ANTONIO MENEGHETTI FACULDADE

INTELIGÊNCIA ARTIFICIAL II

PREVISÃO DE APROVAÇÃO NAS CADEIRAS DO QUARTO SEMESTRE PARA ALUNOS DE SISTEMAS COM INTELIGÊNCIA ARTIFICIAL

Raisson Silveira de Souza

2025/2

23 de setembro de 2025

1. PROBLEMA E OBJETIVO

Atualmente o responsável pelo trabalho atua como desenvolvedor no Tecnoamf, time de desenvolvimento do curso de Sistemas de Informação da Antonio Meneghetti Faculdade. Está trabalhando no projeto "AppAMF", a carteira digital do aluno AMF, neste projeto, o mesmo possui acesso aos dados acadêmicos dos alunos, dados estes como informações de cadeiras, tal qual nomes, períodos letivos, notas e etc, ao visualizar estes dados, percebeu-se a possibilidade da utilização dos mesmos para a aplicação de conceitos de inteligência artificial sendo passados na cadeira de Inteligência Artificial II e Ciência de Dados em minha graduação.

Com estes dados, é possível realizar diversas análises e predições de informações acadêmicas que podem vir a ser úteis à coordenação dos cursos e a diretoria da instituição, informações como a probabilidade de aprovação dos alunos em determinadas cadeiras com base em suas notas gerais, este será o objetivo a ser abordado neste trabalho.

Utilizando-se de um pré-processamento do dataset extraído, removendo informações não úteis para essa análise e predição, é possível utilizar algoritmos de classificação como Random Forest para a predição da aprovação de determinadas cadeiras de cursos para os alunos matriculados, visando compreender e prever o número de alunos que irão avançar para o semestre seguinte.

2. DADOS

O dataset dos alunos da Antonio Meneghetti Faculdade foi obtido após a devida autorização da coordenadoria e direção da instituição, os dados foram extraídos de uma base de dados Postgres populada em ambiente de desenvolvimento do projeto "AppAMF", na qual foi anonimizada durante a extração na execução da query SQL, não capturando colunas de dados pessoais, como CPF, ou dados identificáveis, como RA dos alunos.

Um arquivo .csv (planilha) foi gerado com base na execução da query anteriormente descrita, o dataset gerado contém as seguintes colunas:

- id (ID do registro aluno e disciplina);
- student_id (ID do aluno na base de dados);
- birthdate (data de nascimento do aluno);
- sex (sexo biológico do aluno);
- city (cidade natal do aluno ;
- course (curso do aluno);
- period (período letivo da disciplina referente);
- week_day (dia da semana da disciplina referente);
- discipline (nome da disciplina);
- status (status de aprovação da disciplina);
- g1 (nota da G1 da disciplina);
- g2 (nota da G2 da disciplina);
- final_grade (nota final da disciplina);
- class_skips (número de faltas do aluno na disciplina).

O dataset contém quase 36 mil registros, sendo estes quase 5 mil para o curso de Sistemas de Informação. Foi realizado um tratamento e normalização dos dados do mesmo,

as colunas de aniversário, sexo, cidade natal, dia da semana, período letivo, g1, g2 e faltas foram removidas pois não seriam utilizadas na análise, apenas os registros na qual a coluna referente ao curso fossem de Sistemas de Informação foram filtrados, uma normalização da coluna de nota final foi realizada, assim como o tratamento de notas errôneas, o nome das disciplinas também foi tratado, removendo o número de créditos da mesma, presente em seu nome, as disciplinas realizadas por menos de 10 alunos no total tiveram seus nomes alterados para "Outro", para não interferirem no modelo, e, por fim, a coluna do curso referente foi removida para então ser inicializada a análise.

Setenta e cinco disciplinas foram catalogadas e filtradas entre os registros do curso de Sistemas de Informação como sendo realizadas por alunos deste curso, sendo as mais presentes:

1. Estrutura de Dados (201);
2. Organização e Arquitetura de Computadores (193);
3. Competência Competitiva no Universo da Tecnologia (178);
4. Probabilidade e Estatística (165);
5. Laboratório de Algoritmos II (151).

Verifica-se que a possibilidade dessas maiores presenças se dá por serem as cadeiras iniciais do curso ou pelo alto índice de reprovação, o que acarreta no maior número de matriculados, cabe análise posterior mais aprofundada.

A segurança e integridade do dataset foi garantida por compartilhamento interno do arquivo entre dispositivos pessoais do próprio aluno, assim como a não publicação do mesmo no repositório final do atual trabalho, visando evitar ao máximo qualquer vazamento ou exposição dos dados, mesmo sendo difícil a identificação dos alunos com base nos dados do dataset.

Os algoritmos de Random Forest e Gradient Boosting foram utilizados neste trabalho para a previsão de aprovação dos alunos nas cadeiras do quarto semestre utilizando-se de uma porcentagem de 80% dos registros para treino e 20% para teste, assim como foram criados "targets" com base na coluna binária "disciplina_target" que tem como valor "1" se o aluno foi aprovado naquela disciplina e "0" caso contrário, também, para cada aluno, "targets_df" pega o máximo por aluno, ou seja, se ele foi aprovado naquela cadeira em alguma matrícula (ou tentativa), fica "1".

