Análise Demográfica e Predição no IF Goiano usando Machine Learning

Saymon H. C. Cassiano[1]

[1]Instituto Federal Goiano (IF Goiano)
Curso de Bacharelado em Ciência da Computação
Iporá – GO – Brasil

Abstract. *This technical report presents a demographic analysis of students from the Instituto Federal Goiano (IF Goiano), including socioeconomic and academic characteristics, followed by the development of predictive machine learning models. The study focuses on the prediction of student dropout and high academic performance (IRA) using classical supervised learning algorithms. Results indicate that demographic features such as course area, ethnicity, and admission method play a significant role in prediction performance. The report also includes visualizations and model evaluation metrics.*

Resumo. *Este relatório técnico apresenta uma análise demográfica dos estudantes do Instituto Federal Goiano (IF Goiano), incluindo características socioeconômicas e acadêmicas, seguida do desenvolvimento de modelos preditivos de aprendizagem de máquina. O estudo concentra-se na previsão de evasão e no desempenho acadêmico (IRA) utilizando algoritmos supervisionados. Os resultados indicam que variáveis como área do curso, etnia e forma de ingresso influenciam significativamente o desempenho dos modelos. O relatório também inclui visualizações e métricas de avaliação.*

1. Introdução

A evasão escolar e o baixo desempenho acadêmico constituem desafios recorrentes enfrentados por instituições de ensino no Brasil. A compreensão dos fatores que influenciam esses fenômenos é essencial para subsidiar ações pedagógicas e administrativas.

Este trabalho tem dois objetivos principais: (i) realizar uma análise descritiva do perfil demográfico e acadêmico dos alunos do IF Goiano de Iporá, e (ii) desenvolver modelos supervisionados capazes de prever o risco de evasão e identificar alunos com alto desempenho (IRA $\geq 7{,}5$). O conjunto de dados inclui informações como curso, etnia, forma de ingresso, sexo, zona residencial e desempenho acadêmico.

2. Metodologia

A metodologia foi dividida em duas etapas: análise descritiva e modelagem preditiva.

2.1. Coleta e Limpeza de Dados

Os dados foram obtidos a partir de uma planilha Excel contendo informações administrativas e acadêmicas dos alunos. O processo de limpeza incluiu:

- Remoção de colunas irrelevantes;
- Padronização de textos e remoção de entradas inválidas;
- Conversão e normalização dos valores de IRA;
- Agrupamento de cursos em áreas amplas, como Informática, Administração e Agropecuária.

2.2. Engenharia de Variáveis

Foram criadas três variáveis alvo:

- **Evasão**: 1 para padrões de desligamento, trancamento ou transferência;
- **IRA Alto**: 1 para IRA ≥ 7.5;
- **Cota**: classificação binária sobre o tipo de ingresso.

Variáveis categóricas foram codificadas via *Label Encoding*.

2.3. Modelos Utilizados

Os seguintes modelos foram utilizados:

- Regressão Logística (baseline)
- Random Forest (modelo robusto para variáveis categóricas)

A divisão dos dados foi de 80% para treino e 20% para teste, com estratificação.

