



Trabalho Mineração

APRESENTAÇÃO

PARTE 1

Tópicos

1

Seleção dos Dados

2

Pré-Processamento

3

Transformação dos Dados

4

Mineração dos Dados

5

Interpretação e Análise das
Regras

Seleção e Pré-Processamento

ideia de investigar quais cursos apresentam maior dificuldade em determinadas matérias e verificar a associação entre as dificuldades das matérias

A proposta foi analisar as taxas de reprovação por curso, considerando o total de alunos em todos os anos disponíveis. Para isso, contabilizamos todos os alunos matriculados em cada disciplina dentro de cada curso, a fim de calcular a taxa de reprovação total de cada curso.

Transformação dos Dados

Normalização dos dados

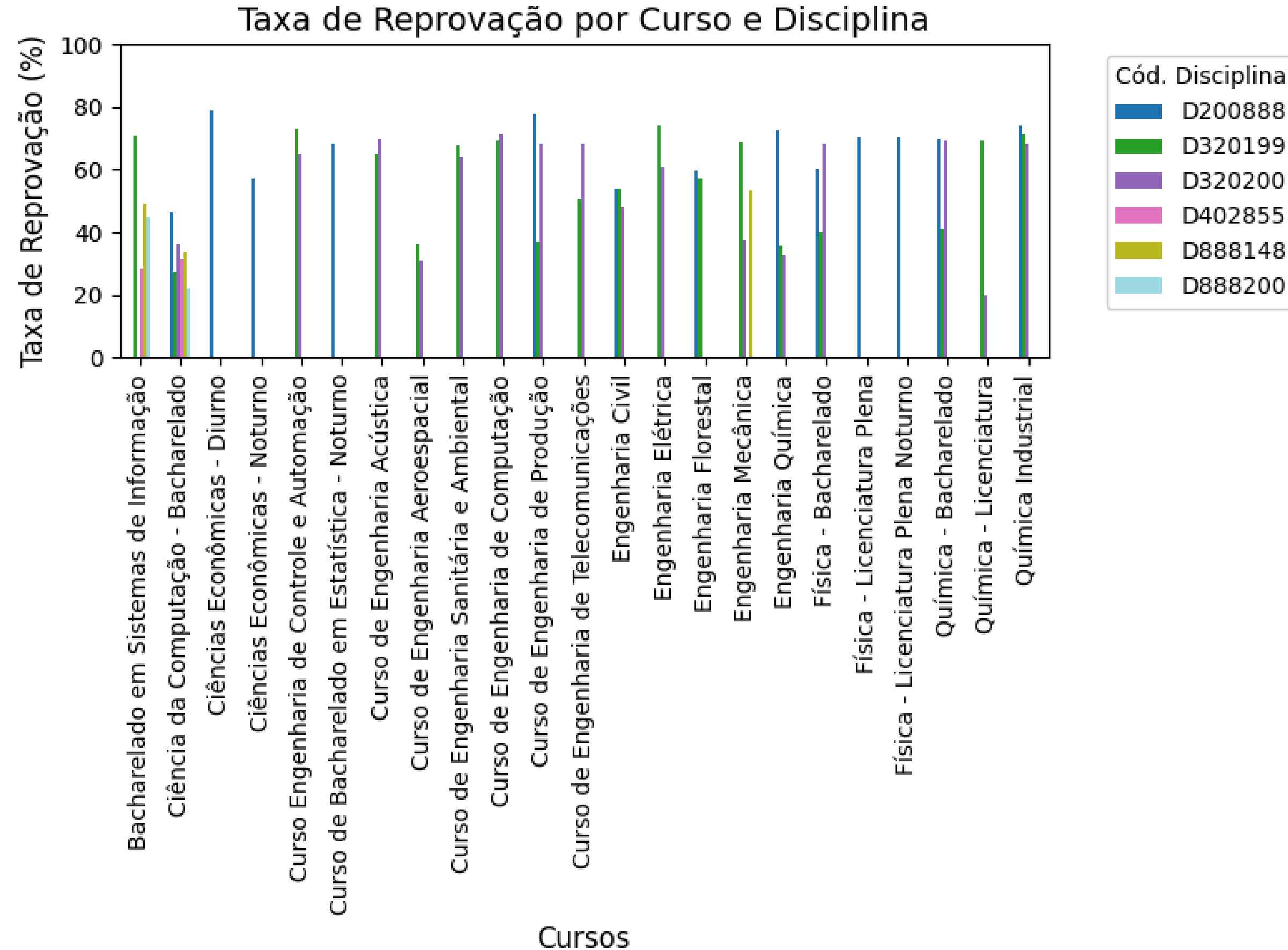
Para simplificar a análise, as situações Reprovado e Reprovado por Frequência foram agrupadas em uma única categoria, representando as reprovações em geral

Além disso, criamos uma coluna que exibia o número total de alunos em cada disciplina e turma.