3. METODOLOGIA

Para a realização da análise deste dataset foi inicializado um repositório com um arquivo .ipynb (notebook python) para a melhor divisão e organização do código, assim como obter uma melhor análise dos resultados, os seguintes passos foram realizados para a construção deste trabalho:

1. Carregamento inicial do dataset, ordenando os registros por "student_id", "period" (período letivo) e "discipline", visando agrupar os dados para uma melhor visualização;
2. Tratamento do dataset removendo colunas não utilizadas, removendo registros de alunos de outros cursos, normalizando a coluna "status" como binária, normalizando a coluna "final_grade" como valor decimal e tratando valores negativos (como "-1"),

removendo os créditos do nome das disciplinas (ex: “Matemática Aplicada - 60” torna-se “Matemática Aplicada”), trocando o nome das disciplinas cursadas no total por menos de 10 alunos para “Outro”, removendo então a coluna “course” após devida filtragem para o curso de Sistemas de Informação. E por fim, foi criado um novo arquivo .csv apenas com os dados finais de análise, tendo como número total de registros 4,6 mil, em comparação com os 35 mil originais de toda a instituição;

3. Para fins de análise preliminar da classificação a ser realizada, foram extraídas informações sobre as disciplinas presentes no dataset, além da geração de um novo arquivo .csv, contendo cerca de 75 disciplinas, como abordado anteriormente na seção 2;
4. No final do notebook são utilizados os algoritmos de Random Forest e Gradient Boosting, ambos úteis para esse problema, pois Random Forest se utiliza de várias árvores de decisão que tentam encontrar padrões entre as notas anteriores (features) e o status de aprovação (target), que por fim tem como resultado o “voto” da maioria das árvores, Random Forest oferece uma maior robustez contra overfitting. Gradient Boosting tem um funcionamento similar, mas tem como foco onde o modelo mais erra, portanto tem maior precisão em dados tabulares e melhor performance com menos dados.

As disciplinas Banco de Dados Aplicado, Engenharia de Software, Liderança nas Profissões Tecnológicas, Língua Inglesa IV, Programação Orientada a Objetos, Redes de Computadores, todas do quarto semestre, terão seus status previstos.

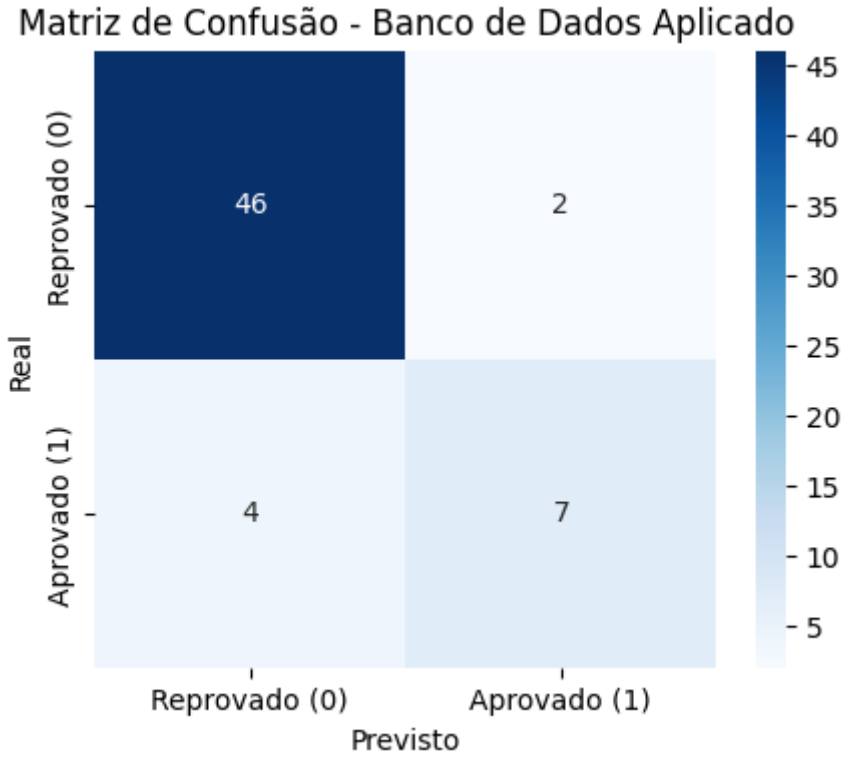
4. EXPERIMENTOS E RESULTADOS

Como resultados obtidos com esse trabalho, informações como acurácia, precisão, recall, f1-score, support e matriz de confusão foram gerados para cada uma das seis disciplinas.

