

Inteligência Artificial II

Raísson Souza

Previsão de aprovação de alunos nas cadeiras do quarto semestre de Sistemas de Informação

Dataset pertencente ao projeto “AppAMF” e devidamente autorizado para uso exclusivamente acadêmico nas disciplinas de Inteligência Artificial II e Ciência de Dados.




Sumário

1. Problema;
2. Dados;
3. Pipeline;
4. Modelos;
5. Resultados;
6. Interpretação;
7. Limitações;
8. Próximos passos;

1. Problema

Possuindo um dataset completo de registros anonimizados de disciplinas cursadas por alunos do curso de Sistemas de Informação, como usar inteligência artificial para prever a aprovação dos mesmos em disciplinas do 4º semestre?

- **Banco de Dados Aplicado;**
 - **Engenharia de Software;**
 - **Liderança nas Profissões Tecnológicas;**
 - **Língua Inglesa IV;**
 - **Programação Orientada a Objetos;**
 - **Redes de Computadores;**
- 

2. Dados

Colunas originais:

- id;
- student_id;
- birthdate;
- sex;
- city;
- course;
- period;
- week_day;
- discipline;
- status;
- g1;
- g2;
- final_grade;
- class_skips;



Após tratamento:

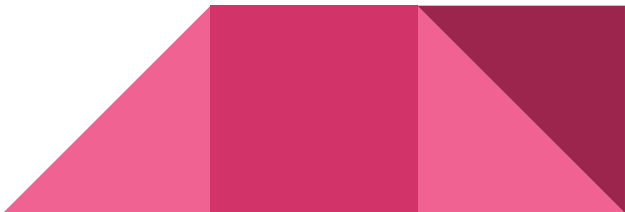
- id;
- student_id;
- discipline;
- status;
- final_grade;



	id	student_id	discipline \
133	248	100004	Gestão da Qualidade de Software
111	249	100004	Laboratório de Algoritmo II
134	247	100004	Organização e Arquitetura de Computadores
653	246	100004	Outro
1918	250	100004	Programação para Dispositivos Móveis
2274	251	100004	Análise e Projeto Orientados a Objetos
2251	253	100004	Desenvolvimento de Sistemas I
2252	252	100004	Gerenciamento de Serviços de TI
2171	254	100004	Laboratório de Inovação I
1205	608	100040	Engenharia de Software
1774	609	100040	Programação para Dispositivos Móveis
3141	610	100040	Análise e Projeto Orientados a Objetos
3027	611	100040	Personalidade e Carreiras Tecnológicas
4808	612	100040	Liderança nas Profissões Tecnológicas
1385	741	100053	Engenharia de Software
1034	743	100053	Gestão de Sistemas de Informação
1035	742	100053	Outro
7850	1049	100094	Desenvolvimento de Sistemas Integrado
7851	1050	100094	Ontopsicologia Aplicada a Sistemas de Informaç...
7849	1048	100094	Outro

	status	final_grade
...		
1035	0	0.50
7850	1	1.00
7851	1	0.97
7849	1	0.90

3. Pipeline

1. Carregamento inicial do dataset;
 2. Ordenação do dataset por id de aluno, semestre e disciplina;
 3. Remoção de colunas não usadas;
 4. Filtragem pelo curso de Sistemas de Informação;
 5. Normalização da coluna status de disciplina;
 6. Normalização e tratamento da coluna de nota final;
 7. Tratamento da coluna de nome de disciplina;
 8. Remoção da coluna do curso referente;
- 

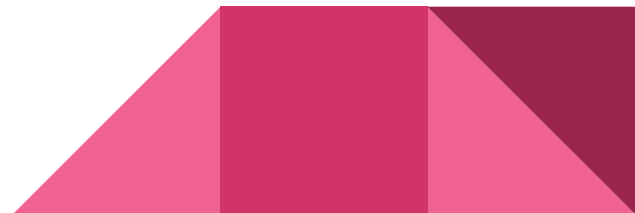
Estatísticas preliminares do dataset:

Total de registros: **4604**

Total de disciplinas: **75**

Número de aprovações: **3011**

Número de reprovações: **1593**



4. Modelos

```
semester4_disciplines = [  
    "Banco de Dados Aplicado",  
    "Engenharia de Software",  
    "Liderança nas Profissões Tecnológicas",  
    "Língua Inglesa IV",  
    "Programação Orientada a Objetos",  
    "Redes de Computadores",  
]  
  
# Criar colunas de target (aprovado em cada disciplina do 4º semestre)  
for disc in semester4_disciplines:  
    df_random_forest[disc + '_target'] = (  
        (df_random_forest['discipline'] == disc) & (df_random_forest['status'] == 1)  
    ).astype(int)  
  
# Features: pivot das notas  
features_df = df_random_forest.pivot_table(  
    index='student_id',  
    columns='discipline',  
    values='final_grade',  
    aggfunc='mean'  
).fillna(0)  
  
# Targets: um target por disciplina do 4º semestre  
targets_df = df_random_forest.groupby('student_id')[  
    [d + '_target' for d in semester4_disciplines]  
].max()
```

