Modelo Preditivo de Evasão Escolar

Público: Aluno com Financiamento Estudantil

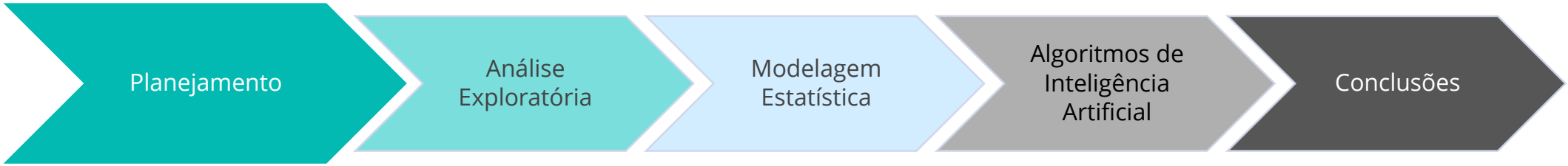
14/06/2021

Pós-Graduação de Análise de Dados, Data Mining e Inteligência Artificial

Nome do Aluno:
Rafael Santos Araújo

Coordenadores:
Profª Drª Alessandra de Álvila Montini
Profª Dr. Adolpho Walter Pimazoni Canton

Metodologia de análise de dados



Definição do problema

- Objetivos
- Conceitos
- Variáveis

Análise preliminar

- Medidas resumo
- Seleção de variáveis

Avaliação das técnicas

- Regressão Logística
- Árvore de Decisão

Avaliação das técnicas

- Random Forest
- Gradient Boost
- XGBoost
- LGBM
- Catboost

Definição da técnica

- Escolha e aplicação do modelo de melhor desempenho

1. Objetivo do Trabalho

O objetivo do trabalho é predizer a evasão ou conclusão escolar dos alunos do ensino superior que possuem financiamento estudantil, parcial ou integral, seja por programas sociais do governo, como o Fies ou Prouni, ou pela própria instituição de ensino superior.

A predição será realizada usando a base pública do Micro Censo Escolar da Educação Superior, fornecida pelo Inep (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira).

Desta forma, é possível identificar as características mais propensas à interrupção ou à conclusão do aluno no curso, possibilitando a análise e aplicação de políticas públicas preventivas para a redução da evasão escolar.

