ELC 1098 - Mineração de Dados

Trabalho 2

Alunos: Augusto Kist Lunardi, Davi de Castro Machado, João Pedro Azenha Righi e Jonathan Weber Nogueira - Universidade Federal de Santa Maria - Centro de Tecnologia - Professor Dr. Joaquim Assunção

Link para o repositório: https://github.com/JonathanWNogueira/data_mining

Preparação dos dados

Na fase de **Seleção dos Dados**, foram escolhidos dois conjuntos de dados principais relacionados à UFSM: um sobre disciplinas e outro sobre os centros. O primeiro conjunto contém informações detalhadas sobre disciplinas, incluindo código da disciplina, código da turma, alunos matriculados e professores responsáveis. O segundo conjunto está relacionado ao fluxo de alunos nos centros da UFSM, abrangendo dados como o número de ingressantes e formados (egressos), desagregados por sexo. O objetivo principal do trabalho é elaborar um plano analítico para identificar padrões e obter insights a partir dos dados selecionados.

Nas etapas de **Pré-Processamento e Transformação**, os dados estavam distribuídos em diversos arquivos nos formatos .xls e .xlsx. Inicialmente, os arquivos do conjunto de dados dos centros foram renomeados para corresponder ao nome de seus respectivos centros, com o intuito de simplificar a manipulação e organização. Além disso, foi necessário corrigir nomes que estavam em formato Unicode para texto normal, garantindo a legibilidade e consistência dos arquivos.

Em seguida, foi realizado o processo de padronização de formato, convertendo todos os arquivos para o formato .xlsx. Durante essa conversão, foi identificado que o arquivo **CE.xlsx** já estava no formato correto, mas, como continha os mesmos dados da sua versão .xls, ele foi sobrescrito. Os demais arquivos em formato .xls foram convertidos para .xlsx sem problemas.

Após a padronização, o próximo passo foi consolidar as informações. As planilhas de um mesmo tipo foram agrupadas em um único arquivo, criando tabelas gerais para cada conjunto de dados. Por exemplo, todas as planilhas relacionadas às disciplinas foram unificadas em uma tabela principal. Para manter a rastreabilidade e identificar a origem dos dados, foi adicionada uma coluna indicando o arquivo de origem. Esse processo de unificação foi realizado tanto para os dados das disciplinas quanto para os dados dos centros, facilitando as análises subsequentes.

Nesse contexto, foram utilizadas algumas estratégias de Mineração de Dados para a extração de informações potencialmente úteis a partir do *dataset* disponibilizado.

