UFPI - CCN - DC
CIÊNCIA DA COMPUTAÇÃO
Tópicos em Computação Aplicada - Prof. Dr. Vitor Cortez

Relatório Técnico

Análise de Dados - Notas dos Alunos do Curso de Ciência da Computação da UFPI

Cayo Cesar Lopes Mascarenhas Pires Cardoso
Fábio Anderson Carvalho Silva
Antonio Geraldo Rego Junior
Icaro Gabryel de Araujo Sillva
Lucas Sales Aguiar Silveira

Teresina
2025

Sumário

1. Introdução

A análise de dados educacionais emergiu como uma ferramenta estratégica fundamental para a gestão e o aprimoramento de instituições de ensino superior. No contexto universitário, a capacidade de coletar, processar e interpretar grandes volumes de dados acadêmicos permite a identificação de padrões de comportamento, fatores de risco para evasão e retenção, e oportunidades para a otimização de currículos e práticas pedagógicas. A aplicação de técnicas de ciência de dados possibilita transformar dados brutos de matrículas, notas e históricos escolares em insights acionáveis que podem guiar decisões administrativas e apoiar o sucesso estudantil.

O curso de Ciência da Computação da Universidade Federal do Piauí (UFPI), como um polo de formação tecnológica na região, enfrenta desafios comuns a cursos de alta demanda e complexidade técnica, incluindo a necessidade de monitorar continuamente o desempenho de seus discentes e garantir uma trajetória acadêmica de qualidade. A compreensão aprofundada dos fatores que influenciam as notas e as taxas de aprovação dos alunos é crucial para o desenvolvimento de políticas internas que visem mitigar dificuldades de aprendizagem, modernizar a estrutura curricular e oferecer suporte direcionado aos estudantes.

Neste contexto, este relatório técnico apresenta uma análise de dados detalhada sobre o desempenho acadêmico dos alunos do curso de Ciência da Computação da UFPI.

O objetivo geral deste trabalho é explorar o conjunto de dados históricos de notas dos alunos para identificar padrões, tendências e correlações que caracterizem o rendimento acadêmico no curso.

Para alcançar este objetivo, foram definidos os seguintes objetivos específicos:

Caracterizar a base de dados: Descrever a estrutura, a origem e o conteúdo dos dados disponíveis.

Avaliar a qualidade dos dados: Realizar uma análise de integridade para identificar e tratar dados ausentes, duplicados ou inconsistentes.

Executar uma análise exploratória: Investigar as características das variáveis por meio de estatísticas descritivas e visualizações, analisando a distribuição das notas e a frequência de resultados.

Desenvolver modelos de dados: Investigar a possibilidade de prever o desempenho em disciplinas futuras ou o risco de reprovação com base em dados históricos.

Este estudo justifica-se pela sua capacidade de fornecer à coordenação do curso, ao corpo docente e aos próprios alunos uma visão quantitativa e qualitativa do cenário

acadêmico, fomentando uma cultura de decisões baseadas em evidências e contribuindo para a excelência do ensino em Ciência da Computação na UFPI. Os resultados aqui apresentados servirão de base para as discussões subsequentes sobre metodologia, análise e conclusões.

2. Metodologia

Nesta seção, são detalhados os procedimentos metodológicos adotados para a realização da análise de dados, desde a origem e descrição da base de dados até as técnicas empregadas para análise, visualização e, eventualmente, modelagem.

2.1. Fonte e Coleta dos Dados

O conjunto de dados utilizado neste estudo foi construído a partir de informações acadêmicas fornecidas voluntariamente por alunos do curso de Ciência da Computação da UFPI. A coleta foi realizada por meio de formulários eletrônicos, nos quais os próprios estudantes submeteram seus boletins ou históricos escolares. Os dados brutos foram, então, processados e anonimizados para garantir a privacidade dos participantes.

Após a coleta, as informações foram organizadas e estruturadas em três arquivos distintos em formato .csv (valores separados por vírgula), seguindo um modelo de dados normalizado para garantir a integridade e evitar redundâncias. A base de dados é composta pelos seguintes arquivos:

alunos_normalizadas.csv: Este arquivo centraliza as informações cadastrais dos discentes. Como os discentes não foram identificados, esse .csv so possui o IDAluno dos mesmos.

disciplinas_normalizadas.csv: Contém o catálogo de disciplinas oferecidas pelo curso. Cada registro neste arquivo corresponde a uma disciplina única, com atributos como código da disciplina e nome.

notas_normalizadas.csv: Este arquivo representa a tabela de fatos do conjunto de dados, conectando alunos e disciplinas. Cada linha registra o desempenho de um aluno em uma disciplina específico. Os atributos esperados incluem o identificador do aluno, o código da disciplina, as notas de cada unidade, a nota final obtida, a frequência e a situação final do aluno (ex: AM, RN, RF, EF).

