

MINERAÇÃO DE DADOS EDUCACIONAIS APLICADA AO CURSO DE ENGENHARIA DE SOFTWARE DA UDESC

¹FERNANDO SANTOS, ²CESAR AUGUSTO FONTICIELHA DE ROSE

¹Universidade do Estado de Santa Catarina (UDESC),

²Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)

<fernando.santos@udesc.br>, <cesar.deroze@pucrs.br>

DOI: 10.21439/conexoes.v19.3882

Resumo. Este trabalho apresenta um estudo de caso com aplicação de mineração de dados educacionais para analisar as matrículas dos acadêmicos do curso de Bacharelado em Engenharia de Software da UDESC. O objetivo é coletar indicadores do curso sobre matrículas, conclusões e evasões, e verificar combinações de disciplinas cursadas simultaneamente que possam maximizar o potencial de aprovação dos acadêmicos. O estudo utilizou dados obtidos do sistema acadêmico da UDESC contendo registros de matrículas desde a implantação do curso em 2014. Gráficos interativos foram criados para visualizar os indicadores do curso. O algoritmo Apriori foi utilizado para obter conjuntos de matrículas frequentes e regras de associação, que foram utilizadas para desenvolver uma prova de conceito de sistema de recomendação de disciplinas. Entre os resultados obtidos neste trabalho, destaca-se que os indicadores revelaram acumulados de conclusão e evasão no curso de 9% e 60,74% respectivamente, e que 77,33% dos formados levam entre 8 e 10 semestres para concluir o curso, sendo que as maiores frequências de reprovações são em disciplinas de programação das fases iniciais e nas matemáticas. O algoritmo Apriori obteve 7.484 regras de associação, contemplando diversas combinações de disciplinas onde houve aprovação. A prova de conceito de sistema foi capaz de recomendar disciplinas, evidenciando a viabilidade de gerar recomendações a partir de regras de associação.

Palavras-chave: ciência de dados; mineração de dados; regras de associação; matrículas.

EDUCATIONAL DATA MINING APPLIED TO THE SOFTWARE ENGINEERING COURSE AT UDESC

Abstract. This paper presents a case study with the application of educational data mining to analyze the enrollments of students on the Bachelor's degree in Software Engineering at UDESC. The goal is to gather course-related indicators on enrollments, graduations, and dropouts, as well as to identify combinations of courses taken simultaneously that may maximize students' likelihood of success. The study utilized data extracted from UDESC's academic system, which contains enrollment records dating back to the program's implementation in 2014. Interactive charts were created to visualize course indicators. The *Apriori* algorithm was employed to identify frequent enrollment sets and association rules, which were then used in the development of a proof-of-concept course recommendation system. Among the results, it was found that the graduation and dropout rates are 9% and 60.74%, respectively. Additionally, 77.33% of graduates complete the program in 8 to 10 semesters, with the highest failure rates occurring in early-stage programming and mathematics courses. The *Apriori* algorithm generated 7,484 association rules, encompassing various combinations of successfully completed courses. The proof-of-concept system successfully recommended courses, demonstrating the feasibility of generating recommendations based on association rules.

Keywords: data science; data mining; association rules; enrollments.



1 INTRODUÇÃO

A Universidade do Estado de Santa Catarina (UDESC) oferece, desde 2014, o curso de Bacharelado em Engenharia de Software (ESO). O curso de ESO é integralizado, de forma regular, em 8 semestres. Entretanto, reprovações em disciplinas frequentemente aumentam o tempo de integralização. Esse fato não é exclusividade do curso de ESO. Segundo Ferreyra *et al.* (2017), na América Latina o tempo de integralização é 36% maior do que a duração regular do curso. Além do tempo de integralização, outro aspecto pertinente é a evasão acadêmica. De acordo com o Instituto Semesp, a evasão acumulada nos cursos de tecnologia da informação é de 65,5% no Brasil (SEMESP, 2023). Até o momento, o Núcleo Docente Estruturante (NDE) do curso não dispõe de indicadores sobre aprovação em disciplinas, tempo de integralização e evasões.

Ao reprovar em disciplina(s), a trajetória do acadêmico é modificada. Entende-se por trajetória de aprendizagem o caminho percorrido pelo estudante ao progredir em direção a um objetivo educacional (Carmo; Gasparini; Oliveira, 2022). No contexto do ensino superior, o objetivo educacional é adquirir as habilidades e competências do perfil do curso, e a trajetória é entendida pelo cumprimento das disciplinas previstas ao longo das fases do curso (Garcia; Gomes, 2020). Mudanças na trajetória por vezes estão relacionadas à evasão acadêmica (Carmo; Gasparini; Oliveira, 2022). Ao se matricular semestralmente, o acadêmico com alguma reprovação escolhe a lista de disciplinas que irá cursar considerando os pré-requisitos e a disponibilidade de horários. A combinação de disciplinas matriculadas, nestes casos, pode diferir da combinação de disciplinas definidas no projeto pedagógico do curso por semestre. A aprovação do acadêmico pode então depender também da complexidade e carga horária das disciplinas matriculadas. Portanto, torna-se relevante ao NDE do curso e aos acadêmicos revelar padrões nas combinações de disciplinas que podem maximizar a aprovação.

Segundo Faceli *et al.* (2021), minerar conjuntos de itens frequentes em bases de dados permite descobrir relações interessantes entre as variáveis em análise. Um exemplo clássico é a análise de cesto de compras — do inglês, *market basket analysis* (MBA) — onde o objetivo é descobrir conjuntos de itens que frequentemente são comprados juntos. Com base nesses conjuntos, pode-se identificar, a partir dos itens comprados, quais outros itens poderiam ser comprados pelo consumidor. Essa identificação é feita a partir de regras de associação, e tem sido usada em diversas áreas, como sistemas de recomendação, detecção de intrusos e bioinformática (Faceli *et al.*, 2021).

Este trabalho apresenta um estudo de caso com aplicação de MBA para analisar as matrículas dos acadêmicos do curso de ESO. O objetivo é coletar indicadores de matrículas e evasão acadêmica, e verificar padrões frequentes na combinação de disciplinas cursadas simultaneamente que maximizam a chance de aprovação do acadêmico. O estudo aplicou o algoritmo *Apriori* em dados obtidos do sistema acadêmico para minerar combinações de matrículas aprovadas e obter regras de associação. A partir das regras de associação, uma Prova de Conceito (PoC) de um sistema de recomendação de disciplinas foi desenvolvida. O sistema indica, a partir de um conjunto de disciplinas selecionadas, em qual(is) outra(s) disciplina(s) há aprovação. Uma limitação da PoC desenvolvida é fazer recomendações apenas de disciplinas da mesma fase.

As contribuições deste trabalho são: (i) indicadores de conclusão, evasão e aprovações de acadêmicos do curso de ESO; (ii) regras que associam a aprovação em certa(s) disciplina(s) com a aprovação em outra(s) disciplina(s); e (iii) uma PoC de sistema de recomendação de disciplinas.

O restante deste artigo está organizado da seguinte forma. A seção 2 contempla a fundamentação teórica sobre o curso de ESO, mineração de dados e de regras de associação. A seção 3 apresenta trabalhos relacionados. A seção 4 expõe a metodologia adotada na realização do trabalho. Os resultados obtidos são apresentados na seção 5 e discutidos na seção 6. Por fim, a seção 7 apresenta considerações finais e sugestões de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta a fundamentação teórica que sustenta o desenvolvimento do trabalho. A seção inicia com a descrição do curso de ESO, seu objetivo e estrutura curricular. Em seguida, a seção apresenta o conceito de mineração de dados e uma metodologia consolidada para sua realização. Por fim, a seção detalha os conceitos e etapas da mineração de regras de associação.