2. Contextualização do Problema

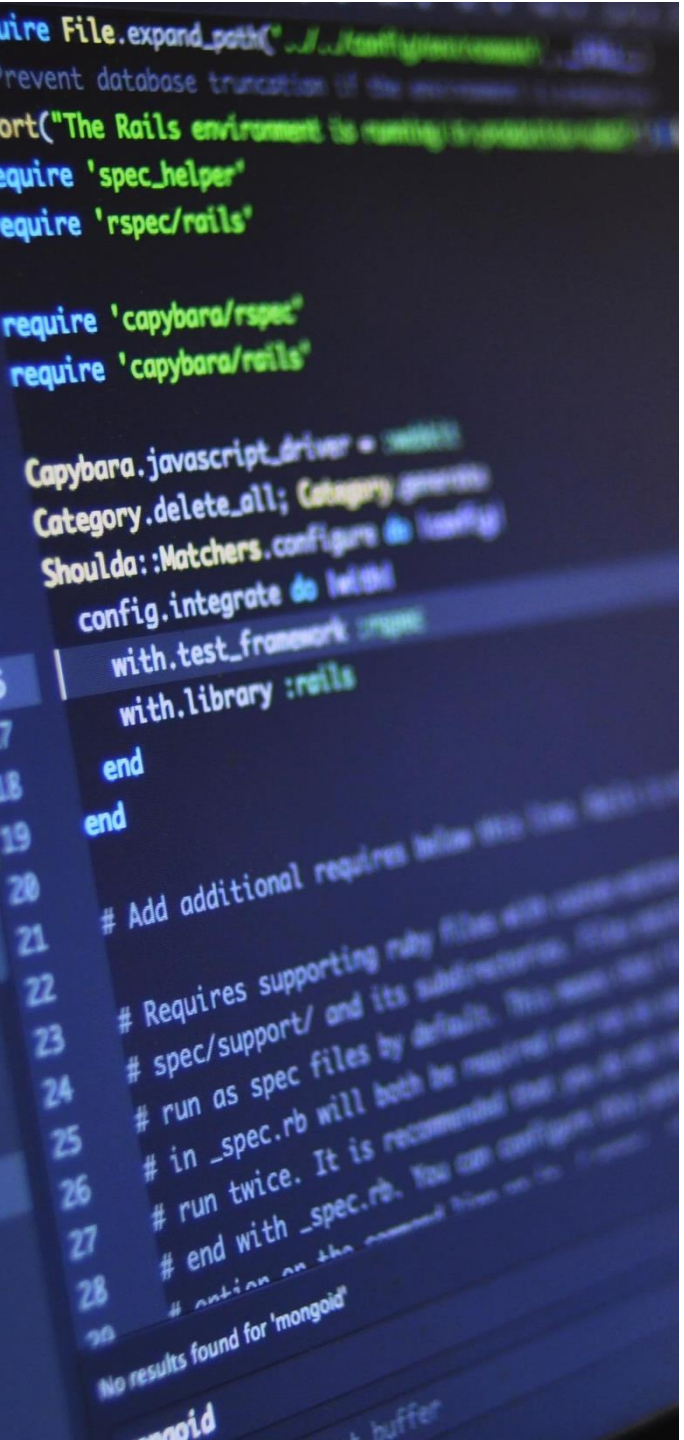
Censo Escolar

O Inep realiza pesquisas para fornecimento e divulgação de dados anonimizados sobre o ensino básico e superior.

Dentre estas pesquisas, temos o Censo da Educação Superior, realizada anualmente pela Diretoria de Estatísticas Educacionais.

Os microdados fornecidos são pertinentes aos alunos, docentes, cursos e instituições de ensino superior (IES).

Este estudo considera a extração dos microdados do Censo da Educação Superior de 2018, com o objetivo de predizer a propensão de evasão escolar do aluno matriculado em um curso do ano seguinte, 2019 (último ano com microdados disponíveis no momento deste projeto).



3. Bases de Dados

Visão da base

- Aluno (caso esteja matriculado em mais de um curso, considera aquele de data de ingresso mais antiga).

Filtros de inclusão

- Alunos que possuem financiamento estudantil
- Que sejam concluintes ou que tenham interrompido o curso.

Período de Análise

- Ano referência: 2018 (T0)
- Previsão: 2019 (T+1)

3.i. Base Original

A base original se encontra na pasta "**/microdados_ed_superior_2018/dados/DM_CURSO.CSV**"

Base auxiliar:
- TB_AUX_CINE_BRASIL.CSV

Link da página das bases anuais:
- https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior

Link de download 1 – Microdados Censo da Educação Superior 2018:
- https://download.inep.gov.br/microdados/microdados_educacao_superior_2018.zip

Link de download 2 – Microdados Censo da Educação Superior 2019:
- https://download.inep.gov.br/microdados/microdados_educacao_superior_2019.zip

3.ii. Filtros

Base de partida

O base bruta, em formato csv, ocupa 3GB.

Redução da base

A leitura da base com somente os alunos com financimento estudantil e terminaram ou interromperam o curso.

Redução da base

Considerando alunos distintos, mtriculados em um curso.

3.iii. Principais variáveis

Variáveis cadastrais

- **TP_CATEGORIA_ADMINISTRATIVA**: indica se a IES é pública ou privada.
- **TP_TURNO**: se o turno do curso é matutino, vespertino ou noturno.
- **TP_GRAU_ACADEMICO**: se o curso é bacharelado, tecnólogo, licenciatura, etc.
- **NU_ANO_INGRESSO**: ano de ingresso no curso.
- **QT_CARGA_HORARIA_TOTAL**: carga horária total do curso.
- **TP_SITUACAO**: motivo da conclusão ou interrupção do curso pelo aluno matriculado.
- **IN_FINANCIAMENTO_ESTUDANTIL:** alunos com financiamento estudantil**.**
- **Variáveis cadastrais do aluno**: cor, gênero, data de nascimento, se é deficiente, se possui bolsa de pesquisa, se é oriundo de escola de ensino médio pública ou privada.
- **CO_CINE_GERAL**: área de conhecimento geral e padronizado do curso.

