

QUALIDADE DO LEITE

Quais as características de um bom e mal leite?



KDD

Descoberta de conhecimento em base de dados

1

Escolha da
Base

kaggle

2

Pré-
processamento



3

Transformação
de dados



4

Mineração
dos dados



5

Interpretação
do resultado



BASE DE DADOS



Número de amostras: **1059**

Número de atributos: **8**

Tipos de atributos: **Numéricos
(com rótulo categórico)**

Dados ausentes: **Não**

Sabor

Dado categórico
0 -> Sabor ruim
1 -> Sabor agradável



Odor

Dado categórico
0 -> Odor desagradável
1 -> Odor padrão



Gordura

Dado categórico
0 -> Baixa gordura
1 -> Muita gordura

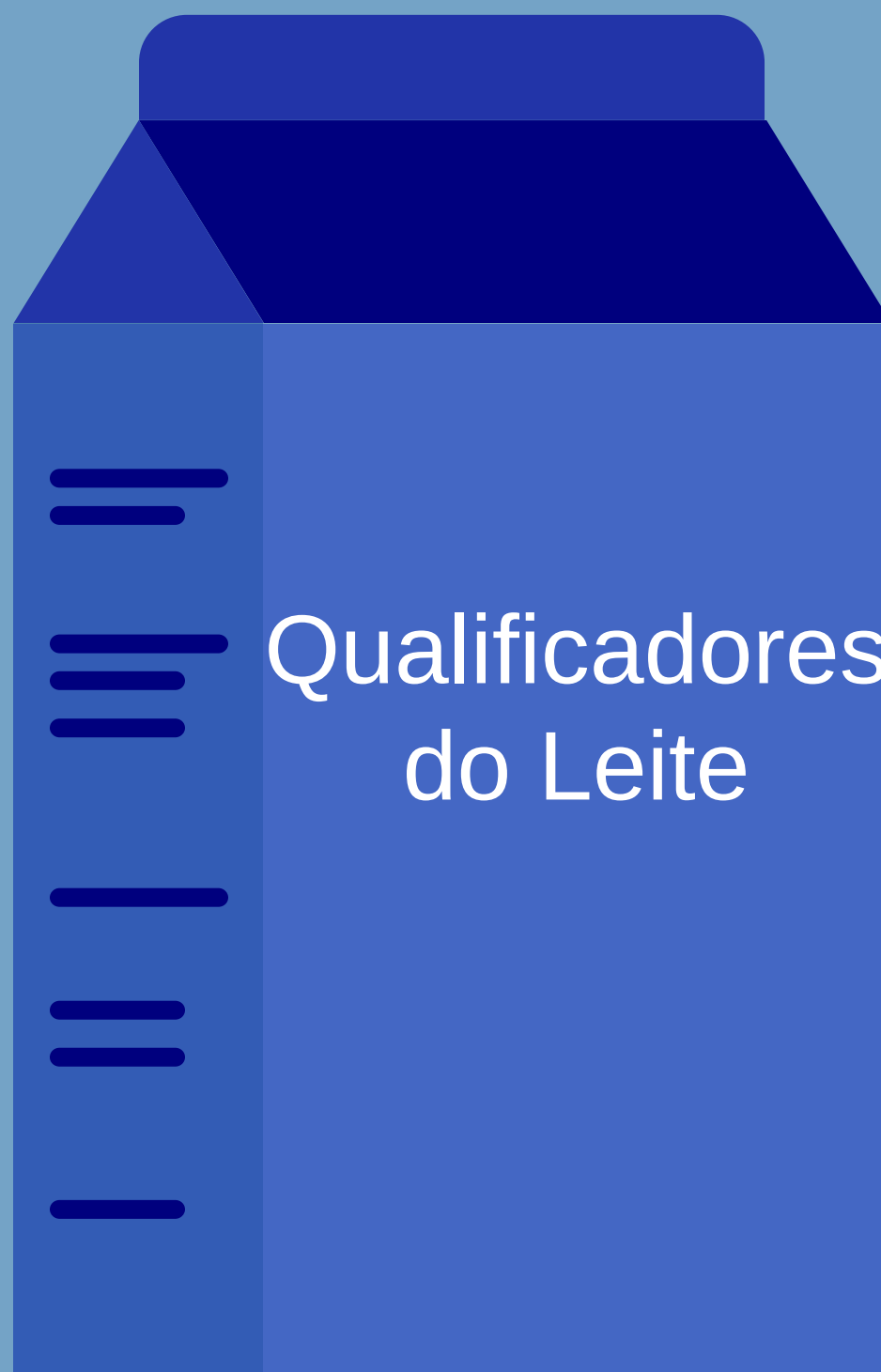


Cor

Varia entre 240-255, e indica a cor do leite



Atributos da Base



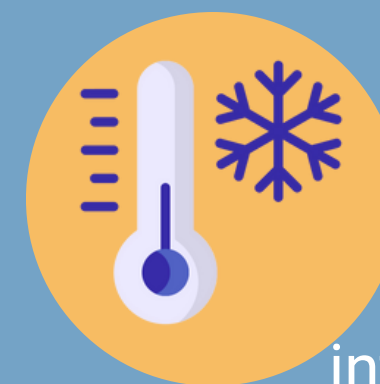
Turbidez

Dado categórico
0 -> Baixa turbidez
1 -> Alta turbidez.



Temperatura

Varia de 34°C a 90°, se a temperatura estiver no intervalo de 34°C a 45,20°C, considera-se boa.

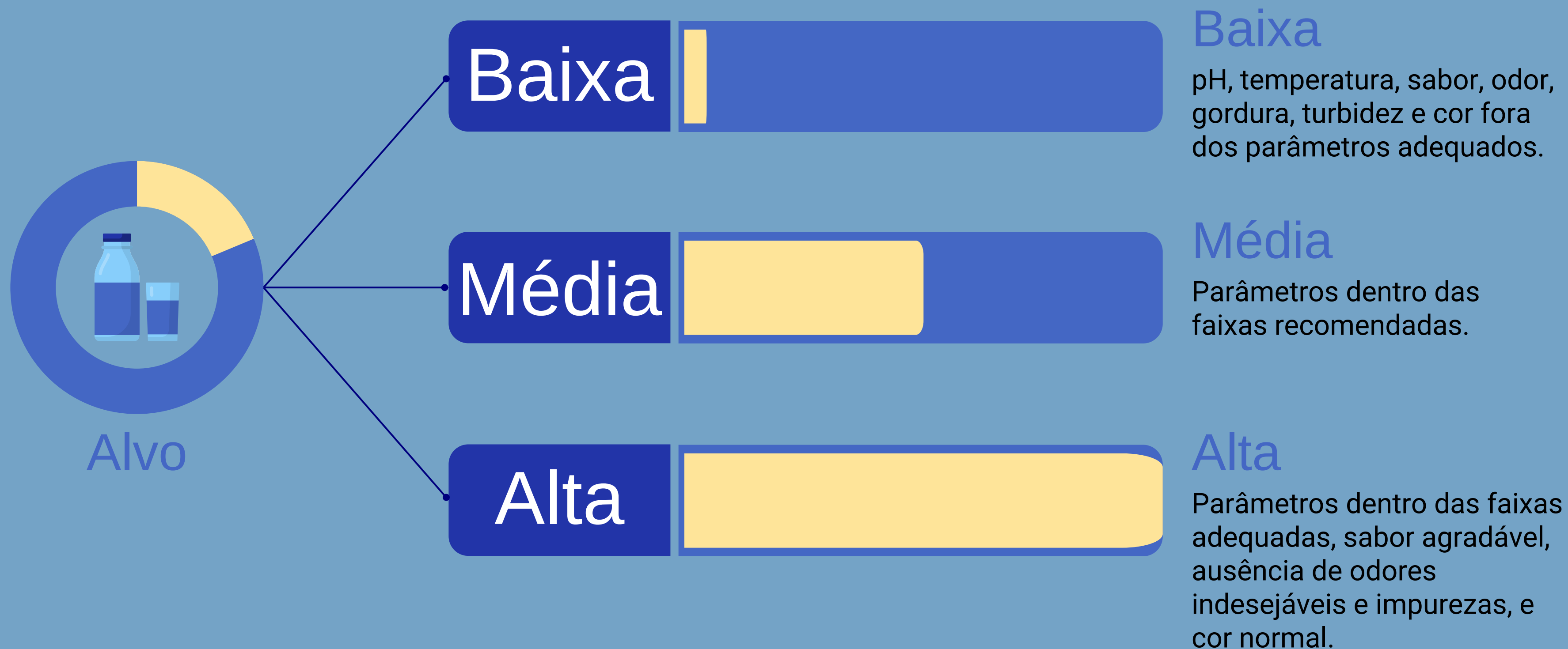


PH

Varia de 3 a 9,5, se estiver no intervalo de 34°C a 45,20°C, considera-se boa.



Atributos Alvo



Seleção e pré-processamento de dados



Informações da Base

INFORMAÇÕES GERAIS DOS DADOS

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 1057 entries, 0 to 1056  
Data columns (total 8 columns):  
#   Column      Non-Null Count  Dtype    
---  -  
0   pH           1057 non-null   float64  
1   Temperature  1057 non-null   int64  
2   Taste        1057 non-null   int64  
3   Odor         1057 non-null   int64  
4   Fat          1057 non-null   int64  
5   Turbidity    1057 non-null   int64  
6   Colour       1057 non-null   int64  
7   Grade        1057 non-null   object  
dtypes: float64(1), int64(6), object(1)  
memory usage: 66.2+ KB  
None
```

VALORES FALTANTES

```
pH           0  
Temperature  0  
Taste        0  
Odor         0  
Fat          0  
Turbidity    0  
Colour       0  
Grade        0  
dtype: int64
```

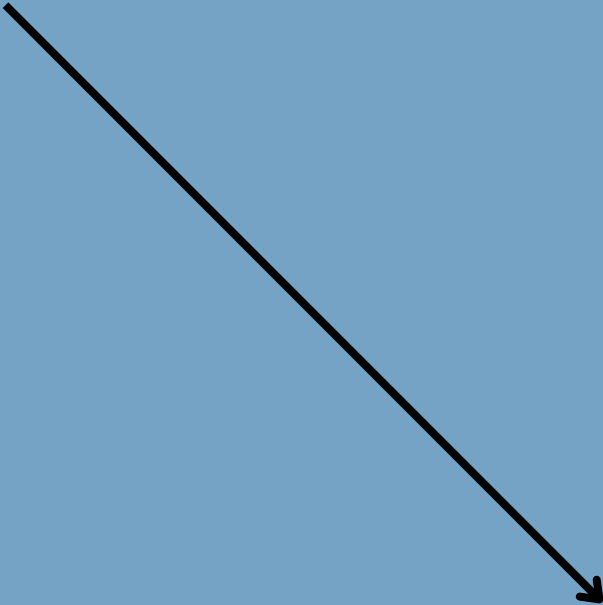
Normalização e redução de dados



2

REDUÇÃO DE DADOS

- Não foi possível reduzir atributos (sao poucos)
- Balanceamento utilizando OVERSAMPLING



Técnica para o desequilíbrio de classes
aumentando artificialmente o numero de
instancias da classe minoritária

Balanceamento da Base

Oversampling

Classe RandomOverSampler
do módulo
imblearn.over_sampling da
biblioteca imbalanced-learn

random_state = 42

```
Quantidade de dados por target antes do balanceamento:  
low      429  
medium   374  
high     254  
Name: Grade, dtype: int64
```

```
Quantidade de dados por target após o balanceamento:  
low      429  
medium   429  
high     429  
Name: Grade, dtype: int64
```