2.1 Bacharelado em engenharia de software da UDESC

O curso de ESO da UDESC *campus* Alto Vale iniciou suas atividades em 2014. Seu objetivo é formar profissionais aptos a produzir sistemas de *software* de alta qualidade, sendo que por alta qualidade, compreende-se *softwares* produzidos aplicando-se técnicas, métodos e ferramentas que permitam produzi-los com propriedades ergonômicas, funcionais, manuteníveis, seguras e de alto desempenho para as diversas áreas de negócio (UDESC, 2016).

O curso de ESO oferece 40 vagas por semestre. A grade curricular é composta por 52 disciplinas, sendo 23 de 2 créditos, 24 de 4 créditos, 4 de 6 créditos, e a disciplina de Trabalho de Conclusão de Curso (TCC) com 5 créditos. Quanto a pré-requisitos, 30 disciplinas não possuem pré-requisito, 16 possuem uma única outra disciplina como pré-requisito, 3 possuem 2 disciplinas como pré-requisitos, e 2 possuem 3 disciplinas como pré-requisitos. Além de cursar as disciplinas e fazer o TCC, o acadêmico realiza estágio obrigatório (UDESC, 2016).

A Tabela 1 relaciona as disciplinas do curso de ESO. A coluna “código” apresenta a sigla de cada disciplina. Essa sigla é formada por dois dígitos seguidos de uma sequência de caracteres (ex.: 15INP, 25PRO1, 35REQ). O primeiro dígito identifica a fase (semestre) em que a disciplina está posicionada na grade curricular. O segundo dígito identifica o curso de ESO e sempre tem o valor 5. A sequência de caracteres identifica a disciplina (ex.: INtrodução a Programação, PROgramação 1, Engenharia de REquisitos).

Tabela 1: Disciplinas do curso de ESO

Fase	Código	Nome	Fase	Código	Nome
1	15ARC	Arquitetura de Computadores	5	55DSW	Desenvolvimento de Sistemas para Web
1	15CEX	Comunicação e Expressão	5	55EQU	Engenharia da Qualidade
1	15FAD	Fundamentos de Administração	5	55MQU	Métodos Quantitativos
1	15FES	Fundamentos de Engenharia de <i>Software</i>	5	55PIN2	Projeto Integrador II
1	15INP	Introdução à Programação	5	55PPR	Padrões de Projeto
1	15MAD	Matemática Discreta	5	55RED	Redes de Computadores
2	25ALG	Álgebra Linear	6	65DSD	Desenvolvimento de Sistemas Paralelos e Distribuídos
2	25GRH	Gestão de Recursos Humanos	6	65EEO	Estratégias de Experimentação e Observação
2	25PRE	Probabilidade e Estatística	6	65GRP1	Optativa I
2	25PRI	Projeto de <i>Interfaces</i>	6	65OPT1	Desenvolvimento para Dispositivos Móveis
2	25PRO1	Programação I	6	65PIN3	Projeto Integrador III
2	25PRS	Processo de <i>Software</i>	6	65TES	Testes de <i>Software</i>
2	25SIS	Sistemas Operacionais	7	75ARS	Arquitetura de <i>Software</i>
3	35BAD	Modelagem de Banco de Dados	7	75GCF	Gerência de Configuração
3	35CDI	Cálculo Diferencial e Integral	7	75GRP2	Gerenciamento de Projetos II
3	35FCO	Fundamentos de Contabilidade	7	75INC	Inteligência Computacional
3	35MPC	Metodologia da Pesquisa Científica	7	75OPT2	Optativa II
3	35PRO2	Programação II	7	75QUA	Qualidade de <i>Software</i>
3	35REQ	Engenharia de Requisitos	7	75TCC1	Orientação para Trabalho de Conclusão de Curso
4	45EMP	Empreendedorismo	8	85ECS	Ética, Computador e Sociedade
4	45EST	Algoritmos e Estrutura de Dados	8	85EEC	Engenharia Econômica
4	45LOG	Lógica Matemática	8	85MAN	Manutenção de <i>Software</i>
4	45PBD	Persistência de Dados	8	85MEF	Métodos Formais
4	45PIN1	Projeto Integrador I	8	85MKT	Marketing
4	45PRJ	Desenho e Projeto de <i>Software</i>	8	85MPS	Melhoria de Processo de <i>Software</i>
			8	85OPT3	Optativa III
			8	85TCC2	Trabalho de Conclusão de Curso

Fonte: UDESC (2016).

Como pode-se verificar, grande parte das fases é composta por 6 disciplinas. A segunda e a sétima fases contêm 7 disciplinas, e a oitava fase é formada por 8 disciplinas, a maioria de apenas 2 créditos, além do TCC. Uma nova grade curricular está em implantação gradual, substituindo a vigente de forma integral em 2027/1 (UDESC, 2023). A nova grade curricular foi proposta para incorporar a extensão universitária, bem como flexibilizar pré-requisitos.

2.2 Mineração de dados

A mineração de dados, ou de forma mais ampla, a descoberta de conhecimento em bases de dados, é um processo que tem por objetivo identificar padrões comprehensíveis, válidos, novos e úteis a partir de grandes conjuntos de dados (Goldschmidt; Passos; Bezerra, 2015). A mineração é realizada a partir da aplicação de algoritmos que recebem como entrada um conjunto de dados e produzem como saída algum padrão encontrado nos dados (Silva; Peres; Boscarioli, 2016). O conjunto de dados pode envolver vários atributos, cada qual descrevendo uma característica do contexto em estudo. A mineração de dados educacionais é a modalidade de mineração na qual os dados pertencem ao contexto educacional (Goldschmidt; Passos; Bezerra, 2015).

Na ciência de dados, a metodologia consolidada para realizar mineração de dados é a *Cross Industry Standard Process for Data Mining (CRISP-DM)* (Wirth; Hipp, 2000). A metodologia é formada por seis fases: compreensão do negócio; compreensão dos dados; preparação dos dados; modelagem; avaliação; e implantação (Provost; Fawcett, 2016). A fase de compreensão do negócio visa compreender o problema no qual pretende-se aplicar a mineração de dados. Na fase de compreensão dos dados, os conjuntos de dados disponíveis para mineração são analisados para identificar sua estrutura e significado, tipos, organização, qualidade, e relação entre os dados. A etapa de preparação dos dados compreende ações de pré-processamento para padronização, limpeza e consolidação dos dados. Na etapa de modelagem é onde se determina qual algoritmo será aplicado para minerar os dados em busca dos resultados desejados, compreendendo também a implementação e calibração do algoritmo escolhido. A etapa de avaliação compreende a análise da qualidade dos resultados obtidos pelo algoritmo de mineração, para verificar se os resultados obtidos estão dentro das expectativas inicialmente definidas. Por fim, a etapa de implantação consiste em colocar em prática os resultados da mineração de dados por meio de ações que impactam o negócio, e realizar o acompanhamento destas ações (Goldschmidt; Passos; Bezerra, 2015).

2.3 Mineração de regras de associação

Regras de associação descrevem associações ou correlações que existem entre os atributos de um conjunto de dados. As associações são extraídas a partir das transações que envolvem os dados, sendo que uma transação representa algum evento envolvendo itens do contexto em estudo, por exemplo, os produtos da cesta em uma compra no supermercado (Silva; Peres; Boscarioli, 2016). As Tabelas 2 e 3 apresentam exemplos de transações que representam compras de itens. A coluna “TID” identifica cada transação. Como pode-se verificar, as transações podem ser representadas por conjuntos, ou através de representação matricial, na qual os valores 0 e 1 indicam, respectivamente, a presença ou ausência do item na transação. Ao minerar as transações pode-se revelar regras de associação entre os itens.

