**Cursos com Maior Taxa de Reprovações e Associações**

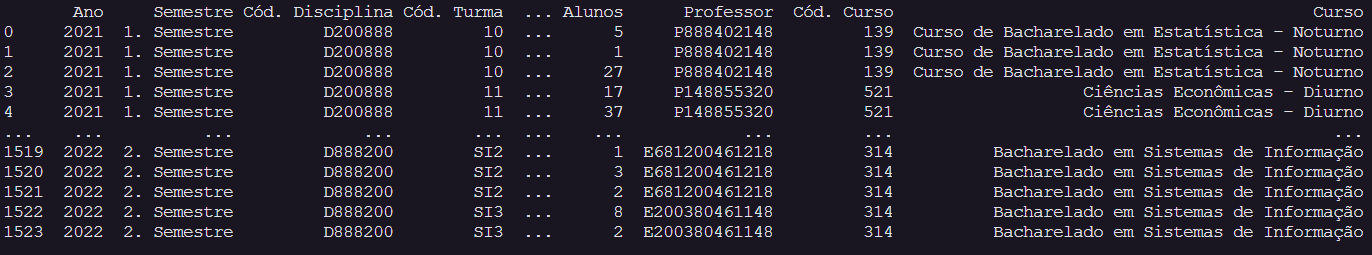
1. **Seleção dos Dados**

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.

1. **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.

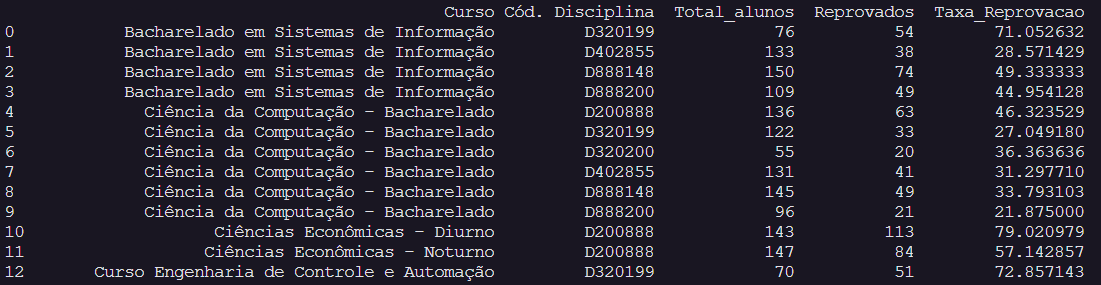


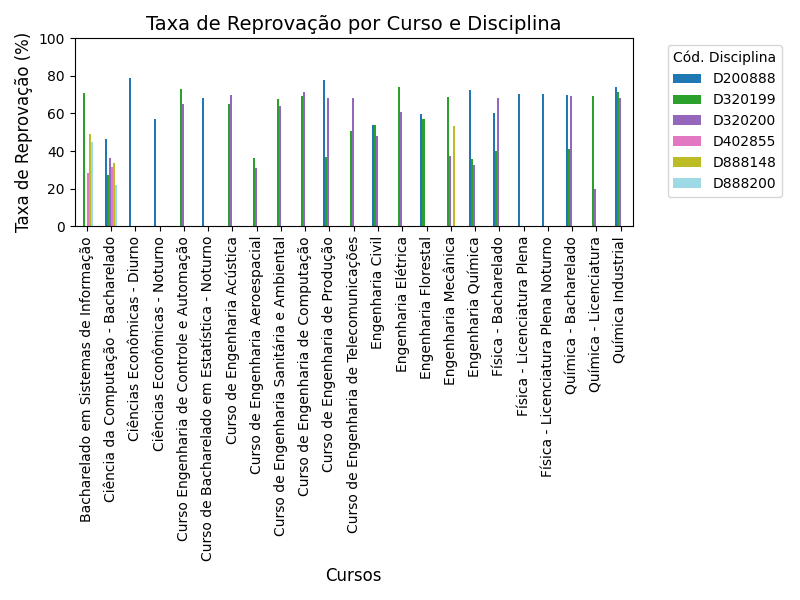
A próxima etapa consistiu na remoção de colunas que não contribuiriam para a análise. As colunas "Cód. Curso", "Professor" e "%" foram eliminadas, pois já existia a coluna "Curso", que contém o nome completo do curso, tornando a coluna de código redundante. A coluna "Professor" foi removida, pois não 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.

1. **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.

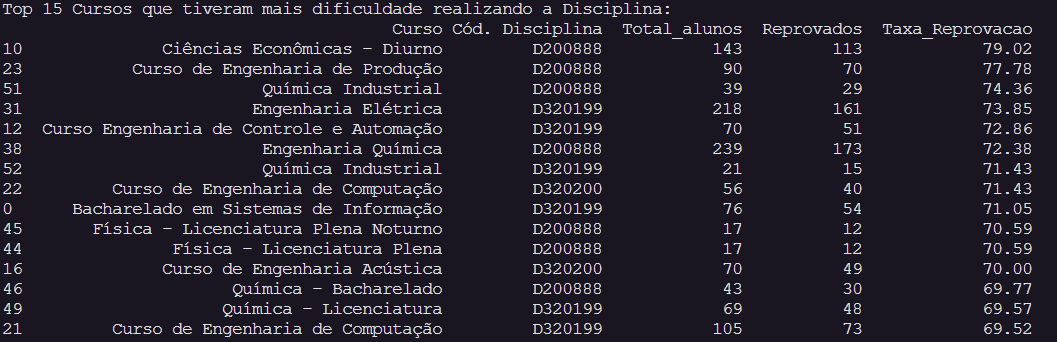
Além disso, criamos uma coluna que exibia o número total de alunos em cada disciplina e turma. Inicialmente, essa coluna refletia apenas o número total de alunos matriculados em uma disciplina, considerando o ano e semestre específicos. Posteriormente, realizamos o agrupamento para que o valor representasse 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.





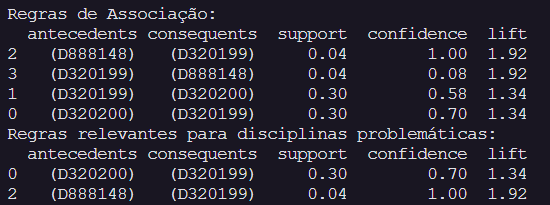
1. **Mineração dos Dados**

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.



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:



### **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.