Técnica de Regressão Linear

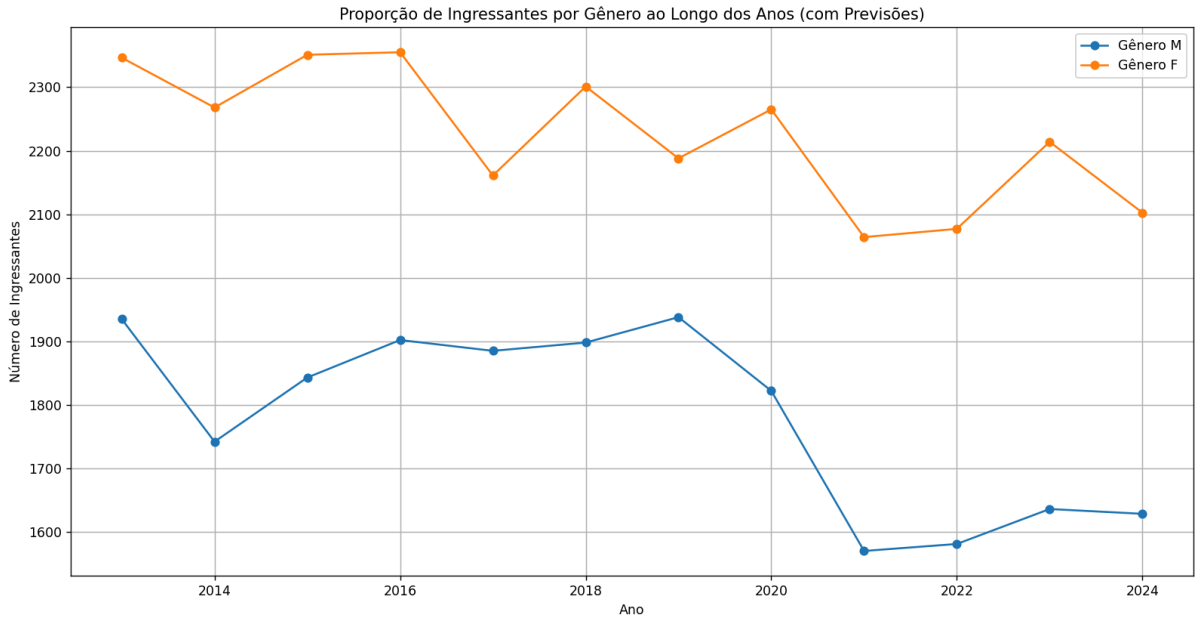
A regressão linear é uma possíveis técnicas utilizadas durante o processo de mineração de dados para a extração de informações potencialmente úteis e previamente desconhecidas. Essa estratégia foi utilizada para modelar a relação linear entre uma variável dependente (aquela que será prevista) e uma ou mais variáveis independentes - aquelas que potencialmente influenciam na variável a ser prevista.

Sendo assim, utilizamos a técnica de regressão linear em séries temporais visto que acreditamos que é uma forma interessante para identificar tendências e padrões ao longo do tempo, podendo ser utilizada tanto potencialmente para dados faltantes quanto para a previsão de dados futuros.

A ideia geral da regressão linear é funcionar da seguinte forma:

- Primeiro, o algoritmo de regressão linear encontra a melhor reta que se ajusta aos dados históricos, ou seja, a reta que minimiza a diferença entre os valores reais e os valores previstos pela reta.
- Após isso, o modelo ajustado pode ser utilizado para prever valores futuros da variável dependente, simplesmente fornecendo novos valores para a variável independente.

Exemplo de utilização:

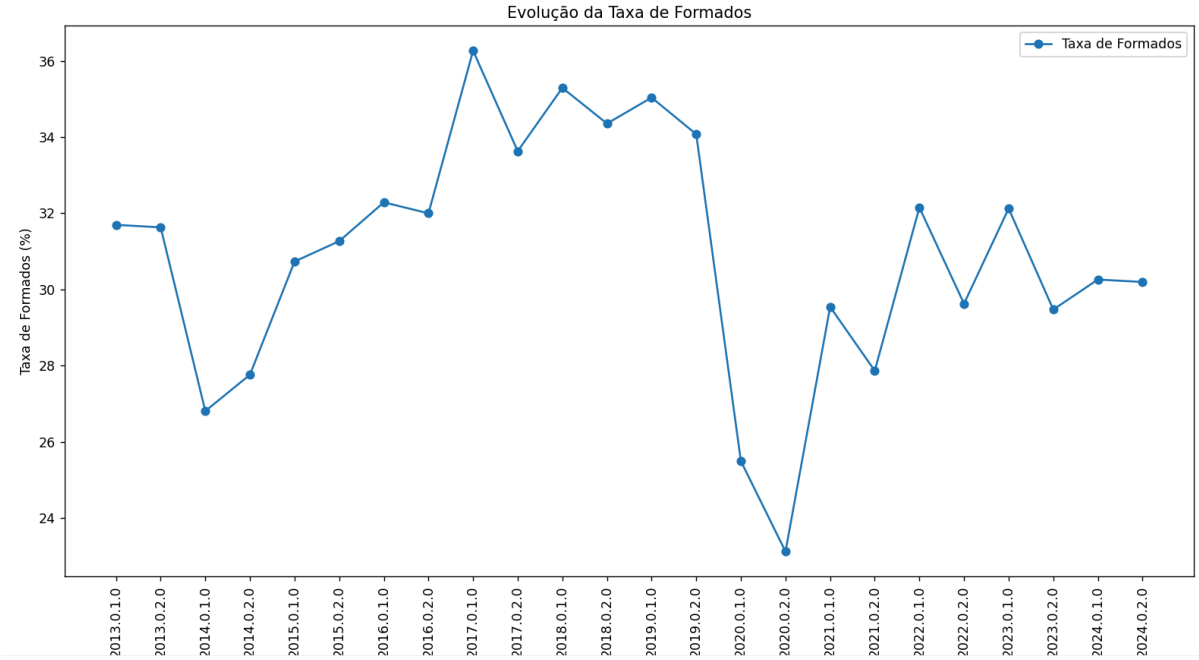


Nesse caso, foram utilizados os dados disponíveis do número total de ingressantes, por gênero durante o período de 2014 à 2023 para estimar essa quantidade no ano de 2013, bem como prever a potencial quantidade de ingressantes (também por gênero em 2024).

Esse tipo de informação é potencialmente útil visto que ele faz uma previsão da quantidade de ingressantes, por gênero, na UFSM. Isso faz com que o desenvolvimento de políticas afirmativas para mulheres ou para a preparação de infraestrutura no geral para os campis seja feita de forma mais planejada, por exemplo. Facilitando a preparação da administração e viabilizando uma maior organização da universidade.

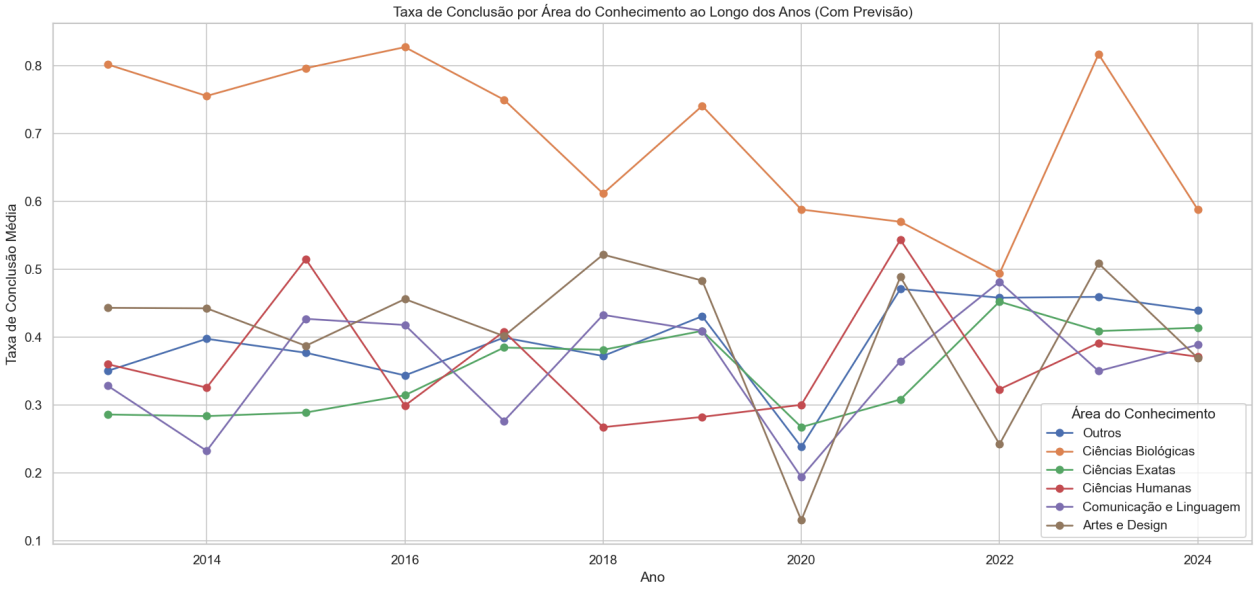
No entanto, é de suma importância salientar que esse tipo de abordagem possui limitações que precisam ser levadas em consideração. E a principal delas seria que esse tipo de modelo é sensível a outliers, ou seja, pontos discrepantes podem influenciar significativamente os resultados. Entretanto, esse problema tende a diminuir conforme a amostragem dos dados aumenta.