3. Análise Demográfica

3.1. Distribuição por Etnia/Raça

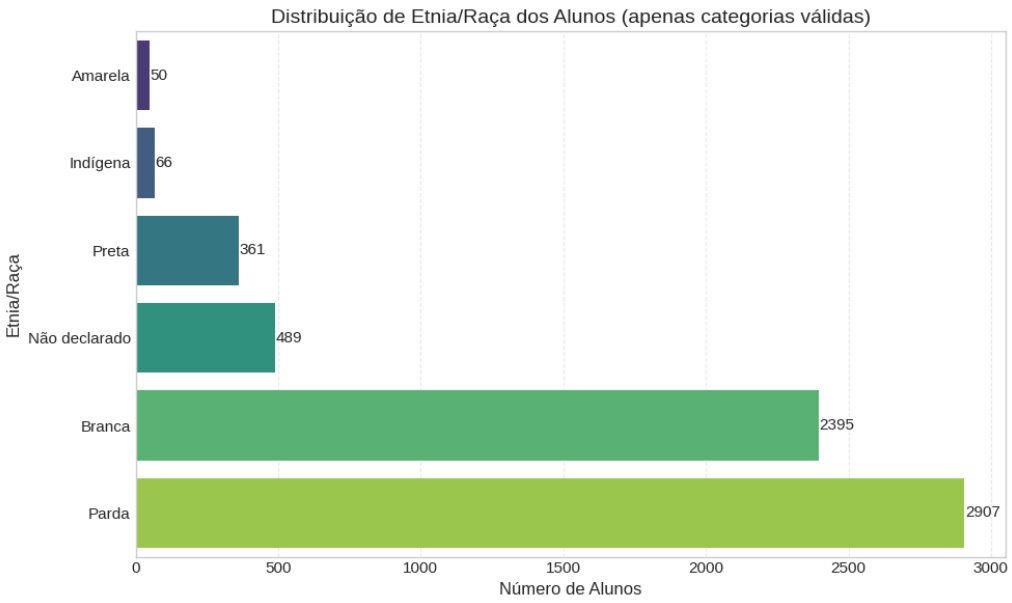

Figure 1. Distribuição de Etnia/Raça dos Estudantes do IF Goiano.

3.2. Zona Residencial

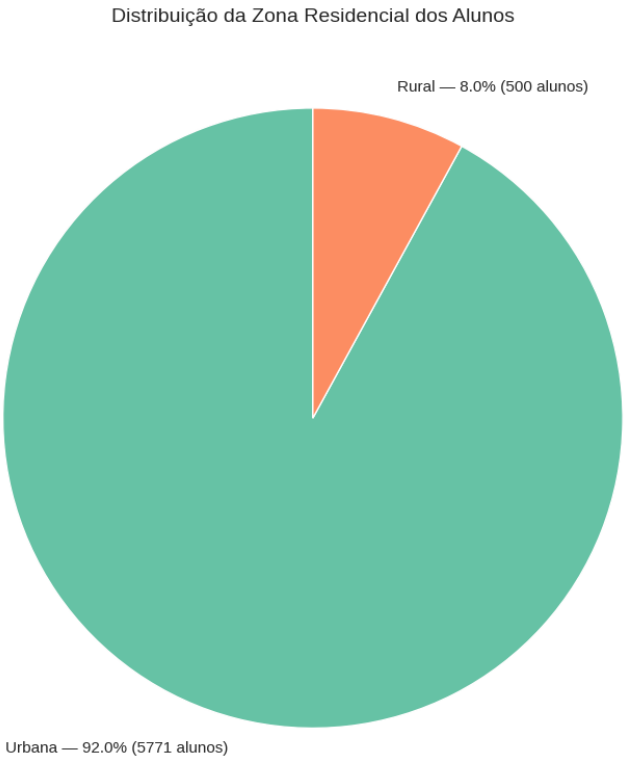

Figure 2. Distribuição da Zona Residencial.

