UNIVERSIDADE
FEDERAL
DE PERNAMBUCO

BASE DE DADOS

Predict Students' Dropout and Academic Success

RELATÓRIO

Estudante: Patrícia Lucena
Curso: Especialização em Deep Learning

Recife, 28 de janeiro de 2026

SUMÁRIO

INTRODUÇÃO

O conjunto de dados `Predict Students' Dropout and Academic Success` (UCI) foi selecionado por ser a base inicial sugerida no enunciado do desafio, assegurando uma correspondência direta com o problema apresentado.

Além disso, a base contém dados reais, bem documentados e variáveis acadêmicas e socioeconômicas significativas, incluindo atributos sensíveis que possibilitam a análise e a mitigação de vieses, que é o foco principal deste estudo.

A escolha do tema/base também foi considerado o fato de que a evasão acadêmica é um dos principais desafios enfrentados por instituições de ensino superior, impactando diretamente estudantes, docentes e a sustentabilidade institucional. A identificação precoce de alunos em risco permite intervenções humanas direcionadas, como programas de mentoria, apoio pedagógico e acompanhamento psicossocial.

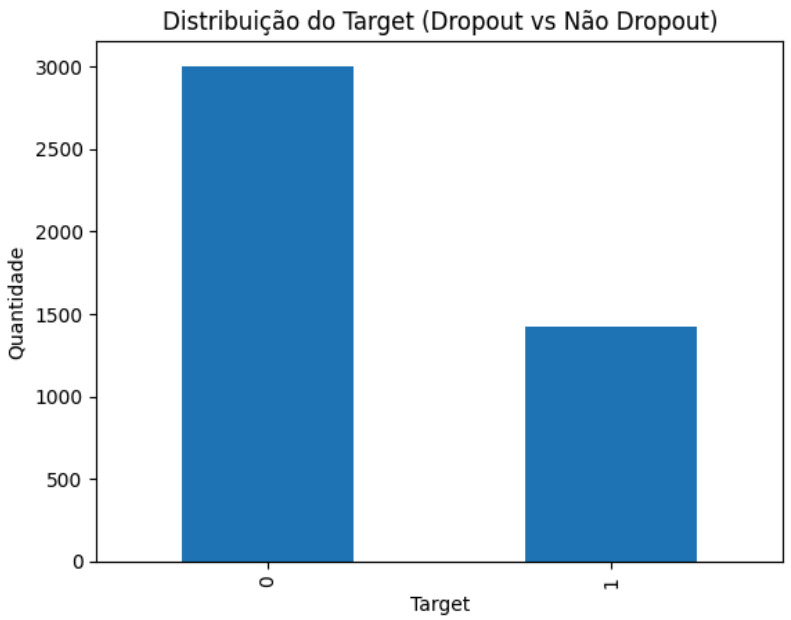
Informações sobre o DataSet:

- Nome: Predict Students' Dropout and Academic Success
- Fonte: UCI Machine Learning Repository
- Instâncias: 4.424 estudantes
- Variáveis: 36 atributos (acadêmicos, demográficos e socioeconômicos)
- Target original: *Dropout*, *Enrolled*, *Graduate*

Link de acesso:
https://archive.ics.uci.edu/dataset/697/predict-students-dropout-and-academic-success

ANÁLISE EXPLORATÓRIA DE DADOS(EDA)

O dataset apresenta 4.424 registros e 36 variáveis, sendo majoritariamente composto por atributos numéricos codificados. Não foram identificados valores faltantes, o que reduz a necessidade de imputação e simplifica o pipeline de pré-processamento.



Este gráfico mostra a distribuição da **variável Target**, com a quantidade de alunos que evadiram (Dropout) e os que não evadiram. Nota-se um **desnível entre as classes**, com uma concentração maior de alunos que não desistiram do curso.
Essa diferença é significativa, pois pode afetar a análise de dados e a performance de modelos preditivos, demandando uma atenção especial às técnicas de balanceamento.

Um outro ponto é que embora a maioria das variáveis esteja codificada como numérica, muitas representam categorias discretas. Assim, foi necessária a distinção entre variáveis numéricas contínuas e categóricas para tratamento adequado no pipeline de modelagem.