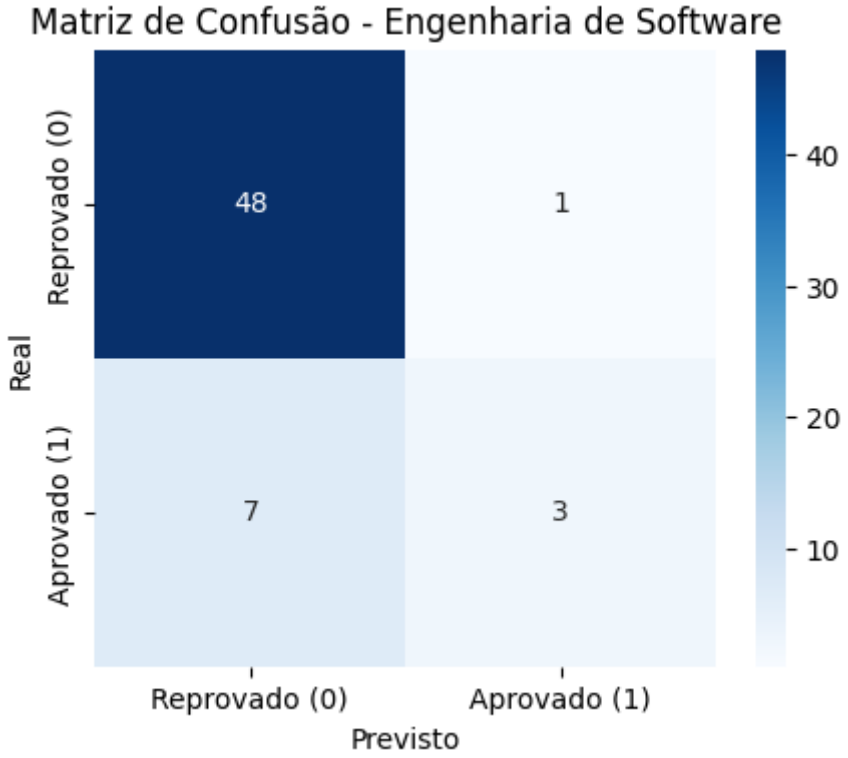
4.1 RANDOM FOREST

Os seguintes resultados foram obtidos para as disciplinas utilizando-se do algoritmo Random Forest:

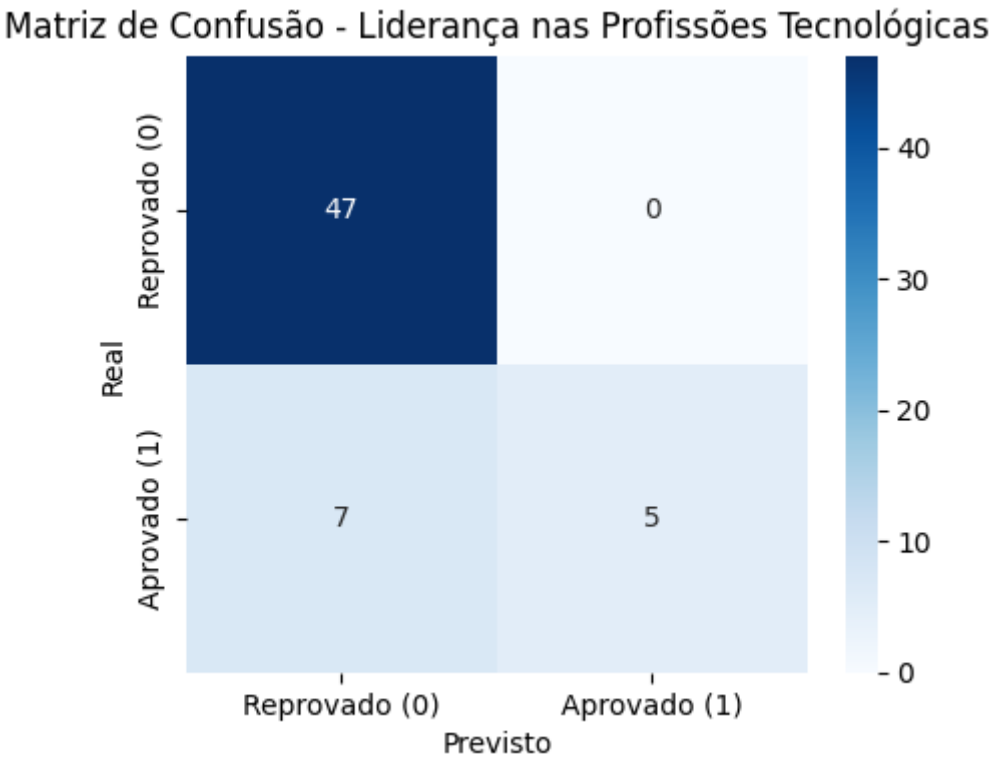
- Banco de Dados Aplicado: Acurácia 0.89



