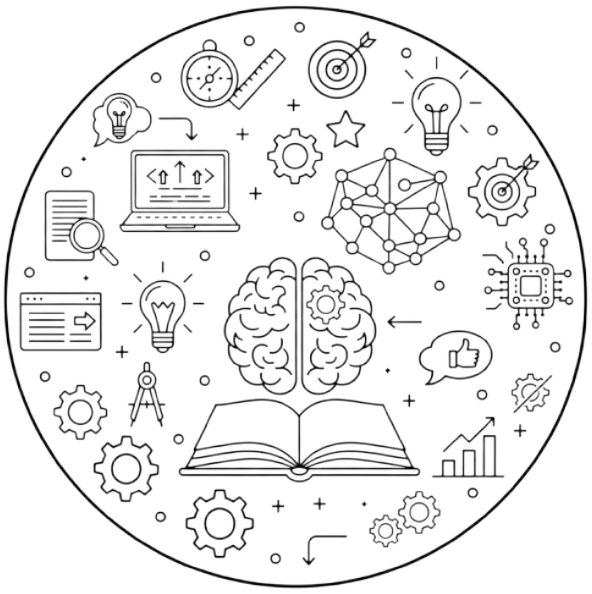


Fatores preditores do nível de senioridade na área de dados

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

12/2025

Dra Erika Pequeno

Florianópolis - SC - Brasil

Versão	Data	Autor(a)	Descrição
1	06/12/2025	Erika Pequeno	Versão inicial do relatório.
2	15/12/2025	Erika Pequeno	Versão final do relatório.

Sumário

1. Entendimento do negócio (Business Understanding)

1.1. CONTEXTUALIZAÇÃO

O volume de anúncios de emprego nas áreas de Engenharia de Dados, Análise de Dados e Ciência de Dados, conforme identificado em uma pesquisa rápida usando o Google, demonstra a alta demanda por profissionais da área (ver Anexo 1 com os primeiros 10 links para sites como LinkedIn Brasil, Glassdoor e Catho com centenas de vagas abertas no dia 06 de dezembro de 2025). Por outro lado, nessa mesma pesquisa Google podemos observar que as vagas são também disponibilizadas para diferentes níveis de maturidade dentro da área de atuação do profissional, desde Junior, passando pelo profissional Pleno, havendo lugar também para a disponibilização de vagas para profissionais Sênior. Para quem se candidata, e para quem está recrutando, características de carreira tais como tempo de experiência em dados e TI, nível de ensino, e conhecimento e uso de ferramentas tecnológicas utilizadas no dia a dia, podem, e devem, ser usadas para indicar o nível de maturidade profissional que se tem, ou que se deseja recrutar. Ademais, muitas vagas informam que a empresa oferece oportunidades iguais a todos os candidatos que atendam os requisitos de conhecimento e competência, independente de qualquer origem socioeconômica, raça, cor, nacionalidade, religião, gênero, idade, estado civil, condição de saúde, acessibilidade reduzida ou qualquer outra situação.

Atendendo ao que foi exposto, pode-se formular a seguinte pergunta:

> *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?*

A pergunta formulada, ao focar-se no gênero, visa especificamente examinar se o mercado de trabalho de dados, embora promova a igualdade de oportunidades, manifesta distinções nos fatores que predizem a senioridade para profissionais do gênero feminino em comparação com o masculino. Essa abordagem não apenas cumpre o requisito de limitação de escopo, mas também gera *insights* valiosos sobre a equidade e a progressão de carreira, tema de relevância no mercado de trabalho brasileiro. Embora a senioridade possa ser analisada sob a ótica de outras divisões (origem socioeconômica, nacionalidade ou condição de acessibilidade), a análise de gênero foi priorizada neste estudo para maximizar a interpretação dos resultados com base nos dados disponíveis na Pesquisa *State of Data Brazil* (“State of Data 2024-2025”, n.d.) (Lages et al. 2025).

1.2. OBJETIVO DO PROJETO

O objetivo do projeto é construir dois modelos de classificação, um para cada gênero (Feminino e Masculino), para prever o nível de senioridade de profissionais da área de dados. Para tal, serão considerados outros dados coletados na mesma pesquisa, tais como dados demográficos, dados sobre carreira e conhecimentos na área de dados. Após a construção dos dois modelos de classificação, pretende-se comparar a relevância dos

fatores considerados na construção dos modelos na determinação da senioridade para mulheres versus homens.

1.3. RESULTADO ESPERADO

Este projeto prevê três resultados:

1) Dois modelos de classificação, um aplicável a pessoas do gênero feminino e outro aplicável a pessoas do gênero masculino, para determinação do nível de senioridade de profissionais da área de dados a partir de outras variáveis coletadas na Pesquisa *State of Data Brazil*;
2) Identificar quais fatores (do conjunto de dados demográficos, de carreira ou de conhecimentos na área de dados) são mais relevantes em cada modelo e, consequentemente, se há fatores com pesos muito distintos de acordo com o gênero do indivíduo;
3) Identificar possíveis disparidades de carreira a partir dos fatores identificados como mais relevantes.

2. Entendimento dos dados (Data Understanding)

Para este projeto foram considerados os dados coletados no âmbito da pesquisa *State of Data Brazil* 2024-2025, um conjunto de dados público disponível no site Kaggle ("Kaggle: Your Machine Learning and Data Science Community", n.d.). Sobre os dados, sabe-se que foram coletados a partir de uma pesquisa aplicada entre novembro e dezembro de 2024, onde mais de 5,2 mil profissionais responderam ao *State of Data Brazil*, uma pesquisa anual sobre o mercado de trabalho em dados no Brasil. A pesquisa foi aplicada pela comunidade Data Hackers em parceria com a Bain & Company. Os dados usados neste projeto correspondem aos dados coletados na 5ª edição da pesquisa, pesquisa esta que teve seu início em 2019.

Nas próximas seções são apresentadas uma descrição geral dos dados e uma análise estatística exploratória dos dados que são usados neste projeto. Mais detalhes sobre a estrutura da pesquisa (listagem das perguntas colocadas e respostas possíveis) podem ser consultadas no arquivo do Anexo 2.

2.1. DESCRIÇÃO DOS DADOS

O questionário aplicado no âmbito da pesquisa *State of Data Brazil* 2024-2025 é particionado em 8 partes, cada uma contendo perguntas e, como resposta, opções de escolha. As 8 partes são:

- Parte 1 - Dados demográficos, contendo 13 perguntas, maioritariamente categóricas;
- Parte 2 - Dados sobre carreira, contendo 19 perguntas, maioritariamente categóricas;
- Parte 3 - Desafios dos gestores de times de dados, contendo 7 perguntas, todas categóricas, algumas com opção de múltipla escolha;
- Parte 4 - Conhecimentos na área de dados, contendo 13 perguntas categóricas, algumas com opções de múltipla escolha;
- Parte 5 - Objetivos na área de dados, contendo 4 perguntas categóricas, algumas com opção de adicionar categoria;

- Parte 6 - Conhecimentos em Engenharia de Dados/DE, contendo 8 perguntas, maioritariamente categóricas, algumas com opções de múltipla escolha;
- Parte 7 - Conhecimentos em Análise de Dados/DA, contendo 4 perguntas categóricas, todas com opções de múltipla escolha;
- Parte 8 - Conhecimentos em Ciências de Dados/DS, contendo 4 perguntas categóricas, todas com opções de múltipla escolha.

A importância de conhecer os dados e compreender a forma como foram coletados, e as implicações que a forma dessa coleta e armazenamento acarretam, é crucial para que se possa avaliar como os dados devem ser analisados e tratados. De forma simplificada, pode-se afirmar que a pesquisa contém 72 duas questões, o que corresponderia a 72 variáveis, majoritariamente categóricas. O grande número de variáveis categóricas que existem nos dados, e para os quais as questões associadas permitem a possibilidade de múltipla escolha (a seleção de mais de uma possibilidade), enriquece a quantidade de informação coletada. Todavia, dificulta o tratamento e análise dos dados, devido à possibilidade combinatória de respostas. Assim, o número de variáveis disponíveis a partir da análise direta dos dados e sem grandes transformações pode ser muito superior. Atendendo a todos estes pontos, a análise exploratória de dados irá considerar um subgrupo e variáveis, selecionadas a partir do que, por intuição e atendendo ao exercício acadêmico aqui apresentado, se pode considerar relevante para o objetivo da pesquisa.

2.2. VARIÁVEIS SELECIONADAS DA PESQUISA PARA A CONSTRUÇÃO DO MODELO

Relembrando, o objetivo da pesquisa é construir dois modelos de classificação, um para cada gênero (Feminino e Masculino), para prever o nível de senioridade de profissionais da área de dados. No conjunto de perguntas da pesquisa, uma das questões colocadas na parte sobre dados demográficos era o gênero do respondente, que poderia escolher uma de 4 respostas: Masculino; Feminino; Prefiro não informar; Outro. Assim, esta pergunta irá fornecer dados sobre a primeira variável a considerar no nosso modelo, que será o gênero. Ainda no conjunto das perguntas sobre dados demográficos (Parte 1), outras questões que podem ser pertinentes considerar nos modelos sobre os respondentes. Estas perguntas permitem definir o seguinte conjunto de variáveis: idade (variável numérica inteira); cor/raça/etnia (variável categórica nominal); pdc (variável categórica nominal); local de residência (variável categórica nominal); nível de ensino (variável categórica ordinal); área de formação (variável categórica nominal).

A partir do conjunto de perguntas relacionadas com os dados sobre carreira (Parte 2) podemos definir o seguinte conjunto de variáveis: situação profissional (variável categórica nominal); setor (variável categórica nominal); dimensão da empresa (variável categórica ordinal); gestor (variável binária/booleana); cargo atual (variável categórica nominal); nível de senioridade - **a nossa variável target** - (variável categórica ordinal); faixa salarial (variável categórica ordinal); tempo de experiência em dados (variável categórica ordinal); tempo de experiência em ti (variável categórica ordinal); modelo de trabalho atual (variável categórica nominal).

Do conjunto de dados relacionados com os desafios dos gestores de times de dados (Parte 3) não foi selecionada nenhuma pergunta cujos dados fossem considerados

relevantes para o objetivo deste projeto (impacto da escolha dos dados partindo do princípio da intuição).

A partir da Parte 4 da pesquisa (Conhecimentos na área de dados), as seguintes variáveis foram escolhidas: função de atuação (variável categórica nominal); número de fontes de dados dia-a-dia (variável numérica inteira); número de linguagens de programação dia-a-dia (variável numérica inteira); linguagem mais usada (variável categórica nominal); usa chatgpt ou copilot no trabalho (variável binária/boobleana).

Do conjunto de dados relacionados com os objetivos na área de dados (Parte 5) não foi selecionada nenhuma pergunta cujos dados fossem considerados relevantes para o objetivo deste projeto (impacto da escolha dos dados partindo do princípio da intuição)

Do conjunto de questões sobre os conhecimentos em Engenharia de Dados(DE) (Parte 6), foram selecionadas as seguintes variáveis: número de rotinas como DE (variável numérica); uso de ferramentas ETL como DE (variável binária/booleana); uso de ferramentas de qualidade de dados (variável binária/booleana); maior tempo gasto como DE (variável categórica nominal).

A partir do conjunto de perguntas sobre conhecimentos em Análise de Dados (DA) (Parte 7), foram selecionadas as seguintes variáveis: número de rotinas como DA (variável numérica); uso de ferramentas ETL como DA (variável binária/booleana); uso de ferramentas de autonomia da área de negócios (variável binária/booleana); maior tempo gasto como DA (variável categórica nominal).

Para finalizar, da Parte 8 da pesquisa, onde as questões estão relacionadas com conhecimentos em Ciências de Dados (DS), foram selecionadas as seguintes variáveis: número de rotinas como DS (variável numérica); técnicas e métodos de DS (variável categórica nominal); uso de tecnologias de DS (variável binária/booleana); maior tempo gasto como DS (variável categórica nominal).

Para as perguntas que buscam saber quais as ferramentas de ETL usadas, ferramentas de qualidades de dados, se usa chatgpt ou não, e outras similares, optou-se por transformar essas informações em informações binárias (*Sim* ou *Não*, *Usa* ou *Não Usa*) já que, muitas vezes, depende da empresa o uso ou não, dessas ferramentas, e que ferramentas podem ser usadas. O foco aqui não é saber o que se usa, mas se usa alguma ferramenta, visando capturar se o indivíduo tem maior ou menor conhecimento sobre rotinas e ferramentas na área de dados , e relacionar isso com o nível de senioridade.

Total de perguntas da pesquisa consideradas neste projeto: 7+10+0+5+0+4+4+4=34
Total de perguntas da pesquisa: 13 + 19 +7 +13 + 4 +8 +4 +4 =72

Neste projeto são usadas informações coletadas de aproximadamente 47% da pesquisa. Assim, a partir do dataset original já serão desconsideradas as informações coletadas a partir de algumas perguntas. Ademais, ao considerar algumas informações coletadas como sim ou não a partir da agregação das respostas dadas, o dataset usado neste projeto terá ainda um número de variáveis menor.

