

# ***IA GENERATIVA: DESAFIOS & OPORTUNIDADES NO BRASIL***

ORIENTADORES: HUGO BASTOS E HAYALA CURTO.

INTEGRANTES: CAIO CÉSAR, THIAGO DOMINGOS, GABRIEL FERNANDES E  
NICOLAS RODRIGUES

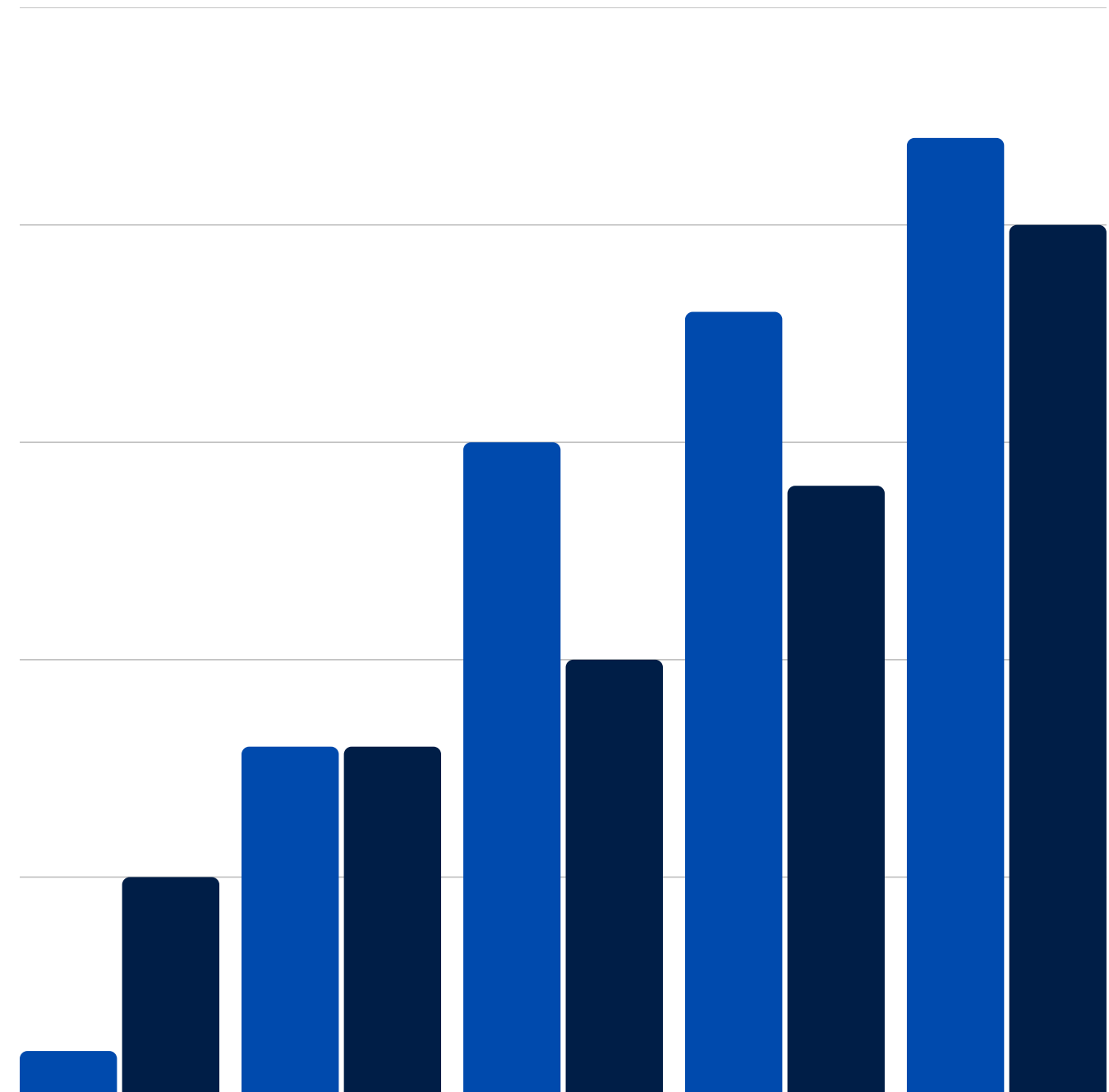
Tecnologias como a IA Generativa e os Large Language Models (LLMs) estão rapidamente remodelando setores, desde a saúde até a educação, transformando a forma como interagimos e trabalhamos. No entanto, sua adoção plena no Brasil ainda enfrenta obstáculos, especialmente para profissionais juniores e microempresas, grupos cruciais para a nossa economia. Nosso objetivo é analisar esses desafios e identificar as oportunidades para superá-los, desbloqueando o vasto potencial da IA Generativa em nosso país.