Nesse outro caso, o gráfico mostra informações, por semestre, em relação à taxa de formandos ao longo dos anos. Sendo acrescentados dados para 2013 e 2024 utilizando a técnica de regressão linear.



Nesse caso, é possível perceber uma queda abrupta do número de formandos em 2020 que deve-se, provavelmente, à pandemia na época. Mas desde então as taxas aumentaram e oscilaram de forma mais branda. E a partir dessas informações estabeleceu-se as previsões citadas anteriormente.

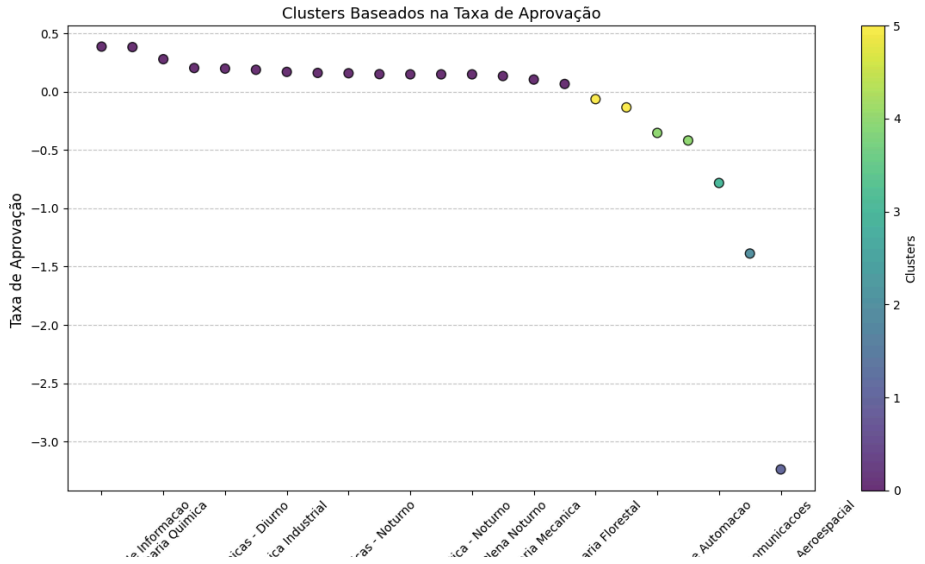
Outro gráfico mais complexo pode ser observado a seguir que relaciona a taxa de conclusão a áreas do conhecimento.



Técnica de Agrupamento: K - means

O K-Means é um algoritmo amplamente utilizado para agrupamento de dados em tarefas de aprendizado não supervisionado. Seu principal objetivo é dividir um conjunto de dados em grupos ou clusters, onde cada ponto de dados pertence ao cluster cujo centróide está mais próximo. Essa proximidade geralmente é calculada usando a distância euclidiana, embora outras métricas possam ser aplicadas dependendo do contexto. Para esta análise, os dados foram previamente normalizados, garantindo que todas as variáveis tivessem a mesma escala e pudessem ser manipuladas adequadamente pelo algoritmo. Além disso, o script utilizado exibe no terminal os dados mais específicos de cada cluster, fornecendo uma visão detalhada dos resultados obtidos.

