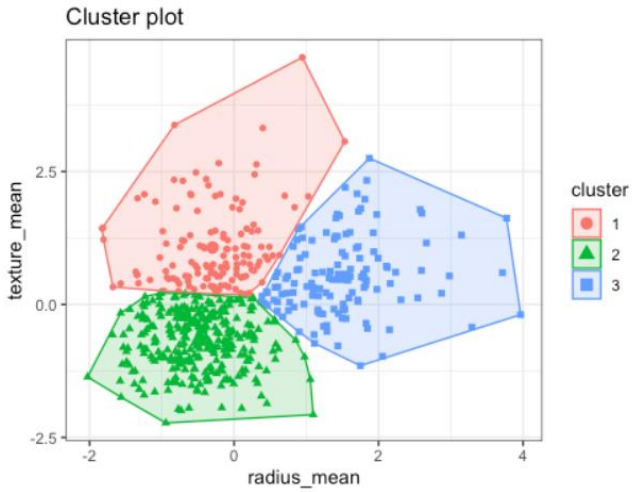
```
df.isnull().sum()
```

```
Área (km²)                                                          0
Densidade demográfica, 2000                                         0
Distância à capital (km)                                            0
Esperança de vida ao nascer, 2000                                   0
Mortalidade até um ano de idade, 2000                               0
Taxa de fecundidade total, 2000                                     0
Percentual de pessoas de 25 anos ou mais analfabetas, 2000          0
Renda per Capita, 2000                                              0
Índice de Gini, 2000                                                0
Intensidade da indigência, 2000                                     0
Intensidade da pobreza, 2000                                        0
Índice de Desenvolvimento Humano Municipal, 2000                    0
Taxa bruta de freqüência à escola, 2000                             0
Taxa de alfabetização, 2000                                         0
Média de anos de estudo das pessoas de 25 anos ou mais de idade, 2000    0
População de 25 anos ou mais de idade, 1991                         0
População de 25 anos ou mais de idade, 2000                         0
População de 65 anos ou mais de idade, 1991                         0
População de 65 anos ou mais de idade, 2000                         0
População total, 1991                                               0
População total, 2000                                               0
População urbana, 2000                                              0
População rural, 2000                                               0
dtype: int64
```

Caracterização dos municípios em grupos

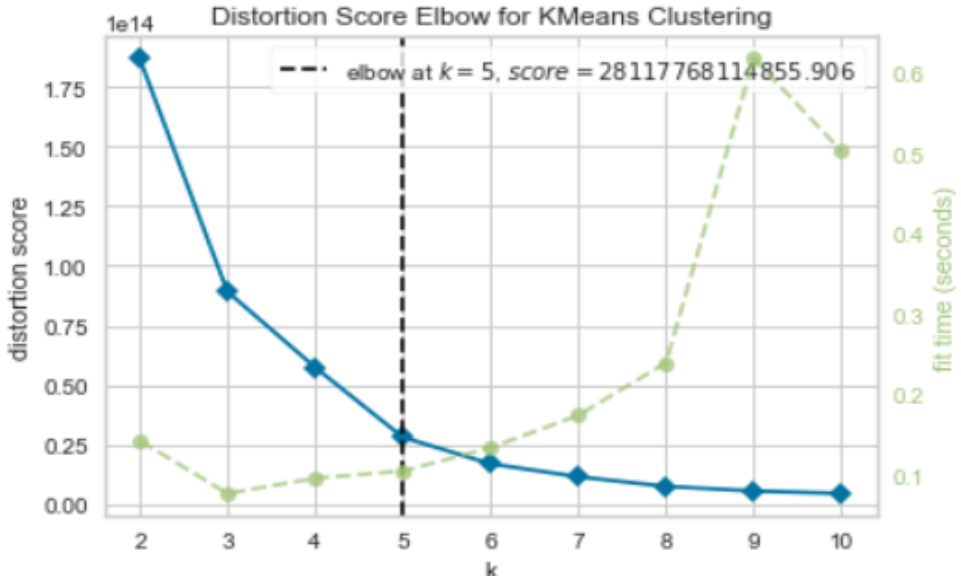
O método escolhido para a caracterização foi o algoritmo de Machine Learning K-Means, O K-Means é um algoritmo do tipo de aprendizado não supervisionado, isto é, observa-se um vetor de regressores, no entanto, não há uma variável dependente(y) que possa supervisionar o aprendizado do modelo.

O algoritmo K-Means agrupa as observações(municípios) em Clusters (grupos) de acordo com as características que são identificadas através dos dados.



