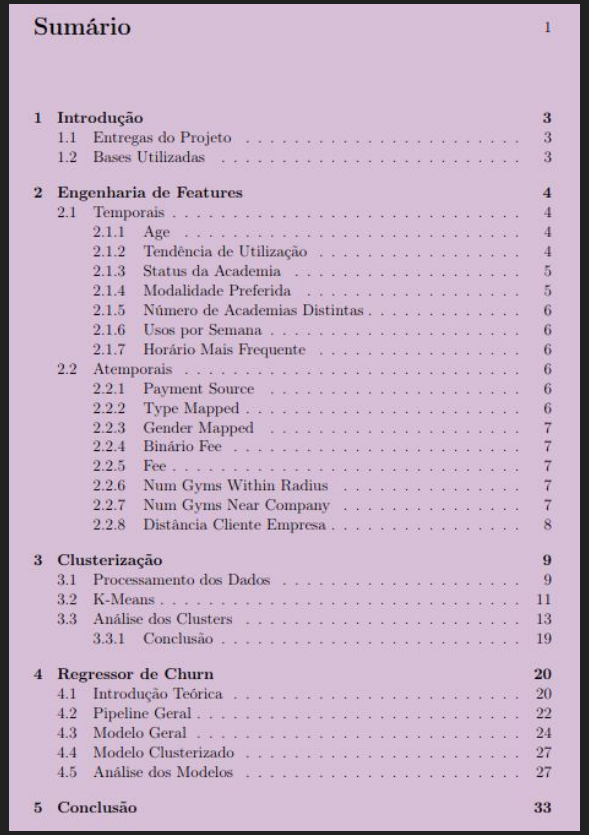
Total Pass + Polijúnior: Construção de um modelo de Machine Learning para a previsão de Churn Rate

Entregas da Poli Júnior

- ’RelatorioPoliJunior.pdf ’
- ’Clusterizacao.ipynb
- ’AnaliseClusterizacao.ipynb’
- ’Tabelas Poli Júnior’
- ’Regressor.ipynb’
- ’AnaliseRegressor.ipynb’

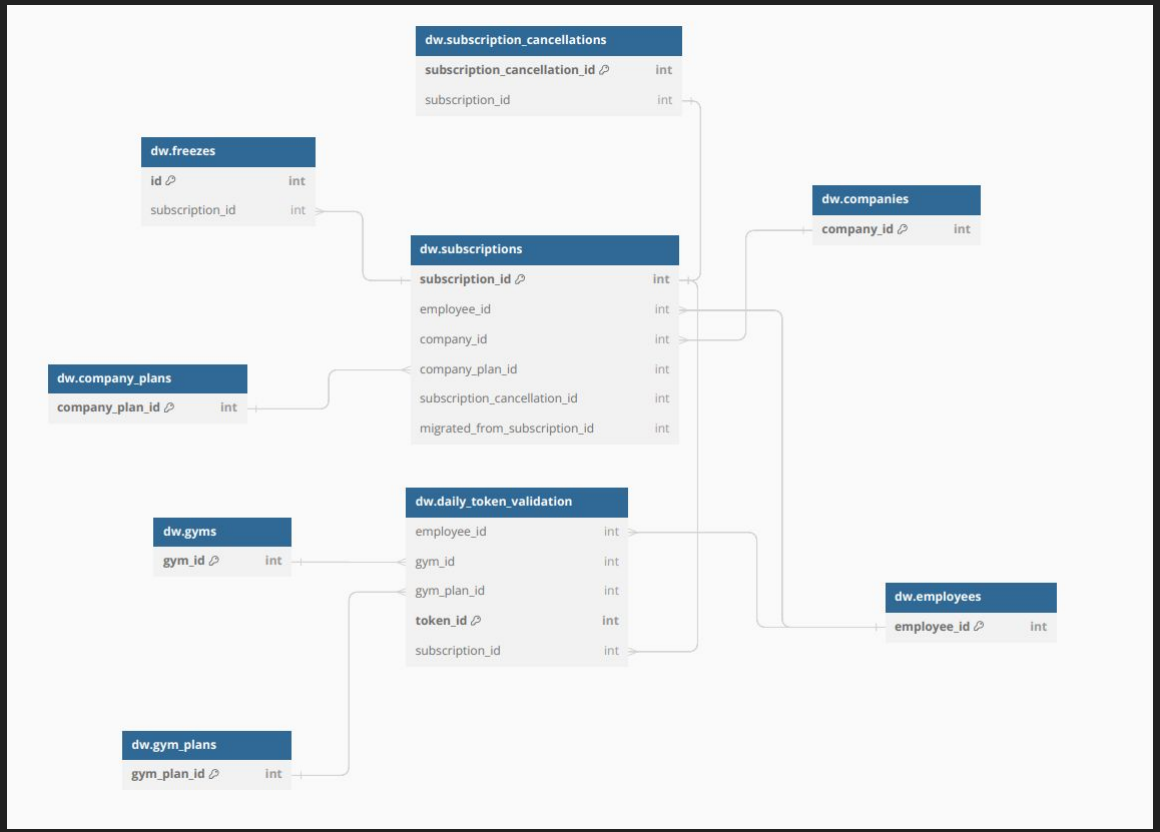
https://drive.google.com/drive/folders/1UYzDJZRai-5fr6G8USkP4EDk0nGGjafn?usp=sharing



