

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
1 Introdução	3
1.1 Entregas do Projeto	3
1.2 Bases Utilizadas	3
2 Engenharia de Features	4
2.1 Temporais	4
2.1.1 Age	4
2.1.2 Tendência de Utilização	4
2.1.3 Status da Academia	5
2.1.4 Modalidade Preferida	5
2.1.5 Número de Academias Distintas	6
2.1.6 Usos por Semana	6
2.1.7 Horário Mais Frequente	6
2.2 Atemporais	6
2.2.1 Payment Source	6
2.2.2 Type Mapped	6
2.2.3 Gender Mapped	7
2.2.4 Binário Fee	7
2.2.5 Fee	7
2.2.6 Num Gyms Within Radius	7
2.2.7 Num Gyms Near Company	7
2.2.8 Distância Cliente Empresa	8
3 Clusterização	9
3.1 Processamento dos Dados	9
3.2 K-Means	11
3.3 Análise dos Clusters	13
3.3.1 Conclusão	19
4 Regressor de Churn	20
4.1 Introdução Teórica	20
4.2 Pipeline Geral	22
4.3 Modelo Geral	24
4.4 Modelo Clusterizado	27
4.5 Análise dos Modelos	27
5 Conclusão	33

Índice

1. Aprofundamento no contexto do projeto

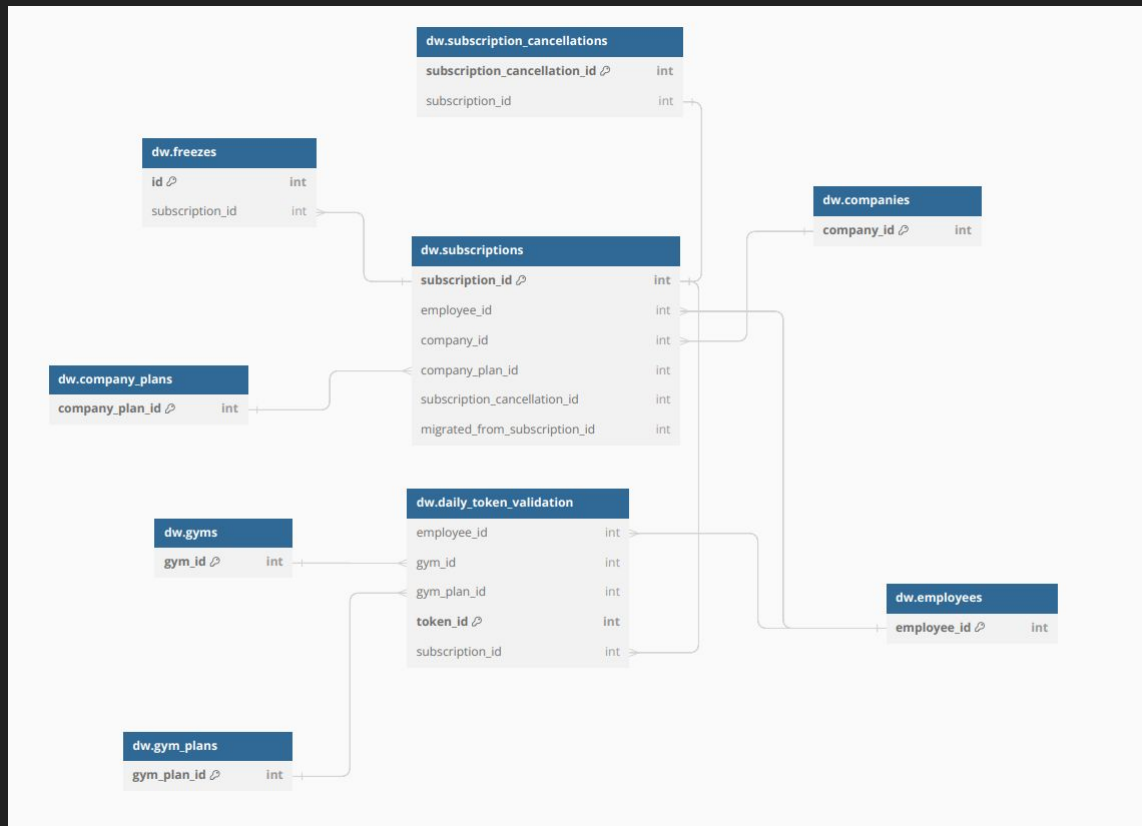
2. Análise Exploratória da Base de Clientes

3. Determinação de Modelos e Pipeline Geral

4. Teste de modelos e tuning de algoritmos

5. Análise de resultados e documentação de insights

Aprofundamento no contexto do projeto



Índice

1. Aprofundamento no contexto do projeto

2. Análise Exploratória da Base de Clientes

3. Determinação de Modelos e Pipeline Geral

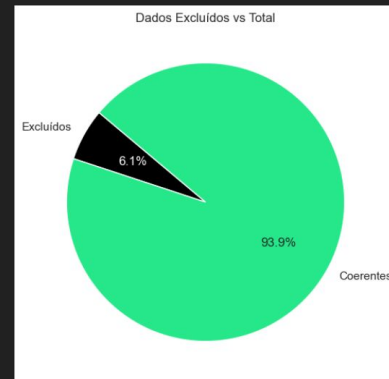
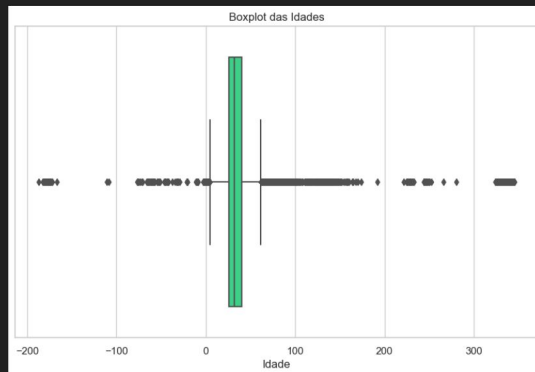
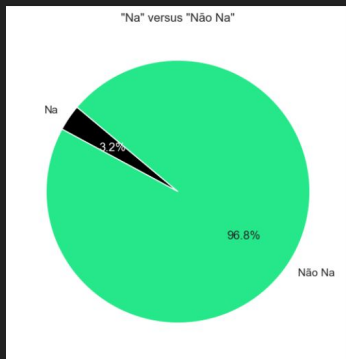
4. Teste de modelos e tuning de algoritmos