Target

- **IN_CONCLUINTE**: se concluiu o curso ou não.

Arquivo de referência, com o dicionário da dados para todas as variáveis:
- ver o Anexo II

3.iv. Processo de redução de variáveis

Base Original

A base original contém os códigos com as características do aluno, IES e curso.

Variáveis novas

Foram criadas variáveis relativas para padronização de métricas, como Anos de estudo e Idade, em vez de data de ingresso e data de nascimento respectivamente.

Remoção de variáveis com missings

Remoção de colunas quase ou totalmente vazias, além de colunas categóricas com alta cardinalidade de valores.

Remoção de variáveis de correlação forte

Durante o estudo foram detectadas variáveis que causavam forte correlação com a variável resposta (target)

Base final.

Base final.

4. Análise Exploratória de Dados

Variáveis pertencentes a IES (Instituição de Ensino Superior)

TP_CATEGORIA_ADMINISTRATIVA	Frequência
4. Privada com fins lucrativos	0,761
5. Privada sem fins lucrativos	0,234
3. Pública Municipal	0,003
7. Especial	0,002

TP_ORGANIZACAO_ACADEMICA	Frequência
1. Universidade	0,435
3. Faculdade	0,294
2. Centro Universitário	0,270

O perfil das IES dos alunos que possuem financiamento:

- **Privada** com fins lucrativos
- Sendo em sua maior parte **Universidade**.

4.i. Análise Exploratória de Dados

Variáveis pertencentes ao Curso

TP_TURNO	Frequência
3. Noturno	0,722
1. Matutino	0,218
4. Integral	0,035
2. Vespertino	0,026

TP_GRAU_ACADEMICO	Frequência
1. Bacharelado	0,696
3. Tecnológico	0,170
2. Licenciatura	0,134

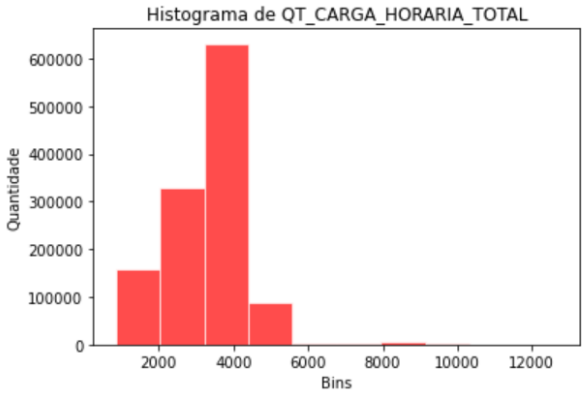
NO_CINE_AREA_GERAL	Frequência
Negócios, administração e direito	0,386
Saúde e bem-estar	0,192
Engenharia, produção e construção	0,140
Educação	0,134
Ciências sociais, jornalismo e informação	0,048
Computação e Tecnologias da Informação e Comunicação (TIC)	0,042
Serviços	0,023
Agricultura, silvicultura, pesca e veterinária	0,016
Artes e humanidades	0,015
Ciências naturais, matemática e estatística	0,004
Programas básicos	0,000

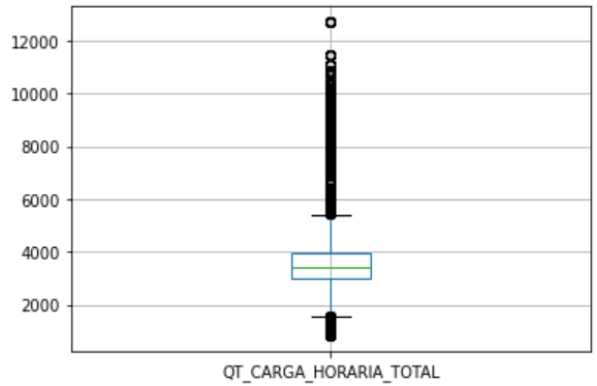
O perfil do Curso dos alunos que possuem financiamento:

- Período **Noturno**
- Curso do formação para **Bacharelado**
- **Presencial**
- Da área geral de Negócios, **administração** e direito

4.ii. Análise Exploratória de Dados

Variáveis pertencentes à Carga Horária do Curso





QT_CARGA_HORARIA_TOTAL	
Média	3.378,4
Desvio Padrão	945,7
Coeficiente de Variação	27,9
Mínimo	840
1º Quartil	3.000
Mediana	3.420
3º Quartil	3.971
Máximo	12.690

A distribuição da carga horária dos alunos que possuem financiamento:

- A maioria estava no princípio do curso.
- Pode-se presumir que, no geral, o aluno que **não concluiu** o curso **evade precocemente**.

4. iii. Análise Exploratória de Dados

Variáveis pertencentes ao Aluno

TP_SEXO	Frequência
1. Feminino	0,566
2. Masculino	0,434

TP_COR_RACA	Frequência
1. Branca	0,366
3. Parda	0,277
0, Aluno não quis declarar cor/raça	0,258
2. Preta	0,074
4. Amarela	0,018
5. Indígena	0,004
9. Não dispõe da informação (Não resposta)	0,004

TP_ESCOLA_CONCLUSAO_ENS_MEDIO	Frequência
1. Pública	0,733
2. Privada	0,260
9. Não dispõe da informação (Não resposta)	0,006

IN_DEFICIENCIA	Frequência
0, Não	0,960
9. Não dispõe de informação (Não resposta)	0,036
1. Sim	0,004

IN_APOIO_SOCIAL	Frequência
0, Não	0,897
1. Sim	0,103

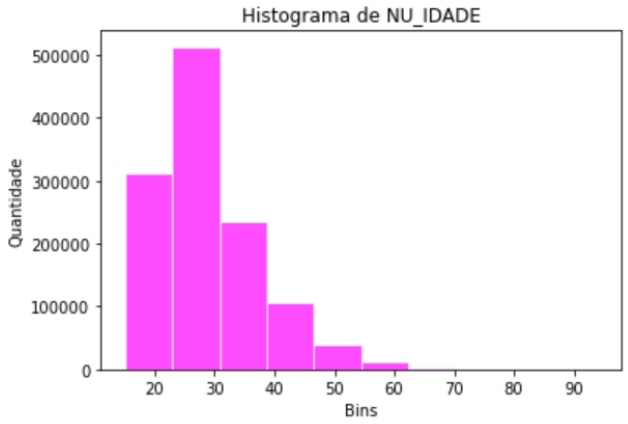
IN_ATIVIDADE_EXTRACURRICULAR	Frequência
0, Não	0,882
1. Sim	0,118

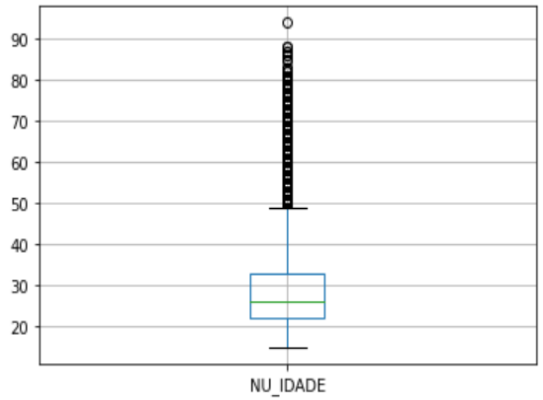
O perfil Aluno que possui financiamento:

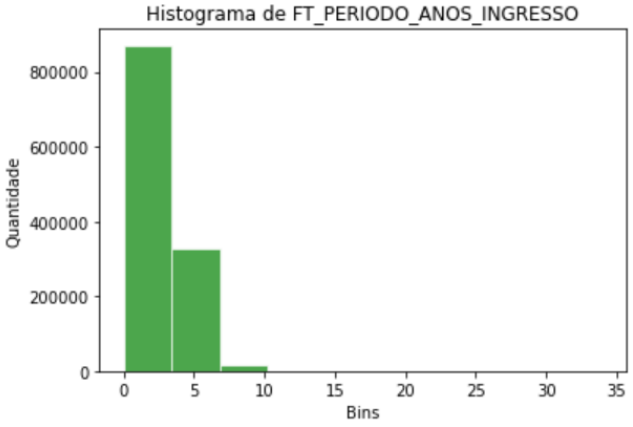
- Maioria são **mulheres**
- **Brancos ou Pardos**. Um quarto não declarou. Pretos e outros representam cerca de 10%
- Maioria advém do **Ensino Público**
- Deficientes representam 3%
- Que recebem algum tipo de apoio social na forma de moradia, transporte, alimentação, material didático e bolsas são 10%
- Que realizam Atividade Extracurricular (estágio não obrigatório, extensão, monitoria e pesquisa) são 11%.