Essa estrutura normalizada permite a realização de análises complexas por meio do cruzamento de informações entre os três arquivos, viabilizando uma visão completa da trajetória acadêmica de cada estudante.

2.2. Dicionário de Dados

A seguir, é apresentado o dicionário de dados detalhado para cada um dos arquivos que compõem a base de dados. O dicionário descreve cada atributo, seu tipo e formato.

Arquivo: alunos_normalizadas.csv

Este arquivo funciona como uma tabela de dimensão para os alunos, garantindo sua anonimização.

Nome do Atributo	Descrição	Tipo do Valor	Formato
AlunoID	Identificador numérico único para cada aluno. Chave primária.	Inteiro	Numérico

Arquivo: disciplinas_normalizadas.csv

Este arquivo serve como uma tabela de dimensão para as disciplinas do curso.

Nome do Atributo	Descrição	Tipo do Valor	Formato
Código	Código único da disciplina. Chave primária.	String	Categórico
Disciplina	Nome completo da disciplina.	String	Categórico

Arquivo: notas_normalizadas.csv

Esta é a tabela de fatos, registrando o desempenho acadêmico. AlunoID e Código são chaves estrangeiras que se conectam aos respectivos arquivos.

Nome do Atributo	Descrição	Tipo do Valor	Formato
AlunoID	Chave de referência para o	Inteiro	Numerico

	aluno.		
Código	Chave de referência para a disciplina.	String	Categórico
Unidade 1	Nota da Primeira Avaliação	Float	Numérico
Unidade 2	Nota da Segunda Avaliação	Float	Numérico
Unidade 3	Nota da Terceira Avaliação	Float	Numérico
Unidade 4	Nota da Quarta Avaliação	Float	Numérico
Unidade 5	Nota da Quinta Avaliação	Float	Numérico
Prova Final	Nota da prova final, caso o aluno tenha feito.	Float	Numérico
Resultado	Média Final do Aluno na Disciplina	Float	Numérico
Faltas	Número total de faltas do aluno.	Inteiro	Numérico
Situação	Status final do aluno na disciplina	String	Categórico

2.3. Análise da Qualidade dos Dados

A análise de qualidade dos dados é uma etapa fundamental para garantir a validade e a confiabilidade das conclusões do estudo. Nesta fase, o conjunto de dados foi inspecionado em busca de dados ausentes, duplicados, inconsistentes e outliers.