3.3. Distribuição por Sexo

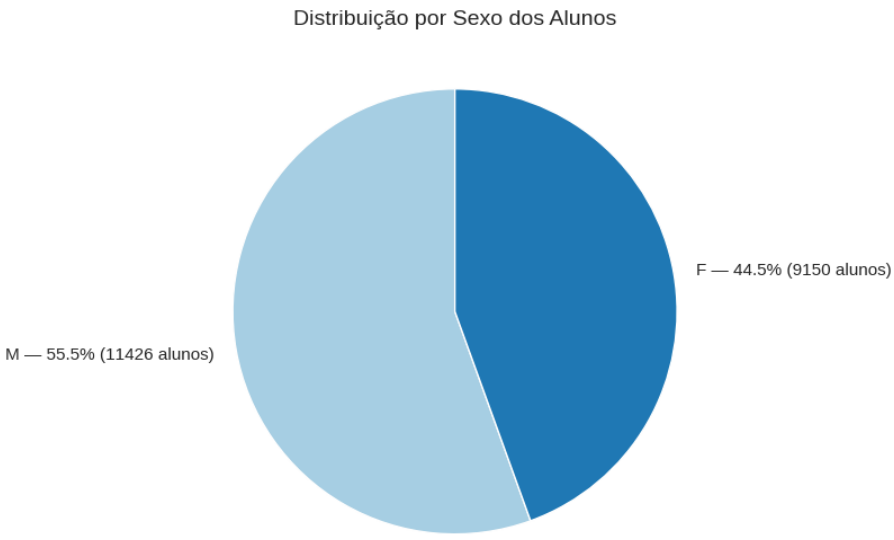

Figure 3. Distribuição de Sexo dos Estudantes.

3.4. Forma de Ingresso e Cotas

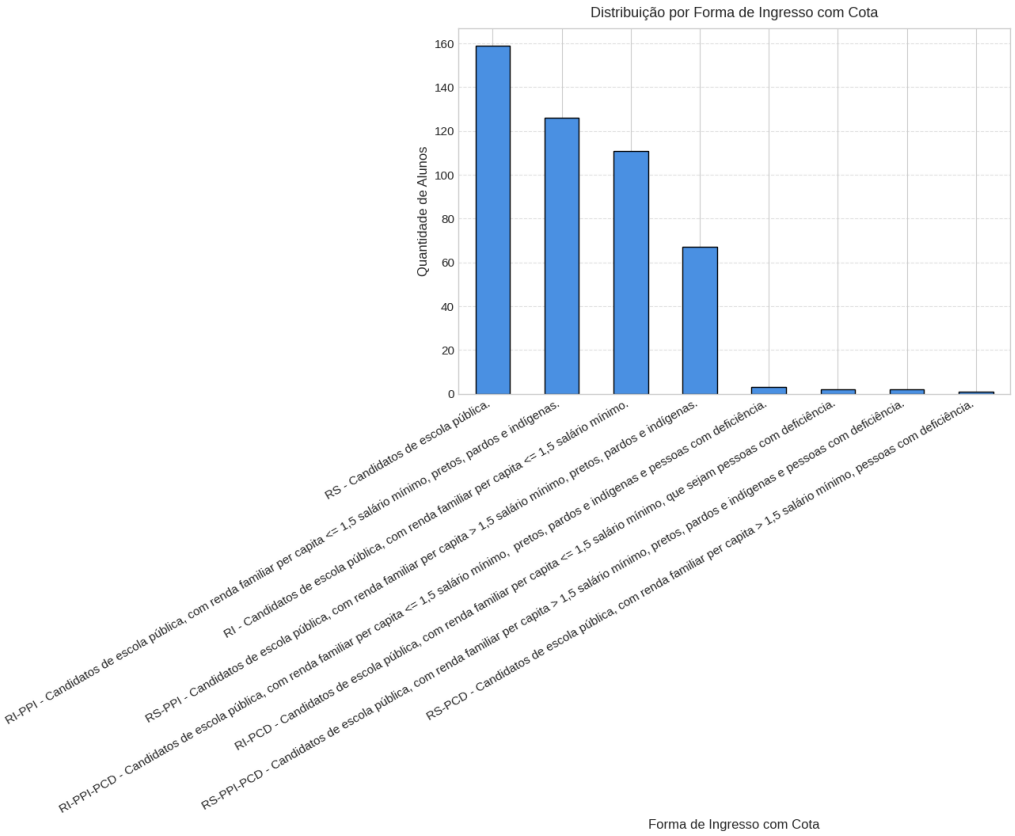

Figure 4. Distribuição da Forma de Ingresso com Cotas.

