MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CIÊNCIAS DA SAÚDE DE PORTO ALEGRE
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIAS DA INFORMAÇÃO E GESTÃO EM SAÚDE

Felipe Augusto Prauchner Christmann

DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS DE CURSOS DE PÓS-GRADUAÇÃO DO UNA-SUS/UFCSPA PARA IDENTIFICAÇÃO DE FATORES QUE CONTRIBUEM NA REINTEGRAÇÃO DOS ALUNOS

Porto Alegre
2023

Felipe Augusto Prauchner Christmann

DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS DE CURSOS DE PÓS-GRADUAÇÃO DO UNA-SUS/UFCSPA PARA IDENTIFICAÇÃO DE FATORES QUE CONTRIBUEM NA REINTEGRAÇÃO DOS ALUNOS

Dissertação no Programa de Mestrado Acadêmico em Tecnologias da Informação e Gestão em Saúde da Universidade Federal de Ciências da Saúde de Porto Alegre.

Orientadora: Profa. Dra. Alessandra Dahmer

Porto Alegre

2023

Catalogação na Publicação

Christmann, Felipe
DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS DE CURSOS DE PÓS-GRADUAÇÃO DO UNA-SUS/UFCSPA PARA IDENTIFICAÇÃO DE FATORES QUE CONTRIBUEM NA REINTEGRAÇÃO DOS ALUNOS / Felipe Christmann. -- 2023.
62 p. : graf., tab. ; 30 cm.

Dissertação (mestrado) -- Universidade Federal de Ciências da Saúde de Porto Alegre, Programa de Pós-Graduação em Tecnologias da Informação e Gestão em Saúde, 2023.

Orientador(a): Alessandra Dahmer.

1. Mineração de Dados. 2. Mineração de Dados Educacionais. 3. Educação a Distância. 4. UNA-SUS. I. Título.

Sistema de Geração de Ficha Catalográfica da UFCSPA com os dados fornecidos pelo(a) autor(a).

Felipe Augusto Prauchner Christmann

DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS DE CURSOS DE PÓS-GRADUAÇÃO DO UNA-SUS/UFCSPA PARA IDENTIFICAÇÃO DE FATORES QUE CONTRIBUEM NA REINTEGRAÇÃO DOS ALUNOS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Tecnologias da Informação e Gestão em Saúde da Universidade Federal de Ciências da Saúde de Porto Alegre como requisito parcial para a obtenção do título de Mestre em Tecnologias da Informação e Gestão em Saúde.

Orientadora: Profa. Dra. Alessandra Dahmer

Aprovada em: _____ de _____ de _____.

BANCA EXAMINADORA:

_____

Prof. Dra. Fernanda Monteiro

Universidade de Brasília

_____

Prof. Dra. Márcia Rosa da Costa

Universidade Federal de Ciências da Saúde de Porto Alegre

_____

Prof. Dr. Sílvio César Cazella

Universidade Federal de Ciências da Saúde de Porto Alegre

AGRADECIMENTOS

Agradeço a Deus, em primeiro lugar.

Minha gratidão também à minha companheira Rosana pelo incentivo e imenso auxílio que permitiu a finalização deste trabalho.

Também agradeço aos meus pais, que permitiram que eu me dedicasse aos estudos quando foi necessário.

Aos meus amigos pela paciência, principalmente à Sandra que foi fundamental na etapa da qualificação.

RESUMO

Introdução: A disponibilidade de dados na educação através de cursos e plataformas online cresce cada vez mais. A oferta de cursos a distância contribui de forma significativa para a geração de dados que podem ser explorados pelo processo de Descoberta de Conhecimento em Base de Dados. **Objetivo**: Analisar as características dos alunos do Curso de Especialização em Saúde da Família do UNA-SUS/UFCSPA, identificando o perfil daqueles que enfrentaram dificuldades ao longo do curso e tiveram sucesso na sua formação através da utilização do processo de descoberta de conhecimento em base de dados. **Método**: A partir das planilhas de dados do Curso de Especialização em Saúde da Família do UNA-SUS/UFCSPA foram selecionados os atributos de interesse e aplicado o processo de Descoberta de Conhecimento em Base de Dados com as técnicas de Árvore de Decisão e Regras de Associação. Após análise dos dados foi possível obter e comparar os perfis dos alunos reintegrados e evadidos. **Resultado**: Obteve-se um conjunto de 16 perfis, 8 para os alunos reintegrados e 8 para alunos evadidos. Na análise por região, quando comparados os perfis de alunos reintegrados e evadidos, encontrou-se diferença entre todos os perfis, porém as características mais comuns foram o acesso à Internet e a presença na aula inaugural. Já na comparação dos perfis gerais (todas as regiões), estes se mostraram iguais. **Conclusões**: O processo de descoberta de conhecimento em base de dados foi aplicado com sucesso na base de dados educacionais do UNA-SUS/UFCSPA. Como resultado, foi possível obter perfis de alunos evadidos e reintegrados e realizar uma breve análise destes perfis, que podem servir de complemento para a criação de estratégias de mitigação da evasão nos cursos de ensino a distância.

Palavras-chave: Mineração de Dados. Mineração de Dados Educacionais. Educação a Distância. UNA-SUS.

ABSTRACT

Introduction*: The availability of data in education through online courses and platforms is growing steadily. The provision of distance learning courses significantly contributes to the generation of data that can be explored through the process of Knowledge Discovery in Databases.* ***Objective****: To analyze the characteristics of students in the Specialization Course in Family Health at UNA-SUS/UFCSPA, identifying the profile of those who faced difficulties during the course and successfully completed their education through the use of the knowledge discovery process in databases.* ***Method****: From the data spreadsheets of the Specialization Course in Family Health at UNA-SUS/UFCSPA, the attributes of interest were selected, and the Knowledge Discovery in Databases process was applied using Decision Tree and Association Rules techniques. After analyzing the data, it was possible to obtain and compare the profiles of reintegrated and dropout students.* ***Result****: A set of 16 profiles was obtained, 8 for reintegrated students and 8 for dropout students. In the regional analysis, when comparing the profiles of reintegrated and dropout students, differences were found among all profiles, but the most common characteristics were internet access and attendance at the opening lecture. In the comparison of overall profiles (all regions), they were found to be similar.* ***Conclusions****: The knowledge discovery process in databases was successfully applied to the educational database of UNA-SUS/UFCSPA. As a result, profiles of dropout and reintegrated students were obtained, and a brief analysis of these profiles was conducted, which can complement the creation of actions and strategies to mitigate dropout in distance learning courses.*

Keywords*: Data Mining. Educational Data Mining. Distance Education. UNA-SUS.*

LISTA DE QUADROS

Quadro 1 – Comparativo entre os estudos relacionados. 18
Quadro 2 – Amostra dos dados disponíveis na base de dados do UNA-SUS/UFCSPA. 22
Quadro 3 – Tipos de dados resultantes após tratamento inicial da base de dados. .28
Quadro 4 – Atributos novos, criados após a utilização da ferramenta Orange. 30
Quadro 5 – Quantidade de instâncias por atributo. 38
Quadro 6 – Regras resultantes após aplicação das Regras de Associação. 40
Quadro 7 – Comparação dos perfis de aluno reintegrado e evadido no Acre. 41
Quadro 8 – Comparação dos perfis de aluno reintegrado e evadido no Amazonas. 42
Quadro 9 – Comparação dos perfis de aluno reintegrado e evadido no Amapá. 43
Quadro 10 – Comparação dos perfis de aluno reintegrado e evadido no Pará. 43
Quadro 11 – Comparação dos perfis de aluno reintegrado e evadido em Roraima. .44
Quadro 12 – Comparação dos perfis de aluno reintegrado e evadido no Rio Grande do Sul. 44
Quadro 13 – Comparação dos perfis de aluno reintegrado e evadido em Sergipe. ..45
Quadro 14 – Comparação dos perfis de aluno reintegrado e evadido no Brasil. 45
Quadro 15 – Comparação dos perfis de alunos reintegrados. 46
Quadro 16 – Comparação dos perfis de alunos evadidos. 46

LISTA DE FIGURAS

Figura 1 – Processo de Mineração de Dados. 6
Figura 2 – Exemplo de Árvore de Decisão para concessão de aposentadoria. 7
Figura 3 – Exemplo de alarme utilizando Redes Bayesianas.....8
Figura 4 – Funcionamento de um perceptron e dos algoritmos SVM.....9
Figura 5 – Estrutura organizacional do Curso de EspSF. 14
Figura 6 – Dinâmica de trabalho do Apoio Acadêmico.....15
Figura 7 – Visualização da ferramenta Orange Data Mining. 24
Figura 8 – Regras utilizadas no Orange para a criação dos novos atributos. 29
Figura 9 – Widget utilizado para gerar as Regras de Associação. 31
Figura 10 – Regras de Associação tendo como alvo alunos evadidos. 32
Figura 11 – Árvore de Decisão geral para toda a amostra. 32
Figura 12 – Árvore de Decisão para toda a amostra, lado esquerdo. 33
Figura 13 – Árvore de Decisão para toda a amostra, lado direito. 33
Figura 14 – Nó da Árvore de Decisão que representa o estado do Acre. 34
Figura 15 – Nó da Árvore de Decisão que representa o estado do Amapá. 35
Figura 16 – Nó da Árvore de Decisão que representa o estado do Amazonas. 35
Figura 17 – Nó da Árvore de Decisão que representa o estado do Pará. 36
Figura 18 – Nó da Árvore de Decisão que representa o estado do Rio Grande do Sul. 36
Figura 19 – Nó da Árvore de Decisão que representa o estado de Roraima. 36
Figura 20 – Nó da Árvore de Decisão que representa o estado de Sergipe. 37
Figura 21 – Gráfico demonstrativo da quantidade de alunos aprovados, reintegrados, reprovados e evadidos. 40

LISTA DE ABREVIATURAS E SIGLAS

EaD – Educação a Distância

AVA – Ambiente Virtual de Aprendizagem

KDD – *Knowledge Discovery in Databases*

MDE – Mineração de Dados Educacionais

MD – Mineração de Dados

DAG – Gráfico Acíclico Dirigido

UNA-SUS – Universidade Aberta do Sistema Único de Saúde

UFCSPA – Universidade Federal de Ciências da Saúde de Porto Alegre

LGPD – Lei Geral de Proteção de Dados Pessoais

SVM – Máquina de Vetor de Suporte

EspSF – Especialização em Saúde da Família

ESF – Estratégia de Saúde da Família

CD – *Compact Disc*

SUMÁRIO

1. INTRODUÇÃO 1
1.1. Pergunta de Pesquisa 2
1.2. Justificativa 2
2. OBJETIVOS 4
2.1. Objetivo Geral 4
2.2. Objetivos Específicos 4
3. CONTEXTO E REFERENCIAIS 5
3.1. Descoberta de Conhecimento em Base de Dados 5
3.1.1. Mineração De Dados Educacionais 10
3.2. UNA-SUS/UFCSPA 13
3.2.1. Curso De Especialização Em Saúde Da Família 13
3.2.2. Apoio Acadêmico 15
3.2.3. Bases de Dados do UNA-SUS/UFCSPA 17
3.3. Estudos Relacionados 18
4. MATERIAIS E MÉTODOS 21
4.1. Materiais 21
4.2. Metodologia 23
4.3. A Ferramenta Orange Data Mining 24
4.4. A Escolha dos Algoritmos Utilizados 25
5. PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS EDUCACIONAL 26
5.1. Pré-processamento 26
5.2. Mineração de Dados 30
5.3. Pós-processamento 37
6. RESULTADOS E DISCUSSÃO 38
7. CONSIDERAÇÕES FINAIS 47

1. INTRODUÇÃO

O uso da Internet aumenta cada dia mais em todos os setores da sociedade, seja para fazer compras em uma loja online, assistir vídeos no Youtube ou para fazer um curso na modalidade Educação a Distância (EaD) em uma instituição de ensino. Como consequência dessa popularização, a geração de dados cresceu de forma explosiva (COSTA et al., 2012). No que tange à educação, um dos responsáveis pela geração de dados de forma abundante é o Ambiente Virtual de Aprendizagem (AVA), presente em muitas instituições educacionais e ferramenta fundamental de ensino na atualidade, principalmente na modalidade EaD.

Estas ferramentas se tornaram ainda mais relevantes durante o período pandêmico da COVID-19, onde a impossibilidade do ensino presencial obrigou as instituições de ensino a migrar para plataformas online para dar continuidade às suas atividades (DOSEA et al., 2020).

No entanto, o EaD também traz algumas questões que são particulares deste cenário. A realização de um curso através de uma plataforma online exige do estudante acesso a um computador e à Internet realidade essa que não é comum em grande parte do território nacional. Algumas regiões mais distantes dos centros urbanos sofrem com problemas de infraestrutura e acesso aos meios de comunicação e à Internet, sendo uma dificuldade extra na hora de realizar um curso na modalidade EaD.

Estas questões acabam agravando os altos índices de abandono de curso que historicamente existem. Segundo o Censo EAD.BR de 2020 (ABED, 2022), estima-se que até 25% dos alunos de cursos EaD evadam. Para combater tal problema, algumas soluções têm sido buscadas por gestores e/ou professores das instituições de ensino. Um exemplo é o mapeamento do perfil dos alunos, onde as suas tendências dentro dos sistemas de ensino possam ser classificadas para que haja o desenvolvimento de ações para mitigar a possível evasão (BEHAR; KELLEN; DA SILVA, 2012). Outro exemplo que pode ser citado é a aferição do estado emocional do estudante, que pode estar apresentando uma deficiência no processo de aprendizagem devido a diversas situações (DA COSTA; CAZELLA; RIGO, 2015).

Uma iniciativa que se enquadra nesse cenário é a Universidade Aberta do SUS (UNA-SUS) (BRASIL; MINISTÉRIO DA SAÚDE, 2023), criada em 2008 pelo

Ministério da Saúde e destinada a suprir as demandas da educação contínua dos profissionais da saúde no Brasil. Este sistema de ensino visa abordar a escassez, uma vez que menos de 2% dos médicos do país possuem qualificações específicas para desempenhar funções como médicos de família e comunitários. Devido à vastidão territorial do Brasil, a oferta dos cursos é feita preferencialmente via EaD, visando assegurar uma abordagem pragmática e dinâmica para esses cursos (AUGUSTO et al., 2018).

A partir de 2009, a Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA) se uniu à Rede UNA-SUS. Nesse contexto, a UNA-SUS/UFCSPA inaugurou a disponibilização do seu programa de Especialização em Saúde da Família (EspSF), voltado exclusivamente para grupos compostos por profissionais que estão ativamente envolvidos na Estratégia de Saúde da Família (ESF) no estado do Rio Grande do Sul, no ano de 2015. No mesmo ano, essa oferta foi estendida para uma turma no Paraná e, posteriormente, em 2012, para turmas no Pará e em Sergipe e, em 2014, os demais estados da região Norte passaram a ser atendidos pelo curso da UNA-SUS/UFCSPA. (PORTELLA, 2019)

O acompanhamento das turmas oferecidas pela UNA-SUS/UFCSPA é feito por uma equipe pedagógica, que coleta informações relacionadas aos alunos e as utiliza para monitorar o seu andamento no curso (COSTA et al., 2015). Ao investigar estas turmas, este trabalho procura mapear o perfil do aluno que obteve sucesso na sua recuperação durante o curso, com o intuito de fornecer informações úteis no desenvolvimento de ações para diminuir a evasão. Para este fim, será utilizado o processo de Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases - KDD*), que possibilita analisar conjuntos de dados para extração de conhecimento relevante.

1.1. Pergunta de Pesquisa

“Como podemos caracterizar o perfil de um aluno reintegrado em um curso na modalidade EaD da área da saúde, utilizando o processo de Descoberta de Conhecimento em Base de Dados?”

1.2. Justificativa

Os cursos EaD possuem altos índices de evasão e as ferramentas tradicionais de gestão da educação não trazem novas soluções para este problema. Este fato associado à grande quantidade de dados dos alunos nas plataformas educativas, oferecem oportunidades para a pesquisa de extração de conhecimento através do processo de Descoberta de Conhecimento em Base de Dados. Neste sentido, é comum encontrar pesquisas que exploram o perfil do aluno evadido, buscando predizer o desfecho desse aluno ou as causas dessa evasão, auxiliando as equipes gestoras e melhorando o desempenho acadêmico. Buscando novas visões sobre o tema, este trabalho procura explorar um perfil diferente de aluno: aquele que enfrentou dificuldades ao longo do curso, no entanto, conseguiu finalizá-lo.

O projeto UNA-SUS/UFCSPA possui um acervo grande de cursos EaD ministrados, contemplando mais de 4000 alunos, oferecendo uma oportunidade para a investigação dos aspectos que tornaram possível esta retomada dos estudos por parte dos alunos com estas características, que concluíram o curso de EspSF. Utilizando o processo de Descoberta de Conhecimento em Base de Dados, espera-se que seja possível conhecer um pouco mais sobre o perfil deste aluno, contribuindo para o desenvolvimento de ações pelos profissionais especializados que administram e aplicam cursos à distância na área da saúde.

Os capítulos dessa dissertação estão distribuídos da seguinte forma: O capítulo 1 apresenta a introdução com o contexto da pesquisa e sua relevância, a definição da pergunta de pesquisa e a justificativa para a realização deste trabalho. O capítulo 2 apresenta os objetivos gerais e específicos. O capítulo 3 aborda o referencial teórico sobre o processo de Descoberta de Conhecimento em Base de Dados. No capítulo 4 é apresentado um resumo dos estudos relacionados a esta pesquisa. O delineamento metodológico e os materiais utilizados nesta pesquisa estão no capítulo 5. O capítulo 6 discorre sobre os resultados e apresenta a discussão dos mesmos e o capítulo 7 apresenta a conclusão da pesquisa.