# SO BRE



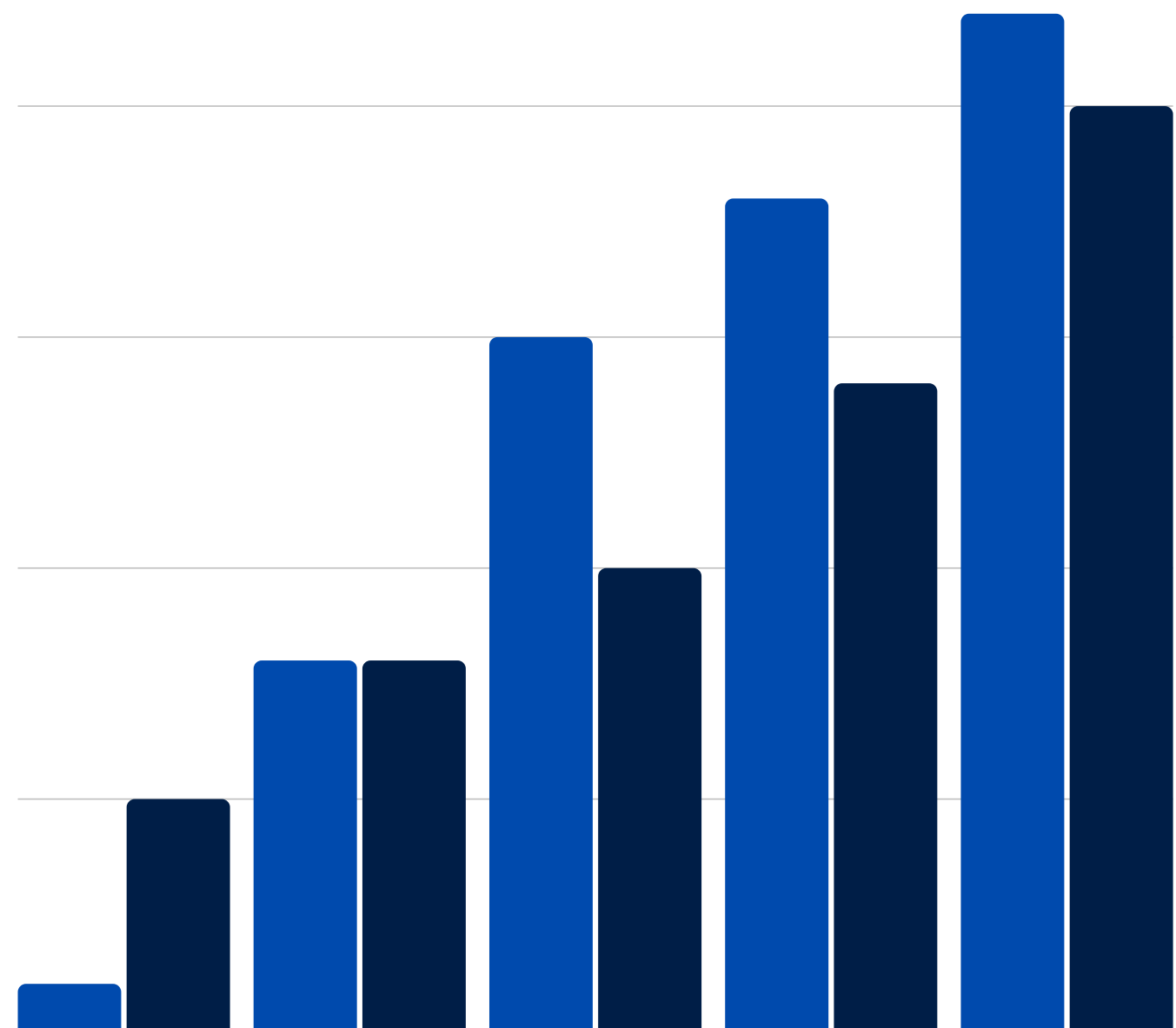
# *STATE OF DATA* **2023**

A base "State of Data Brazil 2023" é a espinha dorsal do nosso estudo. Ela nos trouxe um retrato fiel e atualizado dos profissionais de dados no Brasil, abrangendo demografia, experiência e aspirações de carreira. É por meio dela que pudemos entender as oportunidades e os desafios enfrentados por esses talentos, crucial para desvendar o panorama da IA Generativa em nosso país.