2.3. METADADOS DO PROJETO

O *metadados* do projeto, ou o *dicionário de dados do projeto*, é a informação que descreve, explica e fornece contexto sobre um conjunto de dados. No âmbito do nosso projeto, o dicionário de dados gerado a partir dos dados coletados pela pesquisa e das variáveis selecionadas e descritas na seção anterior é apresentado no Anexo 3. Ao analisar o dicionário de dados identificamos que o projeto aqui apresentado tem como conjunto

de dados um total de 33 variáveis, que contêm informações sobre dados demográficos, de carreira, conhecimentos gerais na área de dados e conhecimentos específicos em Engenharia de Dados, Análise de Dados e Ciência de Dados. A maioria das variáveis são categóricas, existindo também variáveis numéricas e binárias.

Outras informações que se devem considerar antes de realizar a análise exploratória de dados do conjunto de dados do projeto são:

1) Apenas foram considerados dos dados originais as respostas cujo gênero assinalado foi *Masculino* ou *Feminino*;
2) Apenas foram considerados dos dados originais as respostas para as quais a pergunta sobre ser uma pessoa com deficiência foram *Sim* ou *Não*;
3) Apenas foram considerados dos dados originais as respostas relativas ao nível de ensino do respondente estavam contidas na lista *Não tenho graduação formal*, *Estudante de Graduação*, *Graduação/Bacharelado*, *Pós-graduação*, *Mestrado*, *Doutorado ou Phd*;
4) Apenas foram considerados dos dados originais as respostas cuja situação de trabalho assinalada foi *Desempregado, buscando recolocação*; *Empreendedor ou Empregado (CNPJ)*; *Empregado (CLT)*; *Estagiário*; *Freelancer*; *Servidor Público*; *Somente Estudante (graduação)*; *Somente Estudante (pós-graduação)*; *Trabalho na área Acadêmica/Pesquisador*; *Vivo fora do Brasil e trabalho para empresa de fora do Brasil*; *Vivo no Brasil e trabalho remoto para empresa de fora do Brasil*.

Informações adicionais podem ser consultadas no dicionário de dados (anexo 3). No restante do documento, sempre que houver referência ao conjunto de dados, está-se considerando o conjunto de dados correspondente ao dicionário de dados apresentado no Anexo 3.

2.4. CRIAÇÃO DO CONJUNTO DE DADOS PARA O MODELO

Na construção do dataframe do projeto, associado ao dicionário de dados definido no Anexo 3, alguns tratamentos de dados faltantes já foram realizados. Por exemplo, para definir a variável *onde_reside*, foram removidas 3 linhas onde a resposta à pergunta relacionada com a Unidade Federal onde mora era nula e a resposta à pergunta “reside no Brasil” era verdadeira.

Para as colunas *4.b.1, 4.b.2, 4.b.3, 4.b.4, 4.b.5, 4.b.6, 4.b.7* e *4.b.8*, como elas são usadas para calcular a variável *num_fontes_dados*, sempre que nessas colunas houver um valor faltante, ele será substituído por 0.

As colunas *4.d.1, 4.d.2, 4.d.3, 4.d.4, 4.d.5, 4.d.6, 4.d.7,4.d.8, 4.d.9, 4.d.10, 4.d.11, 4.d.12, 4.d.13, 4.d.14* são usadas para contabilizar o número distinto de linguagens de programação que o respondente pode usar no dia a dia. De forma similar ao ocorrido no parágrafo anterior, todos os valores nulos/faltantes são transformados em zero. Ainda, a coluna *4.d.15*, que corresponde a responder “Não utilizo nenhuma das linguagens listadas”, foi desconsiderada, uma vez que o que se busca com esta variável é comparar, ou colocar na mesma categoria, respondentes com características similares, e não há informação da linguagem alternativa que está sendo usada.

Para o cálculo da variável *usa_chatgpt_ou_copilot*, considerados-e as informações das colunas *4.m.1, 4.m.2, 4.m.3, 4.m.4* e *4.m.5*. É importante notar que a presença de 1 na coluna *4.m.1* significa que o respondente concorda com a afirmação “Não uso soluções de

AI Generativa com foco em produtividade". Como a variável do dicionário de dados busca medir o uso de Inteligência Artificial (IA) para aumentar a produtividade, então usaremos apenas as respostas contidas nas colunas *4.m.2, 4.m.3, 4.m.4*, *4.m.5,* que correspondem à concordância do respondente em usar IA de alguma forma para aumentar a produtividade no trabalho. Para tal, e como estas colunas assumem os valores 0.0 e 1.0, será realizada a soma destes valores. Se a soma for superior a 1.0, *usa_chatgpt_ou_copilot* assume o valor *Sim*. Caso contrário, assume o valor *Não*. Também aqui, para tratar os dados faltantes, será atribuído o valor 0.

Para o cálculo das variáveis *num_rotinas_de*, *num_rotinas_da* e *num_rotinas_ds* agiu-se de forma similar ao caso do cálculo da variável *num_fontes_dados,* mas considerando as colunas *6.a.1* a *6.a.8*, *7.a.1* a *7.a.9* e *8.a.1* a *8.a.12*, respectivamente. Também aqui foram desconsideradas as colunas 6.a.9 e 7.a.10, por corresponderem a não executar nenhuma rotina de DE ou DA, respectivamente.

As variáveis *de_usa_ferramentas_etl, de_usa_ferramentas_qualidade_dados, da_usa_ferramentas_etl, da_usa_ferramentas_autonomia_area_de_negocios* e *ds_usa_tecnologias_ds* são calculadas usando uma rotina similar à rotina usada para o cálculo da variável *usa_chatgpt_ou_copilot*, considerando as colunas assinaladas no dicionário e dados.

Finaliza-se esta seção com a descrição geral do dataframe construído para o projeto. O data frame tem um total de 74 variáveis, No Anexo 4 está uma tabela com a relação de-para entre as colunas do dataframe original da Pesquisa (descarregado do site Kaggle) e o dataframe reduzido considerado neste projeto.

2.5. ANÁLISE EXPLORATÓRIA DE DADOS (EDA)

O primeiro resultado que se obtém da análise exploratória está relacionado com as dimensões do dataframe, o número de valores não nulos em cada coluna, e o tipo de dados de cada coluna. O dataframe do projeto tem 3764 registros, 74 colunas. Dentre estas colunas, apenas 29 colunas (1_token, 2_idade, 3_genero, 4_etnia, 5_pcd, 6_onde_reside, 7_nivel_ensino, 9_situacao_profissional, 10_setor_empresa, 11_dimensao_empresa, 12_gestor, 13_cargo_atual, 14_nivel_senioridade, 15_faixa_salarial, 16_tempo_de_experiencia_em_dados, 17_tempo_de_experiencia_em_ti, 18_modelo_de_trabalho_atual, 19_funcao_atuacao, 20_num_fontes_dados, 21_num_linguagens_prog, 23_usa_chatgpt_ou_copilot, 24_num_rotinas_de, 25_de_usa_ferramentas_etl, 26_de_usa_ferramentas_qualidade_dados, 35_num_rotinas_da, 36_da_usa_ferramentas_etl, 37_da_usa_ferramentas_autonomia_area_de_negocios, 47_num_rotinas_ds, 62_ds_usa_tecnologias_ds) não têm valores faltantes. Observa-se também que os tipos de dados da maioria das variáveis precisam ser redefinidos.

<class 'pandas.core.frame.DataFrame'>			
Index: 3764 entries, 0 to 5213			
Data columns (total 74 columns):			
# Column		Non-Null Count	Dtype
---	------	--------------	-----
0	1_token	3764 non-null	object
1	2_idade	3764 non-null	int64
2	3_genero	3764 non-null	object
3	4_etnia	3764 non-null	object
4	5_pcd	3764 non-null	object
5	6_onde_reside	3764 non-null	object
6	7_nivel_ensino	3764 non-null	object
7	8_area_de_formacao	3701 non-null	object
8	9_situacao_profissional	3764 non-null	object
9	10_setor_empresa	3764 non-null	object
10	11_dimensao_empresa	3764 non-null	object
11	12_gestor	3764 non-null	object
12	13_cargo_atual	3764 non-null	object
13	14_nivel_senioridade	3764 non-null	object
14	15_faixa_salarial	3764 non-null	object
15	16_tempo_de_experiencia_em_dados	3764 non-null	object
16	17_tempo_de_experiencia_em_ti	3764 non-null	object
17	18_modelo_de_trabalho_atual	3764 non-null	object
18	19_funcao_atuacao	3764 non-null	object
19	20_num_fontes_dados	3764 non-null	float64
20	21_num_linguagens_prog	3764 non-null	float64
21	22_linguagem_mais_usada	3380 non-null	object
22	23_usa_chatgpt_ou_copilot	3764 non-null	object
23	24_num_rotinas_de	3764 non-null	float64
24	25_de_usa_ferramentas_etl	3764 non-null	object
25	26_de_usa_ferramentas_qualidade_dados	3764 non-null	object
26	27_de_pipeline_python	912 non-null	float64

27	28_de_etl	912 non-null	float64
28	29_de_sql_negocio	912 non-null	float64
29	30_de_integracao_fontes	912 non-null	float64
30	31_de_solucoes_arquitetura_dados	912 non-null	float64
31	32_de_manutencao_repositorios	912 non-null	float64
32	33_de_modelagem_dados	912 non-null	float64
33	34_de_metadados	912 non-null	float64
34	35_num_rotinas_da	3764 non-null	float64
35	36_da_usa_ferramentas_etl	3764 non-null	object
36	37_da_usa_ferramentas_autonomia_area_de_negocios	3764 non-null	object
37	38_da_processametno_análise_python	1652 non-null	float64
38	39_da_dashboards	1652 non-null	float64
39	40_da_sql_negocio	1652 non-null	float64
40	41_da_extracao_api	1652 non-null	float64
41	42_da_modelos_estatisitico	1652 non-null	float64
42	43_da_etl	1652 non-null	float64
43	44_da_modelagem_dados	1652 non-null	float64
44	45_da_planilhas	1652 non-null	float64
45	46_da_analises_estatisticas	1652 non-null	float64
46	47_num_rotinas_ds	3764 non-null	float64
47	48_ds_modelol_regressao	764 non-null	float64
48	49_ds_modelo_classificacao	764 non-null	float64
49	50_ds_sisatema_recomendacao	764 non-null	float64
50	51_ds_metodos_bayesianos	764 non-null	float64
51	52_ds_tecnicas_nlp	764 non-null	float64
52	53_ds_metodos_estatisticos	764 non-null	float64
53	54_ds_markov_hmm	764 non-null	float64
54	55_ds_clusterizacao	764 non-null	float64
55	56_ds_series_temporais	764 non-null	float64
56	57_ds_modelos_RL	764 non-null	float64
57	58_ds_modelos_ML_fraude	764 non-null	float64
58	59_ds_visao_computacional	764 non-null	float64
59	60_ds_deteccao_de_churn	764 non-null	float64

60	61_ds_llms	764 non-null	float64
61	62_ds_usa_tecnologias_ds	3764 non-null	object
62	63_ds_estudos_Ad-hoc	764 non-null	float64
63	64_ds_tratamento_de_dados	764 non-null	float64
64	65_ds_reuniões_entregas	764 non-null	float64
65	66_ds_modelos_ML_producao	764 non-null	float64
66	67_ds_pipelines	764 non-null	float64
67	68_ds_manutencao_ML	764 non-null	float64
68	69_ds_dashboards	764 non-null	float64
69	70_ds_analises_estatisticas	764 non-null	float64
70	71_ds_etl	764 non-null	float64
71	72_ds_mlops	764 non-null	float64
72	73_ds_infra	764 non-null	float64
73	74_ds_llms_negocio	764 non-null	float64

O dataframe foi submetido a um processo de tratamento de dados faltantes e outliers, a fim de preparar os dados para aplicar a técnica de *Machine Learning* planejada. Mais análises exploratórias foram realizadas após o tratamento dos dados.

3. Preparação dos dados

O projeto prevê a criação de dois modelos de classificação, um aplicável a pessoas do gênero feminino e outro aplicável a pessoas do gênero masculino, para determinação do nível de senioridade de profissionais da área de dados. O modelo de Machine Learning que será aplicado será Árvore de Decisão, já que os modelos que se querem construir correspondem a problemas de Classificação Multiclasse (Júnior, Pleno, Sênior). Usando a técnica de Árvore de Decisão, e a partir da interpretação da Árvore de Decisão, será possível identificar quais fatores (colunas) são os mais importantes para determinar o nível de senioridade, o que gera um insight valioso e o resultado esperado para este projeto.

