

# Análise Preditiva de Senioridade em Profissionais de Dados

Foco em Modelos de Machine Learning

Erika Pequeno (Ajuste o seu nome)

Pós-Graduação em Ciência de Dados e Inteligência Artificial (CDIA)

December 16, 2025

# 1. Introdução e Objetivo Central

- **Contexto:** Análise do mercado de trabalho e progressão de carreira na área de Dados.
- **Problema:** Identificar e modelar a relação entre variáveis demográficas/profissionais e o nível de Senioridade (Júnior, Pleno, Sênior).
- **Objetivo do ML:** Construir um modelo de classificação robusto capaz de prever o nível de senioridade de um profissional com base em suas características.

## 2. Metodologia: Dataset e Features

- **Fonte de Dados:** [Mencione a fonte dos dados, e.g., "Pesquisa XYZ sobre Mercado de Dados"].
- **Tamanho da Amostra:**  $N =$  [Insira o N total de amostras]
- **Target (Variável Dependente):** Senioridade (Júnior, Pleno, Sênior).
- **Principais Features (Variáveis Independentes):**
  - 1 Experiência Total (Anos)
  - 2 Número de Tecnologias Dominadas

importante

### 3. Pré-Processamento e Feature Engineering

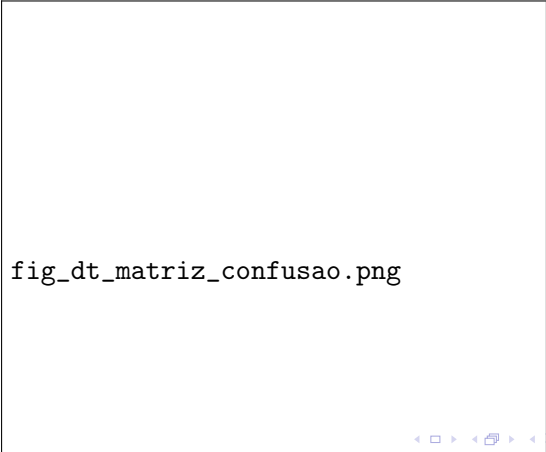
- **Tratamento de Missing Values:** Estratégia de imputação (e.g., Mediana para numéricas, Moda para categóricas).
- **Codificação de Variáveis Categóricas:** Uso de One-Hot Encoding ou Label Encoding.
- **Escalamento:** Aplicação de **StandardScaler** nas variáveis numéricas para padronização.
- **Balanceamento (Opcional):** Estratégia utilizada (se houver, e.g., SMOTE) para mitigar o desequilíbrio das classes.

## 4. Modelo Implementado: Árvore de Decisão (DT)

- **Critério:** Índice Gini ( $\text{Gini} = 1 - \sum_{i=1}^C p_i^2$ ) para medição de impureza.
- **Vantagem:** Alta interpretabilidade e não necessidade de escalonamento.
- **Hiperparâmetros Chave:**
  - 1 'max\_depth' = [Insira o valor usado]
  - 2 'min\_samples\_leaf' = [Insira o valor usado]

## 5. Resultados: Desempenho da Árvore de Decisão

- **Métrica de Avaliação:** Matriz de Confusão e Relatório de Classificação (Precision, Recall, F1-Score).
- **Acurácia Global:**  $\text{Acurácia} = [\text{Valor da DT}] \%$
- **F1-Score Médio:**  $\text{F1-Score (Weighted Avg)} = [\text{Valor da DT}]$



`fig_dt_matriz_confusao.png`

## 6. Resultados DT: Foco na Impureza (Gini)

- **Melhor Desempenho (Classe mais Pura): Sênior** (F1-Score  $\approx$  [Valor]).
- **Pior Desempenho (Maior Impureza): Pleno** (F1-Score  $\approx$  [Valor]).
- **Interpretabilidade:** [Mencione a feature mais importante da DT, e.g., "A Experiência Total foi o split inicial"].

## 7. Modelo Implementado: [Mencione o Modelo 2, e.g., Regressão Logística]

- **Motivação:** Modelo linear para comparação com a DT não linear.
- **Funcionamento:** Utiliza a função logística para modelar a probabilidade de um profissional pertencer a cada classe.
- **Parâmetros Chave:**
  - ① 'penalty' = [l2 ou none]
  - ② 'C' (Inverso da regularização) = [Valor usado]



## 8. Resultados: Desempenho da [Modelo 2]

- **Acurácia Global:**  $\text{Acurácia} = [\text{Valor do Modelo 2}] \%$
- **F1-Score Médio:**  $\text{F1-Score (Weighted Avg)} = [\text{Valor do Modelo 2}]$

fig\_modelo2\_matriz\_confusao.png

## 9. Modelo Implementado: [Mencione o Modelo 3, e.g., Random Forest]

- **Motivação:** Reduzir o \*overfitting\* inerente à DT e aumentar a performance.
- **Funcionamento:** Agregação de múltiplas Árvores de Decisão para tomar uma decisão por "votação".
- **Parâmetros Chave:**
  - 1 'n\_estimators' = [Número de árvores, e.g., 100]
  - 2 'max\_depth' = [Valor usado]

## 10. Resultados: Desempenho da [Modelo 3]

- **Acurácia Global:**  $\text{Acurácia} = [\text{Valor do Modelo 3}]\%$
- **F1-Score Médio:**  $\text{F1-Score (Weighted Avg)} = [\text{Valor do Modelo 3}]$

# 11. Análise Comparativa dos Modelos

Table: Comparação de Desempenho no Conjunto de Teste

Modelo	Acurácia (%)	F1-Score	Interpretabilidade
Árvore de Decisão	[Valor]	[Valor]	Alta
	[Valor]	[Valor]	Modelo 2 > Modelo 1
	[Valor]	[Valor]	[Baixa/Média] Modelo 3 > Modelo 2 > Modelo 1 Baixa

- **Modelo Vencedor:** [Mencione o modelo com melhor F1/Acurácia, e.g., Random Forest].
- **Trade-off:** [Comente o equilíbrio entre Performance vs. Interpretabilidade].

## 12. Importância dos Atributos (Feature Importance)

- Baseado no modelo [Modelo Vencedor].
- **Top 3 Fatores Preditivos:**
  - ① **[Feature 1]** ([Valor de Importância]%)
  - ② **[Feature 2]** ([Valor de Importância]%)
  - ③ **[Feature 3]** ([Valor de Importância]%)
- **Insight:** [Resuma o que essa importância significa, e.g., "A experiência é o principal determinante, seguido por habilidades específicas"].

# 13. Conclusões Finais

- O projeto demonstrou que a senioridade é previsível com um bom nível de acurácia utilizando [Modelo Vencedor].
- A maior limitação encontrada foi a **distinção da classe Pleno**, exigindo atenção em futuros desenvolvimentos.
- O resultado é aplicável para [Mencione o uso prático, e.g., "automatizar a triagem de currículos e mapeamento de carreira"].

## 14. Trabalhos Futuros

- **Experimentação Adicional:** Testar modelos mais avançados (e.g., SVM, XGBoost) ou Deep Learning.
- **Melhoria do Desempenho:** Aplicar técnicas avançadas de \*Hyperparameter Tuning\* (e.g., Grid Search, Random Search).
- **Validação Externa:** Aplicar o modelo em um novo conjunto de dados de validação para medir a capacidade de generalização.

Obrigado!

Perguntas?

Email: [Seu Email]