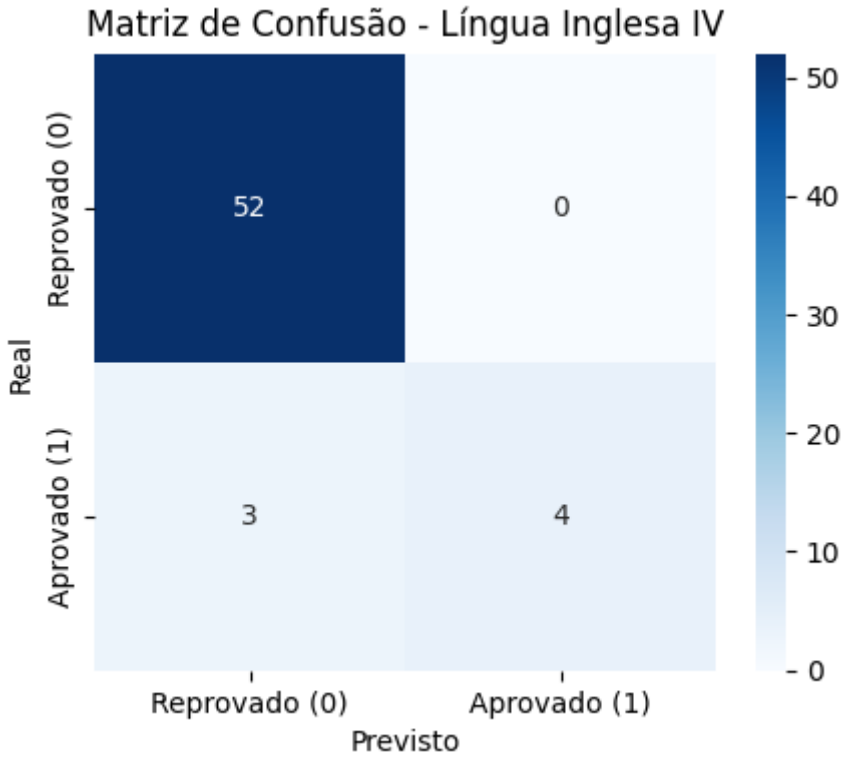
- Engenharia de Software: Acurácia 0.86



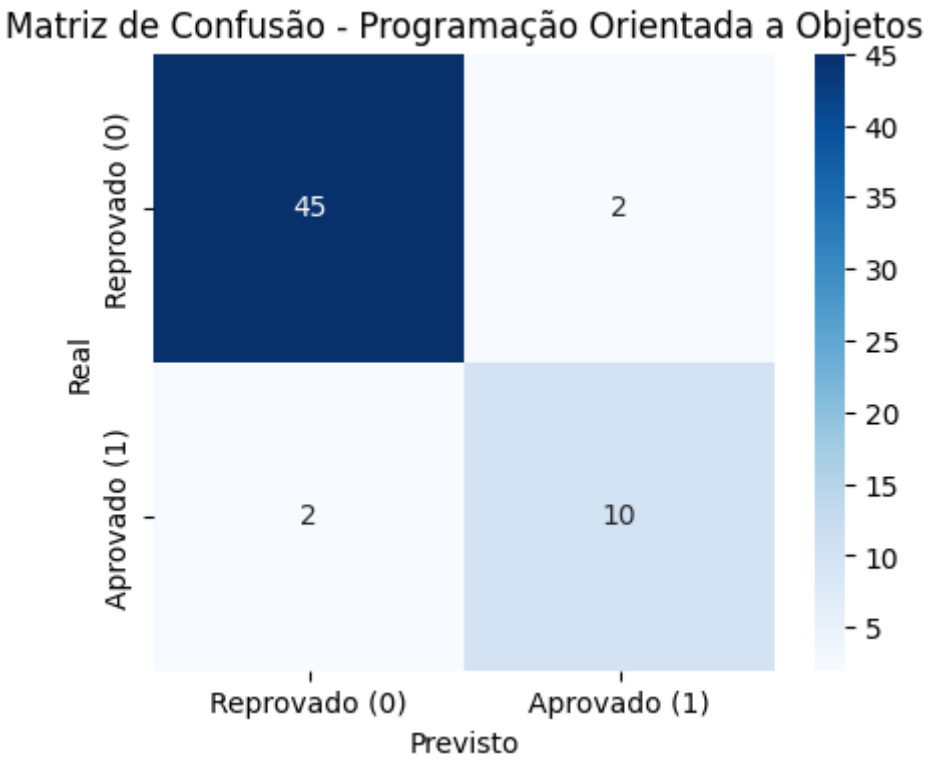
- Liderança nas Profissões Tecnológicas: Acurácia 0.88



