Universidade Federal de Santa Maria

Análise dos Dados de Ingressantes e Formandos por Sexo (Diversos Centros) e Cursos com Maior Taxa de Reprovações e Associações;

Autores: Arthur Bogacki Veríssimo e Leandro Oliveira Galbarino do Nascimento

Santa Maria

2024

PARTE 1 -

Introdução

Nesta análise, investigamos os dados de ingressantes e formandos de diferentes cursos para verificar possíveis correlações entre esses dois fatores. Em particular, buscamos identificar se cursos com maior número de ingressantes apresentam também maior número de formandos, o que poderia sugerir um equilíbrio saudável. Por outro lado, se cursos com mais ingressantes não tiverem uma correspondência significativa em formandos, isso pode indicar uma alta taxa de evasão, possivelmente relacionada à dificuldade do curso.

Além disso, analisaremos os prédios com maior variação entre ingressantes e formandos, com o intuito de identificar áreas específicas mais afetadas pela desistência.

Pré-processamento dos Dados

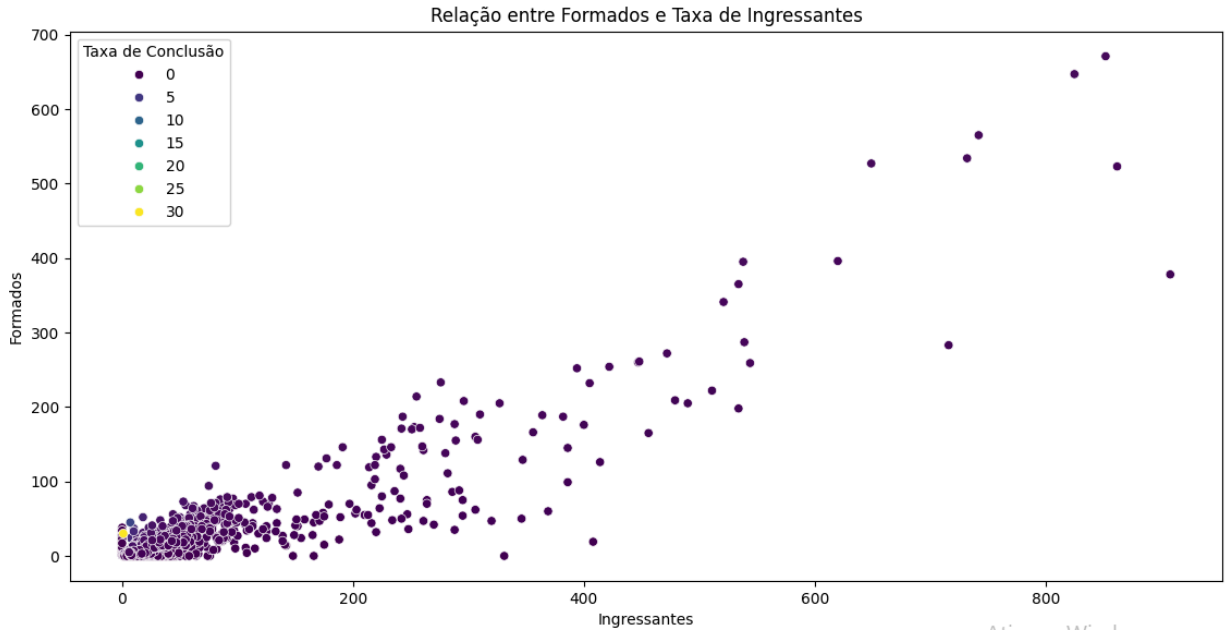
O primeiro passo do processo foi realizar o pré-processamento dos dados. Carregamos os dados de todas as tabelas, que inicialmente estavam separadas por prédio. Com isso, agrupamos todos os dados em uma única tabela, adicionando uma coluna chamada **“nome_predio”** para identificar de qual prédio o curso pertencia. Também calculamos a **Taxa de Conclusão**, utilizando a fórmula:

Taxa de Conclusão = Formandos / Ingressantes

Além disso, removemos a coluna **“NIVEL_CURSO”**, pois todos os cursos tinham o mesmo nível (graduação).