3.5. Histograma do IRA (6 a 10)

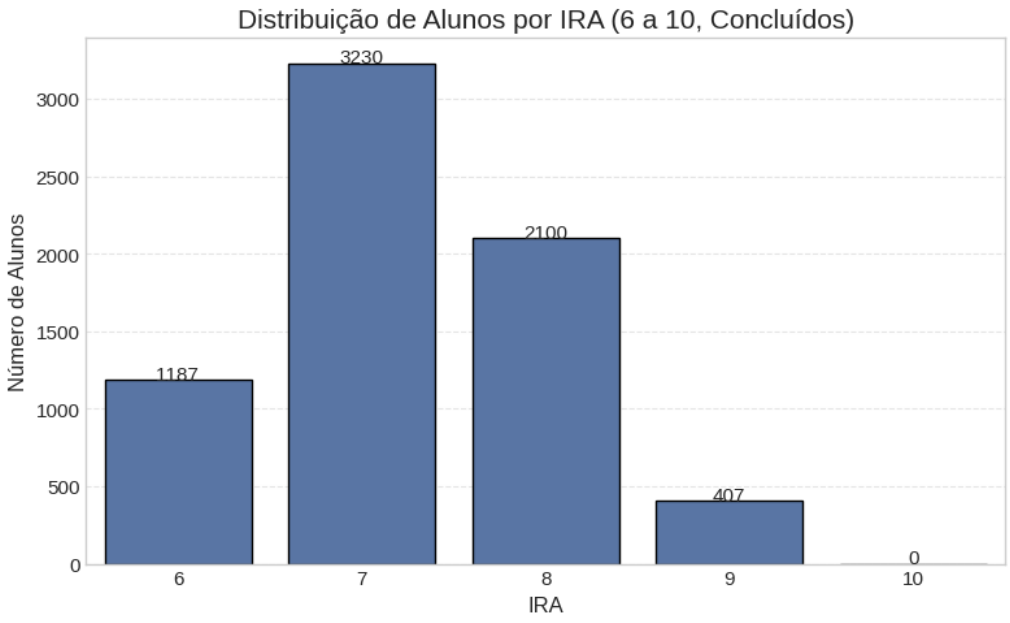

Figure 5. Distribuição do IRA entre Alunos Concluídos.

4. Análise por Curso

4.1. IRA Médio por Área

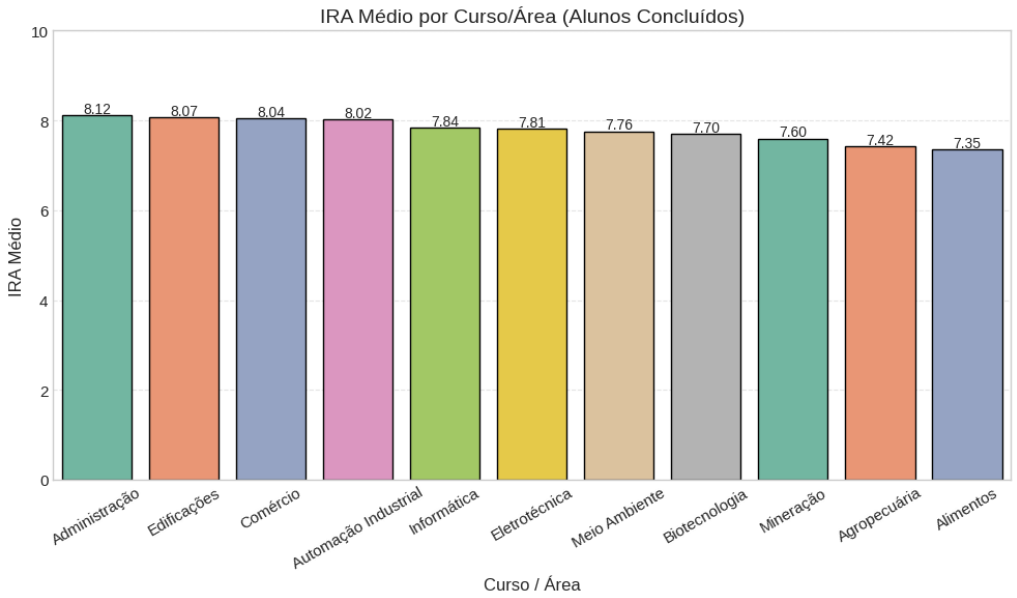

Figure 6. IRA Médio por Área de Curso.