3.1. TRATAMENTO DE DADOS FALTANTES E OUTLIERS

Para atribuir o tipo de variável correta às variáveis do dataframe, a fim de evitar erros de execução, deve-se tratar os valores faltantes de forma adequada. Para as variáveis 8_area_de_formacao, 10_setor_empresa, 11_dimensao_empresa, 12_gestor, 13_cargo_atual, 14_nivel_senioridade, 15_faixa_salarial, 16_tempo_de_experiencia_em_dados, 17_tempo_de_experiencia_em_ti, 18_modelo_de_trabalho_atual, 22_linguagem_mais_usada,
que são variáveis categóricas (ou do tipo texto), os valores faltantes são mantidos como NaN. As variáveis

27_de_pipeline_python, 28_de_etl, 29_de_sql_negocio, 30_de_integracao_fontes, 31_de_solucoes_arquitetura_dados, 32_de_manutencao_repositorios, 33_de_modelagem_dados, 34_de_metadados, 38_da_processametno_análise_python, 39_da_dashboards, 40_da_sql_negocio, 41_da_extracao_api, 42_da_modelos_estatisitico, 43_da_etl, 44_da_modelagem_dados, 45_da_planilhas, 46_da_analises_estatisticas, 48_ds_modelol_regressao, 49_ds_modelo_classificacao, 50_ds_sisatema_recomendacao, 51_ds_metodos_bayesianos, 52_ds_tecnicas_nlp, 53_ds_metodos_estatisticos, 54_ds_markov_hmm, 55_ds_clusterizacao, 56_ds_series_temporais, 57_ds_modelos_RL, 58_ds_modelos_ML_fraude, 59_ds_visao_computacional, 60_ds_deteccao_de_churn, 31_ds_llms, 63_ds_estudos_Ad-hoc, 64_ds_tratamento_de_dados, 65_ds_reuniões_entregas, 66_ds_modelos_ML_producao, 67_ds_pipelines, 68_ds_manutencao_ML, 69_ds_dashboards, 70_ds_analises_estatisticas, 71_ds_etl, 72_ds_mlops, 73_ds_infra, 74_ds_llms_negocio
são variáveis do tipo booleano (Sim/Não, 0/1). Como têm valores faltantes, os valores 0 e 1 foram transformados em 0.0 e 1.0. Visto que são variáveis que representam a execução, ou não, de determinada atividade, onde 0 representa a não realização, estes valores faltantes são bem representados no conjunto de dados por 0. Essa foi a transformação aplicada a estes valores faltantes

Variáveis binárias com valores faltantes devem ser marcadas como 0, onde 0 representa o Não. As variáveis nessas condições são:
23_usa_chatgpt_ou_copilot, 25_de_usa_ferramentas_etl,
26_de_usa_ferramentas_qualidade_dados, 36_da_usa_ferramentas_etl,
37_da_usa_ferramentas_autonomia_area_de_negocios, 62_ds_usa_tecnologias_ds.
A variável 12_gestor é binária, mas do tipo booleano. Para esta, os valores faltantes foram definidos como False.

Analisando os dados usando a função .describe(), primeiro sem considerar as variáveis do tipo “object”, e depois olhando especificamente para este tipo de variáveis, tem-se os seguintes resultados:

index	count	mean	std	min	25%	50%	75%	max
2_idade	3764	31,25	6,83	18	27	30	35	66
20_num_fontes_dados	3764	3,23	1,55	0	2	3	4	8
21_num_linguagens_prog	3764	1,98	1,07	0	2	2	2	8
24_num_rotinas_de	3764	1,04	2,05	0	0	0	0	8
27_de_pipeline_python	3764	0,14	0,35	0	0	0	0	1
28_de_etl	3764	0,03	0,16	0	0	0	0	1
29_de_sql_negocio	3764	0,08	0,28	0	0	0	0	1
30_de_integracao_fontes	3764	0,01	0,08	0	0	0	0	1
31_de_solucoes_arquitetura_dados	3764	0,06	0,23	0	0	0	0	1
32_de_manutencao_repositorios	3764	0,03	0,16	0	0	0	0	1
33_de_modelagem_dado	3764	0,05	0,21	0	0	0	0	1

s								
34_de_metadados	3764	0,02	0,15	0	0	0	0	1
35_num_rotinas_da	3764	1,55	2,10	0	0	0	3	9
38_da_processametno_análise_python	3764	0,13	0,33	0	0	0	0	1
39_da_dashboards	3764	0,24	0,43	0	0	0	0	1
40_da_sql_negocio	3764	0,20	0,40	0	0	0	0	1
41_da_extracao_api	3764	0,01	0,10	0	0	0	0	1
42_da_modelos_estatisitico	3764	0,02	0,14	0	0	0	0	1
43_da_etl	3764	0,02	0,15	0	0	0	0	1
44_da_modelagem_dados	3764	0,03	0,16	0	0	0	0	1

A partir da tabela acima observa-se que os valores que as variáveis numéricas assumem estão de acordo com o que foi definido no dicionário de dados.

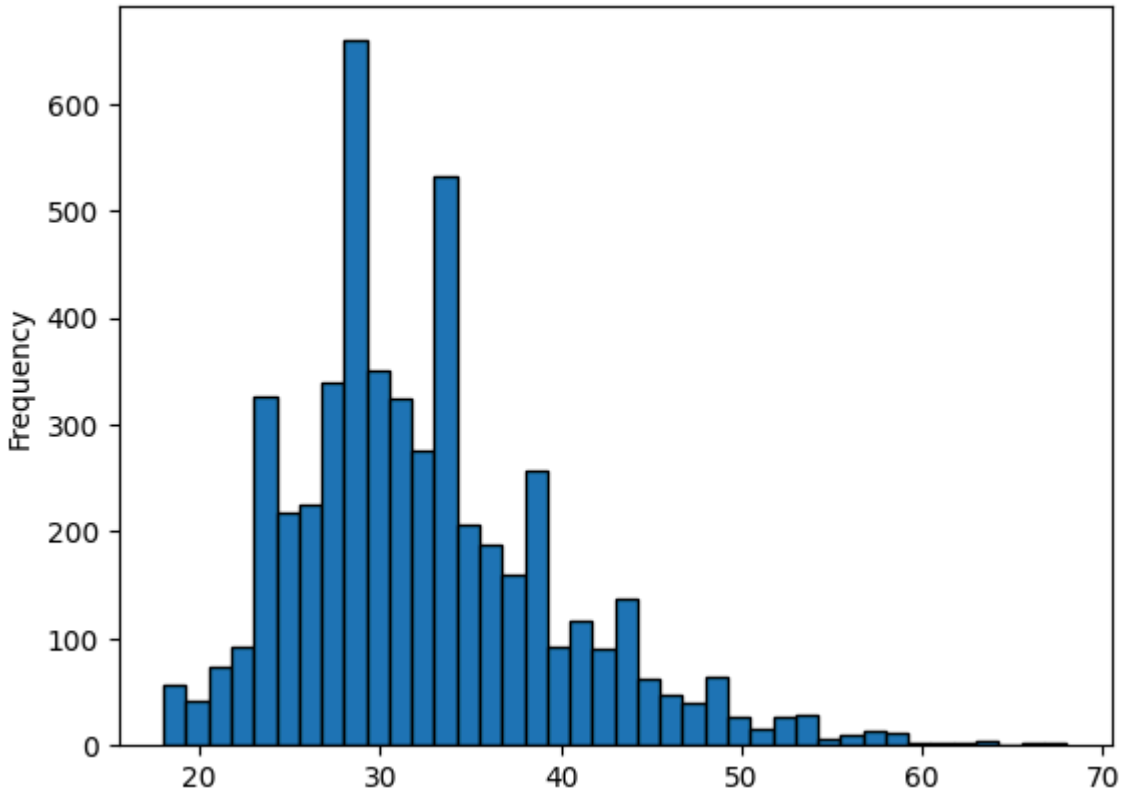
index	count	unique	top	freq
1_token	3764	3762	lb7gt5hdqqxuguv2lb7gto44mpk3ejha	2
3_genero	3764	2	Masculino	2827
4_etnia	3764	7	Branca	2500
5_pcd	3764	2	Não	3651
6_onde_reside	3764	26	SP	1418
7_nivel_ensino	3764	6	Graduação/Bacharelado	1355
8_area_de_formacao	3764	9	Computação / Engenharia de Software / Sistemas de Informação/ TI	1504
9_situacao_profissional	3764	7	Empregado (CLT)	3024
10_setor_empresa	3764	21	Finanças ou Bancos	847
11_dimensao_empresa	3764	8	Acima de 3.000	1841
13_cargo_atual	3764	15	Analista de Dados/Data Analyst	949
14_nivel_senioridade	3764	3	Sênior	1549
15_faixa_salarial	3764	13	de R$ 8.001/mês a R$ 12.000/mês	946
16_tempo_de_experiencia_em_dados	3764	7	de 3 a 4 anos	1216

17_tempo_de_experiencia_em_ti	3764	7	Não tive experiência na área de TI/Engenharia de Software antes de começar a trabalhar na área de dados	2108
18_modelo_de_trabalho_atual	3764	4	Modelo 100% remoto	1830
19_funcao_atuacao	3764	5	Análise de Dados	1652
22_linguagem_mais_usada	3764	15	SQL	1746
23_usa_chatgpt_ou_copilot	3764	2	Sim	3353
25_de_usa_ferramentas_etl	3764	2	Não	2869

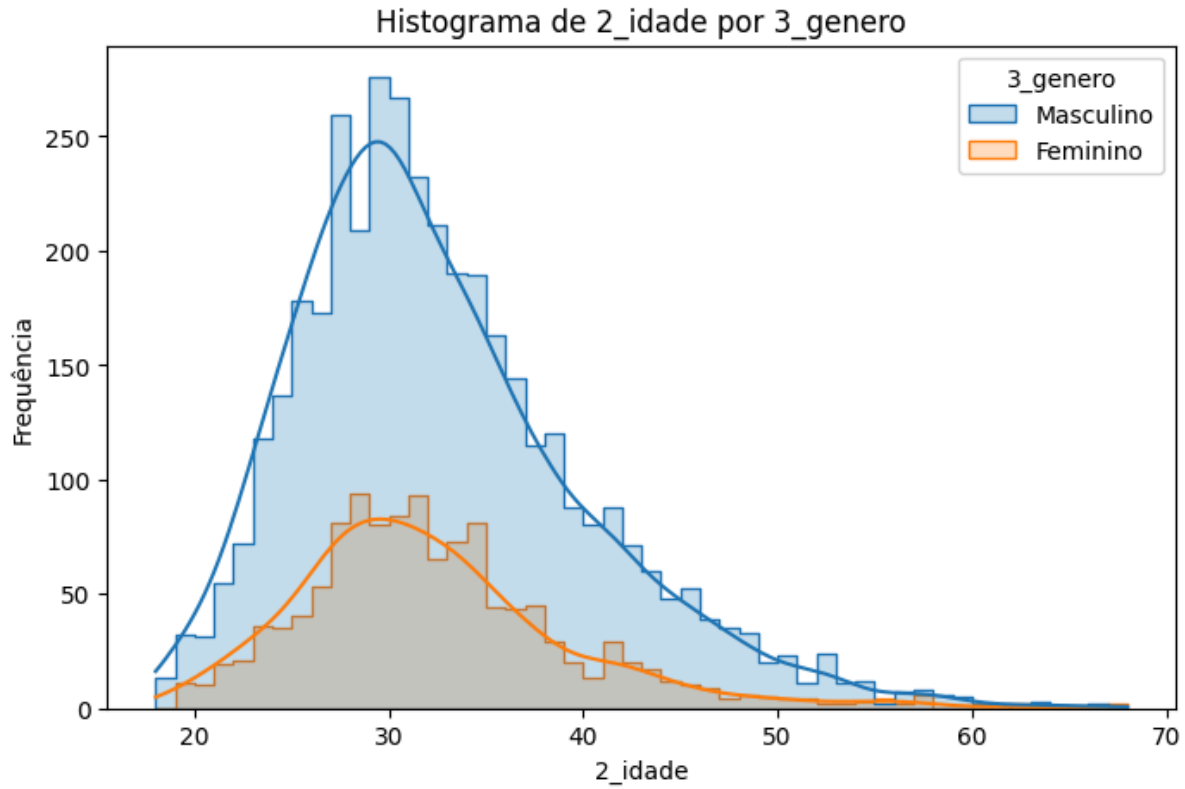
A partir da tabela anterior, observa-se que variável *1_token* aparece duplicada, o que não é esperado, já que se trata de um identificador. Assim, procedeu-se à remoção das duplicatas, mantendo apenas o último cadastro relativo aos tokens duplicados. Todas as restantes variáveis categóricas apresentam valores em concordância com o que foi definido no dicionário de dados.

IMPORTANTE: Devido à remoção das duplicatas, o dataframe do meu projeto está, neste momento, com um total de 5133 registros (linhas).

