Projeto Integrador I: Análise de Desempenho Acadêmico x Presença em Sala

Ike Gabriel Rodrigues Kenard, Leonardo de Lima Amaral,
Marcelo Saraiva Cavalvanti, Vitor Nascimento Franco,
Rafael Mascarenhas Brown de Andrade,
Alessandra de Souza Gonçalves,
Rodrigo Lins Bezerra Magalhaes

2025-06-25

Universidade CEUB

Projeto Integrador I

Grupo Ike - Análise Rendimento Acadêmico

Brasília

2025

Table of contents

Chapter 1

P1 - Análise de Desempenho Acadêmico com Base na Frequência em Sala

Chapter 2

Apresentação Geral

Este relatório tem como objetivo apresentar o desenvolvimento do Projeto Integrador I do curso de Ciência de Dados e Machine Learning do Centro Universitário de Brasília (CEUB). O projeto foi idealizado com foco na automatização da coleta de presença dos alunos em sala de aula e na análise da relação entre frequência e desempenho acadêmico, utilizando conceitos e ferramentas de ciência de dados.

A proposta surgiu da necessidade de tornar o processo de chamada mais eficiente, confiável e capaz de fornecer informações relevantes para a gestão acadêmica. Para isso, foi projetado um sistema que integra hardware (RFID UHF), backend (API para ingestão de dados em tempo real), pipelines de tratamento e dashboards analíticos.

Do ponto de vista analítico, o projeto aplica técnicas de Análise de Séries Temporais para compreender como a frequência dos estudantes evolui ao longo do tempo e de que forma ela se correlaciona com suas notas. Essa abordagem visa oferecer insights acionáveis para professores e coordenadores sobre o comportamento dos alunos.

O trabalho está estruturado em cinco capítulos, além desta introdução:

- **Capítulo 1 – Introdução**: apresenta o problema, os objetivos e a justificativa da proposta.
- **Capítulo 2 – Metodologia**: descreve as etapas de desenvolvimento e as ferramentas utilizadas.
- **Capítulo 3 – Desenvolvimento**: detalha a implementação das tecnicas utilzadas e análises feitas.
- **Capítulo 4 – Resultados**: exibe gráficos, análises e métricas obtidas.
- **Capítulo 5 – Conclusão**: discute os aprendizados e as possibilidades de evolução do projeto.

Espera-se que esta iniciativa contribua para uma gestão educacional mais inteligente, baseada em dados confiáveis e análises precisas.

Chapter 3

1 - Objetivos e Motivações:

Historicamente, a presença dos estudantes em sala de aula está ligada a um desempenho acadêmico superior. Este projeto visa examinar a conexão entre a presença dos alunos e seus resultados finais, bem como provar que a viabilidade do RFID para uma melhor auxílio acadêmico

Através da avaliação de registros históricos de presença e notas, nosso objetivo é produzir percepções que fundamentem e fortifiquem a implementação de um sistema automatizado de controle de frequência nas instituições educacionais.

O primeiro passo emprega algoritmos de aprendizado de máquina, mais especificamente um Random Forest Regressor (RFR) e uma Regressão Linear (RL), para analisar o efeito da frequência no rendimento escolar. Esta avaliação tem como objetivo fundamentar as decisões dos interessados e apoiadores do projeto.

A importância desta pesquisa está na procura por soluções que aprimorem os processos acadêmicos, aprimorem o monitoramento dos estudantes e, consequentemente, auxiliem na melhoria da qualidade da educação.

Chapter 4

2 - Métodologia:

Este projeto fez uso de informações públicas sobre o rendimento de alunos do ensino secundário de duas instituições de ensino portuguesas: Gabriel Pereira (GP) e Mousinho da Silveira (MS). A coleta de dados ocorreu através de relatórios escolares e questionários aplicados aos estudantes.

O dataset possui características associadas a:

- Notas dos três períodos (G1, G2, G3).
- Informações demográficas: idade, gênero, local de residência (urbano ou rural), número de membros da família e outros.
- Fatores sociais: situação familiar, conexão com a internet, conexão amorosa, entre outros.
- Elementos acadêmicos: horas de estudo por semana, atividades extracurriculares, suporte educacional, ausências e reprovações passadas.