Sumário 1

Índice

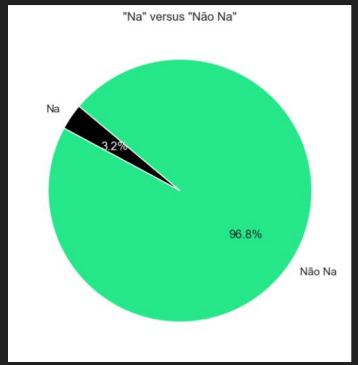
Aprofundamento no contexto do projeto

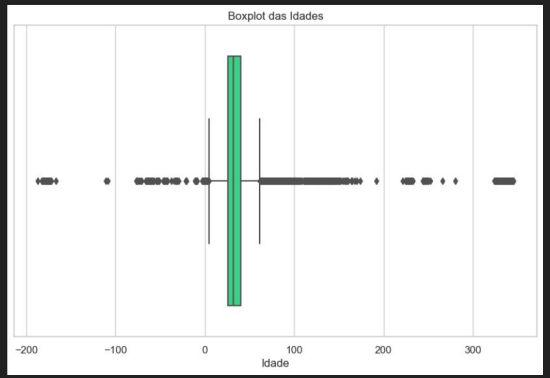


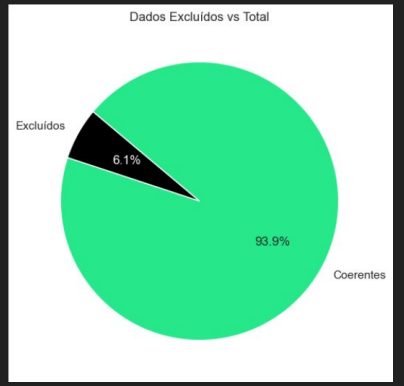
Índice

Análise exploratória da base de clientes e tratamento de dados

```
df_subscriptions['created_at'] = pd.to_datetime(df_subscriptions['created_at'], format="%Y-%m-%dT%H:%M:%S.%fZ")
df_subscriptions['started_at'] = pd.to_datetime(df_subscriptions['started_at'], format="%Y-%m-%dT%H:%M:%S.%fZ")
df_subscriptions['suspended_at'] = pd.to_datetime(df_subscriptions['suspended_at'], format="%Y-%m-%dT%H:%M:%S.%fZ")
df_subscriptions['canceled_at'] = pd.to_datetime(df_subscriptions['canceled_at'], format="%Y-%m-%dT%H:%M:%S.%fZ")
df_subscriptions
]
```







Engenharia de Features

1°- Determinação das métricas mais importantes e úteis para a realização do projeto;

2°- Criação e refinamento das features;

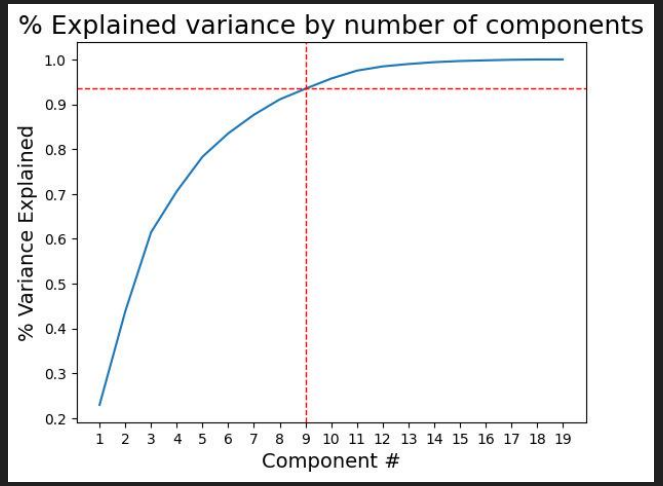
- Tendência de utilização;
- Distância dentro de um raio de 5km entre endereço do cliente e academia.

3°- Funcionalização para o uso no regressor.

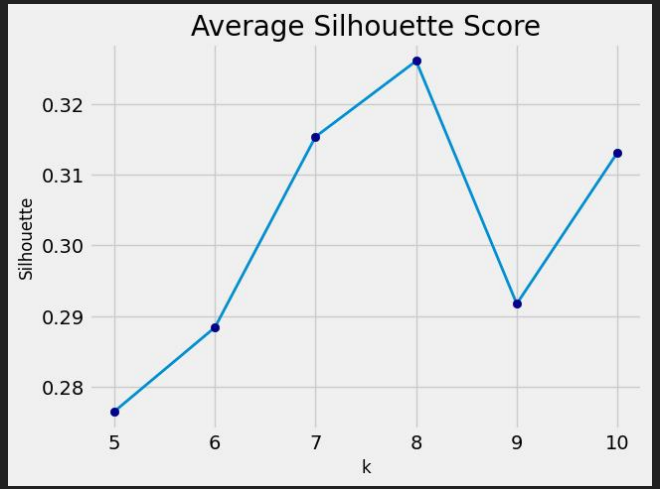
Índice

Clusterização

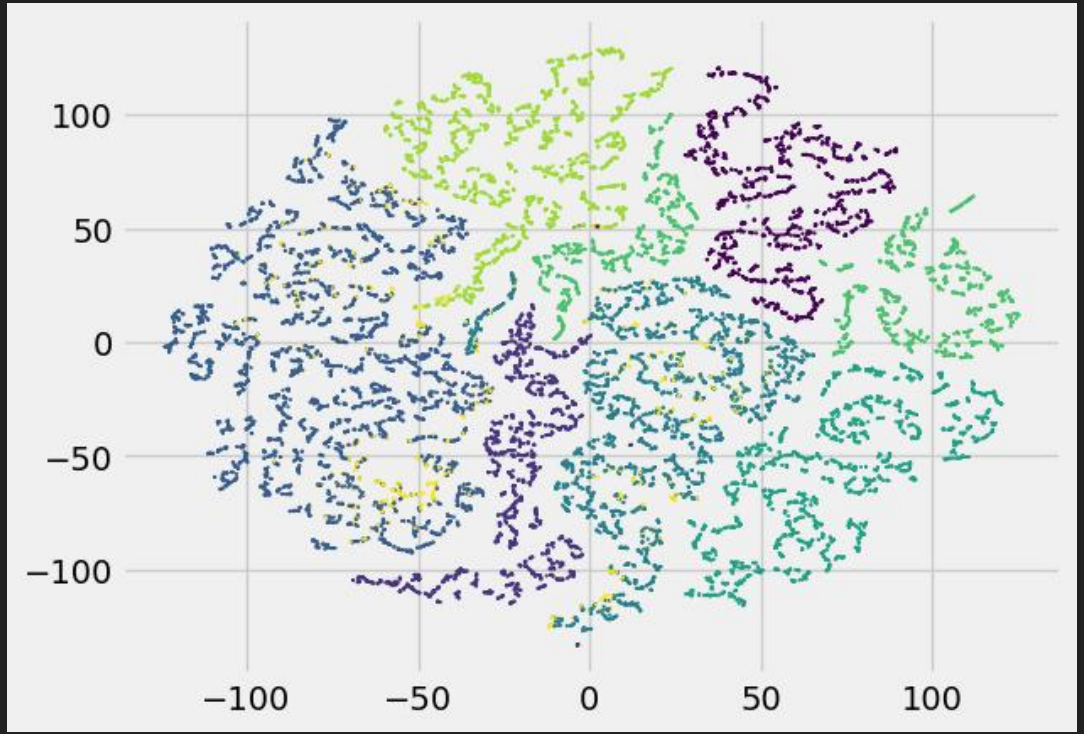
Análise das Componentes Principais (PCA)