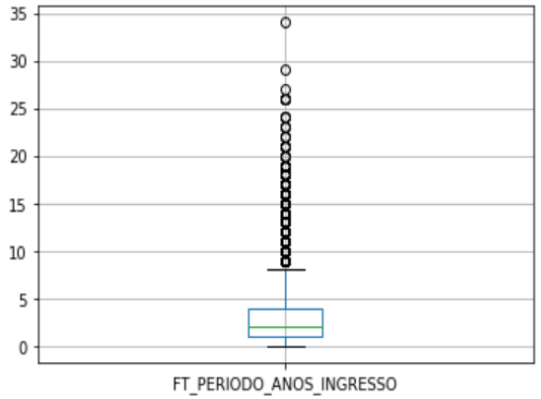
4.iii. Análise Exploratória de Dados

Variáveis pertencentes ao Aluno









NU_IDADE	
Média	28,6
Desvio Padrão	8,4
Coeficiente de Variação	29,3
Mínimo	15
1º Quartil	22
Mediana	26
3º Quartil	33
Máximo	94

FT_PERIODO_ANOS_INGRESSO	
Média	2,2
Desvio Padrão	1,9
Coeficiente de Variação	85,9
Mínimo	0
1º Quartil	1
Mediana	2
3º Quartil	4
Máximo	34

As distribuições indicam que:

- A maioria está na faixa dos **20-30 anos**. A distribuição é assimétrica à direita.
- Até 75% dos alunos cumpriram 4 anos de estudo desde o ingresso. A distribuição é assimétrica à direita.

4.iv. Análise Exploratória de Dados

Variáveis pertencentes ao Tipo de Ingresso do Aluno

IN_INGRESSO_VESTIBULAR	Frequência
1. Sim	0,726
0, Não	0,274

IN_INGRESSO_ENEM	Frequência
0, Não	0,857
1. Sim	0,143

IN_INGRESSO_SELECAO_SIMPLIFICA	Frequência
0, Não	0,950
1. Sim	0,050

IN_INGRESSO_VAGA_REMANESC	Frequência
0, Não	0,869
1. Sim	0,131

IN_INGRESSO_OUTRO	Frequência
0, Não	0,999
1. Sim	0,001

O Aluno ingressou no curso pela forma:

- Pelo **vestibular** da IES, representando a maioria.
- Cerca de 14% pelo **ENEM**
- Outras tipos de ingresso compõem minoria.

Vale observar que o aluno pode ingressar no curso por mais de um forma, por exemplo: pelo Vestibular e pelo ENEM concomitantemente.

4.v. Análise Exploratória de Dados

Variáveis pertencentes ao Tipo de Financiamento do Aluno

IN_FIN_REEMB_FIES	Frequência
0, Não (Aluno possui outro tipo de financiamento)	0,749
1. Sim	0,251

IN_FIN_REEMB_PROG_IES	Frequência
0, Não (Aluno possui outro tipo de financiamento)	0,949
1. Sim	0,051

IN_FIN_REEMB_OUTRO	Frequência
0, Não (Aluno possui outro tipo de financiamento)	0,988
1. Sim	0,012

IN_FIN_NAOREEMB_PROUNI_INTEGR	Frequência
0, Não (Aluno possui outro tipo de financiamento)	0,903
1. Sim	0,097

IN_FIN_NAOREEMB_PROUNI_PARCIAL	Frequência
0, Não (Aluno possui outro tipo de financiamento)	0,963
1. Sim	0,037

