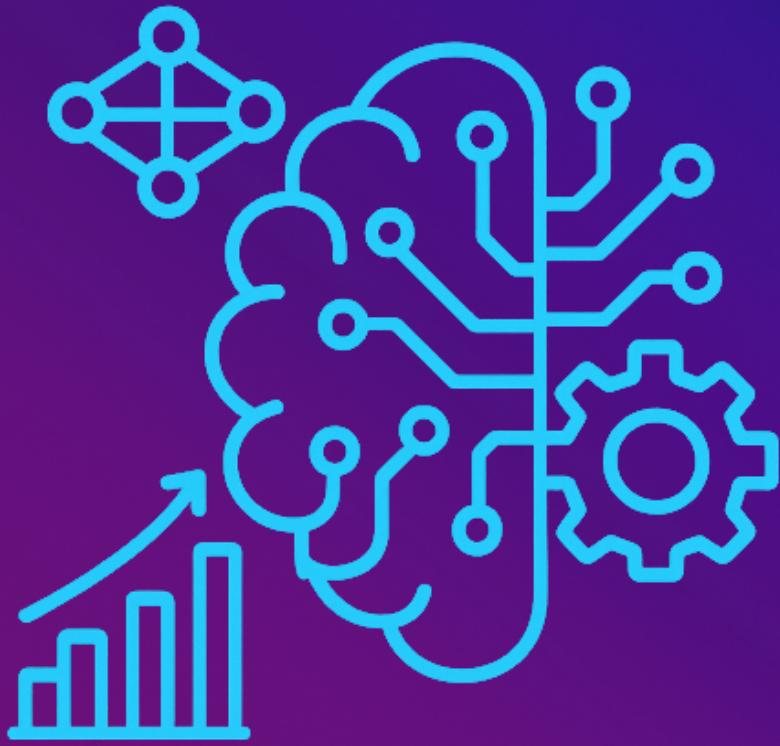


# Projeto em Ciência de Dados I

Grupo 7: nível de qualificação de profissionais



Feito por:

Ian Vinicius Marinho Malta, [ian.malta@sga.pucminas.br](mailto:ian.malta@sga.pucminas.br)  
João Gabriel de Melo Neves, [jgmneves@sga.pucminas.br](mailto:jgmneves@sga.pucminas.br)  
Thiago Couto Ferreira, [thiago.couto@sga.pucminas.br](mailto:thiago.couto@sga.pucminas.br)  
Thales Ribeiro Melo, [thales.melo@sga.pucminas.br](mailto:thales.melo@sga.pucminas.br)

# Introdução + Contextualização

Com a crescente importância da área de dados na tomada de decisões estratégicas, torna-se fundamental que os profissionais que atuam nesse setor sejam altamente qualificados. Nesse cenário, é essencial que os líderes das organizações — como CEOs e diretores — tenham visibilidade sobre a senioridade da equipe de dados, a fim de identificar lacunas e planejar contratações com mais assertividade. Pensando nisso, nossa empresa propõe uma solução que permite identificar, de forma precisa, o nível de experiência dos profissionais (como Júnior, Pleno ou Sênior), oferecendo um suporte confiável para decisões relacionadas à gestão de talentos e estruturação de times.

Nos últimos anos, a Ciência de Dados consolidou-se como uma das áreas mais promissoras dentro do setor de tecnologia, impulsionada pelo aumento exponencial no volume de dados e pela necessidade de transformá-los em insights estratégicos. Empresas de diferentes segmentos passaram a investir fortemente em profissionais de dados — como cientistas, engenheiros e analistas — para aprimorar seus processos, otimizar resultados e sustentar decisões mais embasadas.

Apesar da alta demanda, muitas organizações enfrentam um desafio recorrente: o desequilíbrio na composição de suas equipes, seja pela concentração de profissionais próximos à aposentadoria, sem sucessores preparados, ou pela predominância de perfis júnior, ainda em fase de desenvolvimento técnico.

Diante desse contexto, nossa empresa desenvolveu um sistema inteligente que analisa e apresenta, de forma quantitativa, a distribuição de profissionais por nível de senioridade (Júnior, Pleno e Sênior). O objetivo é oferecer uma ferramenta estratégica para que as empresas possam manter uma equipe equilibrada em termos de experiência, prevenindo riscos futuros, como a escassez de mão de obra qualificada ou a perda de conhecimento institucional.

# Problema

Desenvolver um sistema inteligente capaz de identificar o nível de senioridade predominante entre os profissionais de dados dentro das empresas. O objetivo é fornecer uma ferramenta analítica que permita às organizações avaliar com precisão o grau de capacitação de suas equipes, auxiliando em decisões mais fundamentadas relacionadas à contratação, promoção, definição salarial e planejamento de sucessão.