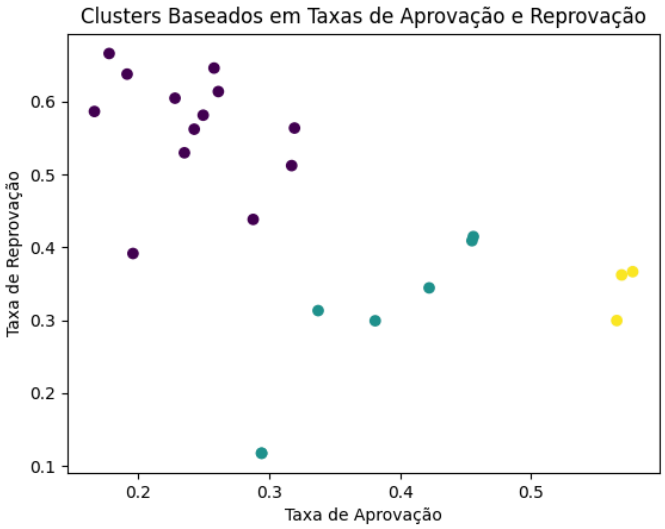
No gráfico apresentado abaixo, os dados foram agrupados em 5 clusters, organizados de forma que cursos com taxas de aprovação semelhantes fossem alocados no mesmo grupo. A análise revelou que a maioria das turmas está no cluster 0, caracterizado por taxas de aprovação positivas (mais aprovados do que reprovados). Por outro lado, os demais clusters representam cursos com predominância de reprovações, com os grupos sendo diferenciados pela distância nas taxas de aprovação.



Relação entre cursos e taxa de aprovação.

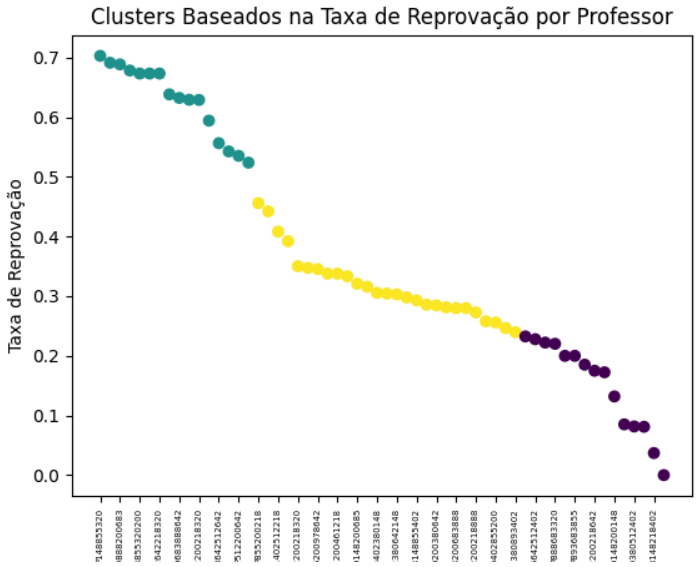
A visualização oferece insights valiosos sobre o desempenho geral dos cursos, permitindo identificar grupos que demandam maior atenção em termos de estratégias pedagógicas ou alocação de recursos.

Outra análise explorou a relação entre as taxas de aprovação e reprovação dos cursos. O agrupamento permitiu observar como essas taxas variam em diferentes contextos e identificar padrões específicos de desempenho acadêmico, como mostrado no gráfico abaixo. Por exemplo, clusters com altas taxas de reprovação podem destacar dificuldades comuns entre os cursos, enquanto clusters com altas taxas de aprovação apontam para áreas de bom desempenho que podem ser usadas como referência ou modelo.



Relação entre taxa de aprovação x reprovação dos cursos.

Por fim, foi analisada a relação entre os professores e as taxas de reprovação. O K-Means agrupou os professores com base no número de reprovações associadas às suas turmas, evidenciando aqueles cujos resultados podem requerer maior atenção. Essa abordagem possibilita ações direcionadas, como treinamentos, revisões de conteúdo e intervenções específicas para melhorar os resultados acadêmicos.



Relação entre os professores e taxa de reprovação.