Tabela 2: Exemplos de transações representadas por conjuntos

TID	Itens
0	{pão, leite}
1	{pão, fralda, cerveja, ovos}
2	{leite, fralda, cerveja, suco}
3	{pão, leite, fralda, cerveja}
4	{pão, leite, fralda, suco}

Fonte: Adaptado de Silva, Peres e Boscarioli (2016).

Uma regra de associação é, formalmente, uma implicação $X \rightarrow Y$, sendo que X e Y são conjuntos de itens tais que $X \cap Y = \emptyset$. Além disso, X e Y são chamados, respectivamente, de antecedente e consequente da regra (Goldschmidt; Passos; Bezerra, 2015). As regras podem ser entendidas como afirmações do tipo *se <antecedente> então <consequente>* (Silva; Peres; Boscarioli, 2016).

Uma aplicação comum de regras de associação é para MBA, onde se busca identificar combinações de itens que ocorrem juntos de forma frequente. Com MBA é possível revelar associações do tipo *se compra X então frequentemente também compra Y*. As regras de associação mineradas são usadas para melhorar aspectos do contexto em estudo, tais como a disposição dos itens à venda visando aproximar os que frequentemente são comprados juntos, ou recomendar outros itens a partir dos que já estão no cesto de compras (Silva; Peres; Boscarioli, 2016).

Tabela 3: Exemplos de transações com representação matricial

TID	cerveja	pão	suco	fralda	ovos	leite
0	0	1	0	0	0	1
1	1	1	0	1	1	0
2	1	0	1	1	0	1
3	1	1	0	1	0	1
4	0	1	1	1	0	1

Fonte: Adaptado de Silva, Peres e Boscarioli (2016).

A primeira etapa na mineração de regras de associação é identificar os subconjuntos de itens que frequentemente ocorrem juntos nas transações. Estes subconjuntos são denominados de *itemsets* frequentes. No exemplo da Tabela 2 nota-se que o *itemset* {leite} é mais frequente que o *itemset* {pão, leite}. O suporte é uma métrica utilizada para avaliar a frequência de *itemsets*. O suporte mede o percentual de transações que possui determinado *itemset*. Ainda no exemplo da Tabela 2, o suporte de {leite} é 0,8 (pois está presente em 4 das 5 transações), enquanto que o suporte de {pão, leite} é 0,6. Um *itemset* é considerado frequente quando seu suporte atende um limiar de suporte mínimo estabelecido pelo cientista de dados (Silva; Peres; Boscarioli, 2016). A partir dos *itemsets* frequentes pode-se aplicar algum algoritmo que identifica as regras de associação existentes. Segundo Silva, Peres e Boscarioli (2016), o *Apriori* é um algoritmo clássico utilizado para identificar as regras de associação.

O *Apriori* gera combinações de *itemsets* para formar regras compostas por antecedentes e consequentes. Para lidar com a ampla quantidade de combinações possíveis, o *Apriori* adota a medida de confiança da regra como critério de inclusão. A confiança de uma regra expressa a sua importância e confiabilidade. Quanto maior a confiança de uma regra, maior é a chance de ocorrência do consequente quando o antecedente ocorrer. A confiança de uma regra $X \rightarrow Y$ é calculada a partir da razão entre o suporte do *itemset* $X \cup Y$ e o suporte do *itemset* do antecedente X (Silva; Peres; Boscarioli, 2016). Em termos estatísticos, a confiança provê uma estimativa da probabilidade condicional $P(Y|X)$. O valor de confiança mínimo para inclusão das regras de associação no *Apriori* é também estabelecido pelo cientista de dados.

A depender do limiar de suporte mínimo e confiança mínima utilizados pelo *Apriori*, grande quantidade de regras de associação são geradas. Sendo assim, aplicar alguma medida de correlação entre os *itemsets* das regras pode melhorar a análise e permitir identificar regras de fato relevantes ao estudo (Silva; Peres; Boscarioli, 2016). Uma medida de correlação frequentemente adotada é o *lift*, que aponta a frequência com que o *itemset* do antecedente ocorre junto com o *itemset* do consequente nas transações. O *lift* de uma regra $X \rightarrow Y$ é determinado pela razão entre a confiança da regra e o suporte do *itemset* consequente Y . Um valor de *lift* igual a 1 indica descorrelação entre X e Y , ou seja, são independentes e, portanto, a ocorrência do antecedente X não tem efeito sobre a ocorrência do consequente Y . Já um valor de $lift < 1$ indica que X e Y são negativamente correlacionados, enquanto $lift > 1$ indica que X e Y são positivamente correlacionados (Faceli *et al.*, 2021).

3 TRABALHOS RELACIONADOS

Esta seção apresenta trabalhos relacionados ao presente estudo de caso. Os trabalhos foram encontrados a partir de buscas no *Google Scholar*, combinando os termos *learning trajectory* e *course trajectory* com *data mining* e *market basket analysis*, tanto em inglês quanto em português.

Almatrafi *et al.* (2016) aplicaram mineração de padrões frequentes nas matrículas de três cursos de graduação: engenharia civil, ciências da computação, e tecnologia da informação. O objetivo foi identificar padrões na trajetória dos acadêmicos, encontrando quais disciplinas frequentemente se matriculam a cada semestre, além de visualizar em formato gráfico a trajetória acadêmica em semestres consecutivos. Os autores analisam os resultados comparando os achados em grupos de alunos com alto e baixo rendimento. Concluem que estudantes com baixo rendimento postergam certas disciplinas para o final do curso, e se matriculam em combinações de disciplinas diferentes das cursadas pelos estudantes com alto rendimento. Os autores sugerem que esses resultados podem ser utilizados por professores e acadêmicos para conscientizar sobre as consequências que as diferentes trajetórias acadêmicas podem ter no desempenho acadêmico.

Reeping *et al.* (2019) aplicaram sequenciamento e visualização de dados para analisar as trajetórias acadêmicas

em cursos de engenharia. O objetivo foi identificar quais são as trajetórias comumente realizadas pelos acadêmicos, e qual a relação destas trajetórias com o desempenho acadêmico em futuras disciplinas. Os autores utilizaram diagramas aluviais (similar a diagramas de conjuntos paralelos ou de *Sankey*) para visualização das trajetórias acadêmicas. Os resultados encontrados sugerem uma grande diversidade nas trajetórias dos acadêmicos. Os autores também sugerem que os achados podem auxiliar professores e acadêmicos na escolha das disciplinas a serem cursadas, objetivando reduzir o tempo de integralização.

Carmo, Gasparini e Oliveira (2022) também buscaram identificar trajetórias de aprendizagem em um curso de ciência da computação. O objetivo foi investigar padrões relacionados à evasão. Para tanto, os autores utilizaram uma abordagem estruturada (baseada em texto), e outra visual (baseada em grafos). Em ambas, o objetivo foi encontrar padrões nas sequências de disciplinas cursadas pelos estudantes evadidos e concluintes. Os resultados também evidenciaram diversidade na trajetória dos acadêmicos, que cursam as disciplinas de maneira diversa, não se atendo à grade curricular proposta. Outro resultado reportado é que grande parte da evasão ocorre ao fim do primeiro semestre, após reprovação em disciplinas iniciais. A partir dos achados, os autores concluem que as disciplinas do primeiro semestre, com alto índice de reprovação, atuam como barreiras para o avanço dos estudantes e levam ao abandono do curso, e ressaltam a necessidade de as instituições de ensino pensarem iniciativas para mitigar a evasão com foco nos primeiros semestres.