Por meio da análise de dados como tempo de experiência, cargos anteriores e histórico profissional, o sistema será capaz de demonstrar o nível de preparo dos colaboradores que atuam com dados na empresa. Essa análise não apenas contribui para decisões mais lógicas e assertivas, mas também oferece uma visão estratégica sobre a maturidade e equilíbrio da equipe como um todo — elementos fundamentais para a sustentabilidade da área de dados a longo prazo.

# Objetivo

Desenvolver um sistema capaz de identificar o nível de qualificação de profissionais da área de dados por meio da análise de atributos específicos, utilizando modelos de aprendizado de máquina, com o intuito de auxiliar empresas na composição de equipes mais equilibradas e estrategicamente estruturadas.



# Justificativas

O avanço da tecnologia e da inteligência artificial tem impulsionado a demanda por profissionais qualificados na área de Ciência de Dados. No entanto, a distribuição desses profissionais pelo território brasileiro ainda não é homogênea, dificultando tanto a alocação estratégica de novas empresas quanto a oferta educacional adequada para formar especialistas nessa área. Nesse contexto, torna-se essencial o desenvolvimento de um sistema inteligente que mapeie geograficamente a densidade de Cientistas de Dados por município, permitindo uma análise aprofundada sobre regiões saturadas ou deficitárias desses profissionais.

-Empresas que buscam expandir seus negócios para novas localidades enfrentam dificuldades para encontrar mão de obra qualificada, especialmente em municípios onde a oferta de Cientistas de Dados é escassa. O sistema proposto facilitará esse processo ao fornecer informações detalhadas sobre a concentração desses profissionais e os municípios com maior ou menor disponibilidade de talentos, possibilitando uma tomada de decisão mais assertiva na escolha de onde estabelecer suas operações.

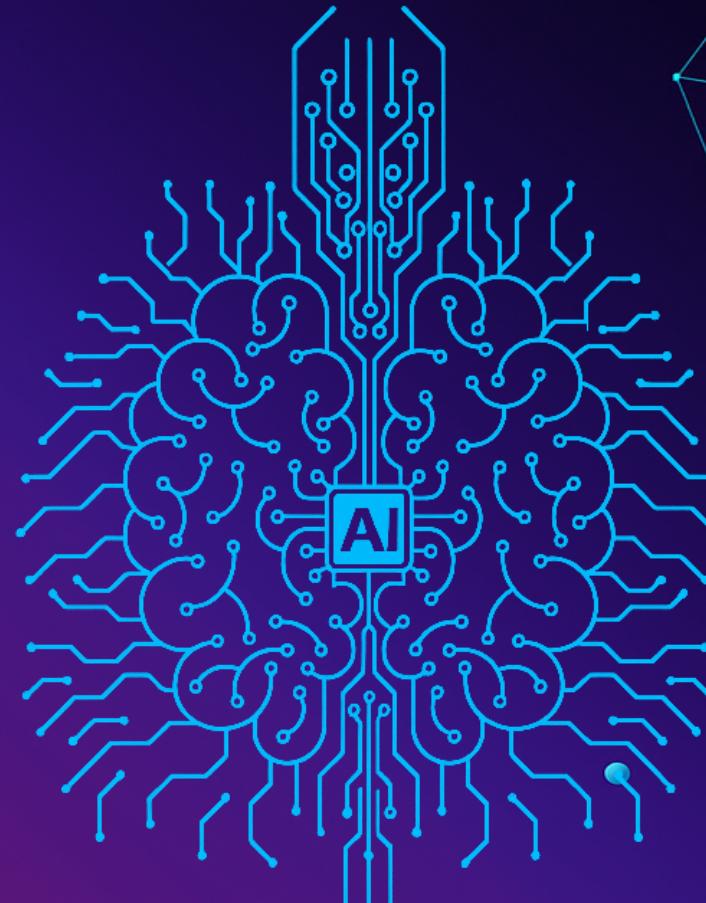
-Além do impacto direto na alocação empresarial, o sistema também contribuirá para o setor educacional. A identificação da demanda por cursos de Ciência de Dados permitirá que universidades e instituições de ensino superior ajustem sua oferta acadêmica, promovendo a expansão do curso para regiões com carência de formação na área. Esse fator é crucial para equilibrar a oferta e a demanda de profissionais no mercado, evitando tanto a saturação quanto a escassez de Cientistas de Dados em determinadas localidades.