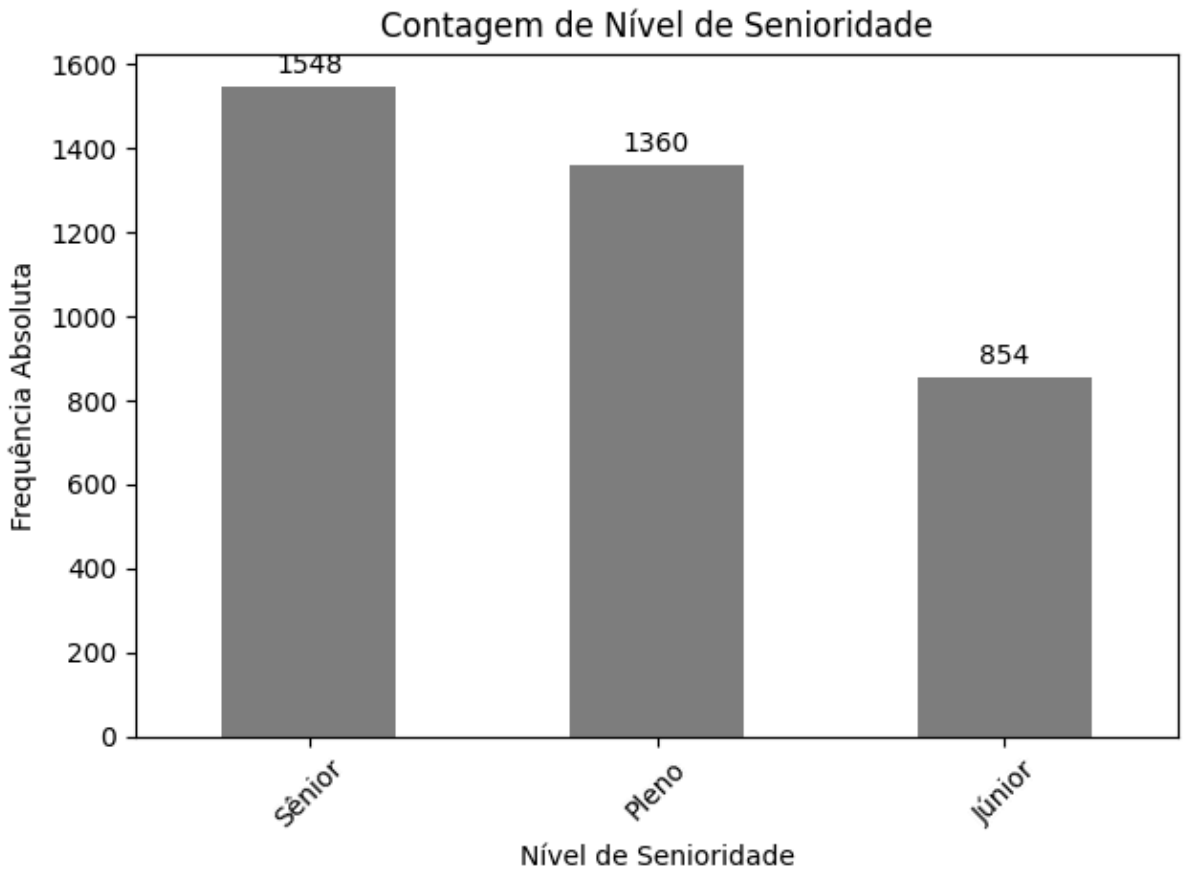
Para finalizar a etapa de tratamento de dados faltantes e outliers, e complementando a análise estatística exploratória, a figura seguinte apresenta um histograma da distribuição das idades dos respondentes no âmbito do nosso projeto:



A partir do histograma da figura xxx, observa-se que a maior parte dos dados está concentrada na faixa de idades mais jovens. Além disso, a cauda do gráfico se estende mais para a direita (idades mais velhas). Isso indica que a média das idades é um pouco maior do que a mediana. De fato, a distribuição é assimétrica à direita (o maior número de observações encontra-se à direita do gráfico). Relativamente a valores de tendência central, observa-se também que existem dois picos: um em torno dos 28 anos; um outro em torno dos 33 anos. Isto pode sugerir que a amostra tem maiores concentrações de indivíduos em torno destas duas idades. Por fim, e usando também as informações que constam da tabela apresentada anteriormente, as idades mínima e máxima são, respectivamente, 18 e 68 anos.



A partir do histograma da figura acima que segmenta a distribuição da idade pelo gênero, observa-se que a amostra é majoritariamente Masculina, dado que a curva de frequência azul atinge um pico de aproximadamente 275, enquanto a curva Laranja (Feminino) atinge um pico próximo de 100. Ademais, para ambos os grupos, a distribuição é assimétrica à direita, com a concentração principal nas idades mais jovens e uma cauda longa se estendendo para as idades mais velhas (até cerca de 70 anos). No que diz respeito à tendência central, o grupo Masculino apresenta seu maior pico de frequência em torno dos 28-29 anos, enquanto o grupo Feminino tem seu pico principal ligeiramente mais tarde, em torno dos 30-32 anos.



Acima podemos observar a frequência de indivíduos em cada uma das categorias de senioridade.

3.2. ENGENHARIA DE ATRIBUTOS

No dicionário de dados, a maioria das variáveis são categóricas nominais. Para preparar os dados a fim de aplicar a técnica de Árvore de Decisão, deve ser considerada a aplicação da técnica de *One-Hot Encoding*, já que esta é a técnica mais indicada para este tipo de variáveis. Relembrando, funciona da seguinte forma:

1. Para cada categoria única na coluna original, é criada uma nova coluna binária (dummy);
2. O valor em cada nova coluna é 1 se a observação pertencer àquela categoria e 0 caso contrário.

A vantagem é que evita que o modelo interprete erroneamente as categorias como tendo uma relação ordinal. Por outro lado, se houver muitas categorias únicas (alta cardinalidade), o *dataframe* pode ficar muito grande e esparso ("maldição da dimensionalidade").

NOTA IMPORTANTE: o conjunto de dados originais já previa a aplicação da técnica de One-Hot Encoding aos dados. Por esse motivo, não foi necessário aplicar essa técnica ao dataframe, restando apenas tratar o formato de dados, que acabou desconfigurando durante o upload dos dados, e tratar os dados faltantes.

Para as variáveis categóricas ordinais (como por exemplo, nível de ensino ou dimensão da empresa), é considerada a codificação *Ordinal Encoding / Label Encoding*, já que irá permitir imputar uma ordem lógica/ hierárquica nos dados. Relembrando, funciona da seguinte forma:

1. Cada categoria única é substituída por um único número inteiro;
2. É crucial que se atribua os números na ordem correta (ex: Graduação/Bacharelado=1, Mestrado=2, Doutorado ou Phd=3).

A codificação das variáveis categóricas ordinais usando *Ordinal Encoding / Label Encoding* ao invés de *One-Hot Encoding* permite ter uma maior eficiência de memória, já que cria apenas uma coluna[1].

3.3. CONJUNTOS DE TREINO E TESTE

A definição do conjunto de treino e teste é realizada a partir da separação dos dados do dataframe criado seguindo a regra 80-20: 80% dos dados serão usados para treino, enquanto 20% dos dados serão usados para teste.

4. Técnica de aprendizado de máquina

4.1. MODELO DE APRENDIZADO DE MÁQUINA ESCOLHIDO

O modelo de aprendizado de máquina escolhido foi Árvore de Decisão, já que o objetivo é fazer uma classificação e o treinamento é supervisionado. A principal razão para escolhê-la é sua simplicidade e interpretabilidade, já que é fácil visualizar e explicar as regras de decisão, o que não acontece com modelos mais complexos. Como o que queremos é comparar dois “classificadores” distintos, esta característica das árvores de decisão torna ainda mais simples a comparação.

4.2. TREINAMENTO DO MODELO DE APRENDIZADO DE MÁQUINA

Todos os detalhes da implementação e treinamento da solução podem ser consultados no arquivo.py, que se encontra documentado. Todavia, as principais informações serão explanadas aqui.

4.3. INTERPRETAÇÃO DO MODELO ESCOLHIDO

No decorrer da execução do trabalho optei por construir 3 árvores de decisão, de acordo com os seguintes critérios:

1) usando todo o dataframe tratado, sem considerar a separação por gênero;
2) usando as linhas do dataframe tratado correspondentes a respondentes do sexo feminino;
3) usando as linhas do dataframe tratado correspondentes a respondentes do sexo masculino.

[1] Há ainda as codificações *Target* e *Frequency Encoding*, aplicáveis quando a variável categórica apresenta muitas categorias.

O foco desta decisão foi tentar identificar, observando as três árvores construídas, se a variável gênero teria um impacto significativo ou não na classificação do nível de senioridade do respondente.

No que se segue são apresentados os resultados para **a árvore de decisão do critério 1**, para uma árvore cujo parâmetro de profundidade foi definido em 5 unidades.

- Nos nós folha finais, para a folha que apresenta a amostra de maior dimensão (137 registros) para a classificação JÚNIOR, o Índice de Gini é 0,014, o que é um valor interessante para o que se busca com este método;
- Nos nós folha finais, para a folha que apresenta a amostra de maior dimensão (377 registros) para a classificação PLENO, o Índice de Gini é 0,395, que não é um valor que demonstre segurança na qualificação;
- Nos nós folha finais, para a folha que apresenta a amostra de maior dimensão (276 registros) para a classificação SÊNIOR, o Índice de Gini é 0,063, o que é um valor interessante para o que se busca com este método;
- Apenas duas features apresentam importância superior a 10% no modelo: a feature que mapeia a faixa salarial e a feature que mapeia o tempo de experiência em dados;
- As duas principais features explicam 90% dos resultados do modelo.

No que se segue são apresentados os resultados para a **árvore de decisão do critério 2**, para uma árvore cujo parâmetro de profundidade foi definido em 5 unidades.

- Nos nós folha finais, para a folha que apresenta a amostra de maior dimensão (80 registros) para a classificação JÚNIOR, o Índice de Gini é 0,025, o que é um valor interessante para o que se busca com este método;
- Nos nós folha finais, para a folha que apresenta a amostra de maior dimensão (120 registros) para a classificação PLENO, o Índice de Gini é 0,385, que não é um valor que demonstre segurança na qualificação;
- Nos nós folha finais, para a folha que apresenta a amostra de maior dimensão (99 registros) para a classificação SÊNIOR, o Índice de Gini é 0,537, que não é um valor que demonstre segurança na qualificação;
- Apenas duas features apresentam importância superior a 10% no modelo: a feature que mapeia a faixa salarial e a feature que mapeia o tempo de experiência em dados;
- As duas principais features explicam 78% dos resultados do modelo.

No que se segue são apresentados os resultados para a **árvore de decisão do critério 3**, para uma árvore cujo parâmetro de profundidade foi definido em 5 unidades.

- Nos nós folha finais, para a folha que apresenta a amostra de maior dimensão (210 registros) para a classificação JÚNIOR, o Índice de Gini é 0,401, o que não é um valor interessante para o que se pretende com este método;
- Nos nós folha finais, para a folha que apresenta a amostra de maior dimensão (269 registros) para a classificação PLENO, o Índice de Gini é 0,386, que não é um valor que demonstre segurança na qualificação;
- Nos nós folha finais, para a folha que apresenta a amostra de maior dimensão (436 registros) para a classificação SÊNIOR, o Índice de Gini é 0,104, o que é um valor interessante para o que se busca com este método;

- Apenas duas features apresentam importância superior a 10% no modelo: a feature que mapeia a faixa salarial e a feature que mapeia o tempo de experiência em dados;
- As duas principais features explicam 88% dos resultados do modelo.

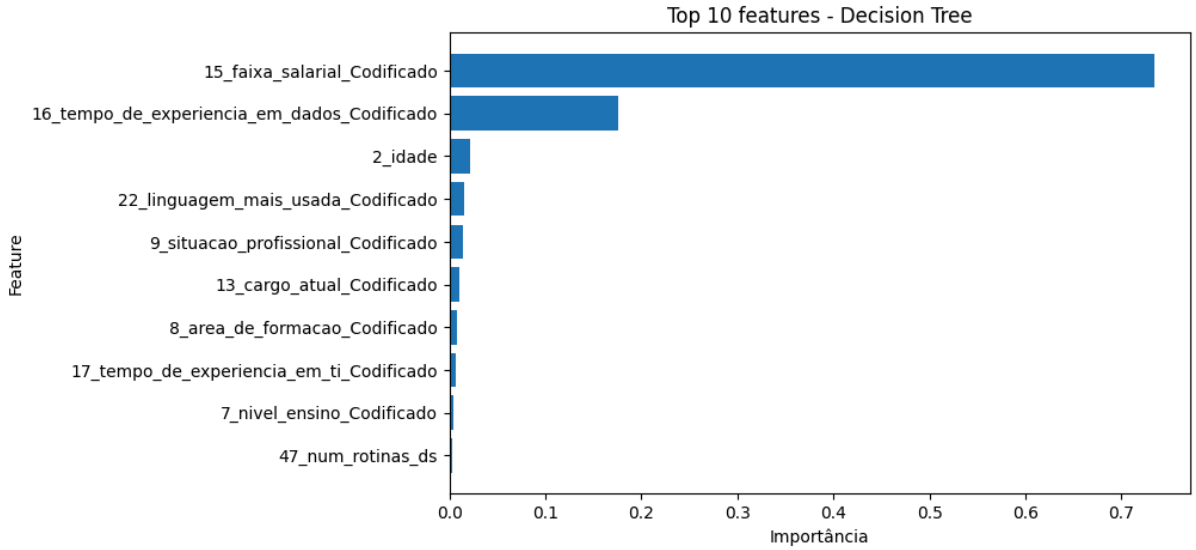
5. Avaliação do modelo de aprendizado de máquina

5.1. MÉTRICAS DE ACURÁCIA E PERFORMANCE

Considerando max_depth = 5 e na árvore de decisão do critério 1 (todo o dataframe criado)

- importância das features:

	feature	importance
12	15_faixa_salarial_Codificado	0.734386
13	16_tempo_de_experiencia_em_dados_Codificado	0.175995
0	2_idade	0.021185
19	22_linguagem_mais_usada_Codificado	0.014744
7	9_situacao_profissional_Codificado	0.014363
11	13_cargo_atual_Codificado	0.009869
6	8_area_de_formacao_Codificado	0.008042
14	17_tempo_de_experiencia_em_ti_Codificado	0.007093
5	7_nivel_ensino_Codificado	0.004089
44	47_num_rotinas_ds	0.003248