Random Forest

```
# Modelo RandomForest
model = RandomForestClassifier(
    n_estimators=1000,
    max_depth=None,
    min_samples_split=2,
    min_samples_leaf=1,
    bootstrap=True,
    class_weight="balanced",
    random_state=42,
    n_jobs=-1
)
```

Gradient Boosting

```
# Modelo Gradient Boosting
model = GradientBoostingClassifier(
    n_estimators=500,
    learning_rate=0.05,
    max_depth=4,
    min_samples_split=5,
    min_samples_leaf=3,
    random_state=42
)
```

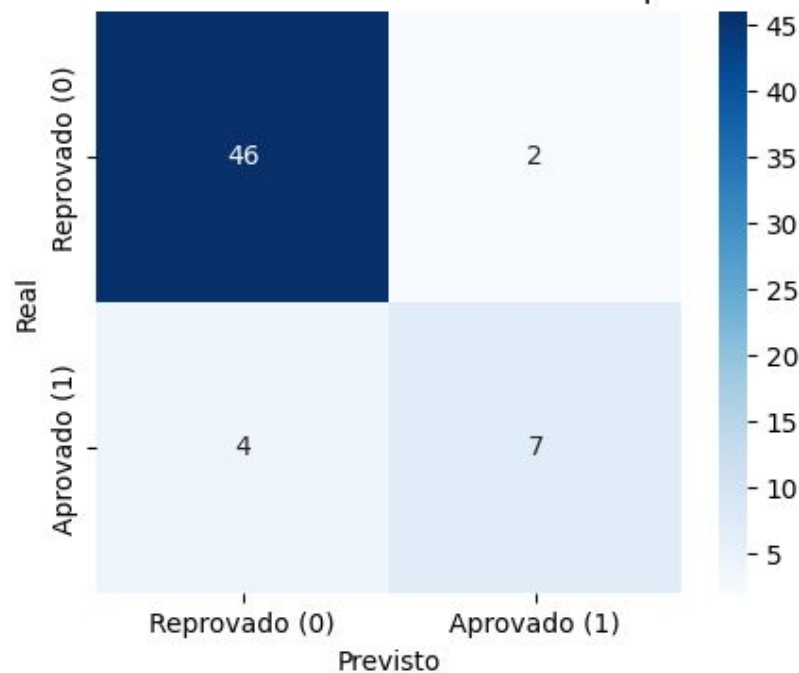
5. Interpretação

≡ RandomForest para Banco de Dados Aplicado ≡

Acurácia: 0.8983050847457628

	precision	recall	f1-score	support
0	0.92	0.96	0.94	48
1	0.78	0.64	0.70	11
accuracy			0.90	59
macro avg	0.85	0.80	0.82	59
weighted avg	0.89	0.90	0.89	59

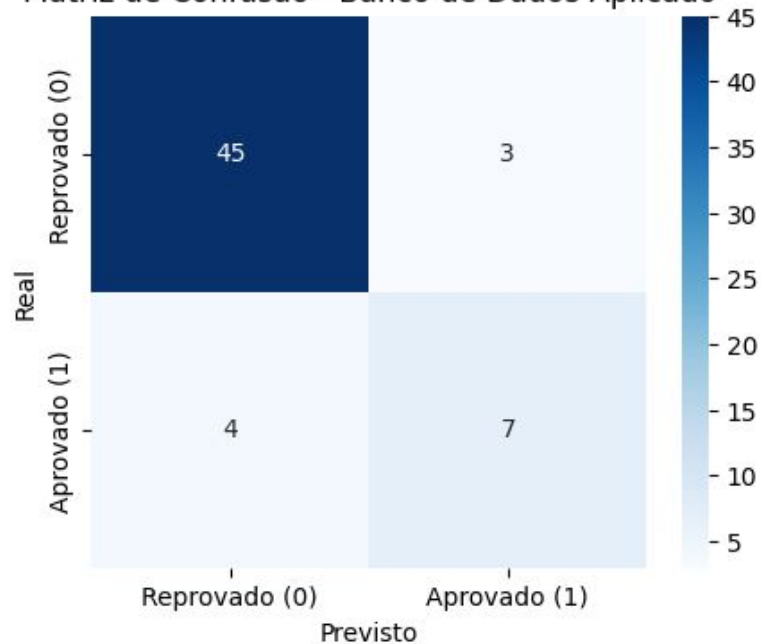
Matriz de Confusão - Banco de Dados Aplicado