Transformação dos Dados

| | Curso | Cód. | Disciplina | Total_alunos | Reprovados | Taxa_Reprovacao |
|----|--|------|------------|--------------|------------|-----------------|
| 0 | Bacharelado em Sistemas de Informação | | D320199 | 76 | 54 | 71.052632 |
| 1 | Bacharelado em Sistemas de Informação | | D402855 | 133 | 38 | 28.571429 |
| 2 | Bacharelado em Sistemas de Informação | | D888148 | 150 | 74 | 49.333333 |
| 3 | Bacharelado em Sistemas de Informação | | D888200 | 109 | 49 | 44.954128 |
| 4 | Ciência da Computação - Bacharelado | | D200888 | 136 | 63 | 46.323529 |
| 5 | Ciência da Computação - Bacharelado | | D320199 | 122 | 33 | 27.049180 |
| 6 | Ciência da Computação - Bacharelado | | D320200 | 55 | 20 | 36.363636 |
| 7 | Ciência da Computação - Bacharelado | | D402855 | 131 | 41 | 31.297710 |
| 8 | Ciência da Computação - Bacharelado | | D888148 | 145 | 49 | 33.793103 |
| 9 | Ciência da Computação - Bacharelado | | D888200 | 96 | 21 | 21.875000 |
| 10 | Ciências Econômicas - Diurno | | D200888 | 143 | 113 | 79.020979 |
| 11 | Ciências Econômicas - Noturno | | D200888 | 147 | 84 | 57.142857 |
| 12 | Curso Engenharia de Controle e Automação | | D320199 | 70 | 51 | 72.857143 |

Transformação dos Dados



Transformação dos Dados

Top 15 Cursos que tiveram mais dificuldade realizando a Disciplina:

| | Curso | Cód. | Disciplina | Total_alunos | Reprovados | Taxa_Reprovacao |
|----|--|------|------------|--------------|------------|-----------------|
| 10 | Ciências Econômicas - Diurno | | D200888 | 143 | 113 | 79.02 |
| 23 | Curso de Engenharia de Produção | | D200888 | 90 | 70 | 77.78 |
| 51 | Química Industrial | | D200888 | 39 | 29 | 74.36 |
| 31 | Engenharia Elétrica | | D320199 | 218 | 161 | 73.85 |
| 12 | Curso Engenharia de Controle e Automação | | D320199 | 70 | 51 | 72.86 |
| 38 | Engenharia Química | | D200888 | 239 | 173 | 72.38 |
| 52 | Química Industrial | | D320199 | 21 | 15 | 71.43 |
| 22 | Curso de Engenharia de Computação | | D320200 | 56 | 40 | 71.43 |
| 0 | Bacharelado em Sistemas de Informação | | D320199 | 76 | 54 | 71.05 |
| 45 | Física - Licenciatura Plena Noturno | | D200888 | 17 | 12 | 70.59 |
| 44 | Física - Licenciatura Plena | | D200888 | 17 | 12 | 70.59 |
| 16 | Curso de Engenharia Acústica | | D320200 | 70 | 49 | 70.00 |
| 46 | Química - Bacharelado | | D200888 | 43 | 30 | 69.77 |
| 49 | Química - Licenciatura | | D320199 | 69 | 48 | 69.57 |
| 21 | Curso de Engenharia de Computação | | D320199 | 105 | 73 | 69.52 |

Mineração dos Dados

Buscamos identificar as regras de associação entre as reprovações.

Consideramos uma disciplina problemática quando a taxa de reprovação é superior a 50%. Com base nessas regras de associação, nosso objetivo foi encontrar outras disciplinas nas quais um curso, que apresenta dificuldade em uma matéria (A), também enfrenta dificuldades em outra (B).