2. OBJETIVOS

2.1. Objetivo Geral

O principal objetivo do trabalho é analisar as características dos alunos do Curso de EspSF do UNA-SUS/UFCSPA através do processo de Descoberta de Conhecimento em Base de Dados, extraindo o perfil daqueles que enfrentaram dificuldades ao longo do curso e tiveram sucesso na sua formação.

2.2. Objetivos Específicos

Os objetivos específicos deste trabalho são listados a seguir:

1. Analisar as fontes de dados do Curso de Especialização em Saúde da Família do UNA-SUS/UFCSPA;
2. Pré-processar os dados;
3. Definir a característica do aluno reintegrado para esta pesquisa;
4. Aplicar as técnicas de Árvores de Decisão e Regras de Associação;
5. Descrever os dados através de modelos a fim de obter o perfil desejado;
6. Avaliar o conhecimento extraído.

3. CONTEXTO E REFERENCIAIS

3.1. Descoberta de Conhecimento em Base de Dados

A Mineração de Dados (MD), é uma das etapas de um processo chamado Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases - KDD*) e inclui o uso de algoritmos e técnicas de análise e estatística para o processamento de informação. De acordo com Fayyad et al. (1996), “KDD é o processo não trivial de identificação de padrões, a partir de dados, que sejam válidos, novos, potencialmente úteis e compreensíveis”.

Segundo Baker et al. (2011), os termos KDD e MD são sinônimos. Portanto, será utilizado neste trabalho o termo Mineração de Dados para se referir a todo o processo de descoberta de conhecimento.

De acordo com Aggarwal (2015), o processo de MD pode ser descrito com as seguintes etapas:

Coleta de dados: Dependendo da fonte, a coleta de dados pode exigir o uso de dispositivos como sensores e pode ser feita manualmente através de pesquisas ou documentos que necessitem de estruturação ou ainda através do auxílio computacional para obtenção dos dados, no caso de aplicações web ou mídias sociais. Esta etapa é específica da plataforma utilizada e geralmente está fora do domínio do analista de mineração de dados. Após a coleta, os dados podem ser armazenados em um banco de dados para processamento.

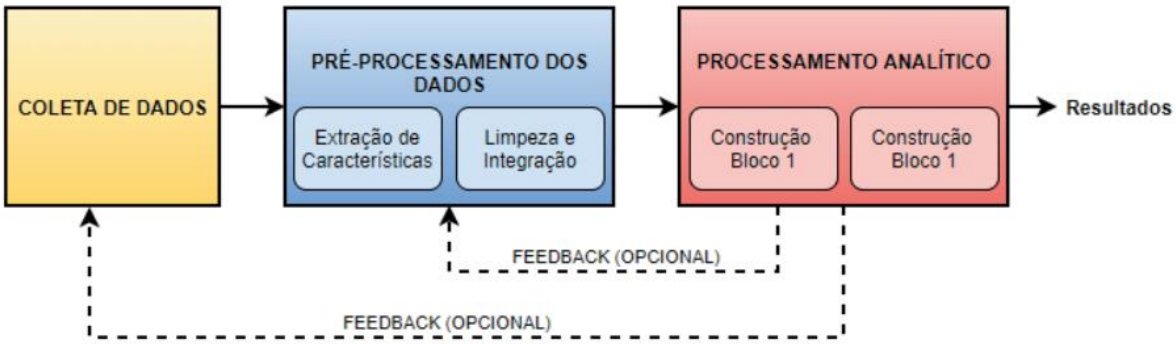
Extração de recursos e limpeza de dados (pré-processamento e transformação): Quando os dados são coletados, eles geralmente não estão em um formato adequado para processamento e, para tal, é essencial transformá-los em um formato que seja interpretável pelos algoritmos de mineração, como multidimensionais, séries temporais ou formato semiestruturado. O formato multidimensional é o mais comum, no qual diferentes campos de dados correspondem às diferentes propriedades medidas, que são chamadas de atributos. A fase de extração de recursos geralmente é realizada em paralelo com a limpeza de dados, onde partes ausentes e incorretas dos dados são estimadas ou corrigidas. Em muitos casos, os dados podem ser extraídos de várias fontes e precisam ser integrados em um formato unificado para processamento. O resultado deste procedimento é um conjunto de dados estruturados, que pode ser efetivamente

usado por um programa de computador. Depois da fase de extração de recursos, os dados podem ser armazenados novamente em um banco de dados para processamento.

Processamento analítico e algoritmos: A parte final do processo de mineração consiste em projetar métodos analíticos eficazes a partir dos dados processados.

A sequência das etapas do processo de MD proposto por Aggarwal (2015) é apresentada na Figura 1.

Figura 1 – Processo de Mineração de Dados.



Fonte: (AGGARWAL, 2015).

A exploração desses dados permite, por exemplo, a geração de sistemas de recomendação, auxilia na tomada de decisão e torna humanamente compreensíveis informações que estavam fora de contexto ou inacessíveis de forma direta.

Em relação à terceira etapa, diversos algoritmos podem ser utilizados, como por exemplo, Árvores de Decisão e Redes Neurais. Problemas diferentes terão resultados diferenciados conforme o algoritmo utilizado, portanto é importante testar as diferentes técnicas para se buscar sempre o melhor resultado. Este processo normalmente é repetido diversas vezes, não só pela melhoria da eficiência e acurácia, mas também para reajustar os parâmetros nas etapas anteriores, como por exemplo, a seleção de variáveis que irão compor a mineração de dados.

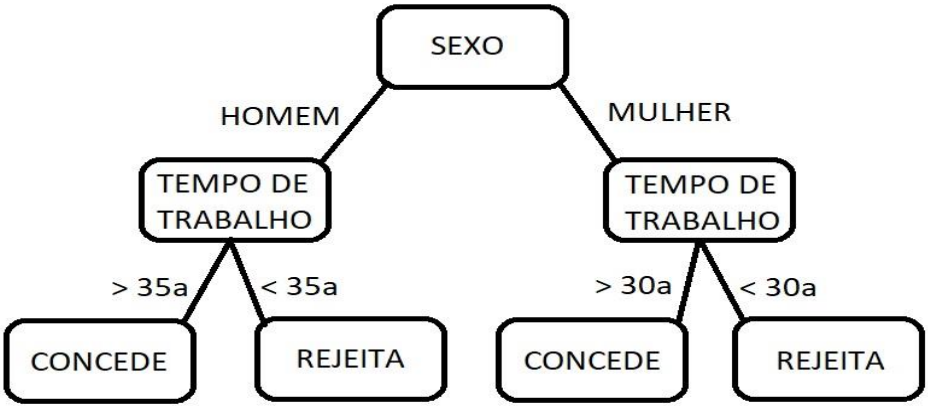
Diversas técnicas de Mineração de Dados estão disponíveis, no entanto, algumas merecem destaque e são listadas a seguir:

K-vizinhos mais Próximos: Este tipo de técnica utiliza a semelhança entre objetos para realizar uma classificação. Dado um objeto qualquer *O*, utiliza a

comparação com outros objetos segundo critérios de semelhança que podem ter um peso, por exemplo. Este processo também é conhecido por clusterização e busca agrupar dados por semelhança. Assim, o algoritmo faz a classificação do objeto alvo baseado nos objetos próximos (semelhantes) a ele.

Árvore de Decisão: Tem formato de uma árvore, onde a raiz contém os objetos a serem classificados e as ramificações (nós) são atributos de semelhança destes objetos. Os objetos vão percorrendo os nós até chegarem na folha (final do caminho), que vai definir a classificação deste objeto. A Figura 2 mostra um exemplo desta técnica, onde se classifica a concessão de aposentadoria para homens e mulheres.

Figura 2 – Exemplo de Árvore de Decisão para concessão de aposentadoria.



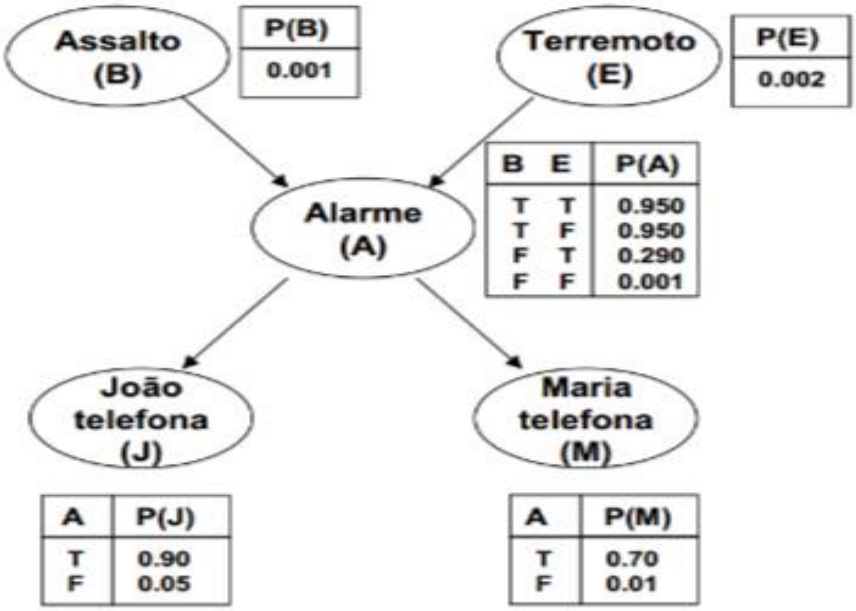
Fonte: Autoria própria (2023).

Redes Bayesianas e Classificadores Bayesianos: Dada uma série de variáveis *A,* com um domínio de valores mutuamente exclusivos que pode ser assumido por *A,* estas variáveis estão ligadas por um Grafo Acíclico Dirigido (DAG) e possuem uma probabilidade de transição associada de uma variável anterior *B* para uma variável seguinte *A* onde P(*A*|*B*1, ... ,*B*n). As probabilidades de cada variável são descritas por uma tabela.

A Figura 3 demonstra o exemplo de um alarme que utiliza Redes Bayesianas. Há dois eventos causais, Assalto (B) e Terremoto (E) com probabilidades de

ocorrência P(B) e P(E). A ocorrência de um ou ambos faz disparar o alarme, com probabilidade P(A|B,E). Este evento encadeia duas possibilidades: João telefona (J) com probabilidade P(J) e Maria telefona (M) com probabilidade P(M). As tabelas

Figura 3 – Exemplo de alarme utilizando Redes Bayesianas.



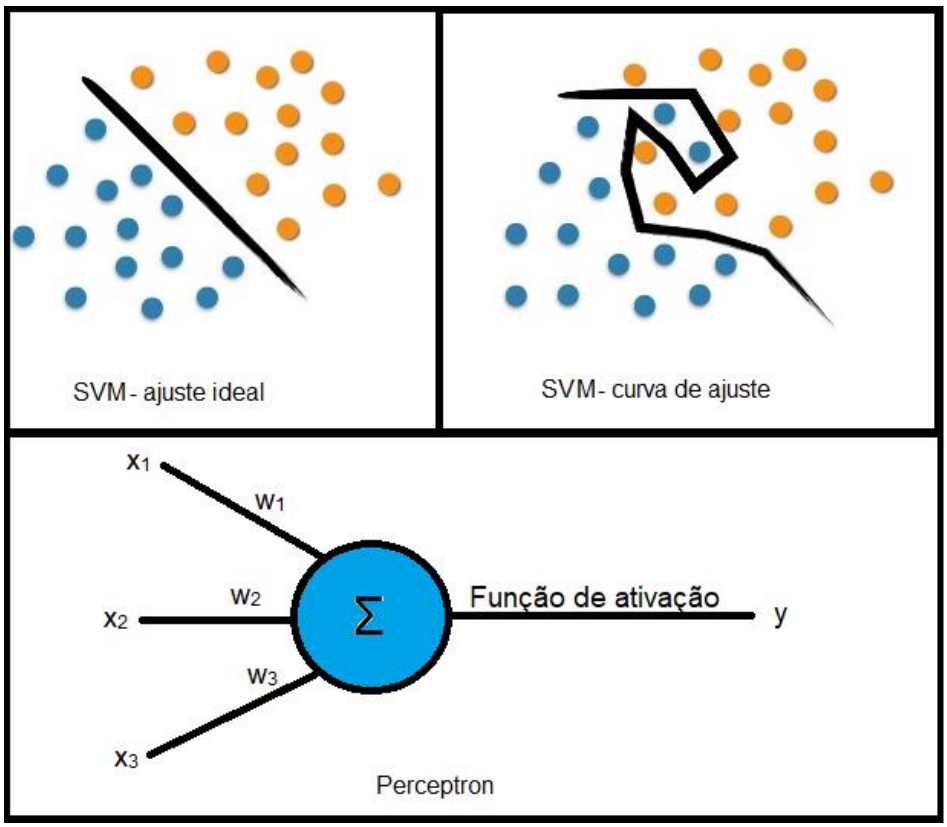
demonstram as probabilidades de ocorrência dos eventos.

Fonte: Adaptado de Russell e Norvig (1995).

***Perceptrons*, Máquina de Vetor de Suporte (SVM) e Redes Neurais Artificiais:** São baseados no funcionamento dos neurônios, sendo os dendritos as entradas (*x*), que possuem um peso (*w*) associado e o axônio a saída (*y*), que propaga o sinal multiplicando os valores dos sinais de entrada pelos pesos (*x.w*) e somando as entradas. A saída (y) possui uma função de ativação, que indicará se o neurônio irá ativar ou não a sua saída. Esta entidade é chamada de *perceptron*. Um *perceptron* pode ser utilizado como um classificador binário. Em Redes Neurais Artificiais, temos vários *perceptrons* dispostos em camadas e que formam uma rede, ligados pelos dendritos e pelos axônios (um axônio liga-se com um dendrito de cada neurônio da próxima camada), com um único *perceptron* na última camada para fornecer um resultado. Já em SVM, o algoritmo procura ajustar uma curva que melhor separa dois grupos de objetos conforme as suas características. Os objetos

são acomodados em lados opostos da curva realizando assim uma categorização dos objetos. A Figura 4 ilustra algoritmos SVM e um *perceptron*.

Figura 4 – Funcionamento de um perceptron e dos algoritmos SVM.



Fonte: Autoria própria (2023).

Os algoritmos podem ser *supervisionados* ou *não supervisionado*s, necessitando ou não da interação humana para a sua execução. A intervenção humana é essencial na análise dos resultados da aplicação dos algoritmos.

Regras de Associação: são frequentemente utilizadas para a fundamentação de campanhas de *marketing*, por exemplo. Consistem, neste caso, em associar o comportamento do consumidor à compra de um produto específico. É utilizada em dados categóricos e é um processo de aprendizado não supervisionado. Formalmente, define-se que “conjuntos de itens frequentes podem ser usados para gerar regras de associação, com o uso de uma medida conhecida como *confiança*. A confiança de uma regra X ⇒ Y é a probabilidade condicional de

que uma transação contém o conjunto de itens Y, dado que contém o conjunto X. A probabilidade é estimada dividindo o suporte do conjunto de itens X ∪ Y pelo do conjunto de itens X.” (AGGARWAL, 2015).

3.1.1. Mineração De Dados Educacionais

A geração de dados com plataformas de ensino online e o uso de recursos computacionais vem crescendo dentro da área de educação. Através de fóruns de discussão e dúvidas, a interação entre alunos e professores e o consumo de material didático oportuniza a exploração destes dados para a melhoria da qualidade no processo de ensino-aprendizagem. Sendo assim, a Mineração de Dados Educacionais (MDE) surge como uma nova área para desenvolver e adaptar métodos e algoritmos de MD para compreensão de dados dentro do contexto educacional.

Segundo a Sociedade Internacional de Mineração de Dados Educacionais (2022):

> A mineração de dados educacionais é uma disciplina emergente, preocupada em desenvolver métodos para explorar os dados únicos e cada vez mais em grande escala que vêm de ambientes educacionais e usar esses métodos para entender melhor os alunos e os ambientes em que eles aprendem.

Com o crescente volume de dados gerados, se faz imprescindível o uso de ferramentas computacionais para a análise e extração de conhecimento. Assim como a MD, a MDE também possui etapas bem definidas para o processo de extração de conhecimento. Segundo Souza e Souza (2021), as seguintes etapas fazem parte da MDE:

- Definição da função da MDE: Define-se nesta etapa o objetivo da MDE, como por exemplo definição do perfil do aluno, sistemas de recomendação etc.;
- Formatação dos dados que serão utilizados: Os dados utilizados podem ser provenientes de diversas fontes e sem uma formatação adequada. Sendo assim, esta etapa prepara os dados para serem analisados pelas ferramentas computacionais;
- Definição das técnicas de MDE: Conforme os objetivos estabelecidos, a escolha da técnica a ser aplicada é de suma importância para o sucesso na extração

de conhecimento. Dentre as técnicas mais utilizadas temos: Estatística Descritiva, Aprendizagem de Máquina e Aprendizagem Profunda.

- Delineamento de como essas técnicas serão aplicadas: Nesta etapa são selecionadas as ferramentas que vão dar suporte ao desenvolvimento de sistemas capazes de processar os dados e gerar os resultados esperados.