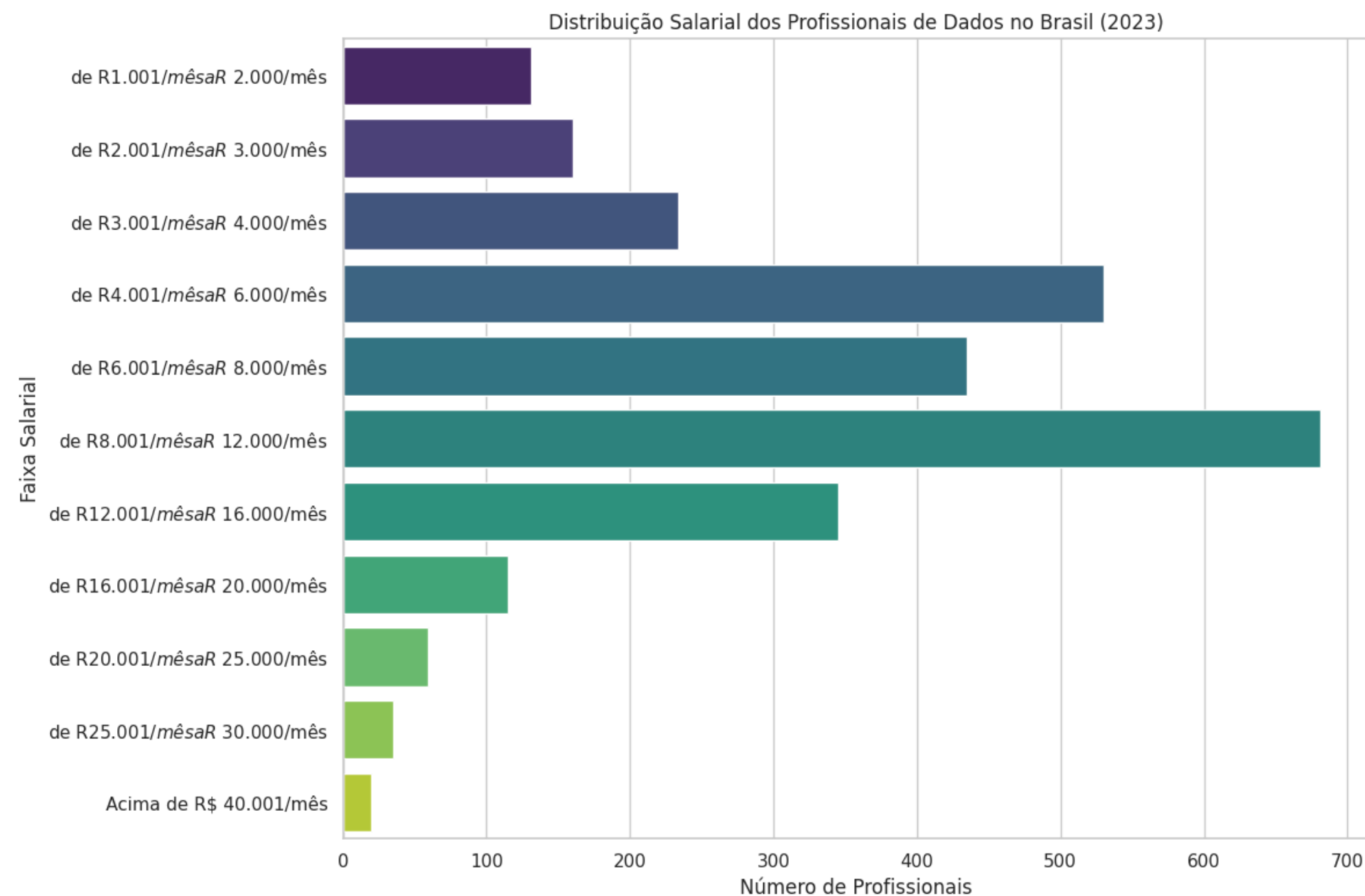


# *MICRODADOS DA EDUCAÇÃO* **2023**

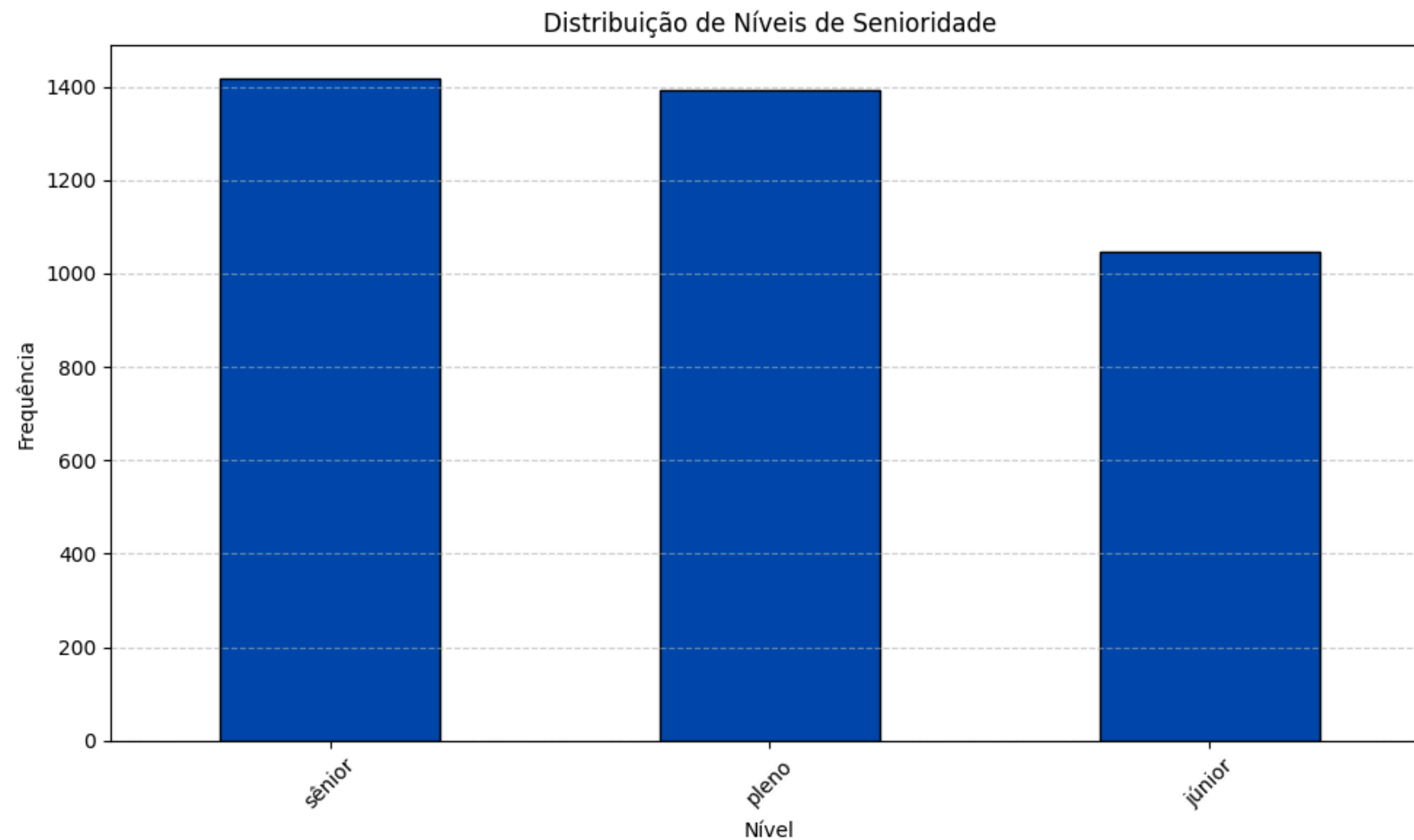
A base "MICRODADOS\_ED\_SUP\_IES 2023" complementou nossa análise, fornecendo um panorama detalhado das instituições de ensino superior no Brasil. Com ela, pudemos mapear a distribuição e a oferta de cursos, especialmente nas áreas ligadas à IA Generativa. Esses dados são essenciais para entender como a formação acadêmica se alinha com as necessidades do mercado e os desafios de adoção da IA.