Regras de Associação:

| | antecedents | consequents | support | confidence | lift |
|---|-------------|-------------|---------|------------|------|
| 2 | (D888148) | (D320199) | 0.04 | 1.00 | 1.92 |
| 3 | (D320199) | (D888148) | 0.04 | 0.08 | 1.92 |
| 1 | (D320199) | (D320200) | 0.30 | 0.58 | 1.34 |
| 0 | (D320200) | (D320199) | 0.30 | 0.70 | 1.34 |

Regras relevantes para disciplinas problemáticas:

| | antecedents | consequents | support | confidence | lift |
|---|-------------|-------------|---------|------------|------|
| 0 | (D320200) | (D320199) | 0.30 | 0.70 | 1.34 |
| 2 | (D888148) | (D320199) | 0.04 | 1.00 | 1.92 |

Interpretação e Análise das Regras

Utilizamos o algoritmo Apriori e a criação de regras de associação para identificar padrões nas taxas de reprovação entre disciplinas.

Se um curso tem a disciplina A com alta taxa de reprovação, então a disciplina B também tende a ter alta taxa de reprovação com uma certa probabilidade.

Se a confiança de uma regra é 0.6, isso significa que 60% dos cursos que apresentam alta taxa de reprovação na disciplina A também têm alta taxa de reprovação na disciplina B. Já o lift mede a força da associação entre o antecedente e o consequente, comparando a probabilidade de ambos ocorrerem juntos com a probabilidade de ocorrerem independentemente



APRESENTAÇÃO

PARTE 2

Tópicos

1

Seleção dos Dados

2

Pré-Processamento

3

Transformação dos Dados

4

Mineração dos Dados

5

Interpretação e Análise das
Regras

Seleção e Pré-Processamento

Primeiro passo realizamos o pré processamento dos dados, excluindo colunas que não serão úteis e adicionando novas colunas úteis

A ideia era analisar algum tipo de relação entre ingressantes e formandos, simular uma taxa de evasão e verificar quais os prédios com maiores taxas de evasão

Analise dos Dados

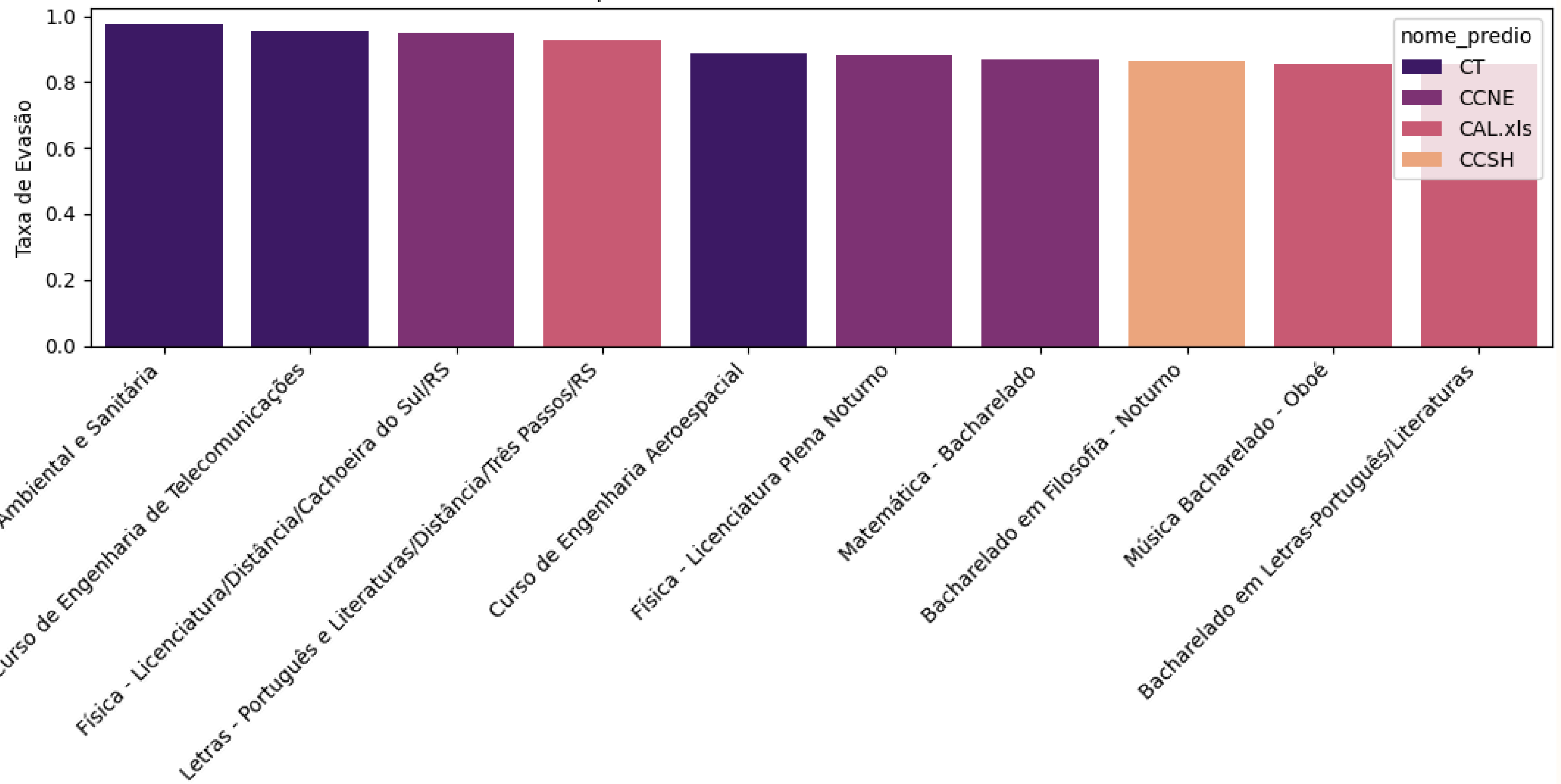
Analizando os resultados obtidos
percebemos uma concentração
maior de desistência nos prédios dos
cursos de exatas como o CT e CCNE