-Por fim, ao integrar variáveis que qualificam um Cientista de Dados especializado, o sistema também possibilitará uma análise mais refinada das competências exigidas no mercado, auxiliando empresas na busca por profissionais com habilidades específicas. Dessa forma, o projeto se justifica não apenas pela inovação tecnológica envolvida, mas também pelo seu potencial impacto no desenvolvimento econômico e educacional do Brasil, promovendo um equilíbrio mais eficiente entre formação acadêmica, demanda do mercado e crescimento empresarial.

# Dados utilizados

State\_of\_data\_BR\_2023\_Kaggle - df\_survey\_2023

Auxiliary\_data\_courses

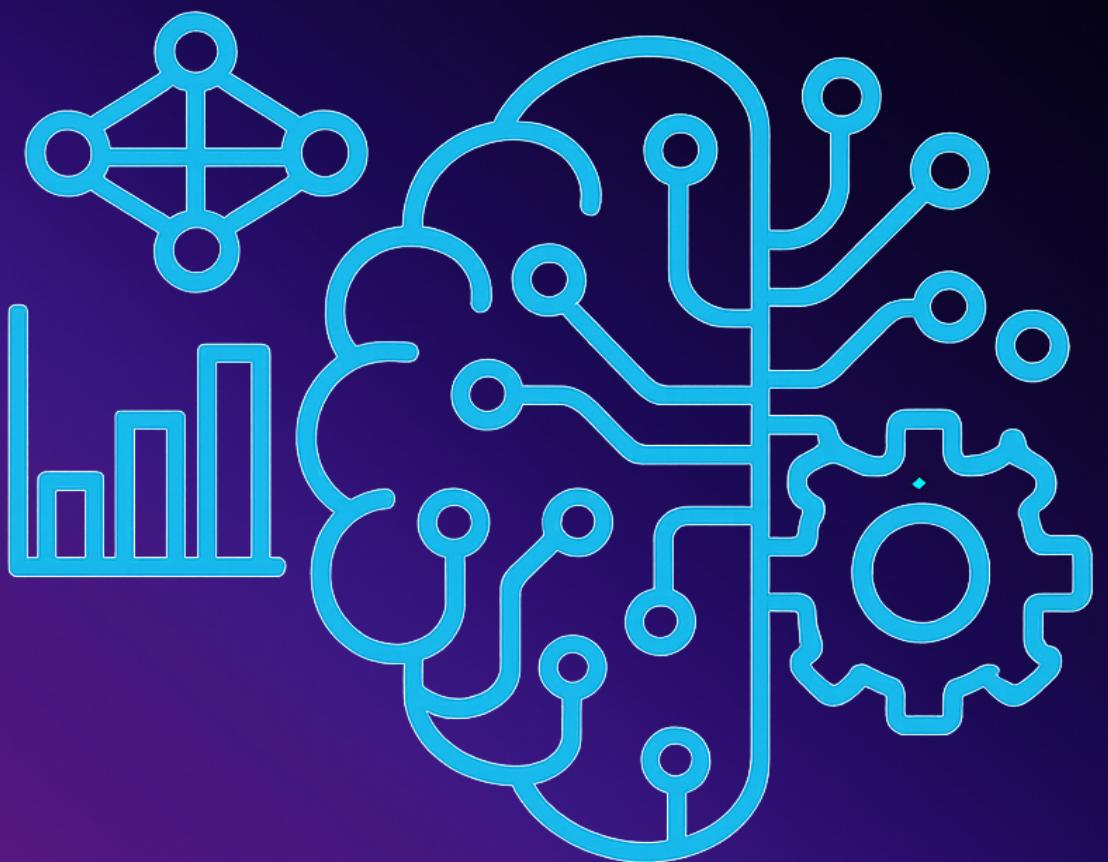