Gráfico: Relação entre Ingressantes e Formados

A seguir, um gráfico que visualiza a relação entre ingressantes e formados, destacando os cursos que possuem uma maior quantidade de ingressantes e formandos:



Cálculo da Taxa de Conclusão e Taxa de Evasão

Com os dados limpos e processados, o próximo passo foi calcular a **Taxa de Conclusão** e a **Taxa de Evasão**:

- **Taxa de Conclusão: Formandos_Total / Ingressantes_Total**

- **Taxa de Evasão: 1 - Taxa de Conclusão**

Além disso, optamos por **ignorar cursos com 0 ingressantes ou 0 formandos**. Isso foi feito para evitar distorções, já que esses dados podem representar erros de coleta ou cursos que já não existem mais.

Análise das Taxas de Evasão e Conclusão

A partir dos cálculos, focamos na identificação dos **cursos com maior taxa de evasão**. Analisando os dados e os gráficos gerados, observamos que cursos de áreas como **exatas** (CT e CCNE) possuem uma maior taxa de desistência. Isso pode estar relacionado a diversos fatores, como a maior dificuldade das disciplinas dessas áreas.

Outro ponto a ser observado foi o **Centro de Artes e Letras (CAL)**, que também apresentou taxas elevadas de evasão, possivelmente devido à desvalorização das artes no Brasil, entre outros fatores.

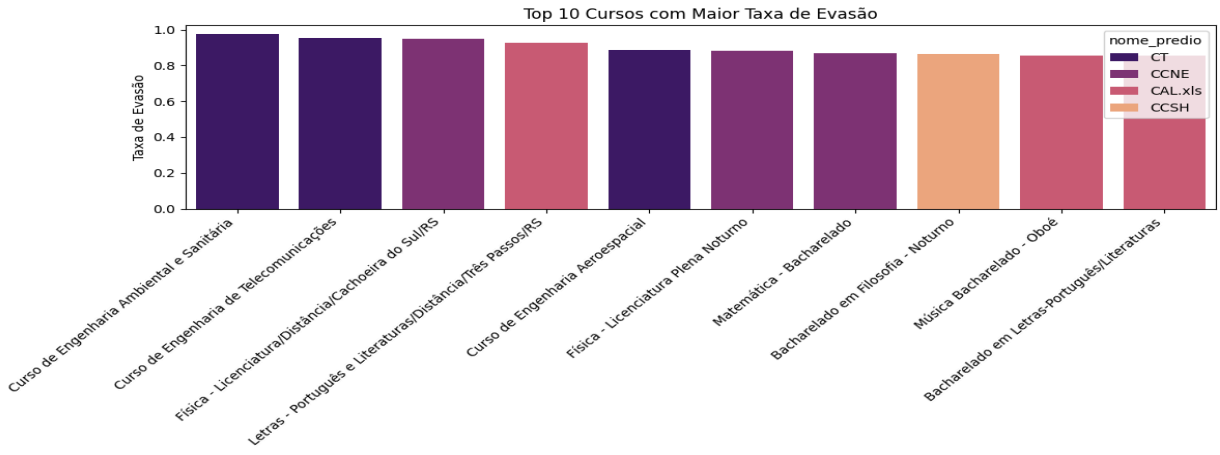
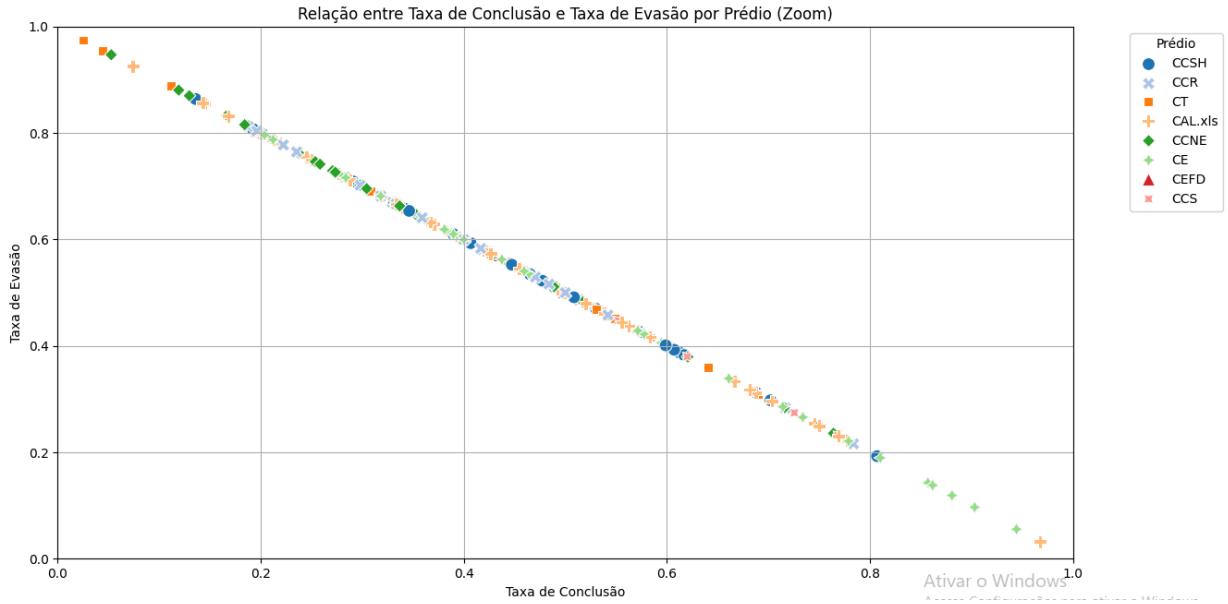