≡≡≡ GradientBoosting para Banco de Dados Aplicado ≡≡≡
Acurácia: 0.8813559322033898

	precision	recall	f1-score	support
0	0.92	0.94	0.93	48
1	0.70	0.64	0.67	11
accuracy			0.88	59
macro avg	0.81	0.79	0.80	59
weighted avg	0.88	0.88	0.88	59

Matriz de Confusão - Banco de Dados Aplicado

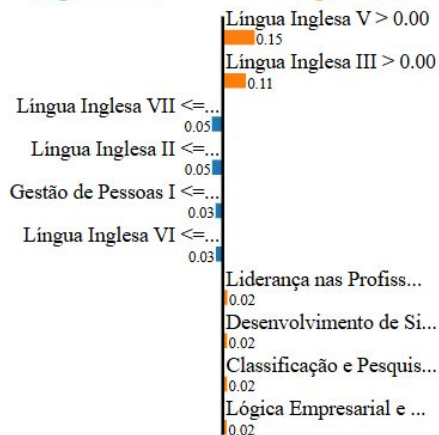


Prediction probabilities



Reprovado

Aprovado

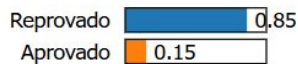


Feature

Value

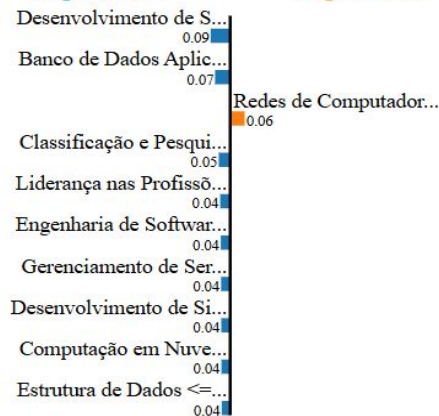
Lingua Inglesa V	0.90
Lingua Inglesa III	0.90
Lingua Inglesa VII	0.00
Lingua Inglesa II	0.00
Gestão de Pessoas I	0.00
Lingua Inglesa VI	0.00
Liderança nas Profissões Tecnológicas	0.92
Desenvolvimento de Sistemas I	0.92
Classificação e Pesquisa de Dados	0.89
Lógica Empresarial e Performance Profissional	0.92

Prediction probabilities



Reprovado

Aprovado



Feature

Value

Desenvolvimento de Sistemas I	0.00
Banco de Dados Aplicado	0.00
Redes de Computadores	0.86
Classificação e Pesquisa de Dados	0.00
Liderança nas Profissões Tecnológicas	0.00
Engenharia de Software	0.00
Gerenciamento de Serviços de TI	0.00
Desenvolvimento de Sistemas II	0.00
Computação em Nuvem	0.00
Estrutura de Dados	0.00

6. Limitações

- Dados da faculdade;
- Desbalanceamento geral entre aprovados x desaprovados (mil de diferença);
- Necessidade de tratamento nos nomes das disciplinas (crédito de disciplina);
- Presença de disciplinas muito pouco cursadas (inutilidade);



7. Próximos Passos

- Realizar outras análises, classificações, regressões ou agrupamentos conforme necessidade real da coordenadoria / diretoria;
- Aplicação em sistema real para tomada de decisão na instituição;



Previsão de aprovação de alunos nas cadeiras do quarto semestre de Sistemas de Informação

Quantidade de registros: 35995

	id	student_id	birthdate	sex	city	\
2138	1	99951	1992-05-21 02:00:00.000 -0300	F	Faxinal do Soturno	
5665	3	99951	1992-05-21 02:00:00.000 -0300	F	Faxinal do Soturno	
9160	2	99951	1992-05-21 02:00:00.000 -0300	F	Faxinal do Soturno	
28710	4	99955	1989-01-06 04:00:00.000 -0200	M	Faxinal do Soturno	
28312	8	99955	1989-01-06 04:00:00.000 -0200	M	Faxinal do Soturno	

	course	period	week_day	\
2138	Ciências Contábeis	2021/1	Quarta	
5665	Ciências Contábeis	2021/2	Segunda	
9160	Ciências Contábeis	2022/1	Terça	
28710	Ciências Contábeis	2025/1	Segunda	
28312	Ciências Contábeis	2025/1	Quarta	

	discipline	status	g1	g2	final_grade	\
2138	Matemática Aplicada - 60	Aprovado	7.5	7.0	7.2	
5665	Contabilidade Intermediária - 60	Aprovado	7.2	8.9	8.0	
9160	Contabilidade Avançada - 60	Trancado	-1.0	-1.0	-1.0	
28710	Contabilidade Introdutória - 60	Aprovado	8.4	8.8	8.6	
28312	Legislação e Ética Profissional - 30	Aprovado	9.8	9.9	9.8	

	class_skips
2138	0
5665	0
9160	0
28710	0
28312	0