Extração dos Dados

Fórmula Usada:
 $1 - (\text{Formandos_Total} / \text{Ingressantes_Total})$

Utilizando os dados de ingressantes e formandos por ano, podemos obter uma taxa de evasão calculando a diferença entre quantos ingressantes e quantos formandos o curso teve. Quando mais perto de 1 a taxa de evasão, maior o numero de alunos ingressantes em relação ao numero de formandos

Top 10 Cursos com Maior Taxa de Evasão



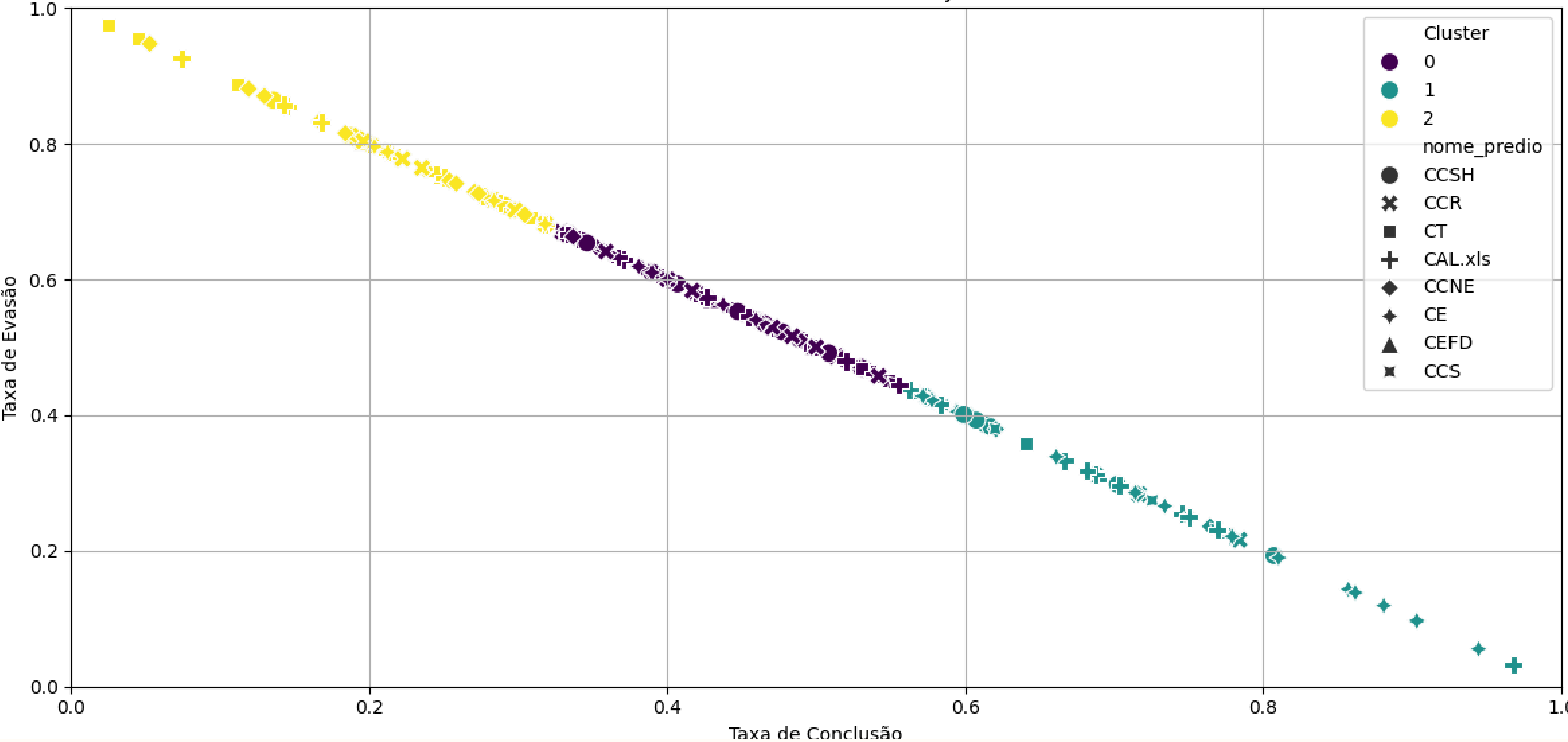
Clustering

Utilizando as técnicas de clustering vistas em aula, separamos os dados em três grupos:

- Grupo 0 (ROXO): Cursos com taxas equilibradas de ingresso e formação. Esses cursos têm números similares de ingressantes e formandos.
- Grupo 1 (AZUL): Cursos onde as taxas de conclusão são maiores que as de ingresso, ou seja, cursos com muitos formandos e poucos ingressantes.

Grupo 2 (AMARELO): Cursos com alta taxa de evasão, onde o número de ingressantes é maior do que o número de formandos.

Clusters de taxa de Conclusão e Evasão (Ajustado entre 0 e 1)



| | nome_predio | Cluster | Quantidade | Porcentagem |
|----|-------------|---------|------------|-------------|
| 0 | CAL.xls | 0 | 21 | 42.857143 |