```
sampler = RandomOverSampler(random_state=42)  
X_resampled, y_resampled = sampler.fit_resample(X, y)
```

NORMALIZAÇÃO

- Técnica escolhida = Min-Max (muitos dados binários)

	pH	Temperature	Taste	Odor	Fat	Turbidity	Colour
0	8.5	70	1	1	1	1	246
1	9.5	34	1	1	0	1	255
2	6.6	37	0	0	0	0	255
3	6.6	37	1	1	1	1	255
4	5.5	45	1	0	1	1	250
5	4.5	60	0	1	1	1	250
6	8.1	66	1	0	1	1	255
7	6.7	45	1	1	0	0	247
8	6.7	45	1	1	1	0	245
9	5.6	50	0	1	1	1	255

[illegible]

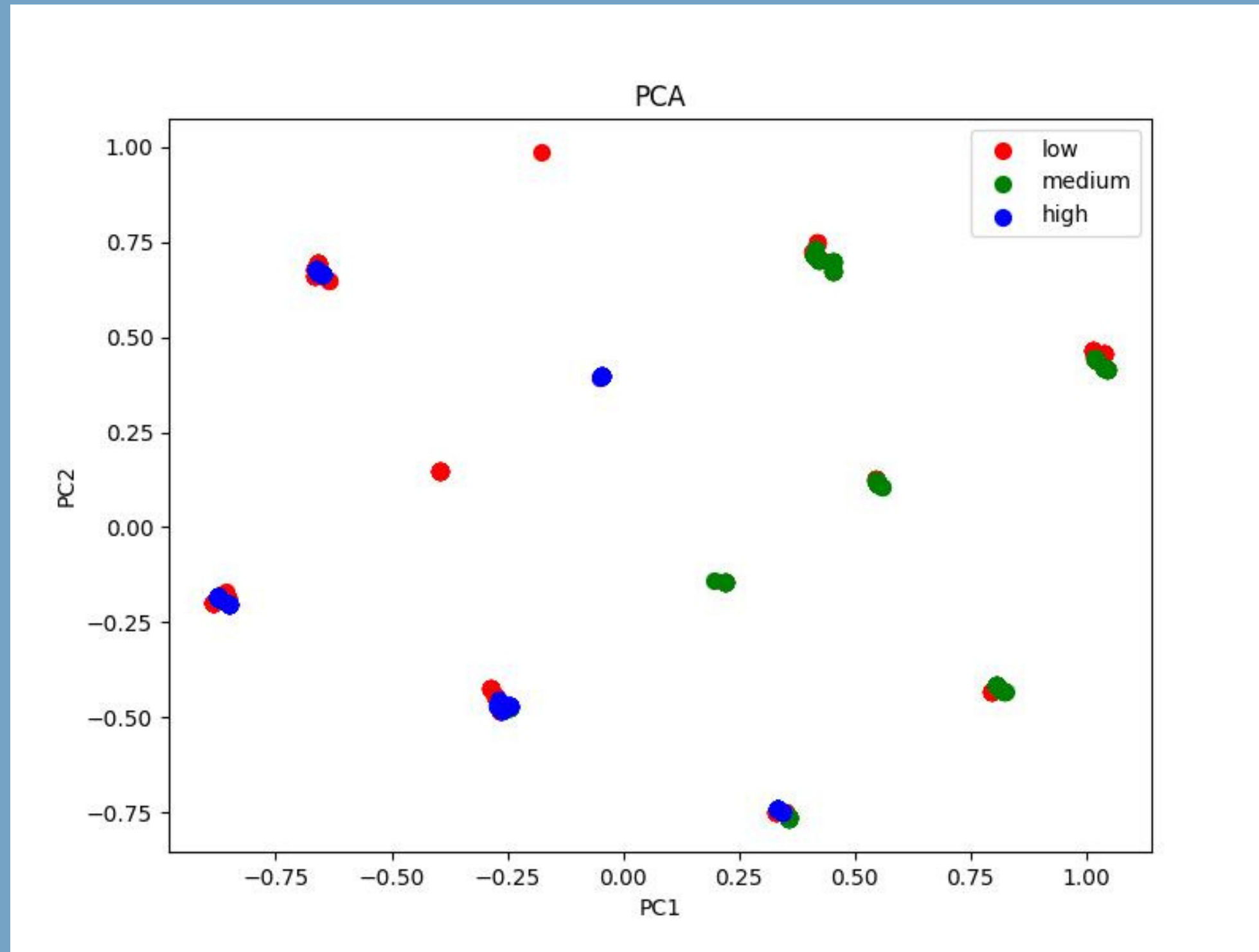
NORMALIZAÇÃO

- Técnica escolhida = Min-Max (muitos dados binarios)

	pH	Temprature	Taste	Odor	Fat	Turbidity	Colour	Grade
0	0.846154	0.642857	1.0	1.0	1.0	1.0	0.400000	low
1	1.000000	0.000000	1.0	1.0	0.0	1.0	1.000000	low
2	0.553846	0.053571	0.0	0.0	0.0	0.0	1.000000	medium
3	0.553846	0.053571	1.0	1.0	1.0	1.0	1.000000	high
4	0.384615	0.196429	1.0	0.0	1.0	1.0	0.666667	low
5	0.230769	0.464286	0.0	1.0	1.0	1.0	0.666667	low
6	0.784615	0.571429	1.0	0.0	1.0	1.0	1.000000	low
7	0.569231	0.196429	1.0	1.0	0.0	0.0	0.466667	medium
8	0.569231	0.196429	1.0	1.0	1.0	0.0	0.333333	medium
9	0.400000	0.285714	0.0	1.0	1.0	1.0	1.000000	low

NORMALIZAÇÃO

PCA -> Encontrar direções ao longo das quais os dados variam

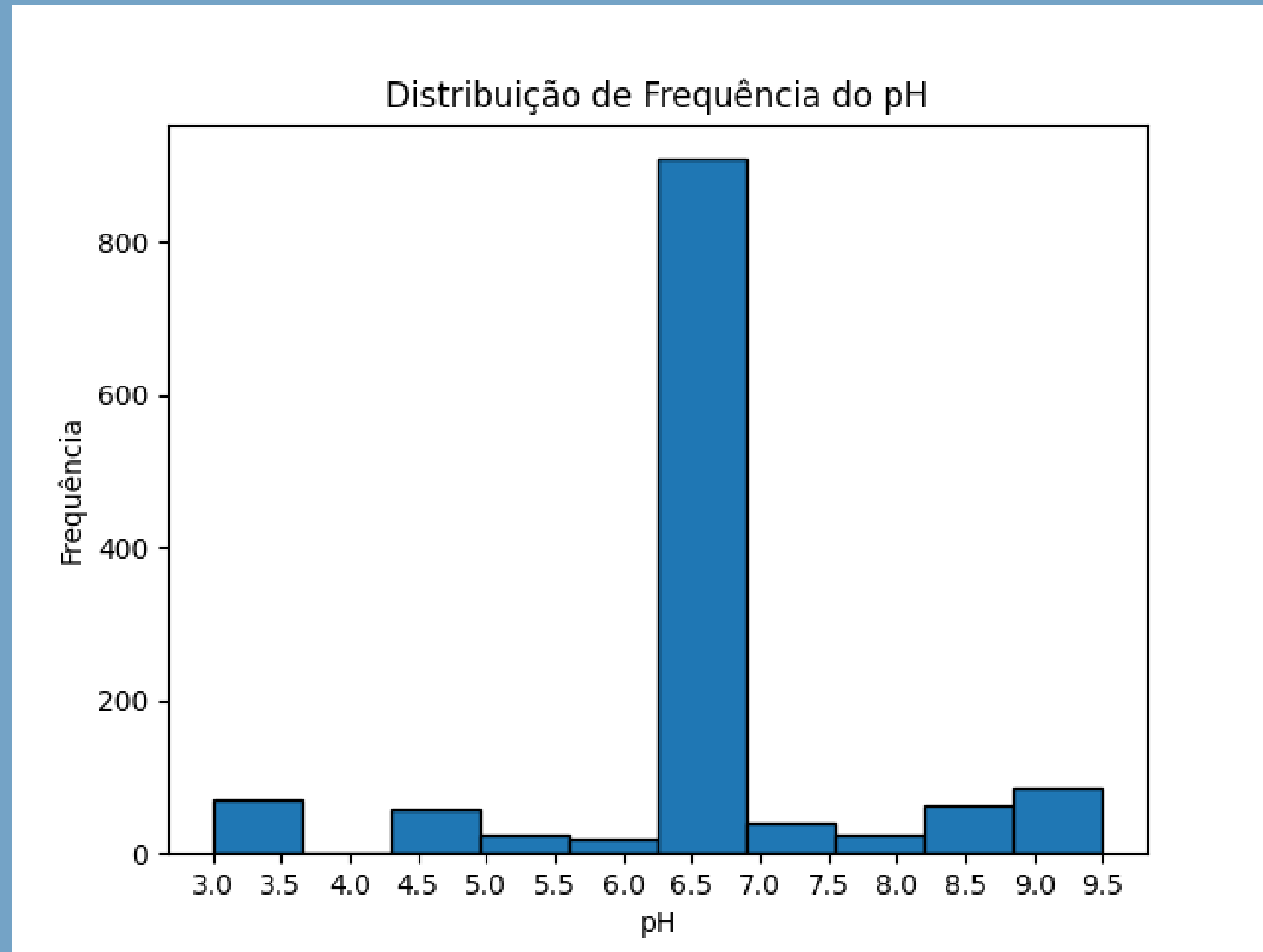


Análise descritiva de dados

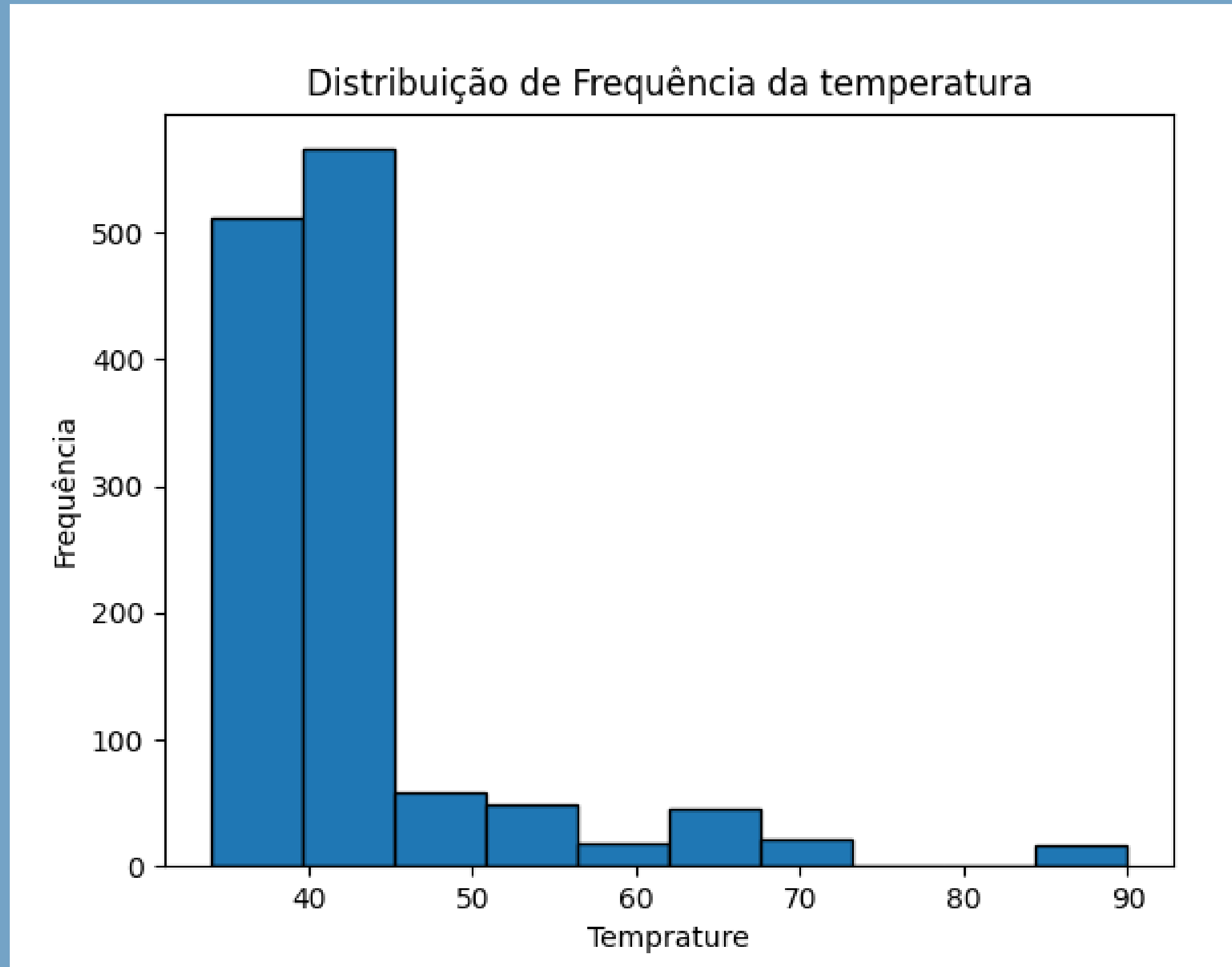
VISUALIZAÇÃO



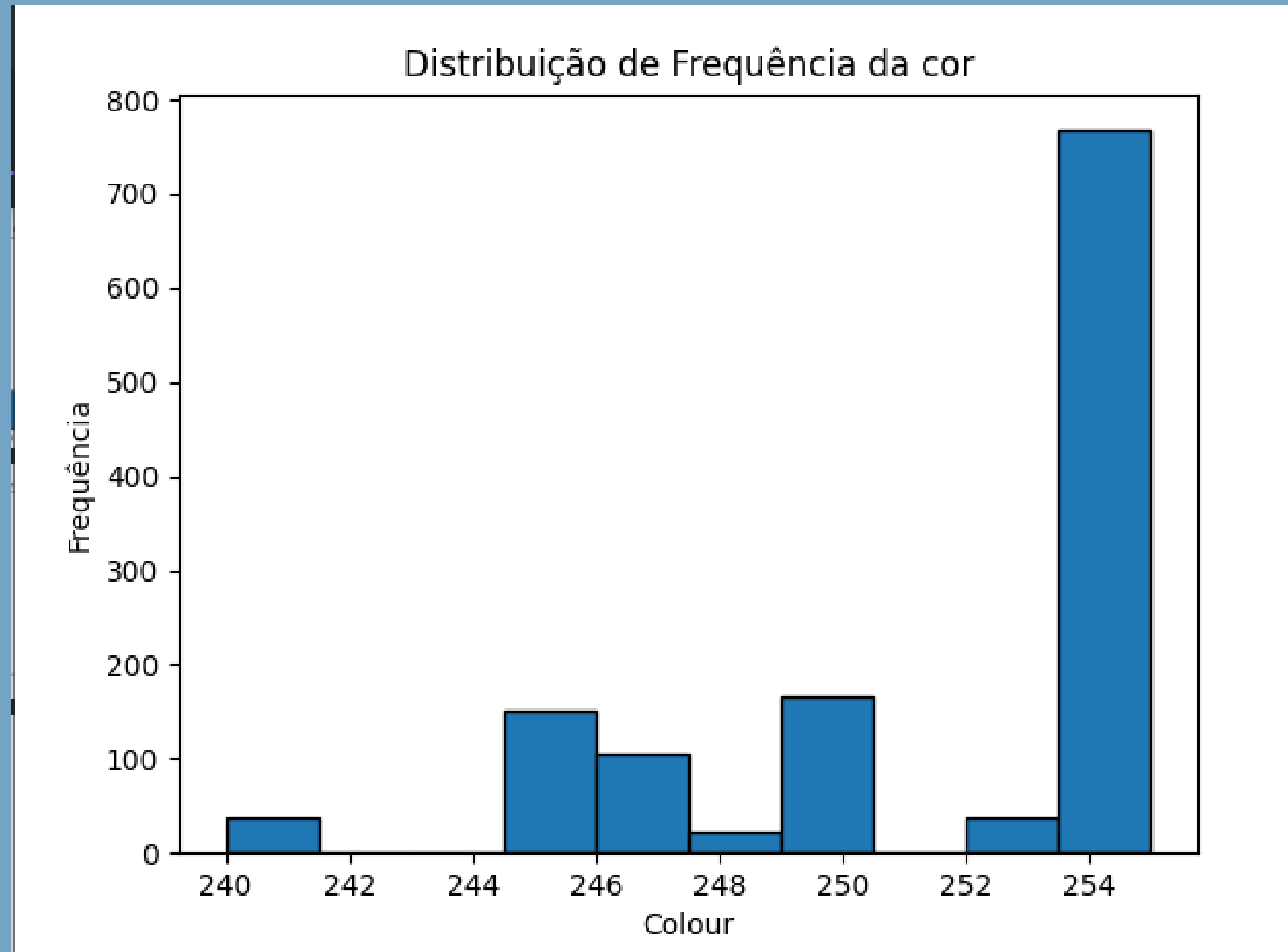
VISUALIZAÇÃO



VISUALIZAÇÃO



VISUALIZAÇÃO



Análise descritiva de dados

MEDIDAS

4

MEDIDAS DE TENDENCIA CENTRAL

MEDIDAS DE TENDENCIA CENTRAL

Temperatura

Media:

43.52836052836053

Mediana:

40.0

Ponto médio:

62.0

Moda:

0 45

Name: Temperature, dtype: int64

pH

Media:

6.639393939393939

Mediana:

6.7

Ponto médio:

6.25

Moda:

0 6.8

Name: pH, dtype: float64

Cor

Media:

251.87412587412587

Mediana:

255.0

Ponto médio:

247.5

Moda:

0 255

Name: Colour, dtype: int64

Gosto

Moda:

0 0

Name: Odor, dtype: int64

Odor

Moda:

0 1

Name: Taste, dtype: int64

Gordura

Moda:

0 1

Name: Fat, dtype: int64

Turbidez

Moda:

0 0

Name: Turbidity, dtype: int64

MEDIDAS DE DISPERSÃO

MEDIDAS DE DISPERSAO

Temperatura

Amplitude:

Desvio padrão:

0.09417428452771771

Variância:

88.6879586630753

Coeficiente de variação:

21.635155421569177

pH

Amplitude:

6.5

Desvio padrão:

0.01270680475317081

Variância:

1.614628870352043

Coeficiente de variação:

19.138500997473155

Cor

Amplitude:

15

Desvio padrão:

0.043051931091782646

Variância:

18.534687707316014

MEDIDAS DE POSIÇÃO RELATIVA

```
-----  
MEDIDAS DE POSICAO RELATIVA  
-----  
Temperatura  
Quartis:  
0.25    38.0  
0.50    40.0  
0.75    45.0  
Name: Temprature, dtype: float64  
Escore-z:  
0         2.810920  
1        -1.011779  
2        -0.693221  
3        -0.693221  
4         0.156268  
  
