

# Fatores preditores do nível de senioridade na área de dados

Projeto Aplicado - Pós-Graduação em Ciência de Dados e Inteligência Artificial

Erika Pequeno, Dra

Centro Universitário SENAI/SC - Campus Florianópolis

Dezembro 2025

# 1. Entendimento do negócio (Business Understanding)

## CONTEXTUALIZAÇÃO, OBJETIVO DO PROJETO, RESULTADO ESPERADO

### Contextualização:

- Alta demanda por profissionais da área de dados;
- Análise do mercado de trabalho;
- Progressão de carreira na área de Dados.

### Objetivo do Projeto

Com base nas características de carreira (tempo de experiência em dados e TI, nível de ensino) e nas ferramentas tecnológicas utilizadas no dia a dia, quais são os principais fatores preditores do nível de senioridade (Júnior, Pleno, Sênior) para profissionais do gênero feminino no mercado de dados brasileiro? A importância desses fatores difere significativamente em comparação com profissionais do gênero masculino?

### Resultado Esperado

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:** Pesquisa State of Data Brazil 2024-2025.
- **Tamanho do dataframe tratado do projeto:** 3764 registros; 74 colunas;
- **Target (Variável Dependente):** Nível de Senioridade (Júnior, Pleno, Sênior).
- **Principais Features (Variáveis Independentes):**
  - 1 Dados demográficos (idade, gênero, UF onde reside,...)
  - 2 Dados sobre carreira (situação profissional, cargo atual,...)
  - 3 Conhecimentos na área de dados (linguagens de programação, ferramentas de ETL,...)
  - 4 Conhecimentos em Engenharia de Dados/DE (rotinas de DE realizadas, ferramentas de qualidade de dados);
  - 5 Conhecimentos em Análise de Dados/DA (rotinas de DA realizadas, ferramentas de autonomia área de negócios)
  - 6 Conhecimentos em Ciências de Dados/DS (rotinas de DS realizadas, tecnologias de DS,...)

### 3. Preparação dos dados

#### TRATAMENTO DE DADOS FALTANTES E OUTLIERS. ENGENHARIA DE ATRIBUTOS.

- **Tratamento de Missing Values:** remoção de linhas onde a variável target era nula; remoção de linhas onde as respostas eram prefiro não responder, ou similares; imputação de valor 0 em variáveis binárias, etc;
- **Codificação de Variáveis Categóricas:** Uso de One-Hot Encoding ou Label Encoding;
- **Escalamento:** Não foi necessário a aplicação de **StandardScaler** nas variáveis numéricas para padronização;

## 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:** simplicidade e interpretabilidade.
- **Hiperparâmetros Chave:**
  - ① 'max\_depth' = 5
  - ② 'min\_samples\_leaf' = default
- **Treino e Teste:** regra 80-20;

## 5. Resultados: Desempenho da Árvore de Decisão

- **Métrica de Avaliação:** Importância das Features; Relatório de Classificação (Precision, Recall, F1-Score); Matriz de Confusão.
- **Acurácia Média modelo 1:** Acurácia = 73%
- **Acurácia Média modelo 2:** Acurácia = 60%
- **Acurácia Média modelo 1:** Acurácia = 71%

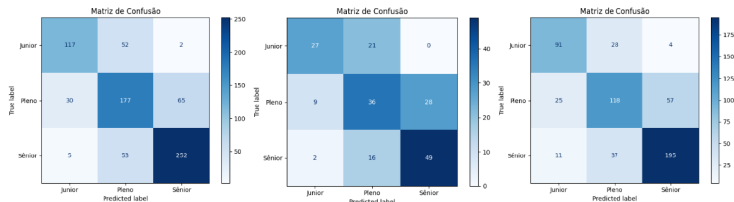


Figure: Matriz de Confusão da Árvore de Decisão.

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

- **Melhor Desempenho (Classe mais Pura): Sênior**  
(F1-Score  $\approx 0,80/0,68/0,78$ ).
- **Pior Desempenho (Maior Impureza): Pleno**  
(F1-Score  $\approx [0,64/0,49/0,62]$ ).
- **Interpretabilidade:** A feature mais importante (em qualquer modelo) da DT, é a faixa salarial, seguida do tempo de experiência em dados.

	feature	importance	feature	importance	feature	importance
12	15_faixa_salarial_Codificado	0.734386	15_faixa_salarial_Codificado	0.631215	15_faixa_salarial_Codificado	0.758688
13	16_tempo_de_experiencia_em_dados_Codificado	0.175995	16_tempo_de_experiencia_em_dados_Codificado	0.150912	16_tempo_de_experiencia_em_dados_Codificado	0.136157
0	2_idade	0.021185	2_idade	0.047845	13_cargo_atual_Codificado	0.019155
19	22_linguagem_mais_usada_Codificado	0.014744	19_experiencia_em_ti_Codificado	0.037225	40_da_sql_negocio	0.010845
7	9_situacao_profissional_Codificado	0.014363	9_situacao_profissional_Codificado	0.024559	8_area_de_formacao_Codificado	0.010254
11	13_cargo_atual_Codificado	0.009069	13_empresa_Codificado	0.020500	de_experiencia_em_ti_Codificado	0.009378
6	8_area_de_formacao_Codificado	0.008042	8_area_de_formacao_Codificado	0.018141	28_de_etl	0.008425
14	17_tempo_de_experiencia_em_ti_Codificado	0.007093	17_tempo_de_experiencia_em_ti_Codificado	0.015849	17_area_de_formacao_Codificado	0.008425
5	7_nivel_ensino_Codificado	0.004089	7_nivel_ensino_Codificado	0.015413	2_idade	0.006786
44	47_num_rotinas_ds	0.003248	48_ds_modelo1_regressao	0.011765	6_onde_reside_Codificado	0.006466

Figure: Importância das Features.

## 13. Conclusões Finais e Trabalhos Futuros

- O projeto demonstrou que a senioridade é fortemente influenciada por apenas duas features: o faixa salarial e o tempo de trabalho com dados;
- De forma geral, os modelos construídos são bons para classificar Seniores;
- A maior limitação encontrada foi a **distinção da classe Pleno**, exigindo atenção em futuros desenvolvimentos.
- A partir da acurácia de 73% do modelo do critério 1, pode-se afirmar que é um bom ponto de partida para criar um modelo, mas também indica que há espaço para melhoria;
- Um modelo melhorado é aplicável para, por exemplo, identificar nível de senioridade de candidatos em processos seletivos;



# Muito Obrigada!

Estou disponível para qualquer esclarecimento!

Email: [erika.pequeno@gmail.com](mailto:erika.pequeno@gmail.com)