Gráfico de Dispersão: Taxa de Conclusão vs. Taxa de Evasão

No gráfico abaixo, cada símbolo desenhado representa um curso, e as cores e formas geométricas indicam os prédios. Ao analisar a dispersão dos pontos, é possível observar que cursos com taxa de evasão próxima de 1 (mais ingressantes que formados) estão concentrados em prédios como o **CT** e o **CCNE**:



Agrupamento com Algoritmo K-Means

Utilizamos o algoritmo **K-Means** para agrupar os cursos em três grupos, considerando aqueles que tiveram taxa de conclusão entre 0 e 1. O agrupamento gerou os seguintes 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.

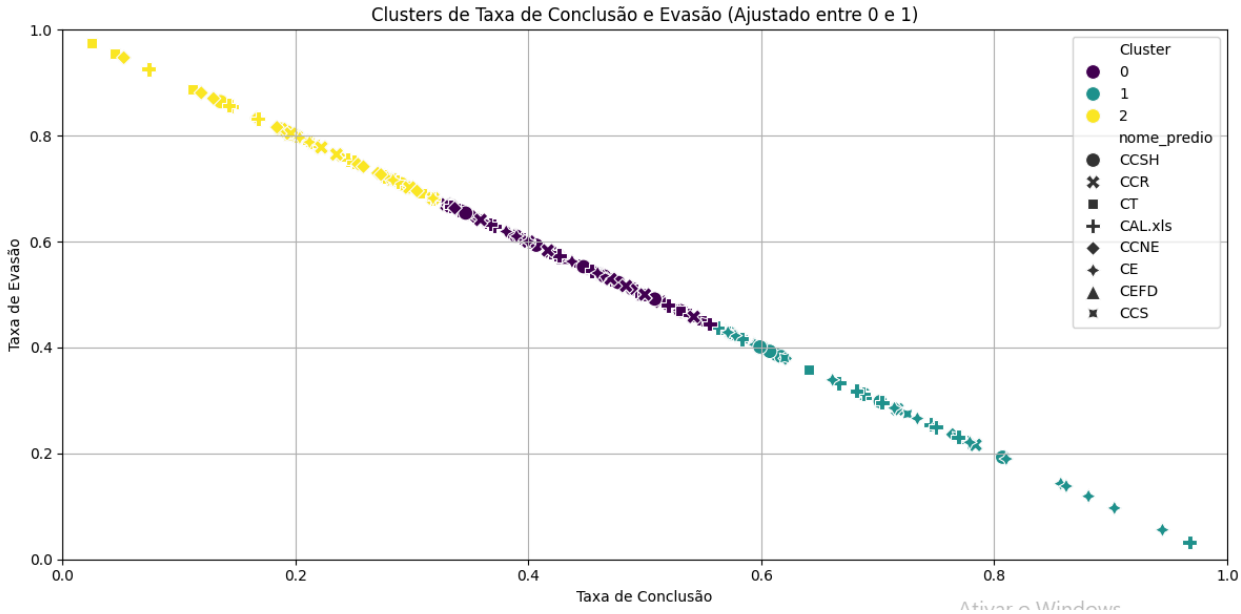


Tabela de Porcentagem por Prédio e Cluster

A tabela a seguir mostra a porcentagem de cursos de cada prédio que pertence a cada grupo de cluster. Podemos observar que o **CT** tem uma distribuição equilibrada entre os três grupos, porém com 40% dos cursos obtendo maiores taxas de evasão, enquanto o **CCNE** tem uma alta concentração de cursos no **Grupo 2**, indicando uma taxa de evasão mais alta.