Além das etapas apresentadas anteriormente, podemos classificar a MDE com a seguinte taxonomia, de acordo com Costa et al. (2012):

- **Predição**
  - Classificação
  - Regressão
  - Estimação de Densidade
- **Agrupamento**
- **Mineração de Relações**
  - Mineração de Regras de Associação
  - Mineração de Correlações
  - Mineração de Padrões Sequenciais
  - Mineração de Causas
- **Destilação de Dados para Facilitar Decisões Humanas**
- **Descobertas com Modelos**

Na área de **predição**, o interesse está em desenvolver modelos a fim de descobrir as variáveis (preditoras) que mais influenciam em um certo aspecto pré-selecionado do problema, que pode ser representado por uma ou mais variáveis (preditoras). A predição exige que as variáveis preditivas sejam codificadas para que seja possível sua relação com as variáveis preditoras. Os principais tipos de predição são: classificação, regressão e estimação de densidade. A predição é largamente utilizada na MDE, principalmente nas questões ligadas à evasão escolar.

Na área de **agrupamento**, o interesse está em agrupar as variáveis em comum, classificando-as em categorias. Estes grupos não são conhecidos previamente e podem ser identificados pela manipulação das características dos dados.

Na área de **mineração de relações**, o objetivo está em se descobrir a força da relação entre variáveis do banco de dados através da mineração. Neste caso, pode ser de quatro tipos:

- Regras de Associação: Neste tipo de mineração, procura-se associar o valor de uma variável X ao valor de outra variável Y. Estes valores devem ser binários, como por exemplo sim ou não, atingiu ou não atingiu, verdadeiro ou falso etc.
- Correlações: Aqui o objetivo é encontrar correlações lineares, como positivo e negativo, entre as variáveis analisadas.
- Sequências: Nesta mineração, busca-se o fator temporal na associação das variáveis analisadas, onde uma sequência de eventos desencadeia alguma variável afirmativa, como por exemplo, o mau comportamento (não prestar atenção na aula, não se preparar para a prova etc.) de um aluno leva ao insucesso na aprendizagem.
- Causa: Desenvolve-se algoritmos e técnicas para verificar se um evento causa outro evento através da análise dos padrões de covariância. Retomando o exemplo anterior, a mineração de causa permite verificar se o aluno não aprende por não se esforçar ou não se esforça por não estar aprendendo. Este tipo de mineração tem especial atenção neste trabalho por se adequar ao tipo de questão abordado.

Na área de **destilação de dados para facilitar decisões humanas**, utiliza-se a mineração para apresentar de forma mais legível um grande conjunto de dados, através de recursos visuais como gráficos e tabelas, para auxiliar na interpretação e significação destes dados.

Neste trabalho, o interesse está em investigar as características do aluno que foi aprovado, porém estava com dificuldades durante o curso. Algumas das técnicas apresentadas podem ser utilizadas para explorar este aspecto dos dados. Neste caso, as técnicas de Agrupamento são as mais indicadas, pois concentram as características ou objetos de estudo em grupos similares, permitindo assim, que se encontre um perfil comum entre os estudantes.

Outra técnica que pode ser útil são as Regras de Associação, que buscam identificar através da frequência de itens ou de características de um objeto se estas ocorrem em outro objeto. Como já citado, as Regras de Associação são frequentemente utilizadas na área de *marketing*, onde se procura associar a compra de um produto a outro produto, ou seja, na ocorrência de um objeto, quantas vezes outro objeto também foi comprado pelo mesmo cliente. No caso desta pesquisa, os atributos que compõem o perfil servirão para fazer uma associação ao resultado do aluno ao final do curso, com foco naqueles classificados como reintegrados por essa

pesquisa.

Para a aplicação destes modelos, destaca-se a ferramenta Weka, muito utilizada no meio acadêmico, que se caracteriza por ser uma ferramenta visual, com uma grande quantidade de algoritmos implementados e boa documentação, permitindo que pesquisadores não tão familiarizados com algoritmos e programação possam desenvolver pesquisas nesta área. Apesar disso, o Weka possui algumas limitações como, por exemplo, ser desenvolvido na linguagem Java, o que limita muitas vezes sua utilização em bases maiores e apresenta baixa performance, em comparação a outras ferramentas.

3.2. UNA-SUS/UFCSPA

O sistema UNA-SUS (Universidade Aberta do Sistema Único de Saúde) disponibiliza cursos de capacitação e especialização para os profissionais da saúde que atuam no Sistema Único de Saúde (SUS). Está presente em todo o Brasil, através de parcerias com as universidades públicas federais. De acordo com seu site (BRASIL; MINISTÉRIO DA SAÚDE, 2023), possui mais de 5,5 milhões de matrículas desde seu início e já ofereceu mais de 370 cursos na modalidade EaD, que são financiados pelo Ministério da Saúde. O sistema UNA-SUS é composto por três elementos: a Rede Colaborativa de Instituições de Ensino Superior, o Acervo de Recursos Educacionais em Saúde - ARES e a Plataforma Arouca.

O Curso de EspSF é oferecido a três categorias de profissionais da saúde: enfermeiros, médicos e dentistas. Estes profissionais fazem parte de programas do governo brasileiro como o PROVAB, criado em 2011 com o objetivo de incentivar médicos a atuarem em áreas carentes de assistência médica, principalmente nas regiões mais remotas e desfavorecidas do país, e o Mais Médicos, criado para suprir a carência de médicos especialmente em áreas remotas, rurais e desfavorecidas do país. Um dos requisitos para a participação dos programas era a realização do Curso de EspSF.

3.2.1. Curso De Especialização Em Saúde Da Família

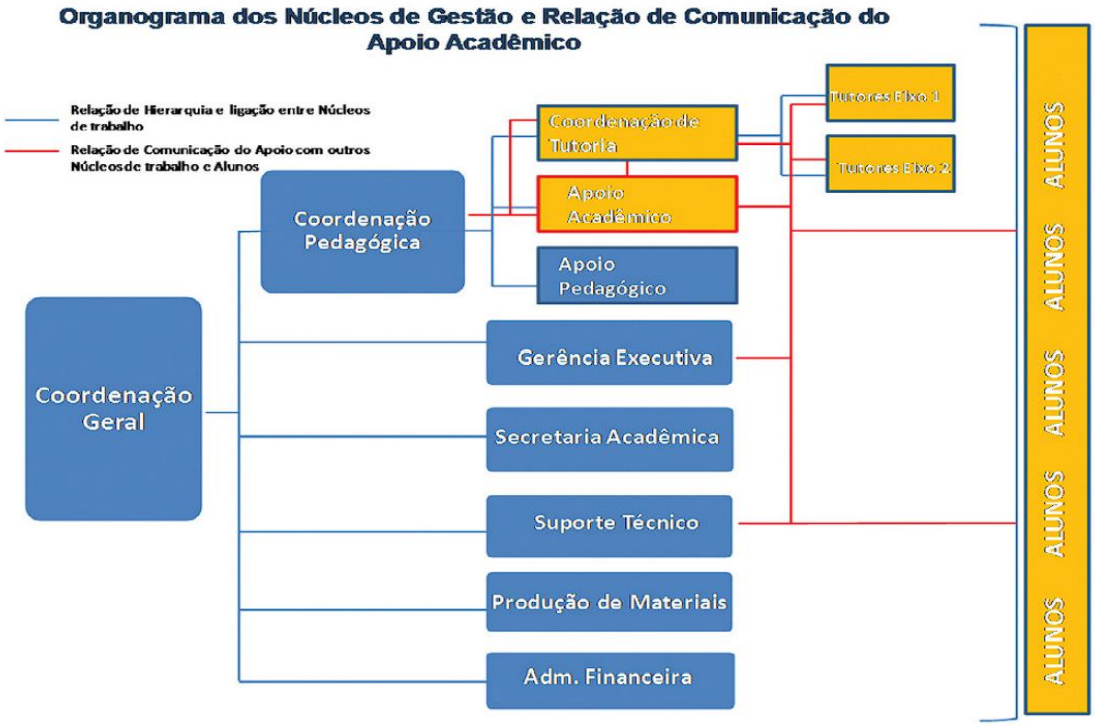
A UFCSPA é parceira do projeto UNA-SUS oferecendo alguns cursos, dentre

eles, o Curso de EspSF. Este curso, conforme legislação educacional, apresenta avaliações a distância e presenciais, realizadas durante o período de 12 meses e está organizado em dois eixos temáticos:

- Eixo 1 - Campo da Saúde Coletiva: Essa área é composta por três unidades distintas: Organização da Atenção à Saúde, Epidemiologia Aplicada e Interações em APS (Atenção Primária à Saúde). Além disso, há uma introdução à Instrumentalização em EaD, onde os alunos se familiarizam com as ferramentas empregadas no ensino a distância.
- Eixo 2 - Núcleo Profissional: Esta área concentra-se em tópicos e conteúdos específicos para cada uma das três profissões (Enfermagem, Medicina e Odontologia). Ela apresenta o material através de Casos Complexos, que abordam questões relacionadas ao cuidado de grupos como Crianças e Adolescentes, Mulheres, Adultos e Idosos, bem como a Saúde Mental e Situações de Urgência na APS. Esses casos são ambientados em cenários fictícios urbanos, representando o cotidiano dos profissionais. Isso permite que os alunos estabeleçam conexões entre os conteúdos e suas próprias experiências, promovendo reflexões e aprimorando suas práticas.

Para a oferta dos cursos, a estrutura é composta pelas coordenações geral e pedagógica, pelos setores de apoio acadêmico e pedagógico, pelos tutores e sua coordenação, gerência executiva, secretaria acadêmica, suporte técnico, produção de materiais e administração financeira. A Figura 5 mostra a estrutura organizacional do curso.

Figura 5 – Estrutura organizacional do Curso de EspSF.



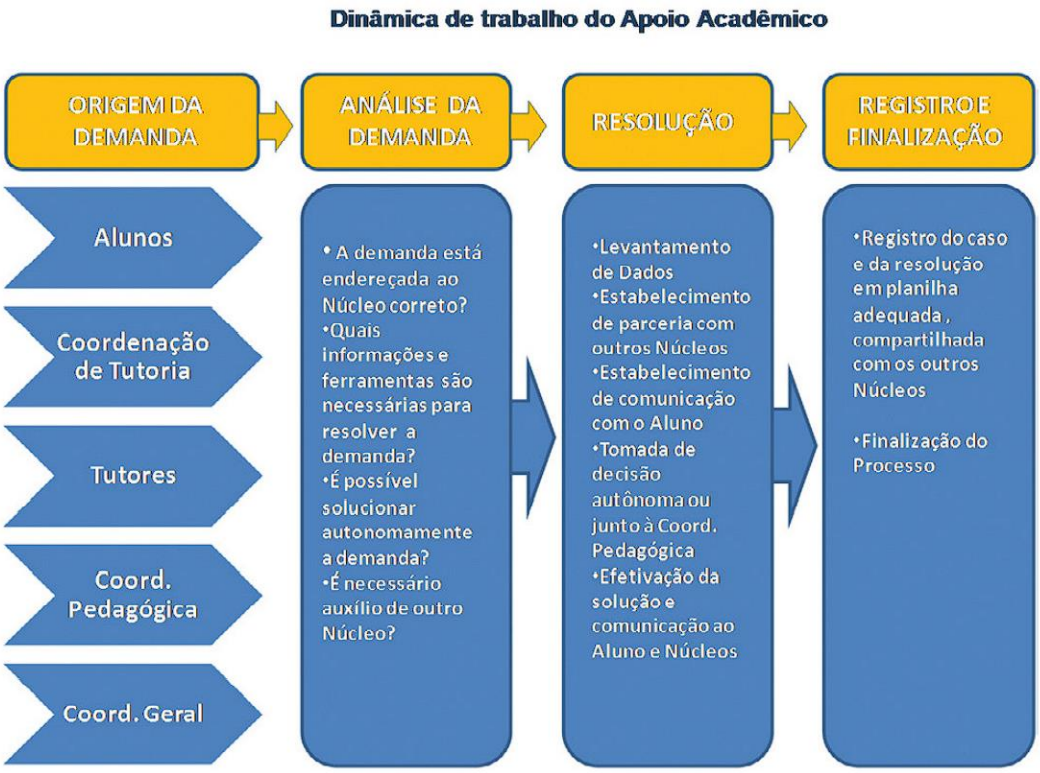
Fonte: (COSTA et al., 2015).

O curso analisado neste estudo teve duração de 1 ano e foi realizado majoritariamente à distância. Havia a exigência da presença do aluno em duas ocasiões: aula inaugural, onde era explicado o funcionamento do curso e a aula do Eixo 2, onde era aplicada uma prova relativa ao primeiro eixo temático do curso. Todos os alunos deveriam ainda elaborar um Trabalho de Conclusão de Curso (TCC), que demandava 20h de atividades.

3.2.2. Apoio Acadêmico

O Apoio Acadêmico do UNA-SUS/UFCSPA é um núcleo de trabalho diretamente vinculado à coordenação pedagógica e possui interação com diversos outros núcleos. Focado na resolução de problemas e demandas de acompanhamento dos alunos, o Apoio Acadêmico segue o fluxo de trabalho demonstrado na Figura 6:

Figura 6 – Dinâmica de trabalho do Apoio Acadêmico.



Fonte: (COSTA et al., 2015).

Segundo Costa et al. (2015), algumas das atribuições do Apoio Acadêmico incluem:

- Acompanhar o desempenho e a frequência dos alunos no curso, realizando contato com os alunos que enfrentam problemas ou dificuldades em relação ao andamento do curso, com o objetivo de auxiliá-los em suas dificuldades;
- Elaborar planos de estudo para alunos que estão com atividades atrasadas, se e quando necessário;
- Resgatar os alunos com baixa frequência ou rendimento com os quais os tutores não conseguem estabelecer comunicação;
- Acolher e encaminhar ao suporte técnico os casos de dificuldades no acesso à plataforma e outros problemas relacionados ao uso das ferramentas do ambiente virtual.

Contando com uma equipe de psicólogos dedicados, o Apoio Acadêmico busca estimular os alunos a refletirem sobre sua relevante contribuição no processo de ensino-aprendizagem. Adicionalmente, o suporte acadêmico segue de perto o desenvolvimento dos alunos, auxiliando na implementação das estratégias delineadas pela Coordenação Pedagógica. Esse acompanhamento é documentado

meticulosamente em planilhas, as quais registram um histórico abrangente de interações e abordagens.

Além deste registro em planilhas, também são disponibilizados aos alunos vários canais de comunicação, entre eles um sistema de abertura de *tickets* para resolução de problemas. Tal abordagem permite a administração eficiente do programa de formação e a organização coerente da base de dados correspondente.

3.2.3. Bases de Dados do UNA-SUS/UFCSPA

O UNA-SUS/UFCSPA disponibiliza o acesso aos cursos através da plataforma Moodle, acrônimo de *Modular Object-Oriented Dynamic Learning Environment* (ambiente modular de aprendizagem dinâmica orientada a objetos). O Moodle é uma aplicação baseada em web, escrita na linguagem PHP e que utiliza um banco de dados MySQL.

Neste banco de dados constam informações relevantes sobre os alunos, como por exemplo, a frequência de acesso à plataforma de estudos, se o aluno acessou os materiais disponíveis e se efetuou as tarefas propostas. Informações relacionadas ao perfil do aluno também são acessíveis por meio deste banco de dados, como idade, local de residência, profissão, entre outras.

Além disso, a equipe de apoio possui planilhas de controle dos alunos com informações importantes, como por exemplo, sua situação e suas dificuldades durante o percurso dentro do curso. Por fim, o software de abertura de chamados OSTicket, utilizado pela equipe de apoio para gerenciar o atendimento aos alunos, também possui uma base de dados própria com o registro dos chamados abertos e sua resolução.

Entretanto, ao considerar estas fontes de dados, constatou-se que explorar todas as opções demandaria um grande esforço para análise e exigiria um tempo consideravelmente maior do que o disponível para uma pesquisa de mestrado. Além disso, este pesquisador enfrentou dificuldades no acesso aos bancos de dados da plataforma Moodle e do software OSTicket, devido a problemas de infraestrutura.

Sendo assim, optou-se por explorar apenas as planilhas utilizadas pelo Apoio Acadêmico e demais setores do curso, que já possuíam informações relevantes e que seriam buscadas nas demais fontes de dados.

Foi autorizado a este pesquisador o acesso aos dados de dezesseis turmas

diferentes oferecidas a partir de 2011, de onde foram selecionadas três para a realização da pesquisa.

3.3. Estudos Relacionados

A presente seção apresenta alguns trabalhos relacionados ao tema desta pesquisa, bem como traz uma breve comparação entre os métodos empregados por cada um e analisa seus resultados de forma sucinta.

Abaixo, é apresentado o Quadro 1, com a listagem dos trabalhos classificados como mais relevantes.

Quadro 1 – Comparativo entre os estudos relacionados.