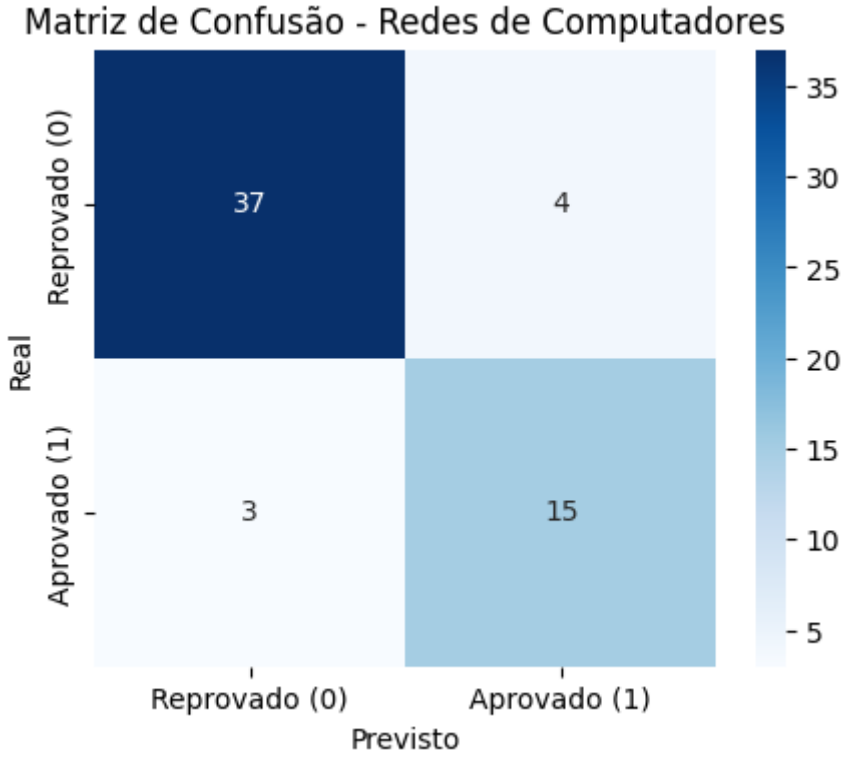
- Língua Inglesa IV: Acurácia 0.94



- Programação Orientada a Objetos: Acurácia 0.93



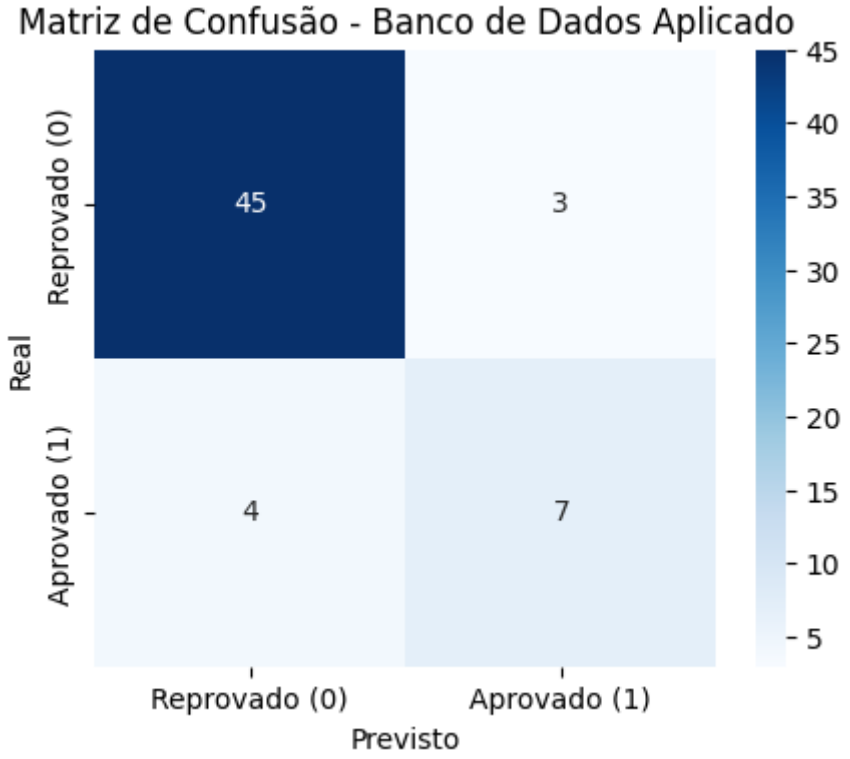
- Redes de Computadores: Acurácia 0.88



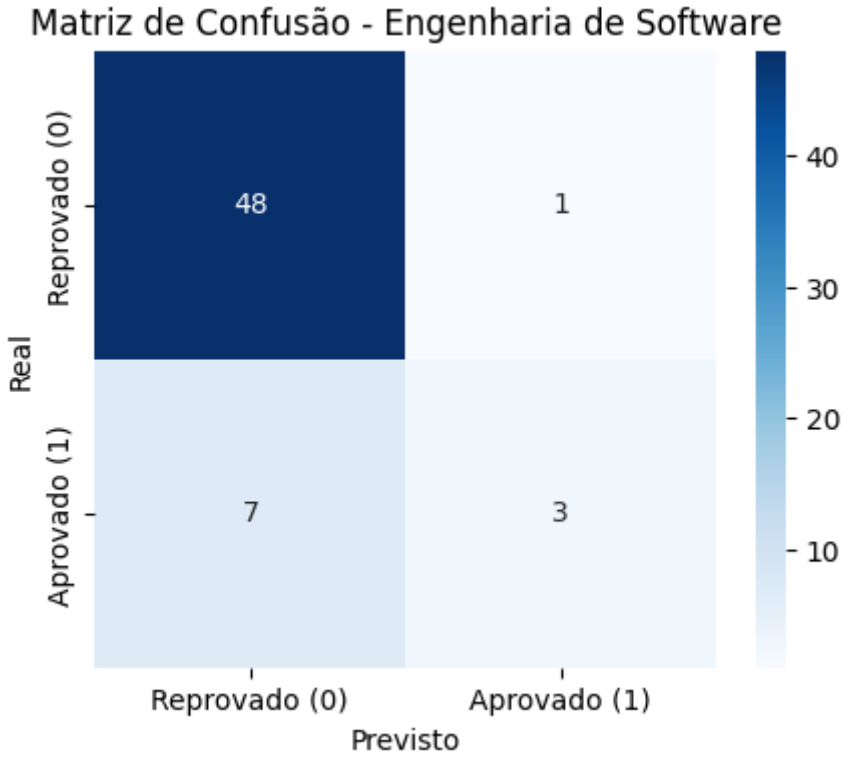
4.2 GRADIENT BOOSTING

Os seguintes resultados foram obtidos para as disciplinas utilizando-se do algoritmo Gradient Boosting:

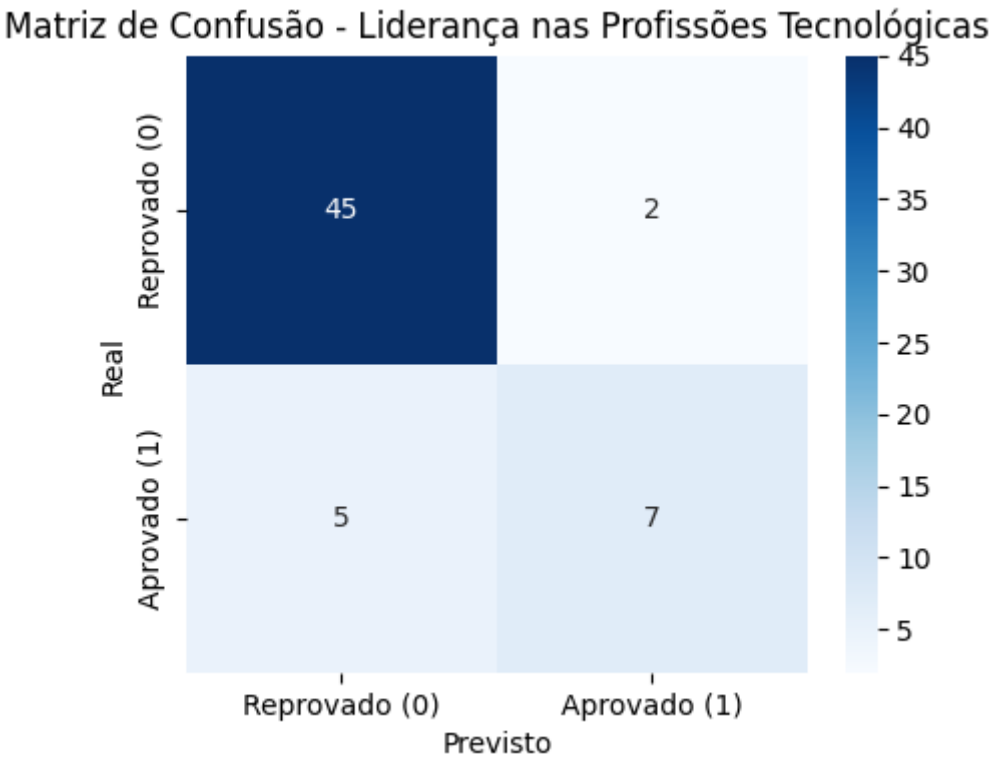
- Banco de Dados Aplicado: Acurácia 0.88