1. Dados Ausentes (Valores Nulos):

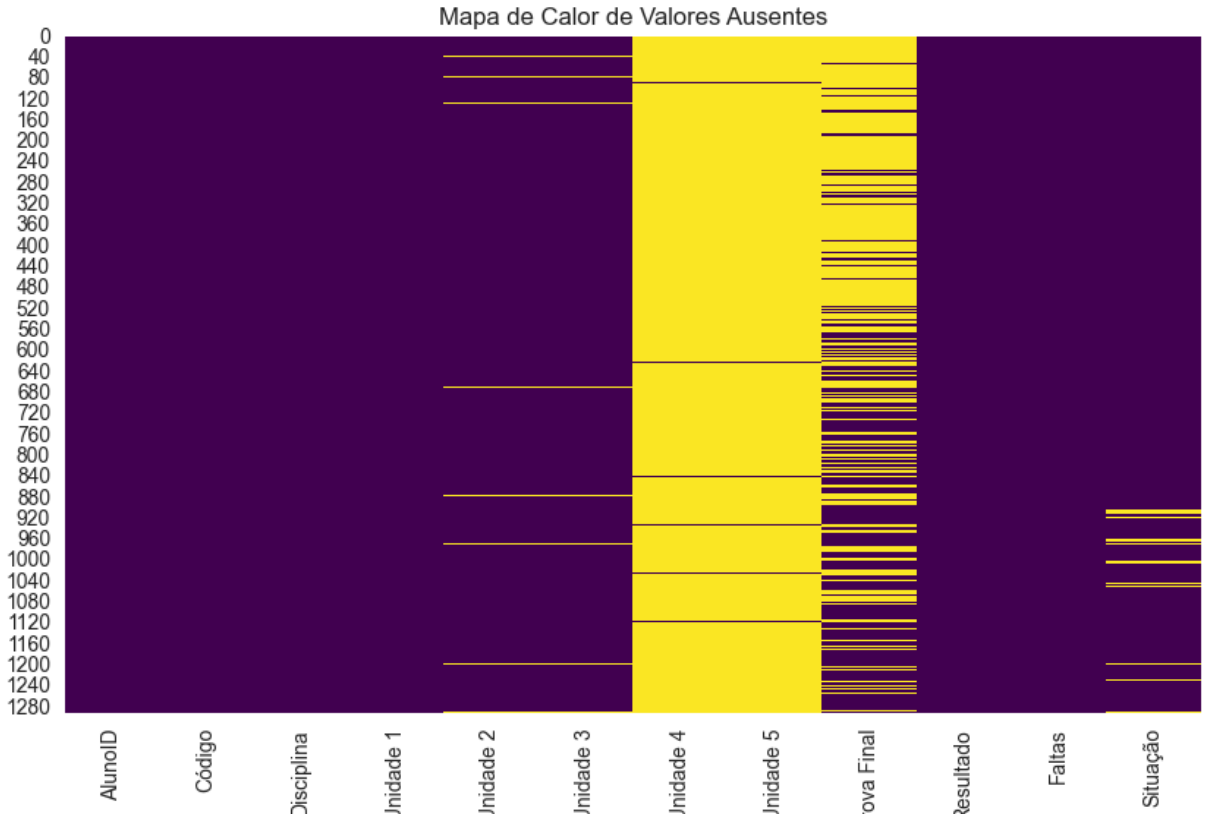
A investigação revelou a presença de valores ausentes em diversas colunas, com destaque para:

Unidade 4 e Unidade 5: Apresentaram uma alta percentagem de ausência (98.68%), o que é esperado, visto que a maioria das disciplinas do curso não possui cinco unidades avaliativas.

Prova Final: Com 56.70% de valores ausentes, reflete os casos em que os alunos foram aprovados por média e não necessitaram realizar a avaliação final.

Unidade 2, Unidade 3 e Situação: Apresentaram percentagens menores de dados ausentes (2.09% e 3.10%, respectivamente), indicando possíveis falhas pontuais na extração ou registro dos dados.

Dado o contexto, os valores ausentes nas colunas de unidades e prova final foram considerados naturais e mantidos como nulos (NaN), enquanto os demais foram tratados durante a análise exploratória.



	Total Ausentes	Porcentagem (%)
Unidade 5	1274	98.683191
Unidade 4	1274	98.683191
Prova Final	732	56.700232
Situação	40	3.098373
Unidade 3	27	2.091402
Unidade 2	27	2.091402

2. Dados Duplicados:

A verificação de registros duplicados no conjunto de dados não identificou nenhuma linha idêntica. Isso indica uma boa integridade na inserção inicial dos dados.

3. Dados Inconsistentes:

Tipos de Dados: Inicialmente, todas as colunas numéricas (notas, faltas) foram carregadas como texto (object). Foi realizado um tratamento para convertê-las aos tipos corretos (float para notas e int para faltas), utilizando a coerção de erros para transformar entradas inválidas em valores nulos.

Valores Inválidos: Foram identificados valores textuais em colunas que deveriam ser numéricas (ex: o texto "Unidade 4" na própria coluna Unidade 4) e na coluna Situação (ex: o texto "Situação"). Estes valores foram tratados como erros de extração e convertidos para nulos durante a limpeza dos dados.

4. Análise de Outliers:

Foi conduzida uma análise de outliers nas colunas de notas utilizando o método do Intervalo Interquartil (IQR). A análise não detectou valores atípicos que justificassem a remoção de registros, sugerindo que as notas se encontram dentro de um intervalo estatisticamente esperado.

Com base nesta análise, o conjunto de dados foi considerado apto para as próximas etapas do estudo, tendo seus principais problemas de qualidade identificados e tratados.

2.4. Análise Exploratória dos Dados

A Análise Exploratória de Dados foi conduzida para investigar as características fundamentais do conjunto de dados, incluindo a distribuição das notas, a relação entre as variáveis, estatísticas descritivas e o desempenho por disciplina.