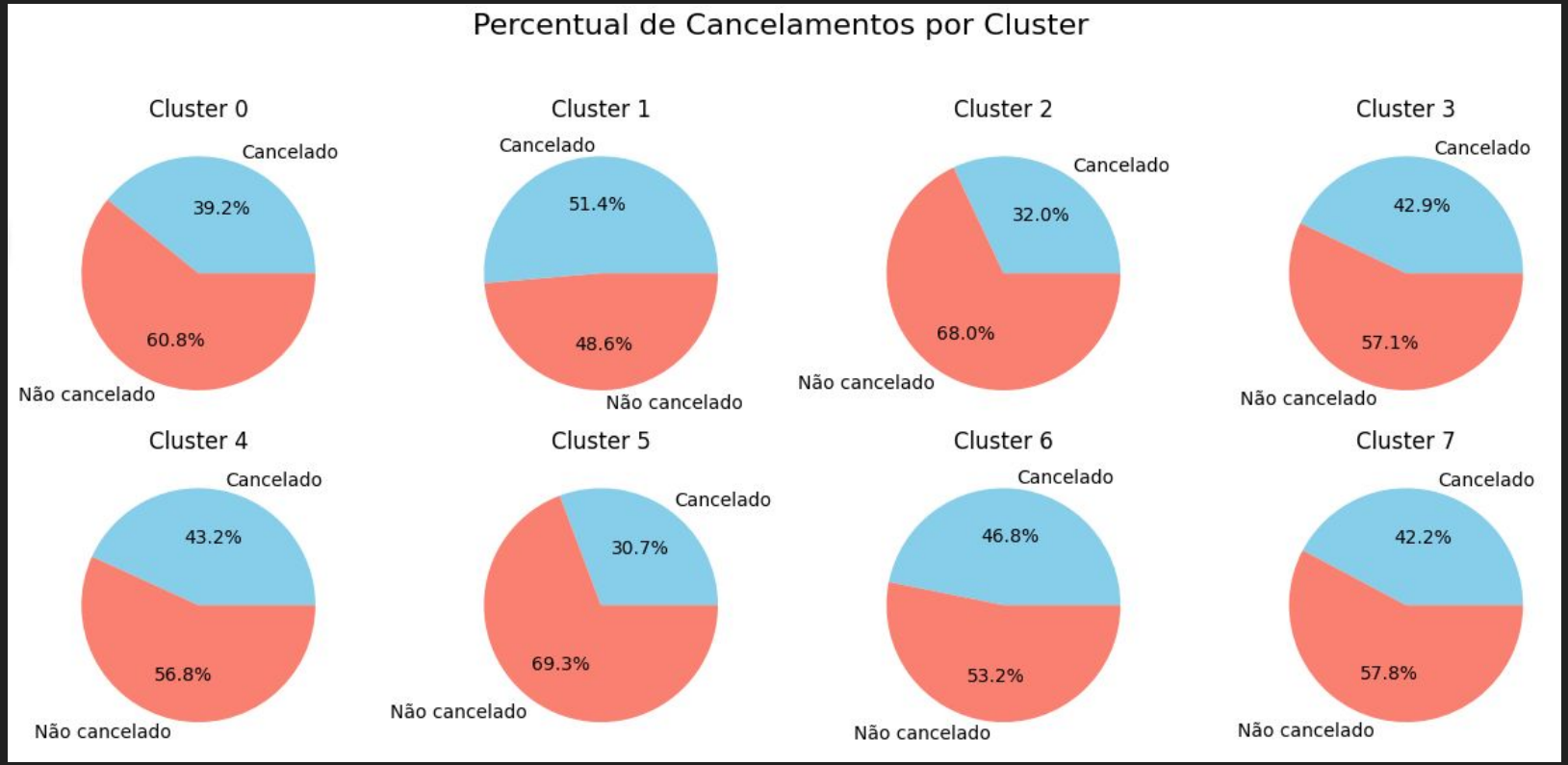
Determinação do Número Ideal de Clusters