Atalla *et al.* (2023) propuseram um sistema de recomendação de disciplinas a serem cursadas. O sistema recomenda disciplinas e atividades extracurriculares a partir de dados do perfil e histórico acadêmico. A sugestão de disciplinas é gerada por algoritmos de recomendação e aprendizagem de máquina. De acordo com os autores, os resultados obtidos indicam a viabilidade de apresentar recomendações aos estudantes com base no perfil acadêmico e trajetórias de aprendizagem similares.

Lucena *et al.* (2024) desenvolveram um sistema de recomendação para recomendar uma matriz horária semestral a estudantes de graduação, com base no histórico estudantil, preferências do estudante e escolhas de seus pares. O sistema não utiliza MBA para as recomendações, e sim um mecanismo de pontuação proposto pelos autores. A pontuação é computada com base nos pré-requisitos entre as disciplinas do currículo e também com base nas disciplinas escolhidas pelos colegas do estudante (princípio de aprendizagem colaborativa). Os autores avaliaram o sistema com um grupo de estudantes, e nos resultados obtidos destacam a avaliação positiva da usabilidade, estabilidade e interesse de utilização do sistema por parte dos estudantes.

Xiao, Hu e Huang (2024) utilizaram *large language models* para extraír trajetórias de aprendizagem a partir do histórico de acadêmicos. As trajetórias extraídas são utilizadas para treinamento de um modelo preditivo, que tem por objetivo recomendar a próxima disciplina a ser cursada pelo acadêmico. Os autores avaliaram o modelo proposto em um *dataset* de *massive open online courses* (MOOC), e afirmam que o modelo proposto teve desempenho superior a outros modelos preditivos.

O estudo de caso apresentado neste trabalho se relaciona com os trabalhos descritos por ter objetivo similar, que é identificar e verificar padrões frequentes nas disciplinas cursadas pelos acadêmicos. Cabe destacar como diferenciais do presente trabalho: (i) o uso de dados concretos do curso de ESO da UDESC para extração de indicadores do curso e mineração de regras de associação; e (ii) o desenvolvimento de uma PoC de sistema de recomendação que poderia ser utilizado por acadêmicos no momento da matrícula.

4 MATERIAIS E MÉTODOS

O objetivo deste trabalho é aplicar MBA para identificar padrões frequentes na combinação de disciplinas cursadas simultaneamente que maximizam a chance de aprovação do acadêmico no curso de ESO. Pretende-se investigar as seguintes questões de pesquisa: (i) Quantos semestres os acadêmicos levam para integralizar o curso, e qual a proporção de acadêmicos? (ii) Quais são os padrões frequentes de matrículas em disciplinas? (iii) Como os padrões frequentes de matrículas se relacionam com a distribuição de disciplinas por fase da grade curricular do curso? (iv) Qual(is) disciplina(s) pode-se recomendar ao acadêmico na realização da matrícula para maximizar sua chance de aprovação?

A metodologia adotada neste trabalho é a CRISP-DM, descrita na seção 2.2. A seguir são descritas as atividades que foram realizadas em cada uma das fases.

4.1 Compreensão do negócio

Nesta fase buscou-se conhecer o contexto em que o processo de mineração de dados é realizado. Conforme já mencionado previamente, o projeto foi realizado na UDESC, especificamente com o curso de graduação em ESO. Nesta fase, em conjunto com a próxima, foram levantados dados gerais sobre indicadores demográficos dos acadêmicos do curso. Os indicadores foram extraídos a partir dos dados fornecidos pelo coordenador de informação e registro discente da pró-reitoria de ensino (administrador do banco de dados). Em 6 de setembro de 2023, o coordenador de informação exportou, em formato de valores separados por vírgula (CSV), dois *datasets* do sistema acadêmico.

O primeiro *dataset*, denominado *eso-alunos-por-situacao-matricula*, contém dados demográficos de todos os estudantes do curso. Já o segundo, denominado *eso-matriculas*, contém dados de todas as matrículas. Ambos contemplam registros desde a implantação do curso de ESO em 2014. Os *datasets* foram processados em Python com pandas¹ e plotly² para extração de indicadores gerais. Todas as implementações foram realizadas e executadas no ambiente *Google Colaboratory*.

4.2 Compreensão dos dados

Nesta fase realizou-se um estudo minucioso das informações disponíveis. O objetivo foi compreender os *datasets* fornecidos, identificando atributos, seus tipos de dados e relacionamentos.

O *dataset* *eso-alunos-por-situacao-matricula* contém atributos que descrevem informações elementares sobre o perfil de cada acadêmico: dados pessoais, sexo, forma de ingresso, tipo de instituição onde cursou ensino médio (privada ou pública). Cada matrícula realizada na trajetória do acadêmico é representada por um registro nesse *dataset* que contém também o semestre atual do acadêmico (cronológico, desde a primeira matrícula), e seu status (formado ou evadido). O *dataset* é composto por 4493 registros de matrículas.

Já o *dataset* *eso-matriculas* é composto por registros de matrículas em disciplinas do curso de ESO. Cada registro contém dados sobre a disciplina matriculada, semestre letivo, situação (aprovado/reprovado), nota final, e dados pessoais dos acadêmicos. O *dataset* é composto por 22.356 registros. Os *datasets* utilizam os mesmos códigos de disciplinas adotados no projeto do curso (UDESC, 2016) e descritos previamente na seção 2.1.

4.3 Preparação dos dados

Nesta fase foram executadas ações de pré-processamento com pandas. Os dados disponibilizados pelo sistema acadêmico já são estruturados. Portanto, foram necessárias somente ações para remoção de registros que não eram de interesse ao trabalho e ações para consolidação de certas disciplinas.

No *dataset* *eso-matriculas* foram removidos registros: de alunos de outros cursos matriculados em disciplinas de ESO (12); com o código da turma ou da disciplina em branco (533); e de disciplinas cursadas em outros cursos (95). Após isso, o *dataset* ficou com 21.716 registros de matrículas.

Para consolidação das disciplinas, todas as optativas de determinada fase foram unificadas em uma única, resultando nas disciplinas 65OPT1, 75OPT2, e 85OPT3 já nomeadas no projeto do curso (UDESC, 2016). Além disso, a disciplina 85ESS (Estágio Obrigatório) foi removida por não ser de interesse ao presente estudo. Também foram removidos registros de matrículas do semestre no qual este estudo foi realizado (2023/2) por ainda não haver resultado de aprovação/reprovação na data em que os dados foram extraídos do sistema acadêmico. O resultado é um *dataset* com 20.117 registros.

4.4 Modelagem

Nessa fase aplicou-se um algoritmo de mineração sobre os dados. O algoritmo *Apriori* foi adotado neste trabalho para identificar os padrões frequentes na combinação de disciplinas cursadas e obter regras de associação. Apesar da existência de outros algoritmos (ex.: *FP-Growth*), cabe ressaltar que o objetivo deste trabalho não é comparar o desempenho de algoritmos, e sim revelar os eventuais padrões frequentes nas combinações de disciplinas. Portanto, apenas o algoritmo *Apriori* foi utilizado devido a sua simplicidade e viabilidade de uso no *dataset* em estudo.

Para utilizar o *Apriori*, os diversos registros de matrículas de cada acadêmico em determinado ano/semestre foram convertidos em transações. Os itens das transações são os identificadores das disciplinas matriculadas. A

representação matricial foi adotada para representar as transações. A partir dessa representação foi possível determinar o suporte de cada disciplina. A etapa de modelagem foi realizada com as bibliotecas pandas¹, plotly² e mlxtend³.

Para o presente estudo, o *dataset* *eso-matriculas* foi dividido em dois subconjuntos. O primeiro é composto apenas pelos registros de matrículas aprovadas, ou seja, onde o acadêmico obteve aprovação na disciplina.