Caracterização dos municípios em grupos

Para obter o número ótimo de clusters, é necessário utilizar o método cotovelo (elbow method), a ideia é rodar o KMeans para vários quantidades diferentes de clusters e dizer qual dessas quantidades é o número ótimo de clusters. O que geralmente acontece ao aumentar a quantidade de clusters no KMeans é que as diferenças entre clusters se tornam muito pequenas, e as diferenças das observações intra-clusters vão aumentando



De acordo com o gráfico, o número ótimo de clusters é 5. Isto é, os municípios serão agrupados em 5 grupos

Caracterização dos municípios em grupos

Correu o algoritmo K-Means com 5 clusters, e tirou-se a média de algumas variáveis com a finalidade de caracterizar os grupos.

	cluster_class	Área (km²)	Densidade demográfica, 2000	Renda per Capita, 2000	Índice de Gini, 2000	Índice de Desenvolvimento Humano Municipal, 2000	População total, 2000
0	1	1544.040144	62.891032	167.338391	0.560509	0.697507	1.872727e+04
1	2	1528.500000	6808.100000	610.040000	0.620000	0.841000	1.043425e+07
2	3	1804.036782	1702.487356	343.485172	0.564713	0.795989	4.046607e+05
3	4	1264.200000	4627.900000	596.650000	0.620000	0.842000	5.857904e+06
4	5	2159.460000	3303.530000	461.766000	0.636000	0.820400	1.702445e+06

Depois, realizou-se uma contagem para saber quantos municípios existem em cada cluster

```
df['cluster_class'].value_counts()
```

```
1    5408
3      87
5      10
4       1
2       1
```

Caracterização dos municípios em grupos

Analisando de uma forma mais detalhada, o cluster 2 e 4 é referente a São Paulo e Rio de Janeiro respectivamente.

```
df_cluster[df_cluster["cluster_class"] == 2]
```

Município	cluster_class	Área (km²)	Densidade demográfica, 2000	Renda per Capita, 2000	Índice de Gini, 2000	Índice de Desenvolvimento Humano Municipal, 2000	População total, 2000
São Paulo (SP)	2	1528.5	6808.1	610.04	0.62	0.841	10434252

```
df_cluster[df_cluster["cluster_class"] == 4]
```

Município	cluster_class	Área (km²)	Densidade demográfica, 2000	Renda per Capita, 2000	Índice de Gini, 2000	Índice de Desenvolvimento Humano Municipal, 2000	População total, 2000
Rio de Janeiro (RJ)	4	1264.2	4627.9	596.65	0.62	0.842	5857904

Caracterização dos municípios em grupos

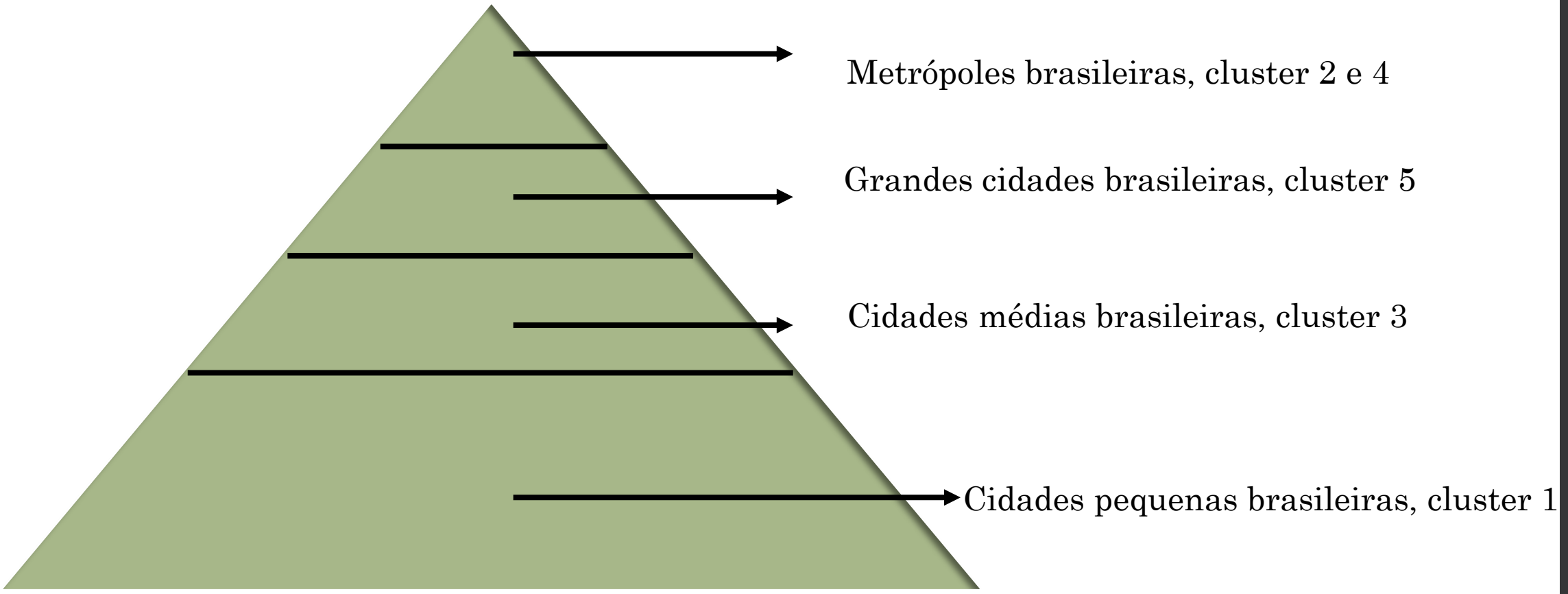
Analisando de uma forma mais detalhada, estas são as cidades presentes no cluster 5

```
df_cluster[df_cluster["cluster_class"] == 5]
```