O uso do K-Means para análise de dados educacionais permite identificar padrões complexos de maneira acessível e visual. Os agrupamentos resultantes revelam insights importantes que podem orientar intervenções específicas, ajudando instituições a melhorar a qualidade do ensino e o desempenho dos estudantes. Ao associar taxas de aprovação, reprovação e variáveis relacionadas, essa técnica fornece uma base robusta para a tomada de decisões estratégicas no contexto educacional. Além disso, a normalização dos dados e a exibição detalhada no terminal tornam o processo mais preciso e transparente para futura análise.

Técina de Apriori

O Apriori foi utilizado para identificar padrões frequentes e gerar regras que associam itens com alta probabilidade de ocorrerem juntos. Sendo assim, a ideia foi identificar padrões de comportamento entre as disciplinas, turmas e resultados dos alunos.

O algoritmo Apriori foi utilizado para gerar regras de associação entre as variáveis (cursos e status de aprovação). Nesse contexto, a confiança e o suporte foram as duas métricas principais utilizadas para classificar as regras geradas.

Observação: O suporte mede a frequência com que uma combinação específica de itens ocorre nos dados, enquanto a confiança avalia a probabilidade de uma situação ocorrer dado um conjunto específico de itens.

Sendo assim, para filtrar regras que pudessem ser interessantes, foram utilizados valores mínimos de 0.3 para confiança e 0.005 para suporte.

O valor de ao menos 0.3 para confiança nos fornece combinações entre professor e curso que quando ocorrem, resultam em um estado ou de aprovação, ou de reprovação, com grau de confiança razoável, superior a 30%.

Para o suporte, foi necessário utilizar um valor tão baixo como 0.005 pois existe um número muito grande de combinações diferentes dentro deste mesmo conjunto de dados. Dessa forma, não existiam casos em que uma combinação de "professor+curso -> situação" ocorresse em mais do que 0,5% das vezes.

Resultados Obtidos:

Regras com Maior Confiança

As regras com maior confiança indicam a probabilidade de o consequente ocorrer, dado que o antecedente esteja presente. Após a execução do algoritmo, as três regras com maior confiança foram:

- Engenharia Química - A292200218320 → Reprovado (Confiança: 66,67%): Alta confiança na reprovação para os alunos deste curso com este professor, indicando a necessidade de investigar fatores como currículo e carga de trabalho.

- Engenharia Aeroespacial - A292200218320 → Aprovado (Confiança: 66,67%): Curiosamente, um outro curso, com o mesmo professor, mas com uma confiança alta para aprovação. Isso poderia indicar que os alunos desse curso têm uma boa taxa de aprovação, o que pode ser interessante para destacar pontos positivos sobre o curso. Além disso, por ser com o mesmo professor, poderia ser analisada a possibilidade de que este professor favorece alunos do curso de Engenharia Aeroespacial, e desfavorece os alunos de Engenharia Química. Entretanto, não podemos afirmar isso apenas com estes dados analisados, visto que somente a taxa de reprovação e aprovação deste professor com os alunos dos cursos não prova isso. Além disso, não são numerosos o suficiente os casos analisados. Embora o suporte seja baixo, essas duas regras ainda são bastante

interessantes e permitem a criação de hipóteses que podem ser investigadas.

- Engenharia Química - P380148200 → Reprovado (Confiança: 66,67%): Esta é outra regra que é interessante, porque reforça a alta taxa de reprovação dos alunos deste curso, mesmo sendo com um professor diferente.