IN_FIN_NAOREEMB_PROG_IES	Frequência
0, Não (Aluno possui outro tipo de financiamento)	0,620
1. Sim	0,380

IN_FIN_NAO_REEMB_OUTRO	Frequência
0, Não (Aluno possui outro tipo de financiamento)	0,956
1. Sim	0,044

A forma de financiamento do Aluno é:

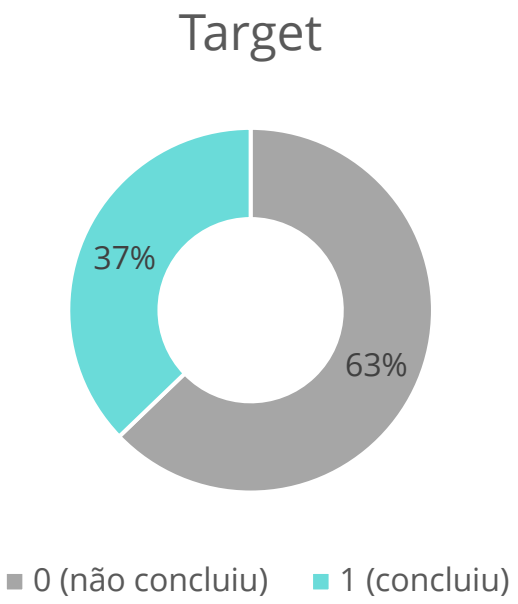
- **Financiamento** estudantil **não reembolsável** administrado pela IES (IN_FIN_NAOREEMB_PROG_IES)
- Seguido pelo **FIES** (IN_FIN_REEMB_FIES)
- Pelo **ProUni,** seja não reembolsável ou parcialmente reembolsável representa um pouco mais de 13%.

4.vi Análise Exploratória de Dados

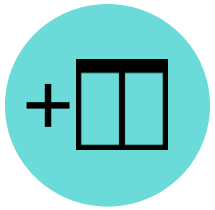
Dos 1.217.795 de alunos matriculados em um curso, 37% concluíram o curso.

Variável Resposta
IN_CONCLUINTE

Target:
1 = Concluiu
0 = Não concluiu

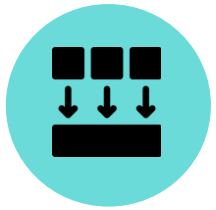


5. Modelagem com Estatística Tradicional



Novas Variáveis

Novas variáveis foram criadas para enriquecer a base. A variável da **Área geral do curso** e **Anos desde o Ingresso** foram criadas, a partir da mescla com a base de cursos com a base de alunos e do ano de ingresso respectivamente.

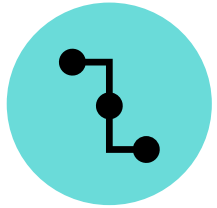


Agrupamento de Variáveis

As variáveis do **Tipo de Ingresso** e **Tipo de Financiamento** foram resumidas para as mais frequentes e os tipos restantes foram agrupados.

Preenchimento de Missings

As variáveis de **Turno** e **Grau Acadêmico** foram preenchidas com uma classificação genérica, para os casos de *missings*.



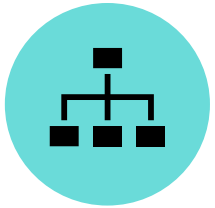
Correlações e Multicolinearidade

Foram analisadas correlações entre variáveis. Além de Information Value (IV) e de Variation Inflation Factor (VIF), para detecção de multicolinearidade.

Divisão Treino-Teste

A proporção da divisão da base em treino e teste foi de 70-30%.

6. Modelagem com Inteligência Artificial



One Hot Encoder

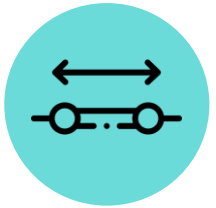
Para as variáveis de múltiplas categorias foi aplicada a técnica de *One Hot Encoder*.
.

Cross-validation

Para os algoritmos usados durante a avaliação, todos tiveram um teste unitário e de validação cruzada estratificada, seguindo a proporção da base: 70-30,

Grid Search

Uma seleção de hiperparâmetros foram testadas para a melhoria do desempenho do modelo escolhido.