Município	cluster_class	Área (km²)	Densidade demográfica, 2000	Renda per Capita, 2000	Índice de Gini, 2000	Índice de Desenvolvimento Humano Municipal, 2000	População total, 2000
Belém (PA)	5	1070.1	1196.0	313.93	0.65	0.806	1280614
Belo Horizonte (MG)	5	331.9	6718.0	557.44	0.62	0.839	2238526
Brasília (DF)	5	5822.1	350.9	605.41	0.64	0.844	2051146
Curitiba (PR)	5	430.9	3682.8	619.82	0.59	0.856	1587315
Fortaleza (CE)	5	313.8	6814.0	306.70	0.66	0.786	2141402
Goiânia (GO)	5	743.0	1467.8	508.30	0.61	0.832	1093007
Manaus (AM)	5	11458.5	122.5	262.40	0.64	0.774	1405835
Porto Alegre (RS)	5	496.1	2741.2	709.88	0.61	0.865	1360590
Recife (PE)	5	218.7	6501.8	392.46	0.68	0.797	1422905
Salvador (BA)	5	709.5	3440.3	341.32	0.66	0.805	2443107

Caracterização dos municípios em grupos

Propôs-se a seguinte classificação dos clusters:



Forma de classificar um novo município entre os grupos.

Nessa seção do trabalho, há a necessidade de se desenvolver um modelo que possa classificar um novo município de acordo com os clusters na parte anterior.

Nessa seção, serão utilizados algoritmos de Machine Learning de aprendizado supervisionado, uma vez que agora se tem a variável dependente (cluster_class).

Os dados tiveram uma preparação que consistiu em separar os mesmos em dados de treino e dados teste. A divisão seguiu da seguinte maneira: 70% dos dados foram considerados como dados de treino, enquanto que 30% dos dados foram considerados como dados teste.

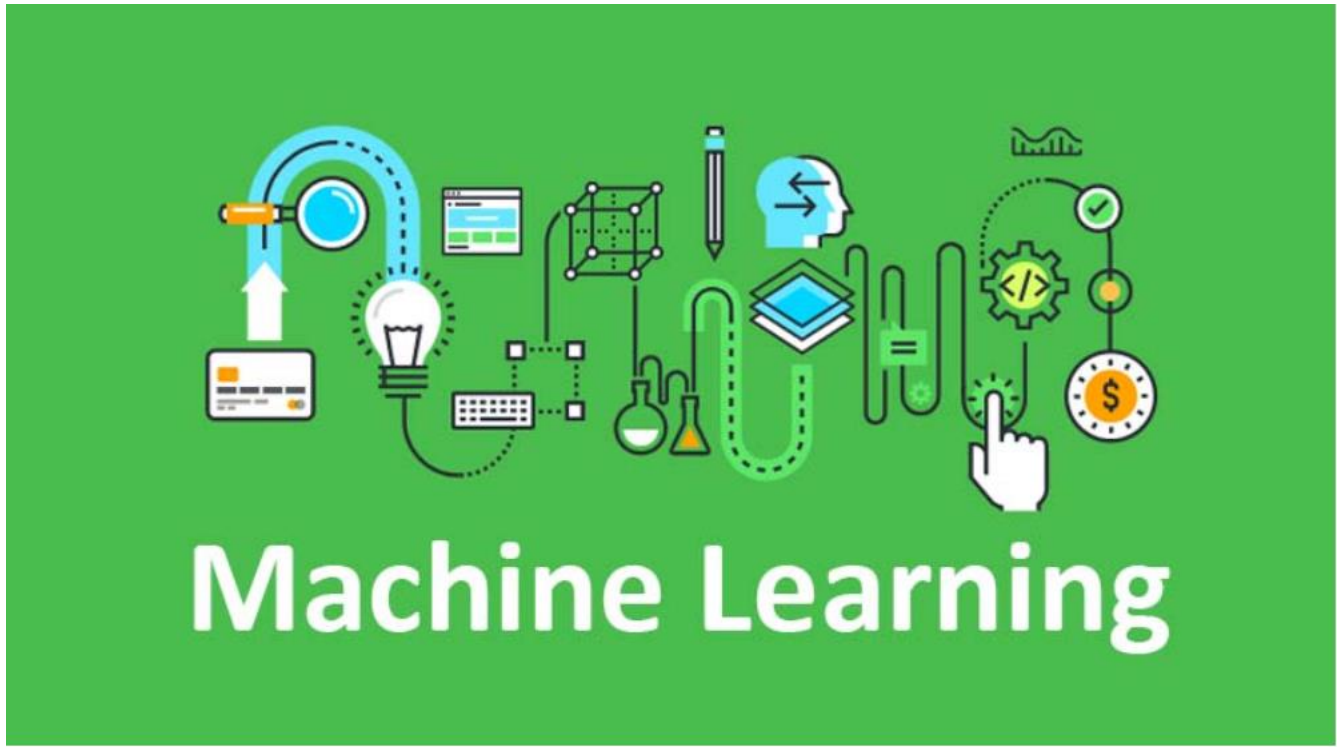
```
# Preparation to run the machine learning models
x=df.drop(['cluster_class'], axis=1)
y=df['cluster_class']
```