4.1 Dicionário de Variáveis

- **school:** escola do estudante (GP ou MS)
- **sex:** gênero (F ou M)
- **age:** idade (15 a 22 anos)
- **address:** tipo de endereço (U - urbano, R - rural)
- **famsize:** tamanho da família (LE3 3, GT3 $>$ 3)
- **Pstatus:** situação de coabitação dos pais (T - juntos, A - separados)
- **reason:** motivo para escolha da escola (home, reputation, course, other)
- **Mjob / Fjob:** profissão da mãe/pai (teacher, health, services, at_home, other)
- **Medu / Fedu:** escolaridade da mãe e do pai (0 a 4)

Código	Nível de Educação dos responsáveis
0	nenhum
1	Fundamenal I
2	Fundamenal II
3	Ensino Médio
4	Ensin Superior

- **guardian:** responsável (mother, father, other)
- **traveltime:** tempo de deslocamento até a escola (1 - <15min a 4 - >1h)
- **studytime:** tempo de estudo semanal (1 - <2h a 4 - >10h)
- **failures:** número de reprovações anteriores (0 a 4)
- **schoolsup:** apoio educacional extra (yes/no)
- **famsup:** apoio da família (yes/no)
- **paid:** aulas extras pagas (yes/no)
- **activities:** atividades extracurriculares (yes/no)
- **nursery:** frequentou pré-escola (yes/no)
- **higher:** desejo de cursar ensino superior (yes/no)
- **internet:** acesso à internet em casa (yes/no)
- **romantic:** possui relacionamento amoroso (yes/no)
- **famrel:** qualidade do relacionamento familiar (1 - ruim a 5 - excelente)
- **freetime:** tempo livre após a escola (1 a 5)
- **goout:** frequência de saída com amigos (1 a 5)
- **Dalc:** consumo de álcool durante a semana (1 a 5)
- **Walc:** consumo de álcool no final de semana (1 a 5)
- **health:** estado de saúde (1 - muito ruim a 5 - muito bom)
- **absences:** número de faltas (0 a 93)

4.2 Variáveis de Notas:

G1: nota do 1º período (0 a 20) G2: nota do 2º período (0 a 20) G3: nota final do ano (0 a 20) — variável alvo do modelo

Observação: A variável alvo G3 possui forte correlação com G1 e G2, pois estas representam avaliações dos períodos anteriores no mesmo ano letivo.

##Tecnologias Empregadas

- RFID: Tecnologia proposta para automatizar o controle de presença.
- Python: Uma linguagem usada para análise e modelagem.
- Bibliotecas: Pandas,Numpy , Matplotlib, Seaborn, Scikit-Learn.
- Aprendizado de Máquina: Retorno de Floresta Aleatória.

4.2.1 Fluxograma do Projeto

1. Obtenção dos dados: Foram achados no Kaggle;

2. Analise e Compreensão dos Dados: Análise das relações entre as variáveis, distribuição das mesmas e interpretação dos dados;
3. Higienização e Pré-tratamento:

- Manipulação de valores incongruentes;
- Ajuste da distribuição das variáveis numéricas utilizando o MinMaxScaler.
- Uso de Dummys nas variaveis binárias:

4. Modelagem Preditiva:

- Divisão em grupos de treinamento e teste (80/20);
- Uso do algoritmo um Random Forest Regrassor para antecipar a pontuação final (G3) com base na frequencia;
- Uso de uma Regressão Linear para afirmar se a numero de faltas afeta positivamente na nota dos alunos.

5. Métricas:

- MAE (Erro Médio Absoluto);
- MSE (Erro Médio Quadrático);
- Coeficiente de Determinação(R2).

Chapter 5

3 - Desenvolvimento

5.1 Análise Exploratória

Realizaram-se análises estatísticas e gráficos, com ênfase em:

5.1.1 1 - A atribuição das pontuações aos estudantes.