Feature Scaling

Para os modelos de regressão logística e de machine learning, as variáveis foram escalonadas.

7. Conclusões

Resultados preliminares

Árvore de Decisão	Treino	Teste
Acurácia	0,999876	0,994862
Sensibilidade/Recall	0,999791	0,992775
Especificidade	0,999927	0,996117
ROC-AUC	0,999859	0,994446
Precisão	0,999878	0,993535
F1	0,999834	0,993155
R2	0,99947	0,978089

Regressão Logística	Treino	Teste
Acurácia	0,992112	0,992246
Sensibilidade/Recall	0,999891	0,99984
Especificidade	0,999927	0,996117
ROC-AUC	0,993664	0,99376
Precisão	0,979525	0,979914
F1	0,989603	0,989777
R2	0,966361	0,96693

- Os modelos estatísticos se ajustaram quase perfeitamente à base, tanto em treino como em teste.
- Houve *overfitting*.
- Assume-se que uma ou mais variáveis possam apresentar características de multicolinearidade.
- As variáveis do modelo foram reavaliadas para adoção de medidas com o objetivo de generalizar os modelos estatísticos.

7.i. Investigando o problema de Multicolinearidade

Passo 1

Correlação

A variável QT_CARGA_HORARIA_INTEG apresenta correlação forte com a variável resposta (target).

Quanto mais horas o aluno cumpre no curso, mais provável sua conclusão.

A variável foi descartada.

Passo 2

VIF

Identificou a variável IN_INGRESSO_VESTIBULAR como possível sinal de multicolinearidade.

Mas ela foi mantida, para efeitos do tema de estudo (decisão de negócio).

Passo 3

Information Value

Indicou que o período desde o Ingresso é suspeita ou boa demais (FT_PERIODO_ANOS_INGRESSO).

Faz sentido, pois quanto mais tempo no curso, maior a chance para concluí-lo

A variável foi mantida, para efeitos do tema de estudo (decisão de negócio).

7.ii. Conclusões

Resultados gerais

Árvore de Decisão	Treino	Teste
Acurácia	0,95006	0,77162
Sensibilidade/Recall	0,90252	0,67369
Especificidade	0,97921	0,83168
ROC-AUC	0,94087	0,75268
Precisão	0,96380	0,71050
F1	0,93215	0,69160
R2	0,78806	0,03077

Regressão Logística	Treino	Teste
Acurácia	0,72491	0,72505
Sensibilidade/Recall	0,55888	0,55982
Especificidade	0,82672	0,82637
ROC-AUC	0,69280	0,69309
Precisão	0,66418	0,66409
F1	0,60700	0,60751
R2	-0,16747	-0,16689

- Com as técnicas para detecção de multicolinearidade, foi possível generalizar um pouco mais os modelos.

7.iii. Conclusões

Resultados gerais

Random Forest	Treino	Teste
Acurácia	0,95004	0,78796
Sensibilidade/Recall	0,92412	0,71189
Especificidade	0,96594	0,83461
ROC-AUC	0,94503	0,77325
Precisão	0,94331	0,72523
F1	0,93361	0,71850
R2	0,78799	0,10012

XGBoost	Treino	Teste
Acurácia	0,79980	0,79640
Sensibilidade/Recall	0,72506	0,72257
Especificidade	0,84563	0,84168
ROC-AUC	0,78535	0,78212
Precisão	0,74228	0,73675
F1	0,73357	0,72959
R2	0,15037	0,13594

CatBoost	Treino	Teste
Acurácia	0,81079	0,80510
Sensibilidade/Recall	0,74162	0,73624
Especificidade	0,85322	0,84732
ROC-AUC	0,79742	0,79178
Precisão	0,75598	0,74727
F1	0,74873	0,74172
R2	0,19701	0,17283

Gradient Boost	Treino	Teste
Acurácia	0,75255	0,75220
Sensibilidade/Recall	0,63907	0,64018
Especificidade	0,82214	0,82089
ROC-AUC	0,73061	0,73054
Precisão	0,68782	0,68669
F1	0,66255	0,66262
R2	-0,05017	-0,05165

LGBM	Treino	Teste
Acurácia	0,77998	0,77870
Sensibilidade/Recall	0,69512	0,69509
Especificidade	0,83201	0,82997
ROC-AUC	0,76357	0,76253
Precisão	0,71730	0,71484
F1	0,70604	0,70483
R2	0,06622	0,06081

- O modelo **Catboost** desempenhou melhor nas métricas gerais.
- A técnica de *cross-validation* obteve resultados semelhantes.
- A técnica de *Grid Search* não melhorou significativamente.