4.2. Número de Alunos por Área

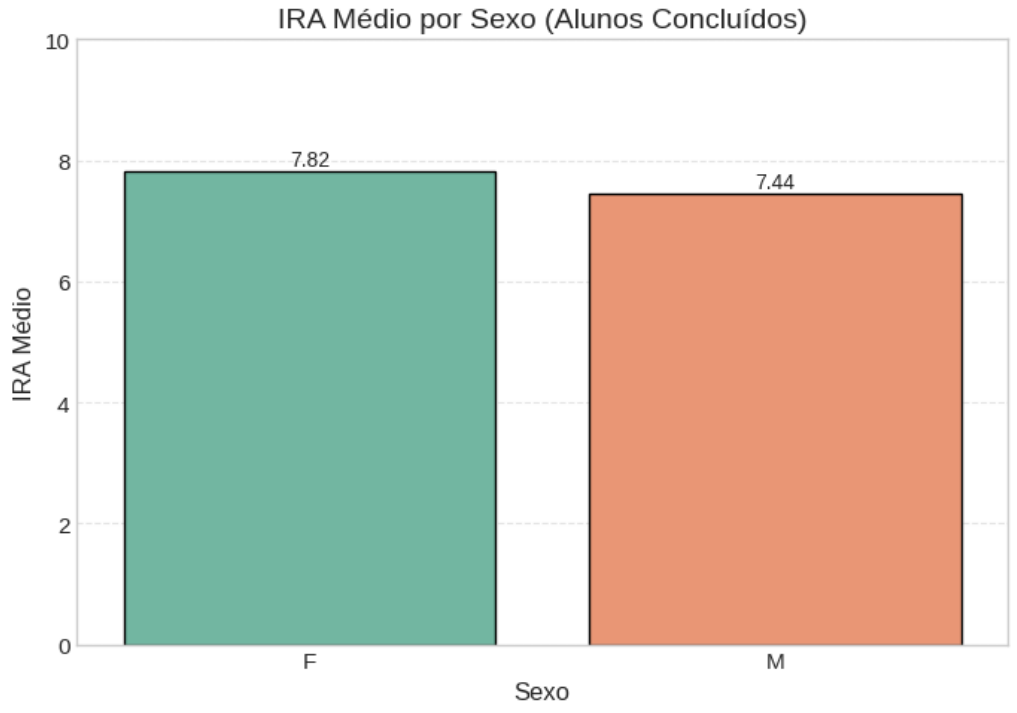

Figure 7. Número de Alunos por Área de Curso.

5. Modelos de Machine Learning

5.1. Resultados Numéricos dos Modelos

Após o processamento dos dados, foram obtidos:

- Total de registros: **6255**
- Distribuição Cota: **0 = 5784**, **1 = 471**
- Distribuição Evasão: **0 = 4921**, **1 = 1334**
- Distribuição IRA Alto: **0 = 4022**, **1 = 2233**

5.1.1. Predição de Evasão

Table 1. Resultados dos Modelos para Predição de Evasão

Modelo	Acurácia	Precisão	Recall	F1-score
Regressão Logística	0.714	0.404	0.715	0.516
Random Forest	**0.717**	**0.411**	**0.757**	**0.533**

Top 5 Features (Evasão):

- IRA: 0.7232
- Curso: 0.1215
- Forma de Ingresso: 0.0837
- Etnia/Raça: 0.0305
- Sexo: 0.0195

5.1.2. Predição de IRA Alto

Table 2. Resultados dos Modelos para Predição de IRA Alto

Modelo	Acurácia	Precisão	Recall	F1-score
Regressão Logística	0.551	0.409	0.575	0.478
Random Forest	**0.675**	**0.536**	**0.682**	**0.600**