...  
1282      -0.268477  
1283      -0.799407  
1284      -0.799407  
1285       0.156268  
1286      -0.587035
```

MEDIDAS DE ASSOCIAÇÃO

Gráfico de Dispersão - Cor vs Turbidez

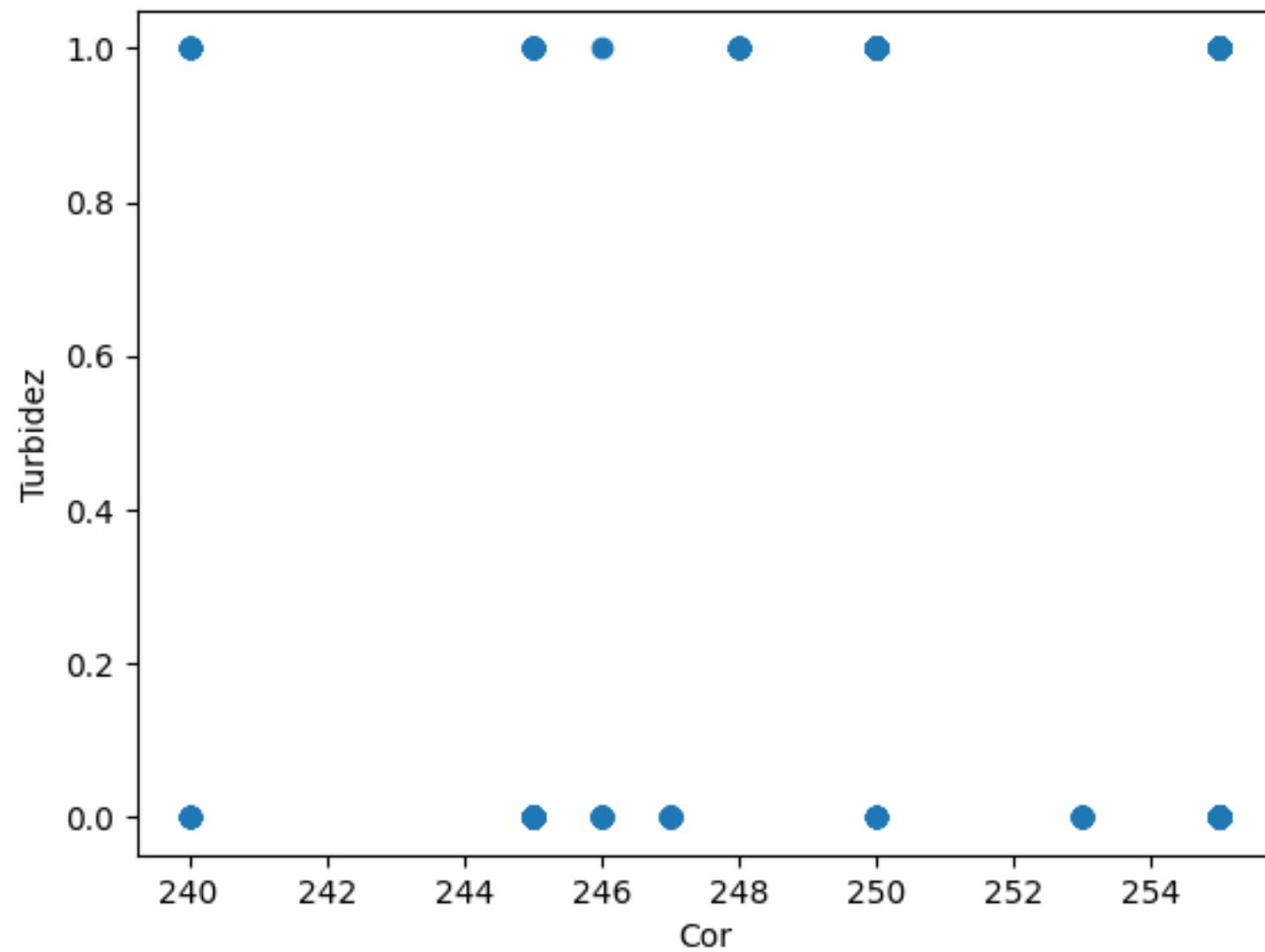
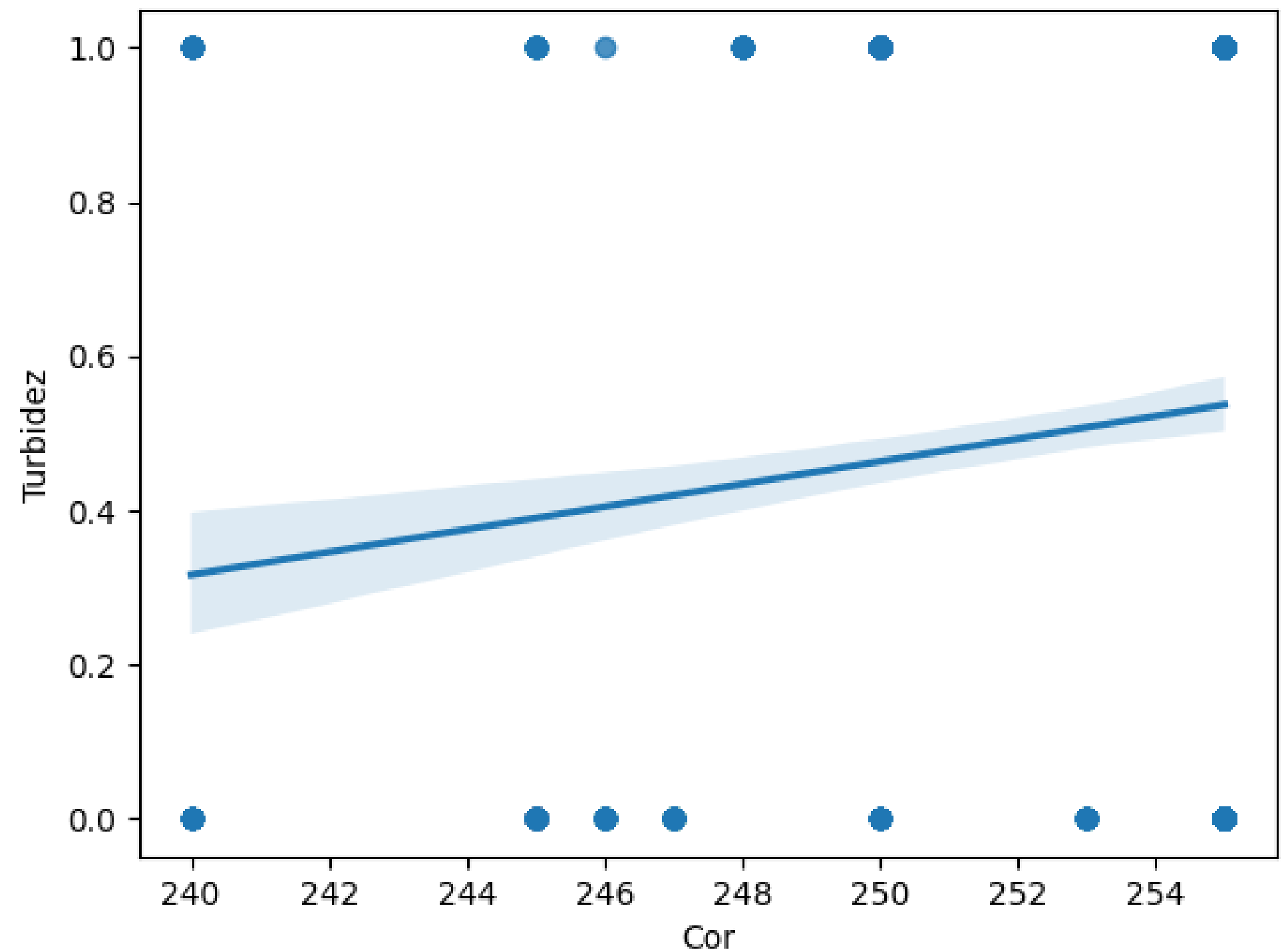


Gráfico de Dispersão - Cor vs Turbidez



MEDIDAS DE ASSOCIAÇÃO

MEDIDAS DE ASSOCIACAO

Covariância entre Cor e Turbidez:

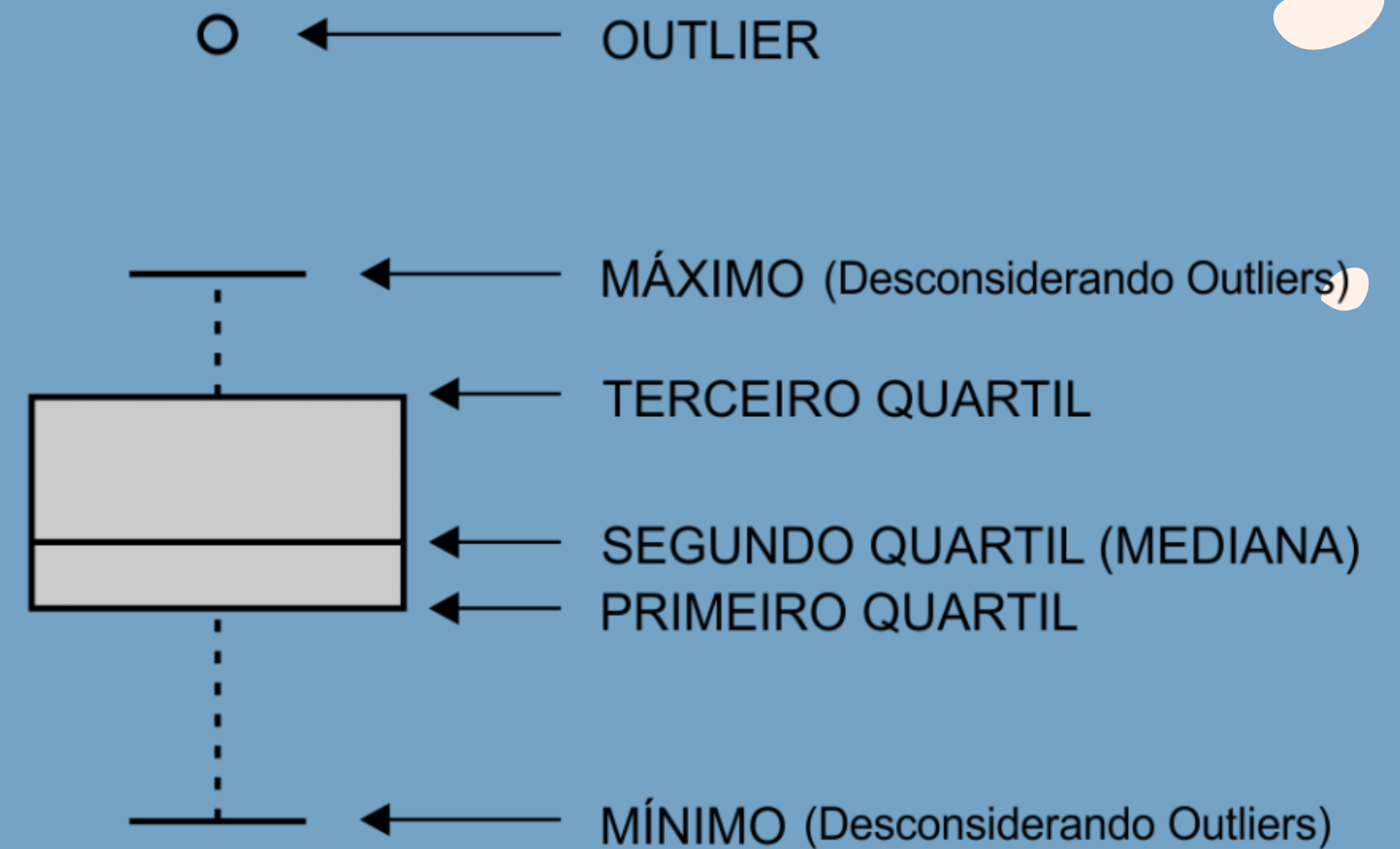
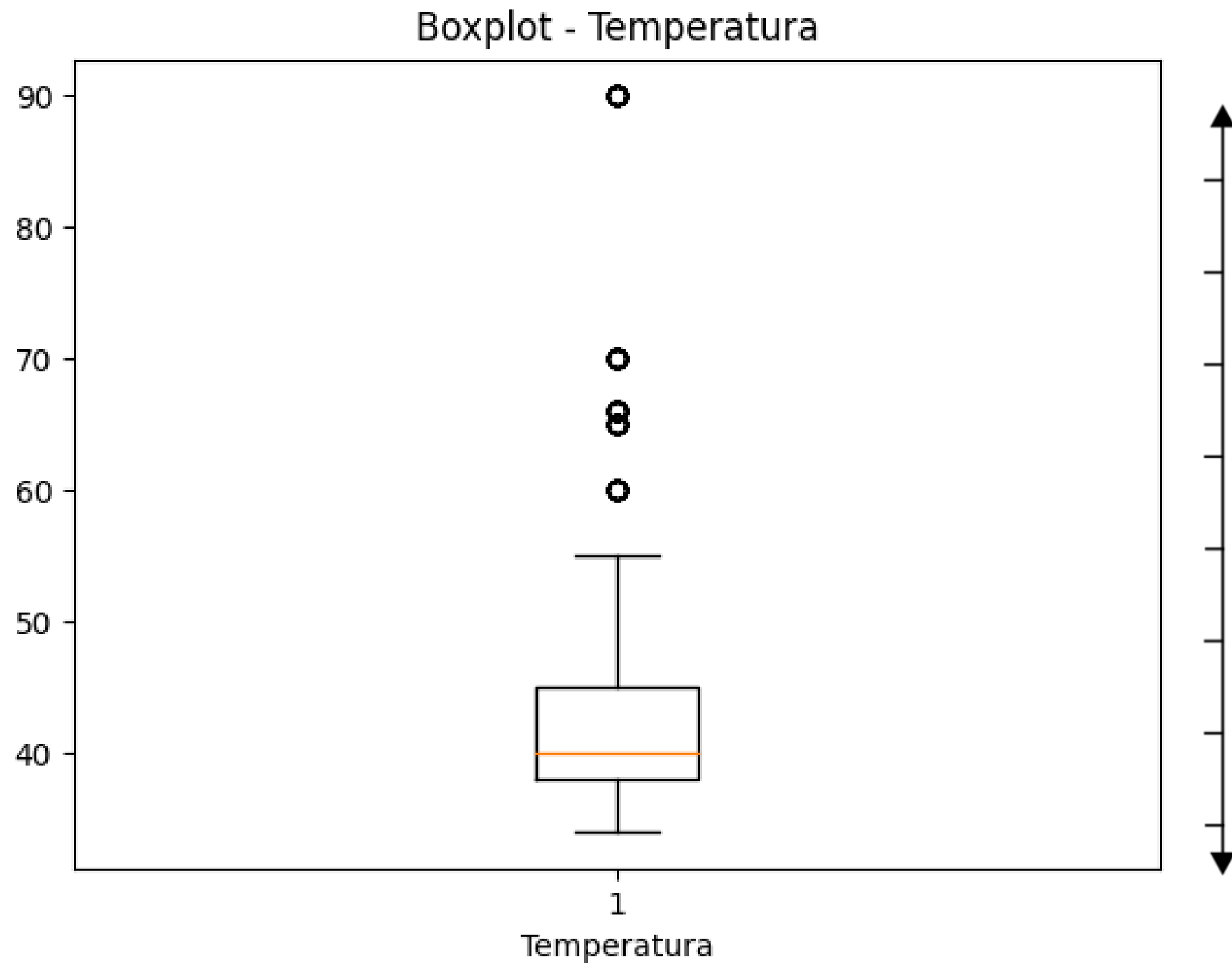
0.2725913278012814

Correlação entre Cor e Turbidez:

0.12660474045084952

MEDIDAS

Diagrama de Caixa

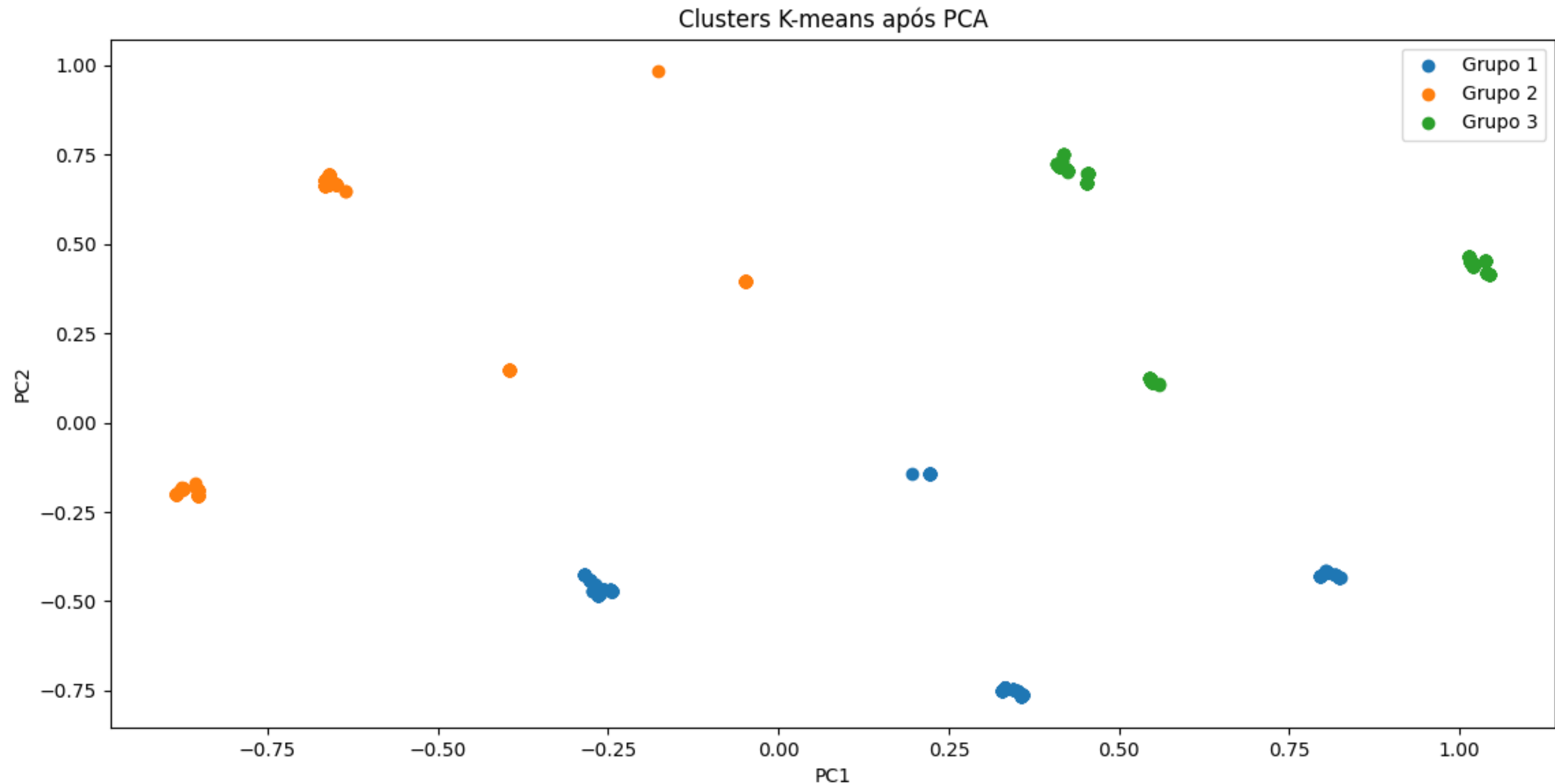


Análise de grupos

5