7.iv. Variáveis Mais Importantes

	Feature	Descrição	Importância
1.	QT_CARGA_HORARIA TOTAL	Carga Horária Total do Curso	33,0
2.	NU_IDADE	Idade do aluno	20,6
3.	IN_FIN_REEMB_FIES	Financiamento do FIES	5,6
4.	TP_TURNO_0,0	EAD	4,2
5.	IN_APOIO_SOCIAL	Apoio Social	3,8
6.	IN_FIN_NAOREEMB_PROG_IES	Financiamento Não Reemboslável da IES	3,0
7.	IN_ATIVIDADE_EXTRACURRICULAR	Atividade Extracurricular	2,8
8.	CO_CINE_AREA_GERAL_9	Área de Saúde e bem-estar	2,2
9.	TP_CATEGORIA_ADMINISTRATIVA_4	IES Privada com fins lucrativos	1,9
10,	IN_FIN_REEMB_PROG_IES	Financiamento Reemboslável da IES	1,5
11.	TP_GRAU_ACADEMICO_1.0	Grau de Bacharelado	1,5
12.	IN_FIN_NAOREEMB_PROUNI_INTEGR	Financiamento do Prouni	1,4

- A **carga horária** do curso e a **idade** do aluno são fatores determinantes para a conclusão do curso.
- As demais variáveis tem pesos menores, mas dão indícios de características que influenciam o resultado do modelo.

7.v Conclusões

Aplicação na base T+1 (2019)

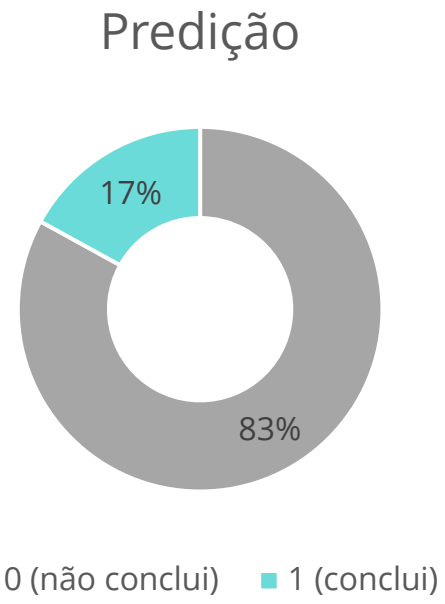
Base T+1 (Alunos sem matrículas ativas)

O modelo Catboost foi aplicado para os alunos concluintes ou não. Em uma base com 1.270.536 registros.

CatBoost (Modelo Escolhido)	Treino
Acurácia	0,78079
Sensibilidade/Recall	0,61881
Especificidade	0,86063
ROC-AUC	0,73972
Precisão	0,68638
F1	0,65084
R2	0,00881

Base T+1 (Alunos cursando) Sem resposta

O modelo Catboost foi aplicado para os alunos ingressantes com até dois anos de curso. Em uma base com 1.647.901 registros.



8. Sugestões para Trabalhos Futuros

Análise do perfil dos alunos

Com a identificação das principais variáveis que determinam a evasão escolar do ensino superior, podemos cortar a base para traçar perfis dos alunos que concluem ou interrompem seu curso.

Entendendo as variáveis e o perfil, tem-se um panorama para a tomada de decisão para incentivo de políticas públicas ou privadas, direcionado de acordo com a necessidade de cada aluno.

Anexos

Anexo I
- Base_amostra.xslx

Anexo II
- Cópia Dicionário de Variáveis.xls

Anexo III
- Notebook com o código do projeto.