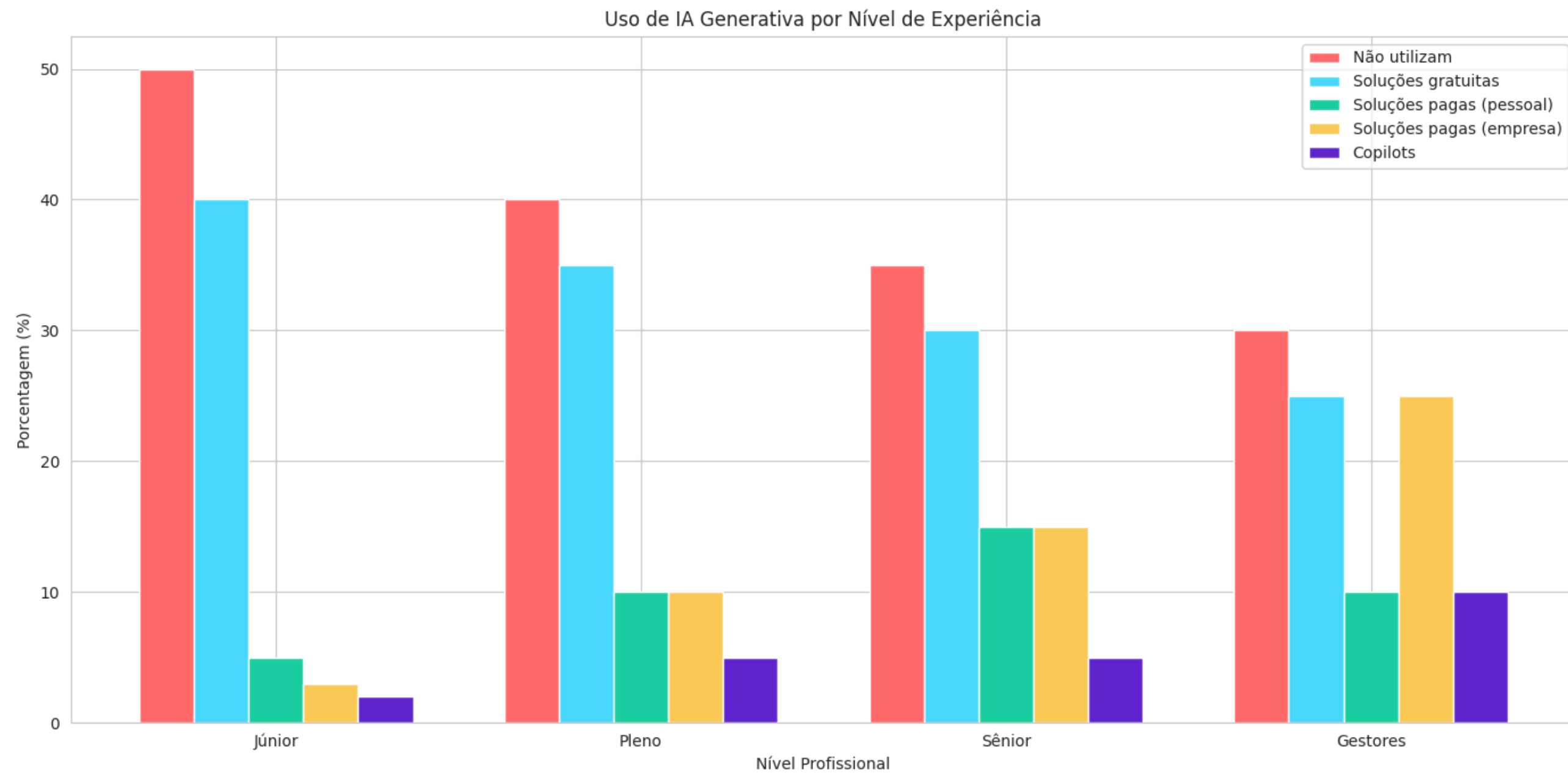
# ANÁLISE EXPLORATORIA



# *ANÁLISE* **EXPLORATORIA**



# ANÁLISE EXPLORATORIA



# *TRATAMENTO DE DADOS* **STATE OF DATA**

1. **Seleção de Colunas:** Foram escolhidas colunas-chave sobre demografia, carreira, satisfação e tecnologias (IA/Cloud).
2. **Tratamento de Dados:** linhas com respostas inválidas sobre satisfação no trabalho foram removidas; Valores ausentes foram preenchidos com 0 ou 'Desconhecido', dependendo da coluna. Os nomes das colunas foram padronizados para um formato simples e legível (ex: idade, genero).
3. **Nova Coluna:** Foi criada a coluna emprego\_status (***'Empregado'/'Desempregado'***).



# *TRATAMENTO DE DADOS* **MICRODADOS**

1. **Seleção de Colunas:** Foram usadas as colunas de **UF (SG\_UF\_IES)**, total de técnicos (**QT\_TEC\_TOTAL**) e total de docentes (**QT\_DOC\_TOTAL**).
2. **Agregação:** Os dados foram agrupados por estado **(UF)** para calcular a soma total de técnicos, docentes e o número de instituições em cada estado.

# JUNÇÃO DE DADOS

- **Método:** Foi realizada uma junção à esquerda (**left join**), mantendo todos os registros da pesquisa "State of Data".
- **Chave:** A união foi feita utilizando a coluna de estado de cada base de dados (**uf\_onde\_mora e uf\_ies**).

**Resultado:** Um arquivo único (**dados\_tratados\_combinados.csv**) onde cada respondente da pesquisa possui os dados educacionais agregados de seu estado.

# ***LIGHTGBM & XGBOOST*** **MODELOS**

O **LightGBM** é um algoritmo de aprendizado de máquina baseado em árvores de decisão, reconhecido por sua alta velocidade e eficiência. No contexto do problema, ele foi implementado dentro de um pipeline que primeiro pré-processa os dados (padronizando valores numéricos e codificando variáveis categóricas) e depois utiliza a técnica SMOTE para criar dados sintéticos da classe minoritária (profissionais "insatisfeitos").

O **XGBoost** é um dos mais robustos e populares algoritmos de gradient boosting, amplamente utilizado em competições e na indústria pela sua precisão e flexibilidade. A abordagem utilizada neste notebook também envolveu um pipeline de pré-processamento similar. No entanto, para tratar o desbalanceamento de classes, em vez de criar dados novos, foi utilizado o parâmetro interno do XGBoost, `scale_pos_weight`.