Estudo/Autor	Algoritmos Utilizados	Foco do Estudo				Identifica Perfil do Aluno
		EaD	Reprovação	Evasão	Retenção	
Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados [Manhães et al., 2011]	*Multilayer Perceptron (MLP)*	Sim	Não	Sim	Não	Não
Minerando dados sobre o desempenho de alunos de cursos de educação permanente em modalidade EaD: Um estudo de caso sobre evasão escolar na UNA-SUS [Da Costa; Cazella; Rigo, 2015]	Árvore de Decisão	Sim	Não	Sim	Não	Sim
Análise de evidências do estado de ânimo desanimado de alunos de um AVEA: uma proposta a partir da aplicação de regras de associação [Santos et al., 2015]	Regras de Associação	Sim	Nâo	Sim	Nâo	Sim
Identificando o perfil de evasão de alunos de graduação através da Mineração de dados Educacionais: um estudo de caso de uma Universidade Comunitária [Paz; Cazella, 2017]	Árvore de Decisão	Não	Não	Sim	Não	Sim
Análise dos problemas da Evasão e Retenção: Uma abordagem através de Mineração de Dados Educacionais [Do Couto; De Santana, 2017]	Arvore de Decisão, *Multilayer Perceptron (MLP),* Redes Bayesianas	Não	Não	Sim	Sim	Não

Utilização de técnicas de Mineração de Dados Educacionais para a predição de desempenho de alunos de EaD em Ambientes Virtuais de Aprendizagem [Rabelo et al., 2017]	Árvore de Decisão	Sim	Sim	Não	Não	Não
Um Modelo para Predição de Reprovação de Aprendizes na Educação a Distância usando Árvore de Decisão [Ferreira et al., 2019]	Árvore de Decisão	Sim	Sim	Não	Não	Sim
Mineração de Dados Educacionais para identificação de Perfil de Retenção em um curso de Ciência da Computação [Rolim; Silva, 2021]	*Support Vector Machine (SVM), KNearest Neighbor (KNN)*	Não	Não	Não	Sim	Sim

Fonte: Autoria própria (2023).

As técnicas encontradas nos estudos variaram, no entanto, em sua maioria foram utilizadas Árvores de Decisão.

No estudo de Manhães et al. (2011), os autores concluíram que a análise realizada é viável e que os experimentos retornaram dados com acurácia média variando entre 75% e 80%, na previsão do risco de evasão.

A pesquisa de Da Costa, Cazella e Rigo (2015), mostra que o treinamento do modelo foi realizado com um conjunto de 248 instâncias e o método de teste foi o *cross-validation*, assumindo o valor de 10 *folds*. O modelo preditivo gerado para evasão classificou corretamente 242 (97,6%) instâncias e incorretamente 6 (2,4%) instâncias.

Como resultado do estudo de Santos et al. (2015), que focava no desânimo dos alunos em relação ao curso, as Regras de Associação geradas apontaram que os alunos que se sentiram desanimados enviaram menos de 50% das atividades e tiveram dificuldades com o conteúdo, o que refletiu no conceito final.

No estudo de Paz e Cazella (2017), através dos experimentos realizados, foi verificado um alto grau de precisão, atingindo acurácia superior a 90%. Além disso, evidencia-se a confirmação da percepção inicial de que o incentivo e o currículo dos alunos estão diretamente relacionados à tendência de evasão, e que estas ocorrem geralmente nos semestres iniciais dos cursos com alunos sem incentivo.

Em outro estudo, Do Couto e De Santana (2017) mostram como resultado que o método *Random Forest* apresentou a melhor acurácia, superior a 87%. No entanto, os autores escolheram como melhor algoritmo o classificador *Bayesian*

Network, pela precisão acima de 85% e por ser capaz de possuir fácil representação de resultados, além de poder expressar o conhecimento do especialista sobre o domínio estudado.

No estudo de Rabelo et al. (2017), os autores concluíram que utilizando as técnicas e metodologia propostas, a acurácia ficou entre 93,9% e 96,5% de precisão para saber se um aluno terá ou não um desempenho satisfatório antes do final da disciplina.

Na pesquisa de Ferreira et al. (2019), a conclusão dos autores mostrou que foi possível gerar uma lista de possíveis reprovados que teriam a oportunidade de serem reintegrados antes do final da disciplina. Para a avaliação do modelo foram comparados os índices dos classificadores e destacou-se o J48 com uma predominância de acurácia de 84,5%, com precisão de 85,52%.

E por fim, Rolim e Silva (2021), através da aplicação dos algoritmos KNN e SVM, obtiveram uma precisão superior a 92%, utilizando como atributos o sexo, idade em que o estudante cursou a disciplina, média final e resultado. Também foram verificados fatores determinantes associados à retenção, representados pela faixa etária dos 19 aos 23 anos, sexo e notas finais.

De modo geral, os estudos encontrados focaram na predição do risco de evasão dos estudantes com base em dados extraídos de bancos de dados do Moodle, sendo em grande parte, dados que não consideravam explicitamente o perfil dos alunos, no que diz respeito a informações socioeconômicas, idade, sexo etc. O foco da maioria dos estudos foi utilizar os dados de notas e interações com o ambiente virtual de aprendizagem, para pré-determinar o desempenho acadêmico com base na mineração realizada sobre estes dados. Não foram encontrados estudos que abordavam especificamente o perfil de um aluno reintegrado, proposta desta pesquisa.

4. MATERIAIS E MÉTODOS

Este capítulo apresenta a metodologia utilizada nesta pesquisa, bem como a descrição dos materiais utilizados, neste caso, dados do Curso de EspSF do UNA-SUS/UFCSPA. Os dados foram obtidos através de acesso autorizado e manipulados em computador com software especializado para este fim.

Este trabalho é classificado como pesquisa aplicada devido à sua possível aplicação prática para as equipes de apoio dos cursos Ead. A abordagem adotada é qualitativa, uma vez que investiga uma perspectiva exploratória e subjetiva. Por fim, seu caráter é o explicativo, pois tem como foco principal retratar os dados de entrada, através da identificação de grupos que compartilham características semelhantes que explicam o fenômeno de evasão ou recuperação.

Primeiramente, é efetuada a coleta dos dados e realizado um tratamento para se adequar estes dados aos propósitos da mineração. A próxima etapa consiste em realizar a mineração de dados com o software específico e, por fim, efetuar a análise dos resultados e a modelagem dos dados obtidos com a mineração. Estas etapas são chamadas de pré-processamento, mineração de dados e pós-processamento, respectivamente (ROMERO; VENTURA, 2013).

4.1. Materiais

Para esta pesquisa, foram extraídos e utilizados os dados contidos em planilhas eletrônicas armazenadas em nuvem, que foram preenchidas pelo Apoio Acadêmico, Setor Pedagógico, tutores e coordenadores do projeto.

Este material foi cedido pela coordenação do curso e a autorização de uso encontra-se no Anexo I deste documento. Os dados foram coletados e tratados conforme a Lei Geral de Proteção de Dados (LGPD) - lei de nº 13.709, de 14 de agosto de 2018 (BRASIL; PRESIDÊNCIA DA REPÚBLICA, 2018).

Estavam disponíveis para consulta os dados de 16 turmas oferecidas a partir de 2011, organizadas em pastas com planilhas diversas, contendo informações dos alunos. No entanto, após uma análise inicial, observou-se que a organização das planilhas foi se modificando com o passar do tempo, e novas informações foram adicionadas ou modificadas. Além disso, algumas poucas planilhas continham um maior número de atributos, incluindo informações a respeito do acesso à Internet dos

alunos. Sendo assim, a escolha de quais turmas seriam pesquisadas se deu pelas planilhas que continham maior número de atributos em comum entre elas e que fossem de interesse dessa pesquisa.

Optou-se, assim, por trabalhar com as turmas 13, 14 e 15. A turma 13 continha um total de 265 alunos, enquanto a turma 14 tinha 380 alunos e a turma 15, 480 alunos. Além de serem as planilhas mais completas, também eram as maiores turmas dentre as 16, totalizando uma amostra de 1125 alunos.

Após a escolha das turmas, procedeu-se à análise dos atributos nas diversas planilhas. Estavam disponíveis informações pessoais dos alunos, notas das atividades do Eixo 1 e do TCC, anotações do Apoio Acadêmico e da Coordenação Pedagógica, informações dos tutores, incluindo a frequência dos alunos nas tutorias, uma planilha separada contendo o acompanhamento dos casos do Eixo 2, e um questionário que foi feito aos alunos com perguntas sobre a infraestrutura de acesso à Internet e a experiência prévia com cursos EaD.

Foram selecionados os atributos considerados mais relevantes para a pesquisa, levando em consideração a anonimidade destes dados, ou seja, buscou-se utilizar o mínimo possível de dados pessoais e que não pudessem identificar os alunos. Dentre as informações escolhidas estão: o gênero do aluno, pois em algumas anotações do Apoio Acadêmico havia casos de gravidez; a sua nacionalidade, já que em diversas anotações os alunos saíam de férias para visitar seus familiares; as respostas do questionário, para elucidar as questões de infraestrutura e conhecimento prévio de plataformas EaD; o estado onde o aluno atua profissionalmente, que fornece informações sobre as diferenças regionais; as suas notas do Eixo 1 para a extração da informação de recuperação do aluno e todas as anotações do Apoio Acadêmico e do Setor Pedagógico em relação ao aluno, que fornecem as demais informações relacionadas com o objetivo da pesquisa. A seguir, o Quadro 2 mostra um exemplo dos dados disponíveis.

Quadro 2 – Amostra dos dados disponíveis na base de dados do UNA-SUS/UFCSPA.

Tipo de Dado	**Amostra**						
Sexo	Masculino	Masculino	Masculino	Feminino	Masculino	Feminino	Feminino
Nacionalidade	Cubana	Cuba	Brasileiro	Cuba	CUBANA	CUBANO	Brasil
Aula inaugural	Sim	Sim	Sim	Sim	Sim	Não	Sim

Aula eixo 2	AUSENTE	SIM	SIM	SIM	SIM	SIM	SIM
Nota UE1	NC	90	40	90	90	90	70
Nota UE2	NC	60	30	70	70	70	80
Nota UE3	NC	90	40	90	90	90	50
Nota RUE1	NC	0	40	0	0	0	0
Nota RUE2	NC	0	60	0	0	0	0
Nota RUE3	NC	0	50	0	0	0	0
Nota EUE1	0	0	60	0	0	0	0
Nota EUE2	0	0	0	0	0	0	0
Nota EUE3	0	0	0	0	0	0	0
Nota TCC	0	71	0	91	78	100	82
Possui experiência em EAD?	Não	Não	Não	Não	Não	Não	Sim
Possui internet no trabalho?	Sim	Não	Não	Não	Não	Não	Sim
Possui internet em casa?	Sim	Sim	Sim	Sim	Não	Não	Sim
Banda larga?	Sim	Não	Não	Sim	Não	Não	Sim
Observações do Apoio Acadêmico	*Aluno estava de férias na 1ª chamada e NÃO realizou prova da 2ª chamada no dia 27/10/2017	NC	NC	NC	NC	NC	aluno defendeu em maio de 2018
Desfecho	RP	AP	AP	AP	AP	AP	AP

Fonte: Autoria própria (2023).

4.2. Metodologia

O processo de Descoberta de Conhecimento em Base de Dados, utilizado para análise dos dados deste trabalho, consiste em três etapas principais: pré-processamento, aplicação dos algoritmos e análise dos resultados.

No capítulo 5, intitulado “Processo de Descoberta de Conhecimento em Base de Dados Educacional”, serão explicadas com detalhes as etapas do processo

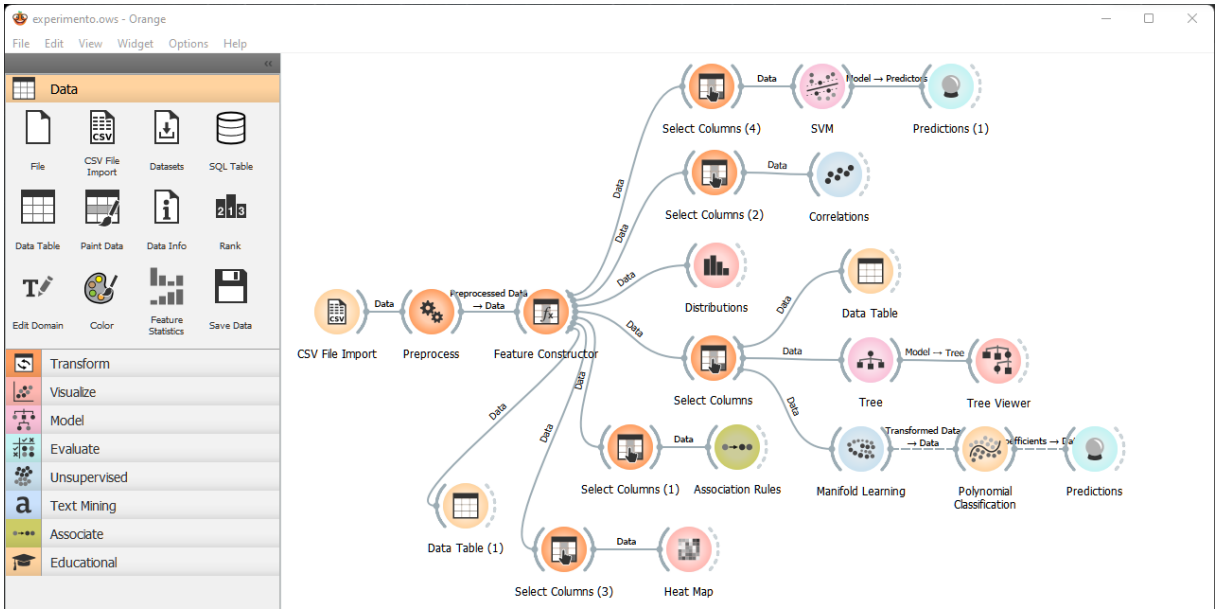
realizado, sendo este dividido em três seções: pré-processamento, mineração de dados e pós-processamento.

4.3. A Ferramenta Orange Data Mining

A área de ciência de dados vem crescendo muito nos últimos anos e, com isso, a disponibilidade de softwares para esse fim também cresce em quantidade e qualidade.

Uma opção interessante para este fim chama-se Orange Data Mining, que possui um grande apelo visual e facilidade de uso, além de uma grande disponibilidade de documentação online. Esta foi a ferramenta escolhida para realizar esta pesquisa. Além dos aspectos mencionados anteriormente, a escolha deste software deu-se pela familiaridade em seu uso, já que ele já havia sido utilizado por este pesquisador em outras oportunidades. A Figura 7 apresenta uma imagem de utilização do Orange Data Mining.

Figura 7 – Visualização da ferramenta Orange Data Mining.



Fonte: Autoria própria (2023).

A ferramenta Orange é escrita na linguagem de programação Python, trazendo grande versatilidade e maior performance na execução dos algoritmos. Esta ferramenta disponibiliza as suas funcionalidades através de *widgets,*

representados pelos ícones na Figura 7. Estas funcionalidades são agrupadas por categorias no canto esquerdo da tela inicial e possuem compatibilidade entre si conforme a entrada e saída de dados. Ao clicar em um *widget*, ele apresenta uma janela de configuração para que o usuário possa estabelecer os parâmetros de execução da funcionalidade, que pode ser uma simples visualização de dados até a aplicação de algoritmos.

4.4. A Escolha dos Algoritmos Utilizados

Verificou-se através da revisão de estudos relacionados que a Árvore de Decisão foi o algoritmo mais utilizado nas pesquisas de mineração de dados. Devido a isso e à disponibilidade de algoritmos na ferramenta escolhida para a aplicação prática da pesquisa, contatou-se que a sua utilização se torna uma escolha viável, pois além de a Árvore de Decisão ser um algoritmo simples de entender, ele fornece uma fácil visualização e entendimento do problema.

Para fins de complementação, definiu-se que as Regras de Associação também se caracterizam por serem algoritmos adequados ao tipo de problema desta pesquisa e foram incluídas na realização da mineração dos dados educacionais, embora não tenham sido muito utilizadas nos estudos encontrados anteriormente.

5. PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS EDUCACIONAL

Este capítulo aborda em detalhes os critérios utilizados para o processamento dos dados, de acordo com a metodologia e definições apresentadas nas seções anteriores.

5.1. Pré-processamento

No contexto da Mineração de Dados, nem sempre os dados são coletados ou disponibilizados no formato adequado para a aplicação dos algoritmos de mineração. Na maioria dos casos, faz-se necessária uma estruturação destes dados para que haja um padrão de preenchimento. Como as planilhas que continham os dados eram manipuladas por diversas pessoas de equipes diferentes, havia muita inconsistência no preenchimento. Além disso, as diferentes informações de um mesmo aluno estavam distribuídas em diversas planilhas, exigindo um trabalho de consolidação dos dados.

Para a realização desta consolidação, foi necessário escolher um atributo em comum entre as planilhas a serem consolidadas e que fosse de valor único. Sendo assim, uma planilha única foi montada com as informações das diversas planilhas e importadas para serem trabalhadas no software *Microsoft Excel*.

Outro aspecto importante no que tange a essa escolha diz respeito ao tipo de dado, que pode ser do tipo categórico, numérico ou texto (ALEEM; GORE, 2020). Um dado categórico caracteriza um elemento qualquer por categorias, por exemplo, ser de nacionalidade brasileira. Já o dado numérico possui valores numéricos de diversos tipos, mas estes valores não podem ser intervalos. Por fim, os dados do tipo texto contêm os dois tipos de dados anteriores e necessariamente precisam de tratamento, a não ser que se utilize um tipo especial de mineração para esse formato de dado.

Após esta escolha, foi feito o cruzamento dos dados para criar uma planilha única com todos os atributos necessários. Finalizada esta etapa, os seguintes atributos compunham a planilha inicial, conforme a nomenclatura original: SEXO, NACIONALIDADE, AULA INAUGURAL, AULA EIXO 2, UE1 (Unidade organizacional 1), UE2 (Unidade organizacional 2), UE3 (Unidade organizacional 3), RUE1

(Recuperação Unidade organizacional 1), RUE2 (Recuperação Unidade organizacional 2), RUE3 (Recuperação Unidade organizacional 3), EUE1 (Exame Unidade organizacional 1), EUE2 (Exame Unidade organizacional 2), EUE3 (Exame Unidade organizacional 3), TCC (Trabalho de Conclusão de Curso), EXPERIÊNCIA EM EAD, STATUS, ACESSO INTERNET NO TRABALHO, ACESSO INTERNET EM CASA, BANDA LARGA e as diversas observações do Apoio Acadêmico que foram concentradas no atributo OBSERVAÇÕES APOIO.