O *dataset* contém 12.976 registros. Já o segundo é composto pelos registros de matrículas não aprovadas, compreendendo matrículas onde houve reprovação por nota, frequência, cancelamento ou trancamento. Esse *dataset* contém 7.141 registros. Portanto, a razão de aprovação é de 64,5% das matrículas realizadas.

Neste estudo a mineração de regras de associação foi realizada apenas no *dataset* de matrículas aprovadas. A escolha deu-se pelo interesse em analisar as combinações de matrículas que resultaram em aprovação para eventual recomendação de disciplinas aos acadêmicos. Para execução do *Apriori*, é necessário definir o valor do limiar de inclusão dos *itemsets* (combinações de disciplinas encontradas nas transações) na mineração das regras de associação. Valores de limiar altos excluem *itemsets* pouco frequentes e produzem poucas regras. Já valores baixos incluem *itemsets* pouco frequentes e produzem muitas regras. O valor deste parâmetro foi avaliado de forma experimental, e o critério utilizado para determinar o valor adotado foi de que todas as disciplinas do curso deveriam ocorrer em ao menos um antecedente ou consequente das regras de associação. O maior valor que atendeu este critério foi 0,019.

4.5 Avaliação

Nesta etapa foi verificada a qualidade dos modelos gerados na etapa anterior. A verificação foi baseada nas métricas adotadas para construção de regras de associação, em particular os valores de suporte e *lift*. O *lift* é uma métrica que define um coeficiente de interesse para uma regra de associação. Um alto valor de *lift* indica uma correlação positiva entre o antecedente e o consequente da regra. Já um valor próximo de 1 indica que o antecedente e o consequente são independentes, e valores abaixo de 1 indicam correlação negativa entre o antecedente/consequente (Faceli *et al.*, 2021). Nessa etapa também foram adotadas estratégias de visualização de dados.

Cabe mencionar que, por serem utilizadas técnicas de aprendizado não supervisionado, a avaliação baseou-se essencialmente na interpretação subjetiva dos padrões revelados pelos dados. Caso um sistema de recomendação de matrículas for viabilizado a partir dos modelos construídos neste estudo, futuramente pode-se realizar análises quantitativas considerando a acurácia das recomendações após a conclusão das disciplinas.

4.6 Implantação

Nesta fase as conclusões obtidas nas fases anteriores são implantadas por meio de ações que impactam o negócio e é feito o acompanhamento dessas ações. Uma PoC foi desenvolvida nessa fase do estudo. Nessa PoC, o conjunto de regras de associação criado a partir das matrículas aprovadas é carregado de um arquivo. Pode-se então especificar a(s) disciplina(s) que o acadêmico pretende se matricular. O sistema identifica a(s) regra(s) de associação cujo antecedente contém a(s) disciplina(s) pretendida(s), e a partir do consequente da(s) regra(s) encontrada(s), sugere outra(s) disciplina(s) que poderia(m) ser cursada(s).

Além da PoC, considera-se que o presente artigo, que evidencia os achados encontrados na mineração de padrões frequentes nas matrículas, seja também um artefato resultante da etapa de implantação. Os resultados deste estudo foram compartilhados com o departamento e com o NDE do curso de ESO, que podem recomendar ajustes na grade curricular, ou aprimorar a construção da grade de horários semestral para oportunizar aos acadêmicos frequentar simultaneamente disciplinas onde há maior potencial de aprovação.

5 RESULTADOS

Esta seção apresenta os resultados obtidos com a aplicação de MBA nas matrículas do curso de ESO. Inicialmente são apresentados resultados referentes aos indicadores demográficos, e de conclusão/evasão. Em seguida, são apresentados resultados obtidos com o algoritmo *Apriori* sobre os grupos de disciplinas que são cursadas em conjunto. Por fim, são apresentados resultados sobre as regras de associação mineradas pelo *Apriori*.

¹<<https://pandas.pydata.org>>

²<<https://plotly.com/>>

³<<https://rasbt.github.io/mlxtend/>>

5.1 Indicadores demográficos e de matrícula/conclusão/evasão

Desde a implantação em 2014 até o momento, 833 acadêmicos se matricularam no curso de ESO. Quanto ao gênero, 80,31% (669) é masculino, 16,69% (139) feminino, e 3% (25) não declarado. Em relação à quantidade acumulada de formados, foram apenas 75 até o momento (ou seja, 9% dos matriculados), sendo 89,33% do gênero masculino (67) e 10,67% feminino (8). Quanto à evasão acumulada, foi registrado um total de 506 acadêmicos que deixaram de frequentar o curso (60,74% dos matriculados). Portanto, frequentam ativamente o curso no momento de realização deste estudo 252 acadêmicos (30,25% do acumulado de matriculados).

Outros aspectos que puderam ser verificados a partir dos *datasets* são o tipo de instituição onde o acadêmico cursou ensino médio e a forma de ingresso. A Tabela 4 apresenta os tipos de instituição registrados. Pode-se verificar que a maior parcela cursou ensino médio em escola pública. A Tabela 4 também apresenta as formas de ingresso registradas. O ingresso através de vestibular, SISU e retorno de diplomados são predominantes.

Tabela 4: Indicadores demográficos

Ensino médio	N	%	Forma de ingresso	N	%
Escola pública	639	77,17	Vestibular	664	77,84
Escola privada	188	22,71	SISU	131	15,36
Sem informação	1	0,12	Retorno de diplomados	38	4,45
			Processo seletivo	10	1,17
			Transferência interna	6	0,70
			Reingresso após abandono	2	0,23
			Estudante convênio	1	0,12
			Transferência externa	1	0,12

Para investigar quantos semestres os acadêmicos levam para concluir ou abandonar o curso, foi elaborado um gráfico interativo que relaciona o número de evasões, matrículas e formaturas por quantidade de semestres cursados. Através do gráfico, apresentado na Figura 1, pode-se inicialmente verificar a amplitude na quantidade de semestres em curso ou cursados pelos acadêmicos. Há casos de acadêmicos vinculados ao curso por 19 semestres. O pico de evasão é registrado com acadêmicos que frequentaram um semestre do curso. Enquadram-se neste perfil aqueles que não se matricularam após frequentar disciplinas da primeira fase, ou se matricularam em alguma disciplina posterior à primeira fase e então cancelaram/abandonaram o curso. Cabe destacar também a evasão de acadêmicos que cancelaram a matrícula sem sequer terminar o primeiro semestre. No geral, observa-se que a evasão diminui à medida que os acadêmicos avançam no curso, mas ainda é observada mesmo após acadêmicos terem frequentado muitos semestres.

Com relação aos acadêmicos atualmente matriculados, verifica-se que os maiores quantitativos são de acadêmicos que cursaram 1 e 2 semestres. Por outro lado, há acadêmicos que se matricularam no curso por mais de 15 semestres, quase o dobro do prazo regular para conclusão do curso. Por fim, com relação aos acadêmicos formados, nota-se que 44% (33 acadêmicos) concluíram o curso após frequentar 8 semestres, portanto, no prazo regular. A grande maioria, 77,33%, levou entre 8 e 10 semestres para concluir o curso.