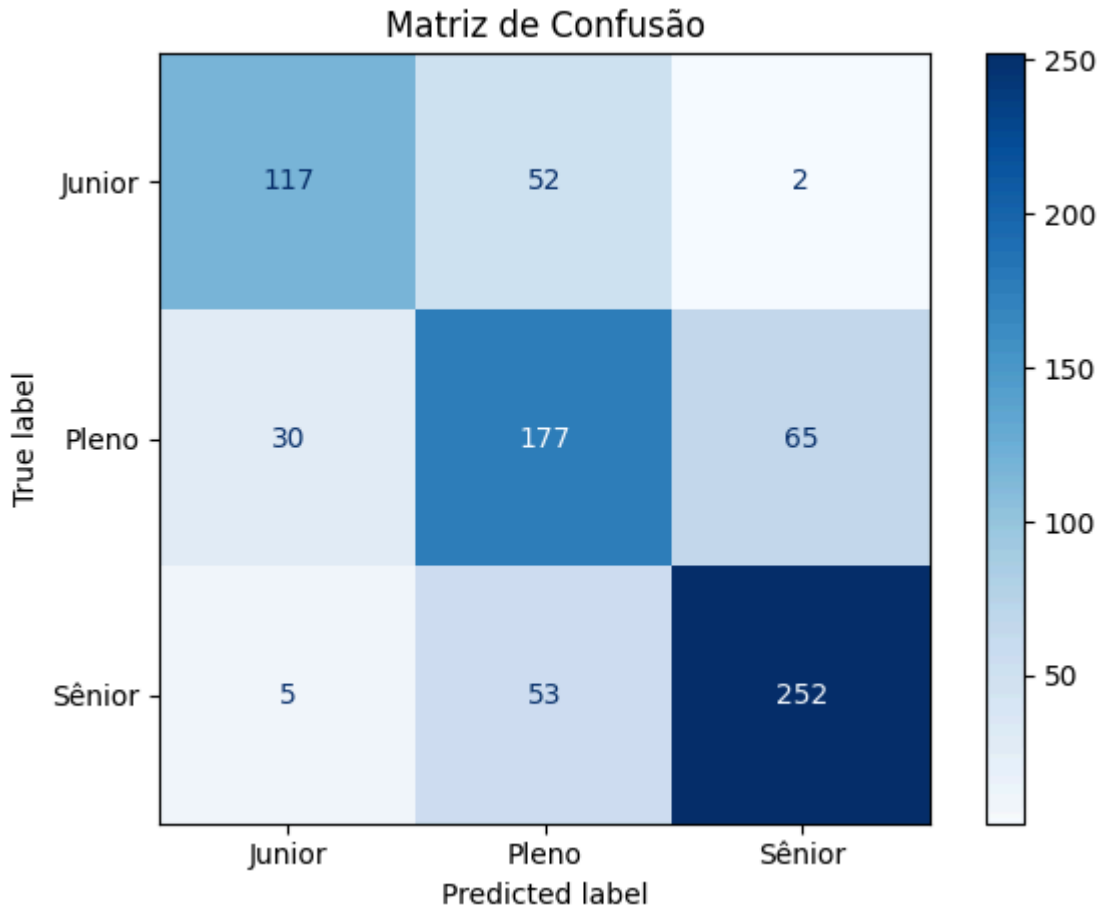
=> Claramente, das 74 features, apenas duas são relevantes.

- Precisão, Recall e F1-score

	precision	recall	f1-score	support
Júnior	0.77	0.68	0.72	171
Pleno	0.63	0.65	0.64	272
Sênior	0.79	0.81	0.80	310
accuracy			0.73	753
macro avg	0.73	0.72	0.72	753
weighted avg	0.73	0.73	0.73	753

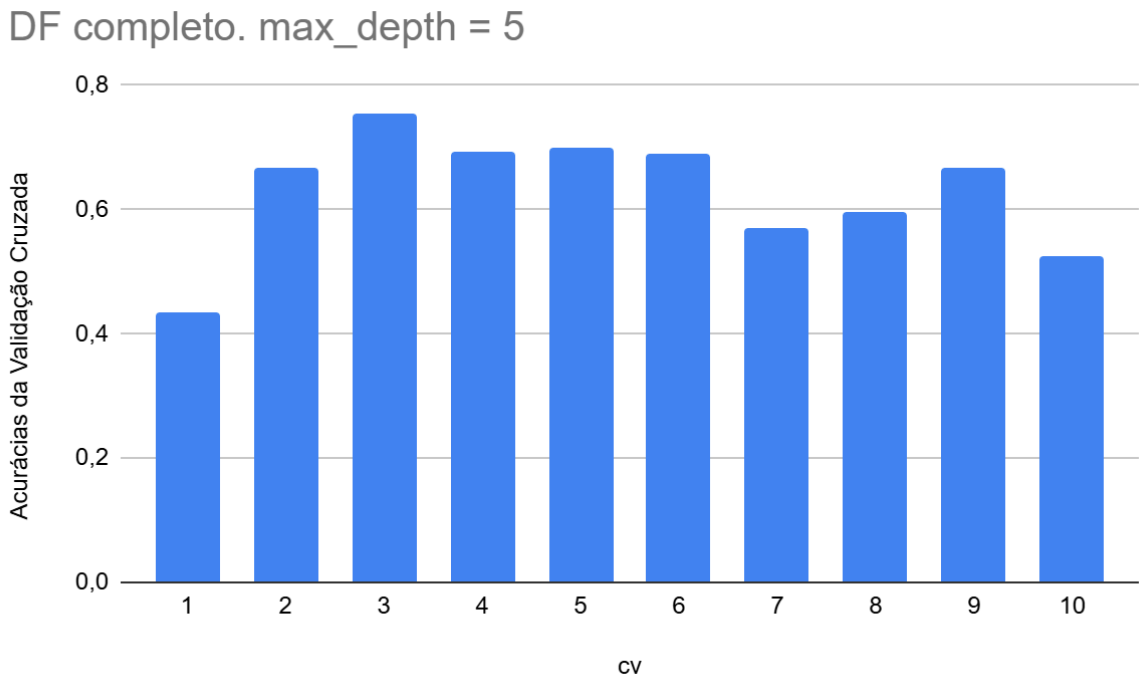
=> PRECISÃO MÁXIMA DE APENAS 77%. O modelo apresenta alguns falsos positivos;
=> RECALL COM VALORES MÉDIOS. O modelo tem dificuldade em classificar adequadamente, deixando de classificar alguns indivíduos que encontrou;
=> O ideal é que o F1-Score seja próximo de 1, o que significaria que o modelo acerta muito e apresenta poucos erros . O MODELO É RAZOÁVEL NA CLASSIFICAÇÃO DOS INDIVÍDUOS DA CLASSE JÚNIOR E SÊNIOR. APRESENTA PERFORMANCE MÉDIA NA CLASSIFICAÇÃO DOS INDIVÍDUOS DA CLASSE PLENO..

- Matriz de confusão



=> MUITAS FALHAS IDENTIFICADAS FORA DA DIAGONAL PRINCIPAL. Principalmente relativamente à classificação como Pleno.

- Validação Cruzada considerando 10 “folds”

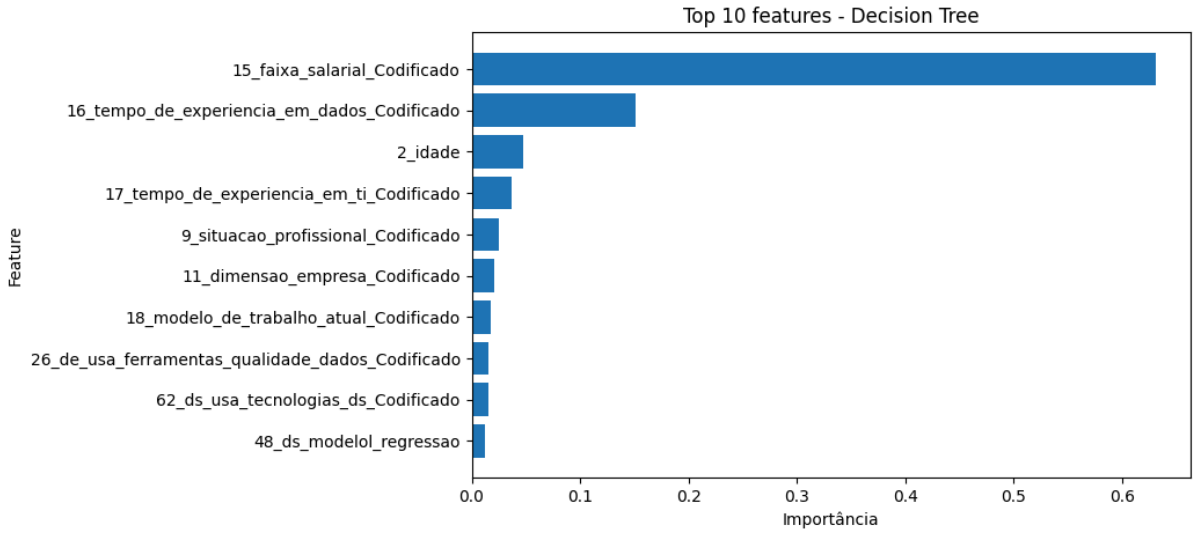


Acurácia Mínima CV: 0.4350
Acurácia Máxima CV: 0.7553
Acurácia Média CV: 0.6300
Desvio Padrão CV: 0.0921
=> GRANDE VARIABILIDADE, o que significa que o modelo apresenta muita incerteza.

Considerando max_depth = 5 e na árvore de decisão do critério 2 (dataframe criado usando dados de pessoas do gênero feminino)

- importância das features:

	feature	importance
12	15_faixa_salarial_Codificado	0.631215
13	16_tempo_de_experiencia_em_dados_Codificado	0.150912
0	2_idade	0.047845
14	17_tempo_de_experiencia_em_ti_Codificado	0.037225
7	9_situacao_profissional_Codificado	0.024559
9	11_dimensao_empresa_Codificado	0.020500
15	18_modelo_de_trabalho_atual_Codificado	0.018141
23	26_de_usa_ferramentas_qualidade_dados_Codificado	0.015849
59	62_ds_usa_tecnologias_ds_Codificado	0.015413
45	48_ds_modelol_regressao	0.011765



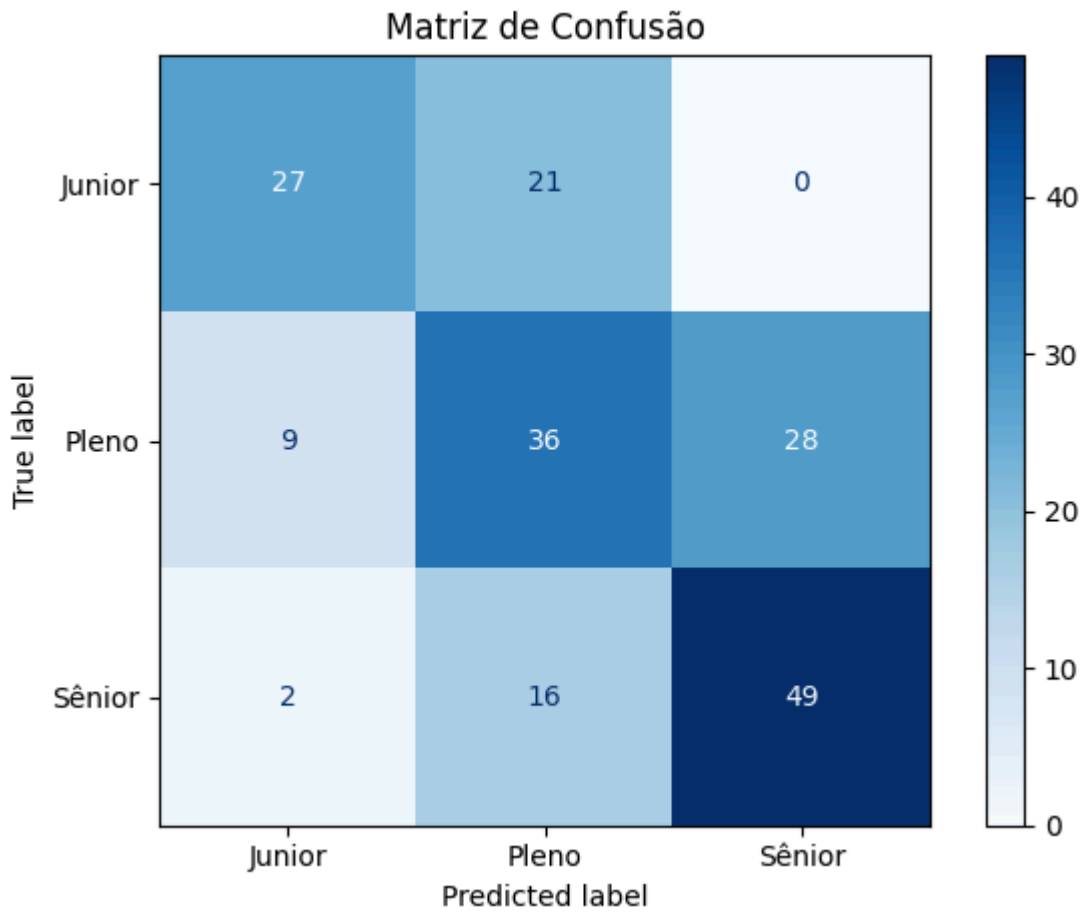
=> Claramente, das 74 features, apenas duas são relevantes.