# Modelos utilizados

• **Modelo 1: Árvore de Decisão**

• **Modelo 2: Random Forest**

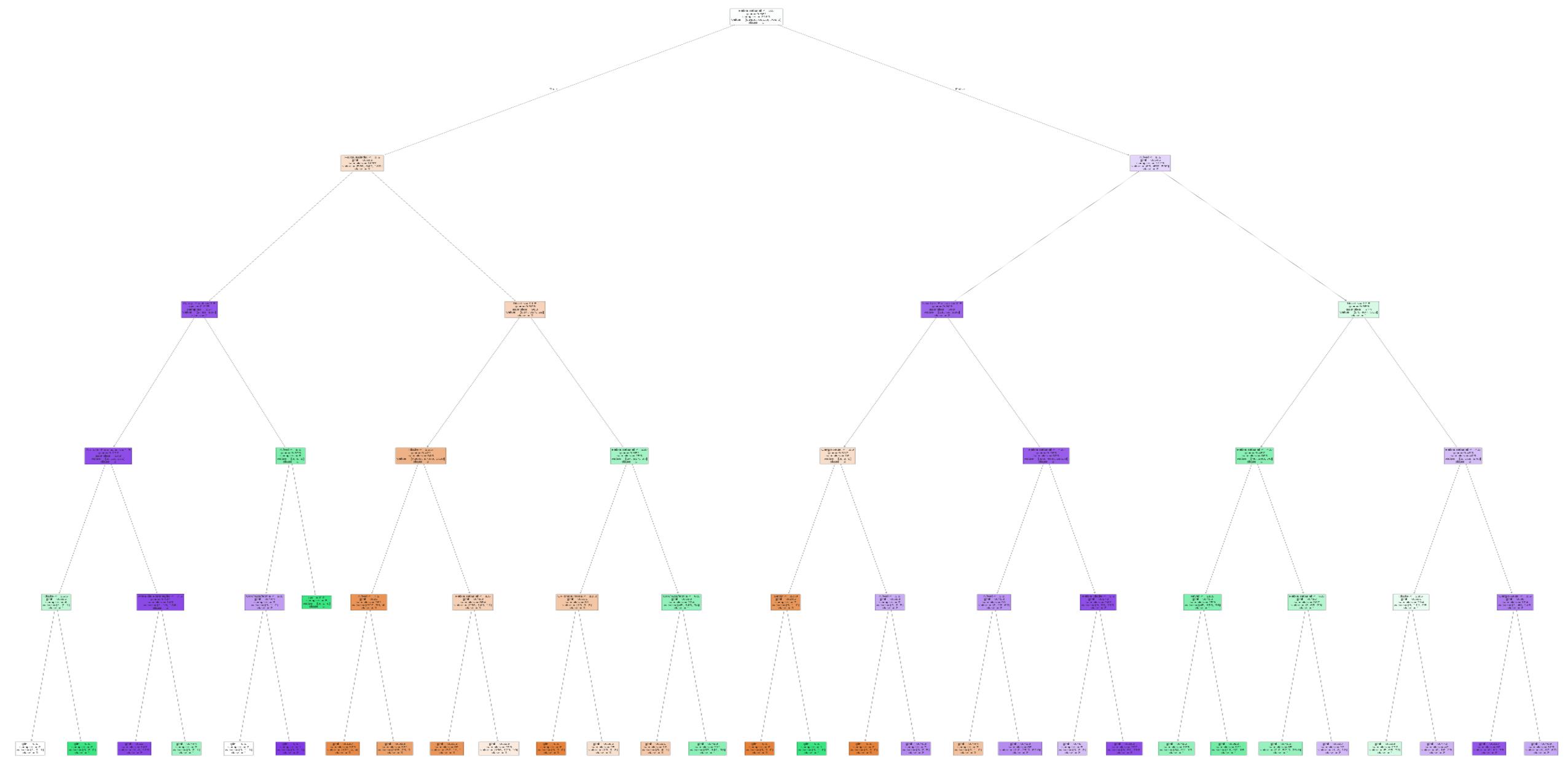
- **Pergunta Orientada a Dados:** Qual o nível predominante dos profissionais que trabalham com dados? (junior, pleno, senior)



# Árvore de Decisão

- 

O modelo Árvore de Decisão foi escolhido porque é fácil de entender e utilizar. Ele mostra de forma clara como as decisões são tomadas com base nos dados. Precisa de pouco tratamento dos dados antes de usar e consegue identificar quais informações são mais importantes para o resultado. Também é útil porque lida bem com situações complexas e pode ser ajustado para evitar erros por excesso de aprendizado.