Os atributos SEXO, NACIONALIDADE, EXPERIÊNCIA EM EAD, ACESSO INTERNET NO TRABALHO, ACESSO INTERNET EM CASA e BANDA LARGA foram obtidos através de um questionário preenchido pelo aluno antes do início do curso.

Os atributos AULA INAUGURAL e AULA EIXO 2 indicam o comparecimento do aluno nas aulas presenciais em cada um dos dois eixos do curso.

Os atributos UE1, UE2, UE3, RUE1, RUE2, RUE3, EUE1, EUE2 e EUE3 são as notas obtidas no Eixo 1 e TCC a nota do Trabalho de Conclusão de Curso.

O atributo STATUS é o desfecho do aluno no curso, sendo classificado como aprovado (AP), reprovado (RP) e evadido (S).

Seguindo com as etapas de pré-processamento e a montagem do banco de dados a ser importado para a ferramenta de mineração, o atributo OBSERVAÇÕES APOIO havia sido preenchido por comentários do Setor Pedagógico e do Apoio Acadêmico em relação ao aluno, e desse texto foi extraída a informação de que o aluno recebeu alguma ajuda do Apoio Acadêmico. Estas anotações continham relatos diversos sobre o contato por telefone, e-mail ou sistema de *tickets*. Vale destacar que esta etapa de extração foi realizada mais de uma vez, já que nem sempre estava claro se houve efetivo auxílio por parte do Apoio Acadêmico. Após a leitura deste conteúdo para cada aluno, foi criado um atributo chamado CONTATO APOIO contendo apenas Sim (S) ou Não (N). Esta foi uma etapa bastante trabalhosa, devido ao fato de não poder ser feita de forma automática.

Já o atributo NACIONALIDADE continha o país de origem do aluno, apresentando uma grande diversidade de valores, como por exemplo: Brasileiro, Venezuela, Cubano, etc. Para fins de simplificação, os valores foram transformados em “E” de estrangeiro ou “B” de brasileiro.

As notas, únicos atributos numéricos, também receberam tratamento, pois continham em alguns registros o texto “NC” ou apresentavam o campo em branco.

Para a mineração de dados, foram atribuídos valores fixos no intervalo de “0” a “100”, onde o texto “NC” e espaços em branco foram substituídos pelo valor “0”.

Os valores dos atributos categóricos foram simplificados para conter o mínimo possível de texto. Por exemplo, o valor “Masculino” foi substituído por “M”, “Sim” por “S” e assim por diante.

Um novo atributo chamado RECUPERAÇÃO foi criado a partir das notas disponíveis no Eixo 1, para determinar se o estudante fez alguma recuperação. Caso o aluno tivesse nota em algum dos atributos RUE1, RUE2, RUE3, EUE1, EUE2 ou EUE3, era marcado como Sim (S) no campo RECUPERAÇÃO. Ao final deste processo, todos os dados pessoais de identificação foram apagados para gerar um banco de dados totalmente anonimizado.

Os tipos de atributos do banco de dados resultante, após o tratamento inicial no software *Microsoft Excel,* e que foram utilizados na ferramenta de mineração de dados, são mostrados no Quadro 3:

Quadro 3 – Tipos de dados resultantes após tratamento inicial da base de dados.

ATRIBUTO	TIPO	AMOSTRA
SEXO	Categórico	M ou F
NACIONALIDADE	Categórico	B ou E
AULA INAUGURAL	Categórico	S ou N
AULA EIXO 2	Categórico	S ou N
UE1	Numérico	0 a 100
UE2	Numérico	0 a 100
UE3	Numérico	0 a 100
RUE1	Numérico	0 a 100
RUE2	Numérico	0 a 100
RUE3	Numérico	0 a 100
EUE1	Numérico	0 a 100
EUE2	Numérico	0 a 100
EUE3	Numérico	0 a 100
TCC	Numérico	0 a 100
EXPERIÊNCIA EM EAD	Categórico	S ou N
ACESSO INTERNET NO TRABALHO	Categórico	S ou N
ACESSO INTERNET EM CASA	Categórico	S ou N
BANDA LARGA	Categórico	S ou N
UF	Categórico	AM, RS, AC, etc.
CONTATO APOIO	Categórico	S ou N
DESFECHO	Categórico	AP, RP ou E
RECUPERAÇÃO	Categórico	S ou N

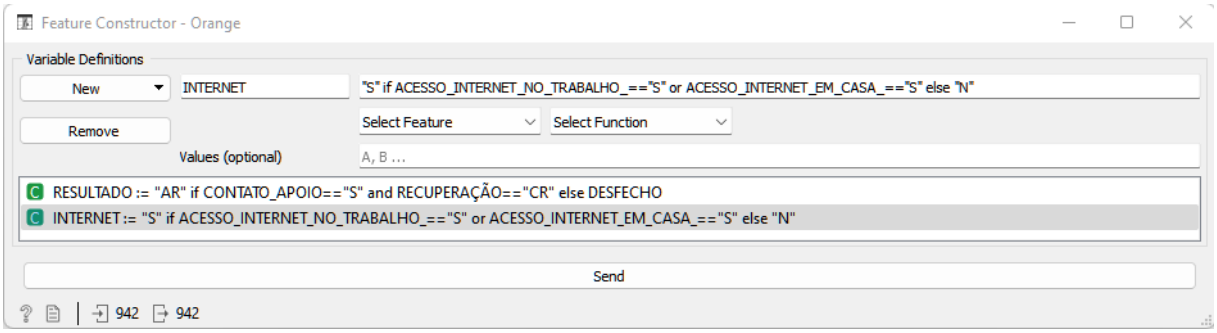
Fonte: Autoria própria (2023).

Após a importação dos dados para a ferramenta Orange, ainda como etapa de pré-processamento, foram removidos todos os alunos com alguma informação faltante. Para isso foi utilizado o *widget* Preprocess com a opção "*remove rows with missing values"*. Com isso, a amostra contabilizou um total de 942 alunos. Destes, 57,64% eram do sexo feminino, 51,70% eram brasileiros, 80,57% não possuíam experiência com cursos em EaD, 95% deles registraram frequência na aula inaugural, 89,92% declararam possuir acesso à Internet em casa ou no trabalho e 70,81% foram aprovados ao final do curso. Os alunos eram residentes dos seguintes estados do Brasil: Acre (3,5%), Amazonas (17,94%), Amapá (2,12%), Pará (29,29%), Roraima (3,08%), Rio Grande do Sul (38,32%) e Sergipe (5,84%).

Em seguida foi utilizado o *widget* Feature Constructor para a criação de novos atributos. O primeiro atributo criado na ferramenta Orange foi o RESULTADO, fruto da combinação do atributo CONTATO APOIO e RECUPERAÇÃO para formar o principal perfil desta pesquisa: Aluno Reintegrado (AR). O aluno que foi considerado reintegrado possui os valores "S" em RECUPERAÇÃO e "S" em CONTATO APOIO. Sendo assim, o atributo original DESFECHO foi reclassificado e copiado para um novo atributo chamado RESULTADO, ficando da seguinte forma: Aprovado Normalmente (AP), Aprovado Reintegrado (AR), Reprovado (RP) e Evadido (E). O aluno que seguiu o fluxo normal do curso foi classificado com AP, o aluno que precisou recuperar alguma nota e teve algum contato por parte do Apoio foi classificado com AR, o aluno reprovado por nota foi classificado com RP e, por fim, o aluno que foi desligado do curso ou o abandonou foi classificado com E. Deve-se destacar que um dos objetivos deste trabalho foi o de definir como obter o perfil do aluno reintegrado, resultando na criação desse atributo.

A criação do atributo INTERNET foi uma simplificação dos três atributos relacionados ao acesso à Internet. Foi considerado que o aluno que tivesse internet em casa ou no trabalho possuía acesso à Internet, não importando se era do tipo banda larga ou não. A Figura 8 representa as regras utilizadas na ferramenta para a criação dos novos atributos.

Figura 8 – Regras utilizadas no Orange para a criação dos novos atributos.



Fonte: Autoria própria (2023).

A tabela final para a aplicação dos algoritmos constitui-se de uma soma dos atributos apresentados nos Quadro 3 e 4, este último mostrado abaixo:

Quadro 4 – Atributos novos, criados após a utilização da ferramenta Orange.

ATRIBUTO	TIPO	AMOSTRA
RESULTADO	Categórico	AP, AR, RP ou E
INTERNET	Categórico	S ou N

Fonte: Autoria própria (2023).

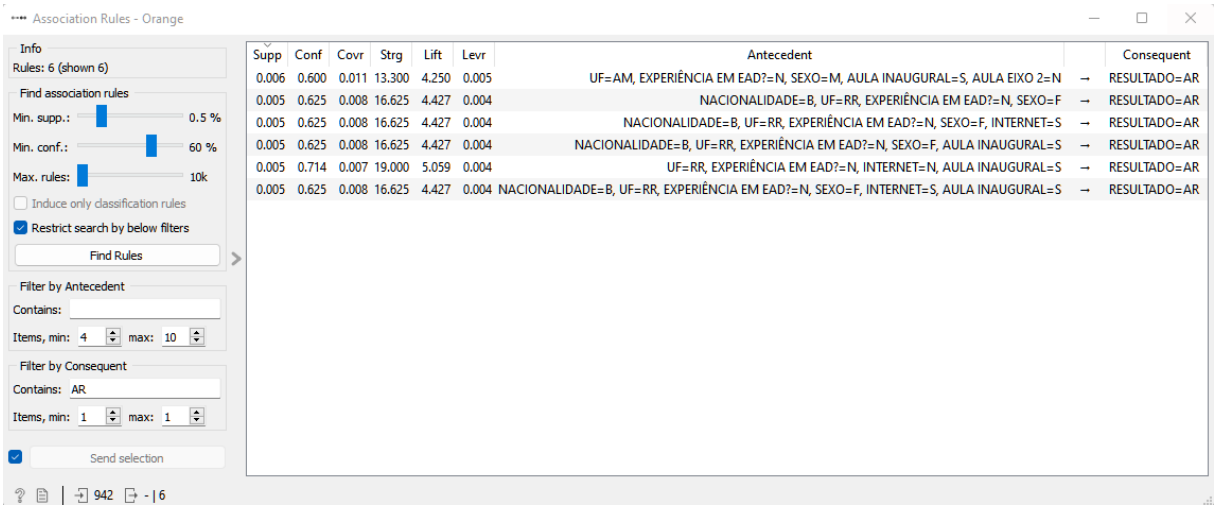
5.2. Mineração de Dados

Neste trabalho estamos interessados em saber se alguma das características do perfil do aluno aparece com frequência nos casos em que o aluno foi marcado como AR e compará-lo com os perfis dos alunos evadidos. A mineração de dados foi realizada com diversas configurações de atributos, buscando sempre o melhor resultado quanto ao número de perfis minerados. Inicialmente levou-se em consideração todos os atributos relacionados ao acesso à Internet dos alunos. No entanto, esta abordagem resultou em um número muito baixo de instâncias. Assim, diversos experimentos foram realizados buscando aumentar o número de instâncias em cada folha. Conseguiu-se uma melhoria unificando os atributos relacionados ao acesso à Internet, resultando na criação do atributo INTERNET e explicado anteriormente no final da sessão 5.1. Apesar dos esforços, o número de instâncias ainda se mostrou baixo para cada folha, resultando em perfis não muito

representativos em relação aos demais. Toda a mineração apresentada levou em consideração os resultados obtidos com o atributo INTERNET.

A Figura 9 mostra o resultado do uso do *widget* Regras de Associação para os alunos reintegrados.

Figura 9 – Widget utilizado para gerar as Regras de Associação.



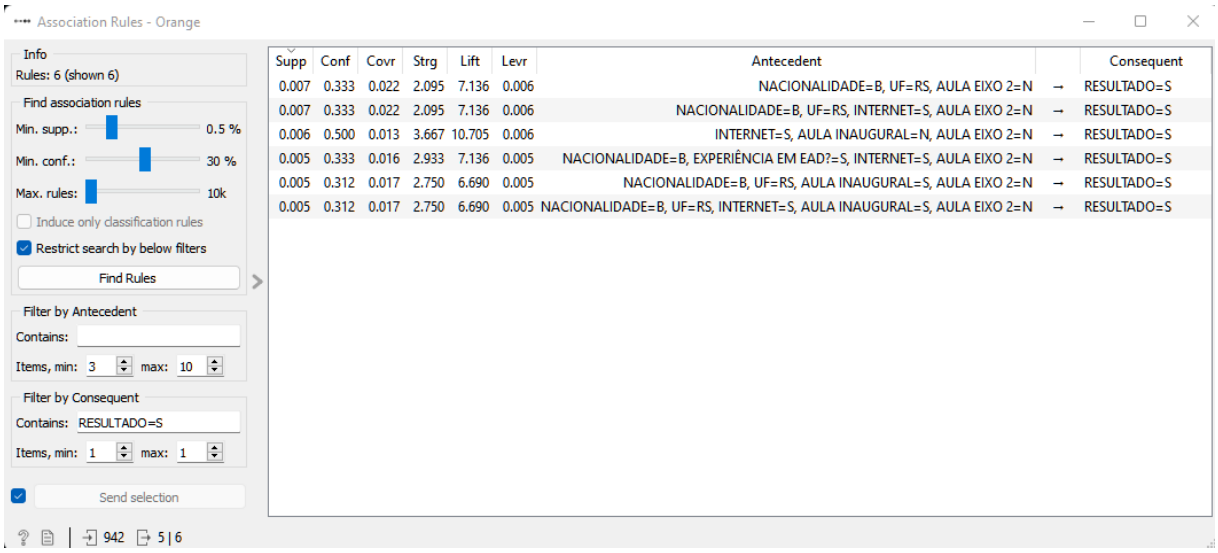
Fonte: Autoria própria (2023).

Os parâmetros utilizados foram: suporte mínimo de 5% e confiança mínima de 60%. Além disso, as regras foram filtradas por consequente na forma Y=AR e que contenha no mínimo 4 antecedentes. Como resultado, obtém-se 6 regras de associação.

Neste caso, podemos destacar dois perfis de alunos com maior confiança para cada estado que apareceram no resultado: com confiança de 71,4% e suporte de 5%, o perfil do aluno do estado de Roraima, sem experiência em EaD, sem acesso à Internet e que participou da aula inaugural. O outro perfil que podemos destacar possui confiança de 60% e suporte de 5% e pertence ao aluno do estado do Amazonas, sem experiência em EaD, do sexo masculino e que participou da aula inaugural, mas não participou da aula do Eixo 2. Estes serão os perfis comparados com o aluno evadido na parte de resultados deste trabalho.

Aplicando as Regras de Associação e escolhendo como alvo os alunos evadidos, temos o seguinte resultado, mostrado na Figura 10:

Figura 10 – Regras de Associação tendo como alvo alunos evadidos.



Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.007	0.333	0.022	2.095	7.136	0.006	NACIONALIDADE=B, UF=RS, AULA EIXO 2=N	→	RESULTADO=S
0.007	0.333	0.022	2.095	7.136	0.006	NACIONALIDADE=B, UF=RS, INTERNET=S, AULA EIXO 2=N	→	RESULTADO=S
0.006	0.500	0.013	3.667	10.705	0.006	INTERNET=S, AULA INAUGURAL=N, AULA EIXO 2=N	→	RESULTADO=S
0.005	0.333	0.016	2.933	7.136	0.005	NACIONALIDADE=B, EXPERIÊNCIA EM EAD?=S, INTERNET=S, AULA EIXO 2=N	→	RESULTADO=S
0.005	0.312	0.017	2.750	6.690	0.005	NACIONALIDADE=B, UF=RS, AULA INAUGURAL=S, AULA EIXO 2=N	→	RESULTADO=S
0.005	0.312	0.017	2.750	6.690	0.005	NACIONALIDADE=B, UF=RS, INTERNET=S, AULA INAUGURAL=S, AULA EIXO 2=N	→	RESULTADO=S

Fonte: Autoria própria (2023).

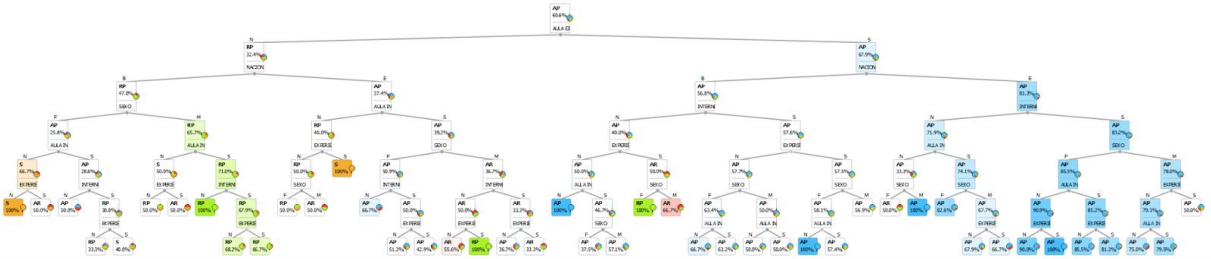
Os parâmetros utilizados foram: suporte mínimo de 5% e confiança mínima de 30%. Além disso, as regras foram filtradas por consequente (Y=S) e que contenham no mínimo 3 antecedentes. Como resultado, obtém-se 6 regras de associação.

Sendo assim, podemos destacar o perfil do aluno do estado do Rio Grande do Sul, sem acesso à internet e que não participou de nenhuma aula presencial. Esta regra de associação possui confiança de 50% e suporte de 5%.