- Precisão, Recall e F1-score

	precision	recall	f1-score	support
Júnior	0.71	0.56	0.63	48
Pleno	0.49	0.49	0.49	73
Sênior	0.64	0.73	0.68	67
accuracy			0.60	188
macro avg	0.61	0.60	0.60	188
weighted avg	0.60	0.60	0.59	188

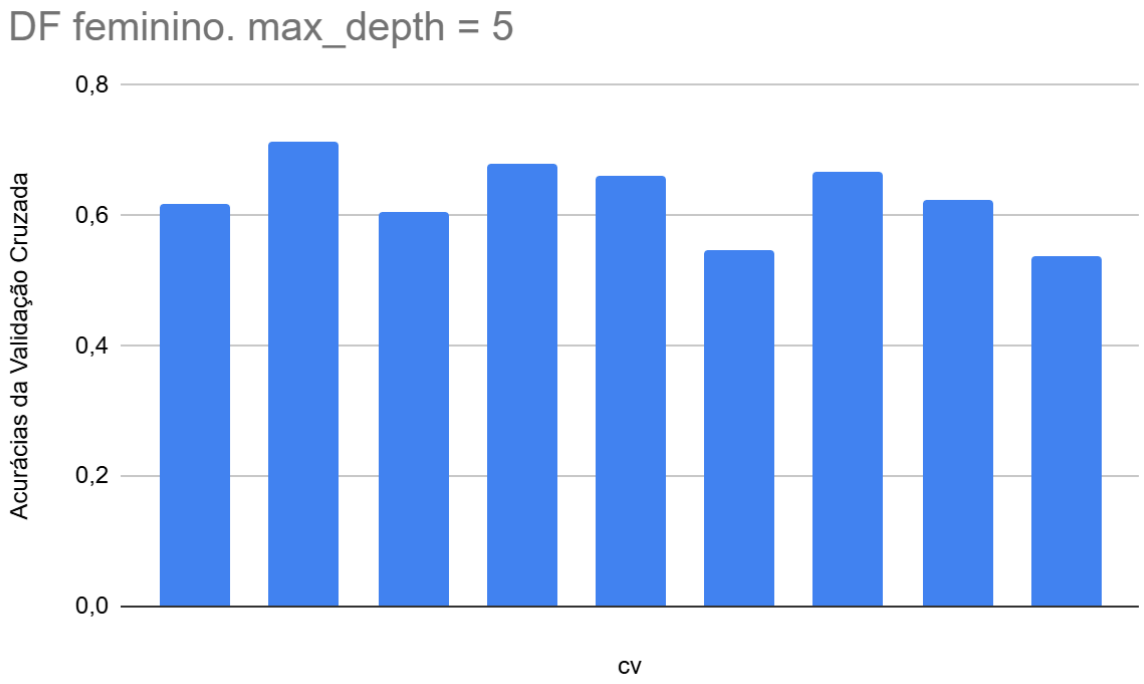
=> PRECISÃO MÁXIMA DE APENAS 71%. O modelo apresenta alguns falsos positivos;
=> RECALL COM VALORES MÉDIOS BAIXOS. O modelo tem dificuldade em classificar adequadamente, deixando de classificar alguns indivíduos que encontrou;
=> O ideal é que o F1-Score seja próximo de 1, o que significaria que o modelo acerta muito e apresenta poucos erros . O MODELO É MEDIANO NA CLASSIFICAÇÃO DOS INDIVÍDUOS DA CLASSE JÚNIOR E SÊNIOR. E AINDA APRESENTA PIOR PERFORMANCE NA CLASSIFICAÇÃO DOS INDIVÍDUOS DA CLASSE PLENO..

- Matriz de confusão



=> MUITAS FALHAS IDENTIFICADAS FORA DA DIAGONAL PRINCIPAL. Principalmente relativamente à classificação como Pleno.

- Validação Cruzada considerando 10 “folds”

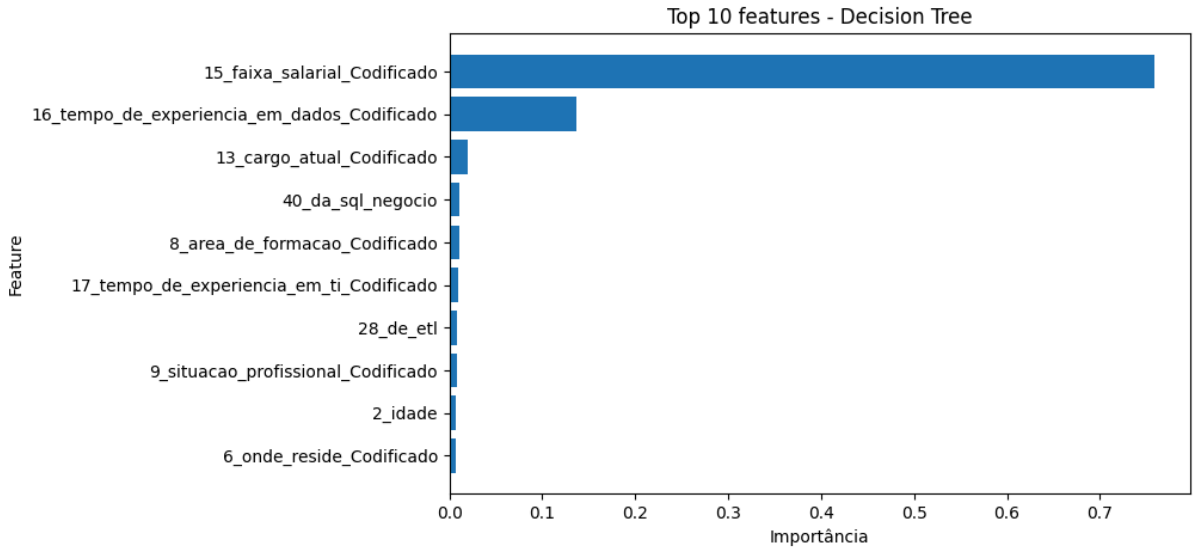


Acurácia Mínima CV: 0.3723
Acurácia Máxima CV: 0.7128
Acurácia Média CV: 0.6025
Desvio Padrão CV: 0.0930
=> GRANDE VARIABILIDADE, o que significa que o modelo apresenta muita incerteza.

Considerando max_depth = 5 e na árvore de decisão do critério 3 (dataframe criado usando dados de pessoas do gênero masculino)

- importância das features:

	feature	importance
12	15_faixa_salarial_Codificado	0.758688
13	16_tempo_de_experiencia_em_dados_Codificado	0.136157
11	13_cargo_atual_Codificado	0.019155
37	40_da_sql_negocio	0.010845
6	8_area_de_formacao_Codificado	0.010254
14	17_tempo_de_experiencia_em_ti_Codificado	0.009378
25	28_de_etl	0.008425
7	9_situacao_profissional_Codificado	0.008425
0	2_idade	0.006786
4	6_onde_reside_Codificado	0.006466



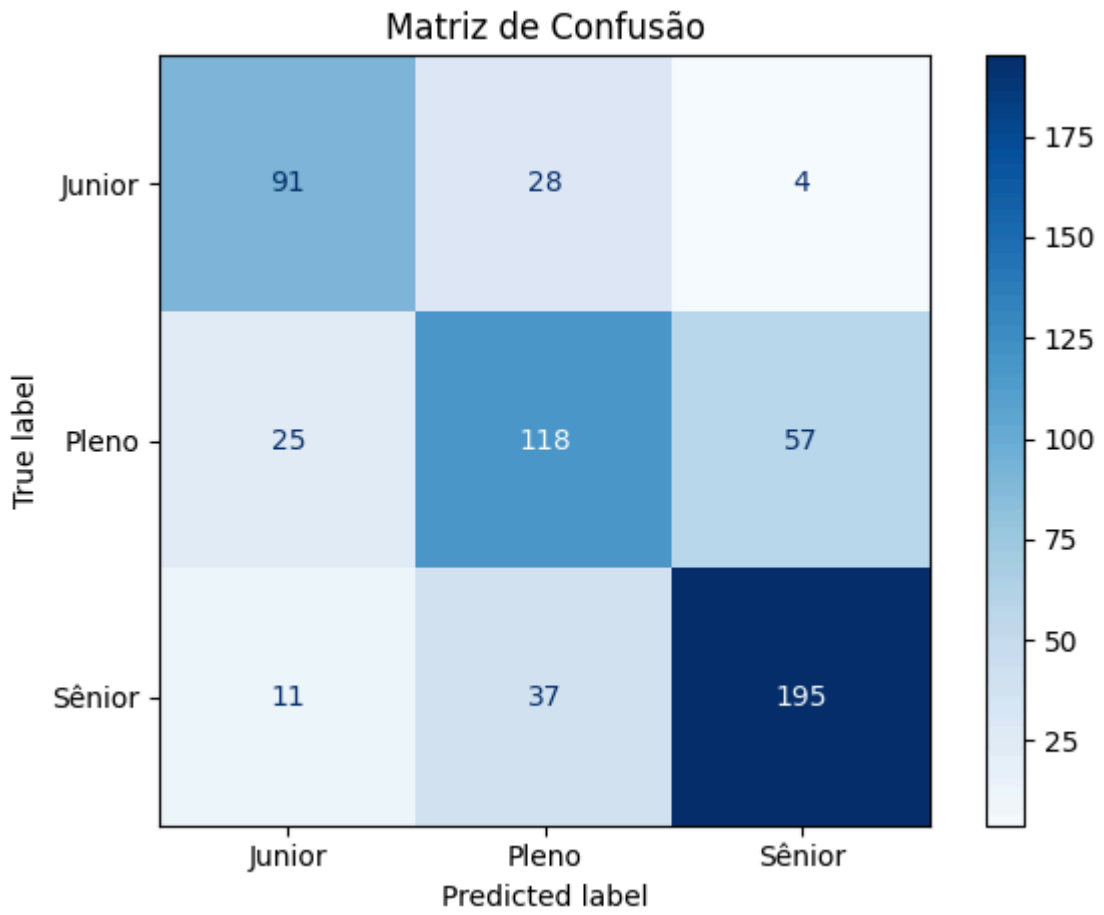
=> Claramente, das 74 features, apenas duas são relevantes.

- Precisão, Recall e F1-score

...	precision	recall	f1-score	support
Júnior	0.72	0.74	0.73	123
Pleno	0.64	0.59	0.62	200
Sênior	0.76	0.80	0.78	243
accuracy			0.71	566
macro avg	0.71	0.71	0.71	566
weighted avg	0.71	0.71	0.71	566

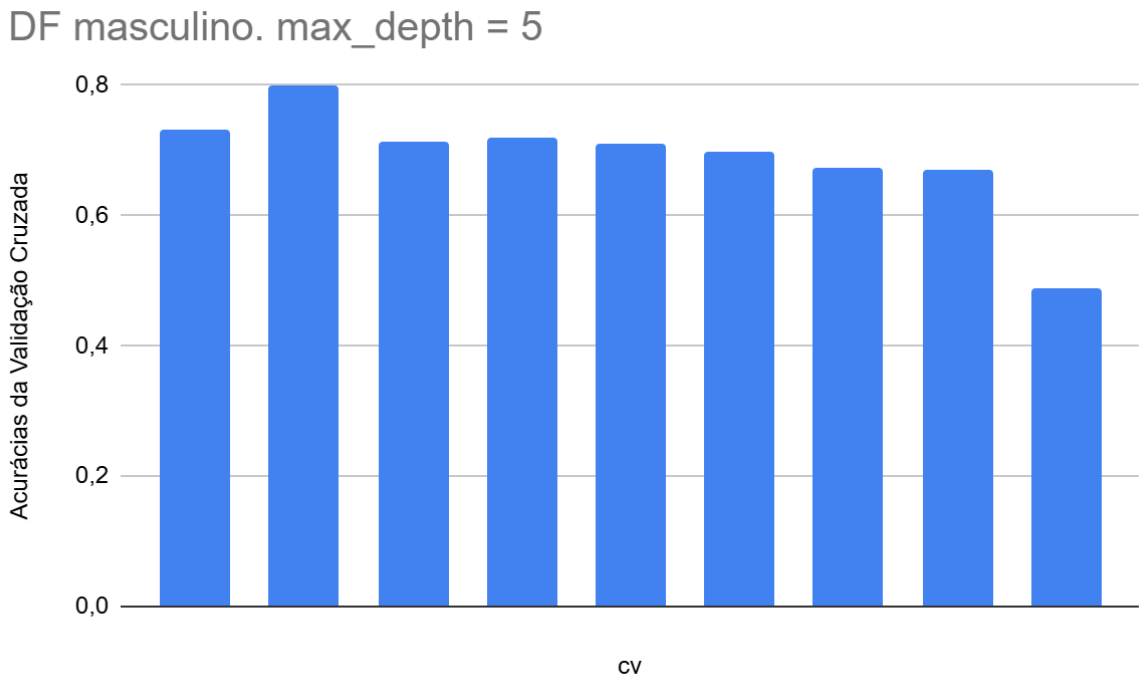
=> PRECISÃO MÁXIMA DE APENAS 76%. O modelo apresenta alguns falsos positivos;
=> RECALL COM VALORES MÉDIOS ALTOS. O modelo tem dificuldade em classificar adequadamente, deixando de classificar alguns indivíduos que encontrou, principalmente da classe PLENO;
=> O ideal é que o F1-Score seja próximo de 1, o que significa que o modelo acerta muito e apresenta poucos erros . O MODELO É MEDIANO NA CLASSIFICAÇÃO DOS INDIVÍDUOS DA CLASSE PLENO. Um pouco melhor na classificação dos indivíduos das outras classes.