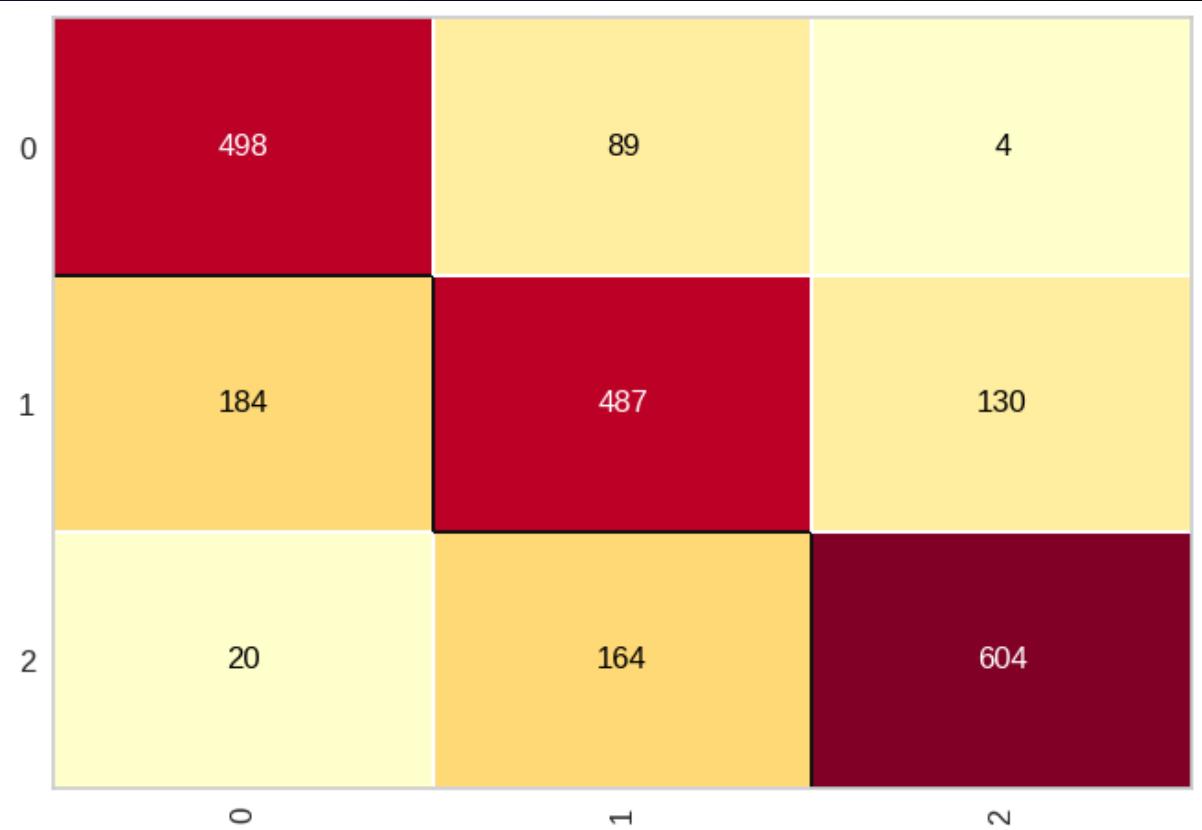
# Treino

## Precisão treino

Classe	Precision	Recall	F1-Score	Support
Júnior	0.71	0.84	0.77	591
Pleno	0.66	0.61	0.63	801
Sênior	0.82	0.77	0.79	788

Acurácia geral: 0.73

	Precision	Recall	F1-Score	Support
Macro Avg	0.73	0.74	0.73	2180
Weighted Avg	0.73	0.73	0.73	2180