As Árvores de Decisão são estruturas visuais que têm formato de uma árvore, onde a raiz contém os objetos a serem classificados e as ramificações (nós) são atributos de semelhança destes objetos. Os objetos vão percorrendo os nós até chegarem na folha (final do caminho), que vai definir a classificação deste objeto.

A Figura 11 mostra a Arvore de Decisão gerada pela ferramenta Orange:

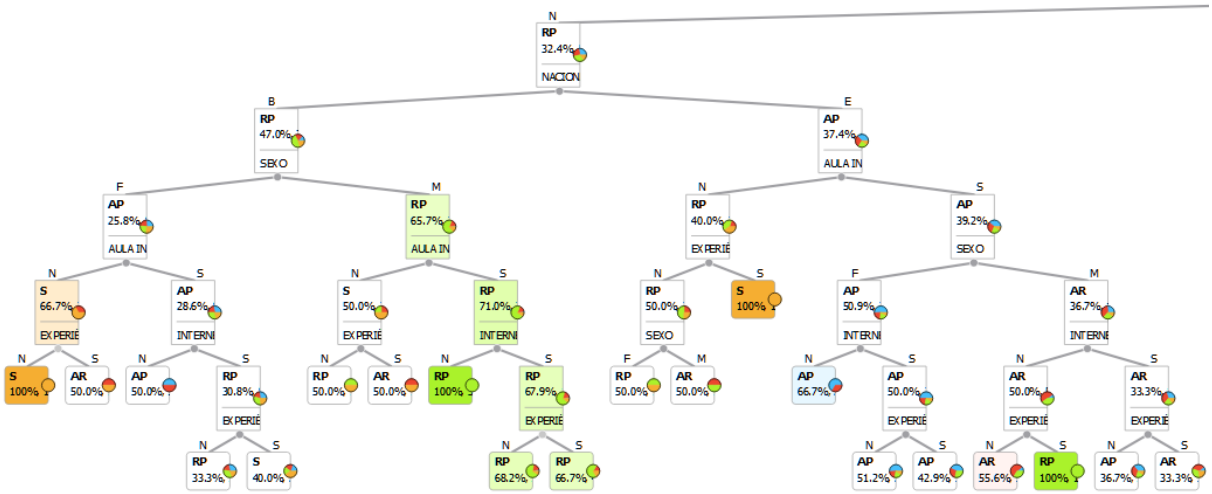
Figura 11 – Árvore de Decisão geral para toda a amostra.



Fonte: Autoria própria (2023).

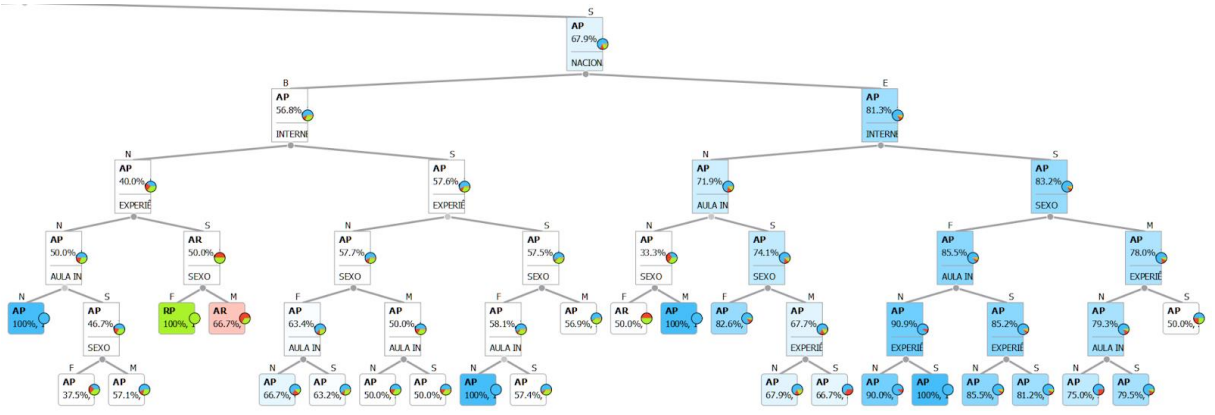
Como a árvore gerada para toda a amostra quando representada inteira em uma única imagem acaba ficando ilegível, a imagem foi dividida em duas partes para melhor visualização. O nó raiz representa o atributo AULA EIXO 2. Na Figura 12, encontra-se o ramo com os alunos que não estavam presentes na segunda aula presencial (lado esquerdo da árvore) e na Figura 13, são mostrados os alunos que estavam presentes.

Figura 12 – Árvore de Decisão para toda a amostra, lado esquerdo.



Fonte: Autoria própria (2023).

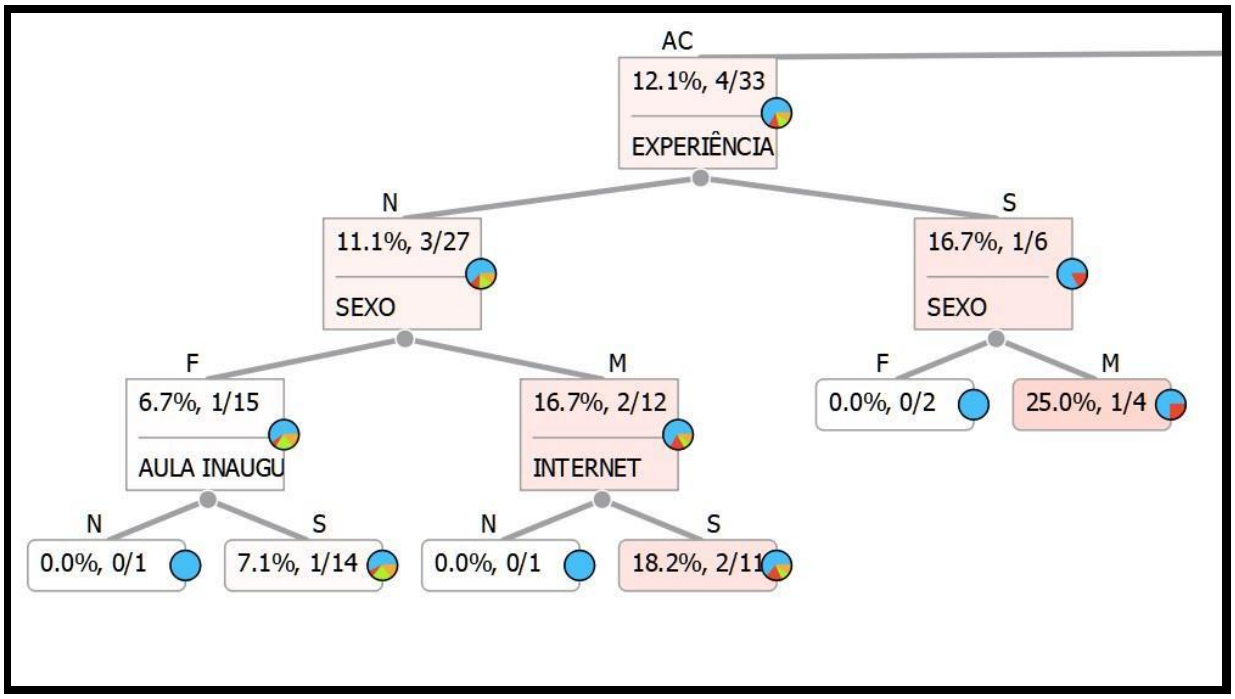
Figura 13 – Árvore de Decisão para toda a amostra, lado direito.



Fonte: Autoria própria (2023).

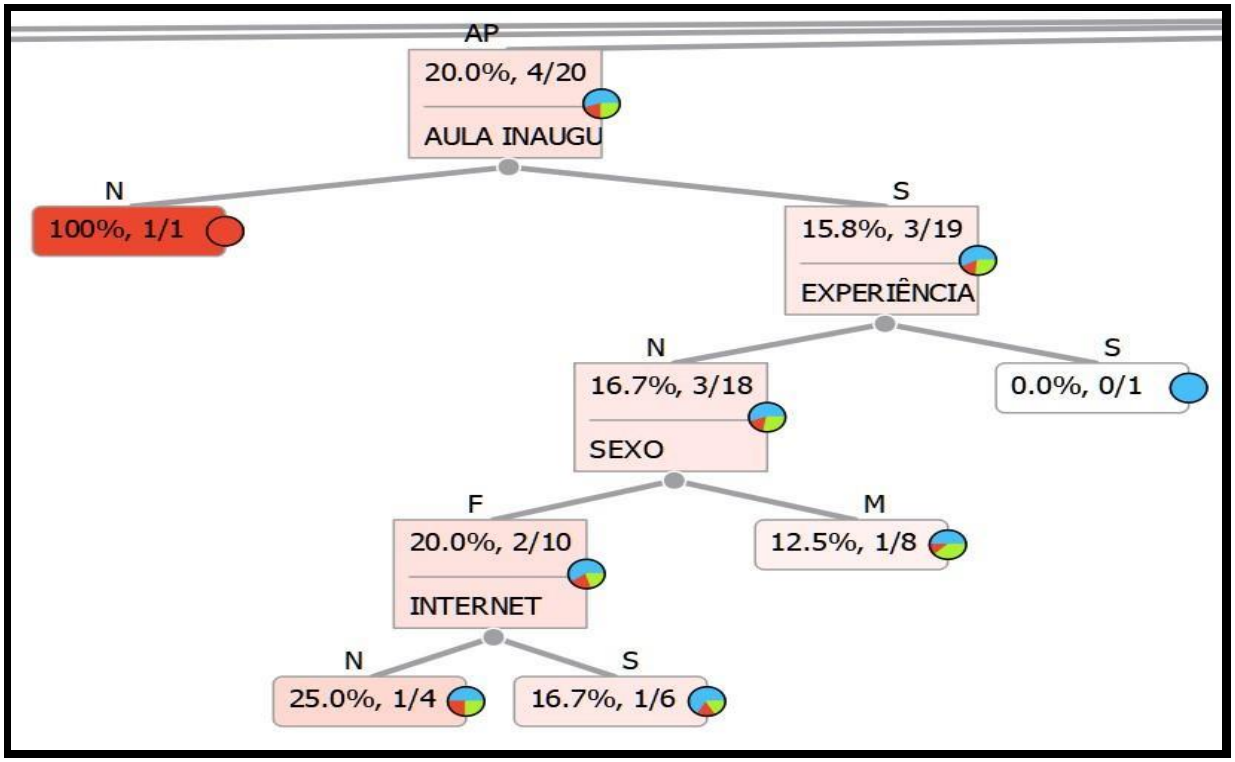
A amostra foi então separada por estado para ser apresentada e analisada. A seguir, nas Figuras de 14 a 20, são mostrados os ramos correspondentes a cada estado e a variável alvo escolhida foi a dos alunos reintegrados (AR). A gradação da cor funciona da seguinte forma: quanto mais vermelho o nó, maior é a quantidade de alunos reintegrados em relação aos demais resultados no nó. Em cada nó aparecem as informações referentes ao nome do atributo, valor de instâncias com o valor escolhido/valor total de instância e a percentagem destas instâncias do total de instâncias no nó.

Figura 14 – Nó da Árvore de Decisão que representa o estado do Acre.



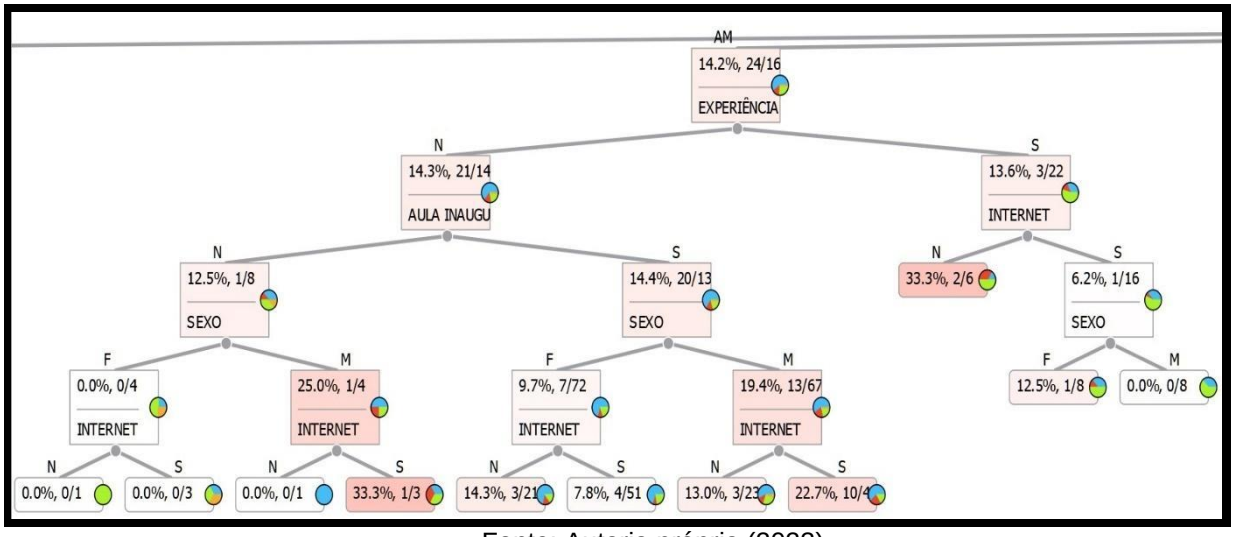
Fonte: Autoria própria (2023).

Figura 15 – Nó da Árvore de Decisão que representa o estado do Amapá.



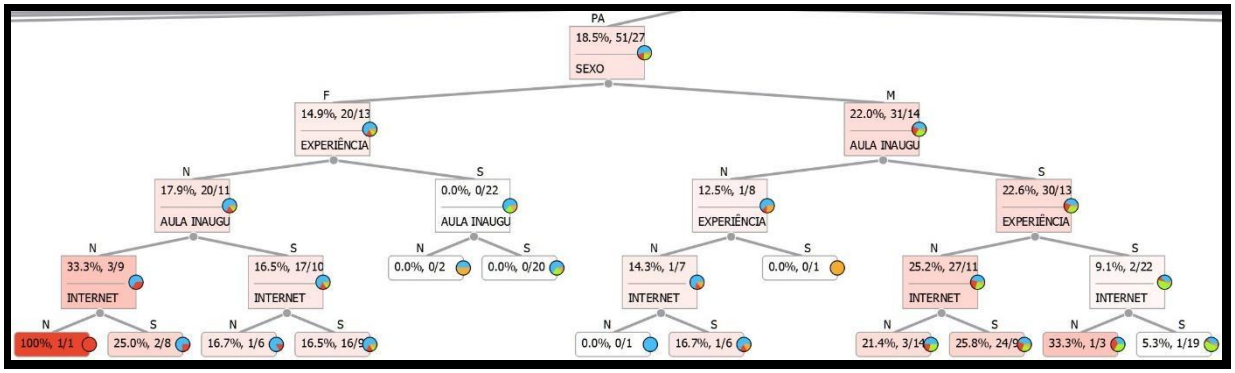
Fonte: Autoria própria (2023).

Figura 16 – Nó da Árvore de Decisão que representa o estado do Amazonas.



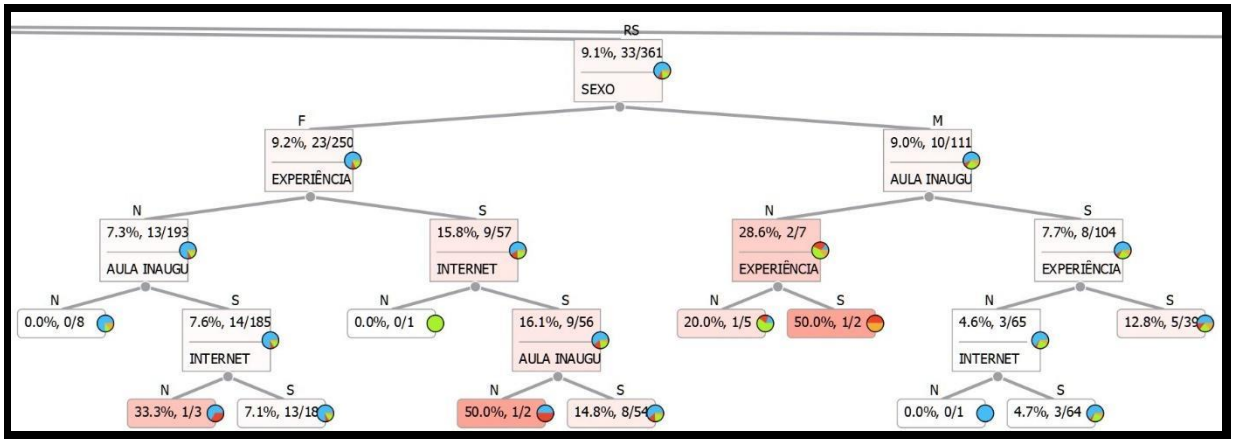
Fonte: Autoria própria (2023).

Figura 17 – Nó da Árvore de Decisão que representa o estado do Pará.