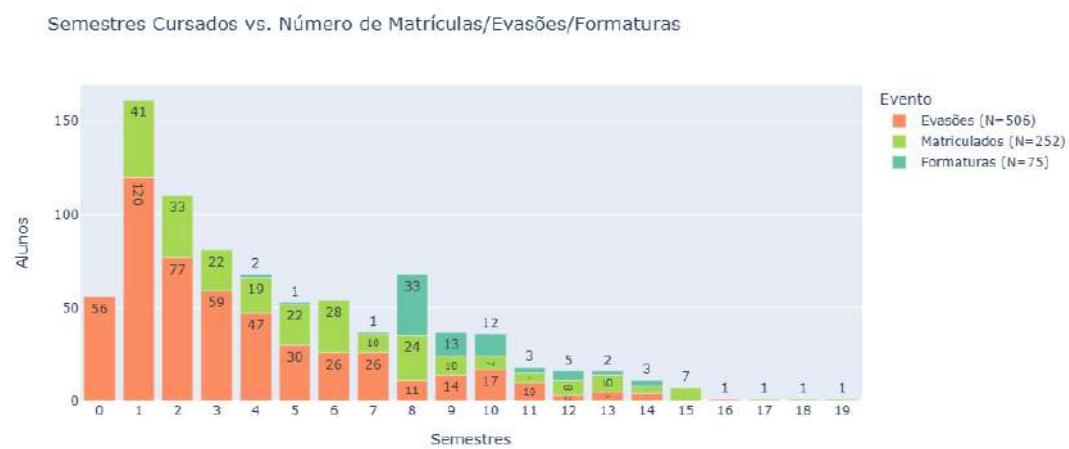
5.2 Aprovações e não aprovações em disciplinas

A partir da representação matricial das transações (disciplinas matriculadas em cada semestre) é possível determinar o suporte de cada disciplina. O valor de suporte representa o percentual de transações que contém a disciplina. A partir dos valores de suporte, apresentados na Figura 2, é possível observar a frequência de aprovação/não aprovação por disciplina.

Ao observar o suporte das aprovações, verifica-se que nas disciplinas iniciais do curso a frequência de aprovação é maior. Em certa medida isso também é verificado nas não aprovações. Isso é causado pela maior quantidade de acadêmicos nas fases iniciais do curso, ocasionando maior volume de transações com disciplinas dessas fases.

Quanto às reprovações, é evidenciado maior frequência nas disciplinas de programação das fases iniciais (15INP e 25PRO1), e nas matemáticas (15MAD, 25ALG, 25PRE e 35CDI).

Cabe destacar que o algoritmo *Apriori* requer um valor de suporte mínimo para considerar a inclusão de itens nos conjuntos. Portanto, a verificação dos valores de suporte das aprovações e não aprovações serviu de heurística

Figura 1: Semestres cursados e quantidade de matrículas/evasões/formaturas.

para experimentar valores de suporte mínimo adotados no *Apriori*.

5.3 Regras de associação

As regras de associação foram geradas a partir do *dataset* de matrículas aprovadas, conforme descrito na metodologia. A execução do *Apriori* resultou em 652 *itemsets* frequentes, sendo criadas 7.484 regras de associação. O tamanho dos antecedentes e consequentes destas regras varia de 1 a 6 disciplinas. O *lift* médio destas regras é 16,25, e seu valor varia de 4,29 (mínimo) a 40,20 (máximo) sendo 13,85 o valor mediano. Estes valores de *lift* indicam correlação positiva entre antecedentes e consequentes em todas as regras. Portanto, essas regras são potencialmente úteis para predição, pois quanto maior o *lift*, mais assertiva é a regra de associação.

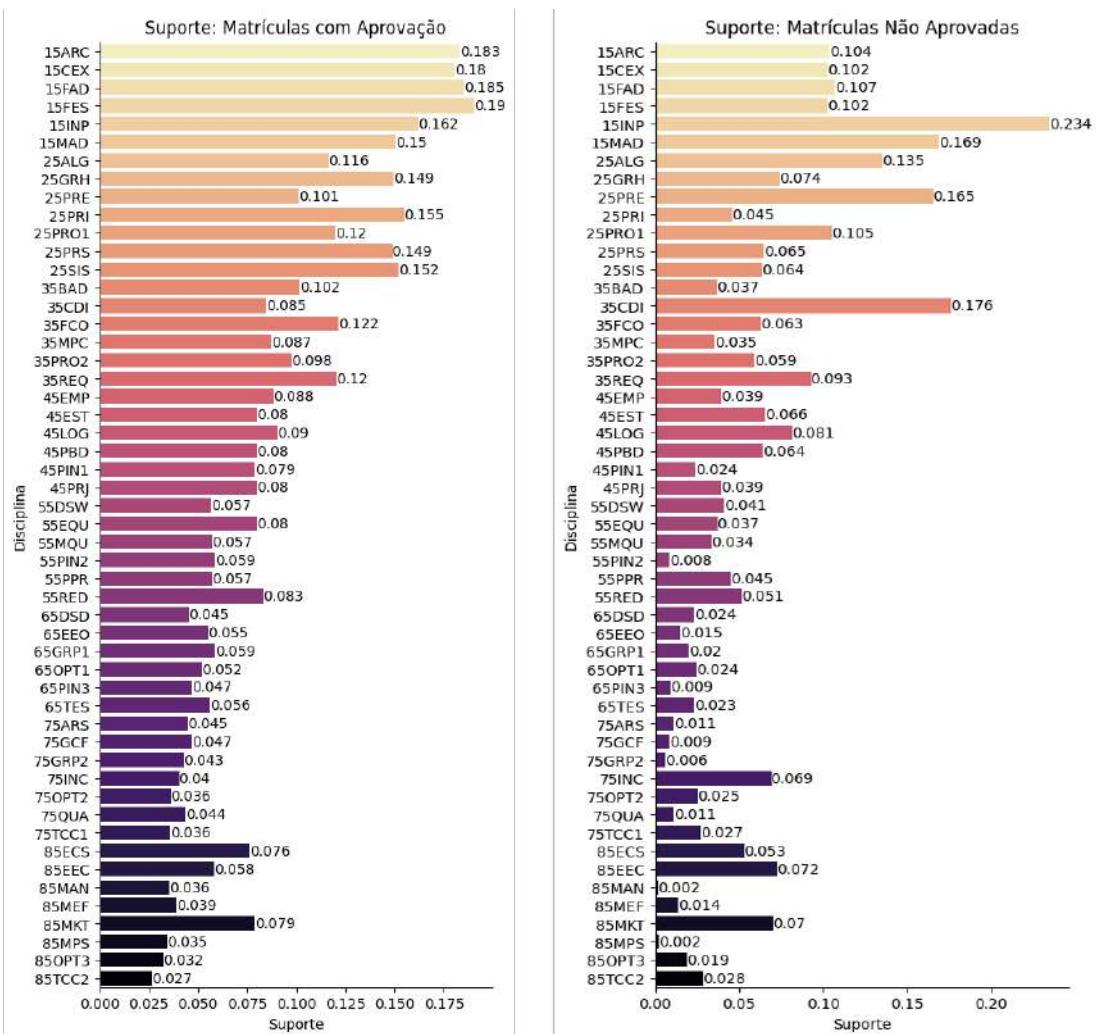
Ao examinar as disciplinas envolvidas em cada regra, constatou-se não haver heterogeneidade de fases nos antecedentes e consequentes. Em todas as regras, as disciplinas do antecedente e consequente são da mesma fase. O elevado número de regras geradas restringe a visualização conjunta de forma efetiva. Portanto, as evidências supracitadas foram extraídas diretamente do conjunto de regras através de manipulações dos dados. Contudo, para ilustrar as regras de associação geradas neste estudo, são apresentadas visualizações gráficas com *heatmaps* de um subconjunto das regras.

A Figura 3 apresenta regras de associação nas quais o antecedente é composto por uma única disciplina, sendo ela da primeira ou última fase do curso. Os consequentes mostrados são aqueles das regras com máximo valor de *lift*. Similarmente, a Figura 4 apresenta as regras de associação nas quais o consequente é composto por uma única disciplina, da primeira e da última fase, e os antecedentes são aqueles das regras com máximo valor de *lift*. Em ambas, o valor entre colchetes nos rótulos dos eixos é o tamanho do *itemset*.