- Matriz de confusão



=> MUITAS FALHAS IDENTIFICADAS FORA DA DIAGONAL PRINCIPAL. Principalmente relativamente à classificação como Pleno.

- Validação Cruzada considerando 10 “folds”



Acurácia Mínima CV: 0.4134
Acurácia Máxima CV: 0.7986
Acurácia Média CV: 0.6620
Desvio Padrão CV: 0.1118
=> GRANDE VARIABILIDADE, o que significa que o modelo apresenta muita incerteza.

5.2. O MODELO ATENDE O OBJETIVO DO PROJETO?

O modelo não atende o objetivo do projeto. O modelo não permite uma classificação em cada uma das 3 classes de senioridade segura, independente do modelo que se esteja analisando. Todavia, o modelo do critério 1 (onde se considera todo o dataframe) apresenta uma performance melhor.

É importante ressaltar que, nos três modelos construídos, as variáveis que apresentam maior importância, ou que explicam melhor os modelos, não têm qualquer relação com as variáveis demográficas coletadas. Este fato pode ser um indício de que o gênero não tem qualquer relevância na definição do nível de senioridade dos indivíduos que desenvolvem as suas atividades na área de dados.

Na seção de conclusões e trabalhos futuros discute-se um algumas ações que podem ser tomadas para buscar melhorar a performance dos modelos e responder à pergunta inicial da pesquisa: 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?

6. Implementação e Entregáveis

A implementação foi realizada usando o Google Colab, e o código está disponível aqui. Todavia, um .py e .ipynb estarão disponíveis no repositório público do GitHub, criado com a finalidade de disponibilizar todo o trabalho desenvolvido no âmbito deste projeto, que inclui:

- Repositório GitHub;
- README.md com descrição do projeto;
- Notebooks do Google Colab;
- Licença de uso e autoria;
- Relatório Técnico (este documento);
- Apresentação Final - slides;
- Apresentação Final - video.

O link para o repositório GitHub é:
Fatores preditores do nível de senioridade na área de dados - Projeto Aplicado PosCDIA (https://github.com/ErikaPequeno/Projeto_Final_CDIA)

7. Conclusões e Trabalhos Futuros

Com este projeto aplicado buscou-se identificar os principais fatores preditores do nível de senioridade (Júnior, Pleno, Sênior) de profissionais da área de dados, usando por base os dados coletados no Âmbito da pesquisa State of Data Brazil 2024-2025.

Foram construídos 3 modelos de classificação usando a técnica de Machine Learning Decision Trees (em português, Árvores de Decisão). A diferença entre os três modelos estava relacionada com o conjunto de dados considerado: um primeiro conjunto de dados considerava todo o dataframe criado a partir dos dados da pesquisa; o segundo conjunto de dados considerava apenas as respostas do indivíduos que responderam pertencer ao gênero feminino; o último conjunto de dados considerava apenas os dados dos respondentes que informaram pertencer ao gênero masculino.

De forma geral, os modelos construídos são bons para classificar Seniores. A maior dificuldade está na classe "Pleno". As características (features) usadas não são boas para separar os indivíduos Pleno dos outros.

Relativamente à acurácia de 73% do modelo do critério 1, pode-se afirmar que é um bom ponto de partida, mas indica que há espaço para melhoria, especialmente na distinção do nível de senioridade Pleno. As melhorias nos modelos de classificação podem considerar ações como: usar dados de anos anteriores; Seleção de variáveis por outros métodos que não a intuição; Avaliar a correlação entre o *nível de senioridade* e a *faixa salarial* e tentar identificar qual a contribuição efetiva no modelo da variável faixa salarial na identificação do nível de senioridade. O mesmo seria interessante para variáveis como *modelo de trabalho atual*, por exemplo.

Como última nota ressalta-se que o sucesso dos modelos construídos têm uma forte relação com a sinceridade das respostas dadas aos questionários aplicados.

Referências

“Kaggle: Your Machine Learning and Data Science Community.” n.d. Kaggle: Your Machine Learning and Data Science Community. Accessed December 13, 2025. https://www.kaggle.com/.

Lages, Gabriel, Felipe Fiamozzini, Monique Femme, and Paulo Vasconcellos. 2025. “State of Data Brazil 2024-2025.” Kaggle. https://www.kaggle.com/datasets/datahackers/state-of-data-brazil-20242025/data.

“State of Data 2024-2025.” n.d. Bain & Company. Accessed December 13, 2025. https://www.bain.com/pt-br/insights/state-of-data-2024/.

Anexo 1

Pesquisa Google

Anexo 2

Estrutura_dos_dados_State_of_Data_Brazil

Anexo 3

Dicionário de dados do projeto

variável	significado	tipo de variável	valores	coluna do dataset original
token	identificador da resposta	alfanumérico	strings	o.a
idade	idade do respondente	variável numérica inteira	1,2,3,...	1.a
genero	gênero do respondente	binária	Masculino Feminino	1.b
etnia	etnia do respondente	variável categórica nominal	Branca Parda Preta Amarela Prefiro não informar Indígena Outra	1.c
pcd	pessoa com deficiência	binária	Sim Não	1.d

onde_reside	unidade federal onde reside, se reside no Brasil. Senão, assume o valor estrangeiro	categórica nominal	AC AL AM AP BA CE DF ES GO MA MG MS MT PA PB PE PI PR RJ RN RO RR RS SC SE SP TO Estrangeiro	transformação usando 1.i.1 e 1.g
nivel_ensino	maior nível de escolaridade do respondente	categórica ordinal	Não tenho graduação formal Estudante de Graduação Graduação/Bacharelado Pós-graduação Mestrado Doutorado ou Phd	1.l
area_de_formacao	àrea de formação do respondente	categórica nominal	Ciências Biológicas/ Farmácia/ Medicina/ Área da Saúde Computação / Engenharia de Software / Sistemas de Informação/ TI Economia/ Administração / Contabilidade / Finanças/ Negócios Estatística/ Matemática / Matemática Computacional/ Ciências Atuariais Marketing / Publicidade / Comunicação / Jornalismo / Ciências Sociais	1.m

			Química / Física Outras Engenharias (não incluir engenharia de software ou TI) Outra opção	
situacao_profissional	situação profissional do respondente	categórica nominal	Desempregado, buscando recolocação Empreendedor ou Empregado (CNPJ) Empregado (CLT) Estagiário Freelancer Servidor Público Somente Estudante (graduação) Somente Estudante (pós-graduação) Trabalho na área Acadêmica/Pesquisador Vivo fora do Brasil e trabalho para empresa de fora do Brasil Vivo no Brasil e trabalho remoto para empresa de fora do Brasil	2.a
setor_empresa	setor a que pertence a empresa do respondente	categórica nominal	Agronegócios Educação Entretenimento ou Esportes Filantropia/ONG's Finanças ou Bancos Indústria Internet/Ecommerce Marketing Outra Opção Seguros ou Previdência Setor Alimentício Setor Automotivo Setor Farmaceutico Setor Imobiliário/ Construção Civil Setor Público Setor de Energia Tecnologia/Fábrica de Software Telecomunicação Varejo Área da Saúde Área de Consultoria	2.b

dimensao_empresa	dimensão da empresa onde trabalha o respondente	categórica ordinal	de 1 a 5 de 6 a 10 de 11 a 50 de 51 a 100 de 101 a 500 de 501 a 1.000 de 1.001 a 3.000 Acima de 3.000	2.c
gestor	o respondente atua como gestor	binária	Sim Não	2.d
cargo_atual	cargo atual do respondente	categórica nominal	Analista de BI/BI Analyst Analista de Dados/Data Analyst Analista de Negócios/Business Analyst Analista de Suporte/Analista Técnico Analytics Engineer Arquiteto de Dados/Data Architect Cientista de Dados/Data Scientist Data Product Manager/ Product Manager (PM/APM/DPM/GPM/PO) Desenvolvedor/ Engenheiro de Software/ Analista de Sistemas Engenheiro de Dados/Data Engineer/Data Architect Engenheiro de Machine Learning/ML Engineer/AI Engineer Estatístico Outra Opção Outras Engenharias (não inclui dev) Professor/Pesquisador	2.f
nivel_senioridade	nível de senioridade do respondente	categórica ordinal	Júnior Pleno Sênior	2.g

faixa_salarial	faixa salarial do respondente	categórica ordinal	Menos de R$ 1.000/mês de R$ 1.001/mês a R$ 2.000/mês de R$ 2.001/mês a R$ 3.000/mês de R$ 3.001/mês a R$ 4.000/mês de R$ 4.001/mês a R$ 6.000/mês de R$ 6.001/mês a R$ 8.000/mês de R$ 8.001/mês a R$ 12.000/mês de R$ 12.001/mês a R$ 16.000/mês de R$ 16.001/mês a R$ 20.000/mês de R$ 20.001/mês a R$ 25.000/mês de R$ 25.001/mês a R$ 30.000/mês de R$ 30.001/mês a R$ 40.000/mês Acima de R$ 40.001/mês	2.h
tempo_de_experiencia_em_dados	tempo de experiência em dados do respondente	categórica ordinal	Não tenho experiência na área de dados Menos de 1 ano de 1 a 2 anos de 3 a 4 anos de 5 a 6 anos de 7 a 10 anos Mais de 10 anos	2.i
tempo_de_experiencia_em_ti	tempo de experiência em ti do respondente	categórica ordinal	Não tive experiência na área de TI/Engenharia de Software antes de começar a trabalhar na área de dados Menos de 1 ano de 1 a 2 anos de 3 a 4 anos de 5 a 6 anos de 7 a 10 anos Mais de 10 anos	2.j
modelo_de_trabalho_atual	modelo de trabalho atual do respondente	categórica nominal	Modelo 100% remoto Modelo híbrido flexível Modelo híbrido com dias fixos de trabalho presencial Modelo 100% presencial	2.r

funcao_atuacao	função de atuação do respondente	categórica nominal	Análise de Dados Buscando oportunidade na área de dados Ciência de Dados Engenharia de Dados Gestor Outra atuação	4.a.1
num_fontes_dados	número de fontes de dados que o respondente pode usar no dia a dia	numérica inteira	número entre 0 e 8	transformação das variáveis 4.b, 4.b.1 a 4.b.8
num_linguagens_prog	número de linguagens de programação que o respondente pode usar no dia-a-dia	numérica inteira	número entre 0 e 15	transformação das variáveis 4.d, 4.d.1 a 4.d.15
linguagem_mais_usada	linguagem mais usada pelo repondente	categórica nominal	.NET C/C++/C# Java JavaScript Julia Matlab Não utilizo nenhuma das linguagens listadas no trabalho PHP Python R SAS/Stata SQL Scala Visual Basic/VBA	4.e
usa_chatgpt_ou_copilot	o respondente usa chatgpt ou copilot no trabalho para melhorar a	binária	Sim Não	transformação das variáveis 4.m, 4.m.1 a 4.m.5

	produtividade			
num_rotinas_de	Quantas rotinas como DE o respondente realiza	numérica inteira	número entre 0 e 9	transformação das variáveis 6.a, 6.a.1 a 6.a.9
de_usa_ferramentas_etl	O respondente usa ferramentas ETL como DE	binária	Sim Não	transformação das variáveis 6.b, 6.b.1 a 6.b.21
de_usa_ferramentas_qualidade_dados	O respondente usa ferramentas de qualidade de dados como DE	binária	Sim Não	6.g
principal_rotina_como_de	rotina com maior tempo gasto como DE	categórica nominal	Desenvolvendo pipelines de dados utilizando linguagens de programação como Python, Scala, Java etc. Realizando construções de ETL\\s em ferramentas como Pentaho, Talend, Dataflow etc. Criando consultas através da linguagem SQL para exportar informações e compartilhar com as áreas de negócio. Atuando na integração de diferentes fontes de dados através de plataformas proprietárias como Stitch Data, Fivetran etc. Modelando soluções de arquitetura de dados, criando componentes de ingestão de dados, transformação e recuperação da informação. Desenvolvendo/cuidando da manutenção de repositórios de dados baseados em streaming de eventos como Data Lakes e Data Lakehouses.	transformação das variáveis 6.h, 6.h.1 a 6.h.9