5. Análise de resultados e documentação de insights

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

1. Aprofundamento no contexto do projeto

2. Análise Exploratória da Base de Clientes

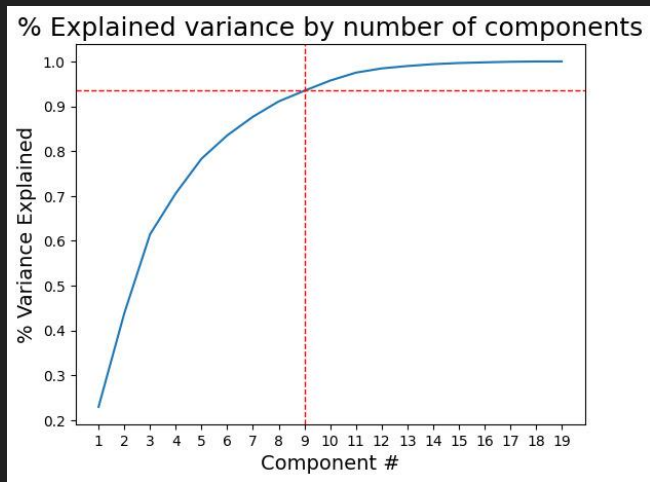
3. Determinação de Modelos e Pipeline Geral

4. Teste de modelos e tuning de algoritmos

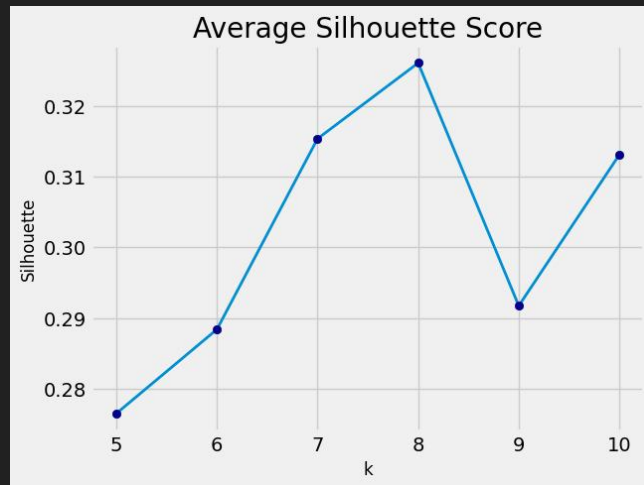
5. Análise de resultados e documentação de insights

Clusterização

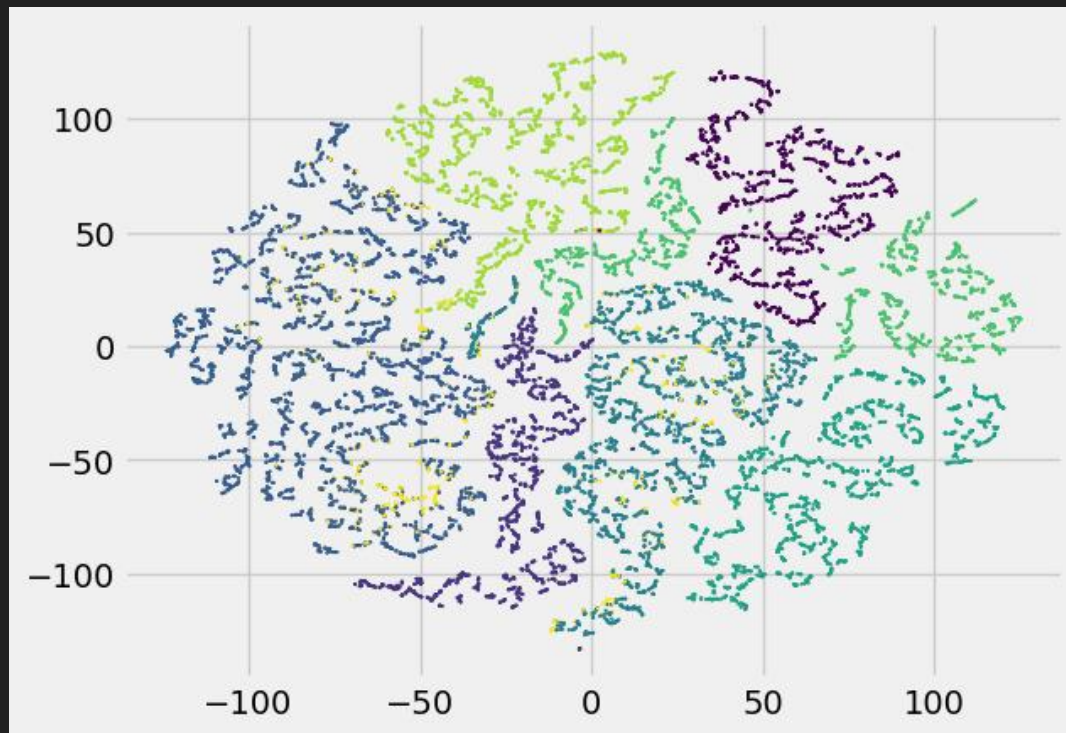
Análise das Componentes Principais (PCA)



Determinação do Número Ideal de Clusters

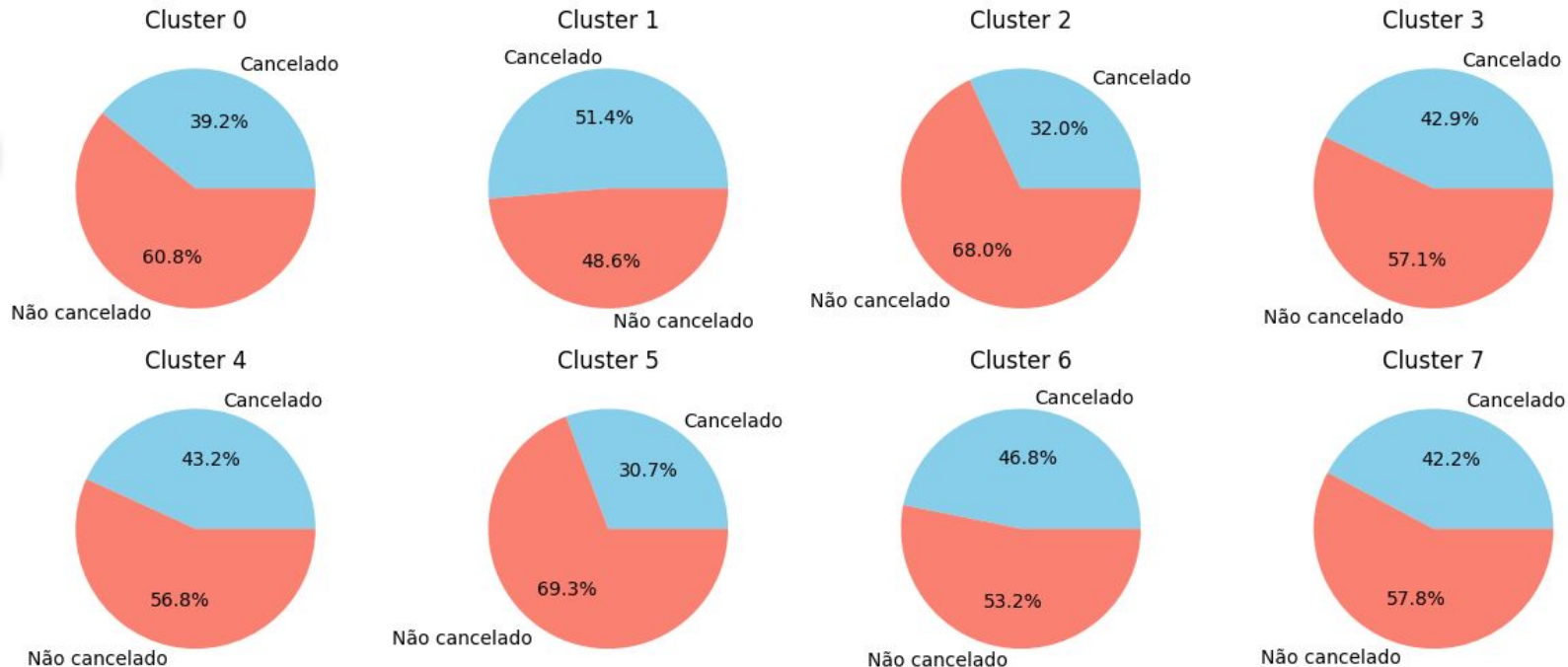


Projeção T-SNE para a Clusterização

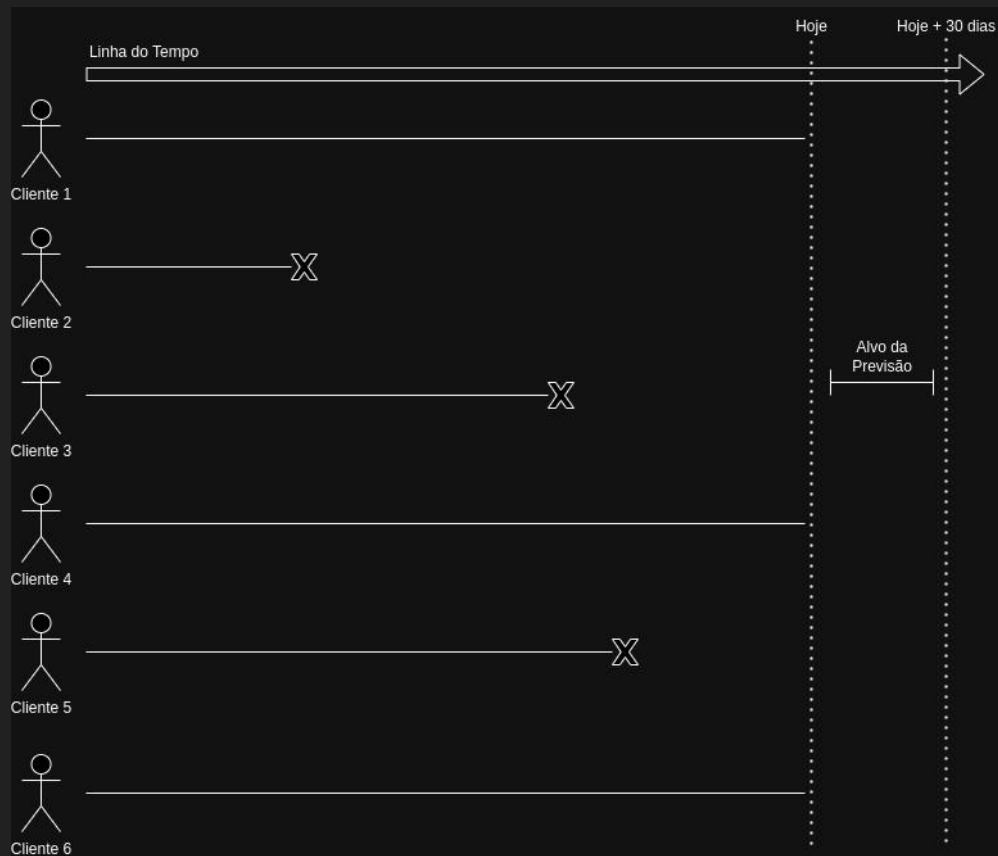


Balanceamento da Clusterização

Percentual de Cancelamentos por Cluster



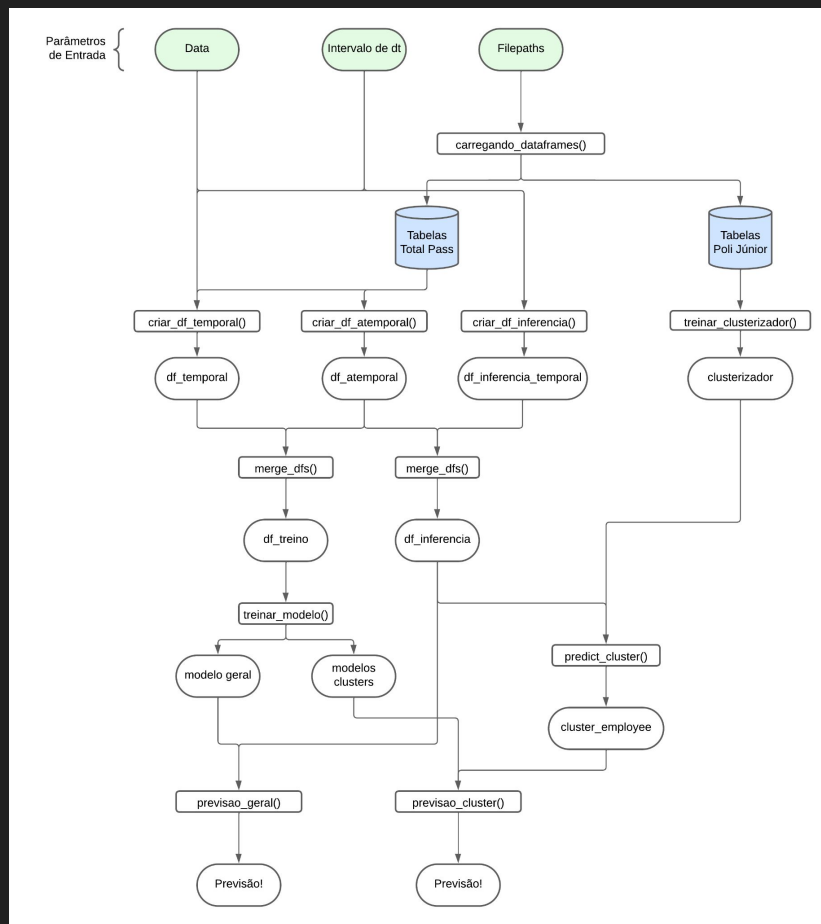
Introdução ao Cut Off Date do Regressor



Introdução ao Cut Off Date do Regressor



Pipeline Geral do Regressor



Índice

1. Aprofundamento no contexto do projeto

2. Análise Exploratória da Base de Clientes

3. Determinação de Modelos e Pipeline Geral

4. Teste de modelos e tuning de algoritmos

5. Análise de resultados e documentação de insights

Teste de modelos e tuning de algoritmos

$$X_{\text{scaled}} = (X - X_{\text{min}}) / (X_{\text{max}} - X_{\text{min}})$$

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

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

1. Aprofundamento no contexto do projeto

2. Análise Exploratória da Base de Clientes

3. Determinação de Modelos e Pipeline Geral

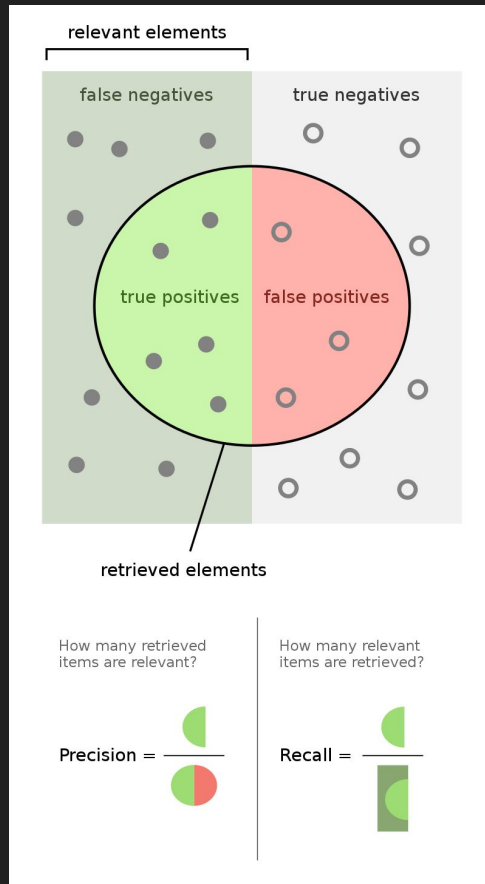
4. Teste de modelos e tuning de algoritmos

5. Análise de resultados e documentação de insights

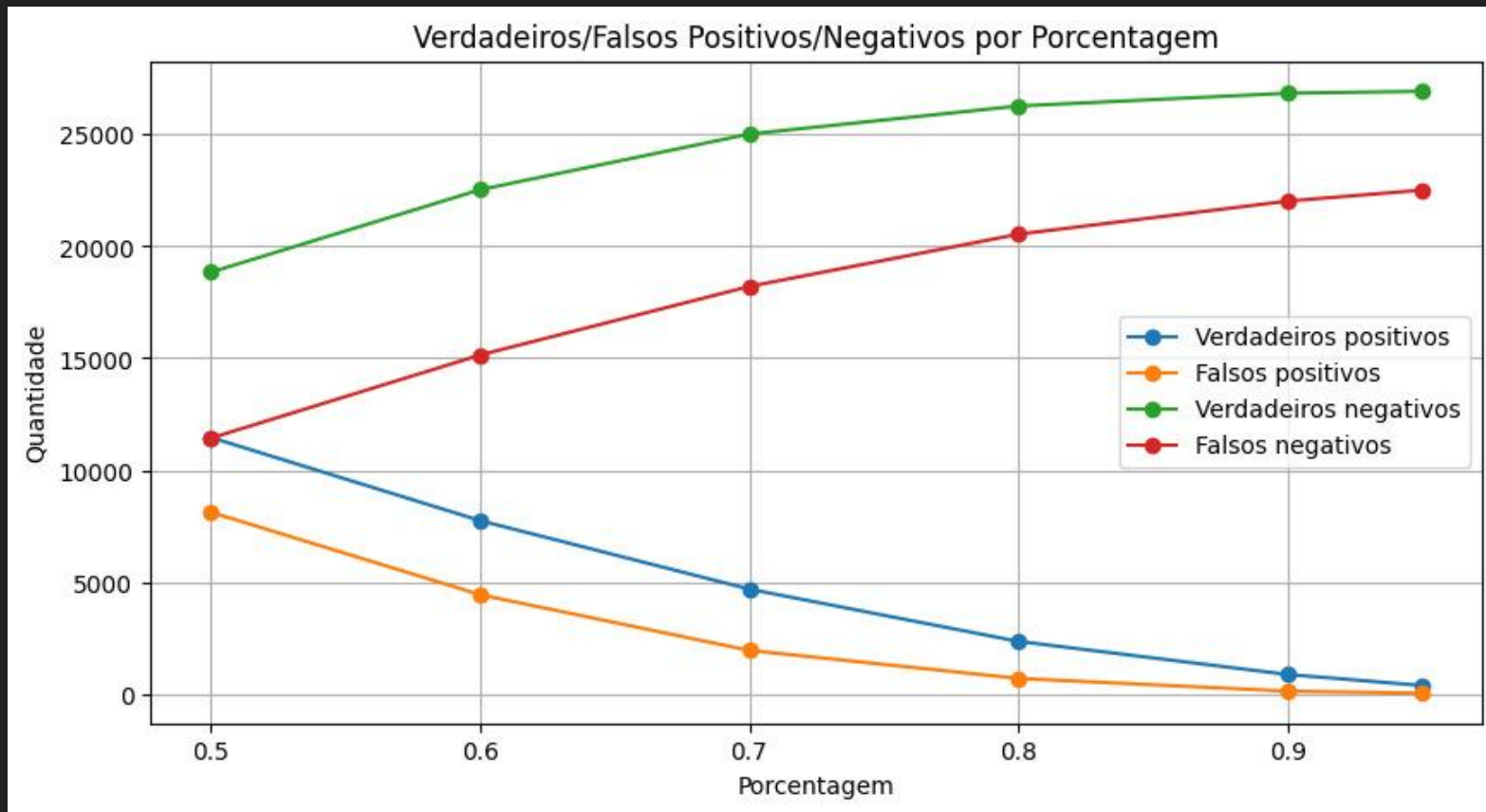
Determinação de Métricas de Avaliação

- 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.

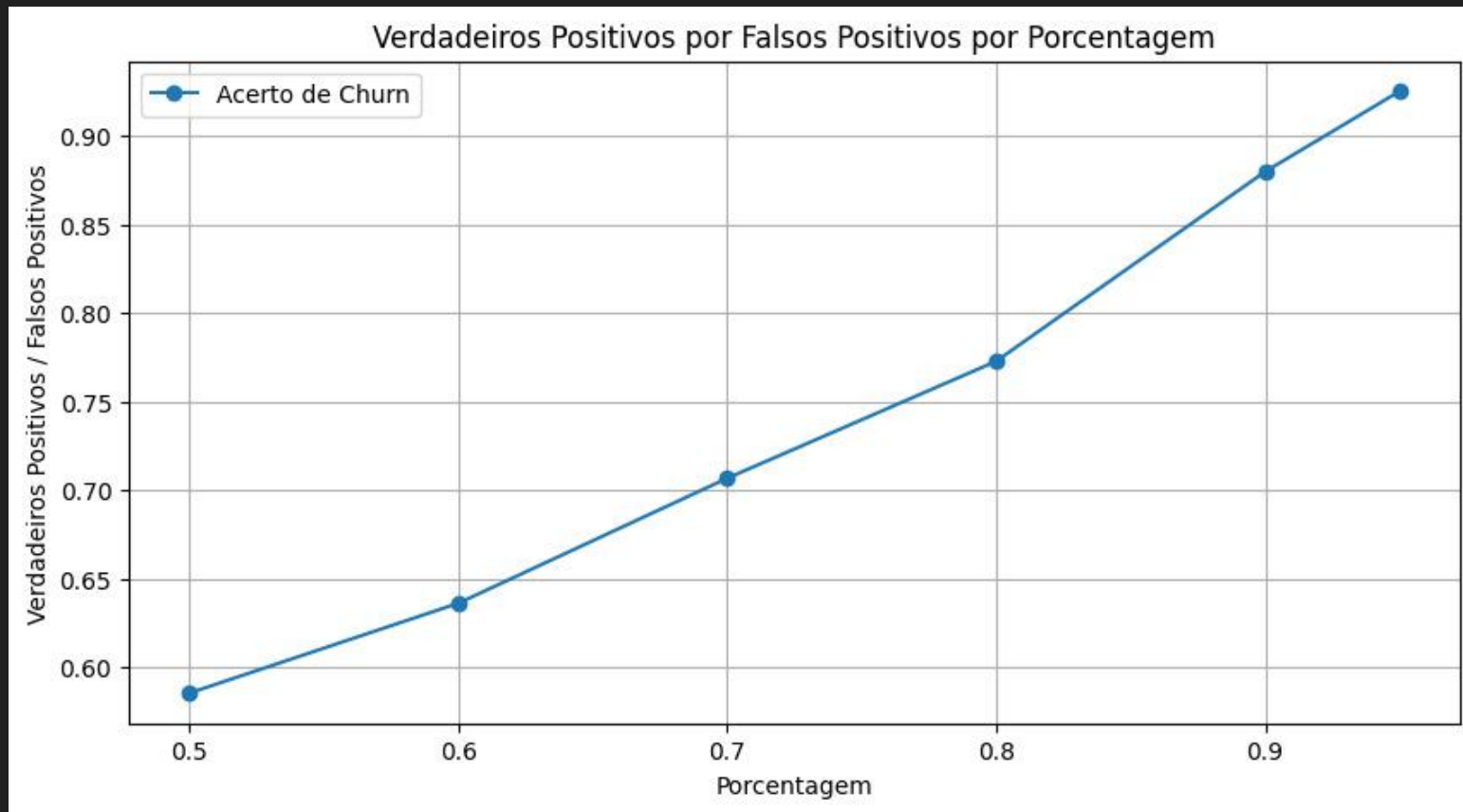
$$\text{recall} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos negativos}}$$
$$\text{precisao} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos positivos}}$$



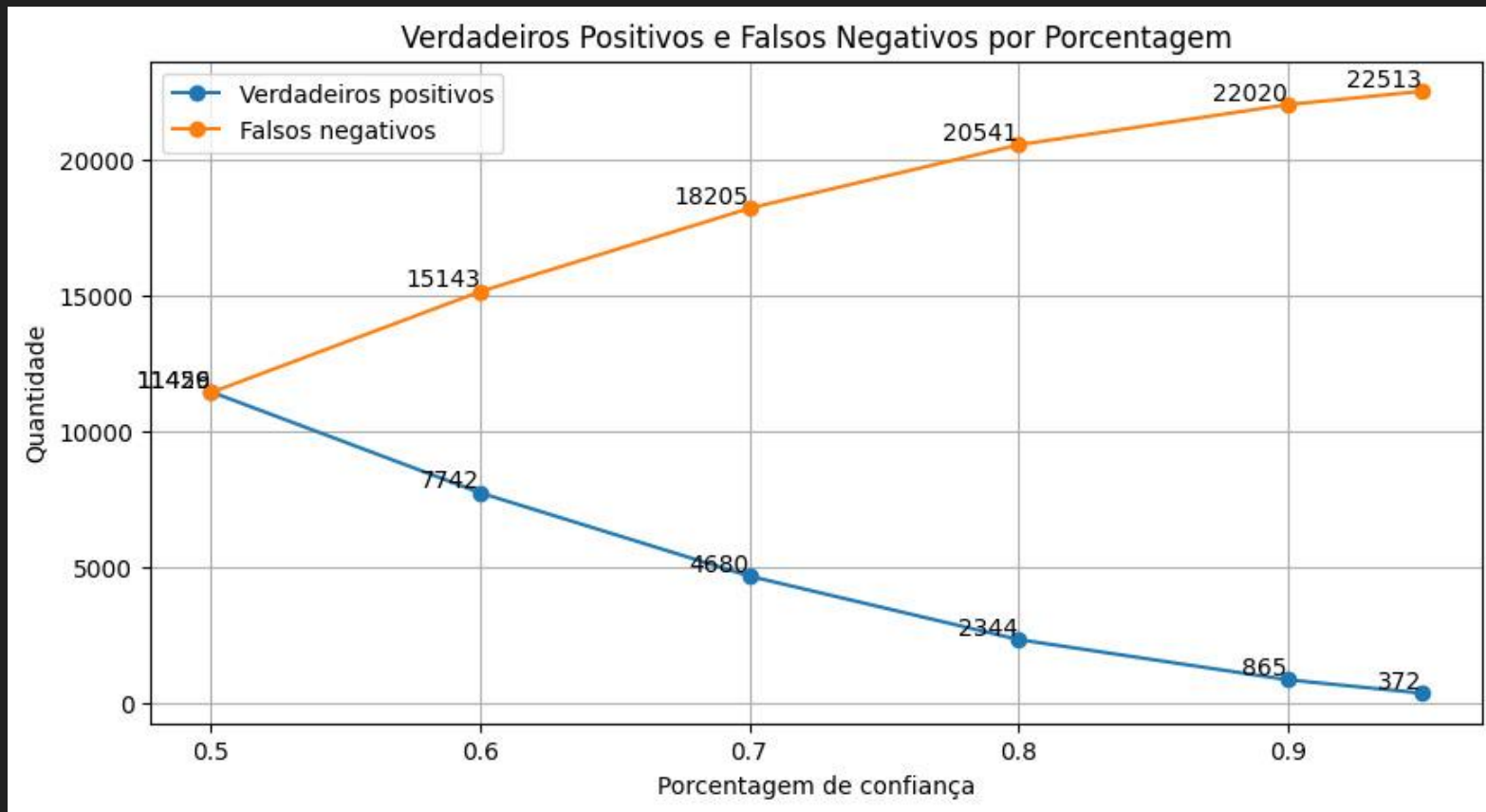
Análise de resultados e documentação de insights



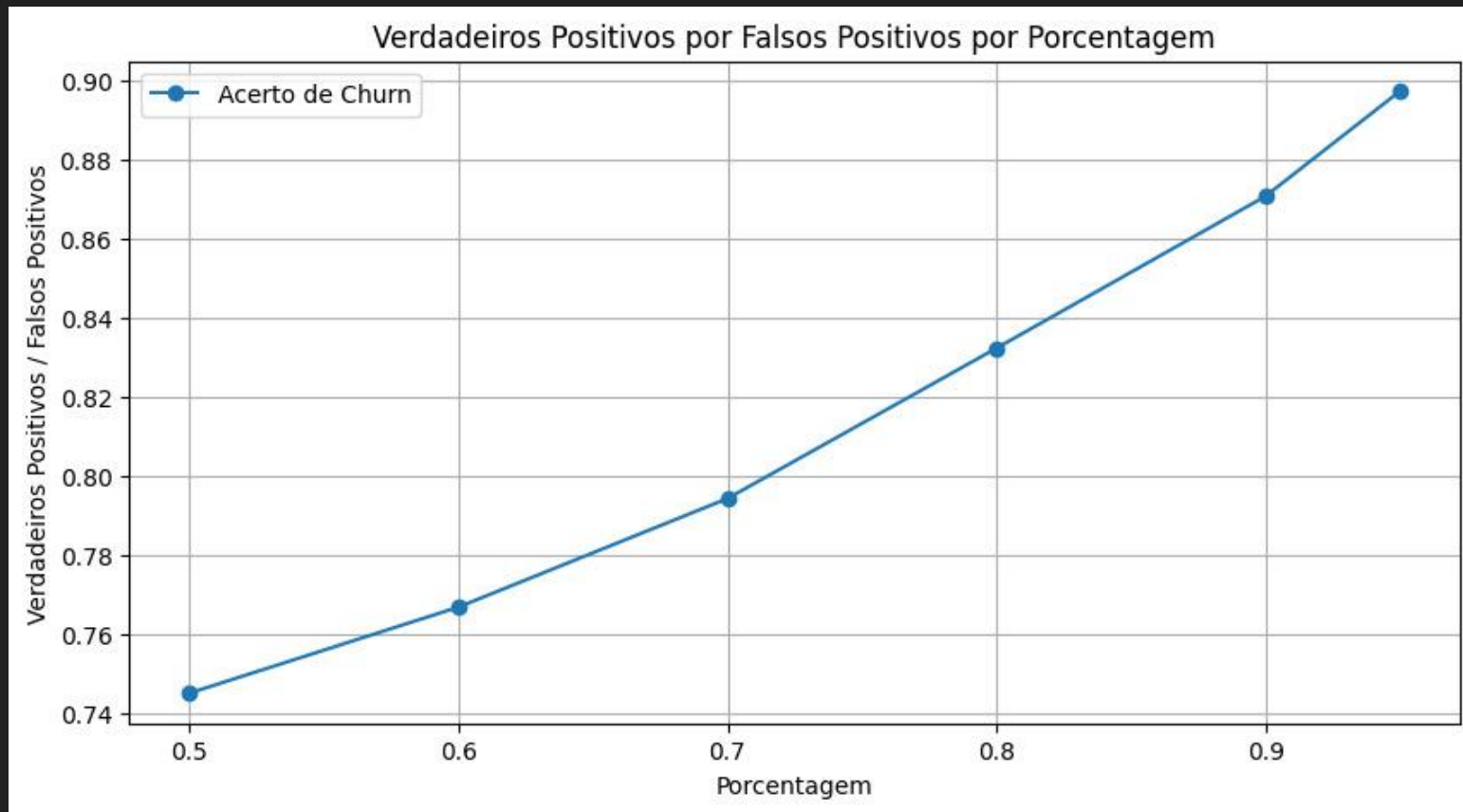
Análise de resultados e documentação de insights



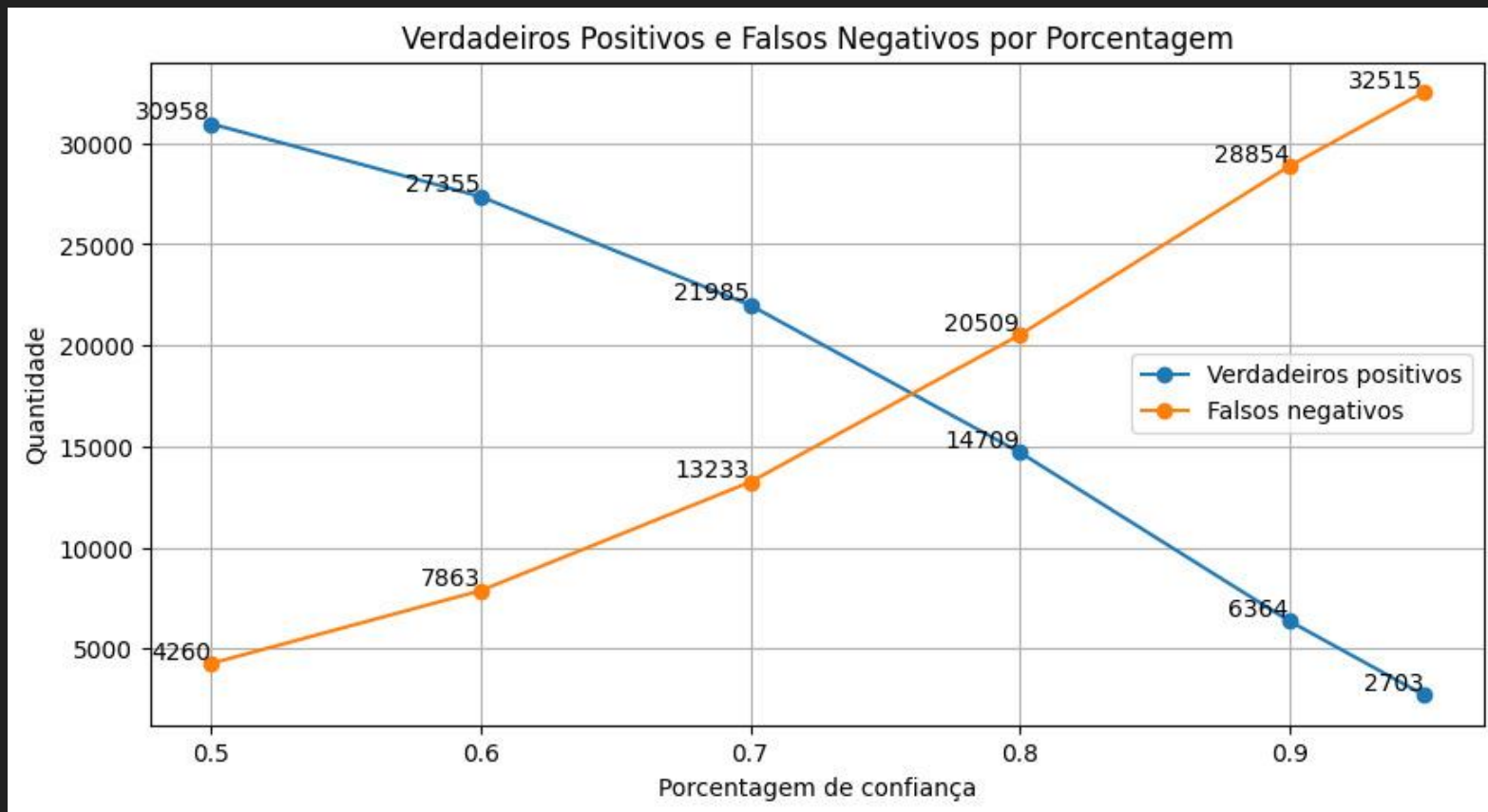
Análise de resultados e documentação de insights