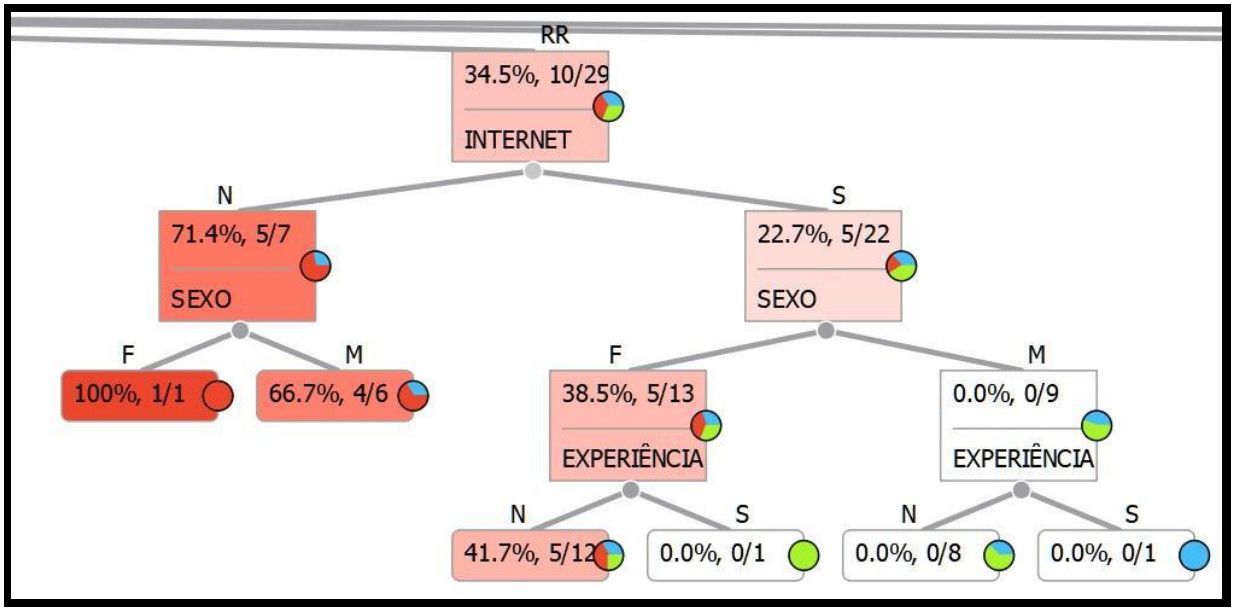
Fonte: Autoria própria (2023).

Figura 18 – Nó da Árvore de Decisão que representa o estado do Rio Grande do Sul.



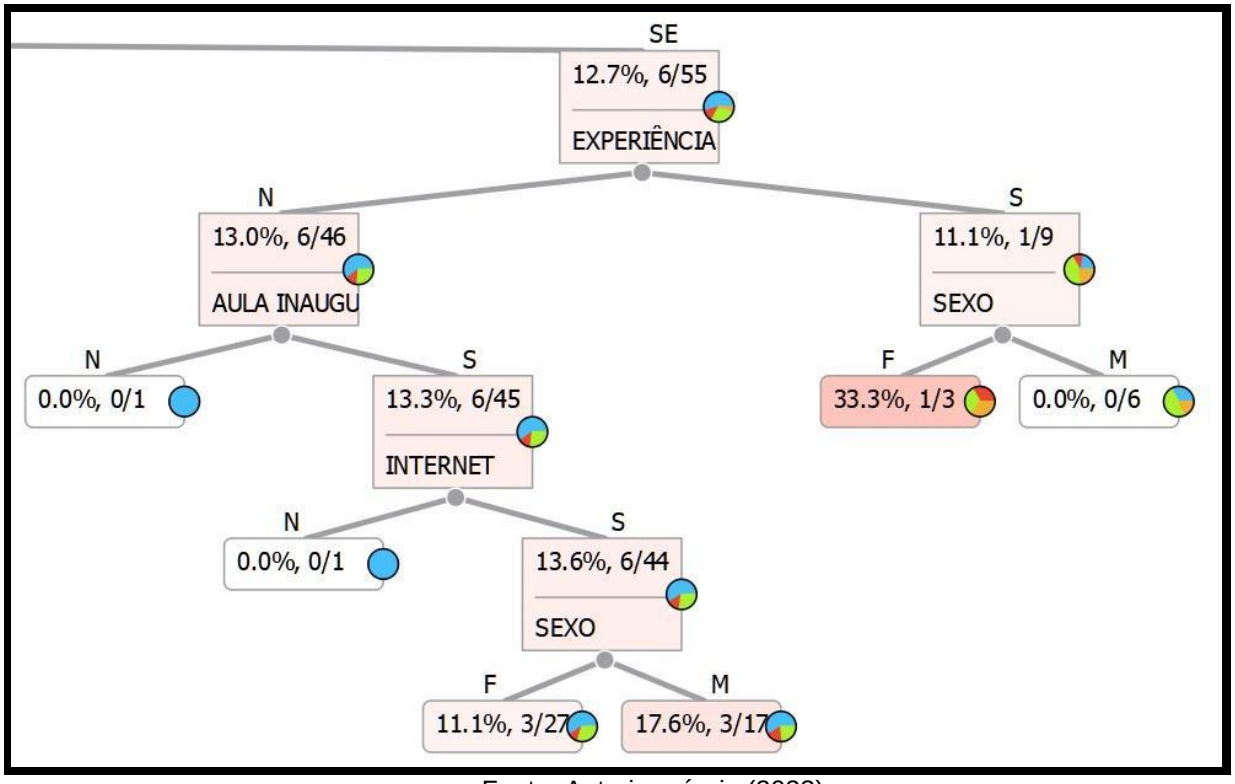
Fonte: Autoria própria (2023).

Figura 19 – Nó da Árvore de Decisão que representa o estado de Roraima.



Fonte: Autoria própria (2023).

Figura 20 – Nó da Árvore de Decisão que representa o estado de Sergipe.



Fonte: Autoria própria (2023).

Devido a algumas limitações de representação da ferramenta utilizada, os atributos NACIONALIDADE e AULA EIXO 2 não são mostrados nas Árvores de Decisão dos estados, pois os nós correspondentes a estes atributos estão acima do nó dos estados, e a representação da árvore se tornaria inviável. Os dados omissos nas figuras estão presentes na aplicação do algoritmo e serão apresentados em formato de quadro em sua totalidade no capítulo de discussão dos resultados.

5.3. Pós-processamento

Na etapa de pós-processamento é realizada a análise da mineração e a extração do conhecimento, no caso de novas informações emergirem. Esta etapa será aprofundada no capítulo seguinte, denominado “Resultados e Discussão”.

6. RESULTADOS E DISCUSSÃO

Neste capítulo serão mostradas as análises da aplicação das técnicas de Mineração de Dados na base do UNA-SUS/UFCSPA do Curso de EspSF. Além disso, serão comparados os resultados das duas técnicas utilizadas e se houve sucesso na obtenção de novos conhecimentos como resultado da pesquisa.

Através do uso das técnicas de Mineração de Dados, buscou-se extrair o perfil do aluno em dificuldades no Curso de Especialização em Saúde da Família oferecido pela UNA-SUS/UFCSPA, que conseguiu terminar o curso com sucesso e compará-lo com os alunos que foram desligados do curso por algum motivo. Esta comparação visa entender se houve fatores que influenciaram na recuperação deste aluno, bem como se é possível atribuir à atuação do Apoio Acadêmico a estes casos de sucesso. Ao final, é feito um comparativo entre os perfis dos diferentes estados, para verificar a consistência dos perfis encontrados durante a etapa de mineração de dados. Esta comparação final também é feita com os perfis dos alunos evadidos. Em seguida, é apresentada a definição do escopo do trabalho e as possibilidades de trabalhos futuros.

Iniciaremos com a análise do banco de dados, que conta com 942 instâncias e 8 atributos que foram utilizados para a mineração de dados. Os atributos utilizados são apresentados no Quadro 5, a seguir:

Quadro 5 – Quantidade de instâncias por atributo.

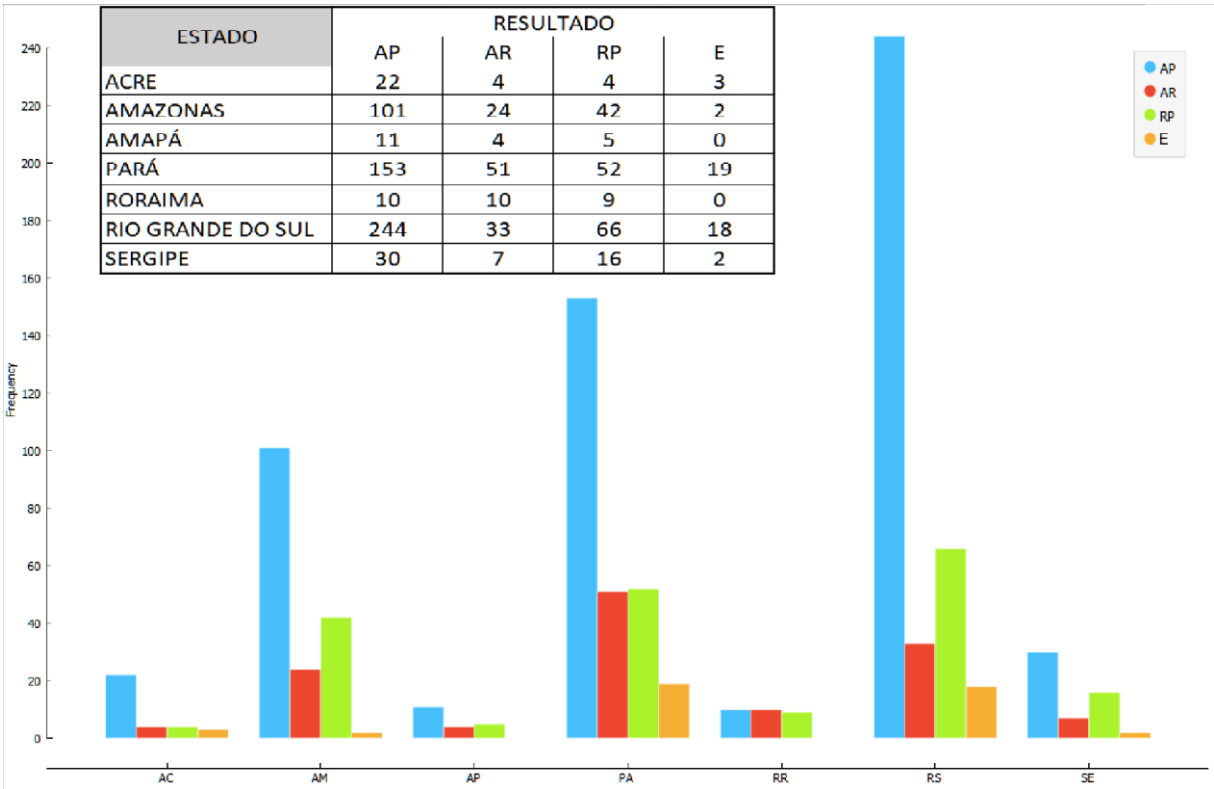
Variáveis	Instâncias		Variáveis	Instâncias
SEXO			**Acesso à Internet**	
Masculino	399		Sim	847
Feminino	543		Não	95
Nacionalidade			**Resultado**	
Brasileiro	487		Aprovados	571
Estrangeiro	455		Recuperados	133
Experiência em EAD			Reprovados	194
Sim	183		Suspensos	44
Não	759		**Estado**	
Aula inaugural			Acre	33
Sim	895		Amazonas	169
Não	47		Amapá	20
Aula Eixo 2			Pará	275
Sim	769		Roraima	29
Não	173		Rio Grande do Sul	361
			Sergipe	55

Fonte: Autoria própria (2023).

É importante destacar que alguns atributos desta amostra estavam bastante desbalanceados, ou seja, um único valor aparecia na maioria das instâncias, como é o caso dos atributos AULA INAUGURAL, AULA EIXO 2, EXPERIÊNCIA EaD e INTERNET. Apesar de existirem técnicas para lidar com este tipo de situação, optou-se por não se aplicar nenhuma técnica de balanceamento para não haver interferência na amostra e não diminuir ainda mais o número de instâncias.

O gráfico da Figura 21 mostra a quantidade de alunos aprovados, reintegrados, reprovados e evadidos, divididos por estado:

Figura 21 – Gráfico demonstrativo da quantidade de alunos aprovados, reintegrados, reprovados e evadidos.



Fonte: Autoria própria (2023).

Conforme visto no capítulo anterior, a aplicação do algoritmo de Regras de Associação resultou em duas regras principais para o aluno reintegrado e uma regra para o aluno evadido, apresentadas no quadro abaixo:

Quadro 6 – Regras resultantes após aplicação das Regras de Associação.

Suporte	Confiança	Antecedente	Consequente
0,006	0,6	UF=AM, EXPERIÊNCIA EM EAD?=N, SEXO=M, AULA INAUGURAL=S, AULA EIXO 2=N	RESULTADO=AR
0,005	0,714	UF=RR, EXPERIÊNCIA EM EAD?=N, INTERNET=N, AULA INAUGURAL=S	RESULTADO=AR
0,005	0,333	INTERNET=S, AULA INAUGURAL=N, AULA EIXO 2=N	RESULTADO=E

Fonte: Autoria própria (2023).

A regra com a maior confiança diz que o aluno de Roraima sem experiência em cursos EaD, sem acesso à Internet e que participou da aula inaugural foi reintegrado pelo Apoio em 71,4% dos casos.

A regra com a segunda maior confiança entre as geradas mostra que o aluno do Amazonas teve características semelhantes aos alunos de Roraima, resultando em 60% dos casos reintegrados pelo Apoio Acadêmico, segundo o critério desta pesquisa.

Os dois estados brasileiros que apareceram nas Regras de Associação mineradas nesta pesquisa se caracterizam por serem estados precários em infraestrutura e de difícil acesso às localidades (COSTA et al., 2015). Vale destacar que em casos em que o aluno tinha dificuldades em acessar o material do curso pela Internet, a equipe do UNA-SUS/UFCSPA enviava o material de estudos gravados em CD's e pendrives para que o aluno pudesse estudar e quando tivesse oportunidade de acessar à Internet, poderia fazer os exercícios de forma mais tranquila (COSTA et al., 2015). Este fato pode explicar o aparecimento destas características neste caso.

Com o algoritmo de Árvores de Decisão foi possível extrair alguns perfis dos alunos no Curso de Especialização em Saúde da Família do UNA-SUS/UFCSPA. Serão apresentados os resultados das folhas mais relevantes em cada estado, seguindo os seguintes critérios: buscou-se as folhas com o maior número absoluto de alunos com a situação alvo. No caso de haver mais de um perfil com o mesmo número de alunos, levou-se em consideração a folha com a maioria dos perfis em relação às outras situações. Um quadro comparativo dos perfis emergidos do aluno reintegrado e do evadido é apresentado a seguir, para cada estado.

Quadro 7 – Comparação dos perfis de aluno reintegrado e evadido no Acre.

ATRIBUTO	RESULTADO	
	AR	**E**
Sexo	M	F
Nacionalidade	B	E
Experiência em EAD	N	N
Aula inaugural	S	S
Aula Eixo 2	S	N
Acesso à Internet	S	S

Fonte: Autoria própria (2023).

No estado do Acre, três atributos se diferenciam entre os perfis: SEXO, NACIONALIDADE e AULA EIXO 2. No caso do atributo SEXO, faltam informações para se chegar a alguma conclusão baseada nos dados investigados nesta pesquisa. Em relação ao outro atributo, a ausência do aluno evadido na segunda aula presencial pode significar que este aluno se desligou cedo do curso, não permitindo uma maior atuação do Apoio Acadêmico. Como no perfil do aluno reintegrado ele estava presente na segunda aula presencial, isso pode ter influenciado na sua recuperação.

Este estado também se caracteriza por ser de difícil acesso e possuir uma infraestrutura precária (COSTA et al., 2015), onde apesar de os alunos terem acesso à internet, a pesquisa mostrou que esse pode não ser um fator relevante na recuperação do aluno.

Quadro 8 – Comparação dos perfis de aluno reintegrado e evadido no Amazonas.

ATRIBUTO	RESULTADO	
	AR	**E**
Sexo	M	F
Nacionalidade	B	B
Experiência em EAD	N	N
Aula inaugural	S	N
Aula Eixo 2	S	N
Acesso à Internet	S	S

Fonte: Autoria própria (2023).

No estado do Amazonas, chama a atenção a ausência dos alunos evadidos nas duas aulas presenciais, mostrando que este aluno nem ao menos começou o curso. Vários fatores podem ter contribuído para esta realidade, como por exemplo, o difícil acesso ao local da aula presencial. Para o aluno reintegrado, apesar dos dados sugerirem que ele esteve presente em todo o curso, outros fatores podem ter influência na dificuldade deste aluno e que a atuação do apoio acadêmico pode ter suprido essa necessidade.

Quadro 9 – Comparação dos perfis de aluno reintegrado e evadido no Amapá.

ATRIBUTO	RESULTADO	
	AR	**E**
Sexo	F	*
Nacionalidade	E	*
Experiência em EAD	N	*
Aula inaugural	N	*
Aula Eixo 2	S	*
Acesso à Internet	S	*

Fonte: Autoria própria (2023).

No estado do Amapá, não houve casos de alunos evadidos, portanto não foi possível fazer um comparativo entre os perfis dos alunos reintegrados e evadidos. Em relação ao perfil dos alunos reintegrados, vale destacar que a ausência na aula inaugural e presença na aula do Eixo 2 mostra que já pode ter havido uma atuação do Apoio Acadêmico para a reintegração desse aluno no curso. Infelizmente este estado teve poucas instâncias, prejudicando a extração de conhecimento com as técnicas utilizadas.

Quadro 10 – Comparação dos perfis de aluno reintegrado e evadido no Pará.