Projeção T-SNE para a Clusterização

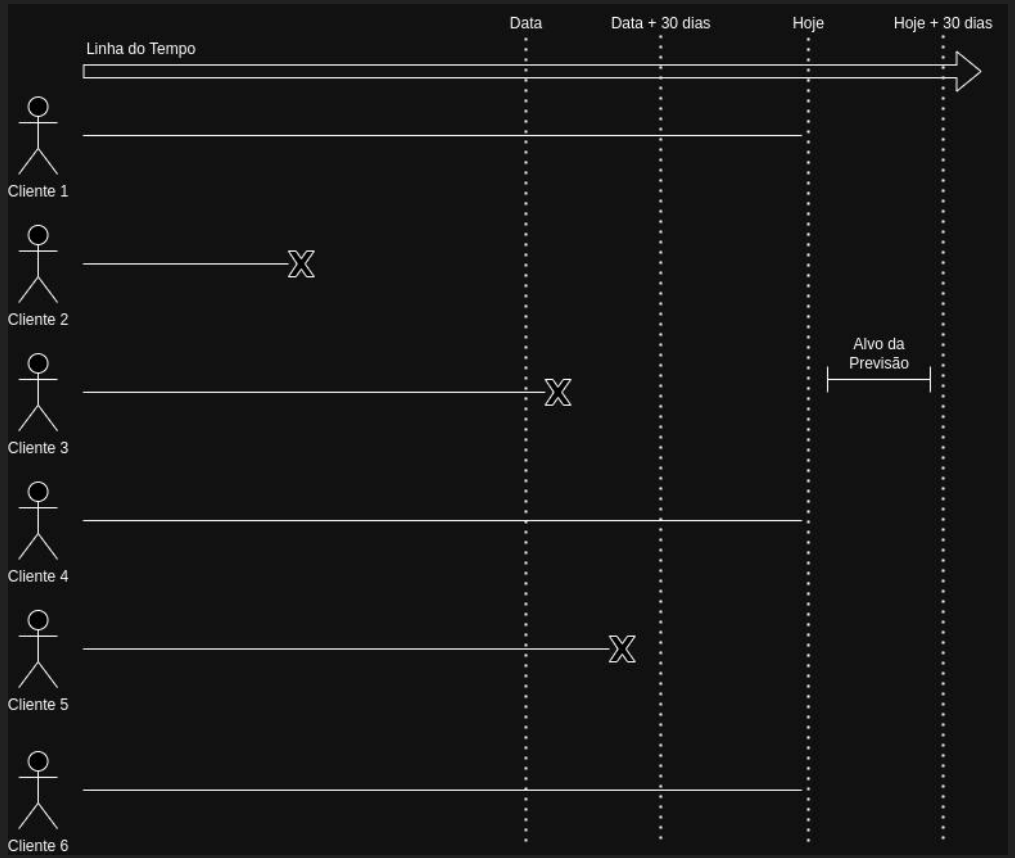


Balanceamento da Clusterização

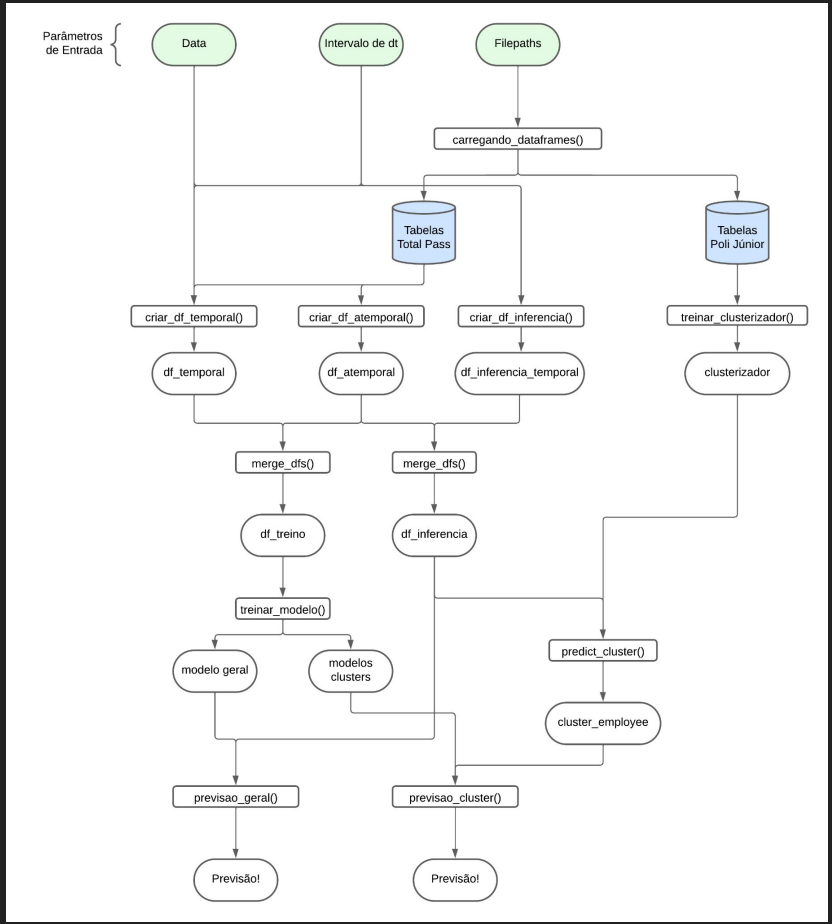


Introdução ao Cut Off Date do Regressor

Introdução ao Cut Off Date do Regressor



Pipeline Geral do Regressor



Índice

Teste de modelos e tuning de algoritmos

X_{scaled} = (X - Xmin)/(Xmin - Xmax)

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Acurácia: 0.6987160346372051

```
[[ 591  616]
 [ 393 1749]]
```

	precision	recall	f1-score	support
0.0	0.60	0.49	0.54	1207
1.0	0.74	0.82	0.78	2142
accuracy			0.70	3349
macro avg	0.67	0.65	0.66	3349
weighted avg	0.69	0.70	0.69	3349

========== Undersampling ==========

Acurácia: 0.6616900567333532

Matriz de Confusão:

```
 [[ 862  345]
 [ 788 1354]]
```

Relatório de Classificação:

	precision	recall	f1-score	support
0.0	0.52	0.71	0.60	1207
1.0	0.80	0.63	0.71	2142
accuracy			0.66	3349
macro avg	0.66	0.67	0.65	3349
weighted avg	0.70	0.66	0.67	3349

========== Oversampling ==========

Acurácia: 0.6906539265452374

Matriz de Confusão:

```
 [[ 668  539]
 [ 497 1645]]
```

Relatório de Classificação:

	precision	recall	f1-score	support
0.0	0.57	0.55	0.56	1207
1.0	0.75	0.77	0.76	2142
accuracy			0.69	3349
macro avg	0.66	0.66	0.66	3349
weighted avg	0.69	0.69	0.69	3349

========== SMOTE ==========

Acurácia: 0.680501642281278

Matriz de Confusão:

```
 [[ 731  476]
 [ 594 1548]]
```

Relatório de Classificação:

	precision	recall	f1-score	support
0.0	0.55	0.61	0.58	1207
1.0	0.76	0.72	0.74	2142
accuracy			0.68	3349
macro avg	0.66	0.66	0.66	3349
weighted avg	0.69	0.68	0.68	3349