	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.
2. **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.

Considerações Finais

Este estudo oferece uma visão importante sobre a relação entre ingressantes, formandos e evasão nos cursos, com implicações para o planejamento e políticas educacionais. A análise por clusters pode ajudar na identificação de cursos que exigem atenção especial em termos de taxa de evasão, e os resultados podem ser utilizados para estratégias de retenção de alunos.

PARTE 2 -

Cursos com Maior Taxa de Reprovações e Associações

1. Seleção dos Dados

Na etapa de análise dos dados presentes nos arquivos relacionados às disciplinas, surgiu a 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. A taxa de reprovação foi obtida pela divisão do número de alunos reprovados pelo número total de alunos matriculados na disciplina, considerando todos os anos de cada curso específico.

Para a associação, seriam utilizadas regras de associação para verificar se, quando um curso apresenta uma alta taxa de reprovação em uma determinada matéria(Matéria A), isso indica uma alta taxa de reprovação também em outra matéria (Matéria B). Ou seja, seria possível identificar se há uma correlação entre as dificuldades em diferentes disciplinas, permitindo uma análise mais aprofundada sobre os cursos e as matérias que apresentam maiores desafios para os alunos.

2. Pré-Processamento

Para realizar esse processo, o primeiro passo foi combinar todos os arquivos Excel em um único DataFrame. Essa etapa foi essencial para consolidar os dados de diferentes anos e disciplinas, permitindo uma análise abrangente e consistente.

	Ano	Semestre	Cód. Disciplina	Cód. Turma	...	Alunos	Professor	Cód. Curso	Curso
0	2021	1. Semestre	D200888	10	...	5	P888402148	139	Curso de Bacharelado em Estatística - Noturno
1	2021	1. Semestre	D200888	10	...	1	P888402148	139	Curso de Bacharelado em Estatística - Noturno
2	2021	1. Semestre	D200888	10	...	27	P888402148	139	Curso de Bacharelado em Estatística - Noturno
3	2021	1. Semestre	D200888	11	...	17	P148855320	521	Ciências Econômicas - Diurno
4	2021	1. Semestre	D200888	11	...	37	P148855320	521	Ciências Econômicas - Diurno
...	...	...	...	...	...	...	...	...	...
1519	2022	2. Semestre	D888200	SI2	...	1	E681200461218	314	Bacharelado em Sistemas de Informação
1520	2022	2. Semestre	D888200	SI2	...	3	E681200461218	314	Bacharelado em Sistemas de Informação
1521	2022	2. Semestre	D888200	SI2	...	2	E681200461218	314	Bacharelado em Sistemas de Informação
1522	2022	2. Semestre	D888200	SI3	...	8	E200380461148	314	Bacharelado em Sistemas de Informação
1523	2022	2. Semestre	D888200	SI3	...	2	E200380461148	314	Bacharelado em Sistemas de Informação