Análise Univariada e Estatísticas Descritivas: A análise da distribuição das notas das unidades e dos resultados finais revelou uma assimetria negativa, indicando uma concentração de notas altas, próximas do valor máximo (10.0). A Unidade 1 apresentou a menor dispersão entre as notas (desvio padrão de 1.53), enquanto a Unidade 2 mostrou a maior variabilidade (desvio padrão de 2.07).

Distribuição das Médias: A distribuição da média geral dos alunos (considerando todas as disciplinas cursadas) mostrou-se concentrada em torno da nota 8.0, reforçando a tendência de bom desempenho geral entre os alunos da amostra.

Análise de Correlação: A matriz de correlação entre as notas das unidades indicou uma correlação positiva, embora de intensidade moderada. Isso sugere que há uma tendência de que alunos com bom desempenho em uma unidade também se saiam bem nas outras, mas essa relação não é forte o suficiente para que uma nota determine completamente as demais.

Ranking de Disciplinas por Desempenho e Reprovação: A análise por disciplina revelou insights importantes. Disciplinas como "Trabalho de Conclusão de Curso I e II" e "Banco de Dados" apresentaram as maiores médias, frequentemente com nota máxima (10.0). Por outro lado, foi elaborado um ranking das disciplinas com maiores taxas de reprovação, identificando aquelas que representam maiores desafios para os alunos e que podem ser pontos de atenção para a coordenação do curso.

Os insights gerados nesta etapa foram cruciais para a compreensão do perfil de desempenho dos alunos e para orientar a subsequente etapa de modelagem de dados.

2.5. Modelagem dos Dados

Nesta fase do estudo, foi desenvolvido um modelo de aprendizado de máquina com o objetivo de prever o desempenho acadêmico futuro dos alunos. O problema foi enquadrado como uma tarefa de regressão, onde o objetivo é prever as notas (valores contínuos) de um conjunto de disciplinas de um determinado período com base nas notas dos períodos anteriores.

1. Preparação dos Dados para Modelagem:

Para adequar os dados ao formato exigido pelos modelos de regressão, foi realizado um processo de transformação. Primeiramente, calculou-se a média das notas das unidades para cada disciplina cursada por um aluno. Em seguida, os dados foram reestruturados através de uma operação de pivot, resultando em um DataFrame onde cada linha representa um aluno (AlunoID) e cada coluna representa uma disciplina. Os valores da tabela correspondem à média do aluno naquela disciplina. Para casos de reprovação, onde um aluno cursou a mesma disciplina mais de uma vez, foi utilizada a menor nota como valor, representando o cenário de maior dificuldade.

2. Definição do Problema Alvo:

O alvo da predição é um conjunto de notas de múltiplas disciplinas de um período subsequente. Por exemplo, utilizando as notas das disciplinas do 1º período como variáveis de entrada (features), o modelo prevê as notas das disciplinas do 2º período (targets).

3. Seleção de Atributos e Alvo (Features & Target):

As disciplinas obrigatórias do curso foram agrupadas por período (P1 a P7). O treinamento foi realizado de forma sequencial:

Modelo 1: Features = P1, Target = P2

Modelo 2: Features = P1 + P2, Target = P3

... e assim sucessivamente.

As disciplinas de TCC e Estágio foram excluídas por possuírem métodos de avaliação distintos.

4. Modelo Utilizado e Justificativa:

O modelo escolhido foi o Random Forest Regressor, encapsulado por um MultiOutputRegressor. A justificativa para essa escolha baseia-se em:

Natureza do Problema: Trata-se de um problema de regressão multioutput, pois o modelo precisa prever múltiplas notas simultaneamente.

Não Linearidade: As relações entre o desempenho em diferentes disciplinas não são necessariamente lineares, tornando o Random Forest mais adequado que modelos lineares.

Robustez: O Random Forest é um algoritmo robusto contra overfitting e eficaz para lidar com dados complexos e não lineares.

5. Treinamento e Avaliação:

Para cada etapa de predição (ex: prever P2), os dados foram divididos em conjuntos de treino (80%) e teste (20%). A principal métrica utilizada para avaliar o desempenho foi o Erro Médio Absoluto (MAE), que mede a média das diferenças absolutas entre as notas previstas e as reais. Além disso, foi aplicada a técnica de Validação Cruzada (Cross-Validation) com 5 folds para garantir que o desempenho do modelo não dependa de uma divisão específica dos dados.