antecedents	consequents	confidence
frozenset({'Curso de Engenharia Aeroespacial - A292200218320'})	frozenset({'Aprovado'})	0.6666666666666666
frozenset({'Engenharia Química - A292200218320'})	frozenset({'Reprovado'})	0.6666666666666666
frozenset({'Engenharia Química - P380148200'})	frozenset({'Reprovado'})	0.6666666666666666
frozenset({'Ciência da Computação - Bacharelado - C218200683888'})	frozenset({'Aprovado'})	0.5
frozenset({'Ciência da Computação - Bacharelado - B200380461148'})	frozenset({'Aprovado'})	0.5
frozenset({'Engenharia Elétrica - B888642512642'})	frozenset({'Aprovado'})	0.5
frozenset({'Química Industrial - P380148200'})	frozenset({'Aprovado'})	0.5
frozenset({'Curso de Engenharia Aeroespacial - B200380461148'})	frozenset({'Reprovado'})	0.5
frozenset({'Engenharia Elétrica - B888642512642'})	frozenset({'Reprovado'})	0.5
frozenset({'Bacharelado em Sistemas de Informação - C402380642148'})	frozenset({'Aprovado'})	0.5
frozenset({'Bacharelado em Sistemas de Informação - E200380461148'})	frozenset({'Aprovado'})	0.5

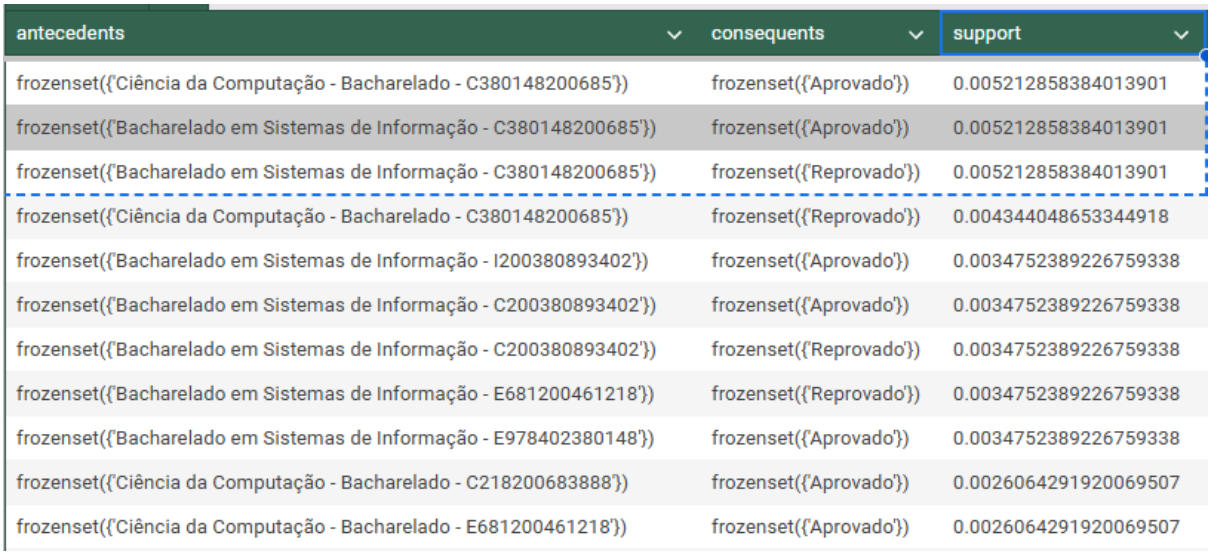
Regras com Maior Suporte

As regras com maior suporte indicam as combinações de antecedentes e consequentes que ocorreram com maior frequência no conjunto de dados. Além disso, as três regras com maior suporte mostraram as seguintes associações:

- Bacharelado em Sistemas de Informação - C380148200685 → Aprovado (Suporte: 0.0521): Alta incidência de aprovação entre os alunos desse curso com esse professor.

- Ciência da Computação - Bacharelado - C380148200685 → Aprovado (Suporte: 0.0521): Similar ao Bacharelado em Sistemas de Informação, com uma boa taxa de aprovação com o mesmo professor.