Gŕaico de barras comparando notas estudantes de cada escola

5.1.2 2 - A distribuição da frequência.

Histograma de frequencia de cada escola

5.1.3 3 - Relação entre presença e pontuação.

Heatmap de presença x pontuação por escola

Essas avaliações evidenciaram uma tendência evidente: estudantes que frequentam mais frequentemente tendem a alcançar melhores desempenhos acadêmicos.

5.2 Hiperparâmetros do Random Forest Regressor:

- Quantidade de estimadores: 100
- Profundidade máxima (max_depth): Não definida (árvores se desenvolveram de forma autônoma)
- Estado Aleatório: 42 (para replicabilidade)

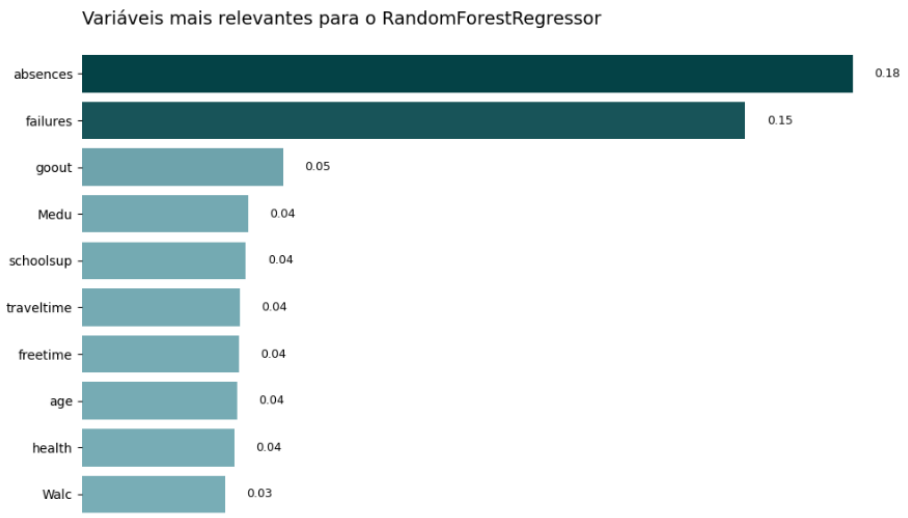
5.3 Dicionário de Dummys

Variável	Valor	Significado
schoolsup	1	Aluno recebe suporte educacional extra
famsup	1	Aluno recebe suporte extra da família
paid	1	Aluno paga aulas extras
activities	1	Aluno participa de atividades extracurriculares
nursery	1	Aluno frequentou o berçário
higher	1	Aluno pretende fazer curso superior
internet	1	Aluno tem acesso à internet em casa
romantic	1	Aluno está em um relacionamento romântico