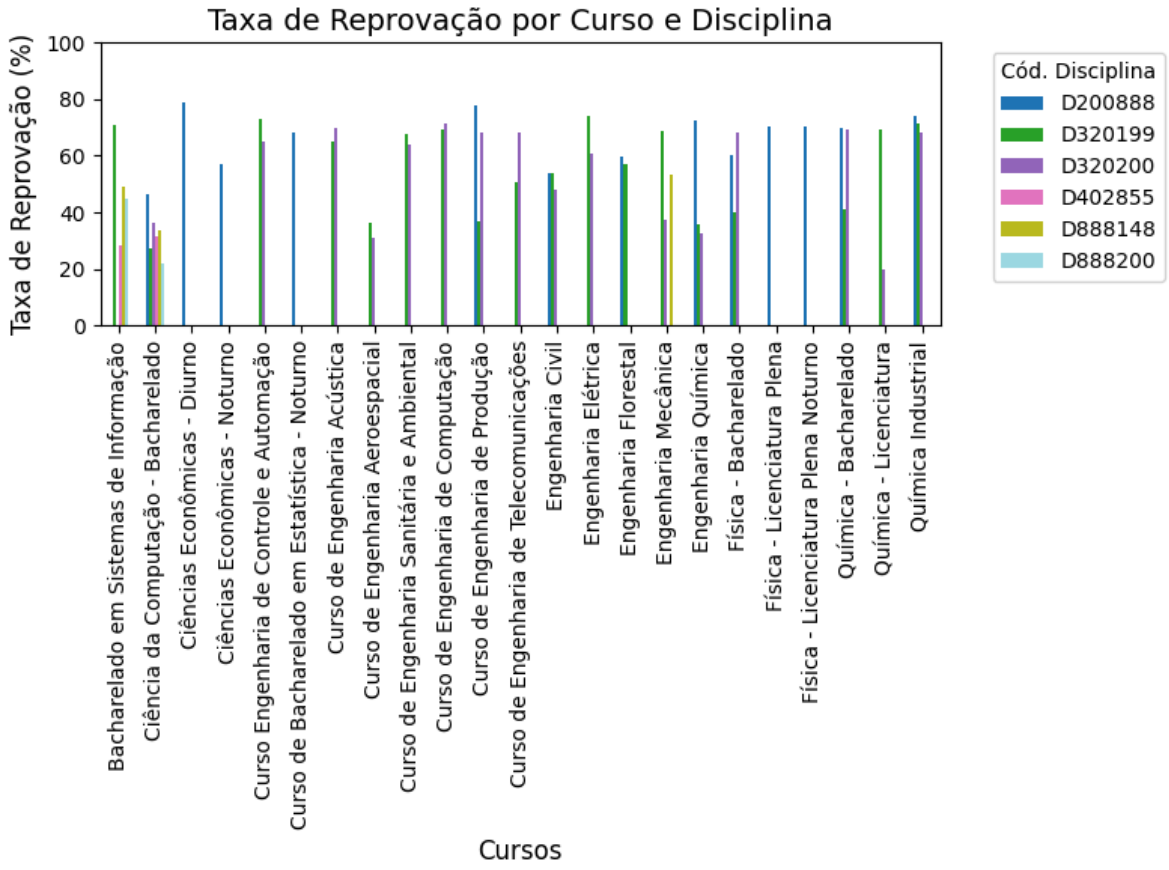
A próxima etapa consistiu na remoção de colunas que não contribuiriam para a análise. As colunas "Cód. Curso", "Professor" e "%" foram eliminadas, pois já existia a coluna "Curso", que contém o nome completo do curso, tornando a coluna de código redundante. A coluna "Professor" foi removida, pois não seria utilizada na nossa mineração em específico. A coluna "%" foi descartada porque seria necessário criar uma outra variável, a taxa de reprovação, que seria calculada pela divisão entre o número de reprovados e o total de alunos. No entanto, a coluna "%" não seria de grande ajuda nesse cálculo.

3. Transformação dos Dados

Em seguida, realizamos a normalização dos dados, consolidando todas as categorias relacionadas à reprovação. A coluna Situação, presente no DataFrame, contém os seguintes valores possíveis: Aprovado, Reprovado, Tr. Parcial, Não Concl., Dispensado, CancMatricula e Repr.Freq. 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. Inicialmente, essa coluna refletia apenas o número total de alunos matriculados em uma disciplina, considerando o ano e semestre específicos. Posteriormente, realizamos o agrupamento para que o valor representasse a soma do total de alunos de todos os anos de um curso. Com isso, foi possível calcular a taxa de reprovação, que é o número de alunos reprovados na disciplina dividido pelo número total de alunos inscritos na matéria, para cada curso específico.

	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



4. Mineração dos Dados

Na etapa anterior, conseguimos identificar quais cursos apresentavam as maiores taxas de reprovação. Isso nos permitiu observar em quais disciplinas os cursos enfrentam maior dificuldade.

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

Além dessa análise, 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).

Usando os algoritmos Apriori e association_rules da biblioteca mlxtend do Python, conseguimos gerar as regras de associação e encontramos as seguintes regras:

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

5. 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. As regras geradas foram do tipo: Antecedente: (D320200) (disciplina A com alta taxa de reprovação), Consequente: (D320199) (disciplina B com alta taxa de

reprovação). Isso significa que, 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.

A confiança (Confidence) e o lift são duas métricas importantes para interpretar essas regras. A confiança indica a probabilidade de o consequente ocorrer dado que o antecedente já ocorreu. Por exemplo, 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. Um lift maior que 1 sugere que existe uma associação positiva entre as disciplinas, ou seja, elas ocorrem juntas mais frequentemente do que seria esperado por acaso.

Em resumo, as regras de associação não indicam causalidade, mas sim correlação. Elas mostram que cursos com alta taxa de reprovação em uma disciplina tendem a ter alta taxa de reprovação em outras disciplinas associadas, com base nos dados observados.