6. Resultados da Modelagem:

A avaliação dos modelos demonstrou alta capacidade preditiva. O MAE médio, obtido através da validação cruzada para a previsão de cada período, foi consistentemente baixo, conforme apresentado na tabela abaixo:

```
MAE médio para cada período:
Modelo que preve o período 2: 0.44
Modelo que preve o período 3: 0.69
Modelo que preve o período 4: 0.40
Modelo que preve o período 5: 0.49
Modelo que preve o período 6: 0.40
Modelo que preve o período 7: 0.34
```

2.6. Ferramentas Utilizadas

Para a execução deste projeto, foram utilizadas as seguintes tecnologias e bibliotecas da linguagem Python:

- **Pandas:** Para manipulação e análise dos dados.

- **Scikit-learn:** Para a criação, treinamento e avaliação dos modelos de aprendizado de máquina (RandomForestRegressor, MultiOutputRegressor, train_test_split, cross_val_score).
- **Matplotlib:** Para a geração de visualizações e gráficos.
- **Jupyter Notebook:** Como ambiente de desenvolvimento interativo.

3. Resultados e Discussão

Nesta seção, são apresentados e discutidos os resultados obtidos na etapa de modelagem de dados. O foco é a análise do desempenho do modelo preditivo e a interpretação de suas implicações.

3.1. Desempenho do Modelo de Regressão

A avaliação do modelo Random Forest Regressor demonstrou uma notável capacidade de prever as notas dos alunos em períodos futuros. A performance, medida pelo Erro Médio Absoluto (MAE) através de validação cruzada, foi consistentemente baixa, indicando uma alta precisão nas previsões.

A tabela a seguir resume o MAE médio para a previsão de cada período letivo:

Modelo (Previsão do Período)	Erro Médio Absoluto (MAE)
Periodo 2	0,44
Periodo 3	0,69
Periodo 4	0,40
Periodo 5	0,49
Periodo 6	0,40
Periodo 7	0,34

A análise dos resultados permite extrair as seguintes observações:

- **Alta Acurácia:** Os valores de MAE, variando entre 0.34 e 0.69, são considerados excelentes em uma escala de notas de 0 a 10. Isso significa que, em média, as previsões do modelo erram por menos de 0.7 pontos em relação à nota real do aluno.

- **Variação de Desempenho:** O erro foi ligeiramente maior na previsão do 3º período (MAE = 0.69). Isso pode ser atribuído à natureza das disciplinas deste período, que podem ser mais desafiadoras ou ter menor correlação com o desempenho dos períodos iniciais. Em contrapartida, a previsão para o 7º período apresentou o menor erro (MAE = 0.34), sugerindo que, em estágios mais avançados do curso, o desempenho do aluno tende a ser mais estável e, portanto, mais previsível.
- **Implicações Práticas:** A eficácia do modelo valida a hipótese de que o histórico de desempenho de um aluno é um forte preditor de seu sucesso futuro. Essa ferramenta pode ser utilizada pela coordenação do curso para identificar, de forma proativa, alunos com risco de dificuldades em disciplinas futuras, permitindo a implementação de ações de apoio e intervenção pedagógica precoce.

4. Conclusão

Este relatório técnico detalhou o processo de análise e modelagem de dados de desempenho acadêmico dos alunos de Ciência da Computação da UFPI. A partir de um conjunto de dados coletado voluntariamente, foi possível desenvolver e validar um modelo de aprendizado de máquina com alta capacidade preditiva.

O principal resultado deste estudo é a comprovação de que é viável prever, com um grau de erro muito baixo (MAE < 0.7), as notas que um aluno obterá em períodos futuros com base em seu histórico. O modelo Random Forest Regressor se mostrou robusto e adequado para essa tarefa.

Limitações e Trabalhos Futuros:

A principal limitação deste estudo reside na natureza da amostragem, que foi voluntária e pode não representar a totalidade do corpo discente, além de um tamanho amostral relativamente pequeno. Como sugestões para trabalhos futuros, recomenda-se:

- **Ampliação da Base de Dados:** Utilizar um conjunto de dados mais abrangente, com registros de um maior número de alunos e de um período de tempo mais longo.
- **Inclusão de Novas Features:** Adicionar outras variáveis ao modelo, como dados sociodemográficos (anonimizados), informações sobre o colégio de origem ou o tipo de ingresso na universidade, para verificar se aumentam o poder preditivo.

- **Desenvolvimento de uma Ferramenta Interativa:** Implementar o modelo em uma aplicação que possa ser utilizada pela coordenação para realizar simulações e obter previsões em tempo real, auxiliando na gestão acadêmica e no acompanhamento individual dos alunos.