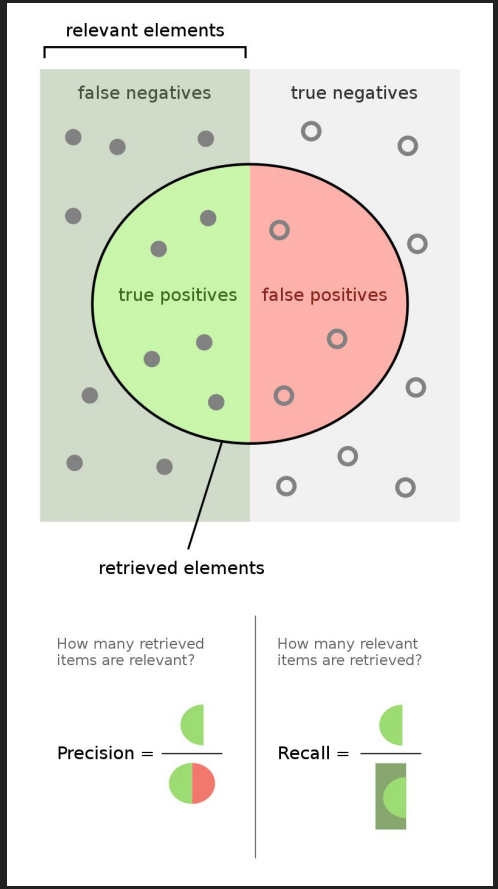
Índice

Determinação de Métricas de Avaliação

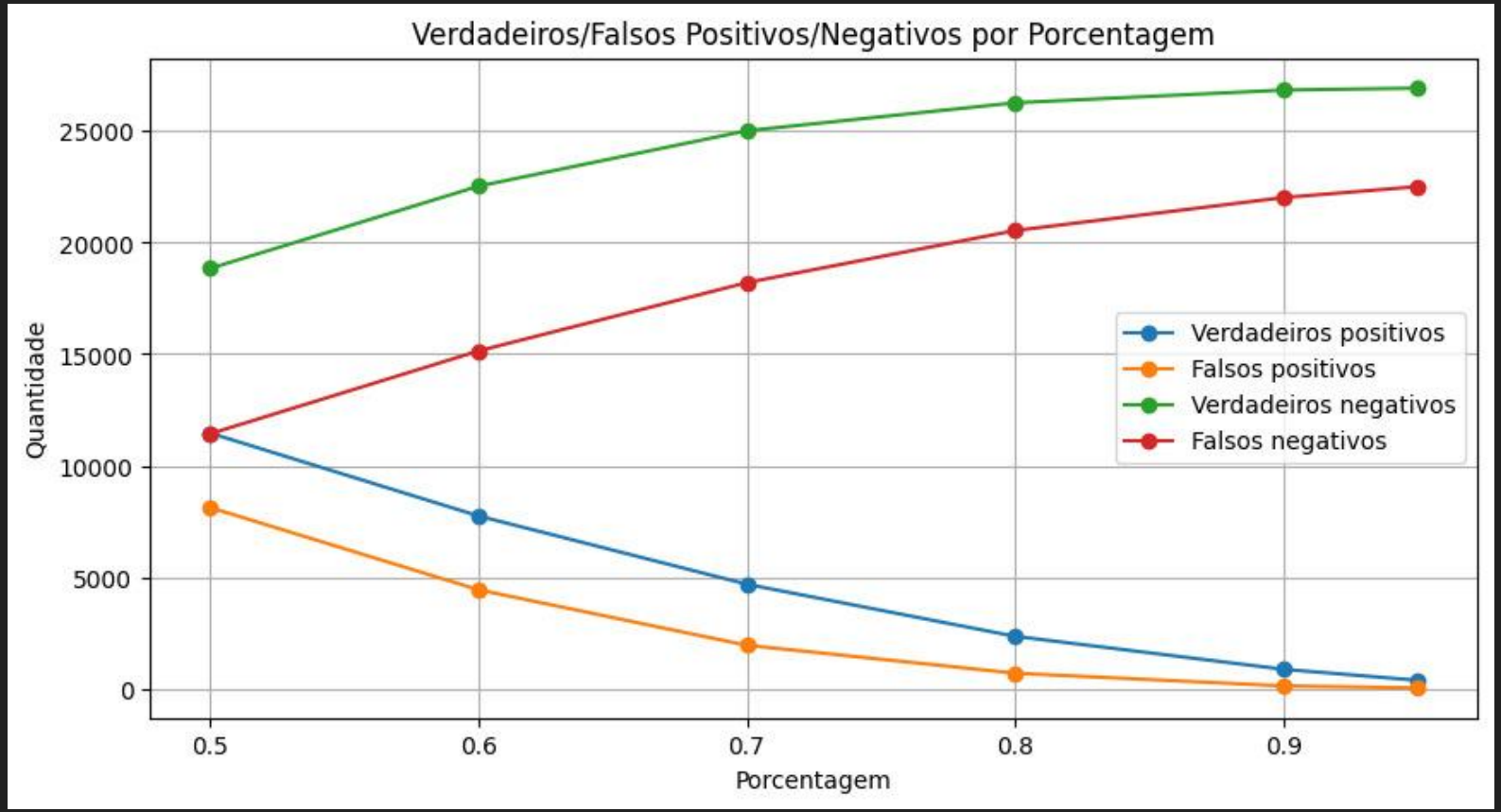
- Verdadeiros positivos: Clientes que cancelaram e o modelo previu cancelamento. Nessa situação, a Total Pass teria a possibilidade de evitar o cancelamento.
- Falsos positivos: Clientes que não cancelaram e o modelo previu cancelamento. Nessa situação, a Total Pass perderia capital, atuando sobre um cliente que não cancelaria.
- Verdadeiros negativos: Clientes que não cancelaram e o modelo previu não cancelamento. Nessa situação, a Total Pass não gasta capital desnecessariamente.
- Falsos negativos: Clientes que cancelaram e o modelo previu não cancelamento. Nessa situação, a Total Pass perdeu a chance de investir na retenção desses clientes.

$$recall = \frac{verdadeiros\ positivos}{verdadeiros\ positivos + falsos\ negativos}$$