De forma geral, verifica-se que o *lift* das regras envolvendo fases iniciais é menor do que das fases finais. Isso pode ser explicado pelo maior volume de transações nas fases iniciais em função da maior quantidade de acadêmicos nelas matriculados.

Isso faz com que a aprovação nessas disciplinas seja algo frequente, levando a correlações positivas de menor intensidade. Já nas fases finais, onde o volume de transações/acadêmicos é menor, os altos valores de *lift* indicam correlação positiva de maior intensidade entre o antecedente e o consequente.

Outro aspecto que pode ser verificado são os diferentes tamanhos dos antecedentes/consequentes nas regras com *lift* máximo. Nota-se que muitas regras não sugerem aprovação conjunta em todas as disciplinas da mesma fase (em uma analogia, “comprar todos os itens da mesma prateleira”). Apesar de curioso, esse fato não causaria problemas no sistema de recomendação, já que há outras regras com menores valores de *lift* que contemplam a aprovação conjunta em todas as disciplinas.

Figura 2: Suporte das disciplinas com aprovação e não aprovadas.

5.4 PoC do Sistema

Uma PoC de um sistema de recomendação de matrículas foi desenvolvida em um *notebook Google Colaboratory*. Nesta PoC é possível especificar a(s) disciplina(s) que o acadêmico pretende se matricular, e o sistema sugere outra(s) disciplina(s) que poderia(m) ser realizadas. O *notebook* está disponível e pode ser executado *online*.⁴

A Figura 5 apresenta as saídas geradas pela PoC do sistema ao se indicar matrícula na disciplina 65DSD. O sistema apresenta outras disciplinas que acadêmicos matriculados em 65DSD também foram aprovados — Figura 5(A). A PoC do sistema também apresenta um gráfico com as regras de associação que fundamentaram a indicação e seus respectivos valores de *lift* — Figura 5(B).

6 DISCUSSÃO

A partir dos indicadores de conclusão e evasão coletados, verificou-se que o percentual acumulado de evasão no curso (60,74%) é superior à média nacional das instituições de ensino superior públicas (40,3%) apontada pelo *Mapa do Ensino Superior no Brasil* (SEMESP, 2023). Contudo, ao se considerar apenas os cursos de tecnologia da informação, o mesmo mapa indica que a evasão acumulada é de 65,5%, sendo que nos cursos de Engenharia de

⁴<<https://colab.research.google.com>>

Figura 3: Regras de associação fases 1 e 8 com disciplina única no antecedente.

Regras de Associação Fases 1 e 8 [len(antec)=1; max(lift(conseq))]



Software este valor é ainda maior.

Considerando os cursos de Engenharia de Software, a evasão nacional acumulada fica em 52,5%, portanto apenas ligeiramente inferior à evasão registrada no curso de ESO da UDESC. Considerando os formados, observou-se que 50,6% (38) levaram mais tempo do que o previsto (8 semestres) para se formar. Este indicador é superior aos 36% registrados para a América Latina (Ferreyra *et al.*, 2017). No que diz respeito às reprovações, os resultados revelados neste trabalho estão alinhados com os de Carmo, Gasparini e Oliveira (2022), que relataram alto índice de reprovações em disciplinas do primeiro semestre e sugeriram que isso atua como barreira para o avanço dos estudantes. Os indicadores levantados respondem à questão de pesquisa (i).

A partir desses indicadores, é pertinente ao departamento de ESO investigar as causas e mitigar a evasão e o tempo de integralização do curso. Uma ação neste sentido foi a reformulação da matriz curricular realizada recentemente (UDESC, 2023). Essa reformulação eliminou vários pré-requisitos e disciplinas de 2 créditos, visando melhorar a fluidez acadêmica e possivelmente reduzir o índice de evasão.

O algoritmo *Apriori* foi capaz de minerar regras de associação de forma ágil, já que o *dataset* de matrículas utilizado no estudo contém apenas 12.976 registros. Contudo, em função do limiar utilizado para incluir todas as disciplinas, uma grande quantidade de regras (7.484) foi minerada, restringindo a visualização conjunta. Em função da limitação de espaço, apenas um pequeno conjunto dessas regras foi mostrado neste artigo. Outras regras, bem como os *itemsets* mais frequentes estão disponíveis em um repositório criado com material suplementar.⁵

O repositório contém gráficos interativos criados com as bibliotecas *plotly* e *d3blocks*⁶, que permitem visualizar *itemsets* com maior suporte e regras de associação com maior *lift* por disciplina.

A mineração das regras de associação foi realizada apenas no *dataset* de matrículas aprovadas, caracterizando uma limitação do estudo. Portanto, a resposta à questão de pesquisa (ii) limita-se aos padrões frequentes em matrí-

⁵Link omitido para avaliação duplo-cega

⁶<<https://d3blocks.github.io/d3blocks>>

Figura 4: Regras de associação fases 1 e 8 com disciplina única no consequente.

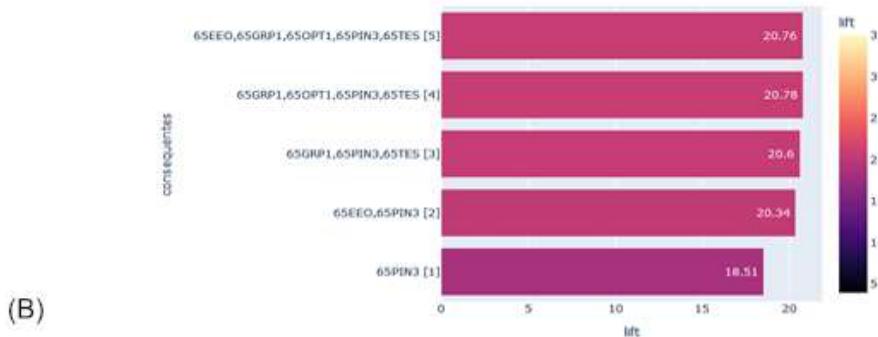
Regras de Associação Fases 1 e 8 [max(lift(antec)); len(conseq)=1]


Figura 5: PoC do sistema de recomendação de matrículas.

Acadêmicos aprovados em ['65DSD'] também foram aprovados em ['65OPT1', '65GRP1', '65TES', '65PIN3']. Os aprovados em ['65DSD'] também foram aprovados nas seguintes outras disciplinas:
 (65PIN3)
 (65EEO, 65PIN3)
 (65GRP1, 65TES, 65PIN3)
 (65OPT1, 65GRP1, 65TES, 65PIN3)
 (65EEO, 65OPT1, 65TES, 65PIN3, 65GRP1)

(A)

Regras de associação com max(lift(conseq)) e antecedente = ['65DSD']


(B)

culas aprovadas. Observou-se uma peculiaridade nas regras de associação mineradas: antecedentes e consequentes envolvem disciplinas sempre da mesma fase. A ocorrência de aprovação simultânea em disciplinas de diferentes fases não é significativa a ponto de produzir regras de associação entre elas. Isso sugere que os padrões frequentes de aprovações em disciplinas refletem a grade curricular do curso, respondendo à questão de pesquisa (iii).

A PoC do sistema de recomendação desenvolvido utiliza as regras de associação mineradas para indicar disciplina(s) que poderia(m) ser cursada(s) pelo acadêmico conjuntamente, respondendo à questão de pesquisa (iv). No entanto, em função das regras mineradas somente associarem disciplinas da mesma fase, a recomendação é útil apenas para acadêmicos que pretendem se matricular em disciplinas de uma mesma fase. A recomendação de disciplinas em fases distintas envolveria nova análise dos *itemsets* frequentes, com eventual redução no limiar para inclusão nas regras de associação.

