

# Modelo Preditivo Suicídio

## Uma análise do modelo

Com base no resultado e na natureza dos seus dados, o principal problema parece ser o **desequilíbrio de classes**.

1. **Desequilíbrio de classe:** Conjunto de dados altamente desequilibrado, conforme indicado pelo relatório de classificação:

- A classe 0 (*não-suicídio*) tem um grande número de amostras (155.221) .
- A classe 1 (*suicídio*) tem um número muito menor de amostras (1.427) .

Este desequilíbrio pode levar o modelo a inclinar-se para a classe majoritária (classe 0) e pode resultar num fraco desempenho de previsão para a classe minoritária (classe 1).

2. **Impacto no treino do modelo:**

- **Precisão:** A precisão global de 96,4% parece elevada, mas é distorcida pela classe dominante. É crucial olhar para além da precisão em conjuntos de dados desequilibrados.
- **Precisão e recuperação:** Para a classe 1 (suicídio), a precisão (0,06) e a recuperação (0,21) são baixas.

Isto sugere que o modelo identifica apenas uma pequena fração de instâncias positivas reais (baixa recuperação) e, quando prevê positivo, está frequentemente incorreto (baixa precisão).

3. **Consequências potenciais:**

- **Misclassificação:** O modelo pode classificar incorretamente os casos de suicídio (falsos negativos), perdendo potencialmente oportunidades de intervenção.
- **Preconceito em relação à classe majoritária:** O modelo pode ser demasiado otimista na previsão da classe majoritária (não suicida), negligenciando a classe minoritária.

4. **Exploração adicional:**

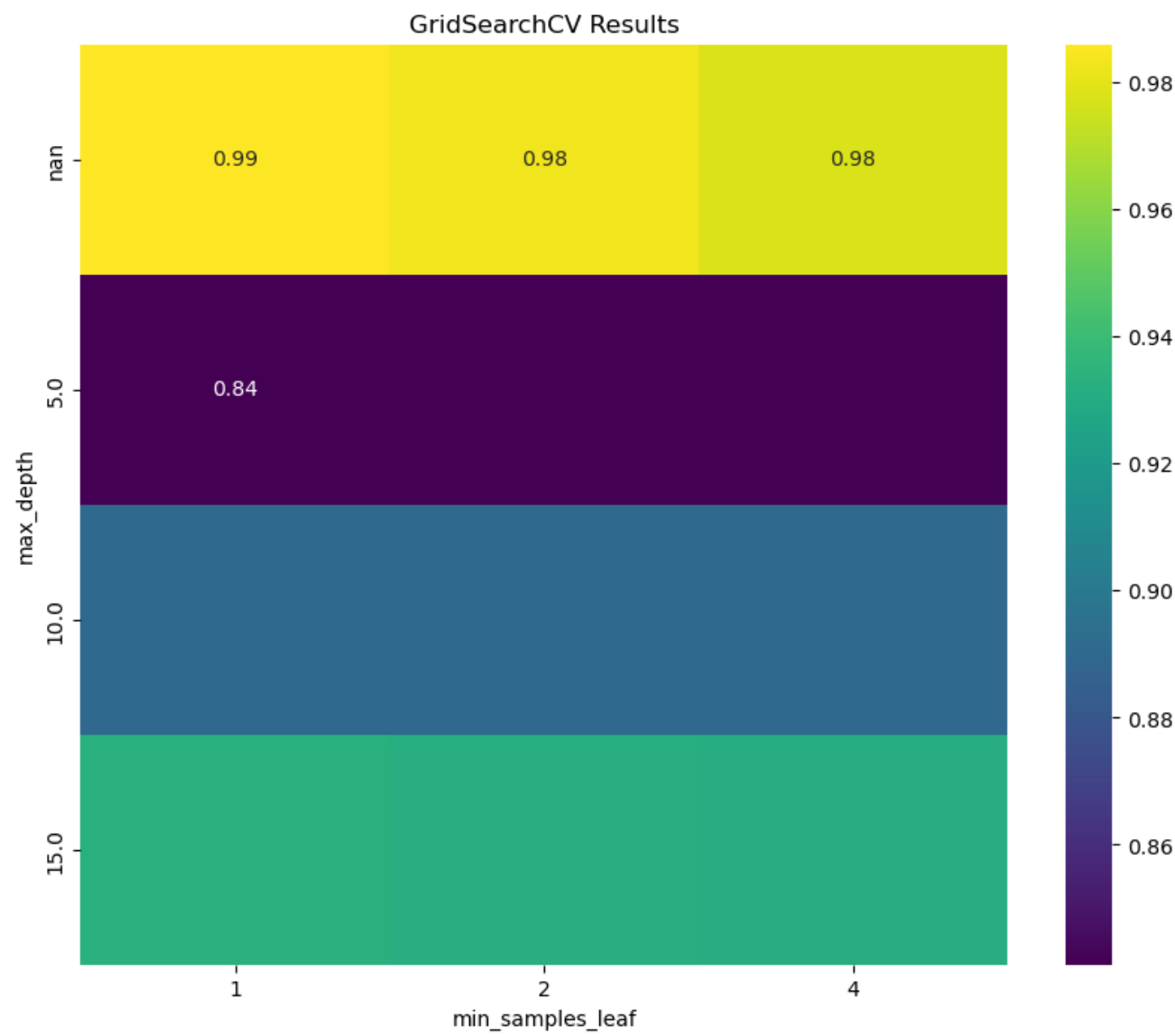
- Investigar a importância das características para garantir que o modelo está a concentrar-se em preditores relevantes.
- Considerar métodos de conjunto ou algoritmos diferentes que tratem melhor o desequilíbrio das classes, como o aumento do gradiente ou o conjunto de classificadores.

## Resultado

## Validação Parâmetros de Profundidade e Folhas

| Parâmetros      | Valores                                    |
|-----------------|--|
| Best parameters | {'max_depth': None, 'min_samples_leaf': 1} |
| Accuracy        | 0.9642638271794086                         |

Gráfico 1:

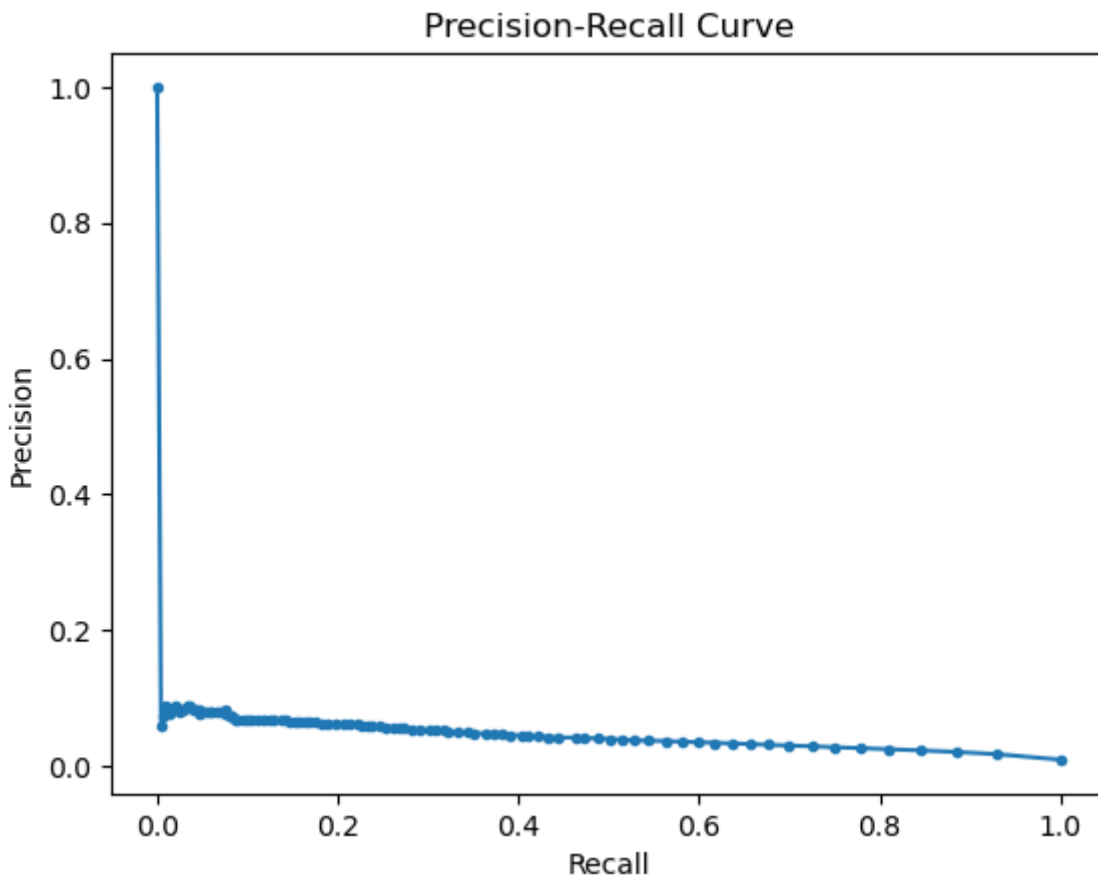


## Validação Cruzada

| Métricas      |           | Valores  |          |         |                               |
|---------------|-----------|--|----------|---------|-------------------------------|
| Accuracy      |           | 0.96   |          |         |                               |
| Macro avg     |           | 0.53 (Precision) / 0.59 (Recall) / 0.54 (F1-Score) |          |         |                               |
| Weighted avg  |           | 0.98 (Precision) / 0.96 (Recall) / 0.97 (F1-Score) |          |         |                               |
| Total Support |           | 156648   |          |         |                               |
| Class         | Precision | Recall   | F1-Score | Support |                               |
| 0             | 0.99      | 0.97   | 0.98     | 155221  | * Quantidade de Não Suicídios |

| Class | Precision | Recall | F1-Score | Support                        |
|-------|-----------|--------|----------|--------------------------------|
| 1     | 0.06      | 0.21   | 0.09     | 1427 * Quantidade de Suicídios |

Gráfico 2:



Em resumo, a questão principal é o desequilíbrio entre o número de amostras nas duas classes (suicidas vs. não suicidas), que afecta a capacidade do modelo para generalizar e prever eficazmente para a classe minoritária (Suicídio → Não (0) Sim (1)).

Isso é evidenciado pela baixa precisão, recall e f1-score para a classe minoritária (classe 1), enquanto a classe majoritária (classe 0) tem desempenho muito superior

## O que podemos fazer?

Existe uma grande diferença no número de exemplos entre as classes preditivas – que pode levar a modelos que tendem a favorecer a classe majoritária em detrimento da classe minoritária.

- Implicações nos Resultados:** O modelo parece ter uma alta precisão para a classe 0 (Não Suicídio) (99%), mas uma baixa recall (21%) para a classe 1 (Suicídio). Isso significa que o modelo identifica corretamente a maioria das instâncias da classe 0, mas perde muitas instâncias da classe 1, resultando em um f1-score baixo para a classe 1 (9%).
- Soluções para Desequilíbrio de Classes:**
  - **Reamostragem dos Dados:** Oversampling (aumento de exemplos da classe minoritária) e undersampling (redução de exemplos da classe majoritária).

Aqui podemos pegar os outros casos de suicídio da Bahia (BA) e Paraná (PR) registrados em anos anteriores aos do dataframe utilizado nesse modelo (2022, 2021, 2020 e assim por diante)

- **Peso das Classes:** Atribuir pesos diferentes às classes para penalizar erros na classe minoritária durante o treinamento.

Consigo imaginar o que isso significa, mas preciso estudar um pouco mais a respeito de como aplicar isso tecnicamente.

- **Métricas Apropriadas:** Além de accuracy, usar métricas como precision, recall, f1-score que são mais informativas em casos de desequilíbrio.
3. **Seleção do Modelo:** Dependendo do contexto e da importância de cada classe, pode ser necessário ajustar a métrica de avaliação ou escolher um modelo que seja mais robusto em lidar com classes desbalanceadas.

Pode ser que esse seja uma última opção

Considerando esses pontos, avaliar e ajustar o tratamento do desequilíbrio de classes pode melhorar significativamente o desempenho do seu modelo, especialmente na capacidade de detectar instâncias da classe minoritária com mais precisão e recall.