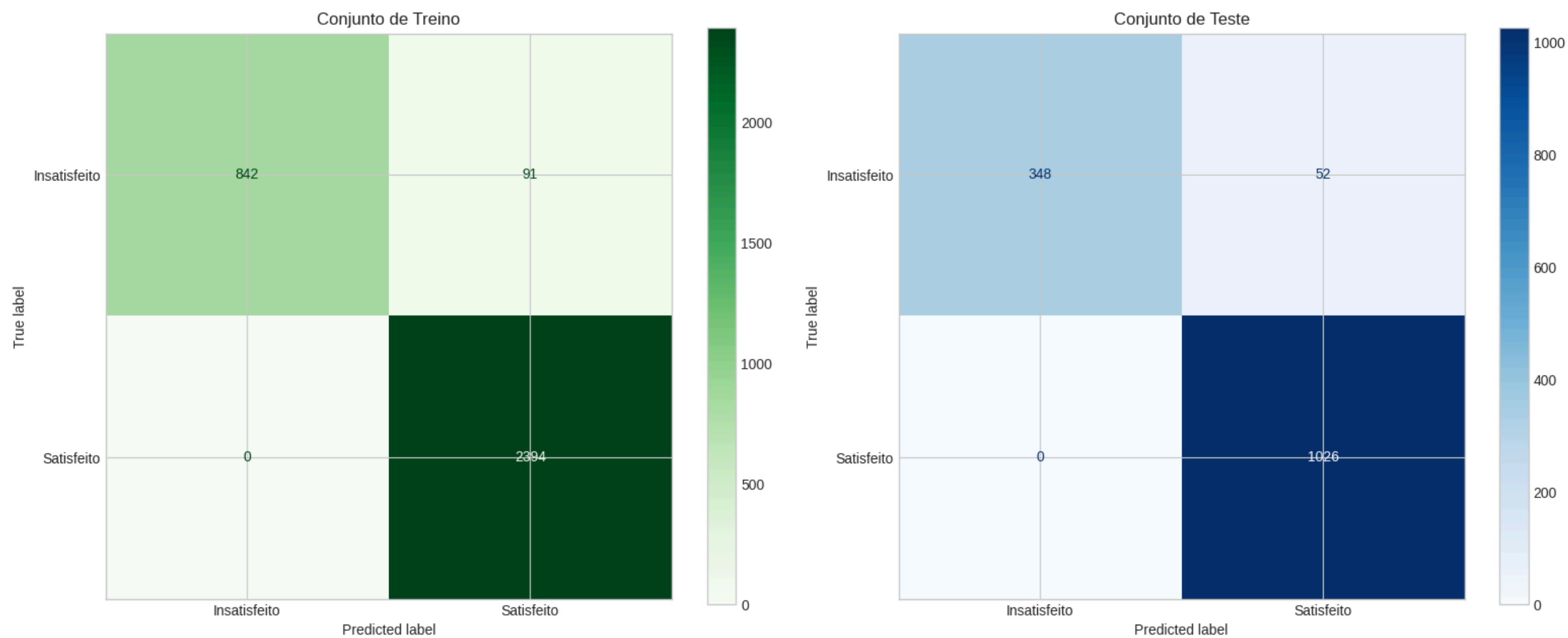
# ANÁLISE DE RESULTADOS

## LightGBM

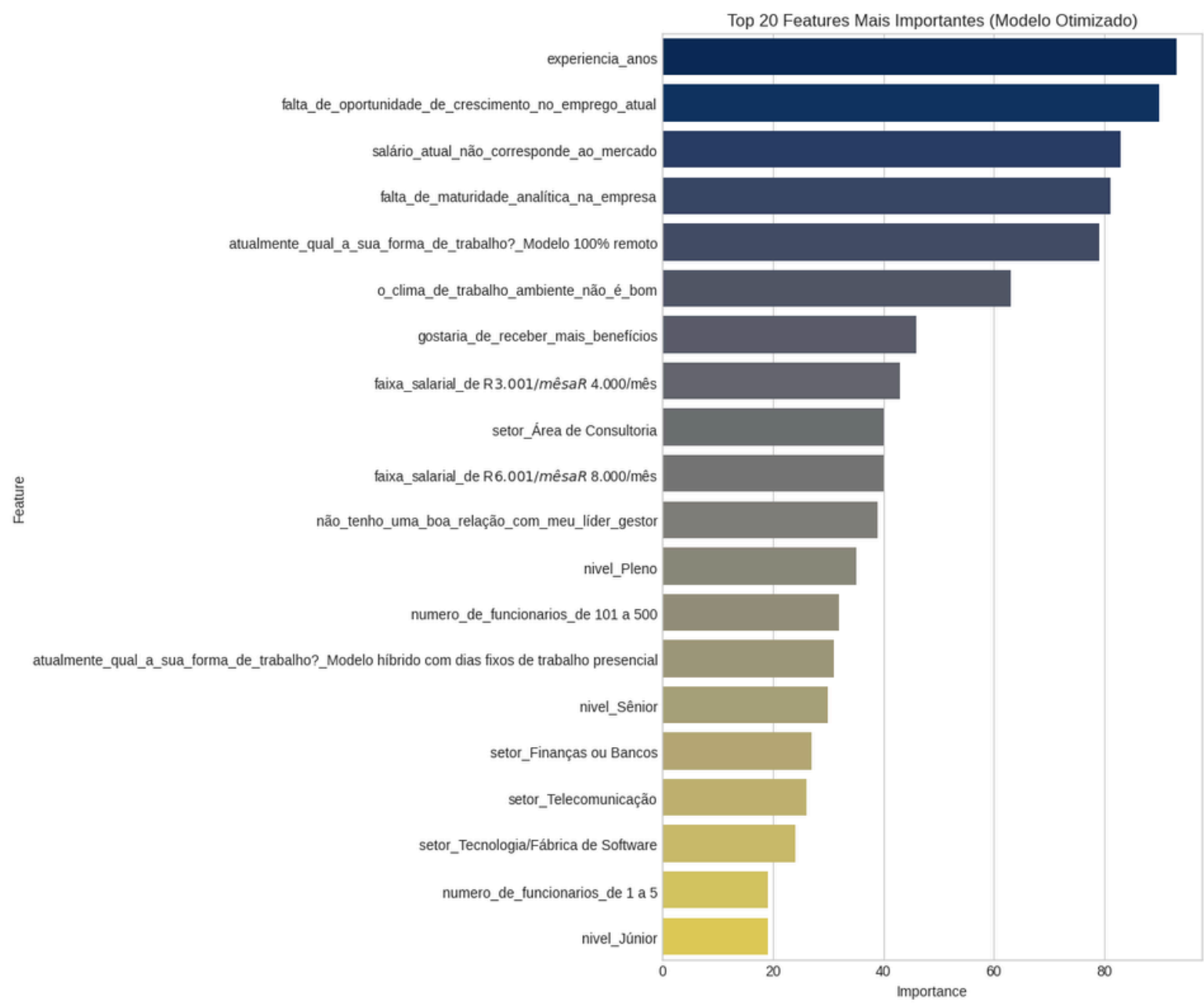
Métrica	Classe	Treino	Teste
Acurácia Geral	N/A	97%	96%
AUC-ROC	N/A	0.982	0.955
Precisão	Insatisfeito	1.00	1.00
Precisão	Satisfeito	0.96	0.95
Recall	Insatisfeito	0.90	0.87
Recall	Satisfeito	1.00	1.00
F1-Score	Insatisfeito	0.95	0.93
F1-Score	Satisfeito	0.98	0.98
F1-Score (Macro Avg)	N/A	0.97	0.95