Ao longo do desenvolvimento deste trabalho, percebeu-se que para visualizar um grande volume de regras de decisão e extrair *insights* de forma eficiente seria pertinente o uso de *dashboards* interativos que permitissem filtrar antecedentes e consequentes dinamicamente. Uma tentativa foi realizada utilizando a biblioteca `arulespy`⁷. Essa biblioteca faz integração do Python com o R, e oferece funções para criar um *dashboard* de exploração interativa de regras. Contudo, a tentativa foi malsucedida por limitações na integração Python/R para conversão de *dataframes* do pandas para tipos de dados R.

O uso de apenas um algoritmo de mineração (*Apriori*) pode levantar suspeita de vieses na mineração de dados. Quanto a isso, cabe destacar que quando os mesmos parâmetros (suporte, *lift*, etc.) são utilizados, diferentes algoritmos de mineração geram as mesmas regras (Hossain; Sattar; Paul, 2019; Idris *et al.*, 2022). Além disso, as regras geradas pelo *Apriori* são consideradas válidas devido à própria natureza do algoritmo, que se baseia no princípio da antimonotonicidade do suporte para gerar *itemsets* frequentes (Goldschmidt; Passos; Bezerra, 2015). Neste sentido, argumenta-se por descartar vieses na mineração das regras feitas pelo *Apriori*.

7 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Este artigo apresentou um estudo de caso com a aplicação de MBA em matrículas de disciplinas do curso de ESO da UDESC Alto Vale. Indicadores demográficos, de conclusão, de evasão, e de aprovações foram extraídos a partir de um dataset com todas as matrículas realizadas desde o início do curso em 2014. Estes indicadores revelaram que os acumulados de conclusão e evasão no curso são de 9% e 60,74% respectivamente, e que 77,33% dos formados levam entre 8 e 10 semestres para concluir o curso. As maiores frequências de reprovações são em disciplinas de programação das fases iniciais e nas matemáticas.

O algoritmo *Apriori* foi aplicado para minerar combinações frequentes e regras de associação entre as disciplinas com matrículas aprovadas. Ao todo, 7.484 regras foram mineradas, permitindo o desenvolvimento de uma PoC de sistema de recomendação. Nessa PoC, ao selecionar uma (ou mais) disciplina(s), o sistema consulta as regras de associação e indica qual(is) disciplina(s) há aprovação.

Como trabalhos futuros, sugere-se aprofundar a investigação da evasão, segmentando a análise em períodos, para investigar o possível efeito de fatores externos, por exemplo, a pandemia de Covid-19. Quanto à mineração de regras de associação, sugere-se realizá-la considerando simultaneamente as matrículas aprovadas e não aprovadas. Isso pode revelar associações entre disciplinas com e sem aprovação do tipo aprovados em x reprovam em y. Outra sugestão é considerar o perfil e o histórico acadêmico na mineração. Isto poderia identificar fatores que implicam em aprovações ou reprovações, bem como poderia resultar em regras mais assertivas para cada perfil de acadêmico. Uma sugestão adicional seria realizar uma comparação entre diferentes algoritmos para geração de regras de associação, como por exemplo o FP-Growth, a partir do conjunto de dados utilizado neste trabalho. Por fim, pode-se também realizar uma avaliação das recomendações geradas pelo sistema. Essa avaliação poderia ser realizada inicialmente com técnicas de validação cruzada. Posteriormente, a avaliação poderia observar, por um razoável período de tempo, as aprovações/reprovações dos acadêmicos que se matricularam seguindo as recomendações para identificar a efetividade destas recomendações.

⁷Integração Python/R para mineração de regras

REFERÊNCIAS

- ALMATRAFI, O.; JOHRI, A.; RANGWALA, H.; LESTER, J. Identifying course trajectories of high achieving engineering students through data analytics. *In: 2016 ASEE Annual Conference & Exposition Proceedings.* ASEE: Conferences, 2016.
- ATALLA, S. *et al.* An intelligent recommendation system for automating academic advising based on curriculum analysis and performance modeling. *Mathematics*, v. 11, n. 5, 2023.
- CARMO, E. P. d.; GASPARINI, I.; OLIVEIRA, E. H. T. d. Identificação de trajetórias de aprendizagem em um curso de graduação e sua relação com a evasão escolar. *In: Anais do XXXIII Simpósio Brasileiro de Informática na Educação (Sbie 2022).* Brasil: Sociedade Brasileira de Computação, 2022. p. 323–333.
- FACELI, K.; LORENA, A. C.; GAMA, J. A.; ALMEIDA, T. A. D.; CARVALHO, A. C. P. L. F. D. **Inteligência artificial:** Uma abordagem de aprendizado de máquina. Rio de Janeiro: Ltc, 2021.
- FERREYRA, M. M.; AVITABILE, C.; ÁLVAREZ, J. B.; PAZ, F. H.; URZÚA, S. **At a crossroads:** Higher education in Latin America and the Caribbean. EUA: World Bank, Washington, Dc, 2017. ISBN 978-1-4648-1014-5.
- GARCIA, L. M. L. d. S.; GOMES, R. S. Visualização e análise da trajetória de aprendizagem realizada no currículo no ensino superior. *In: Anais do XXXI Simpósio Brasileiro de Informática na Educação (Sbie 2020).* Brasil: Sociedade Brasileira de Computação, 2020. p. 1593–1602.
- GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. **Data mining:** Conceitos, técnicas, algoritmos, orientações e aplicações. Rio de Janeiro: Elsevier, 2015.
- HOSSAIN, M.; SATTAR, A. H. M. S.; PAUL, M. K. Market basket analysis using apriori and fp growth algorithm. *In: Proceedings of the 22nd International Conference on Computer and Information Technology (Iccit).* Bangladesh: Ieee, 2019. p. 1–6.
- IDRIS, A. I. *et al.* Comparison of apriori, apriori-tid and fp-growth algorithms in market basket analysis at grocery stores. *The International Journal of Informatics and Computer Science*, v. 6, p. 107–112, 2022.
- LUCENA, T. B. D. S.; GOMES, P. S.; MUSA, D. L.; OLIVEIRA, T. D. Sistema de recomendação para auxiliar a montagem de matriz horária semestral de estudantes de graduação. *Renote*, v. 22, n. 1, p. 544–555, jul. 2024.
- PROVOST, F.; FAWCETT, T. **Data science para negócios.** Rio de Janeiro: Alta Books, 2016.
- REEPING, D.; KNIGHT, D. B.; GROHS, J. R.; CASE, S. W. Visualization and analysis of student enrollment patterns in foundational engineering courses. *International Journal of Engineering Education*, v. 35, p. 142–155, 2019.
- SEMESP. **Mapa do ensino superior do Brasil - 13ª edição.** Brasil: Instituto Semesp, 2023. Disponível em: <https://www.semesp.org.br/mapa>.
- SILVA, L. A.; PERES, S. M.; BOSCAROLI, C. **Introdução à mineração de dados com aplicações em R.** Rio de Janeiro: Elsevier, 2016.
- UDESC. **Projeto pedagógico do curso de bacharelado em engenharia de software.** SC: Universidade do Estado de Santa Catarina, 2016. Disponível em: <http://www.ceavi.udesc.br/>.
- UDESC. **Projeto pedagógico do curso de bacharelado em engenharia de software.** SC: Universidade do Estado de Santa Catarina, 2023. Disponível em: <https://www.udesc.br/arquivos>.
- WIRTH, R.; HIPP, J. Crisp-dm: Towards a standard process model for data mining. *In: Anais da 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.* Manchester: Practical Application Company, 2000.
- XIAO, Z.; HU, W.; HUANG, X. Highlighting learning pathways for enhanced course recommendation using large language models. *In: Proceedings of the 10th International Conference on Systems and Informatics (Icsai).* Shanghai: Ieee, 2024. p. 1–8.