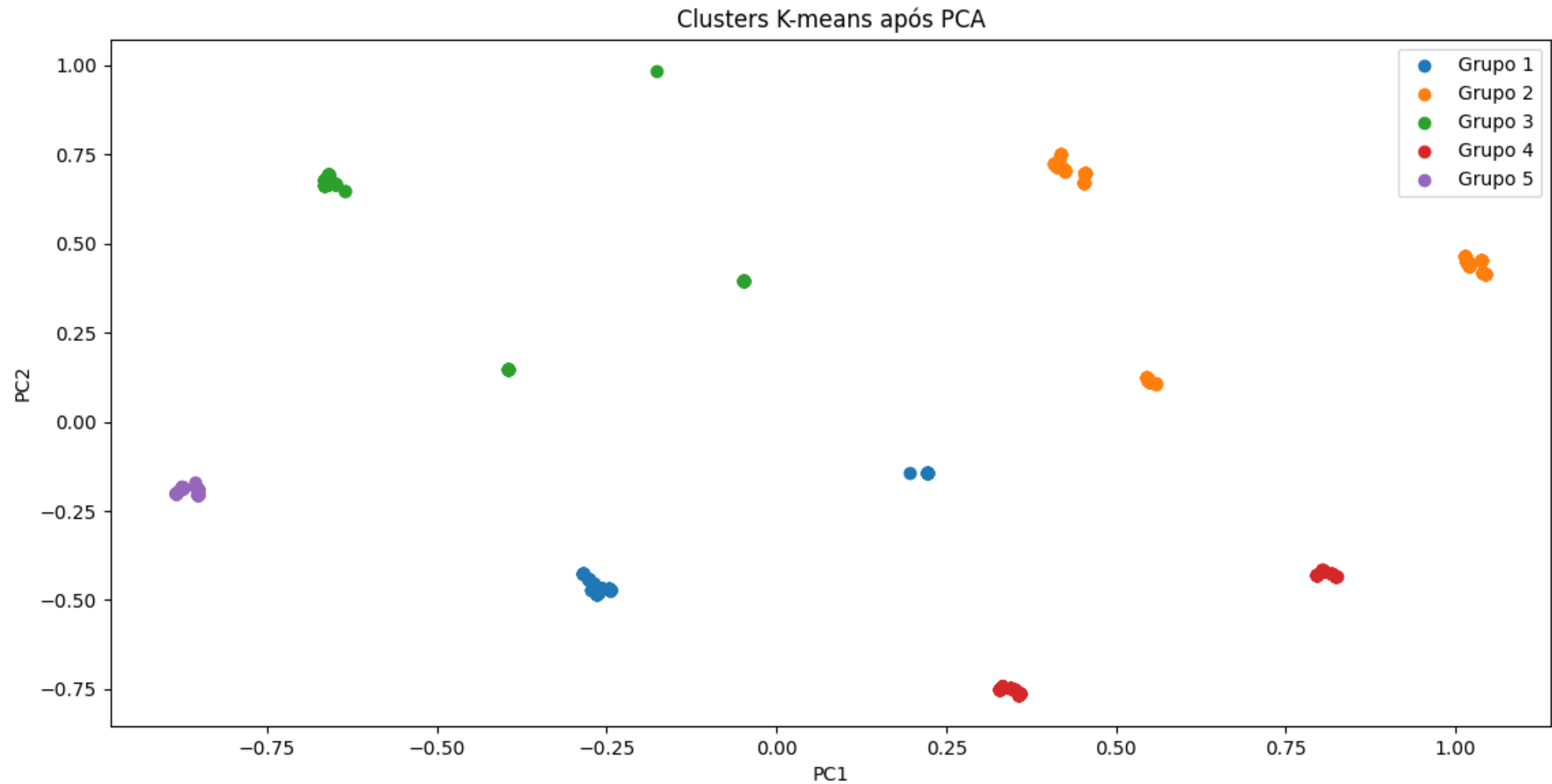
K-MENS - Distancia euclidiana

K = 3



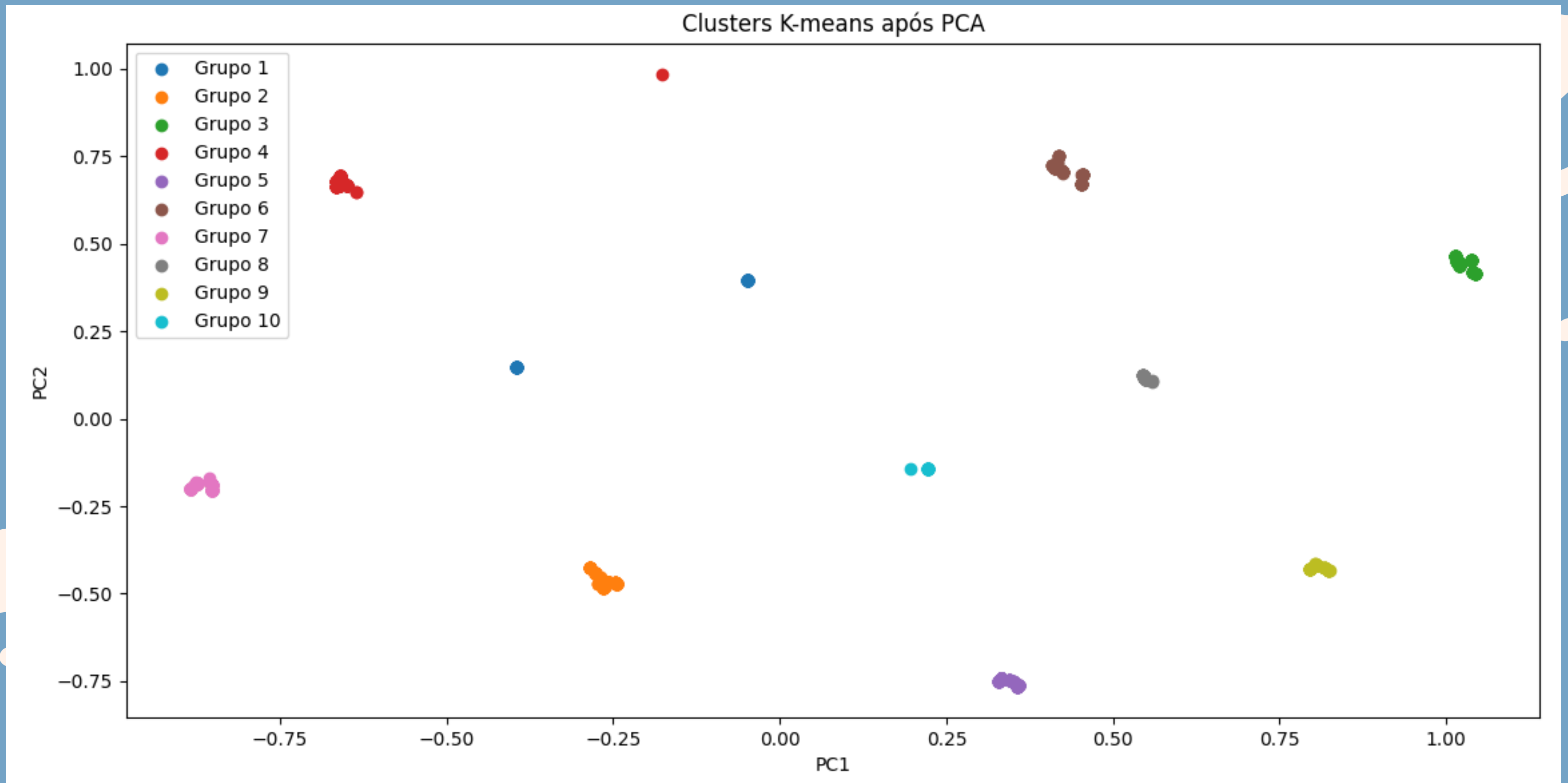
K-MENS - Distancia euclidiana

K = 5



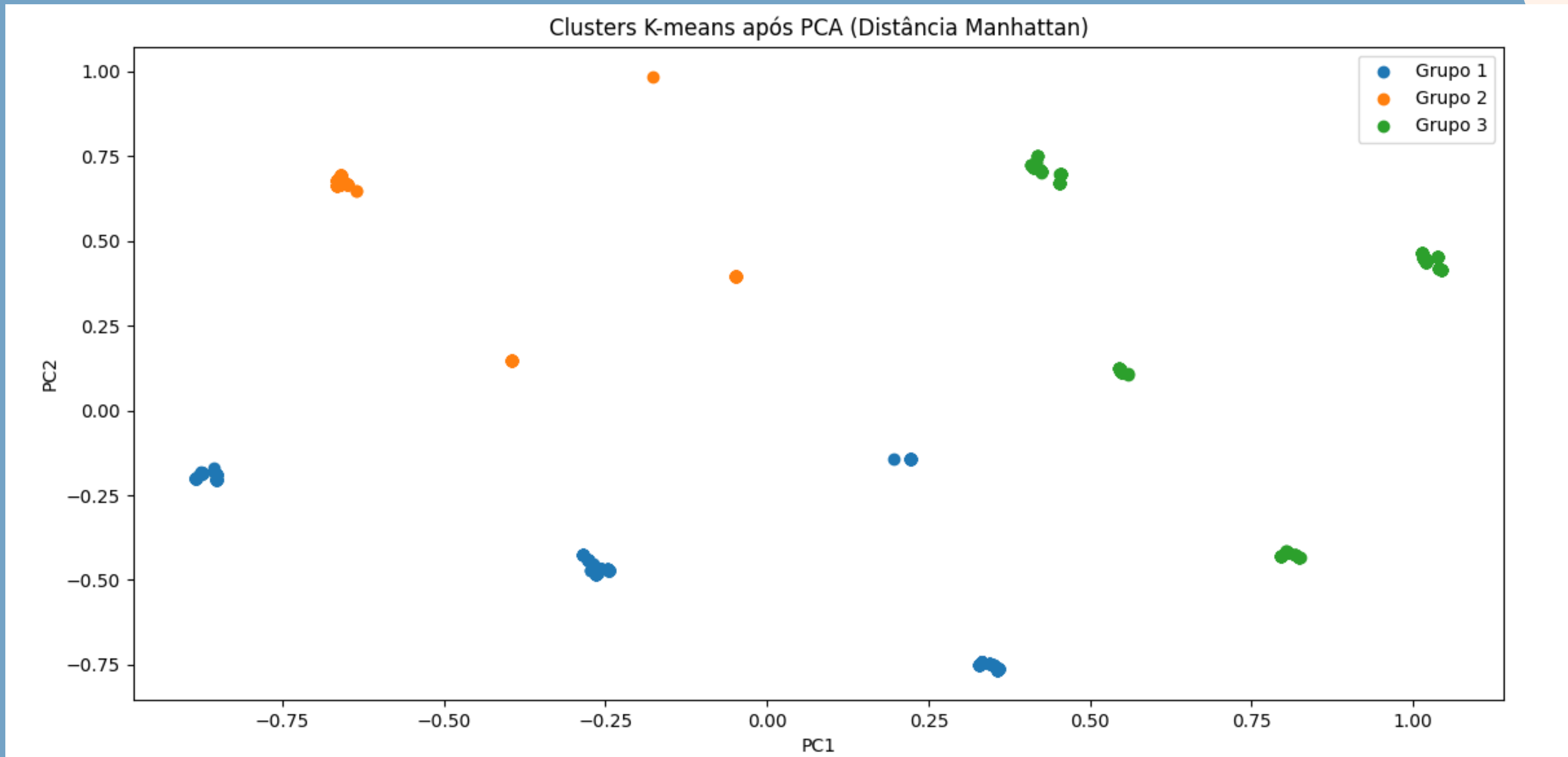
K-MENS - Distancia euclidiana

K = 10



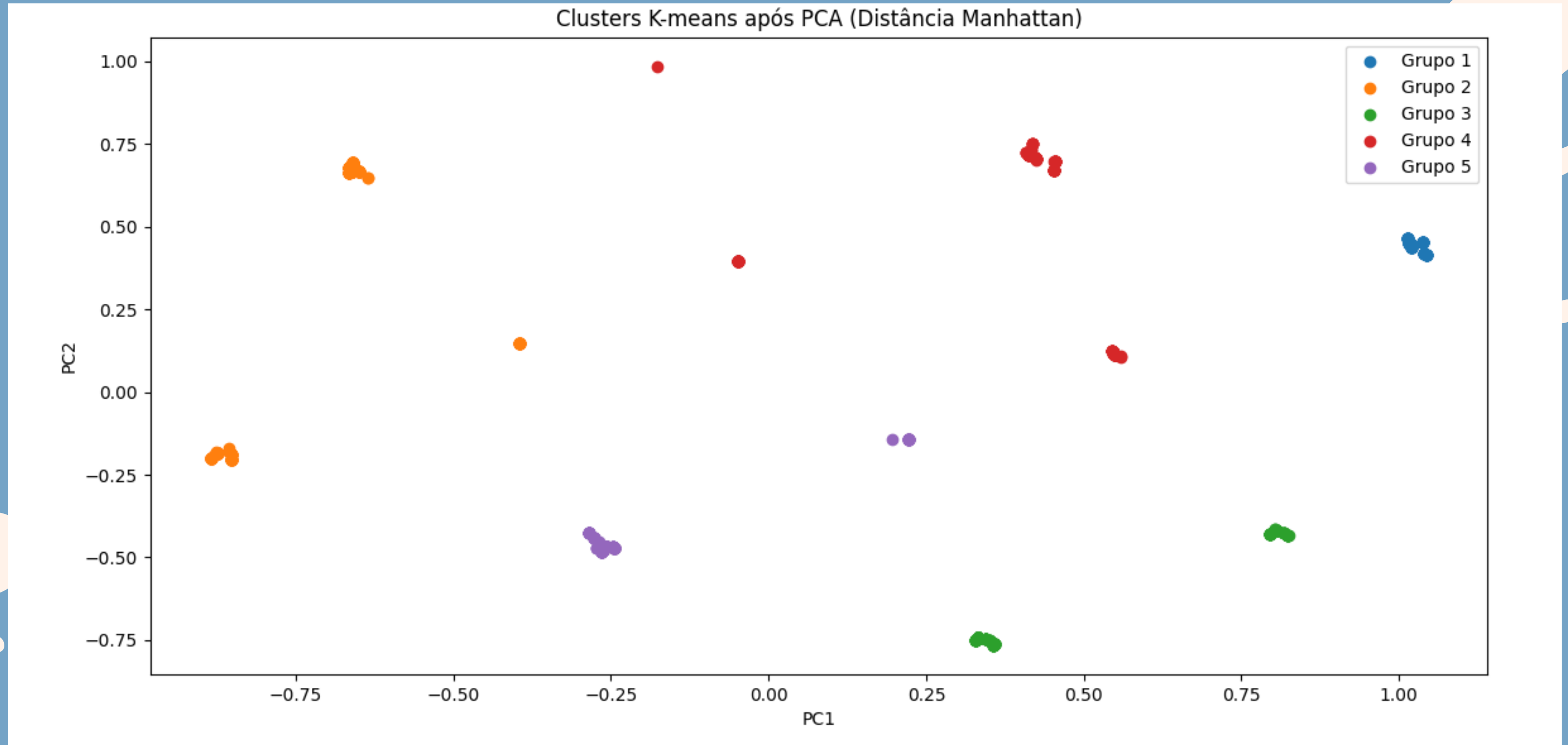
K-MENS - Distancia Manhattan

K = 3



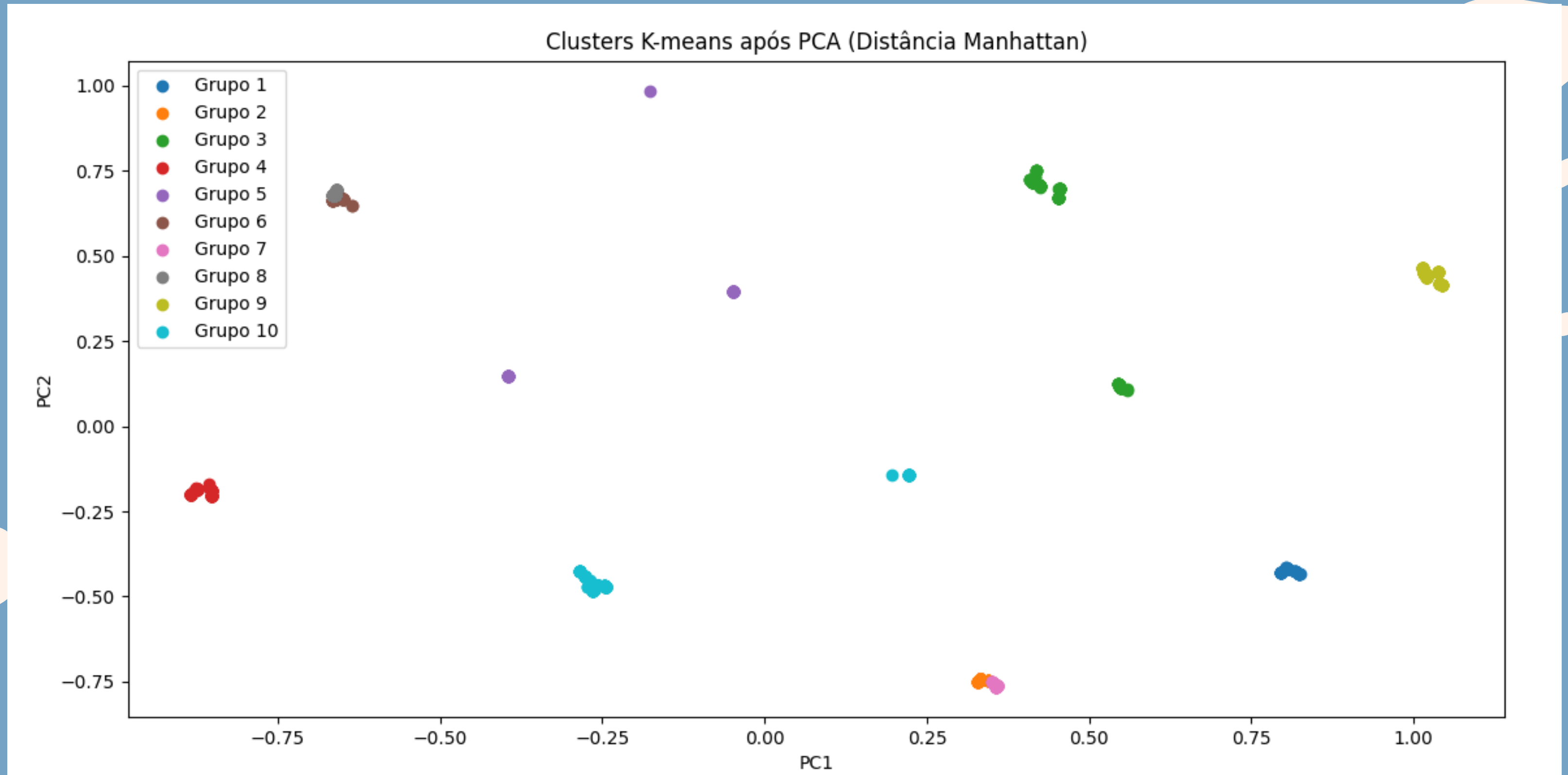
K-MENS - Distancia Manhattan

K = 5



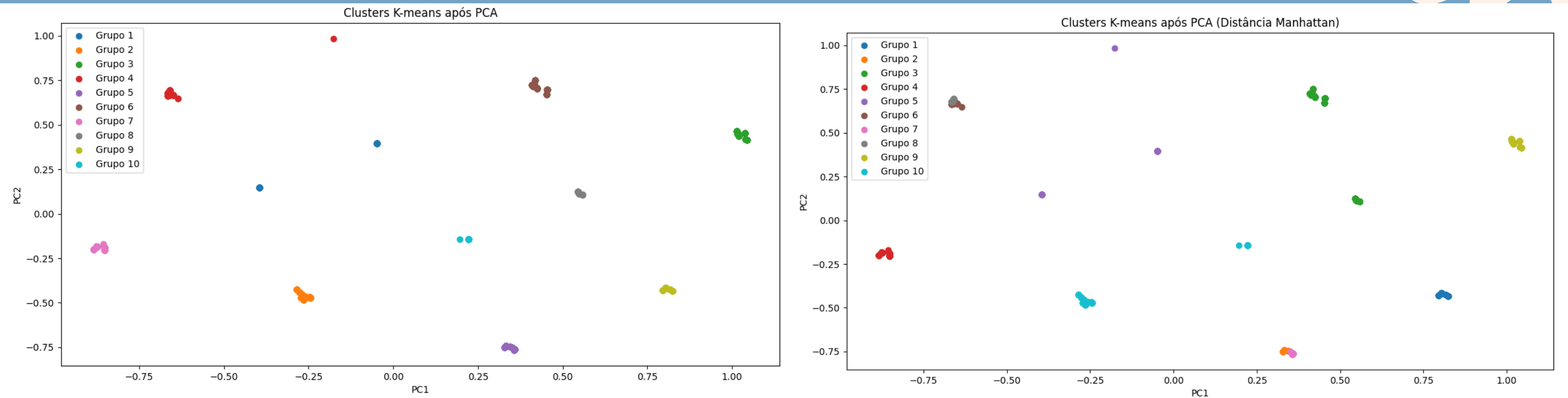
K-MENS - Distancia Manhattan

K = 10



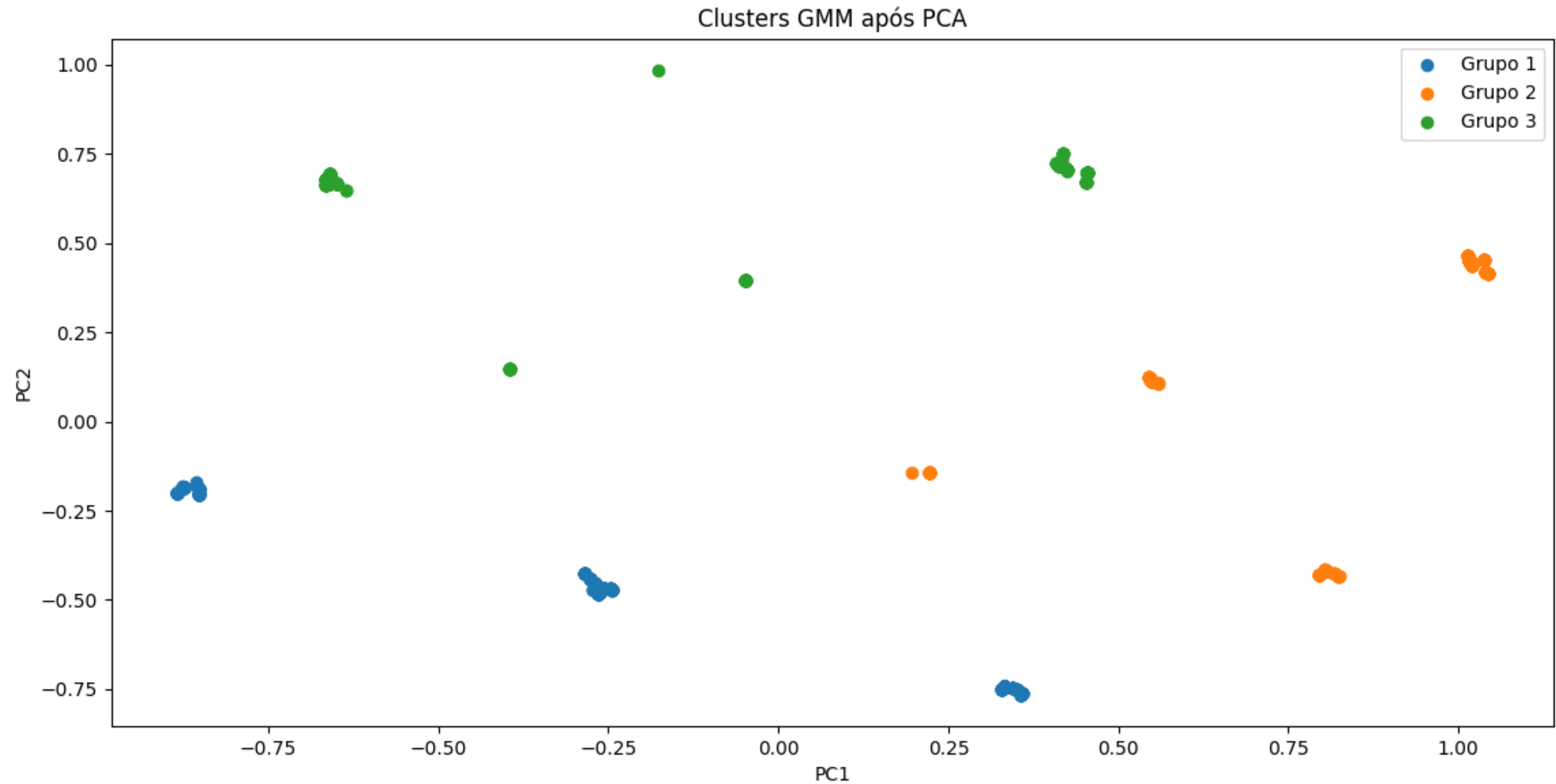
K-MENS - Euclidiana vs Manhattan

K = 10



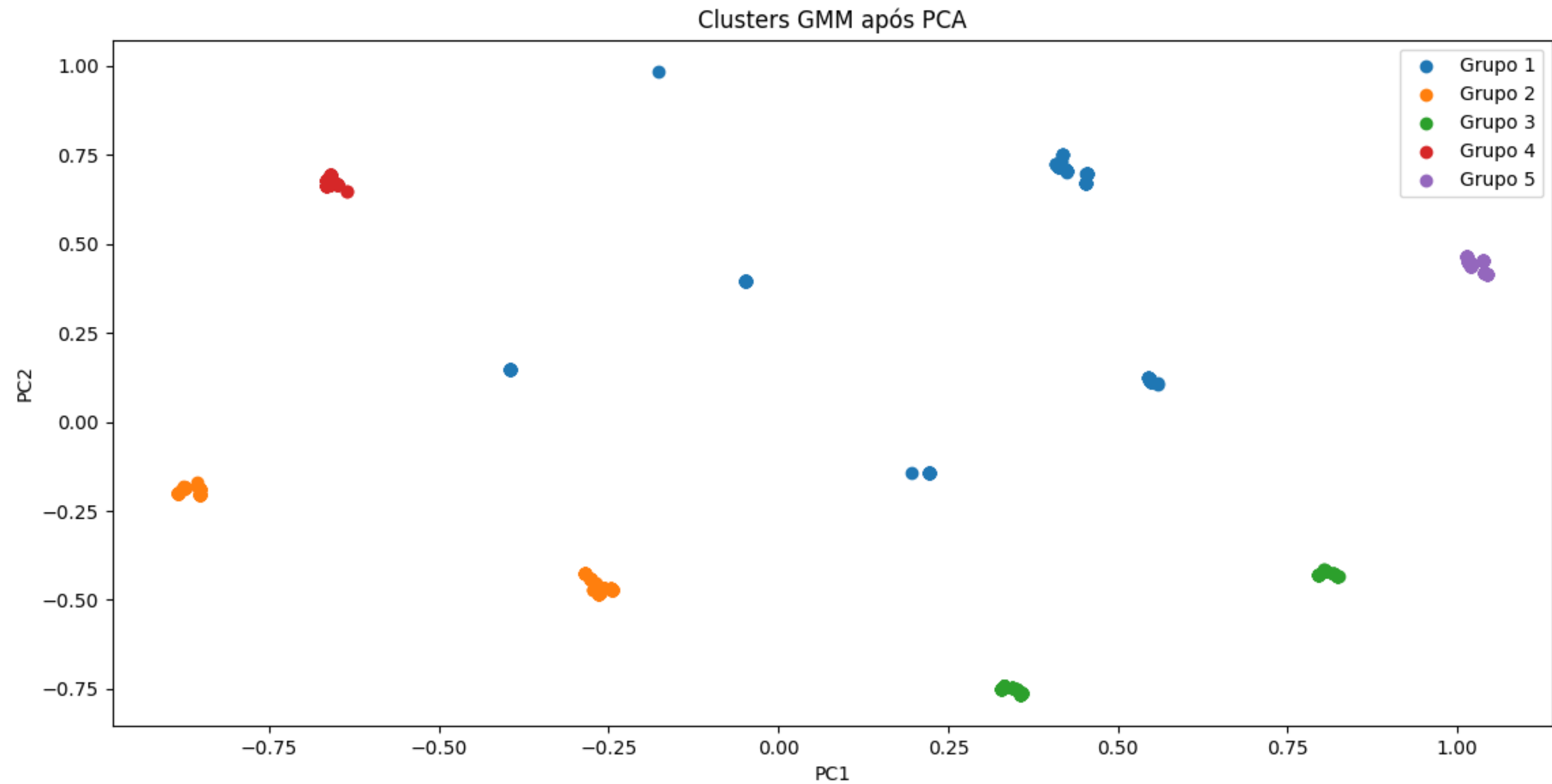
GMM

$K = 3$



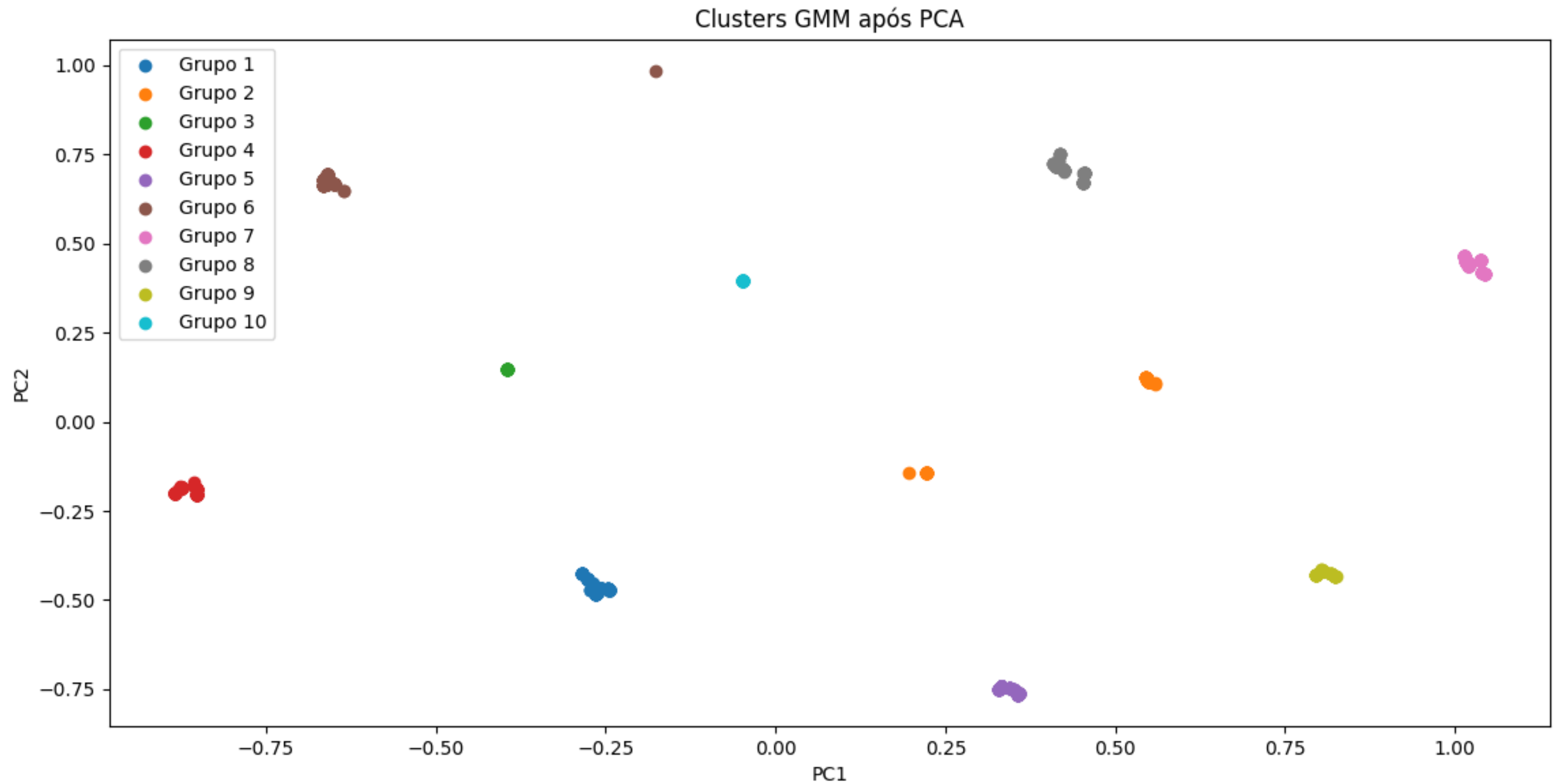
GMM

$K = 5$



GMM

$K = 10$



QUALIDADE DOS AGRUPAMENTOS

Coeficiente de forma

Kmeans - Euclidana

K3 = 0.40246724529144123

K5 = 0.20268239032562355

K10 = 0.023440377867055884

Kmeans - Manhattan

K3 = 0.45895422253875

K5 = 0.224443811142217

K10 = 0.1038507034678613

GMM

K3 = 0.4105691464559676

K5 = 0.3373034798290252

K10 = 0.09551553478201857

QUALIDADE DOS AGRUPAMENTOS

HOMOGENIDADE

Kmeans - Euclidana

K3 = 0.2766691913417482

K5 = 0.30356149881449795

K10 = 0.36749783082413234

Kmeans - Manhattan

K3 = 0.24936000363501326

K5 = 0.24441188763277777

K10 = 0.35451597076560104

GMM

K3 = 0.20860604862337456

K5 = 0.2620263757465329

K10 = 0.36749783082413245

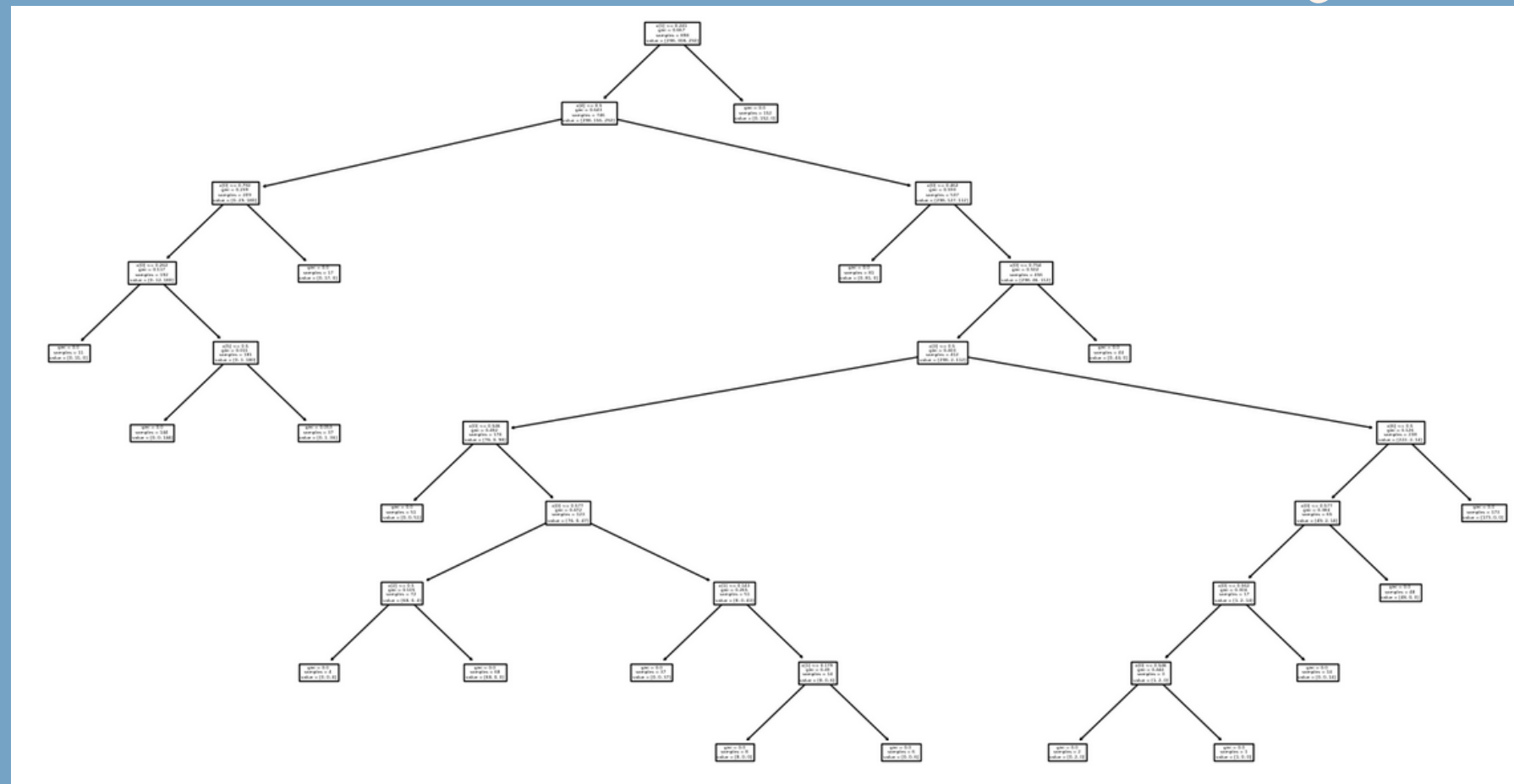
Classificadores



Árvore de Decisão

6.1

Árvore de Decisão

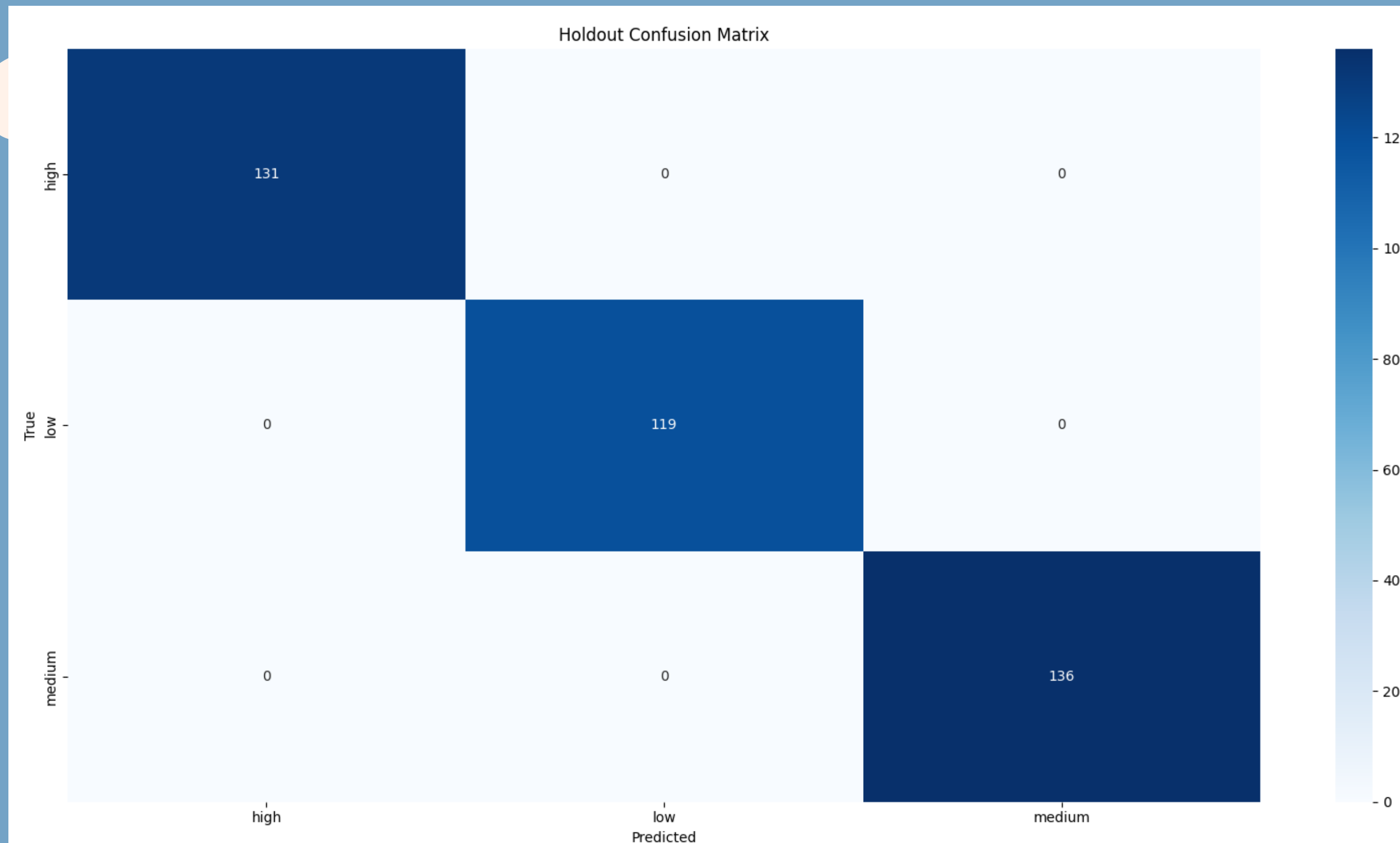


Classe


DecisionTreeClassifier
do módulo sklearn.tree
do scikit-learn.

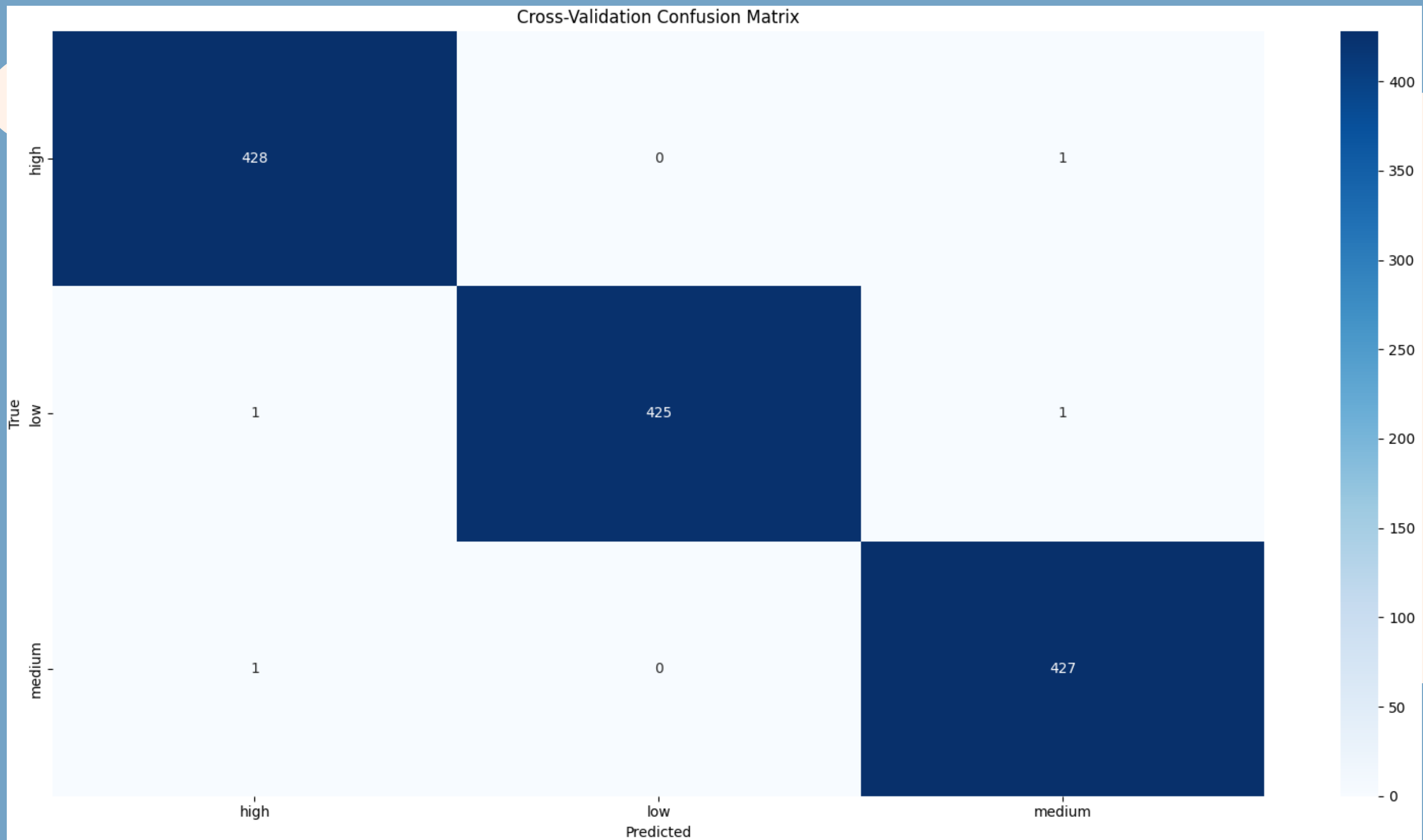
Base de dados
Pré-processada e
Balanceada

18 nós



 Acuracia:100%

 F1-Score: 1



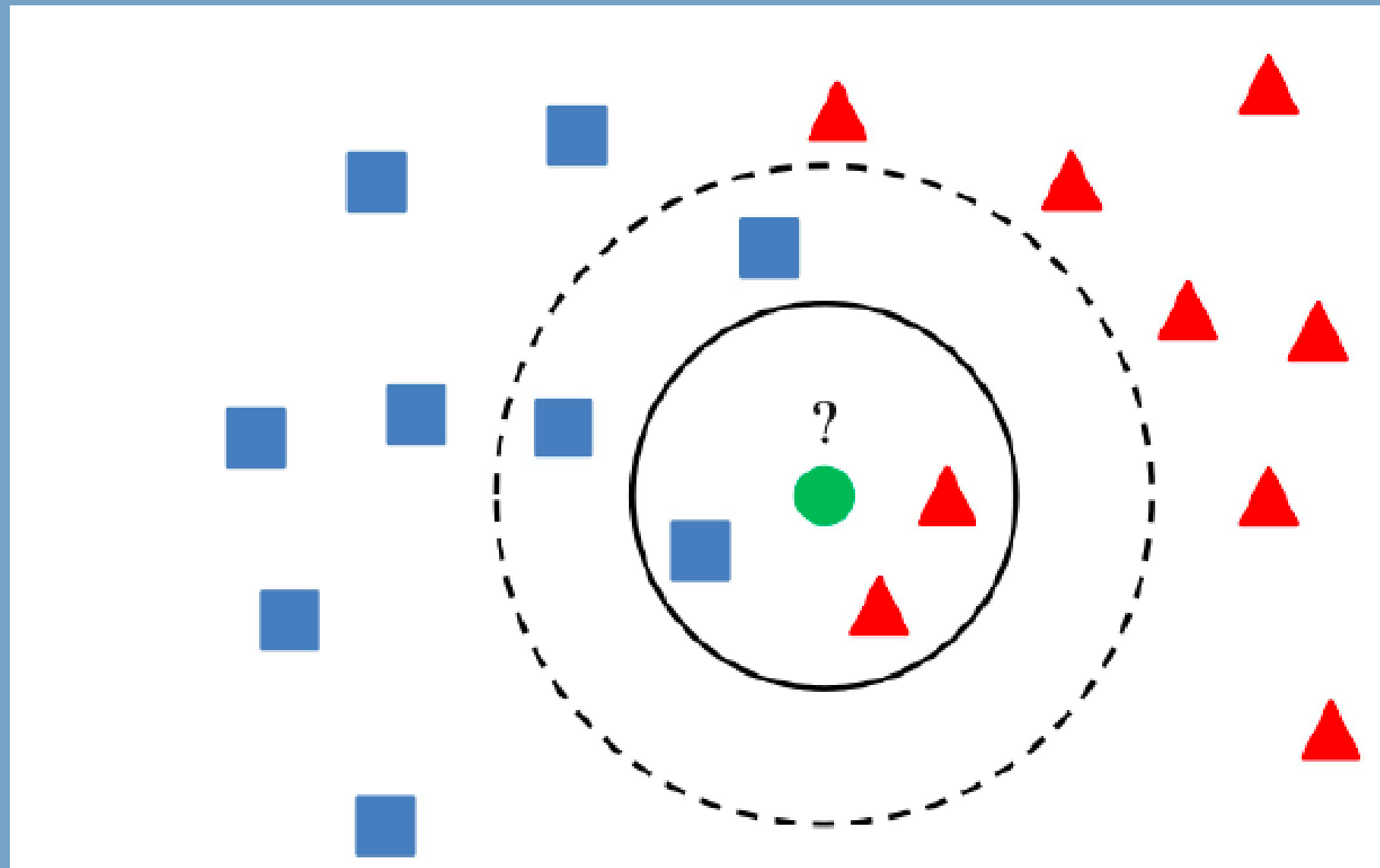
 Acuracia: 99,69%

 F1-Score: 0.99688

K-Nearest Neighbors (KNN)

6.2

KNN

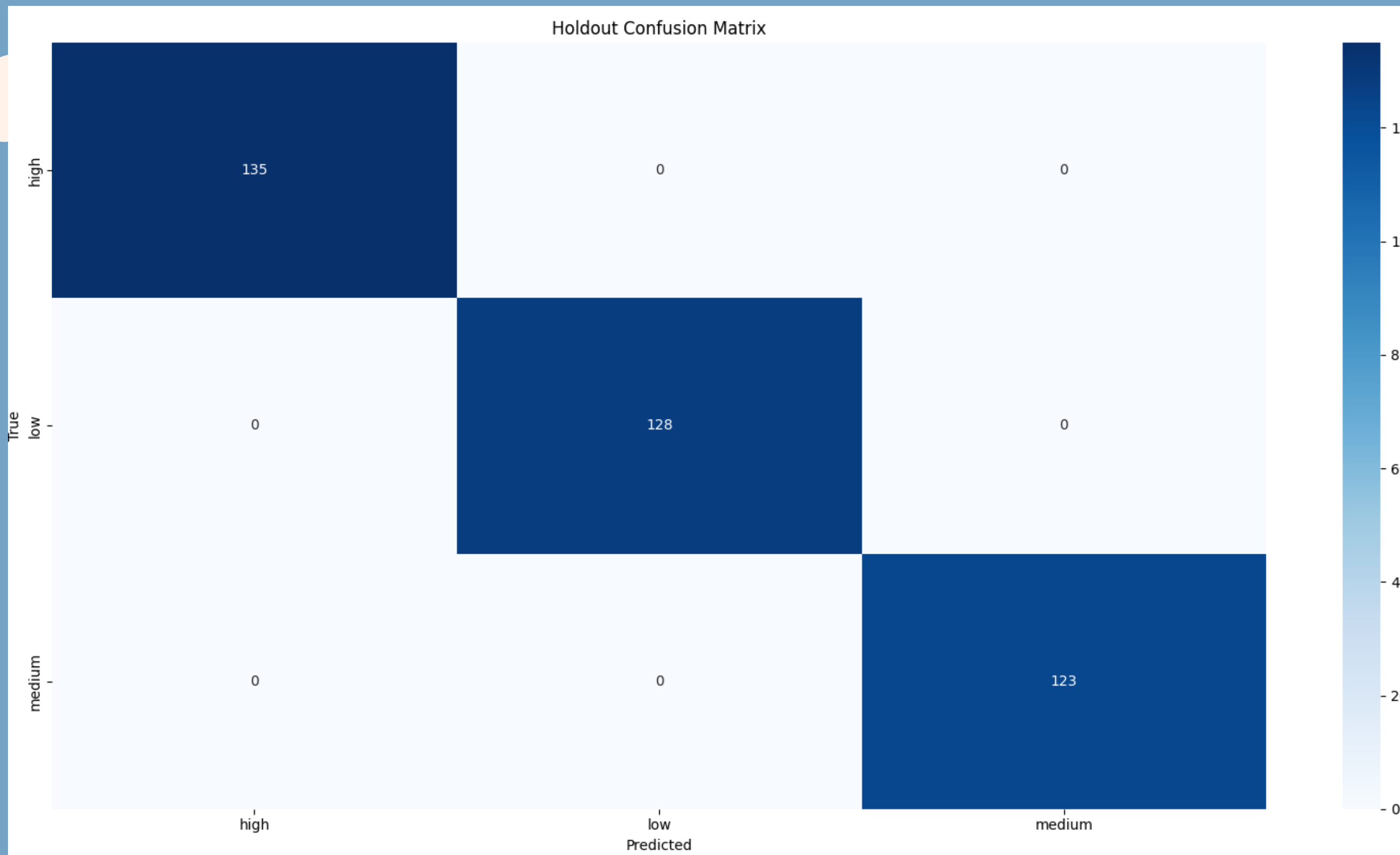


Classe


KNeighborsClassifier do
módulo sklearn.neighbors
do scikit-learn.

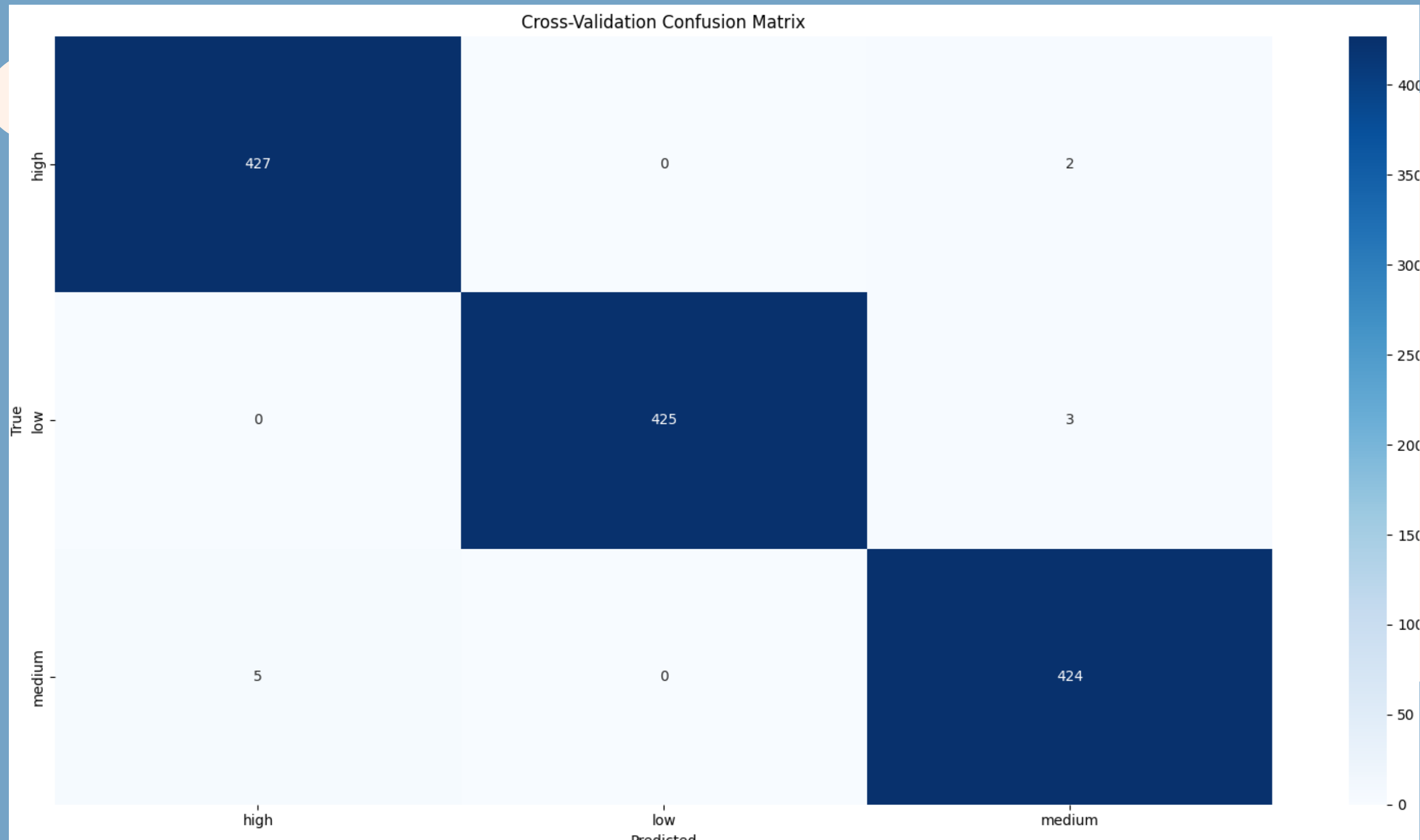
Base de dados Pré-
processada e
Balanceada

K=5



 Acuracia: 100%

 F1-Score: 1



 Acuracia:99.1437%

 F1-Score: 0.9915

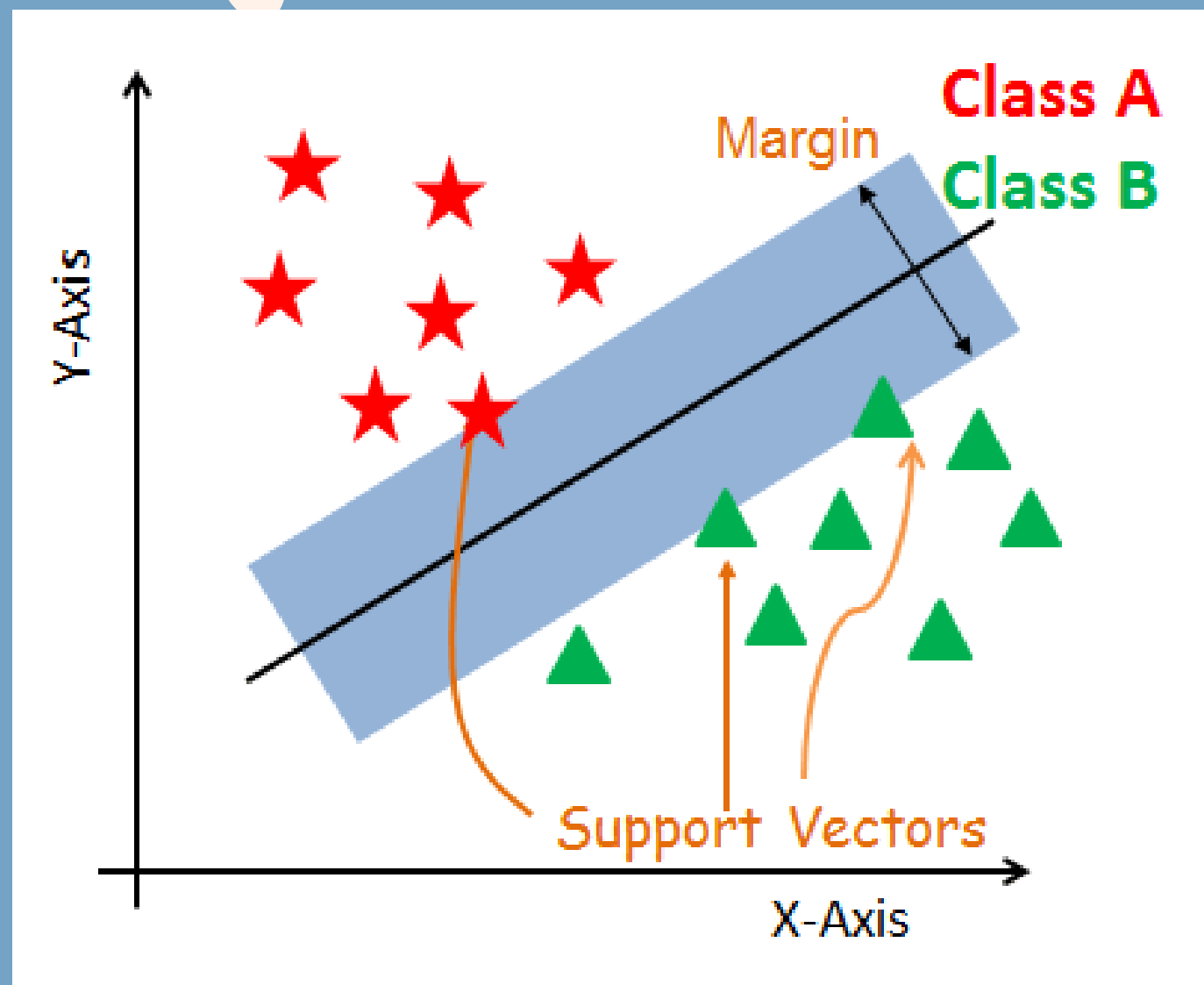
Support Vectors Machine (SVM)



6.3



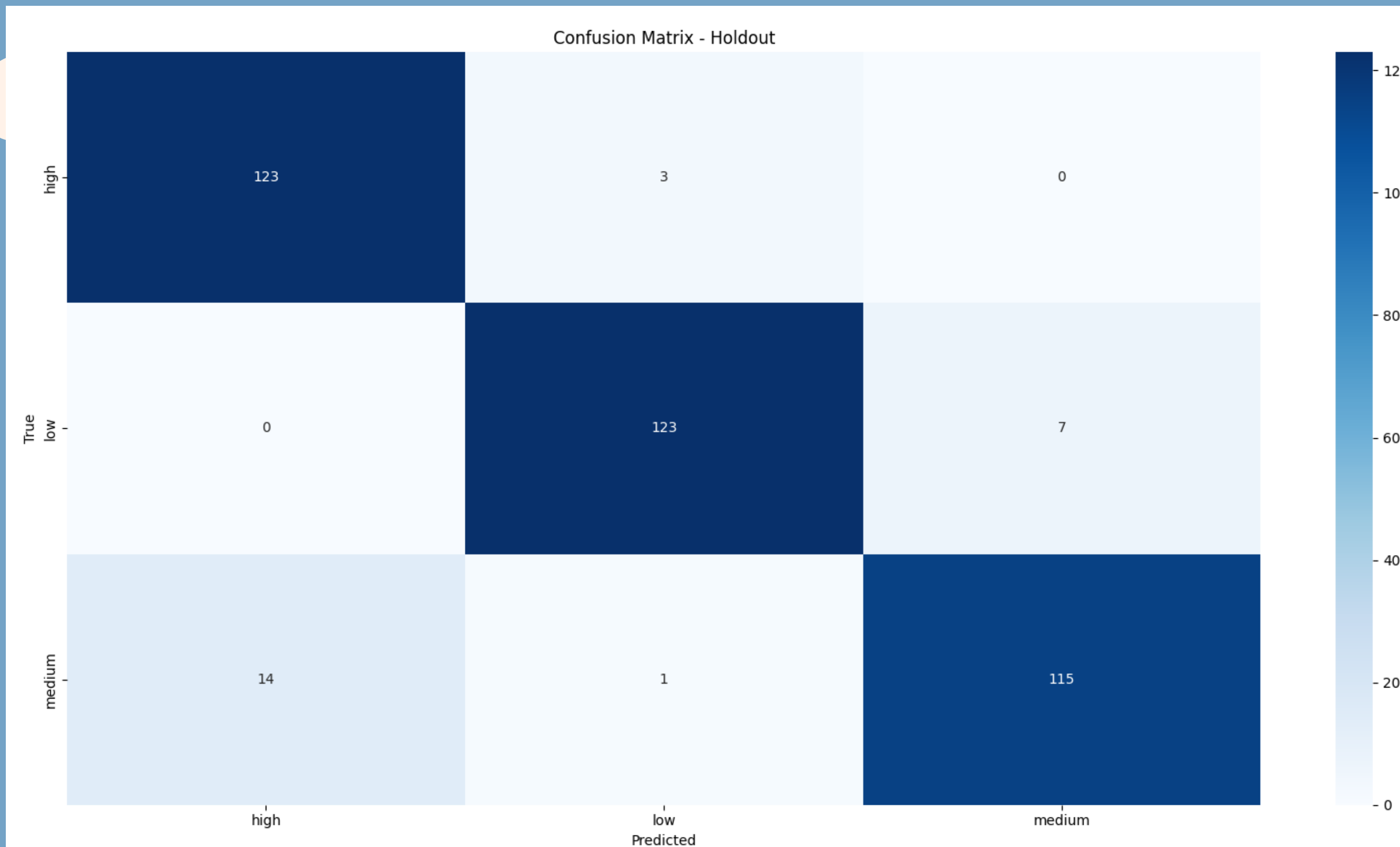
SVM



Classe SVC do módulo
sklearn.svm do scikit-
learn.

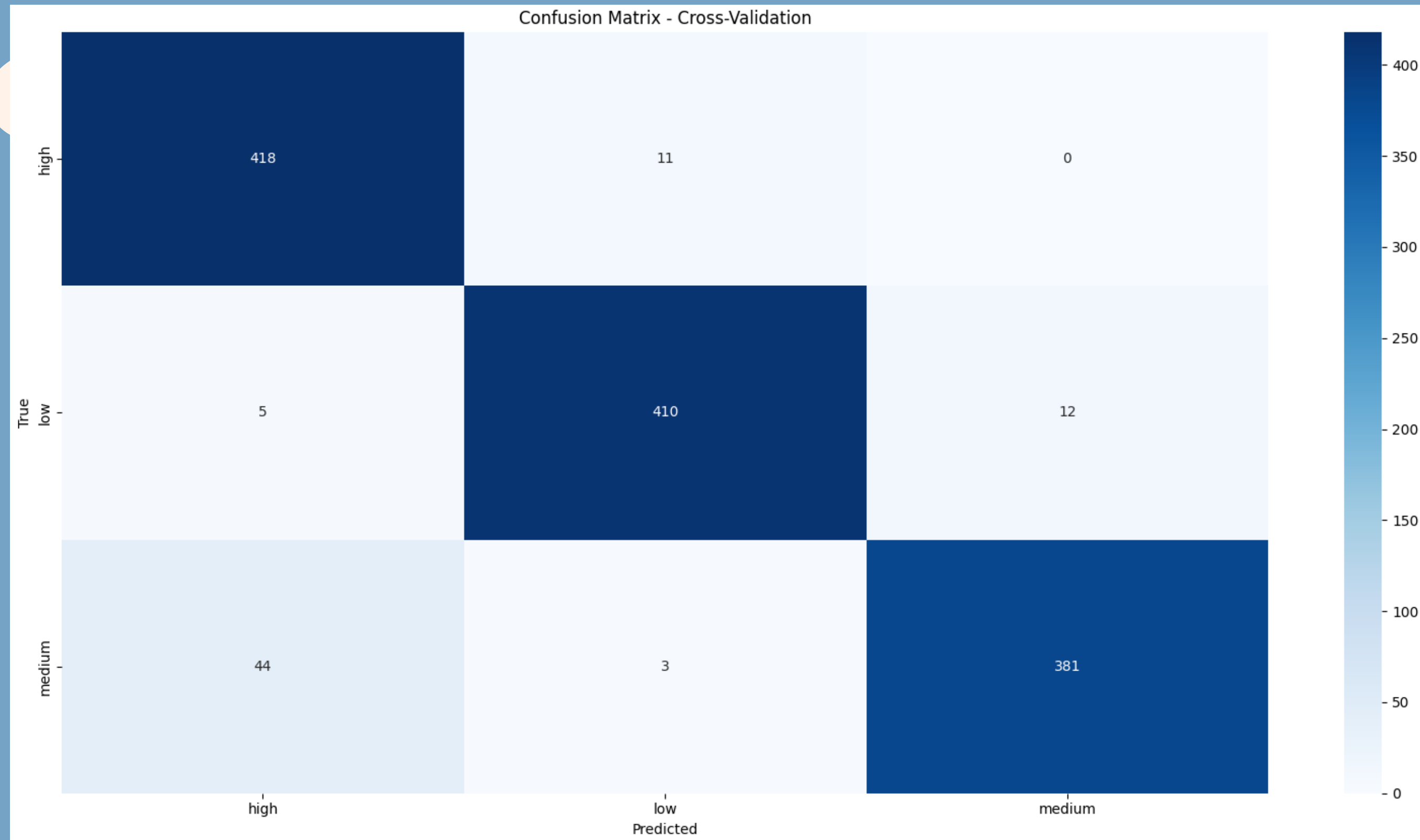
Base de dados
Pré-processada e
Balanceada

Kernel Polinomial
(poly, rbf, linear)



 Acuracia: 93,52%

 F1-Score: 0.94



 Acuracia: 94.16%

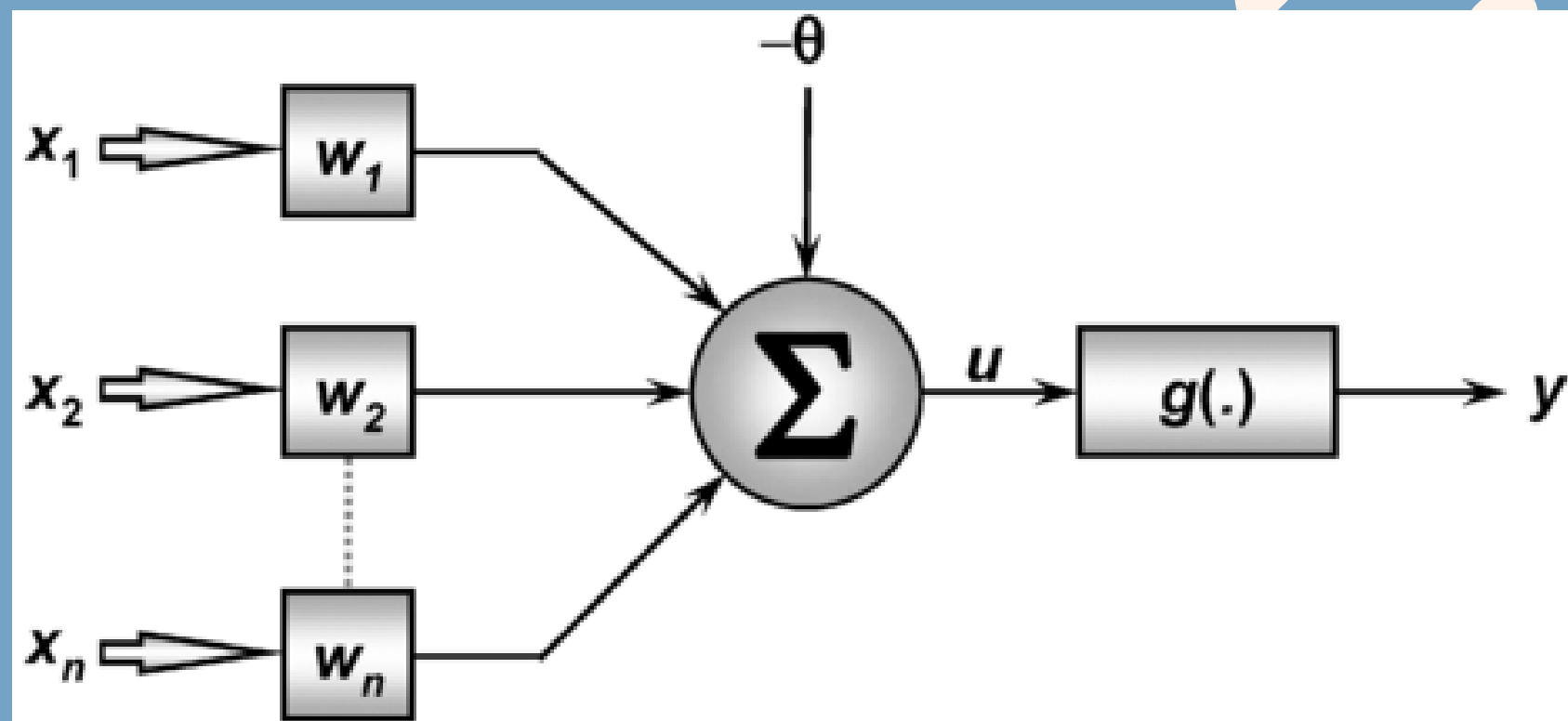
 F1-Score: 0,94

Rede Neural Multilayer Perceptron (MLP)



6.4

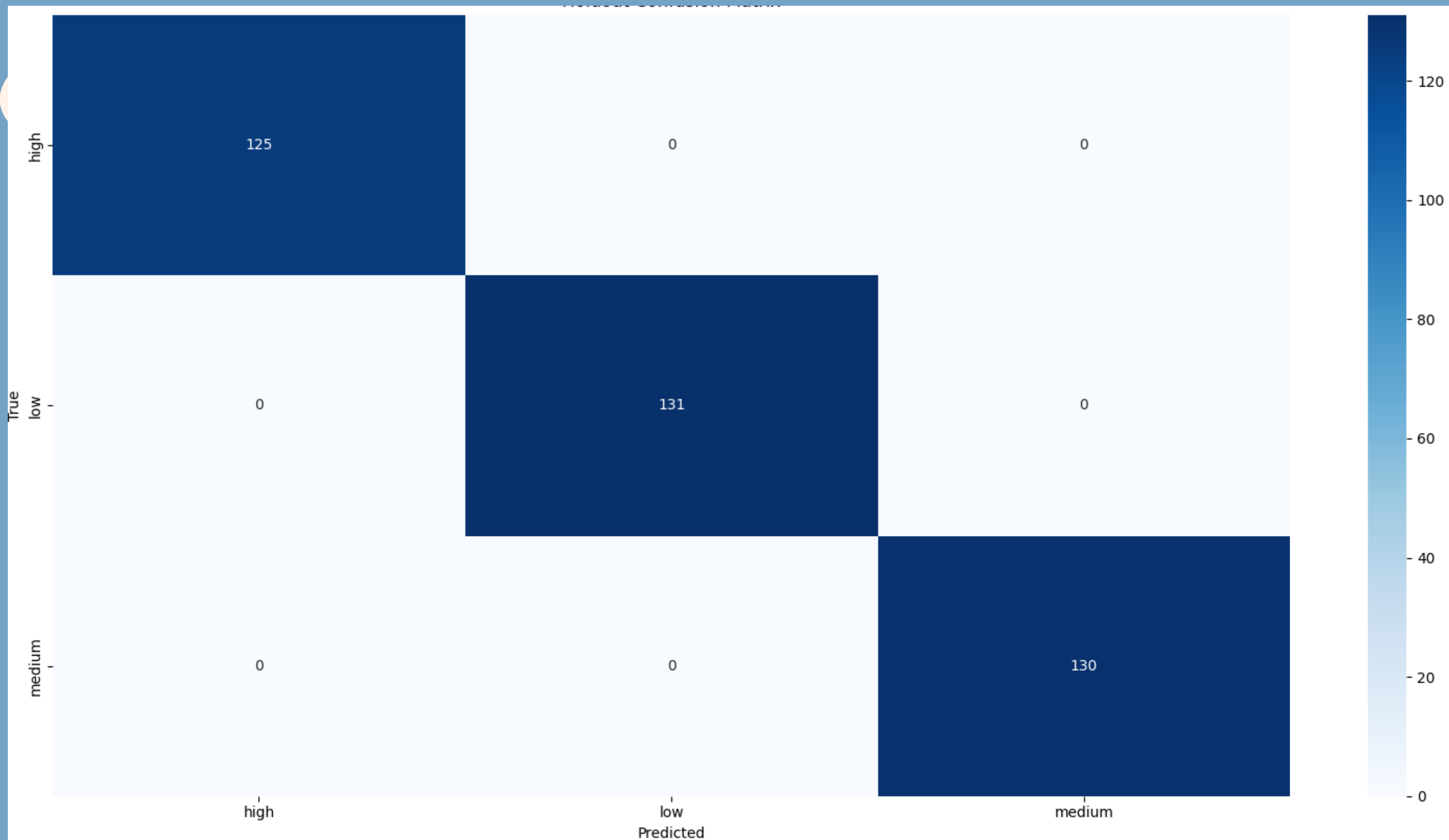
MLP




Classe MLPClassifier do
módulo
sklearn.neural_network do
scikit-learn.

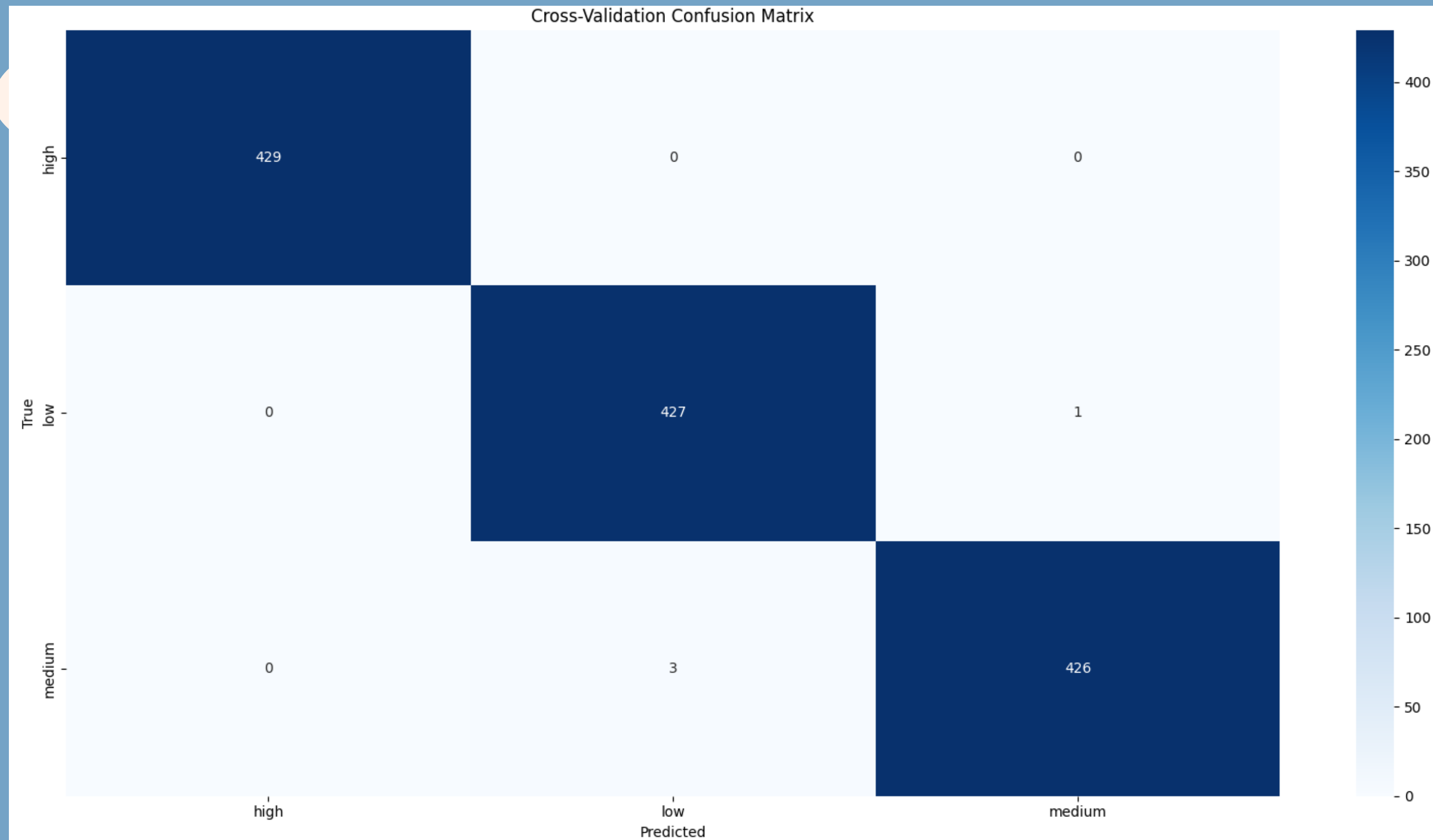
Base de dados
Pré-processada e
Balanceada

2 camadas ocultas,
onde cada camada tem 20
neurônios



 Acuracia:100%

 F1-Score: 1



 Acuracia:99.6890%

 F1-Score: 0.9969

Conclusão



TABELA DE RESULTADOS CLASSIFICAÇÃO

Métodos	Holdout		Cross-Validation	
Métricas	Acurácia	F1-Score	Acurácia	F1-Score
Árvore de Decisão	100%	1	99,69%	0.99688
K-Nearest Neighbors(KNN)	100%	1	99.1437%	0.9915
Support Vectors Machine(SVM)	93.5233%	0.9351	94.1589%	0.9416
Rede Neural Multilayer Perceptron (MLP)	100%	1	99.6890%	0.9969