			Atuando na modelagem dos dados, com o objetivo de criar conjuntos de dados como Data Warehouses, Data Marts, Datasets etc. Cuidando da qualidade dos dados, metadados e dicionário de dados. Nenhuma das opções listadas refletem meu dia a dia.	
num_rotinas_da	Quantas rotinas como DA o responden te realiza	numérica inteira	número entre 0 e 10	transforma ção das variáveis 6.a, 6.a.1 a 6.a.10
da_usa_ferramentas _etl	O responden te usa ferrament as ETL como DA	binária	Sim Não	transforma ção das variáveis 7.b, 7.b.1 a 7.b.21
da_usa_ferramentas _autonomia_area_d e_negocios	O responden te usa ferrament as de de autonomia da área de negócios como DA	binária	Sim Não	transforma ção das variáveis 7.c, 7.c.1 a 7.c.6
principal_rotina_co mo_da	rotina com maior tempo gasto como DA	categórica nominal	Processando e analisando dados utilizando linguagens de programação como Python, R etc. Realizando construções de dashboards em ferramentas de BI como PowerBI, Tableau, Looker, Qlik etc. Criando consultas através da linguagem SQL para exportar informações e compartilhar com as áreas de negócio. Utilizando APIs para extrair dados e complementar minhas análises., Realizando experimentos e	transforma ção das variáveis 7.d, 7.d.1 a 7.d.10

			estudos utilizando metodologias estatísticas como teste de hipótese, modelos de regressão etc. Desenvolvendo/cuidando da manutenção de ETLs utilizando tecnologias como Talend, Pentaho, Airflow, Dataflow etc., Atuando na modelagem dos dados, com o objetivo de criar conjuntos de dados como Data Warehouses, Data Marts, Datasets etc. Desenvolvendo/cuidando da manutenção de planilhas para atender as áreas de negócio. Utilizando ferramentas avançadas de estatística como SAS, SPSS, Stata etc, para realizar análises de dados. Nenhuma das opções listadas refletem meu dia a dia.	
num_rotinas_ds	Quantas rotinas como DS o respondente realiza	numérica inteira	número entre 0 e 12	transformação das variáveis 8.a, 8.a.1 a 8.a.12
ds_usa_tecnicas_e_metodos_ds	O respondente usa técnicas e métodos de DS como DS	categórica nominal	Utilizo modelos de regressão (linear, logística, GLM). Utilizo redes neurais ou modelos baseados em árvore para criar modelos de classificação. Desenvolvo sistemas de recomendação (RecSys). Utilizo métodos estatísticos Bayesianos para analisar dados. Utilizo técnicas de NLP (Natural Language Processing) para análisar dados não-estruturados. Utilizo métodos estatísticos clássicos (Testes de hipótese, análise multivariada, sobrevivência, dados longitudinais, inferência estatistica) para analisar	8.b, 8.b.1 a 8.b.14

			dados. Utilizo cadeias de Markov ou HMM\\s para realizar análises de dados. Desenvolvo técnicas de Clusterização (K-means, Spectral, DBScan etc). Realizo previsões através de modelos de Séries Temporais (Time Series). Utilizo modelos de Reinforcement Learning (aprendizado por reforço). Utilizo modelos de Machine Learning para detecção de fraude. Utilizo métodos de Visão Computacional. Utilizo modelos de Detecção de Churn. Utilizo LLMs para solucionar problemas de negócio.,	
ds_usa_tecnologias_ds	O respondente usa tecnologias como DS	binária	Sim Não	transformação das variáveis 8.c, 8.c.1 a 8.c.11
principal_rotina_como_ds	rotina com maior tempo gasto como DA	categórica nominal	Estudos Ad-hoc com o objetivo de confirmar hipóteses, realizar modelos preditivos, forecasts, análise de cluster para resolver problemas pontuais e responder perguntas das áreas de negócio. Coletando e limpando dos dados que uso para análise e modelagem. Entrando em contato com os times de negócio para definição do problema, identificar a solução e apresentação de resultados. Desenvolvendo modelos de Machine Learning com o objetivo de colocar em produção em sistemas (produtos de dados). Colocando modelos em	transformação das variáveis 8.d, 8.d.1 a 8.d.12

			produção, criando os pipelines de dados, APIs de consumo e monitoramento. Cuidando da manutenção de modelos de Machine Learning já em produção, atuando no monitoramento, ajustes e refatoração quando necessário. Realizando construções de dashboards em ferramentas de BI como PowerBI, Tableau, Looker, Qlik, etc. Utilizando ferramentas avançadas de estatística como SAS, SPSS, Stata etc, para realizar análises. Criando e dando manutenção em ETLs, DAGs e automações de pipelines de dados. Criando e gerenciando soluções de Feature Store e cultura de MLOps. Criando e mantendo a infra que meus modelos e soluções rodam (clusters, servidores, API, containers, etc.) Treinando e aplicando LLMs para solucionar problemas de negócio.	

Anexo 4

Relação de-para entre as colunas originais da pesquisa e o dataframe do projeto.

Coluna dataframe original	pergunta original	nome variável no projeto	nome da coluna no dataframe
0.a	token	token	token
1.a	idade	idade	idade
1.b	genero	genero	genero
1.c	cor/raca/etnia	etnia	etnia
1.d	pcd	pcd	pcd
1.l	nivel_de_ensino	nivel_ensino	nivel_ensino
1.m	área_de_formação	area_de_formacao	area_de_formacao

2.a	situação_de_trabalho	situacao_profissional	situacao_profissional
2.b	setor	setor_empresa	setor_empresa
2.c	numero_de_funcionarios	dimensao_empresa	dimensao_empresa
2.d	atua_como_gestor	gestor	gestor
2.f	cargo_atual	cargo_atual	cargo_atual
2.g	nivel	nivel_senioridade	nivel_senioridade
2.h	faixa_salarial	faixa_salarial	faixa_salarial
2.i	tempo_de_experiencia_em_dados	tempo_de_experiencia_em_dados	tempo_de_experiencia_em_dados
2.j	tempo_de_experiencia_em_ti	tempo_de_experiencia_em_ti	tempo_de_experiencia_em_ti
2.r	modelo_de_trabalho_atual	modelo_de_trabalho_atual	modelo_de_trabalho_atual
4.a.1	atuacao_em_dados	funcao_atuacao	funcao_atuacao
4.e	linguagem_mais_usada	linguagem_mais_usada	linguagem_mais_usada
6.h.1	Desenvolvendo pipelines de dados utilizando linguagens de programação como Python, Scala, Java etc.	principal_rotina_como_de	de_pipeline_python
6.h.2	Realizando construções de ETL\\s em ferramentas como Pentaho, Talend, Dataflow etc.	principal_rotina_como_de	de_etl
6.h.3	Criando consultas através da linguagem SQL para exportar informações e compartilhar com as áreas de negócio.	principal_rotina_como_de	de_sql_negocio
6.h.4	Atuando na integração de diferentes fontes de dados através de plataformas proprietárias como Stitch Data, Fivetran etc.	principal_rotina_como_de	de_integracao_fontes
6.h.5	Modelando soluções de arquitetura de dados, criando componentes de ingestão de dados, transformação e recuperação da informação.	principal_rotina_como_de	de_solucoes_arquitetura_dados
6.h.6	Desenvolvendo/cuidando da manutenção de repositórios de dados baseados em streaming de eventos como Data Lakes e Data Lakehouses.	principal_rotina_como_de	de_manutencao_repositorios
6.h.7	Atuando na modelagem dos dados, com o objetivo de criar conjuntos de dados como Data Warehouses, Data Marts, Datasets etc.	principal_rotina_como_de	de_modelagem_dados

6.h.8	Cuidando da qualidade dos dados, metadados e dicionário de dados.	principal_rotina_como_de	de_metadados
7.d.1	Processando e analisando dados utilizando linguagens de programação como Python, R etc.	principal_rotina_como_da	da_processametno_análise_python
7.d.2	Realizando construções de dashboards em ferramentas de BI como PowerBI, Tableau, Looker, Qlik etc.	principal_rotina_como_da	da_dashboards
7.d.3	Criando consultas através da linguagem SQL para exportar informações e compartilhar com as áreas de negócio.	principal_rotina_como_da	da_sql_negocio
7.d.4	Utilizando APIs para extrair dados e complementar minhas análises.,	principal_rotina_como_da	da_extracao_api
7.d.5	Realizando experimentos e estudos utilizando metodologias estatísticas como teste de hipótese, modelos de regressão etc.	principal_rotina_como_da	da_modelos_estatisitico
7.d.6	Desenvolvendo/cuidando da manutenção de ETLs utilizando tecnologias como Talend, Pentaho, Airflow, Dataflow etc.,	principal_rotina_como_da	da_etl
7.d.7	Atuando na modelagem dos dados, com o objetivo de criar conjuntos de dados como Data Warehouses, Data Marts, Datasets etc.	principal_rotina_como_da	da_modelagem_dados
7.d.8	Desenvolvendo/cuidando da manutenção de planilhas para atender as áreas de negócio.	principal_rotina_como_da	da_planilhas
7.d.9	Utilizando ferramentas avançadas de estatística como SAS, SPSS, Stata etc, para realizar análises de dados.	principal_rotina_como_da	da_analises_estatisticas
8.b.1	Utilizo modelos de regressão (linear, logística, GLM).	ds_usa_tecnicas_e_metodos_ds	ds_modelol_regressao
8.b.2	Utilizo redes neurais ou modelos baseados em árvore para criar modelos de classificação.	ds_usa_tecnicas_e_metodos_ds	ds_modelo_classificacao
8.b.3	Desenvolvo sistemas de recomendação (RecSys).	ds_usa_tecnicas_e_metodos_ds	ds_sisatema_recomendacao
8.b.4	Utilizo métodos estatísticos Bayesianos para analisar dados.	ds_usa_tecnicas_e_metodos_ds	ds_metodos_bayesianos
8.b.5	Utilizo técnicas de NLP (Natural Language Processing) para análisar dados não-estruturados.	ds_usa_tecnicas_e_metodos_ds	ds_tecnicas_nlp

8.b.6	Utilizo métodos estatísticos clássicos (Testes de hipótese, análise multivariada, sobrevivência, dados longitudinais, inferência estatistica) para analisar dados.	ds_usa_tecnicas_e_metodos_ds	ds_metodos_estatisticos
8.b.7	Utilizo cadeias de Markov ou HMM\\s para realizar análises de dados.	ds_usa_tecnicas_e_metodos_ds	ds_markov_hmm
8.b.8	Desenvolvo técnicas de Clusterização (K-means, Spectral, DBScan etc).	ds_usa_tecnicas_e_metodos_ds	ds_clusterizacao
8.b.9	Realizo previsões através de modelos de Séries Temporais (Time Series).	ds_usa_tecnicas_e_metodos_ds	ds_series_temporais
8.b.10	Utilizo modelos de Reinforcement Learning (aprendizado por reforço).	ds_usa_tecnicas_e_metodos_ds	ds_modelos_RL
8.b.11	Utilizo modelos de Machine Learning para detecção de fraude.	ds_usa_tecnicas_e_metodos_ds	ds_modelos_ML_fraude
8.b.12	Utilizo métodos de Visão Computacional.	ds_usa_tecnicas_e_metodos_ds	ds_visao_computacional
8.b.13	Utilizo modelos de Detecção de Churn.	ds_usa_tecnicas_e_metodos_ds	ds_deteccao_de_churn
8.b.14	Utilizo LLMs para solucionar problemas de negócio.,	ds_usa_tecnicas_e_metodos_ds	ds_llms
8.d.1	Estudos Ad-hoc com o objetivo de confirmar hipóteses, realizar modelos preditivos, forecasts, análise de cluster para resolver problemas pontuais e responder perguntas das áreas de negócio.	principal_rotina_como_ds	ds_estudos_Ad-hoc
8.d.2	Coletando e limpando dos dados que uso para análise e modelagem.	principal_rotina_como_ds	ds_tratamento_de_dados
8.d.3	Entrando em contato com os times de negócio para definição do problema, identificar a solução e apresentação de resultados.	principal_rotina_como_ds	ds_reuniões_entregas
8.d.4	Desenvolvendo modelos de Machine Learning com o objetivo de colocar em produção em sistemas (produtos de dados).	principal_rotina_como_ds	ds_modelos_ML_producao
8.d.5	Colocando modelos em produção, criando os pipelines de dados, APIs de consumo e monitoramento.	principal_rotina_como_ds	ds_pipelines
8.d.6	Cuidando da manutenção de modelos de Machine Learning já em produção, atuando no monitoramento, ajustes e refatoração quando necessário.	principal_rotina_como_ds	ds_manutencao_ML

8.d.7	Realizando construções de dashboards em ferramentas de BI como PowerBI, Tableau, Looker, Qlik, etc.	principal_rotina_como_ds	ds_dashboards
8.d.8	Utilizando ferramentas avançadas de estatística como SAS, SPSS, Stata etc, para realizar análises.	principal_rotina_como_ds	ds_analises_estatisticas
8.d.9	Criando e dando manutenção em ETLs, DAGs e automações de pipelines de dados.	principal_rotina_como_ds	ds_etl
8.d.10	Criando e gerenciando soluções de Feature Store e cultura de MLOps.	principal_rotina_como_ds	ds_mlops
8.d.11	Criando e mantendo a infra que meus modelos e soluções rodam (clusters, servidores, API, containers, etc.)	principal_rotina_como_ds	ds_infra
8.d.12	Treinando e aplicando LLMs para solucionar problemas de negócio.	principal_rotina_como_ds	ds_llms_negocio
	onde_reside	onde_reside	onde_reside
	num_fontes_dados	num_fontes_dados	num_fontes_dados
	num_linguagens_prog	num_linguagens_prog	num_linguagens_prog
	usa_chatgpt_ou_copilot	usa_chatgpt_ou_copilot	usa_chatgpt_ou_copilot
	num_rotinas_de	num_rotinas_de	num_rotinas_de
	num_rotinas_da	num_rotinas_da	num_rotinas_da
	num_rotinas_ds	num_rotinas_ds	num_rotinas_ds
	de_usa_ferramentas_etl	de_usa_ferramentas_etl	de_usa_ferramentas_etl
	de_usa_ferramentas_qualidade_dados	de_usa_ferramentas_qualidade_dados	de_usa_ferramentas_qualidade_dados
	da_usa_ferramentas_etl	da_usa_ferramentas_etl	da_usa_ferramentas_etl
	da_usa_ferramentas_autonomia_area_de_negocios	da_usa_ferramentas_autonomia_area_de_negocios	da_usa_ferramentas_autonomia_area_de_negocios
	ds_usa_tecnologias_ds	ds_usa_tecnologias_ds	ds_usa_tecnologias_ds