Chapter 6

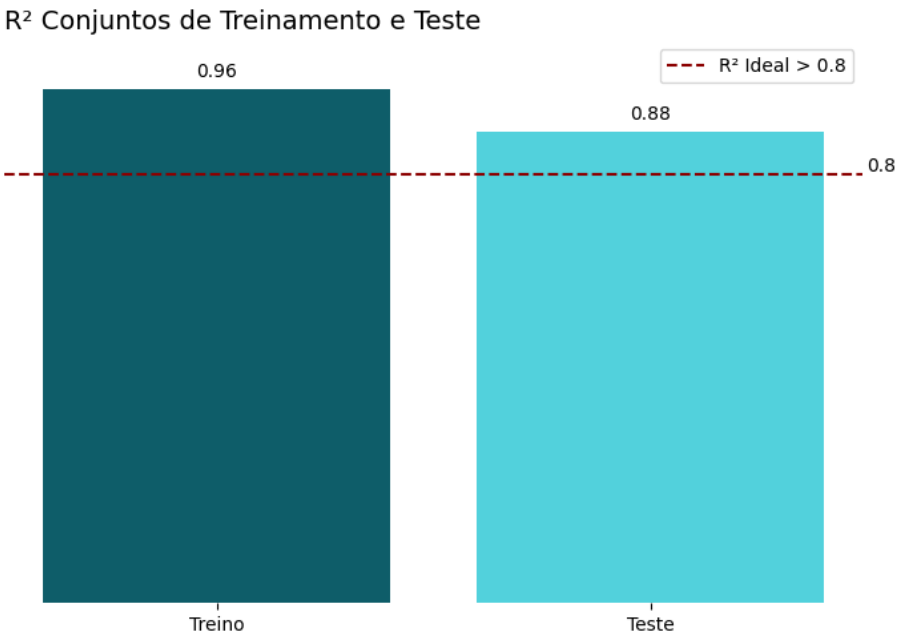
4 - Resultados

6.1 Variaveis com Maior Relevância:



Interpretação das variaveis: É notoória como para o modelo a variavel *Absence* (falatas/aussncia) se mostrou como a mais relevante para a resolução do problema

6.2 Métricas Random Forest Regressor:



Média R2: 0.82 Indica Bom ajuste; MAE: 1.25 Indica de baixa precisão média; MSE: 2.79 Indica que o mododelo está dentro de uma faixa tolerável, melhorando a precisão das previsões.

Esses achados corroboram a hipótese inicial: existe uma correlação significativa entre a presença dos estudantes e suas notas.

Análise de Overfitting (RandomForest)

Observe se o R^2 de treinamento é similar ao de teste, indicando menor propensão a overfiting

6.3 Métricas Regressão liniear :

- **Mean Absolute Error (MAE):** 3.41
- **Mean Squared Error (MSE):** 20.27
- **R-squared (R^2):** 0.10

6.4 Interpretação dos modelos:

Estratégias voltadas para o aumento da frequência dos alunos podem levar a melhorias significativas no desempenho dos alunos. Essas percepções foram

cruciais para demonstrar aos interessados a viabilidade da implementação do sistema RFID.

Chapter 7

5 - Conclusão

Nossa análise mostrou que **faltar às aulas** (*Absence*") é um fator que mais pesa na previsão do desempenho dos alunos, com uma importância de cerca de 0.18. Isso significa que quanto mais um aluno falta, pior tende a ser seu desempenho. Nossos testes iniciais da Regressão Linear também apontam para essa relação, mesmo que o resultado do R^2 tenha dado 0.10 sugira que há mais a ser explorado.

Além disso, conseguimos criar um modelo preditivo usando Random Forest que explica em média 82% da variação nas notas dos alunos. Isso reforça muito a ideia de que estar presente na escola é crucial para ir bem nos estudos.

A tecnologia RFID surge como uma solução promissora para melhorar o controle de presença, coletar dados automaticamente para análises futuras e ajudar na administração da escola e os alunos. Com base nisso, os próximos passos incluem criar um modelo experimental do sistema RFID e coletar dados reais de presença com ele para deixar nosso modelo ainda mais preciso.

É importante destacar que, apesar dessas descobertas claras, o desempenho dos alunos é algo complexo e exige mais investigação. Por isso, é fundamental expandir nossas análises para incluir outros fatores, como a participação em atividades extras, o desempenho em provas contínuas, o histórico escolar completo e, especialmente, questões socioeconômicas dos alunos.

Em resumo, este projeto prova a importância do controle automático de presença para o avanço acadêmico. Ele oferece uma base sólida para futuras ações, mas também sublinha a necessidade de pesquisar mais a fundo para entender todos os elementos que influenciam o sucesso educacional.

Em resumo, com o RFID avançaremos na análise de dados por que passaríamos a capitar algo algo que uma chamada normal não apresenta:

1 - Se o aluno aproveita o tempo todo de cada aula do ano ou não (deixamos de

ficar limitados a condição 'binária' de Presença ou Falta;

2 - E assim testar se os alunos, em geral, chegam pontualmente, no contexto do ceub às 8h e saem as 10:50h, apresentam melhor rendimento do que outros que têm o costume de chegar atrasado (e o nivel de atraso), ou quem têm o costume de sair bem mais cedo.

Chapter 8

Referências

As referências estão organizadas no arquivo `.bib` e serão exibidas automaticamente.