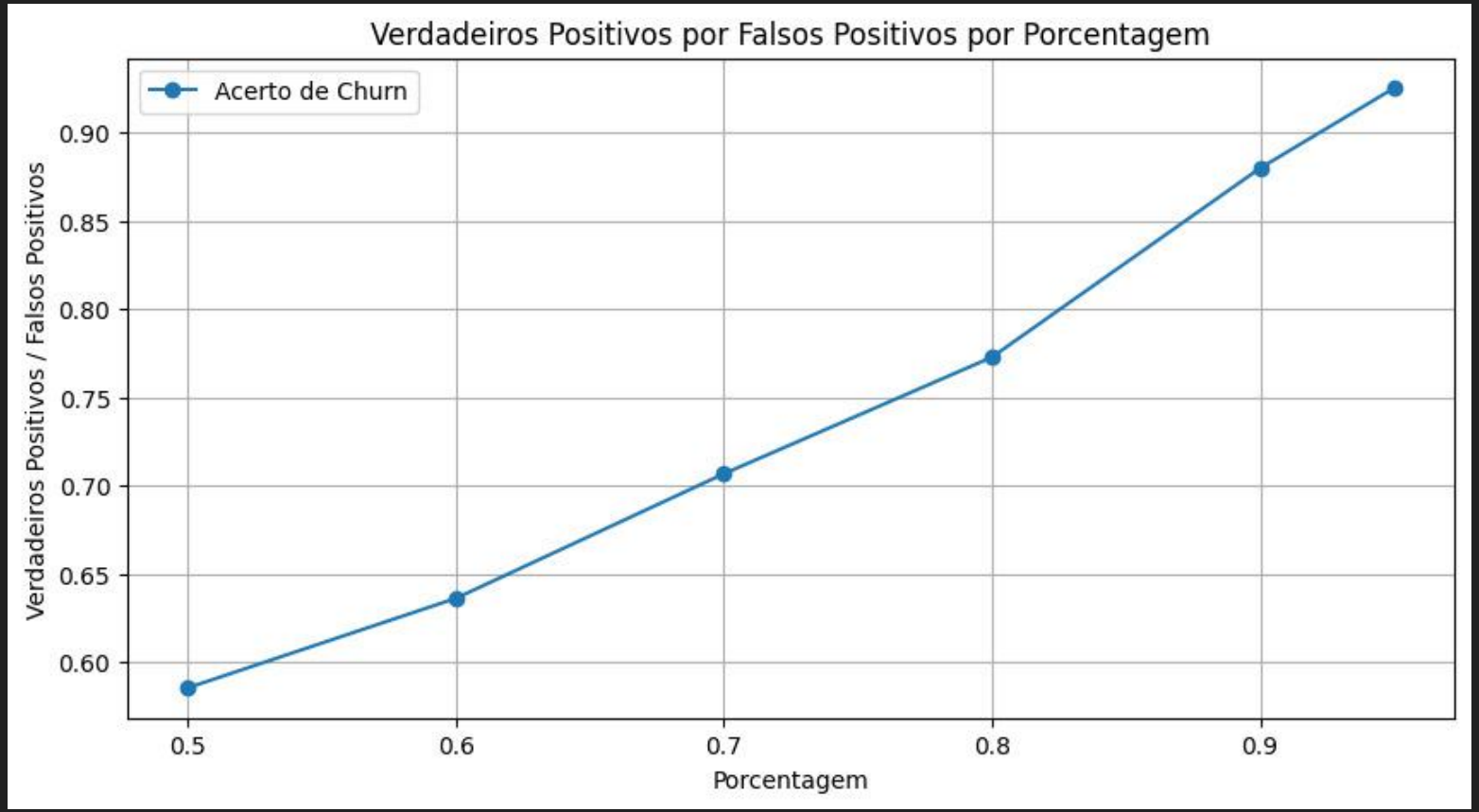
$$precisao = \frac{verdadeiros\ positivos}{verdadeiros\ positivos + falsos\ positivos}$$



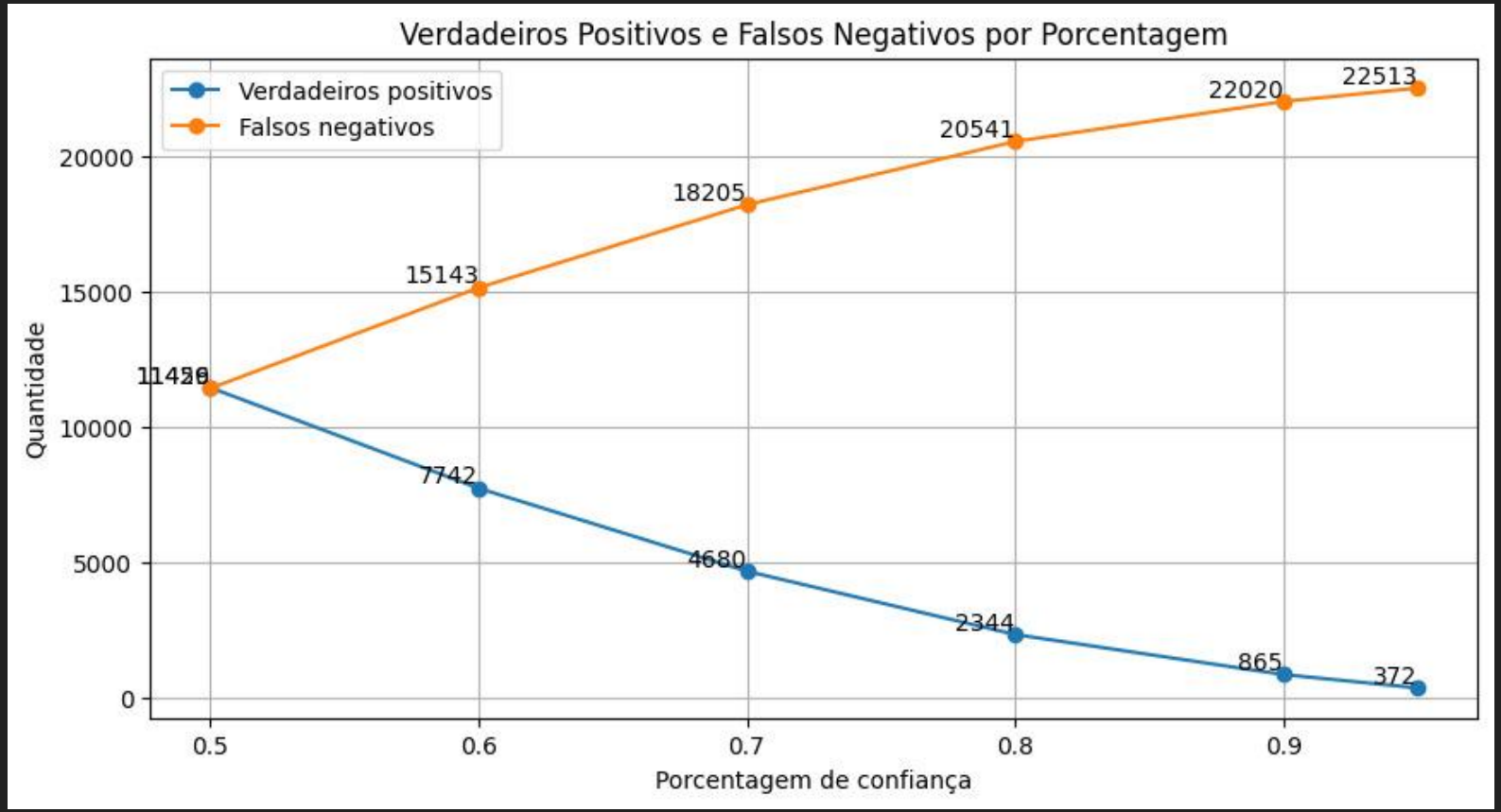
Análise de resultados e documentação de insights