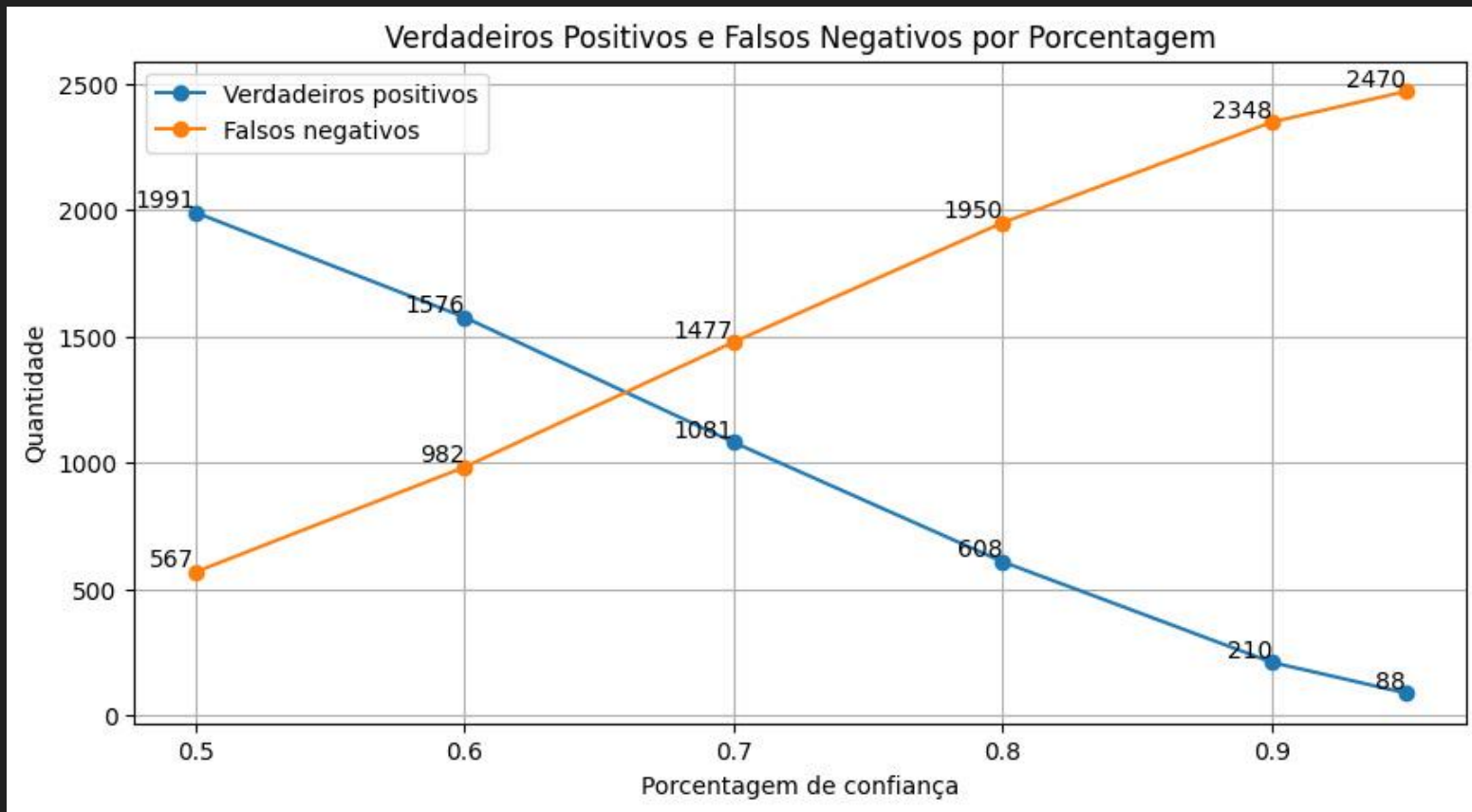
Análise de resultados e documentação de insights



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.