- Bacharelado em Sistemas de Informação - C200380893402 → Aprovado (Suporte: 0.0521): Também observa uma boa taxa de aprovação nesse curso com este outro professor.



antecedents	consequents	support
frozenset({'Ciência da Computação - Bacharelado - C380148200685'})	frozenset({'Aprovado'})	0.005212858384013901
frozenset({'Bacharelado em Sistemas de Informação - C380148200685'})	frozenset({'Aprovado'})	0.005212858384013901
frozenset({'Bacharelado em Sistemas de Informação - C380148200685'})	frozenset({'Reprovado'})	0.005212858384013901
frozenset({'Ciência da Computação - Bacharelado - C380148200685'})	frozenset({'Reprovado'})	0.004344048653344918
frozenset({'Bacharelado em Sistemas de Informação - I200380893402'})	frozenset({'Aprovado'})	0.0034752389226759338
frozenset({'Bacharelado em Sistemas de Informação - C200380893402'})	frozenset({'Aprovado'})	0.0034752389226759338
frozenset({'Bacharelado em Sistemas de Informação - C200380893402'})	frozenset({'Reprovado'})	0.0034752389226759338
frozenset({'Bacharelado em Sistemas de Informação - E681200461218'})	frozenset({'Reprovado'})	0.0034752389226759338
frozenset({'Bacharelado em Sistemas de Informação - E978402380148'})	frozenset({'Aprovado'})	0.0034752389226759338
frozenset({'Ciência da Computação - Bacharelado - C218200683888'})	frozenset({'Aprovado'})	0.0026064291920069507
frozenset({'Ciência da Computação - Bacharelado - E681200461218'})	frozenset({'Aprovado'})	0.0026064291920069507

Portanto, o Apriori permitiu identificar padrões de aprovação e reprovação em alguns cursos, proporcionando *insights* importantes sobre a relação entre as disciplinas e os resultados dos alunos. As regras de maior confiança destacam cursos com alta probabilidade de reprovação ou aprovação, enquanto as de maior suporte indicam cursos com maior frequência de sucesso entre os alunos. Esses resultados podem ser usados para investigar mais a fundo os fatores que influenciam o desempenho acadêmico e ajustar currículos ou estratégias de ensino, por parte da UFSM e corpo docente.

Avaliação e conclusão

As técnicas utilizadas no projeto possibilitaram a extração de informações relevantes e previamente desconhecidas a respeito de cursos e alunos da UFSM. Através da regressão linear, foi possível prever com razoável precisão o número de ingressantes por gênero nos próximos anos, além de identificar tendências nas taxas de formação. A análise dos dados revelou um impacto significativo da pandemia de 2020, que refletiu na queda do número de formandos, mas também indicou uma recuperação nos anos subsequentes. Essas informações são cruciais para o planejamento estratégico da universidade.

A utilização do K-Means para a segmentação dos dados, agrupando cursos e turmas com base nas taxas de aprovação e reprovação, forneceu uma visão clara das áreas com melhores e piores desempenhos. A separação dos dados em clusters ajudou a identificar padrões específicos, como a concentração de turmas com altas taxas de aprovação em um grupo e turmas com maior índice de reprovação em outros. Esse tipo de análise facilita a tomada de decisões em relação à melhoria de cursos ou intervenções em turmas com resultados abaixo da média.

Por fim, o uso do algoritmo Apriori para a geração de regras de associação possibilitou a identificação de padrões frequentes entre disciplinas, cursos e resultados de aprovação. As regras geradas destacaram não apenas os cursos com maior taxa de reprovação, como também aqueles com melhores taxas de aprovação, oferecendo *insights* valiosos para uma possível revisão curricular ou abordagem pedagógica. Além disso, ao focar na confiança e suporte como métricas de avaliação, foi possível filtrar as combinações mais significativas e relevantes, contribuindo para a personalização da análise.

Portanto, através desse trabalho foi possível observar que a mineração de dados pode ser uma ferramenta útil para identificar padrões e tendências em grandes volumes de dados acadêmicos. Embora os modelos utilizados apresentem boas previsões, é importante ressaltar que as limitações dos métodos, como a sensibilidade a outliers na regressão linear e a escolha adequada dos parâmetros para o algoritmo Apriori, precisam ser consideradas para garantir a robustez das conclusões.