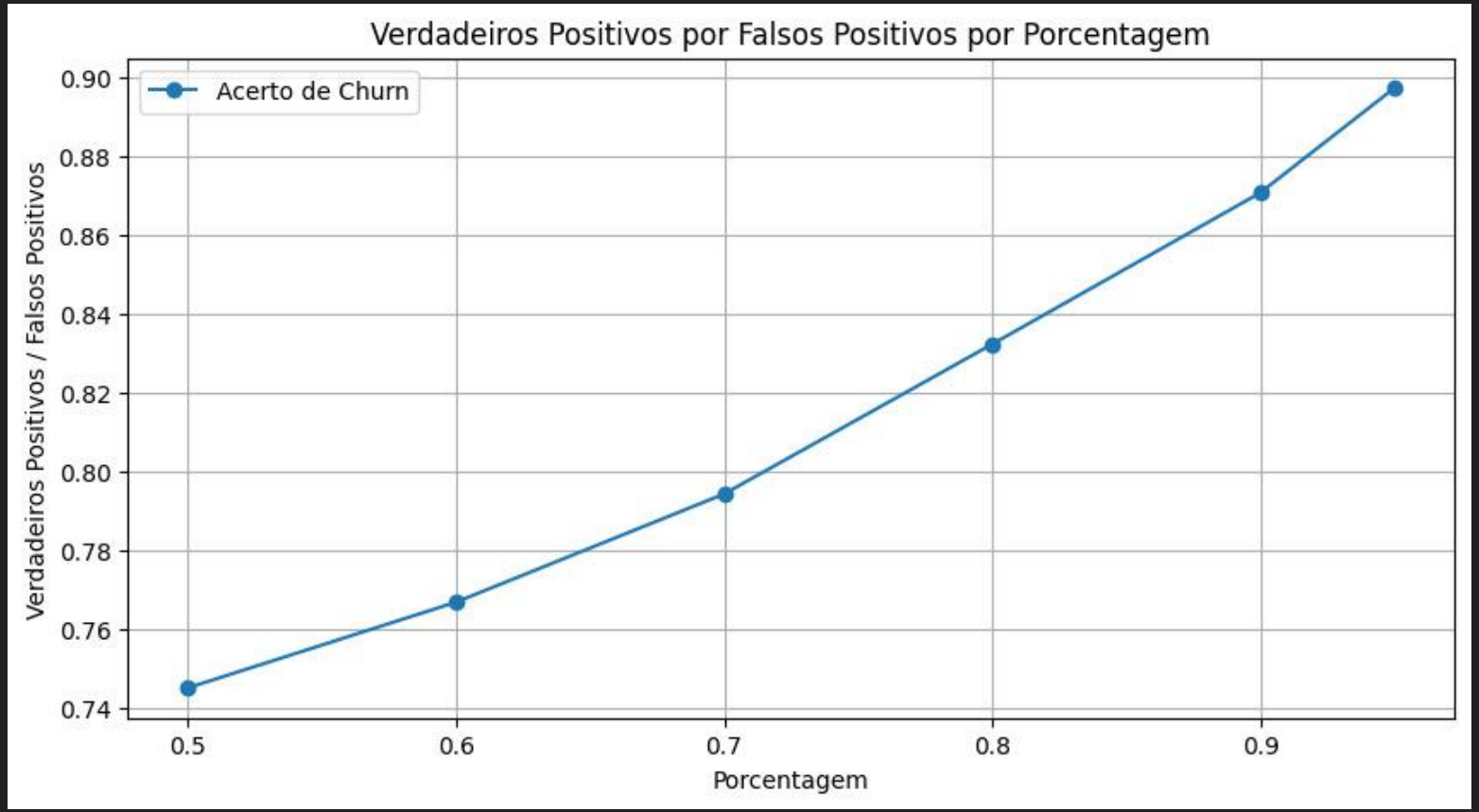
Análise de resultados e documentação de insights