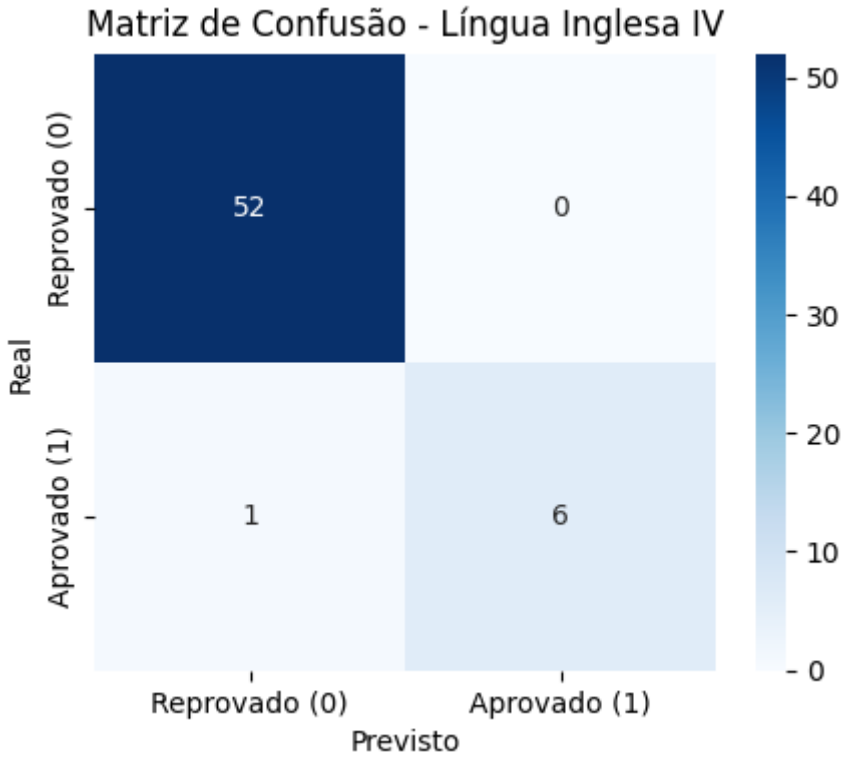
- Engenharia de Software: Acurácia 0.86



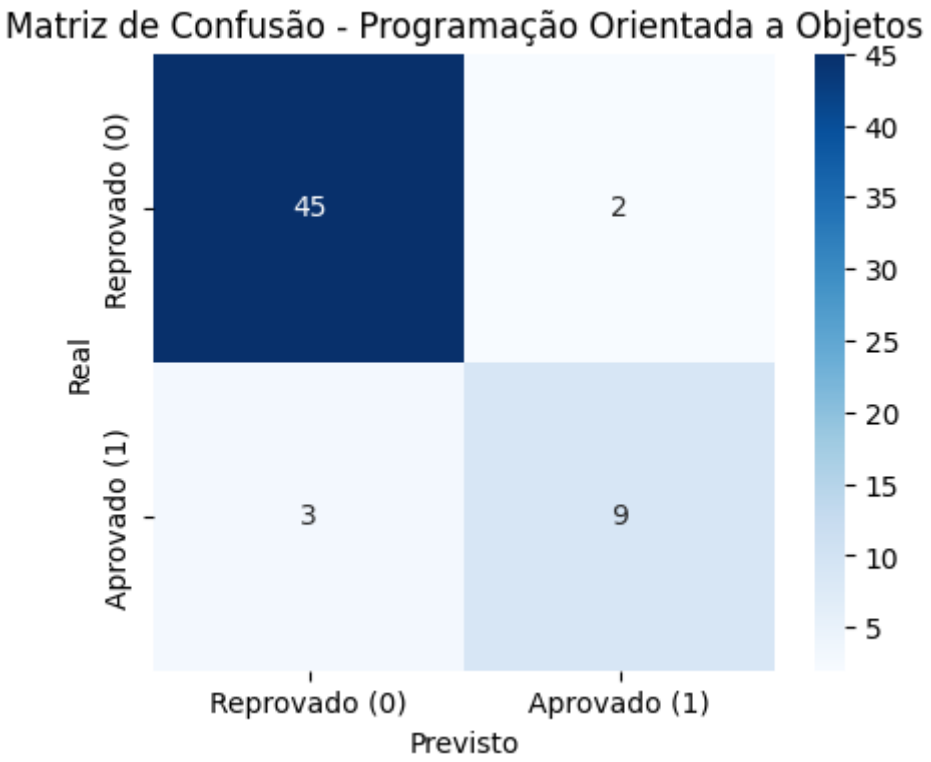
- Liderança nas Profissões Tecnológicas: Acurácia 0.88



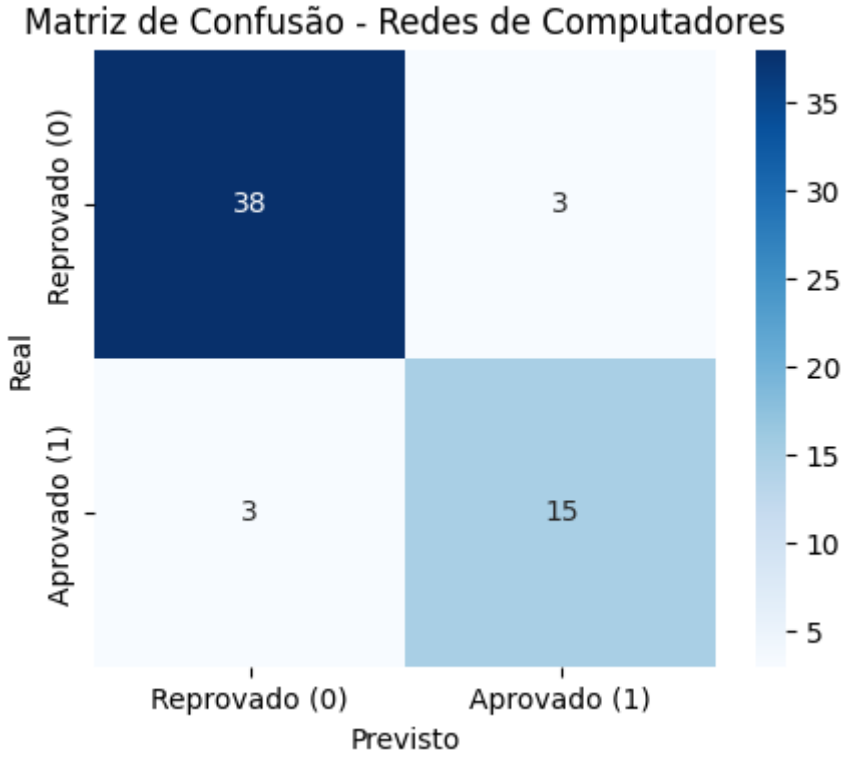
- Língua Inglesa IV: Acurácia 0.98



- Programação Orientada a Objetos: Acurácia 0.91



- Redes de Computadores: Acurácia 0.89



5. INTERPRETAÇÃO

Utilizando como base de análise a disciplina de “Banco de Dados Aplicado” com o algoritmo RandomForest, que compartilha resultados semelhantes com as outras disciplinas

também previstas do mesmo semestre, verifica-se na figura abaixo uma acurácia alta, mas um desbalanceamento visível nos dados utilizados pelo modelo, classe "0", que são os reprovados é muito mais alta que a classe "1" que são os aprovados.

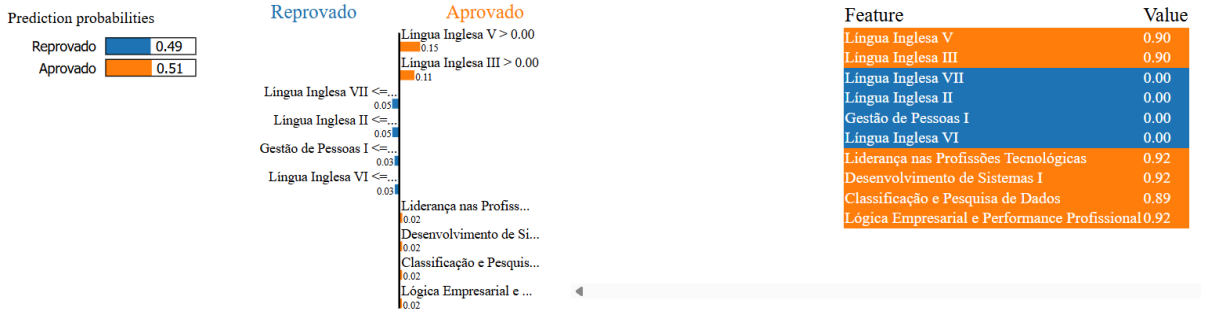
```
=== RandomForest para Banco de Dados Aplicado ===
Acurácia: 0.8983050847457628
              precision    recall  f1-score   support

           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
```

Ao confirmar os valores gerais de alunos aprovados verifica-se que dentre os mais de quatro mil registros utilizados, cerca de três mil e onze são de alunos aprovados e mil e quinhentos e noventa e três reprovados, portanto, constata-se que o desbalanceamento na matriz de confusão se dá pela alta distribuição de registros de alunos reprovados dentro do conjunto de teste, porém, esse comportamento se dá pela tentativa de balanceamento do modelo, levando em conta que RandomForest tenta realizar um balanceamento entre as classes claramente desproporcionais, o modelo dá mais relevância a classe minoritária.

Nas figuras abaixo estão listadas explicações locais de algumas previsões do modelo geradas com LIME, definindo a maneira pela qual o modelo decidiu se as determinadas cadeiras listadas seriam dadas como aprovadas ou não para o aluno escolhido.



6. REFINAMENTOS

Não foi realizado nenhum refinamento no modelo ou pipeline para avaliação de mudanças ou melhorias.

7. CONCLUSÕES E PRÓXIMOS PASSOS

Os resultados obtidos com este trabalho indicam um modelo bastante balanceado e treinado para realizar previsões de aprovações de alunos nas cadeiras do 4° semestre, cabe futuramente ser utilizado com estratégia pela coordenadoria e diretoria, até mesmo com outras aplicações como análise de perfil individual de alunos, visando um possível indicativo ou auxílio para a determinação de uma nova grade ou alteração pontual de ementa de uma cadeira

Devido ao tempo de espera para a aprovação da solicitação de uso deste dataset, seria interessante em uma oportunidade futura, realizar um melhor tratamento do mesmo e aplicação de diferentes algoritmos para resolver diferentes problemas ou gerar análises ainda mais úteis.

8. ÉTICA E LIMITAÇÕES

O dataset pertence ao banco de dados do ambiente de desenvolvimento do projeto “AppAMF”, a carteira digital do aluno AMF, sendo de propriedade da instituição Antonio Meneghetti Faculdade, que autorizou o aluno a usar para fins acadêmicos apenas nas cadeiras de Inteligência Artificial II e Ciência de Dados. O dataset contém dados anonimizados de alunos e disciplinas, fornecendo mais de 30 mil registros passíveis de análise dentro da ciência de dados e inteligência artificial.

O dataset contém algumas limitações claras em sua utilização quanto ao curso de Sistemas de Informação, como por exemplo, abrange um número muito maior de registros de alunos aprovados do que reprovados, contém “sujeira” nos nomes das disciplinas, o que necessita de tratamento geral e correções especiais para casos especiais, disciplinas muito pouco cursadas, e além disso, contém um número relativamente pequeno de dados, totalizando menos de cinco mil registros que contém para cada aluno uma ou mais matrículas de uma mesma disciplina.

9. REPRODUTIBILIDADE

O repositório “g1-ia2” contém o atual trabalho, para a execução do mesmo será necessário o arquivo “global.csv” que contém o dataset utilizado, porém, o mesmo não foi comitado para o GitHub por motivos de segurança e um termo de responsabilidade, será necessário solicitar o mesmo ao responsável pelo trabalho até o fim da disciplina (dezembro de 2025), com base nisso, possuindo o dataset, basta realizar os seguintes passos:

1. Clonar o repositório em “https://github.com/raisson-souza/g1-inteligencia-artificial-2”;
2. Inicializar um “venv” Python rodando “python -m venv venv”;
3. Instalar as dependências necessárias para o projeto: “pip install -r requirements.txt”;
4. Introduzir o dataset como arquivo chamado “global.csv” na pasta “datasets”;
5. Rodar todo o trabalho em sequência no arquivo “classification.ipynb”.

10. REFERÊNCIAS

Não há referências de pesquisas.

11. APÊNDICE

A seguir foram disponibilizadas amostras parciais do dataset sendo utilizado em código.

```python
import pandas as pd

dataset_path = "datasets/global.csv"

df = pd.read_csv(dataset_path)
df = df.sort_values(by=["student_id", "period", "discipline"])

print(f"Quantidade de registros: {len(df)}\n")
print(df.head())
```

```
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
```

```
Quantidade de registros: 4604

        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                               Psicologia do Líder
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                             Sistemas Distribuídos
7850  1049      100094              Desenvolvimento de Sistemas Integrado
7851  1050      100094  Ontopsicologia Aplicada a Sistemas de Informaç ...
7849  1048      100094                Trabalho de Conclusão de Curso II

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