# ANÁLISE DE RESULTADOS

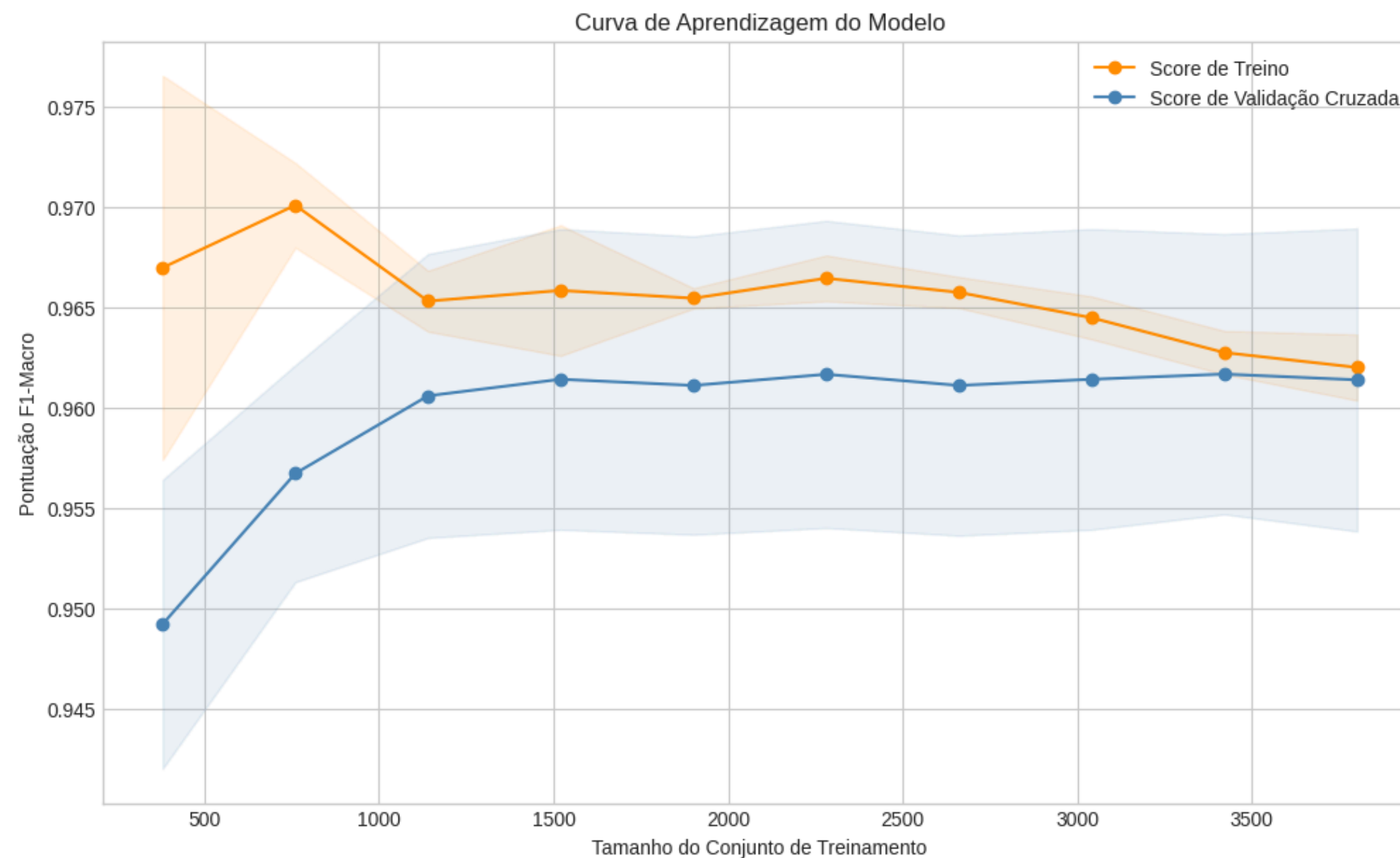
Matrizes de Confusão Comparativas



# ANÁLISE DE RESULTADOS



# ANÁLISE DE RESULTADOS



# ANÁLISE DE RESULTADOS

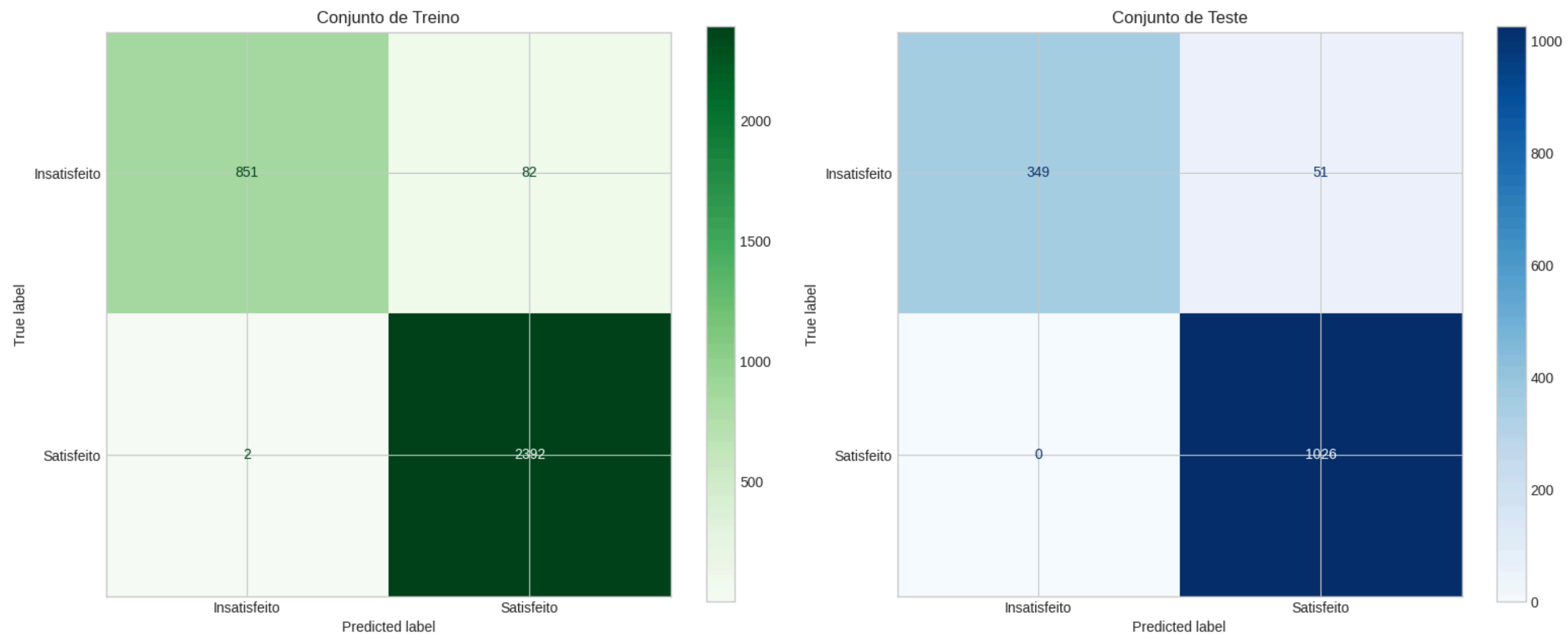
## XGBoost

Métrica	Classe	Treino	Teste
Acurácia Geral	N/A	97%	96%
AUC-ROC	N/A	0.988	0.961
Precisão	Insatisfeito	1.00	1.00
Precisão	Satisfeito	0.97	0.95
Recall	Insatisfeito	0.91	0.87
Recall	Satisfeito	1.00	1.00
F1-Score	Insatisfeito	0.95	0.93
F1-Score	Satisfeito	0.98	0.98
F1-Score (Macro Avg)	N/A	0.97	0.95