| 1 | CAL.xls | 1 | 13 | 26.530612 |
| 2 | CAL.xls | 2 | 15 | 30.612245 |
| 3 | CCNE | 0 | 8 | 25.806452 |
| 4 | CCNE | 1 | 7 | 22.580645 |
| 5 | CCNE | 2 | 16 | 51.612903 |
| 6 | CCR | 0 | 12 | 52.173913 |
| 7 | CCR | 1 | 3 | 13.043478 |
| 8 | CCR | 2 | 8 | 34.782609 |
| 9 | CCS | 0 | 2 | 28.571429 |
| 10 | CCS | 1 | 5 | 71.428571 |
| 11 | CCSH | 0 | 15 | 50.000000 |
| 12 | CCSH | 1 | 9 | 30.000000 |
| 13 | CCSH | 2 | 6 | 20.000000 |
| 14 | CE | 0 | 20 | 42.553191 |
| 15 | CE | 1 | 21 | 44.680851 |
| 16 | CE | 2 | 6 | 12.765957 |
| 17 | CEFD | 0 | 2 | 66.666667 |
| 18 | CEFD | 2 | 1 | 33.333333 |
| 19 | CT | 0 | 6 | 40.000000 |
| 20 | CT | 1 | 3 | 20.000000 |
| 21 | CT | 2 | 6 | 40.000000 |

Conclusões

A análise dos dados revelou padrões interessantes sobre as taxas de evasão e conclusão nos cursos de diferentes centros. Os resultados indicam que:

1. Áreas de Exatas (CT e CCNE): Apresentam uma maior taxa de evasão, o que pode estar relacionado à dificuldade das disciplinas.

Análise de Agrupamento: O K-Means ajudou a identificar padrões em relação à taxa de evasão e conclusão, agrupando os cursos em três categorias com base nas suas características.



Agradeço!