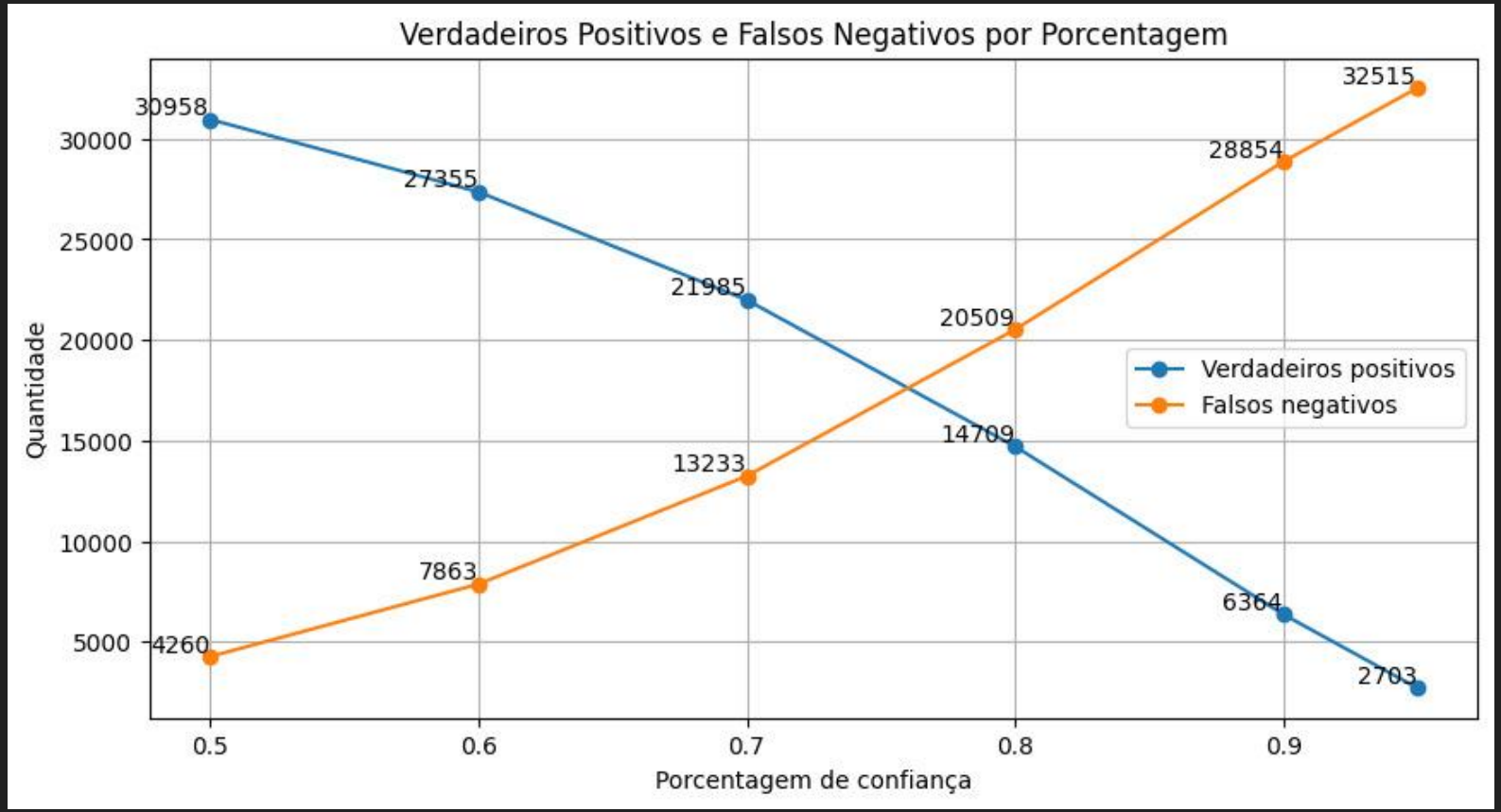
Análise de resultados e documentação de insights



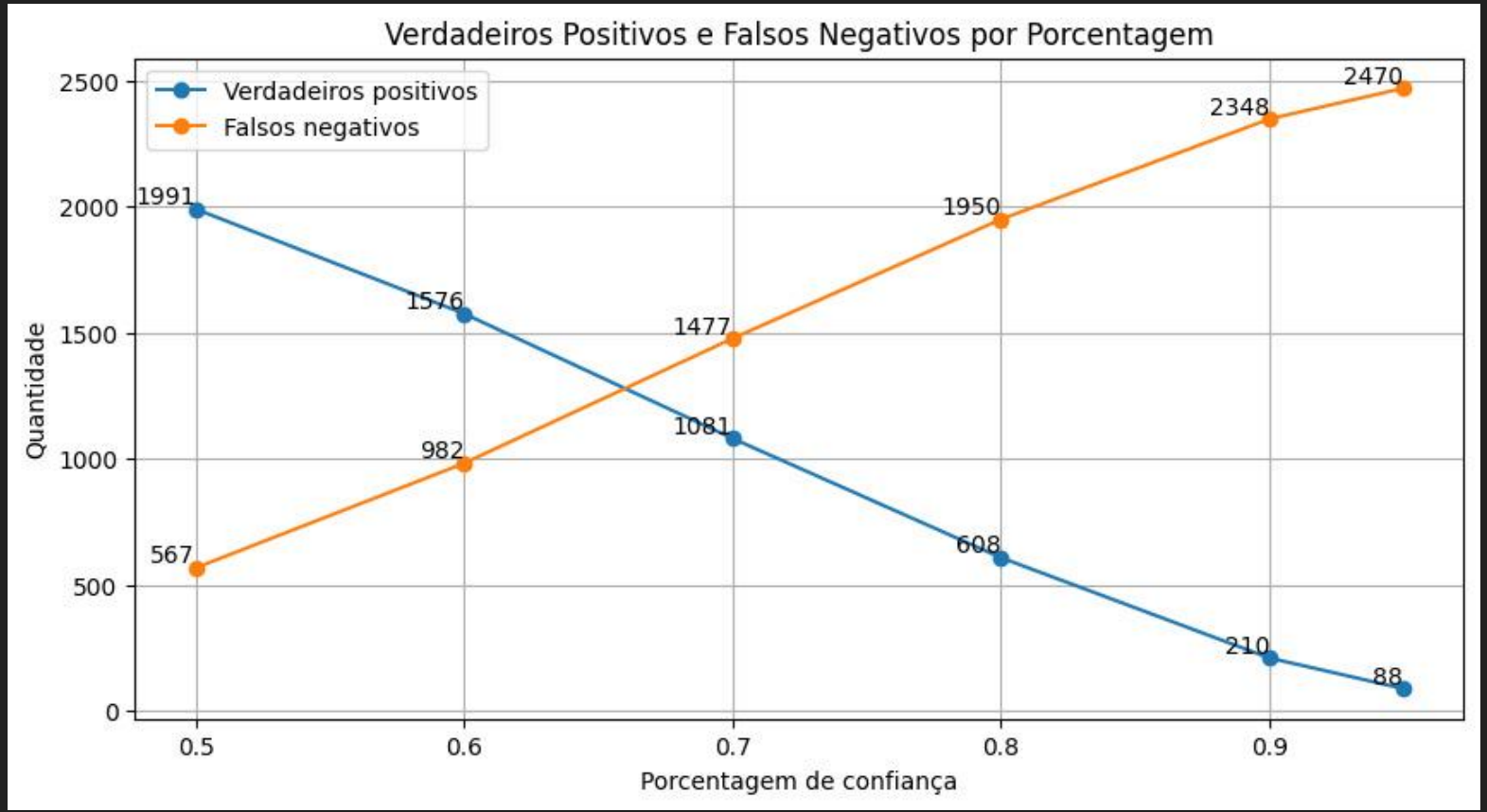
Análise de resultados e documentação de insights



Exemplo de cluster balanceado



Exemplo de cluster desbalanceado



Como definir a Porcentagem de Confiança escolhida?

1°- Custo da ação.

2°- Perda em reter um falso positivo.

3°- Tempo de duração da ação.