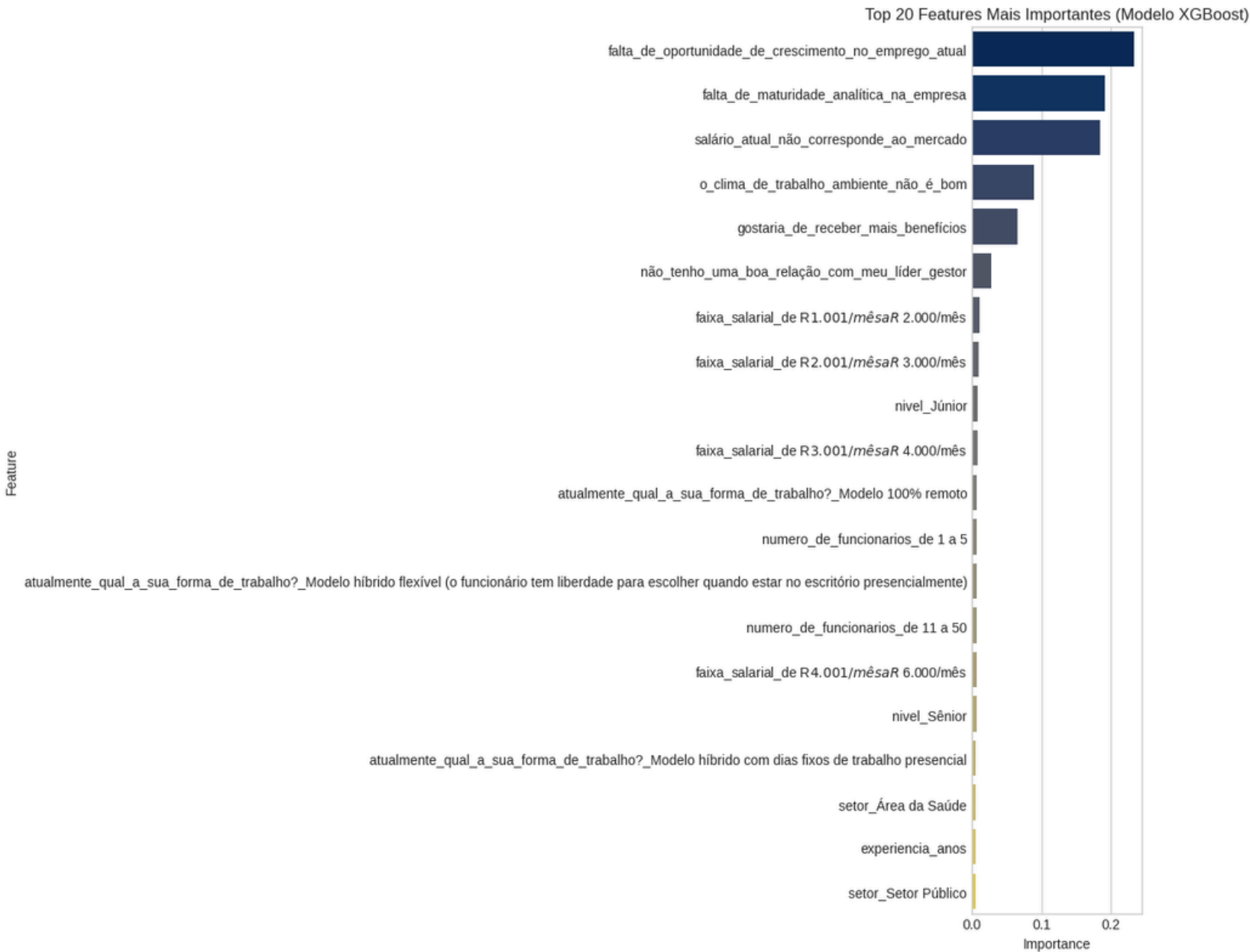


# ANÁLISE DE RESULTADOS

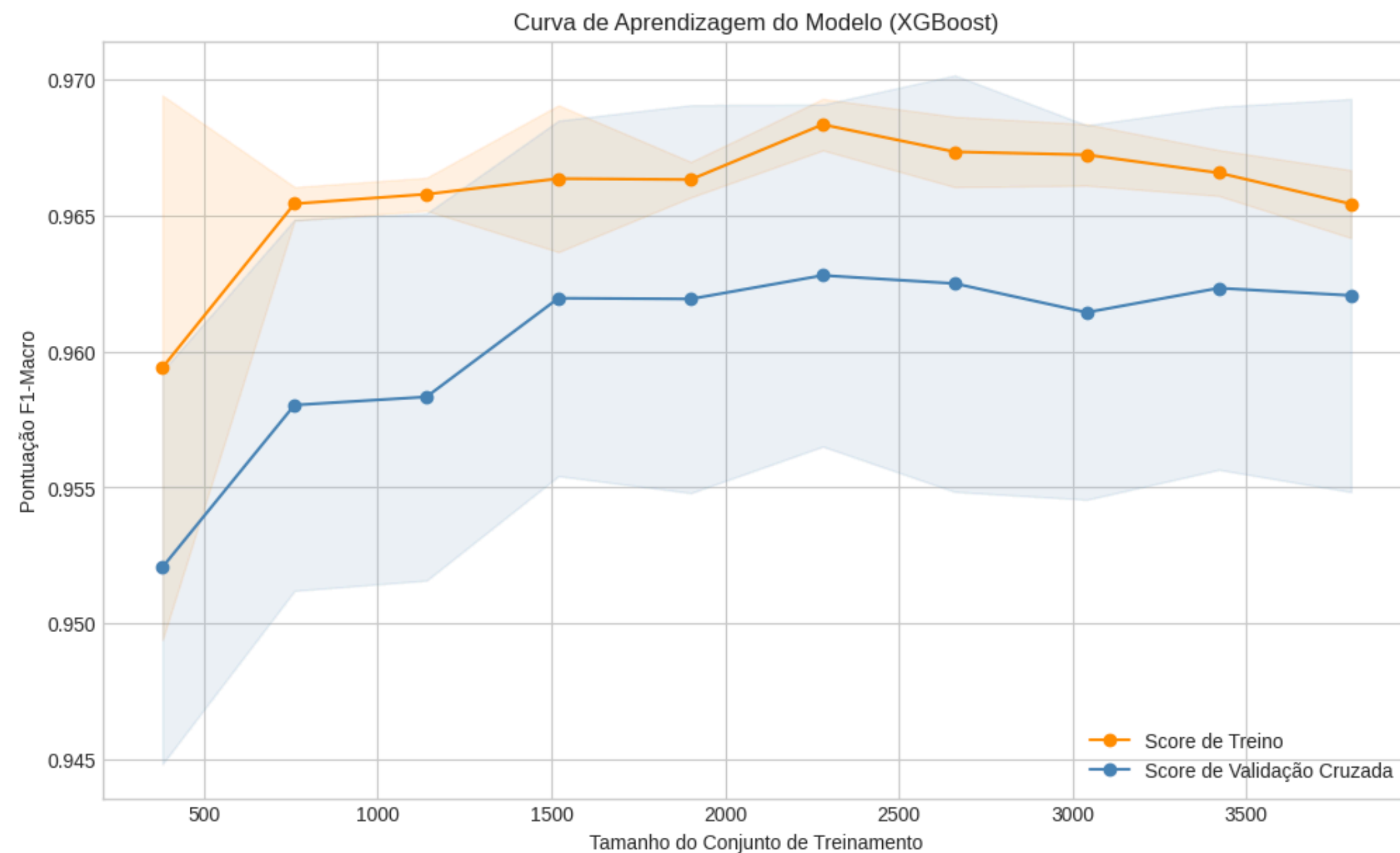
Matrizes de Confusão Comparativas (XGBoost)



# ANÁLISE DE RESULTADOS



# ANÁLISE DE RESULTADOS



## ***LIGHTGBM***

Nosso modelo utilizou o LightGBM para prever a satisfação dos profissionais de dados. No entanto, acreditamos que alguns fatores podem ter prejudicado o desempenho real do modelo. O principal deles é o possível vazamento de informação, pois features como os “motivos de insatisfação” estão altamente correlacionadas com a variável alvo, facilitando demais a tarefa de classificação e reduzindo o valor do modelo em aplicações práticas. Além disso, o uso do SMOTE para balanceamento pode ter criado exemplos sintéticos que não representam bem a população, impactando a robustez do modelo. Por fim, o conjunto de dados apresenta um viés de amostragem – como a maioria dos respondentes são de níveis pleno e sênior, o modelo pode não capturar adequadamente as particularidades dos profissionais juniores, limitando sua capacidade de generalização.