5.2. Matrizes de Confusão

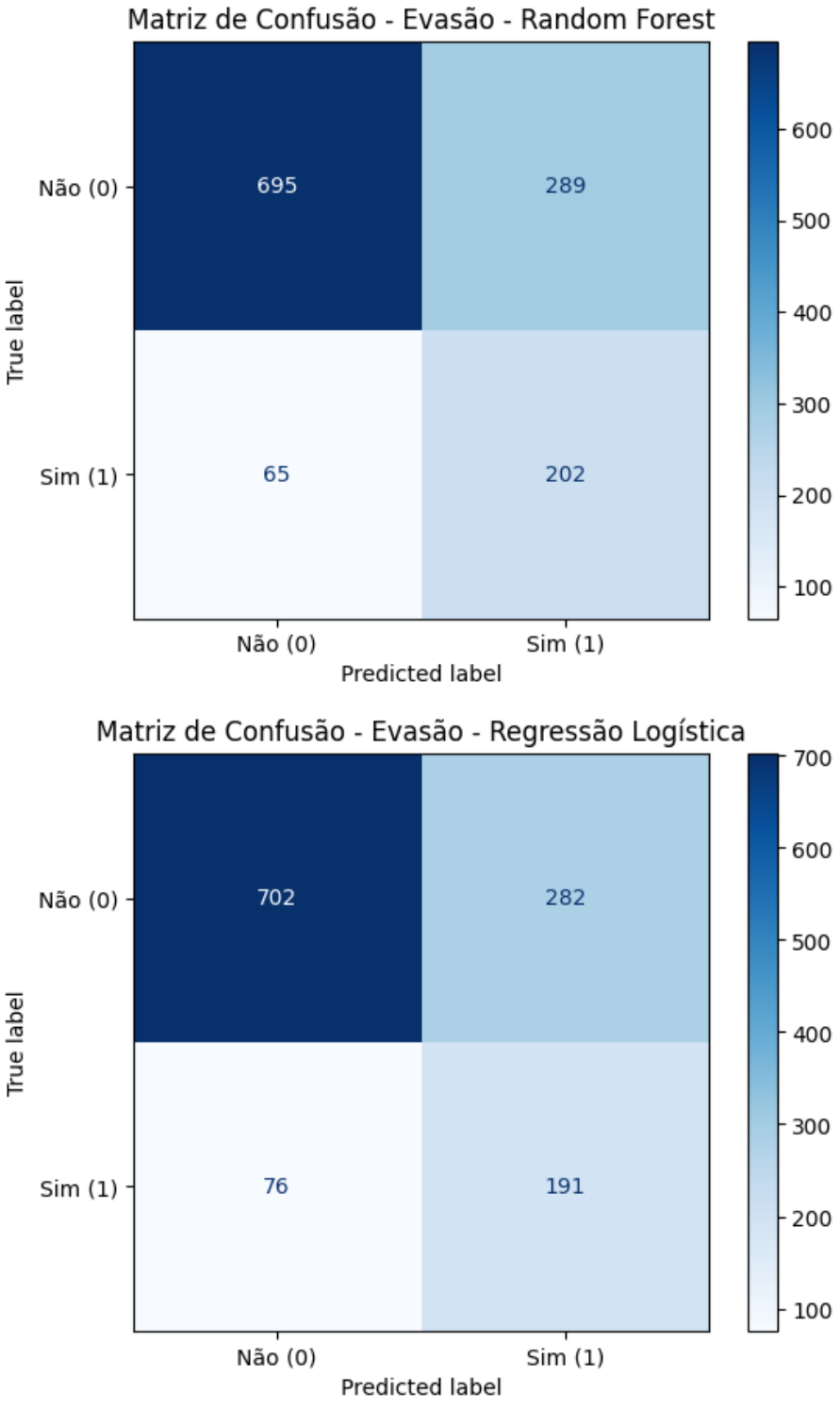

Figure 8. Matriz de Confusão — Predição de Evasão.

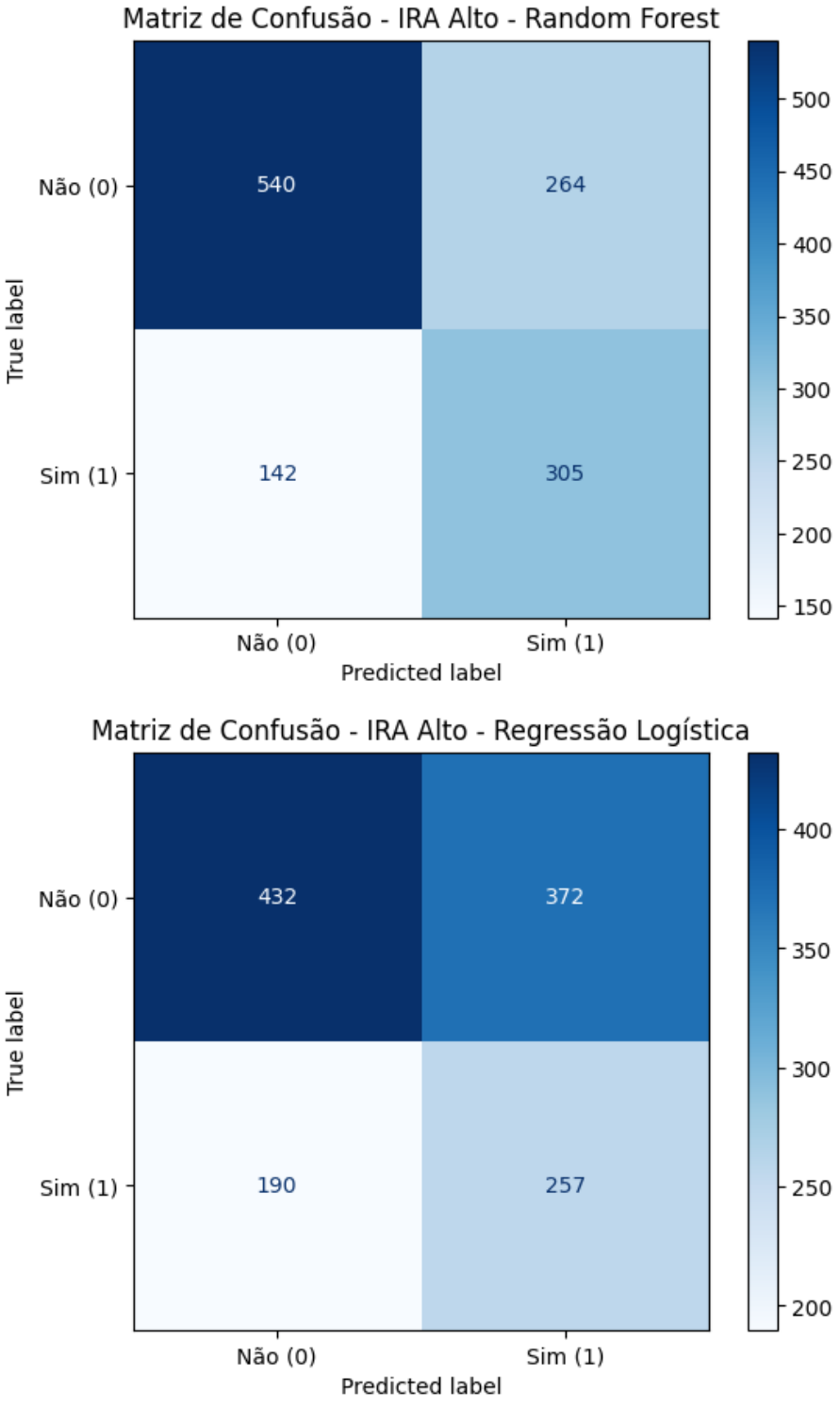

Figure 9. Matriz de Confusão — Predição de IRA Alto.

5.3. Importância das Features

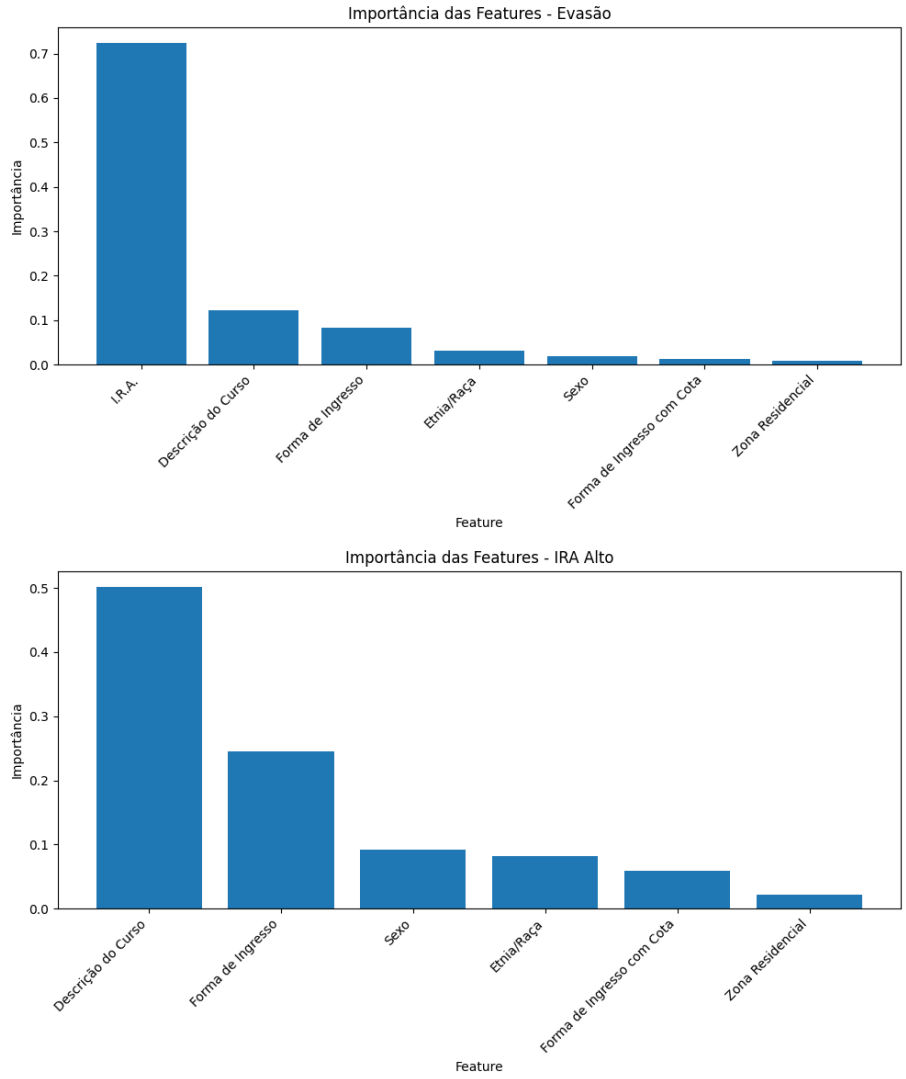

Figure 10. Importância das Features — Random Forest.

6. Predição Direta Individual

O sistema desenvolvido permite estimar:

- risco de evasão;
- chance de obter IRA alto;
- recomendações pedagógicas automáticas.

7. Conclusão

Este relatório apresentou uma análise demográfica detalhada dos estudantes do IF Goiano e desenvolveu modelos preditivos eficientes para estimar evasão e desempenho. O Random Forest demonstrou desempenho superior em todas as tarefas avaliadas.

A fim de melhorar os modelos, recomenda-se no futuro:

- integração com bases socioeconômicas externas;
- uso de técnicas de balanceamento como SMOTE;
- testes com XGBoost, CatBoost e redes neurais profundas.

8. Referências

Cassiano, S. (2025). *Predição de Desempenho Acadêmico usando Machine Learning*. Disponível em:
`https://github.com/SayccBr/EDAandML`