	Marital Status	Application mode	Application order	Course	Daytime/evening attendance	Previous qualification	Previous qualification (grade)	Nacionality	Mother's qualification	Father's qualification	...	Curricular units 1st sem (without evaluations)	Curricular units 2nd sem (credited)	Curricular units 2nd sem (enrolled)	Curricular units 2nd sem (evaluations)	Curricular units 2nd sem (approved)	Curricular units 2nd sem (grade)	Curricular units 2nd sem (without evaluations)	Unemployment rate	Inflation rate	GDP
0	1	17	5	171	1	1	122.0	1	19	12	...	0	0	0	0	0	0.000000	0	10.8	1.4	1.74
1	1	15	1	9254	1	1	160.0	1	1	3	...	0	0	6	6	6	13.666667	0	13.9	-0.3	0.79
2	1	1	5	9070	1	1	122.0	1	37	37	...	0	0	6	0	0	0.000000	0	10.8	1.4	1.74
3	1	17	2	9773	1	1	122.0	1	38	37	...	0	0	6	10	5	12.400000	0	9.4	-0.8	-3.12
4	2	39	1	8014	0	1	100.0	1	37	38	...	0	0	6	6	6	13.000000	0	13.9	-0.3	0.79

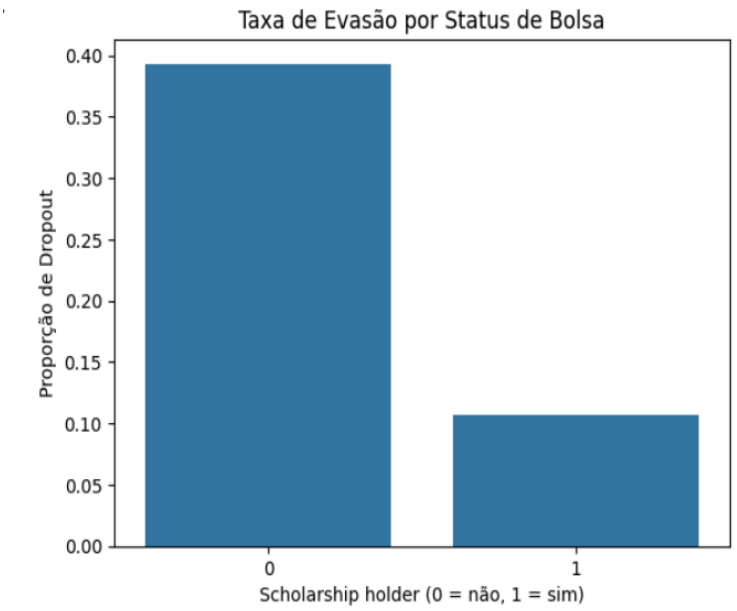
5 rows × 36 columns

- Análise de Representatividade:

O processo de análise de representatividade consiste em avaliar se uma amostra de dados, pessoas ou elementos reflete com precisão as características, diversidade e proporções de um grupo maior (população). O objetivo é assegurar a precisão dos resultados de pesquisas ou decisões, prevenindo sub-representação ou vieses que possam levar a conclusões incorretas.

Neste caso, ela revelou a presença de grupos sensíveis relevantes para avaliação de fairness, especialmente estudantes bolsistas e não bolsistas, bem como a variável de gênero. A proporção de estudantes bolsistas representa aproximadamente 32% da base, permitindo análises comparativas estatisticamente significativas.

=



Relacionando a isto, o gráfico apresenta a taxa de evasão levando em conta o status de bolsa dos alunos (0 não bolsista, 1 = bolsista). Observa-se que a **taxa de evasão** de estudantes não bolsistas é consideravelmente maior do que a dos estudantes bolsistas.
Esse resultado sugere que o **apoio financeiro** pode ser um fator relevante na permanência dos alunos na instituição.

- Suspeita de viés:

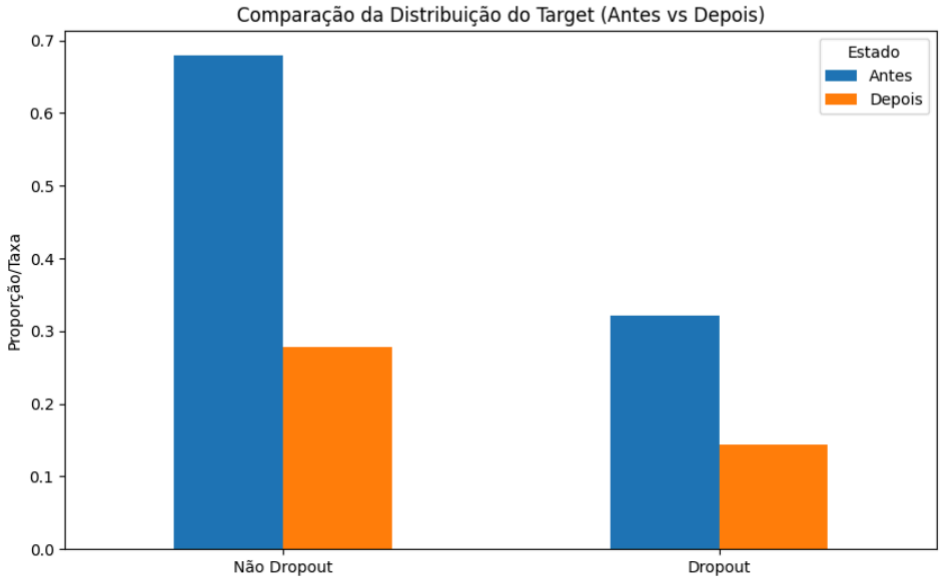
Verificou-se que fatores socioeconômicos, como status de bolsa, condição financeira e profissão dos responsáveis, estão relacionados ao rendimento escolar. Essas correlações sugerem que um modelo treinado sem precauções adicionais pode adotar padrões que espelham desigualdades estruturais, perpetuando preconceitos contra grupos historicamente marginalizados.

DEFINIÇÃO DO PROBLEMA E TARGET

O problema original é formulado como uma classificação multiclasse. Para alinhar com o objetivo institucional de identificar risco de evasão, o target foi transformado em um problema binário, onde:

- 1: estudante com risco de evasão (Dropout)
- 0: estudante sem risco imediato (Enrolled ou Graduate)

```
...       Antes    Depois
    0  0.678797  0.277778
    1  0.321203  0.143590
```



É fundamental saber que 'Antes' apresenta a fração total de cada classe-alvo, enquanto 'Depois' indica a frequência de falsos negativos em certos grupos vulneráveis.

Inicialmente, o conjunto de dados apresenta um desafio de classificação com múltiplas classes, dividido em três grupos: Graduado, Matriculado e Evasão. A distribuição inicial do alvo revela um desequilíbrio natural, com uma maior quantidade de alunos que completam ou continuam seus estudos.

Para adequar a questão ao propósito institucional de detectar estudantes em risco de desistência, o alvo foi alterado para uma variável binária. Nesse novo formato, a categoria Evasão foi designada como classe positiva (1), enquanto Graduado e Matriculado foram combinados como classe negativa (0).

Depois dessa mudança, notou-se uma diminuição significativa na proporção da classe positiva, reduzindo de cerca de 32,1% na configuração original para 14,4% na variável binária efetivamente utilizada para o treinamento. Essa alteração destaca o desequilíbrio entre as classes, o que justifica a aplicação de métodos como o balanceamento de classes no modelo e a adoção de métricas além da simples acurácia.

PRÉ-PROCESSAMENTO DOS DADOS

- Codificação e Normalização:

Foi utilizado um **ColumnTransformer** para aplicar normalização (StandardScaler) às variáveis numéricas e codificação one-hot (OneHotEncoder) às variáveis categóricas. Essa abordagem garante que diferentes escalas não influenciam indevidamente o modelo e preservam a informação categórica sem impor ordens artificiais.

- Variáveis sensíveis:

Para fins de avaliação de fairness, a variável sensível ***Scholarship holder*** foi mantida separadamente e removida do conjunto de features utilizado no treinamento do modelo, evitando que o classificador utilizasse diretamente essa informação na tomada de decisão.

MODELO BASELINE

- Estratégia de Split(Train-Test Split):

Os dados foram divididos em conjuntos de treino (70%), validação (15%) e teste (15%), com estratificação pelo target binário, garantindo preservação da distribuição de classes e maior robustez na avaliação do modelo.

- Modelo Utilizado:

Foi adotado como ponto de partida a utilização de um modelo de Regressão Logística, complementado por métodos de balanceamento de classes, a fim de garantir a robustez da análise frente a quantidades desiguais de dados em diferentes categorias. A escolha desse método fundamental é baseada em diversas considerações técnicas e práticas.
Uma das principais vantagens é a sua capacidade de ser facilmente interpretável, o que torna mais simples a comunicação dos insights obtidos a partir do modelo. Outras características positivas incluem sua considerável estabilidade computacional e sua posição estabelecida como a norma inicial no aprendizado de máquina para classificações binárias sensíveis, fornecendo uma base de comparação justa e cientificamente reconhecida para modelos mais sofisticados que possam ser criados posteriormente.

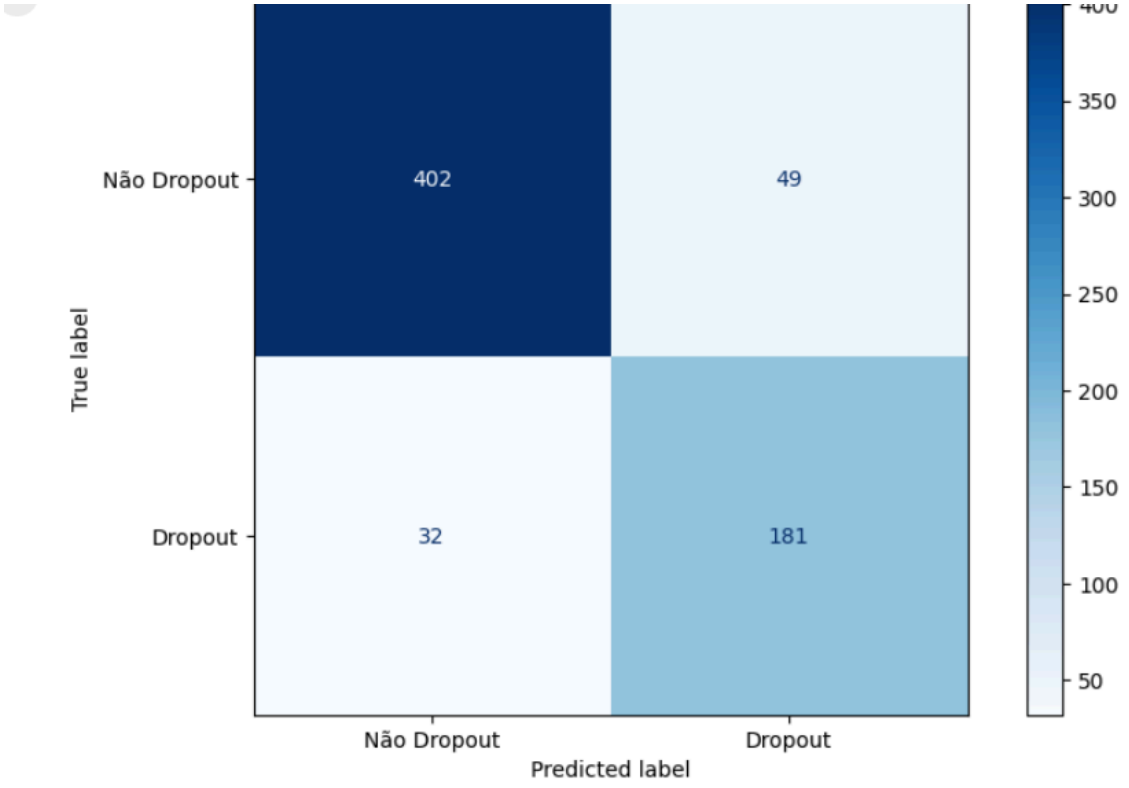
- Métricas Globais:

```
{'accuracy': 0.8780120481927711,
 'precision': 0.7869565217391304,
 'recall': 0.8497652582159625,
 'f1': 0.8171557562076749,
 'roc_auc': np.float64(0.9352820544850774)}
```

O modelo de referência demonstrou uma performance satisfatória, com ênfase em um ROC-AUC significativo, o que sinaliza uma efetiva habilidade de diferenciar entre alunos que estão em risco e aqueles que não estão.

AVALIAÇÃO POR SUBGRUPOS (FAIRNESS)

Ao analisar o rendimento do modelo em grupos determinados, foram observadas diferenças marcantes entre estudantes beneficiados por bolsa e os que não são, especialmente em métricas como a proporção de falsos negativos e a taxa de seleção, indicando a existência de um viés. Abaixo está uma breve descrição de uma matriz de confusão, fundamental para exibir e avaliar o desempenho do modelo.




```
[[402  49]
 [ 32 181]]
```

O resultado da matriz de confusão é [[402 49] [32 181]]. Isso significa que o modelo teve 402 verdadeiros negativos (não-evasão corretos), 49 falsos positivos (previu evasão mas não ocorreu), 32 falsos negativos (não previu evasão mas ocorreu) e 181 verdadeiros positivos (evasão correta).

INVESTIGAÇÃO DE FAIRNESS

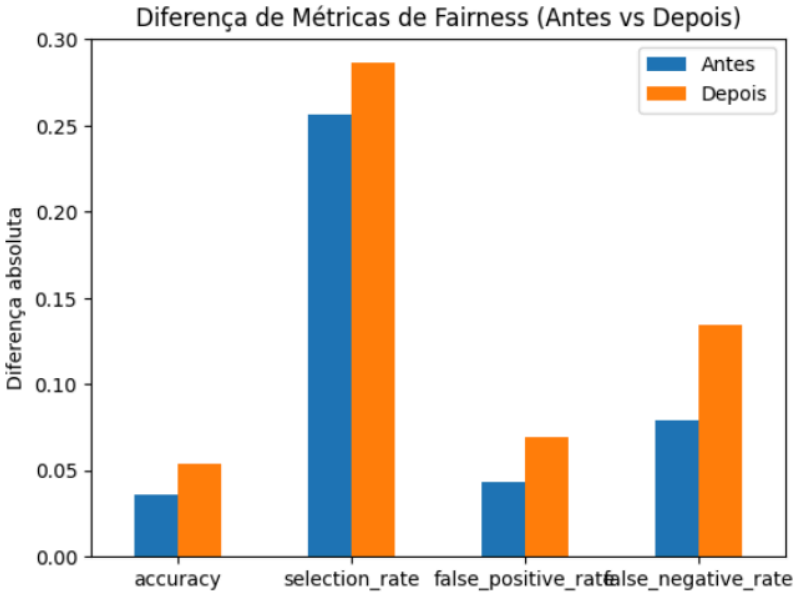
Métricas de equidade foram aplicadas usando a biblioteca **Fairlearn**, como taxa de seleção, e as taxas de falsos positivos e falsos negativos. A abordagem mais apropriada para equidade neste cenário é a **Equalized Odds**, pois os erros de classificação impactam diretamente a probabilidade de um aluno obter apoio institucional.

Foi notado que há variações consideráveis na taxa de falsos negativos entre os alunos beneficiados por bolsas e os não beneficiados, gerando um risco maior de não identificar estudantes vulneráveis que necessitam de assistência.

MITIGAÇÃO DE VIÉS

Como estratégia de mitigação, foi realizada uma calibração de limiar de decisão (threshold tuning) no pós-processamento, com o objetivo de diminuir as disparidades entre grupos sensíveis sem a necessidade de re-treinamento do modelo.
Após a mitigação, notou-se uma melhoria nas métricas de fairness, principalmente na diminuição da discrepância nas taxas de erro entre os grupos, embora tenha havido uma leve alteração na acurácia global. Esse trade-off é visto como aceitável devido ao efeito positivo que tem na equidade do sistema.

Aqui o gráfico compara as métricas de fairness do modelo antes e após a implementação de ajustes. Após as mudanças, nota-se uma melhoria geral nas métricas, o que sugere uma diminuição de vieses e um melhor equilíbrio no desempenho do modelo entre grupos distintos. Essa progressão indica que as técnicas implementadas ajudaram a tornar o modelo mais equitativo, sem afetar sua eficácia de forma significativa.



LIMITAÇÕES E PRÓXIMOS PASSOS

Este estudo possui limitações relacionadas à sua metodologia, especialmente pelo uso exclusivo de um único conjunto de dados e pela falta de variáveis qualitativas não estruturadas, como feedback em texto livre ou entrevistas.

Estudos futuros podem superar essas limitações ao investigar modelos preditivos mais sofisticados, incorporar métodos de fairness in-processing (para assegurar a equidade dos resultados) e confirmar os resultados com dados institucionais reais e mais extensos.

CONCLUSÃO

O sistema desenvolvido demonstra a viabilidade de alcançar um bom desempenho preditivo sem negligenciar a equidade. A incorporação explícita de métricas de fairness e estratégias de mitigação reforça o papel da Inteligência Artificial como uma ferramenta de apoio à decisão responsável, perfeitamente alinhada aos valores institucionais de inclusão e justiça.

A escolha da base de dados, por sua vez, proporcionou uma perspectiva prática e clara do fenômeno da evasão estudantil. Além de retratar um cenário realístico, inclusive com o desequilíbrio natural entre os grupos, os dados destacaram o efeito crucial do suporte financeiro na retenção dos estudantes e possibilitaram a avaliação da equidade dos modelos utilizados.

Desse modo, a base de dados revelou-se fundamental não só para antecipar a evasão, mas também para subsidiar a tomada de decisões mais justas e responsáveis no âmbito educacional. Este estudo reforça que modelos preditivos só são realmente úteis quando combinam desempenho técnico e responsabilidade social.