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.30,
                                                    random_state=101)
```

Forma de classificar um novo município entre os grupos.

Os seguintes algoritmos foram utilizados nesta seção:

- KNN
- Random Forest
- Support Vector Machine
- Decision Tree



Forma de classificar um novo município entre os grupos.

- KNN

	precision	recall	f1-score	support
1	1.00	1.00	1.00	1617
3	0.93	0.84	0.89	32
4	0.00	0.00	0.00	1
5	0.75	1.00	0.86	3
accuracy			1.00	1653
macro avg	0.67	0.71	0.69	1653
weighted avg	0.99	1.00	0.99	1653

- Random Forest

	precision	recall	f1-score	support
1	1.00	1.00	1.00	1617
3	1.00	0.94	0.97	32
4	0.00	0.00	0.00	1
5	0.75	1.00	0.86	3
accuracy			1.00	1653
macro avg	0.69	0.73	0.71	1653
weighted avg	1.00	1.00	1.00	1653

Forma de classificar um novo município entre os grupos.

- Support vector machine

	precision	recall	f1-score	support
1	1.00	1.00	1.00	1617
3	1.00	0.94	0.97	32
4	0.00	0.00	0.00	1
5	0.75	1.00	0.86	3
accuracy			1.00	1653
macro avg	0.69	0.73	0.71	1653
weighted avg	1.00	1.00	1.00	1653

- Decision Tree

	precision	recall	f1-score	support
1	0.99	1.00	1.00	1617
3	0.77	0.53	0.63	32
4	0.00	0.00	0.00	1
5	0.00	0.00	0.00	3
accuracy			0.99	1653
macro avg	0.44	0.38	0.41	1653
weighted avg	0.98	0.99	0.99	1653

Forma de classificar um novo município entre os grupos.

Algoritmo	F1 Score
KNN	1
Random Forest	1
SVM	1
Decision Tree	0,99

Todos os algoritmos tiveram uma boa qualidade para classificar os clusters presente nos dados teste de acordo com a sua classificação original. No entanto, tem que se ter um pouco de cuidado para ao avaliar a precisão dos algoritmos, pois os dados são não balanceados, devido ao fator de ter uma enorme disparidade entre os municípios presentes no cluster 1 em relação aos demais clusters. No entanto, recomenda-se a utilização do SVM para classificar novos municípios.

Quais municípios devem ser a porta de entrada para a empresa?

- A sugestão dos municípios para a entrada no país são os clusters 2 e 4, São Paulo e Rio de Janeiro respectivamente, uma vez que são as metrópoles brasileiras. Também, de acordo com os dados, são as cidades de maior população no país, assim como são as cidades mais ricas e desenvolvidas.
- Após a entrada nas cidades mencionadas anteriormente, o próximo passo é ingressar nas cidades do cluster 5, cluster no qual foi considerado como as grandes cidades brasileiras, cidades nas quais possuem uma grande população, um bom desenvolvimento e uma boa renda per capita.