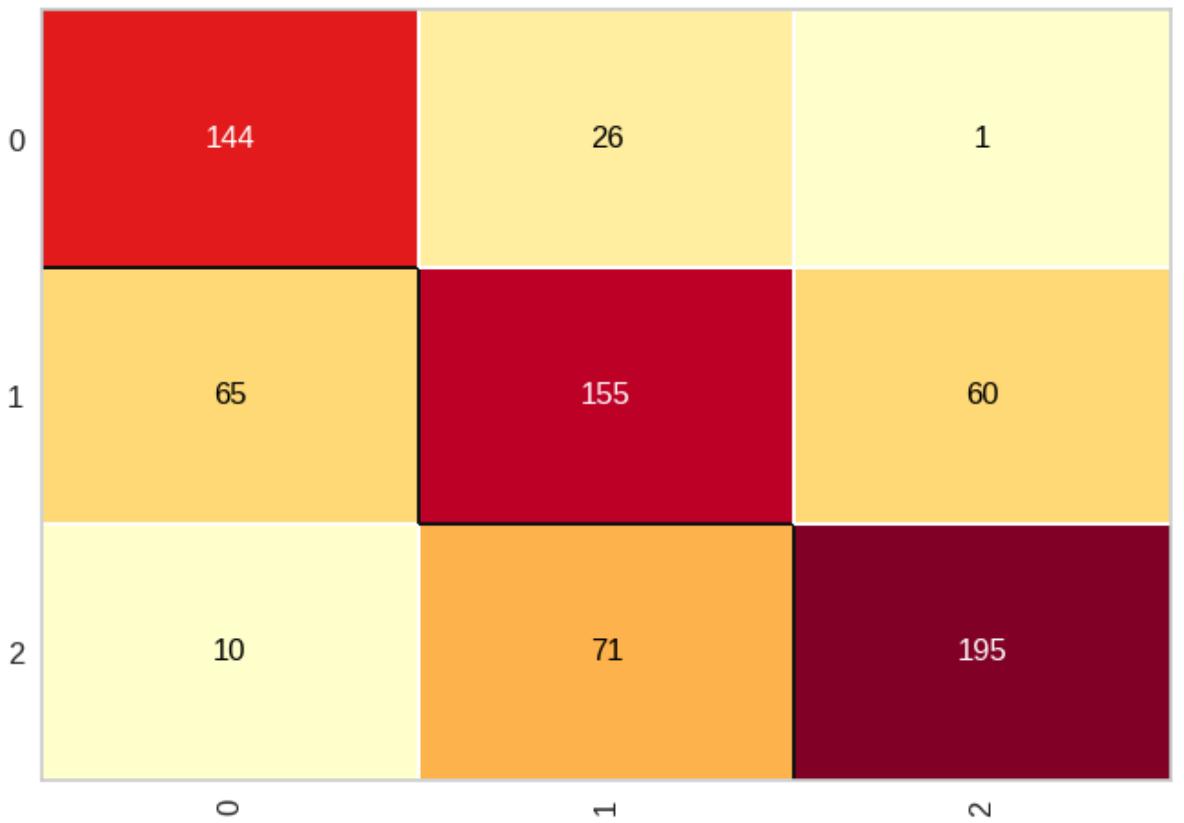
# Teste

## Precisão teste

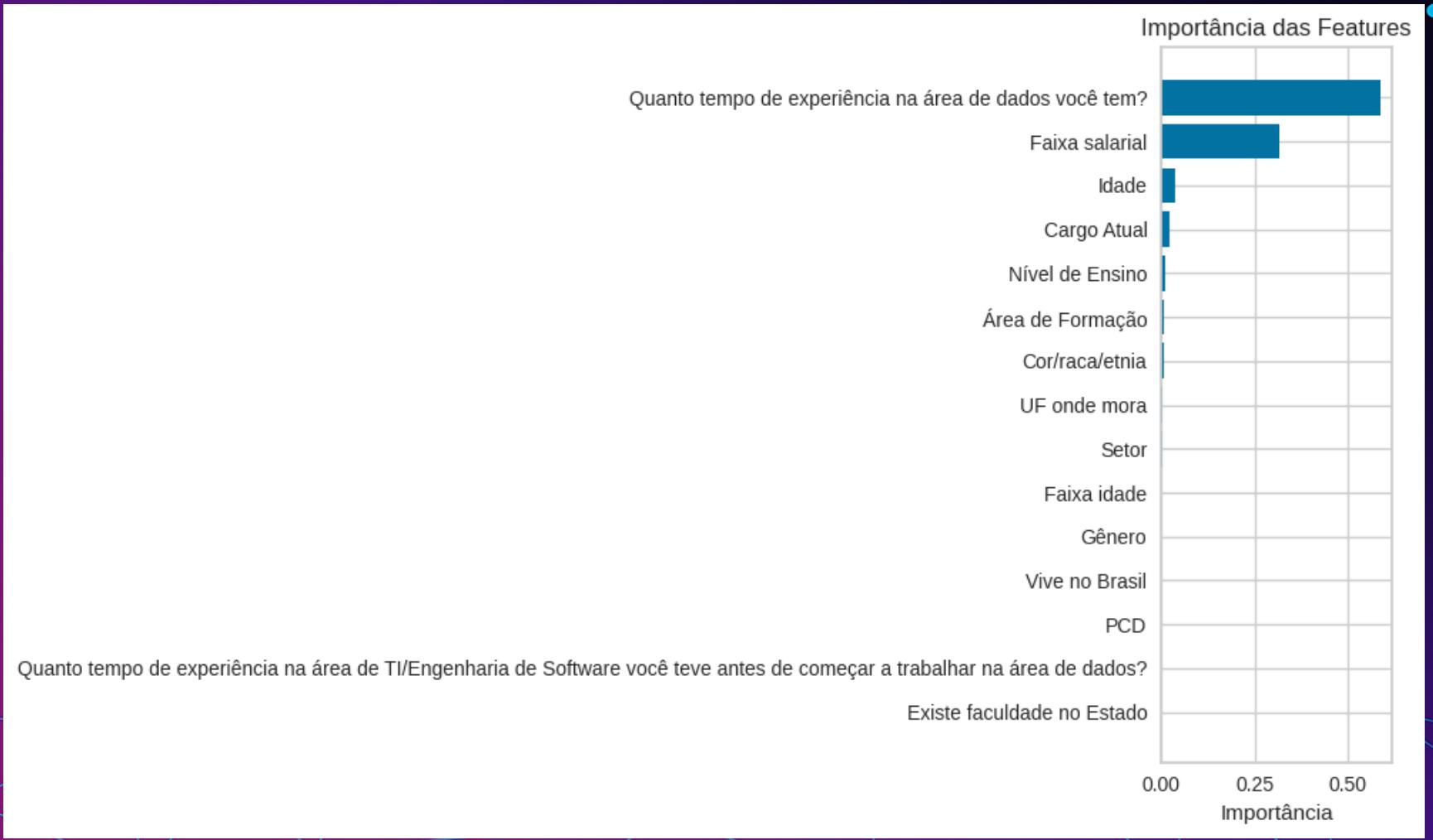
Classe	Precision	Recall	F1-Score	Support
Júnior	0.66	0.84	0.74	171
Pleno	0.62	0.55	0.58	280
Sênior	0.76	0.71	0.73	276

Acurácia geral: 0.68

	Precision	Recall	F1-Score	Support
Macro Avg	0.68	0.70	0.68	727
Weighted Avg	0.68	0.68	0.68	727



# Feature Importance



# Random Forest

- 

O modelo Random Forest foi escolhido porque ele possui características que o tornam uma das técnicas mais robustas e eficazes em problemas de classificação por: "Alta capacidade preditiva" em que o modelo combina múltiplas árvores de decisão, o que reduz significativamente a variância do modelo e melhora a generalização, evitando problemas de overfitting comuns; "Robustez a ruídos e outliers" onde devido ao processo de agregação (bagging), o modelo é menos sensível a dados ruidosos e a outliers, garantindo previsões mais estáveis e confiáveis; "Facil implementação no modelo de arvore de decisão" que para fazer o modelo arvore de decisão para random forest é bem facil.



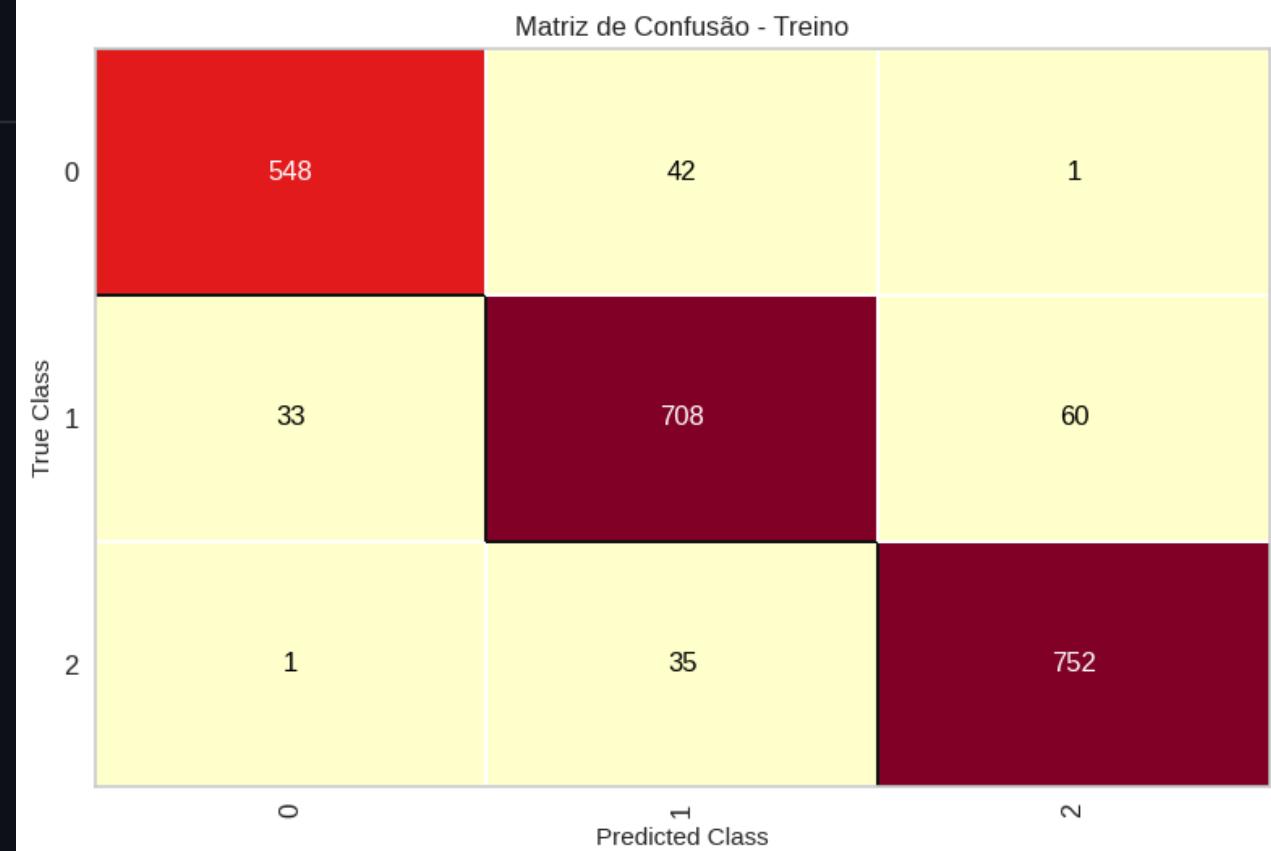
# Treino

## Precisão treino

Classe	Precision	Recall	F1-Score	Support
Júnior	0.94	0.93	0.93	591
Pleno	0.90	0.88	0.89	801
Sênior	0.92	0.95	0.94	788

Acurácia geral: 0.92

	Precision	Recall	F1-Score	Support
Macro Avg	0.92	0.92	0.92	2180
Weighted Avg	0.92	0.92	0.92	2180



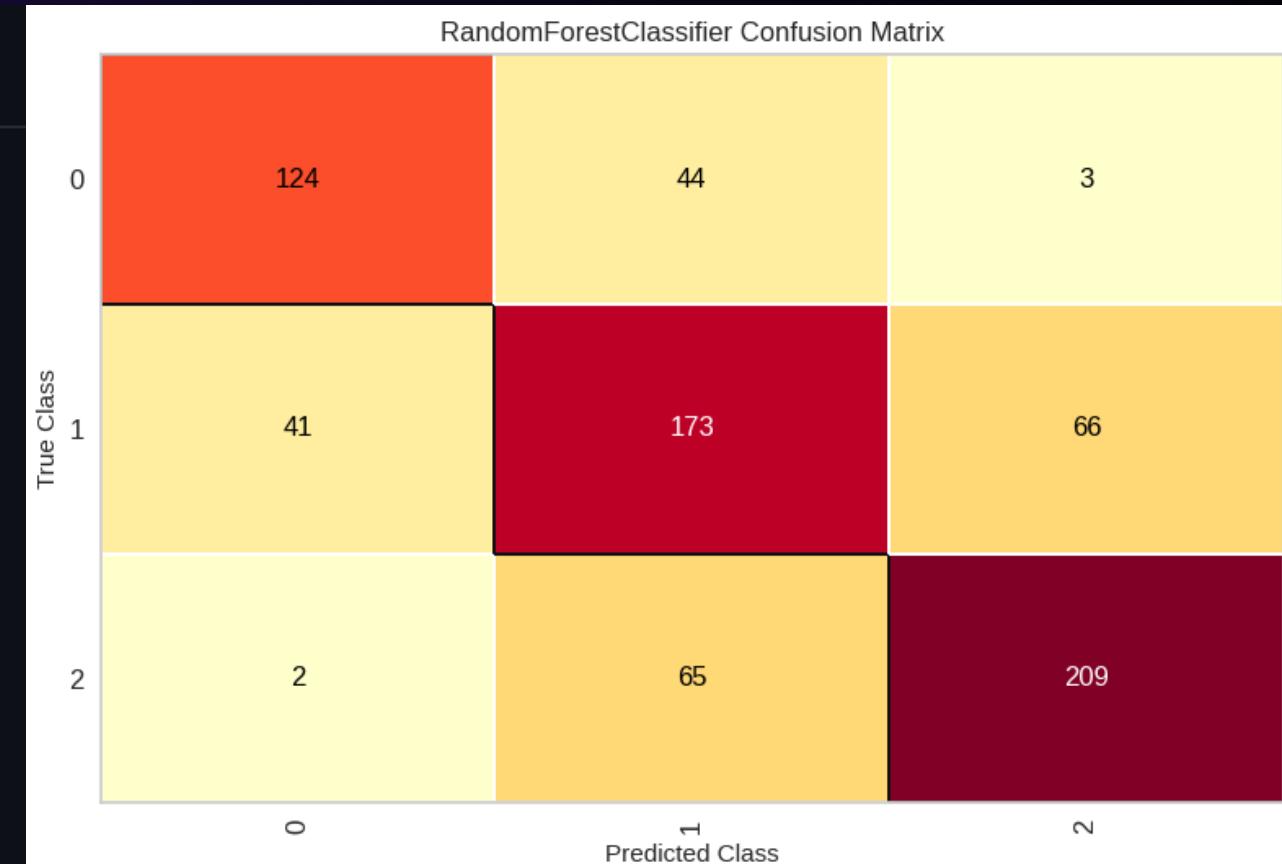
# Teste

## Precisão teste

Classe	Precision	Recall	F1-Score	Support
Júnior	0.74	0.73	0.73	171
Pleno	0.61	0.62	0.62	280
Sênior	0.75	0.76	0.75	276

Acurácia geral: 0.67

	Precision	Recall	F1-Score	Support
Macro Avg	0.69	0.66	0.67	727
Weighted Avg	0.68	0.66	0.67	727



# Feature Importance