# *DIS CUS SÃO*



## ***XGBOOST***

No caso do modelo XGBoost, apesar do bom desempenho quantitativo, identificamos potenciais limitações que podem ter afetado sua performance e confiabilidade. O balanceamento entre as classes foi tratado via parâmetro de peso (*scale\_pos\_weight*), mas isso pode não ser suficiente quando há forte desbalanceamento, prejudicando a aprendizagem sobre a classe minoritária. O modelo também pode estar sofrendo com vazamento de informação, já que utiliza variáveis diretamente relacionadas à satisfação, o que limita sua utilidade em contextos reais de previsão. Adicionalmente, a distribuição desigual dos níveis de senioridade nos dados de entrada pode fazer com que o XGBoost aprenda padrões muito específicos de plenos e seniors, deixando de captar fatores importantes para os profissionais juniores.

**DIS  
CUS  
SÃO**

# CONCLUSÃO

Apesar dos bons resultados quantitativos dos nossos modelos, aprendemos que interpretar e confiar em previsões de satisfação profissional vai muito além de apenas olhar para as métricas. Fatores como perguntas que já entregam a resposta e a falta de diversidade nos perfis dos respondentes podem dar uma falsa impressão de acerto. Isso mostra que, para gerar valor real com ciência de dados, precisamos olhar sempre para a qualidade dos dados, buscar representatividade e garantir que o modelo realmente ajude a tomar decisões melhores no mundo real.