ATRIBUTO	RESULTADO	
	AR	**E**
Sexo	M	M
Nacionalidade	E	B
Experiência em EAD	N	N
Aula inaugural	S	S
Aula Eixo 2	S	S
Acesso à Internet	S	S

Fonte: Autoria própria (2023).

No estado do Pará, houve diferença apenas na nacionalidade do aluno, onde o perfil do aluno reintegrado é estrangeiro. Neste caso, os dados não fornecem informações suficientes para uma conclusão. Uma possível explicação é a questão de adaptação dos estrangeiros a realidade brasileira.

Quadro 11 – Comparação dos perfis de aluno reintegrado e evadido em Roraima.

ATRIBUTO	RESULTADO	
	AR	**E**
Sexo	F	*
Nacionalidade	B	*
Experiência em EAD	N	*
Aula inaugural	S	*
Aula Eixo 2	S	*
Acesso à Internet	S	*

Fonte: Autoria própria (2023).

No estado de Roraima, novamente não houve casos de alunos evadidos, sendo assim não foi possível fazer um comparativo entre os perfis dos alunos reintegrados e evadidos. O estado de Roraima possuía uma amostra muito pequena de alunos reintegrados, tornando essa análise em particular também inconclusiva. No entanto, este foi um dos estados que apareceu na mineração de dados utilizando a técnica de Regras de Associação, o que pode ser explicado pelo pequeno número de instâncias, já que relativamente, cada indivíduo tem uma influência maior sobre as outras instâncias.

Quadro 12 – Comparação dos perfis de aluno reintegrado e evadido no Rio Grande do Sul.

ATRIBUTO	RESULTADO	
	AR	**E**
Sexo	F	M
Nacionalidade	E	B
Experiência em EAD	N	S
Aula inaugural	S	S
Aula Eixo 2	N	S
Acesso à Internet	S	S

Fonte: Autoria própria (2023).

O estado do Rio Grande do Sul mostra um perfil atípico para o aluno evadido, já que as características sugerem que este aluno teve condições de concluir o curso.

Este perfil pode indicar que o aluno desistiu perto do final do curso. Já o perfil do aluno reintegrado pode indicar que o Apoio atuou para manter o aluno no curso, apesar da ausência na segunda aula presencial.

Quadro 13 – Comparação dos perfis de aluno reintegrado e evadido em Sergipe.

ATRIBUTO	RESULTADO	
	AR	**E**
Sexo	M	F
Nacionalidade	B	B
Experiência em EAD	N	S
Aula inaugural	S	S
Aula Eixo 2	S	N
Acesso à Internet	S	S

Fonte: Autoria própria (2023).

Para o estado de Sergipe, percebe-se que o perfil do aluno reintegrado participou das duas aulas presenciais, o que mostra que a dificuldade talvez estivesse em outros fatores que não somente as notas baixas ou atrasos de entrega. Já o perfil do aluno evadido que apesar de possuir experiência em cursos à distância, os estudantes já haviam desistido do curso antes do segundo módulo.

Quadro 14 – Comparação dos perfis de aluno reintegrado e evadido no Brasil.

ATRIBUTO	RESULTADO	
	AR	**E**
Sexo	M	M
Nacionalidade	B	B
Experiência em EAD	N	N
Aula inaugural	S	S
Aula Eixo 2	S	S
Acesso à Internet	S	S

Fonte: Autoria própria (2023).

Quando analisamos os perfis sem a distinção por estado, como mostrado no Quadro 14, o resultado do perfil emergido foi o mesmo para os alunos reintegrados e

evadidos. Este perfil é igual ao emergido nos estados de Sergipe e Amazonas, no caso dos alunos reintegrados e no caso dos alunos evadidos, este mesmo perfil apareceu no estado do Pará. Sendo assim, nenhuma conclusão podemos tirar deste conjunto de dados.

Pode-se concluir que houve uma grande variedade nos perfis emergidos, mostrando que os fatores regionais podem ser determinantes e que os dados não foram consistentes o suficiente para convergir o perfil do aluno. Para efeito de comparação, os quadros seguintes mostram os perfis separados por estado:

Quadro 15 – Comparação dos perfis de alunos reintegrados.

ATRIBUTO	RESULTADO: REINTEGRADO							
	AC	AM	AP	PA	RR	RS	SE	BRASIL
Sexo	M	M	F	M	F	F	M	M
Nacionalidade	E	B	E	E	B	E	B	B
Experiência em EAD	N	N	N	N	N	N	N	N
Aula inaugural	S	S	N	S	S	S	S	S
Aula Eixo 2	S	S	S	S	S	N	S	S
Acesso à Internet	S	S	S	S	S	S	S	S

Fonte: Autoria própria (2023).

Neste comparativo do Quadro 15, dois atributos estavam com o mesmo valor em todos os perfis, onde todos os alunos tinham acesso à Internet e não possuíam experiência em cursos à distância. Essa característica pode ser explicada pelo desbalanceamento dos dados nestes dois atributos.

Quadro 16 – Comparação dos perfis de alunos evadidos.

ATRIBUTO	RESULTADO: EVADIDO							
	AC	AM	AP	PA	RR	RS	SE	BRASIL
Sexo	F	F	*	M	*	M	F	M
Nacionalidade	E	B	*	B	*	B	B	B
Experiência em EAD	N	N	*	N	*	S	S	N
Aula inaugural	S	N	*	S	*	S	S	S
Aula Eixo 2	N	N	*	S	*	S	N	S
Acesso à Internet	S	S	*	S	*	S	S	S

Fonte: Autoria própria (2023).

No caso dos alunos evadidos, mostrados no Quadro 16, dois estados não tinham nenhuma instância, pois o banco de dados contava com apenas 44

ocorrências deste tipo de aluno, dificultando a comparação dos resultados. Esta baixa amostragem pode se dar pelo trabalho diferenciado realizado neste curso através do apoio acadêmico. Outro fator que pode contribuir com isso é a obrigatoriedade da inscrição no curso, que vinculava a conclusão à continuidade nos programas Mais Médicos e PROVAB. Outrossim, observa-se que o atributo de acesso à Internet foi o único presente em todos os perfis, provavelmente pelo desbalanceamento deste atributo dentro da amostra de dados.

Com estes resultados não foi possível obter informações sólidas a respeito das características dos alunos que participaram do curso EspSF. No entanto, podem ser úteis para direcionar trabalhos futuros e melhorias na coleta e tratamento dos dados, conforme limitações apontadas durante a pesquisa. Estes dados ficam disponíveis para sua utilização pelas equipes pedagógica e do apoio acadêmico.

7. CONSIDERAÇÕES FINAIS

O objetivo da pesquisa foi parcialmente cumprido, não sendo possível identificar o perfil do aluno reintegrado utilizando as técnicas de Descoberta de Conhecimento em Base de Dados e a amostra usada nesta pesquisa. Os perfis emergidos foram diferentes em cada estado, tanto para o aluno reintegrado como para o aluno evadido. O Brasil é um país de dimensões continentais e as diferentes características de cada região podem ter influenciado os diferentes perfis do aluno. Além disso, foi apontado durante a pesquisa que as características dos dados também contribuíram para que houvesse essa diferenciação, como o desbalanceamento de alguns atributos e a baixa representatividade dos perfis emergidos.

Em resumo, a grande maioria dos alunos possuía acesso à Internet, o que mostrou que este fato sozinho não explica a diferença entre um aluno aprovado e um evadido, como era de se esperar. A falta de experiência anterior com cursos à distância parece também não ser um fator decisivo para essa diferenciação. A frequência nas aulas presenciais, o gênero e o fato de o aluno ser estrangeiro ou brasileiro foram os fatores que mais influenciaram na diferenciação dos perfis.

A hipótese norteadora da pesquisa era de que a metodologia utilizada no curso Curso de EspSF do UNA-SUS/UFCSPA através de uma equipe de Apoio

Acadêmico (COSTA et al., 2015) seria muito importante para a realização de cursos na modalidade EaD, ainda mais em se tratando de públicos tão abrangentes como foi o caso. A mineração dos perfis pode permitir uma aferição das ações que estão funcionando e permite a criação de novas iniciativas para auxiliar os alunos a terminarem o curso. O perfil de aluno reintegrado é de suma importância neste contexto, pois mostra as características que devem ser observadas na hora de se pensar as estratégias para evitar a alta evasão. Apesar de não haver uma conclusão a respeito de um perfil único do aluno, a definição de aluno reintegrado pode ser utilizada em trabalhos futuros para a obtenção deste perfil.

A comparação dos perfis reintegrados com os perfis evadidos pode mostrar que, em caso de diferença entre eles, as características evidenciadas podem elucidar sobre a abordagem das ações do Apoio Acadêmico.

No caso de emergirem os mesmos perfis, a explicação da diferença entre evadidos e reintegrados não se encontra nos atributos escolhidos ou os dados não foram suficientes para evidenciar essa diferenciação.

Em relação às limitações da pesquisa, o escopo dos dados ficou limitado a três turmas de 16 disponíveis. Além disso, não houve uma diferenciação entre a origem dos comentários do atributo ANOTAÇÕES APOIO, que tiveram origem em diferentes setores do curso e não apenas do Apoio Acadêmico e que foi fundamental na definição do perfil do aluno reintegrado. Também não foi definida nenhuma ação de apoio ao perfil de aluno descoberto, podendo ser explorado em trabalhos futuros.

No que tange a uma possível continuidade deste trabalho, a pesquisa pode ser estendida com uma maior quantidade de turmas do UNA-SUS/UFCSPA e com a escolha de outros atributos disponíveis nas planilhas e outras fontes de dados. A nota dos alunos também é um material que não foi muito explorado nesta pesquisa e pode fornecer novas informações sobre o perfil do aluno que frequentou este curso.

Como subproduto desta pesquisa, o banco de dados anonimizado resultante deste trabalho pode ser disponibilizado publicamente para consulta.

REFERÊNCIAS

ABED - ASSOCIAÇÃO BRASILEIRA DE EDUCAÇÃO A DISTÂNCIA. **Censo EAD.BR 2020 - Relatório analítico da aprendizagem a distância no Brasil**. Disponível em: < https://abed.org.br/arquivos/CENSO_EAD_2020_PORTUGUES.pdf >. Acesso em: 2 jul. 2023.

AGGARWAL, C. C. **Data Mining: The Textbook**. Springer International Publishing, 2015.

ALEEM, A.; GORE, M. M. Educational data mining methods: A survey. **Proceedings - 2020 IEEE 9th International Conference on Communication Systems and Network Technologies, CSNT 2020**, p. 182–188, 2020.

AUGUSTO, D. K. et al. Quantos médicos de família e comunidade temos no Brasil? **Revista Brasileira de Medicina de Família e Comunidade**, v. 13, n. 40, p. 1–4, 7 jun. 2018.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 02, 31 ago. 2011.

BEHAR, P. A.; KELLEN, K.; DA SILVA, A. Mapeamento de Competências: Um foco no aluno da Educação a Distância. **Novas Tecnologias na Educação**, v. 10, n. 3, dez. 2012.

BRASIL; PRESIDÊNCIA DA REPÚBLICA. **Lei Geral de Proteção de Dados Pessoais (LGPD)**. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>. Acesso em: 2 jul. 2023.

COSTA, E. et al. Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. **Jornada de Atualização em Informática na Educação - JAIE 2012**, v. d, p. 1–29, 2012.

COSTA, M. R. et al. **Encurtando distâncias: o papel do apoio acadêmico em um curso de especialização em saúde da família**. Disponível em: < https://ares.unasus.gov.br/acervo/html/ARES/24803/1/Encurtando%20a%20dist%C3%A2ncia_Cap2.pdf >. Acesso em: 2 jul. 2023.

DA COSTA, S. S.; CAZELLA, S.; RIGO, S. J. Minerando dados sobre o desempenho de alunos de cursos de educação permanente em modalidade EAD: Um estudo de caso sobre evasão escolar na UNA-SUS. **Renote**, v. 12, n. 2, 2015.

DO COUTO, D. DA C.; DE SANTANA, A. L. **Análise dos problemas da Evasão e Retenção - Uma abordagem através de Mineração de Dados Educacionais**. XIII Encontro Anual de Computação - EnAComp 2017. Universidade Federal de Goiás, 2017.

DOSEA, G. S. et al. MÉTODOS ATIVOS DE APRENDIZAGEM NO ENSINO ONLINE: A OPINIÃO DE UNIVERSITÁRIOS DURANTE A PANDEMIA DE COVID-

19. **Interfaces Científicas - Educação**, v. 10, n. 1, p. 137–148, 6 set. 2020.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **Knowledge Discovery and Data Mining: Towards a Unifying Framework**. Disponível em: <www.aaai.org>. Acesso em: 2 jul. 3023.

FERREIRA, J. L. C. et al. **A model for predicting disapproval of apprentices in distance education using decision tree**. ACM International Conference Proceeding Series. Association for Computing Machinery, 20 maio 2019.

MANHÃES, M. L. B. et al. **Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados**. Anais do XXII SBIE - XVII WIE. 21 a 25 nov. 2011.

BRASIL; MINISTÉRIO DA SAÚDE. **UNA-SUS - Universidade Aberta do SUS**. Disponível em: <https://www.unasus.gov.br/>. Acesso em: 2 jul. 3023.

PAZ, F.; CAZELLA, S. **Identificando o perfil de evasão de alunos de graduação através da Mineração de dados Educacionais: um estudo de caso de uma Universidade Comunitária**. Anais dos Workshops do VI Congresso Brasileiro de Informática na Educação (CBIE 2017). Brazilian Computer Society (Sociedade Brasileira de Computação - SBC), 10 nov. 2017.

PORTELLA, F. F. **EXPLORANDO A RELAÇÃO ENTRE O PERFIL DO ALUNO E SEU DESEMPENHO ACADÊMICO NA MODALIDADE EAD: ANÁLISE COM BASE NOS DADOS DO CURSO DE ESPECIALIZAÇÃO EM SAÚDE DA FAMÍLIA DA UNA-SUS/UFCSPA**. Dissertação de Mestrado. Universidade Federal de Ciências da Saúde de Porto Alegre. Porto Alegre, 2019.

RABELO, H. et al**. Utilização de técnicas de mineração de dados educacionais para predição de desempenho de alunos de EaD em ambientes virtuais de aprendizagem.** Anais do XXVIII Simpósio Brasileiro de Informática na Educação (SBIE 2017), v. 1, n. Cbie, p. 1527, 2017a.

RABELO, H. et al. **Utilização de técnicas de mineração de dados educacionais para predição de desempenho de alunos de EaD em ambientes virtuais de aprendizagem**. Anais do XXVIII Simpósio Brasileiro de Informática na Educação (SBIE 2017). Brazilian Computer Society (Sociedade Brasileira de Computação - SBC), 10 nov. 2017b.

ROLIM, J. P.; SILVA, R. C. **Mineração de Dados Educacionais para identificação de Perfil de Retenção em um curso de Ciência da Computação** Em: ESCOLA REGIONAL DE INFORMÁTICA DE GOIÁS (ERI-GO), 9. , 2021, Evento Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2021 . p. 195-204.

ROMERO, C.; VENTURA, S. Data mining in education. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 3, n. 1, p. 12–27, jan. 2013.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence: A modern approach.** Egnlewood Cliffs: Prentice-Hall, 1995.

SANTOS, F. D. et al. **Análise de Evidências do Estado de Ânimo Desanimado de**

Alunos de um AVEA: uma Proposta a partir da Aplicação de Regras de Associação. Anais dos Workshops do IV Congresso Brasileiro de Informática na Educação (CBIE 2015). Sociedade Brasileira de Computação - SBC, 26 out. 2015a.

SANTOS, F. D. et al. **Análise de Evidências do Estado de Ânimo Desanimado de Alunos de um AVEA: uma Proposta a partir da Aplicação de Regras de Associação**. Anais dos Workshops do IV Congresso Brasileiro de Informática na Educação (CBIE 2015). Sociedade Brasileira de Computação - SBC, 26 out. 2015b.

SOCIEDADE INTERNACIONAL DE MINERAÇÃO DE DADOS EDUCACIONAIS. **Educational Data Mining**. Disponível em: <https://educationaldatamining.org/>. Acesso em: 2 jul. 2023.

SOUZA, V. F. DE; SOUZA, M. F. DE. ANÁLISE SOBRE A EVOLUÇÃO DA MINERAÇÃO DE DADOS EDUCACIONAIS SEGUNDO A PERSPECTIVA DOS PRINCIPAIS ESTUDOS DA ÁREA: UMA REVISÃO NARRATIVA. Em: **Ciência da Computação: tecnologias emergentes em computação - Volume 2**. Editora Científica Digital, 2021. p. 27–46.

ANEXO I – TERMO DE AUTORIZAÇÃO

TERMO DE AUTORIZAÇÃO PARA USO DE DADOS ARMAZENADOS EM NUVEM

Pelo presente termo, eu Aline Corrêa de Souza, coordenadora do projeto UNA-SUS/UFCSPA, autorizo expressamente a utilização dos dados armazenados em nuvem relativos ao projeto, pelo período de maio de 2021 a julho 2023, para fins exclusivos de ensino e pesquisa.

Declaro que estou ciente e concordo que os dados sejam acessados pelo aluno do PPGTIG da UFCSPA Felipe Augusto Prauchner Christmann para fins acadêmicos, científicos ou técnicos e que nenhum dado identificado será divulgado.

Declaro, ainda, que os dados coletados pelo projeto foram obtidos mediante consentimento dos participantes dos cursos, e que os mesmos foram informados sobre a forma como os dados seriam utilizados e as garantias de sigilo e privacidade dos dados.

Local e data: Porto Alegre, 05 de maio de 2021.

Assinatura: