Unioeste - Universidade Estadual do Oeste do Paraná
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
Colegiado de Matemática
Curso de Licenciatura em Matemática

Classificação de Dados em Busca de Padrões que Evidenciem a Evasão no Curso de Licenciatura em Matemática

Gabriela Artini da Silva

CASCAVEL
2020

GABRIELA ARTINI DA SILVA

CLASSIFICAÇÃO DE DADOS EM BUSCA DE PADRÕES QUE EVIDENCIEM A EVASÃO NO CURSO DE LICENCIATURA EM MATEMÁTICA

Monografia apresentada como requisito parcial para obtenção do grau de Licenciatura em Matemática, do Centro de Ciências Exatas e Tecnológicas da Universidade Estadual do Oeste do Paraná - Campus de Cascavel

Orientadora: Profa. Dra. Rosangela Villwock

CASCAVEL
2020

GABRIELA ARTINI DA SILVA

CLASSIFICAÇÃO DE DADOS EM BUSCA DE PADRÕES QUE EVIDENCIEM A EVASÃO NO CURSO DE LICENCIATURA EM MATEMÁTICA

Monografia apresentada como requisito parcial para obtenção do Título de Licenciado em Matemática, pela Universidade Estadual do Oeste do Paraná, Campus de Cascavel, aprovada pela Comissão formada pelos professores:

Profa. Dra. Rosangela Villwock (Orientadora)
Colegiado de Matemática, UNIOESTE

Profa. Dra. Simone Aparecida Miloca
Colegiado de Matemática, UNIOESTE

Prof. Dr. Amarildo de Vicente
Colegiado de Matemática, UNIOESTE

Cascavel, 22 de julho de 2021

AGRADECIMENTOS

Primeiramente agradeço aos meus pais, por todo amor, auxílio e incentivo que me ofereceram diante todos os obstáculos enfrentadas.

Agradeço especialmente a minha professora orientadora, Rosangela Villwock, por todo o conhecimento compartilhado durante a realização deste trabalho e no decorrer do curso.

A todos os professores que colaboraram nesse percurso e nos encorajaram a continuar.

Aos meus colegas de turma, por todas as risadas, choros, conselhos, ajudas, mas principalmente pelo crescimento que tivemos juntos.

Ao meu colega e amigo Guilherme Gasparini Lovatto, pela imensa parceria no decorrer dos últimos anos.

À Universidade Estadual do Oeste do Paraná, por todas as oportunidades oferecidas nesses anos de graduação.

E por fim, a todas as pessoas que passaram na minha vida nos últimos anos e que de algum modo colaboraram para o meu crescimento.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

IES	Instituicão de Ensino Superior
MEC	Ministério da Educação
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
SISU	Sistema de Selecão Unificada
Unioeste	Universidade Estadual do Oeste do Paraná
KDD	Knowledge Discovery in Databases (Descoberta de Conhecimento em Bases de Dados)
EDM	Educational Data Mining (Mineração de Dados Educacionais)
CEP	Comitê de Ética em Pesquisa com Seres Humanos
WEKA	Waikato Environment for Knowledge Analysis

Sumário

Resumo

Diante do elevado número de evasão dos alunos do curso de Licenciatura em Matemática, é necessária uma análise que possa identificar quais fatores influenciam tal decisão. Para isso, a Mineração de Dados Educacionais pode ser utilizada de modo a extrair informações úteis a partir de base de dados educacionais. Desta forma, objetiva-se com esse trabalho encontrar informações que descrevam alunos propensos a evasão por meio da Mineração de Dados Educacionais. Para tanto, foi criada uma base de dados a partir de dados obtidos nos históricos escolares dos ingressantes de 2010 a 2012 do curso de Licenciatura em Matemática da Universidade Estadual do Oeste do Paraná, *campus* Cascavel. Os dados coletados foram: número de vezes que o aluno cursou cada uma das disciplinas, se foi aprovado, se prestou exame quando foi aprovado e a maior frequência entre as vezes que cursou cada uma das disciplinas. Para a Mineração de Dados Educacionais, foi utilizada a tarefa de Classificação, que busca modelos para predizer o atributo escolhido como classe baseando-se nos demais atributos. Na classificação dos dados o algoritmo utilizado foi o J48, que é uma implementação do algoritmo C4.5 no WEKA. Trabalhos correlatos revelaram que o algoritmo J48 teve acurácia de mais de 90% na previsão de alunos propensos a evasão. Com este trabalho, foi possível notar que entre os anos de 2010 a 2012, 78 alunos evadiram do curso. A disciplina com maior índice de reprovações foi Álgebra, ofertada no terceiro ano de curso, a qual foi cursada em média 1,79 vez. A partir da Mineração de Dados, com o algoritmo J48, foi possível notar que no primeiro ano do curso a disciplina Geometria Analítica e Vetorial é a disciplina que mais influencia para que os alunos abandonem o curso, sendo que o modelo observado teve precisão de 92,5%. Essa mesma disciplina foi apontada como a disciplina do primeiro ano do curso com maior índice de reprovação, cursada em média 1,67 vez. Dentre todas as disciplinas do curso, a disciplina de Álgebra, ofertada no terceiro ano do curso, é apontada neste trabalho como a que mais influencia na evasão. A mesma, também possui maior índice de reprovações dentre todas as disciplinas

do curso, cursada em média 1,79 vez.

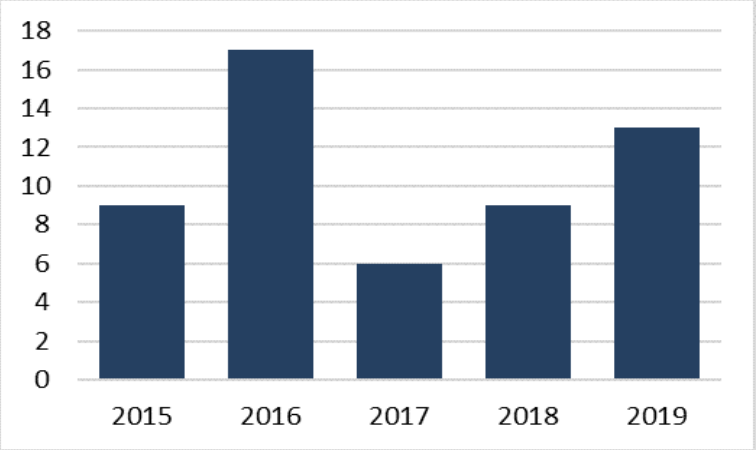
Palavras-chave: Evasão. Mineração de dados. Classificação. Árvore de decisão.

Capítulo 1

Introdução

Os cursos de graduação têm por objetivo formar profissionais capacitados e que possam exercer a profissão, entretanto, nos últimos anos o número de formandos têm sido uma preocupação para vários cursos, principalmente para o curso de Licenciatura em Matemática da Universidade Estadual do Oeste do Paraná (Unioeste), *campus* Cascavel. A Figura 1.1 mostra o número de formandos no curso de Licenciatura em Matemática nos últimos cinco anos, a partir de dados obtidos no sistema Academus[1].

Figura 1.1: Número de formandos no curso de Licenciatura em Matemática



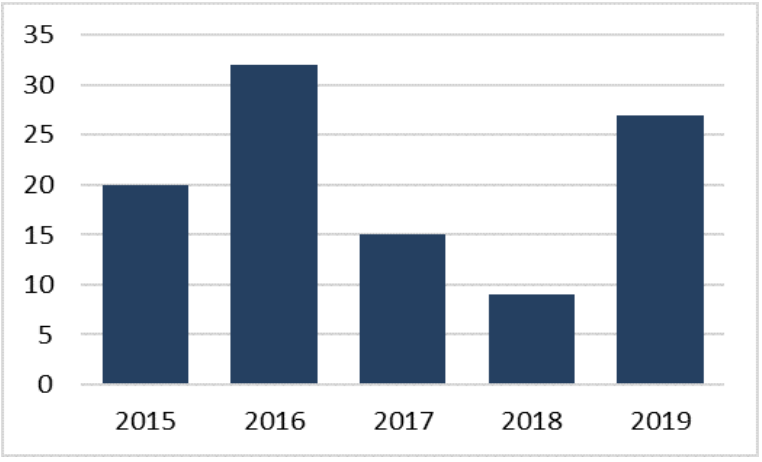
Fonte: Elaborado pela autora com dados obtidos no sistema Academus

Um dos principais motivos dessa preocupação é o alto índice de abandono durante o curso. "Tal problema possui diversas origens e afeta tanto a Instituição de Ensino Superior (IES) quanto o aluno e gera graves consequências sociais e financeiras" (SOUZA, 2016). A Figura 1.2 mostra

[1]O Academus é o sistema de gestão acadêmica da UNIOESTE que gerencia desde o ingresso de acadêmicos em cursos de graduação até a sua desvinculação da universidade.

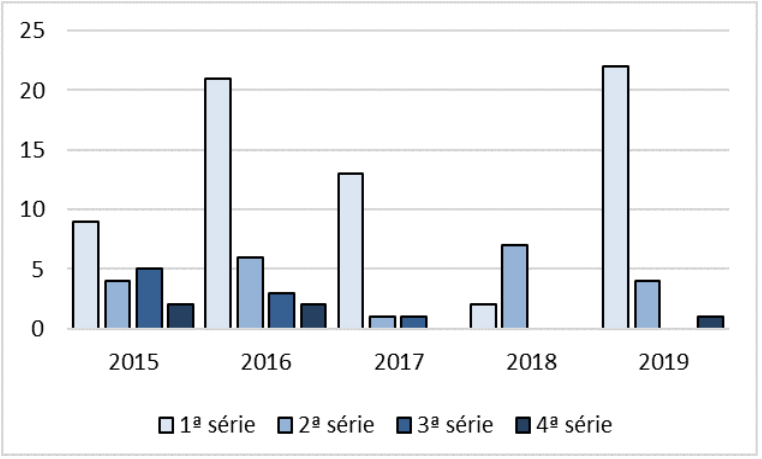
o número de abandono no curso de Licenciatura em Matemática nos últimos cinco anos. Com a Figura 1.3 que mostra o número de abandono por série entre 2015 a 2019, é possível notar que, exceto em 2018, a maior parte dos abandonos no curso de Licenciatura em Matemática, foi no primeiro ano do curso.

Figura 1.2: Número de abandonos no curso de Licenciatura em Matemática



Fonte: Elaborado pela autora com dados obtidos no sistema Academus

Figura 1.3: Número de abandonos por série no curso de Licenciatura em Matemática



Fonte: Elaborado pela autora com dados obtidos no sistema Academus

A evasão escolar é um problema que vem atingindo várias instituições de ensino de maneira negativa, e isso não seria diferente no ensino superior brasileiro. "Segundo Braga, Miranda-Pinto e Cardeal (1996), esse tema configurava-se como preocupação das universidades públicas e do MEC desde 1972" (POLYDORO, 2000, p. 45)

"Entre 2001 e 2005, de acordo com cálculos feitos com base em dados do Inep, a taxa média de evasão no ensino superior brasileiro foi de 22%, com pouca oscilação, mas mostrando tendências de crescimento" (SILVA FILHO et al., 2007, p. 658). Mais precisamente, no curso de Matemática, esse índice de evasão é ainda superior, visto que em 2005 essa taxa foi de 44% (SILVA FILHO et al., 2007).

Silva Filho et al. utilizam o termo evasão total para indicar os alunos que ingressaram um curso, em uma instituição de ensino superior e não obteve o diploma após um certo número de anos.

Pode-se destacar também as seguintes definições de Brasil (1997, p. 20):

> evasão de curso: quando o estudante desliga-se do curso em situações diversas tais como: abandono (deixa de matricular-se), desistência (oficial), transferência (mudança de curso), exclusão por norma institucional; evasão da instituição: quando o estudante desliga-se da instituição na qual está matriculado; evasão do sistema - quanto o estudante abandona de forma definitiva ou temporária o ensino superior.

Para Gaioso (2005 apud APPIO, 2012), evasão é definida como interrupção do ciclo de estudos e pode ocorrer de forma indireta, como pela incompatibilidade de horários. A mesma autora mostra em sua pesquisa alguns problemas que causaram a evasão apontados pelos próprios alunos, sendo elas,

> [...] falta de orientação vocacional, imaturidade do estudante, reprovações sucessivas, dificuldades financeiras, falta de perspectiva de trabalho, ausência de laços afetivos na universidade, ingresso na faculdade por imposição familiar, casamentos não planejados e nascimento de filhos (GAIOSO, 2005 apud APPIO, 2012, p. 7).

Entretanto, além dos fatores socioeconômicos apresentados, Baggi e Lopes (2011) salientam a necessidade da compreensão de fatores de ordem acadêmica que desestimulem o aluno a concluir o curso de graduação. Diante disso, alguns trabalhos relacionados a evasão resultante de reprovações sucessivas e outros fatores são apresentados abaixo.

Castro (2013) busca compreender em seu trabalho, o problema de evasão nos cursos de licenciatura na Universidade Estadual do Oeste do Paraná, *campus* Cascavel, através de análise

de documentos e questionários. Com esse estudo, foi possível notar que entre os fatores intrainstitucionais que causaram a evasão se destacam as reprovações e a possibilidade de não terminar o curso no tempo mínimo estabelecido. Já nos fatores extrainstitucionais, destacam-se a opção ou priorização de outro curso e o trabalho, no qual quase 21% dos evadidos trabalhavam e cursavam o curso de graduação simultaneamente. A autora também entrevista coordenadores dos cursos estudados, os quais indicaram que os principais indícios de evasão foram faltas, reprovações/notas baixas e desinteresse de modo geral.

Souza (2008) busca em seu trabalho encontrar razões para a evasão no ensino superior através da mineração de dados, trabalhando com dados dos cursos de Engenharia da Universidade Federal Fluminense. Analisando os alunos que cancelaram o curso, ou seja, evadiram, foi possível analisar que a disciplina Cálculo Diferencial e Integral Aplicado I foi a que obteve maior número de reprovações. Além disso, a reprovação recorrente nas disciplinas de base dos cursos estudados pode influenciar decisões relacionadas a abandonar o curso.

Hoed (2016) analisa a evasão dos cursos da área de Computação da Universidade de Brasília, a partir de dados fornecidos pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) e dados da própria instituição. Nessa pesquisa, o autor aponta que a reprovação em disciplinas pontuais é um dos motivos de evasão, visto que os alunos que reprovam nas disciplinas de Algoritmos e de Cálculo 1, realizadas nos primeiros semestres do curso, não concluem o curso.

Entretanto, as informações oferecidas pelo INEP não indicam de forma detalhada as taxas de evasão. Para isso é necessário a realização de um cálculo aproximado levando em conta número de matriculados, ingressantes e concluintes por ano, que são encontrados apenas em trabalhos acadêmicos relacionados ao tema (BAGGI; LOPES, 2011).

1.1 Justificativa

No curso de Licenciatura em Matemática da Universidade Estadual do Oeste do Paraná, *campus* Cascavel, são oferecidas 40 vagas para ingressantes, sendo 20 vagas oferecidas por meio do Vestibular e 20 vagas por meio do Sistema de Seleção Unificada (SISU). Entretanto, Andreta (2013) analisa os casos de evasão e formandos entre os anos de 2008 a 2012, mostrando que são em média 15 alunos formandos e 26 evasões por ano. Desse mesmo modo, com os dados

da Figura 1.1 e da Figura 1.2, podemos notar que entre os anos de 2015 a 2019, foram em média 11 alunos formandos e 21 evasões por ano.

Além disso, Appio (2012) e Andreta (2013) são trabalhos que motivaram a realização dessa pesquisa, visto que pesquisaram sobre a evasão no curso de Licenciatura em Matemática e mostraram o quão preocupante é o alto número de evadidos. Para verificar os principais fatores que contribuíram para a evasão, Andreta (2013) utilizou um questionário socioeconômico e Appio (2012) realizou uma pesquisa documental, analisando dados acadêmicos sobre as disciplinas cursadas em uma turma do curso de Licenciatura em Matemática. Similar a Appio (2012), essa pesquisa foi realizada analisando três turmas do curso, de maneira mais ampla.

Diante desses altos números de evasão no ensino superior que constam nos trabalhos acima, faz-se necessário algum método que possa identificar as principais características de alunos que possuem maior risco de evasão, baseando-se no desempenho acadêmico, ou seja, nas disciplinas com maior índice de reprovação e outros fatores relacionados às disciplinas cursadas, para que assim, algumas medidas educativas possam ser elaboradas e aplicadas.

1.2 Objetivos

1.2.1 Objetivo geral

Este trabalho teve como objetivo geral encontrar padrões que descrevam alunos com maior propensão ao abandono, observando dados referentes a vida escolar dos ingressantes no curso de Licenciatura em Matemática nos anos de 2010 a 2012 da Universidade Estadual do Oeste do Paraná – Unioeste, *campus* Cascavel.

1.2.2 Objetivos específicos

Os objetivos específicos do trabalho foram:

- Estudar o abandono do curso de Licenciatura em Matemática nos últimos anos;
- Identificar quais disciplinas possuem maior número de reprovação entre os alunos do curso;
- Buscar informações relevantes para identificar alunos do curso propensos ao abadono.

1.3 Estrutura do Trabalho

Este trabalho está organizado em 4 capítulos. O Capítulo 1 é composto pelo tema, o problema e a justificativa do trabalho, bem como os objetivos gerais e específicos. No Capítulo 2 são apresentadas as etapas operacionais da Descoberta de Conhecimento em Bases de Dados e as tarefas da Mineração de dados, com ênfase na Tarefa de Classificação, também é descrito o algoritmo C4.5 e as árvores de decisão. Além disso contém uma revisão da literatura, apresentando alguns trabalhos na área da Mineração de Dados Educacionais. No Capítulo 3 foi apresentado a coleta dos dados utilizados para a base de dados e o *software* WEKA que foi utilizado para a mineração de dados. O Capítulo 4 apresenta os resultados obtidos e as discussões, já o Capítulo 5 apresenta as considerações finais dessa pesquisa.

Capítulo 2

Referencial teórico

Diante do crescimento de dados e informações armazenadas, tornaram-se necessárias ferramentas computacionais capazes de analisar dados contidos em grandes bancos de dados. Para que isso seja possível, existe uma área chamada *Knowledge Discovery in Databases* (KDD), traduzida como Descoberta de Conhecimento em Bases de Dados, que analisa grandes quantidades de dados, sendo capaz de analisar, interpretar e relacionar esses dados, os quais podem apresentar tendências úteis para a descoberta de um conhecimento (GOLDSHMIDT; PASSOS; BEZERRA, 2015).

Muitas vezes a noção de encontrar padrões úteis em bases de dados recebe nomes distintos, como KDD e Mineração de Dados, mas de acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996), o KDD é o processo de descoberta de conhecimento, já Mineração de Dados é uma etapa do processo KDD. Além disso, os autores definem o KDD como um processo não trivial de identificar um padrão válido, novo e útil, a partir de um conjunto de dados, para que se torne compreensível.

De modo geral, o KDD é um processo utilizado para coleta, análise e interpretação de dados, em busca de uma relação entre as variáveis, para a extração do conhecimento.

2.1 Etapas Operacionais do KDD

A partir da escolha de um problema a ser submetido ao processo KDD descrito acima, iniciam-se as três etapas operacionais: o Pré-Processamento, a Mineração de Dados e o Pós-Processamento, as quais serão explicadas a seguir.

A etapa do Pré-Processamento tem por objetivo preparar os dados para a etapa de Mineração

de dados e é composta pela captação, organização e tratamento dos dados. Essa etapa identifica as informações relevantes para o processo KDD na base de dados, tais informações também podem ser chamadas de atributos. (GOLDSHMIDT; PASSOS; BEZERRA, 2015).

Além disso, nessa etapa são analisados e corrigidos eventuais erros que podem comprometer a qualidade dos modelos de conhecimento extraídos no final do processo de KDD (GOLDSHMIDT; PASSOS; BEZERRA, 2015).

A etapa de Mineração de Dados é considerada a principal do processo de KDD, pelo fato de que é nessa etapa que um conhecimento útil é descoberto. A Mineração de Dados é composta pela aplicação de tarefas que têm por objetivo explorar dados para construir um modelo de conhecimento, que é definido como um padrão ou conjunto deles, que possui como propósito descrever, com linguagem formal, um conjunto de dados (GOLDSHMIDT; PASSOS; BEZERRA, 2015).

Goldshmidt, Passos e Bezerra (2015) exploram as tarefas de KDD que são utilizadas na Mineração de Dados. Algumas tarefas de KDD são: Associação, Classificação, Clusterização/Agrupamento.

Cada uma dessas tarefas da mineração de dados possui um objetivo, que são predição ou descrição. A predição busca encontrar um modelo para prever valores desconhecidos em novas situação, já a descrição busca encontrar um modelo para descrever comportamento dos dados (GOLDSHMIDT; PASSOS; BEZERRA, 2015).

A etapa do Pós-Processamento tem como objetivo analisar e interpretar os resultados da Mineração de Dados. Como esses resultados, muitas vezes, não são de fácil compreensão, se faz necessário simplificar o modelo de conhecimento, tornando-o menos complexo, mas sem perda de informações relevantes. Por fim, é realizada a organização e apresentação dos resultados para o problema (GOLDSHMIDT; PASSOS; BEZERRA, 2015).

2.2 Tarefas da Mineração de Dados

A tarefa de Associação é utilizada para descobrir padrões que descrevam carcterísticas associadas nos dados, e possui por objetivo extrair os padrões mais interessantes de uma forma eficiente (TAN; STEINBACH; KUMAR, 2009). De acordo com Goldshmidt, Passos e Bezerra (2015), um dos principais fatores de motivação para a tarefa de Associação, é a capacidade de

incrementar as vendas a partir de um conjunto de regras de associação extraído de bases de dados de segmentos comerciais.

A Tabela 2.1 abaixo, apresenta um conjunto de dados referente a cestas de compras.

Tabela 2.1: Cestas de compras

TID	Itens
1	{Pão, Leite}
2	{Pão, Fraldas, Cerveja, Ovos}
3	{Leite, Fraldas, Cerveja, Cola}
4	{Pão, Leite, Fraldas, Cerveja}
5	{Pão, Leite, Fraldas, Cola}

Fonte: Tan, Steinbach e Kumar (2009)

Nesse exemplo, a tarefa de Associação é capaz de extrair a seguinte regra do conjunto de dados:

$$\{Fraldas \rightarrow Cerveja\}$$

Essa regra sugere que há um forte relacionamento entre a venda de fraldas e de cerveja, ou seja, muitos clientes que compraram fraldas também compraram cerveja (TAN; STEINBACH; KUMAR, 2009). A partir dessa regra de associação, varegistas podem potencializar a venda de seus produtos.

A tarefa de Classificação pode ser definida como a busca de uma função que associe cada registro do conjunto de dados a um único rótulo categórico chamado classe. Nessa tarefa, os atributos do conjunto de dados são divididos em dois grupos. Um grupo contém apenas o atributo-alvo, que é o atributo para o qual se deve fazer a predição e o outro contém os atributos previsores ou atributos de predição (GOLDSHMIDT; PASSOS; BEZERRA, 2015).

Para Tan, Steinbach e Kumar (2009, p.172), "classificação é a tarefa de aprender uma função alvo f que mapeie cada conjunto de atributos x para um dos rótulos de classes y pré-determinados". A tarefa de Classificação é a tarefa utilizada neste trabalho e será aprofundada na seção a seguir.

A tarefa de Clusterização ou Agrupamento é utilizada para agrupar registros de dados em grupos (clusters), sendo que em cada grupo, os elementos possuem propriedades que os distingam dos outros grupos e, se difere da Classificação pois os objetos considerados como entrada não possuem rótulos associados (GOLDSHMIDT; PASSOS; BEZERRA, 2015). Esta tarefa,

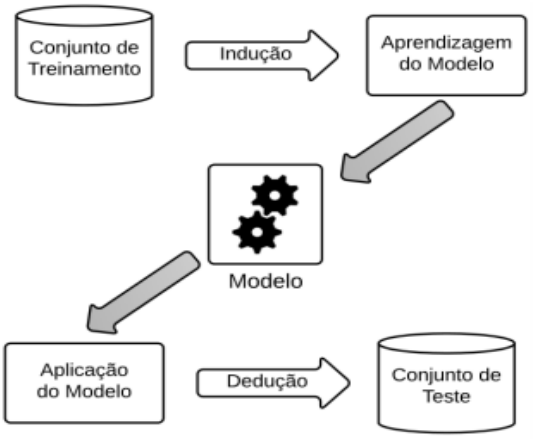
tem o intuito de que os objetos dentro de um grupo sejam semelhantes entre si e diferentes de outros objetos dos outros grupos, sendo que quanto maior a semelhança entre os objetos de cada grupo e menor semelhança entre os grupos, mais distinto será o agrupamento (TAN; STEINBACH; KUMAR, 2009).

2.3 Classificação

A Classificação é um processo que ocorre em duas etapas. A primeira é chamada etapa de aprendizagem, na qual um modelo de classificação é construído. Já a segunda é a etapa de classificação, na qual o modelo encontrado anteriormente é utilizado para prever rótulos dos dados fornecidos (HAN; KAMBER; PEI, 2012).

Segundo Costa et al. (2012), o modelo classificador possui como entrada um conjunto de amostras de dados onde a classe já é conhecida, chamada de conjunto de treinamento. Nesse conjunto de dados, é induzido um modelo classificador, o qual é testado junto a um conjunto de testes, formado por um conjunto de amostras cujas classes são ocultadas, mas que precisam ser preditas pelo modelo. A Figura 2.1 descreve o funcionamento de um modelo classificador.

Figura 2.1: Funcionamento de um modelo classificador



Fonte: Costa et al. (2012)

Desse modo, o modelo gerado deve se adaptar bem aos dados do conjunto de treinamento e prever com precisão os rótulos das classes de registros do conjunto de testes (TAN; STEINBACH; KUMAR, 2009).

Para avaliar o desempenho de um modelo de classificação, é utilizado o número de registros previstos correta e incorretamente pelo modelo. Essa contagem é dada em uma tabela, chamada matriz de confusão (TAN; STEINBACH; KUMAR, 2009)

Além disso, alguns cálculos podem ser realizados para facilitar a comparação de desempenhos de diferentes modelos, destacando-se a medida de precisão e a taxa de erro. A medida de precisão de um modelo é o quociente entre o número de previsões corretas e o número total de previsões. Já a taxa de erro é o quociente entre o número de previsões erradas pelo número total de previsões (TAN; STEINBACH; KUMAR, 2009).

2.3.1 Árvore de decisão e algoritmo C4.5

Há vários algoritmos para criação de modelos classificadores, dentre eles, destacam-se: árvores de decisão, classificadores baseados em regras, classificadores Bayesianos, Redes Neurais, entre outros.

A árvore de decisão é formada a partir de uma abordagem recursiva de particionamento do conjunto de dados (GOLDSHMIDT; PASSOS; BEZERRA, 2015). Tan, Steinbach e Kumar (2009) evidenciam que uma árvore de decisão é organizada de forma hierárquica e é formada por um nodo raiz, nodos internos e nodos folhas:

- Um nodo raiz não possui arestas chegando e zero ou mais arestas saindo;
- Nodos internos possuem apenas uma aresta chegando e duas ou mais saindo;
- Nodos folhas ou terminais têm uma aresta chegando e nenhuma saindo.

Para mostrar o funcionamento de métodos baseados em árvore de decisão, de acordo com Goldshmidt, Passos e Bezerra (2015) consideramos o esquema $X(A_1, A_2, A_3, ..., A_n, C)$, onde A_i é um atributo previsor e C o atributo-alvo, podendo assumir valores $\{c_1, c_2, ..., c_k\}$ chamados classes do problema.

De modo geral, em uma árvore de decisão, a raiz da árvore contém todo o conjunto de dados, em seguida escolhe-se um predicado, chamado ponto de separação, que é a condição que melhor discrimina as classes. Esse predicado induz a divisão do conjunto de dados em dois ou mais subconjuntos disjuntos, sendo que cada um é associado a um nó filho. Cada novo nó abrange um subconjunto do conjunto de dados que é recursivamente separado até que o subconjunto

associado a cada nó folha consista inteiramente ou predominantemente de registros de uma mesma classe (GOLDSHMIDT; PASSOS; BEZERRA, 2015).

Um dos algoritmos clássicos utilizados para construção de árvores de decisão é o C4.5 (GOLDSHMIDT; PASSOS; BEZERRA, 2015). O algoritmo C4.5 realiza a avaliação dos pontos de separação calculando a entropia $H(T)$, que é dada pela Equação 2.1.

$$H(T) = -\sum_{j=1}^{k} \frac{freq(c_j, T)}{|T|} \cdot \log_2 \frac{freq(c_j, T)}{|T|} \tag{2.1}$$

onde:

- T é o conjunto de registros de entrada;
- $freq(c_j, T)$ é a quantidade de registros da classe c_j em T;
- $|T|$ é o número total de classes do conjunto T;
- k indica o número de classes distintas que ocorrem em registros de T.

A entropia de um conjunto de dados assume valores entre 0 e 1, esse valor representa a pureza do conjunto de dados. De acordo com Tan, Steinbach e Kumar (2009, p. 187) "quanto menor o grau de impureza, mais distorcida é a distribuição de classes", ou seja, se a entropia é 0, todos os registros pertencem a uma mesma classe, já se a entropia é 1, cada registro pertence a uma classe distinta.

O próximo passo realizado pelo algoritmo C4.5 é avaliar o atributo previsor mais adequado para associar a um nó da Árvore de Decisão. Para isso é necessário realizar o cálculo da $info_{Ai_{(T)}}$, que é o valor esperado da entropia uma vez que T é dividido de acordo com os valores de A_i. O cálculo de $info_{Ai_{(T)}}$ é realizado pela Equação 2.2.

$$info_{Ai_{(T)}} = \sum_{j=1}^{n} \frac{|T_j|}{|T|} \cdot H(T_j) \tag{2.2}$$

onde:

- T é o conjunto de registros de entrada;
- T_j é cada um dos subconjuntos da partição induzida por A_i sobre T;

- $|T_j|$ é a quantidade de registros em T_j.

Depois do cálculo da $info_{Ai_{(T)}}$, é realizado o cálculo de $GInfo(A_i, T)$, chamado ganho de informação, dado pela Equação 2.3. Esse cálculo deve ser realizado para todos os atributos previsores ainda remanescentes a cada iteração.

$$GInfo(A_i, T) = H(T) - info_{Ai_{(T)}} \tag{2.3}$$

O algoritmo C4.5 seleciona o atributo com maior ganho de informação para o nó da árvore. A partir disso, o conjunto T é subdividido e o procedimento é repetido até que cada conjunto T_j seja associado totalmente ou predominantemente a uma classe (ou seja, vai se tornar folha).

2.3.2 Exemplo de Árvore de Decisão

Para exemplificar o processo descrito acima, será construída a árvore de decisão a partir de um conjunto de dados. Esse problema, de Goldshmidt, Passos e Bezerra (2015), consiste em construir uma Árvore de decisão considerando as condições climáticas de um determinado dia e a decisão de jogar golfe ou não jogar golfe. As informações estão dispostas na Tabela 2.2.

Tabela 2.2: Conjunto de dados

Aparência	Temperatura (ºF)	Umidade (%)	Vento	Jogar Golfe
Ensolarada	75	70	Sim	Sim
Ensolarada	80	90	Sim	Não
Ensolarada	85	85	Não	Não
Ensolarada	72	95	Não	Não
Ensolarada	69	70	Não	Sim
Nublada	72	90	Sim	Sim
Nublada	83	78	Não	Sim
Nublada	64	65	Sim	Sim
Nublada	81	75	Não	Sim
Chuvosa	71	80	Sim	Não
Chuvosa	65	70	Sim	Não
Chuvosa	75	80	Não	Sim
Chuvosa	68	80	Não	Sim
Chuvosa	70	96	Não	Sim

Fonte: Goldshmidt, Passos e Bezerra (2015)

Neste problema, nota-se que o atributo-alvo da Classificação é Jogar Golfe, o qual possui duas classes, Sim e Não, com nove e cinco ocorrências respectivamente.

O cálculo da entropia de T é dado pela Equação 2.4.

$$H(T) = -\Big(\frac{9}{14} \cdot \log_2 \frac{9}{14}\Big) - \Big(\frac{5}{14} \cdot \log_2 \frac{5}{14}\Big) = 0,940 \tag{2.4}$$

Em seguida, o C4.5 calcula a $info_{Ai_{(T)}}$. A Equação 2.5 calcula a $info(Aparência, T)$.

$$\begin{aligned} info(Aparência, T) =& \frac{5}{14} \cdot \Big(-\frac{2}{5} \cdot \log_2 \frac{2}{5} - \frac{3}{5} \cdot \log_2 \frac{3}{5}\Big) \\ &+ \frac{4}{14} \cdot \Big(-\frac{4}{4} \cdot \log_2 \frac{4}{4} - \frac{0}{4} \cdot \log_2 \frac{0}{4}\Big) \\ &+ \frac{5}{14} \cdot \Big(-\frac{3}{5} \cdot \log_2 \frac{3}{5} - \frac{2}{5} \cdot \log_2 \frac{2}{5}\Big) \\ info(Aparência, T) =& 0,347 + 0 + 0,347 = 0,694 \end{aligned} \tag{2.5}$$

Na segunda parcela da Equação 2.5 observa-se um argumento zero para a função log. Segundo Goldshmidt, Passos e Bezerra (2015), a definição de entropia considera neste caso o valor resultante igual a zero.

A equação 2.6 calcula a $info(Temperatura, T)$.

$$\begin{aligned} info(Teperatura, T) =& \frac{4}{14} \cdot \Big(-\frac{2}{4} \cdot \log_2 \frac{2}{4} - \frac{2}{4} \cdot \log_2 \frac{2}{4}\Big) \\ &+ \frac{10}{14} \cdot \Big(-\frac{7}{10} \cdot \log_2 \frac{7}{10} - \frac{3}{10} \cdot \log_2 \frac{3}{10}\Big) \\ info(Teperatura, T) =& 0,286 + 0,629 = 0,915 \end{aligned} \tag{2.6}$$

A equação 2.7 calcula a $info(Umidade, T)$.

$$\begin{aligned} info(Umidade, T) =& \frac{9}{14} \cdot \Big(-\frac{5}{9} \cdot \log_2 \frac{5}{9} - \frac{4}{9} \cdot \log_2 \frac{4}{9}\Big) \\ &+ \frac{5}{14} \cdot \Big(-\frac{4}{5} \cdot \log_2 \frac{4}{5} - \frac{1}{5} \cdot \log_2 \frac{1}{5}\Big) \\ info(Umidade, T) =& 0,637 + 0,258 = 0,895 \end{aligned} \tag{2.7}$$

A equação 2.8 calcula a $info(Vento, T)$.

$$\begin{aligned} info(Vento, T) =& \frac{6}{14} \cdot \Big(-\frac{3}{6} \cdot \log_2 \frac{3}{6} - \frac{3}{6} \cdot \log_2 \frac{3}{6}\Big) \\ &+ \frac{8}{14} \cdot \Big(-\frac{6}{8} \cdot \log_2 \frac{6}{8} - \frac{2}{8} \cdot \log_2 \frac{2}{8}\Big) \\ info(Vento, T) =& 0,429 + 0,0464 = 0,893 \end{aligned} \tag{2.8}$$

A partir disso, conforme as equações abaixo, é possível fazer o cálculo do ganho de informação de cada um dos atributos.

$$GInfo(Aparência, T) = 0,940 - 0,694 = 0,246 \tag{2.9}$$

$$GInfo(Temperatura, T) = 0,940 - 0,915 = 0,025 \tag{2.10}$$

$$GInfo(Umidade, T) = 0,940 - 0,895 = 0,045 \tag{2.11}$$

$$GInfo(Vento, T) = 0,940 - 0,893 = 0,047 \tag{2.12}$$

O atributo previsor com maior ganho de informação é selecionado como nó raiz da árvore, nesse caso, o atributo Aparência. Depois disso, subdivide-se o conjunto T e o procedimento é repetido para cada novo nó gerado. Informações adicionais sobre o processo de avaliação de pontos de separação podem ser pesquisados em Goldshmidt, Passos e Bezerra (2015).

A partição final do conjunto de dados está nas tabelas abaixo.

- Aparência = Ensolarada:

  Umidade $\leqslant$ 75: Jogar = Sim

Tabela 2.3: Aparência Ensolarada/Jogar Sim

Aparência	Temperatura (ºF)	Umidade (%)	Vento	Jogar Golfe
Ensolarada	75	70	Sim	Sim
Ensolarada	69	70	Não	Sim

Fonte: Goldshmidt, Passos e Bezerra (2015)

  Umidade $>$ 75: Jogar = Não

Tabela 2.4: Aparência Ensolarada/Jogar Não

Aparência	Temperatura (ºF)	Umidade (%)	Vento	Jogar Golfe
Ensolarada	80	90	Sim	Não
Ensolarada	85	85	Não	Não
Ensolarada	72	95	Não	Não

Fonte: Goldshmidt, Passos e Bezerra (2015)

- Aparência = Nublada:

  Aparência = Nublada: Jogar = Sim

Tabela 2.5: Aparência Nublada/Jogar Sim

Aparência	Temperatura (ºF)	Umidade (%)	Vento	Jogar Golfe
Nublada	72	90	Sim	Sim
Nublada	83	78	Não	Sim
Nublada	64	65	Sim	Sim
Nublada	81	75	Não	Sim

Fonte: Goldshmidt, Passos e Bezerra (2015)

- Aparência = Chuvosa:

  Vento = Sim: Jogar = Não

Tabela 2.6: Aparência Chuvosa/Jogar Não

Aparência	Temperatura (ºF)	Umidade (%)	Vento	Jogar Golfe
Chuvosa	71	80	Sim	Não
Chuvosa	65	70	Sim	Não

Fonte: Goldshmidt, Passos e Bezerra (2015)

  Vento = Não: Jogar = Sim

Tabela 2.7: Aparência Chuvosa/Jogar Sim

Aparência	Temperatura (ºF)	Umidade (%)	Vento	Jogar Golfe
Chuvosa	75	80	Não	Sim
Chuvosa	68	80	Não	Sim
Chuvosa	70	96	Não	Sim

Fonte: Goldshmidt, Passos e Bezerra (2015)

Com isso, a árvore de decisão da Figura 2.2 pode ser lida a partir das seguintes regras de decisão:

Aparência = Ensolarada:

  Umidade $\leqslant 75$: Jogar = Sim

  Umidade > 75: Jogar = Não

Aparência = Nublada:

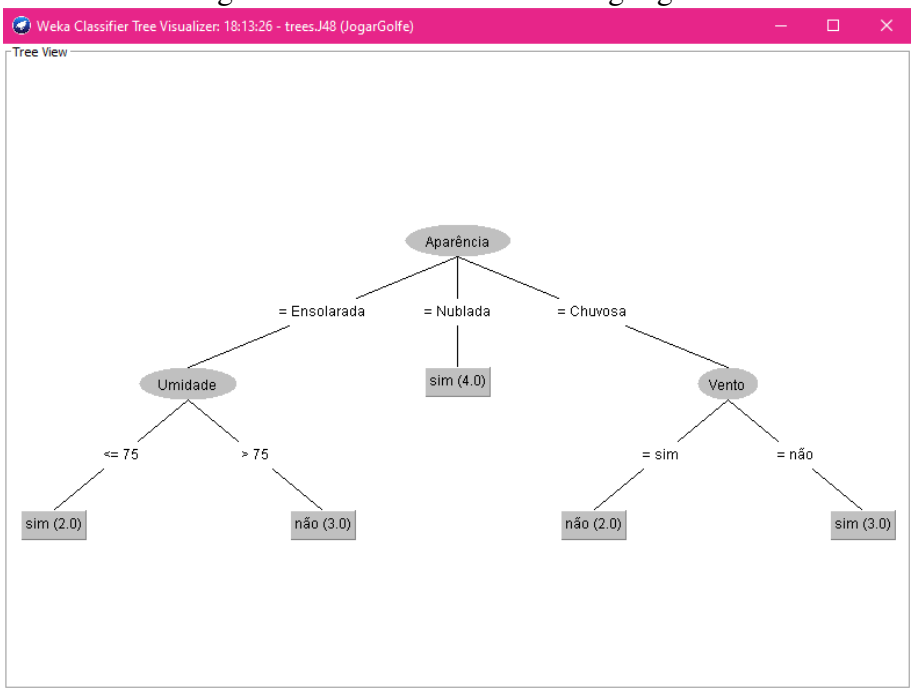
  Jogar = Sim

Aparência = Chuvosa:

Vento = Sim: Jogar = Não

Vento = Não: Jogar = Sim

Figura 2.2: Árvore de decisão Jogar golfe



Fonte: Acervo da autora

2.3.3 Avaliação do modelo de classificação

Há dois tipos de erros cometidos por um modelo de classificação, sendo eles, o erro de treinamento e o erro de generalização, que são respectivamente, o número de erros de classificação ocorridos nos registros de treinamento e o número de erros esperados do modelo em registros não visto anteriormente (conjunto de teste) (TAN; STEINBACH; KUMAR, 2009).

No *overfitting* um modelo é muito apropriado para o conjunto de treinamentos, entretanto, possui um grande erro de generalização, ou seja, é ineficaz para prever novos resultados (conjunto de teste). Já no *underfitting*, tanto o erro de treinamento quanto o erro de generalização são grandes, ou seja, não se garante acurácia relevante no conjunto de treinamento e também não garante acurácia relevante no conjunto de testes (TAN; STEINBACH; KUMAR, 2009).

2.3.4 Validação Cruzada

Alguns dos métodos utilizados para avaliar o desempenho de um classificador são: Método *holdout*, Sub-Amostragem Aleatória, Validação Cruzada e Bootstrap (TAN; STEINBACH; KUMAR, 2009). Explicaremos a seguir o método de Validação Cruzada.

No método de Validação Cruzada cada registro é utilizado o mesmo número de vezes para treinamento e uma vez para teste. A validação cruzada de duas partes, por exemplo, ocorre particionando o conjunto de dados em dois subconjuntos de mesmo tamanho. Escolhe-se um para ser o conjunto de treinamento e outro para conjunto de teste e em seguida, o conjunto de treinamento passa a ser o conjunto de teste e vice-versa. O erro é calculado pela soma dos erros de ambas as execuções (TAN; STEINBACH; KUMAR, 2009).

Na validação cruzada utilizada neste trabalho, o conjunto de dados é dividido em 10 partições. Escolhe-se um subconjunto para teste e os outros nove serão para treinamento e isso será repetido até que todos os subconjuntos sejam utilizados como teste uma vez.

Generalizando, na Validação Cruzada toma-se k partições iguais, em cada uma das execuções uma das partições é escolhida para teste e as outras para treinamento, repete-se k vezes. Nesse caso, o erro será a soma dos erros das k execuções (TAN; STEINBACH; KUMAR, 2009).

2.4 Trabalhos Correlatos

A Mineração de Dados Educacionais, traduzida de Educational Data Mining (EDM), recebe este nome pelo fato de os dados utilizados para o processo KDD serem extraídos de contextos educacionais, também, essa área "procura desenvolver ou adaptar métodos e algoritmos de mineração existentes, de tal modo que se prestem a compreender melhor os dados em contextos educacionais" (COSTA et al., 2012, p.4).

Romero e Ventura (2010) apontam que a Mineração de Dados Educacionais converte dados provenientes de sistemas educacionais, transformando em informações úteis, combinando três áreas do conhecimento: Computação, Educação e Estatística. Além disso, os mesmos autores citam alguns exemplos para a utilização do EDM, como para prever desempenho dos estudantes, detectar comportamentos indesejáveis como a evasão, visualizar a evolução do progresso dos alunos e etc., utilizando as tarefas de Classificação e Clusterização.

Alguns trabalhos que utilizam a Mineração de Dados Educacionais e suas tarefas para diagnosticar causas da evasão em cursos de graduação serão descritos a seguir, bem como os resultados obtidos com essas pesquisas.

Oliveira Júnior, Noronha e Kaestner (2017) realizaram uma abordagem genérica a partir de dados do sistema acadêmico da Universidade Tecnológica Federal do Paraná, analisando quais atributos são mais relevantes para prever a evasão através da mineração de dados. A pesquisa revelou que os novos atributos criados foram os que mais contribuíram na tarefa de previsão de evasão, sendo que o atributo 'dificuldade média das disciplinas cursadas pelo aluno' melhorou a acurácia dos algoritmos de classificação.

Couto e Santana (2017) utilizaram nove algoritmos de classificação para diagnosticar as causas de evasão e retenção dos cursos de graduação da Universidade Federal do Pará, através dos registros acadêmicos obtidos pelo Sistema Integrado de Gestão de Atividades Acadêmicas. Pelo algoritmo Bayesian Network, com precisão de 86%, foi possível concluir que o desempenho do estudante depende diretamente do número de trancamentos, do Índice de Eficiência Acadêmico e do percentual de reprovações durante o curso.

Para identificar estudantes em risco de evasão Lanes e Alcântara (2018) utilizaram o algoritmo de classificação J48, com os registros do sistema acadêmico da Universidade Federal do Rio Grande. Com acurácia de 90,7%, foi possível perceber que os alunos que possuíam baixo coeficiente de rendimento evadiram, já alunos com alto coeficiente de rendimento e que são bolsistas concluíram o curso.

Paz e Cazella (2017) desenvolveram um estudo de caso aplicando o processo de Descoberta de Conhecimento em Bases de Dados. Para isso, analisaram 4697 alunos matriculados em curso de graduação da Universidade Comunitária do Rio Grande do Sul no segundo semestre de 2016, e com a árvore de decisões do algoritmo classificador J48 obtiveram acurácia de 90% nos resultados obtidos. Os resultados obtidos apontaram que os casos de evasão estão diretamente relacionados com os incentivos fornecidos (bolsas) e o currículo dos alunos.

O trabalho de Quinot (2018) utiliza a Mineração de Dados Educacionais para avaliar informações socioeconômicas das escolas frente ao desempenho das mesmas na Prova Brasil. Com a utilização do algoritmo J48 foi possível identificar atributos relevantes em escolas que possuem alto desempenho, esses atributos são: existência de auditório, brinquedoteca e poucos

sinais de depredação. Já em escolas com pouco desempenho o atributo relevante é a falta de linha telefônica, a falta de internet para os alunos e falta de computadores para os professores.

Capítulo 3

Metodologia

A pesquisa teve por finalidade explicar e analisar os fatores que determinam a evasão dos acadêmicos ingressantes nos anos 2010 a 2012 do curso de Licenciatura em Matemática da Universidade Estadual do Oeste do Paraná, *campus* Cascavel.

Foram analisadas quais disciplinas mais influenciam para um aluno evadir, de modo que seja possível prever alunos em risco de evasão.

3.1 Coleta de Dados

Para realizar a pesquisa documental, por se tratar de uma pesquisa que envolve o ser humano de forma indireta, o projeto do trabalho foi encaminhado ao Comitê de Ética em Pesquisa com Seres Humanos (CEP) da Universidade Estadual do Oeste do Paraná. A pesquisa foi aprovada pelo CEP, conforme Parecer Consubstanciado do CEP número 4.281.363. Os documentos foram requeridos e liberados pela Pró-Reitoria de Graduação.

Os documentos utilizados para formar a base de dados são históricos informais de alunos ingressantes nos anos de 2009 a 2012. A escolha dos ingressantes nesses anos foi dada pelo fato de que o prazo máximo para conclusão do curso de Licenciatura em Matemática é de até 7 anos, prorrogável para mais um ano. Entretanto, no ano de 2009 o Projeto Político Pedagógico do curso sofreu alterações sendo estas implementadas em 2010. Desta forma, os dados referentes aos ingressantes no ano de 2009 foram excluídos desta pesquisa.

Para formar a base de dados, as seguintes informações foram obtidas dos históricos informais: número de vezes que o aluno cursou cada uma das disciplinas, se foi aprovado, se prestou exame quando foi aprovado em cada disciplina e a maior frequência entre as vezes que cursou

cada uma das disciplinas.

Na etapa de pré-processamento, alguns dados foram removidos da base de dados. Os dados removidos foram de alunos que possuíam aproveitamento, pelo fato de que não era possível obter o valor correto para os atributos, visto que não havia o número de vezes cursadas na disciplina, bem como se realizou exame e qual a frequência obtida em cada disciplina. Esses alunos que possuiam aproveitamento totalizaram 19 instâncias excluídas da base de dados. Além disso, um aluno ingressante 2011 foi excluído da base de dados por ainda estar matriculado no curso.

Quanto às classes desse modelo, tínhamos inicialmente 4 classes: Cancelado por abandono, Cancelado, Tranferido para outra Instituição de Ensino Superior e Formado. Entretanto, ao inicar a classificação dos dados, o modelo de classificação não classificou nenhuma instância como Cancelado ou Tranferido para outra Instituição de Ensino Superior, por isso, optou-se pela remoção das instâncias classificadas como tais. Além disso, essas instâncias foram removidas por não serem relevantes para essa pesquisa.

Com isso, a base de dados ficou composta por 88 instâncias, as quais correspondem aos dados de 88 alunos que compõem esta pesquisa.

Para a Mineração de Dados Educacionais, foi utilizada a tarefa de Classificação, que busca modelos para predizer a classe baseando-se nos demais atributos para um dado registro. Na classificação dos dados o algoritmo utilizado foi o J48, que é uma implementação do algoritmo C4.5 no WEKA.

3.2 WEKA

Waikato Environment for Knowledge Analysis - WEKA é um *software* livre de aprendizagem de máquina de código aberto, criado na Universidade de Waikato, Nova Zelândia. Algumas características dessa ferramenta de KDD, o WEKA, é que ele está implementado na linguagem Java e pode ser utilizado em diferentes plataformas, como *Windows*, *Linux*, MAC OS e outras (WAIKATO, 2010).

O WEKA fornece inúmeras implementações, ou seja, algoritmos para realizar as tarefas de KDD. Além das tarefas de mineração de dados, o *software* realiza o pré-processamento, o pós-processamento e é capaz de analisar o desempenho do algoritmo aplicado ao banco de dados (WITTEN; FRANK, 2005).

Para a leitura dos dados pelo WEKA, é necessário que os dados estejam em um arquivo no formato ARFF. O arquivo ARFF é constituído por duas partes, a primeira parte possui uma lista com todos os atributos, já na segunda parte é composta pelos registros, ou seja, os dados a serem minerados com o valor de cada atributo (DAMASCENO, 2010).

No caso do exemplo da Tabela 2.2, a Figura 3.1 mostra o arquivo ARRF referente a esse conjunto de dados, no qual a primeira parte é composta pelos atributos, precedidos por @atributte e a segunda parte é precedida por @data e contém os registros com o valor de cada atributo separado por vírgula, sendo que cada linha corresponde a um registro.

Figura 3.1: Formato ARFF

```
@relation JogarGolfe

@attribute Aparência {Ensolarada, Nublada, Chuvosa}
@attribute Temperatura real
@attribute Unidade real
@attribute Vento {sim, não}
@attribute Jogar Golfe {sim, não}

@data
Ensolarada,75,70,sim,sim
Ensolarada,82,90,sim,não
Ensolarada,85,85,não,não
Ensolarada,72,95,não,não
Ensolarada,69,70,não,sim
Nublada,72,92,sim,sim
Nublada,83,78,não,sim
Nublada,64,65,sim,sim
Nublada,81,75,não,sim
Chuvosa,71,80,sim,não
Chuvosa,65,70,sim,não
Chuvosa,75,82,não,sim
Chuvosa,68,80,não,sim
Chuvosa,78,96,não,sim
```

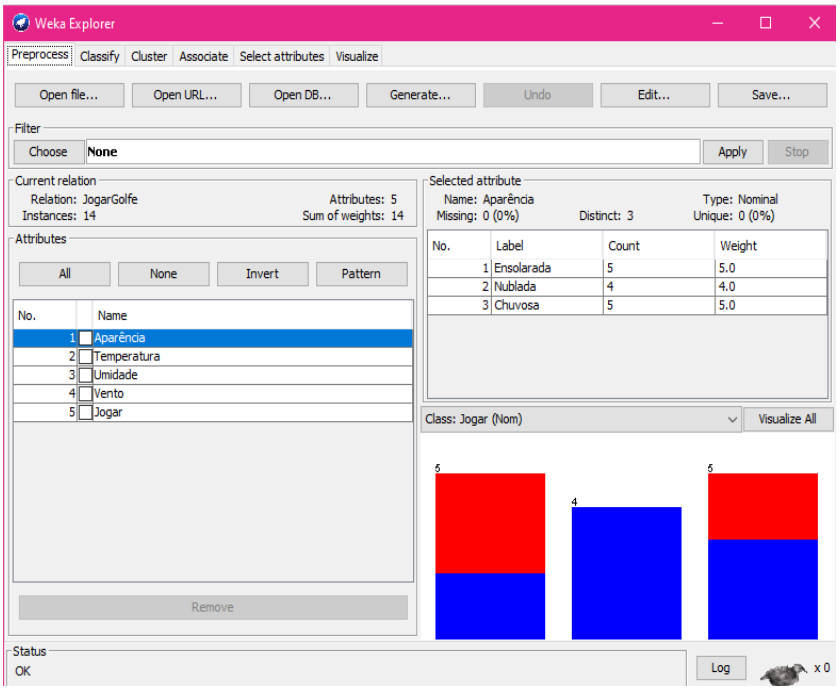
Fonte: Acervo da autora

A interface do WEKA que será utilizada para a realização do trabalho é o *Explorer*, nela há seis painéis: pré-processamento, classificação, clusterização, associação, seleção de atributos e visualização (WITTEN; FRANK, 2005). Nessa interface será possível realizar todas as etapas do processo KDD, inclusive analisar a acurácia dos resultados obtidos.

Para exemplificar, abrindo o arquivo ARFF da Figura 3.1 na interface *Explorer*, o arquivo é

direcionado ao painel de pré-processamento e algumas informações são exibidas: o número de atributos, o número de instâncias, uma análise estatística da variável selecionada e um gráfico de colunas com informações da mesma. Neste caso, tem-se a tela conforme Figura 3.2.

Figura 3.2: Explorer

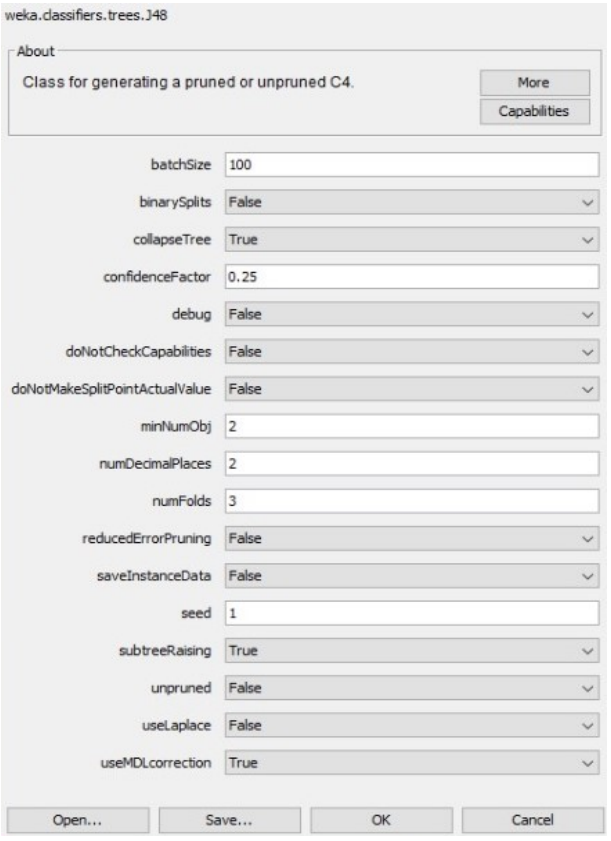


Fonte: Acervo da autora

No painel Classificação, entre os algoritmos disponíveis para classificação no software WEKA, destaca-se o J48, que gera árvores de decisão a partir de uma implementação do algoritmo C4.5, desenvolvido por J. Ross Quinlan (HAN; KAMBER; PEI, 2012).

O algoritmo J48 possui a configuração de parâmetros e respectivos valores descritos na Figura 3.3. Tais parâmetros são explicados em Quinot (2018) e Appio (2012).

Figura 3.3: Parâmetros J48



Fonte: Acervo da autora

Optou-se pela utilização dos parâmetros do algoritmo J48 no WEKA, exceto o parâmetro NumMinObj (que define o número mínimo de instâncias por folha). Esta escolha se deve ao fato de que uma regra de decisão para um número menor de 10 alunos seria muito restrita. Para melhor desempenho do classificador também optou-se pelo uso da Validação Cruzada.

Além disso, nesta etapa, é utilizada a validação cruzada com 10 folds. Isso significa que o conjunto total de dados foi dividido em 10 subconjuntos do mesmo tamanho e mutuamente exclusivos, em seguida, um subconjunto é utilizado para o teste e os outros 9 subconjuntos são utilizados para calcular a acurácia. Esse processo é realizado 10 vezes alternando o subconjunto utilizado para realizar o teste.

Capítulo 4

Resultados e discussões

Primeiramente, foi traçado a partir dos resultados de um questionário socioeducacional aplicado aos alunos durante a inscrição para o vestibular, um perfil dos ingressantes nos anos de 2010 a 2012, totalizando 119 alunos (e não 120 como era previsto). Essas informações foram obtidas no Portal Unioeste.

4.1 Perfil dos alunos ingressantes entre 2010 e 2012

Abaixo segue o perfil do aluno ingressante de 2010 a 2012 no curso de Licenciatura em Matematica. Observe que essas informacões foram coletadas quando os acadêmicos ingressantes no curso prestaram o vestibular.

Sobre a renda mensal da família dos acadêmicos, verifica-se que 41% possuíam renda total mensal de até 2 salários mínimos. Dentre os acadêmicos do curso, 118 residiam no Paraná e 49% moravam com os pais. Também, 79 alunos trabalhavam, sendo que 34 trabalhavam no setor de comércio e 29 no setor de prestação de serviços.

Quanto a questões educacionais, mais de 85% dos alunos cursaram o Ensino Fundamental e Médio em escolas públicas e 76% não frequentaram cursos preparatórios para o vestibular. Além disso, 47% dos alunos optaram prestar vestibular na Unioeste por ser uma universidade pública que atende às condicões socioeconômicas da família e do próprio aluno.

Considerando os dados utilizados na base de dados, antes da etapa de pré processamento, 36 dos alunos eram cotistas e 62 ingressaram na instituição como não cotista.

4.2 Resultados preliminares

Realizando uma análise preliminar dos dados, foi possível notar nos anos de 2010 e 2012, 78 alunos tiveram sua matrícula cancelada por abandono, conforme tabela 4.1.

Tabela 4.1: Cancelados por abandono

Ano de ingresso	Cancelamentos por abandono
2010	33
2011	24
2012	21

Fonte: Elaborado pela autora a partir base de dados

Observando as disciplinas do primeiro ano do curso, entre os anos de 2010 a 2012, a média do número de vezes que os alunos cursaram cada disciplina foi:

- Geometria Analítica e Vetorial - 1,67
- Cálculo Diferencial e Integral I - 1,65
- Complementos da Matemática - 1,60
- Fundamentos da Matemática - 1,46
- Geometria Euclidiana I - 1,40
- Desenho Geométrico - 1,34
- Laboratório de Ensino - 1,34

Dentre todas as disciplinas do curso é possível notar que as disciplinas com maior número de reprovações (proporcional a quantidade de alunos que cursaram a disciplina) são do primeiro ano do curso, entretanto, a disciplina de Álgebra, ofertada no terceiro ano do curso, foi a disciplina com maior média de reprovações, cerca de 1,79 vez cursada.

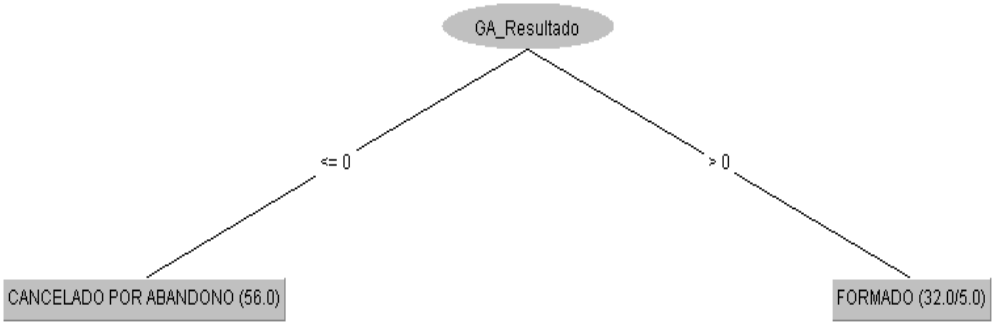
4.3 Resultados obtidos com a Mineração de Dados Educacionais

Na aplicação do algoritmo para classificação optou-se inicialmente pela utilização de apenas variáveis relativas às disciplinas do primeiro ano, dado que é onde ocorre o maior número de

cancelados por abandono.

A Base de Dados 1 ficou composta portanto pelas 7 disciplinas cursadas no primeiro ano do curso, totalizando 31 atributos e 88 registros. A Figura 4.1 mostra a árvore de decisão gerada para a base de dados 1.

Figura 4.1: Árvore de Decisão para o melhor modelo - Base de dados 1



Fonte: Acervo da autora

Com essa árvore de decisão pode-se notar que a disciplina Geometria Analítica e Vetorial, ofertada no primeiro ano do curso, é a disciplina que mais influencia para que os alunos não concluam o curso. Também é válido relembrar que Geometria Analítica e Vetorial é a disciplina do primeiro ano com maior número de reprovações.

Essa regra foi válida para 56 alunos, que representam aproximadamente 63,6% dos alunos. Além disso, esse modelo classificou 32 alunos como formados, sendo que 5 foram classificados incorretamente. Neste caso os alunos aprovaram em Geometria Analítica e Vetorial e não formaram.

A matriz de confusão desse modelo é a representada na Figura 4.2. Sabemos que a precisão é dada pelo quociente dos número de previsões (classes previstas corretamente) corretas pelo número total de previsões, logo, a precisão da classe Cancelada por Abandono é dada por $\frac{56}{56} = 1$, portanto a precisão dessa classe é 100% de classes previstas corretamente. Já a precisão da classe Formado é dada por $\frac{27}{32} = 0,844$, portanto, a precisão dessa classe é 84,4%.

Figura 4.2: Matriz confusão

```
=== Confusion Matrix ===

  a  b   <-- classified as
 56  5 |  a = CANCELADO POR ABANDONO
  0 27 |  b = FORMADO
```

Fonte: Acervo da autora

A precisão do modelo é calculada a partir da precisão de cada classe, ou seja,

$$\frac{1 \cdot 61 + 0,844 \cdot 27}{88} = 0,925 \tag{4.1}$$

No WEKA também é realizado o cálculo do Recall. O recall é dado pela razão entre o número de classes classificadas corretamente, pelo total de classes. O recall da classe Cancelada por Abandono é calculada por $\frac{56}{61} = 0,918$, o recall da classe Formado é $\frac{27}{27} = 1$.

Do mesmo modo, o recall do modelo é calculado por

$$\frac{0,918 \cdot 61 + 1 \cdot 27}{88} = 0,943 \tag{4.2}$$

A figura abaixo é fornecida pelo WEKA e mostra a precisão e o recall das classes e do modelo, encontrada nas equações acima.

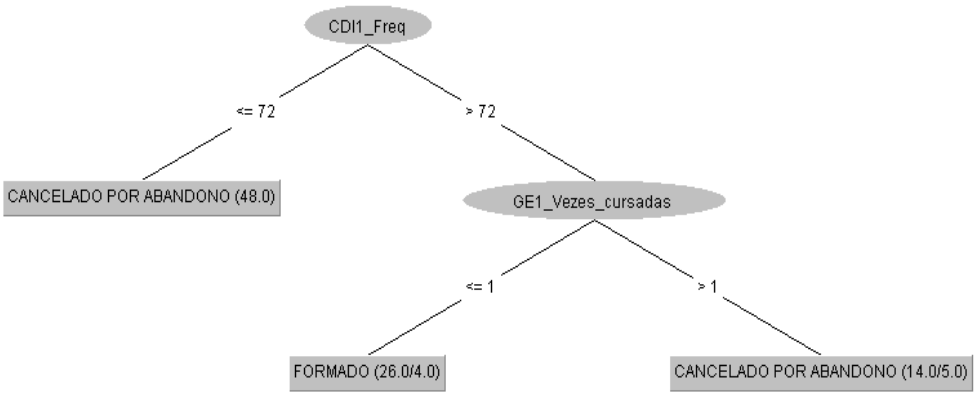
Figura 4.3: Precisão e Recall

Precision	Recall	Class
1,000	0,918	CANCELADO POR ABANDONO
0,844	1,000	FORMADO
0,952	0,943	

Fonte: Acervo da autora

Utilizando a mesma base de dados, mas removendo o atributo resultado de cada disciplina, temos a árvore de decisão da Figura 4.4.

Figura 4.4: Árvore de Decisão para o melhor modelo - Base de dados 1 sem o atributo Resultado

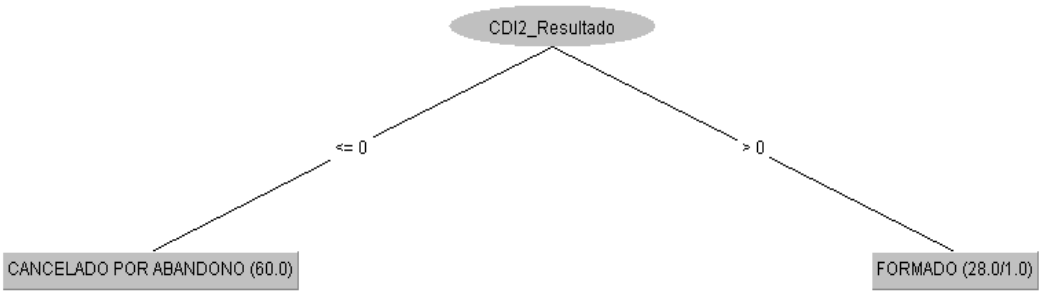


Fonte: Acervo da autora

Neste modelo, com precisão média de 89,7%, o fator mais decisivo foi a frequência na disciplina de Cálculo Diferencial e Integral 1. Os alunos que possuem frequência menor ou igual a 72% tiveram suas matrículas canceladas por abandono. Já os alunos com frequência maior que 72% e que cursaram Geometria Euclidiana uma única vez, se formaram. Essa regra foi alcançada por 26 alunos, sendo que para 4 a classificação foi incorreta. Por fim, quem teve frequência maior que 72% e cursou Geometria Euclidiana mais de uma vez, teve sua matrícula cancelada por abandono, ou seja, quem não abandonou Cálculo Diferencial e Integral 1 mas cursou Geometria Euclidiana 1 mais de uma vez teve a matrícula cancelada por abandono. Essa regra foi válida para 14 alunos e para 5 alunos a classificação foi incorreta.

Considerando agora uma base de dados composta por dados relativos aos dois primeiros anos de curso, temos a Base de Dados 2, composta pelas 7 disciplinas do primeiro ano e 8 disciplinas do segundo ano, totalizando 63 atributos e 88 instâncias. No modelo obtido através da árvore de decisão da Figura 4.5, a disciplina mais relevante para um aluno evadir, é Cálculo Diferencial e Integral 2. O modelo obtido possui precisão média de 98,9%.

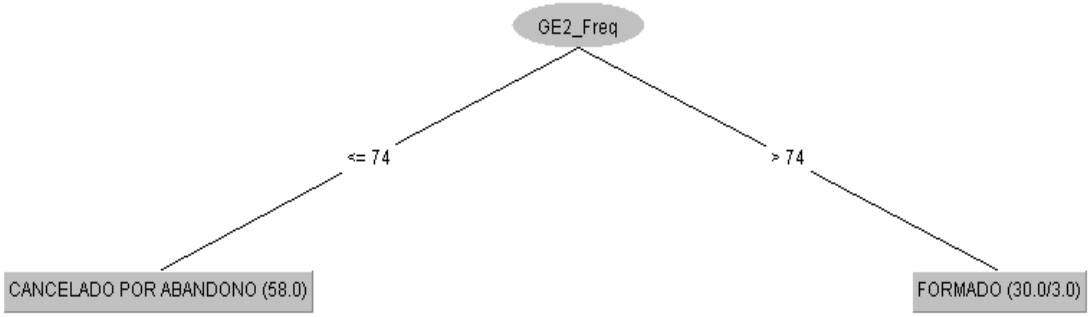
Figura 4.5: Árvore de Decisão para o melhor modelo - Base de dados 2



Fonte: Acervo da autora

Com a Base de Dados 2, removendo o atributo Resultado, restaram 48 atributos. A regra gerada possui precisão média de 94,4% e é a seguinte: se o aluno teve frequência na disciplina de Geometria Euclidiana 2 menor ou igual a 74% é cancelado por abandono e se a frequência é maior que 74% é formado, conforme Figura 4.6. Vale ressaltar que 75% é a frequência mínima exigida para que o acadêmico seja aprovado na disciplina. A classificação foi incorreta apenas para três alunos que foram classificados como formado, mas tiveram a matrícula cancelada por abandono.

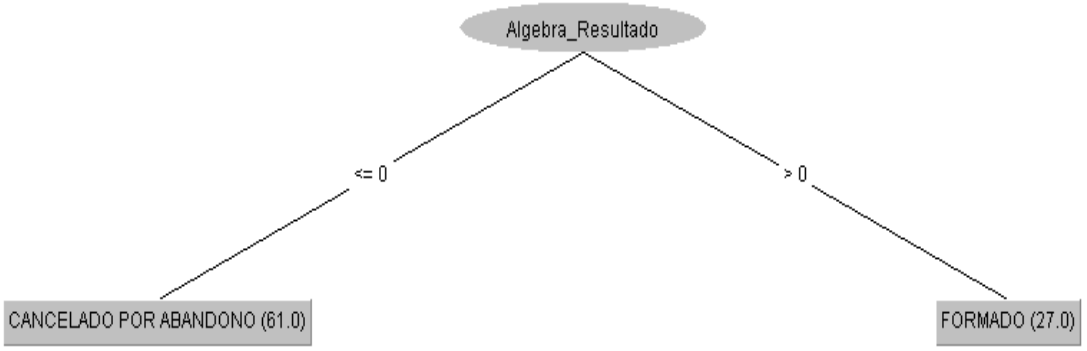
Figura 4.6: Árvore de Decisão para o melhor modelo - Base de dados 2 sem o atributo Resultado



Fonte: Acervo da autora

A Base de Dados 3, é composta pelas 7 disciplinas do primeiro ano, 8 do segundo ano e 7 do tereiro ano, totalizando 91 atributos. Com essa base de dados, foi possível analisar que entre todos esses atributos, o mais relevante para a evasão foi a discplina de Álgebra, ofertada no terceiro ano do curso (ver Figura 4.7). O modelo teve precisão média de 98,9%.

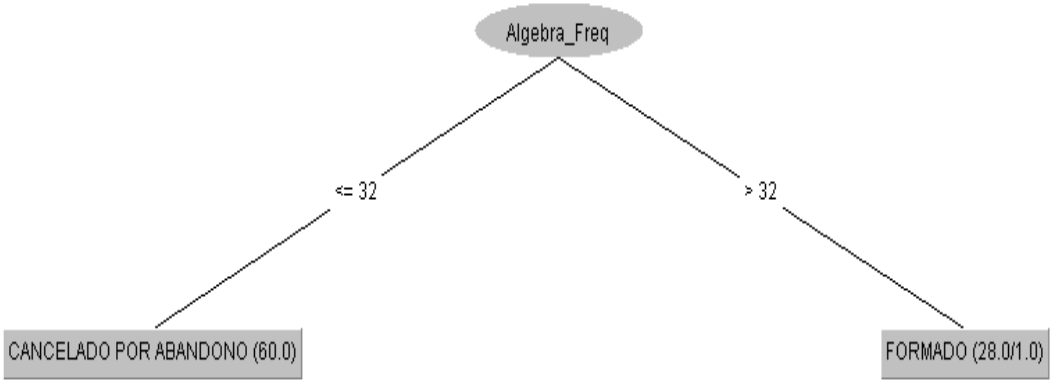
Figura 4.7: Árvore de Decisão para o melhor modelo - Base de dados 3



Fonte: Acervo da autora

Removendo o atributo Resultado, a seguinte regra pode ser observada: se o aluno teve frequência menor ou igual a 32% em Álgebra, então o aluno foi cancelado por abandono, caso contrário, o aluno se formou. Essa regra foi válida para quase todos os alunos, exceto um que foi classificado como formado, mas teve sua matrícula cancelada por abandono, conforme árvore da Figura 4.8. O modelo obtido possui precisão média de 96,7%.

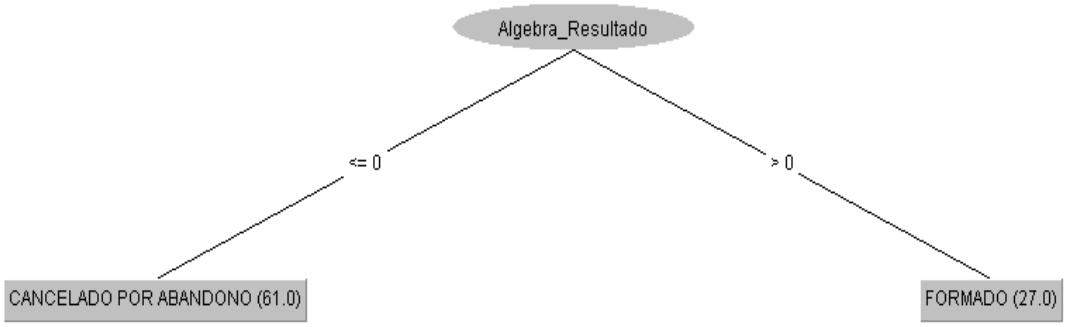
Figura 4.8: Árvore de Decisão para o melhor modelo - Base de dados 3 sem o atributo Resultado



Fonte: Acervo da autora

A base de dados 4, é composta por 123 atributos, sendo que são analisadas todas as disciplinas dos 4 anos de curso e 88 instâncias. A árvore de decisão da Figura 4.9 corresponde a classificação dos dados dessa base de dados.

Figura 4.9: Árvore de Decisão para o melhor modelo - Base de dados 4



Fonte: Acervo da autora

Com a árvore de decisão gerada, Álgebra é a disciplina do curso que mais influencia para a evasão. Esse modelo foi gerado com precisão média de 98,9%.

Capítulo 5

Considerações finais

Neste trabalho foi realizada uma pesquisa documental com os ingressantes 2010 a 2012 do curso de Licenciatura em Matemática da Universidade Estadual do Oeste do Paraná, campus Cascavel. A partir disso, buscamos encontrar padrões que descrevam alunos com maior propensão ao abandono, observando dados referentes a vida escolar dos ingressantes no curso de Licenciatura em Matemática nos anos de 2010 a 2012 da Universidade Estadual do Oeste do Paraná – Unioeste, *campus* Cascavel.

Foi possível notar que a evasão ocorre de maneira mais evidente no primeiro ano do curso e analisando as 7 disciplinas que compõem o primeiro ano do curso, foi possível perceber que a disciplina Geometria Analítica e Vetorial é a disciplina que mais influencia para que os alunos abandonem o curso. O modelo que encontrou essa regra teve precisão de 92,5%. Já na análise preliminar, foi possível observar que no primeiro ano do curso a disciplina que foi cursada mais vezes foi Geometria Analítica e Vetorial, em média 1,67 vezes.

Com a base de dados 2, notou-se que Cálculo Diferencial e Integral 2 é a disciplina que mais contribui para o abandono. Além disto, entre as disciplinas do segundo ano do curso, a mesma possui o maior número de vezes cursadas, em média 1,51 vez.

De modo geral, a disciplina de Álgebra, ofertada no terceiro ano do curso, é apontada neste trabalho como a que mais influencia na evasão. A mesma, também possui maior índice de reprovações dentre todas as disciplinas do curso, cursada em média 1,79 vez.

Da mesma forma que autores como Appio (2012), Castro (2013), Souza (2008), Hoed (2016), também pode-se concluir aqui que reprovações sucessivas pode ser uma das causas da evasão.

Vale ressaltar que pelo fato de terem sido analisadas três turmas ingressantes, fatores como

professor da disciplina, desempenho da turma, entre outros, não devem ser motivadores dos resultados apresentados neste trabalho.

Com isso, faz-se necessário medidas institucionais de combate a evasão, visto que, no curso de Licenciatura em Matemática esse índice é muito elevado.

Referências Bibliográficas

ANDRETA, A. A. *Influência de Fatores Socioeconômicos na Evasão de Acadêmicos do Curso de Matemática da Unioeste – Campus de Cascavel*. Dissertação (Monografia) — Universidade Estadual do Oeste do Paraná, 2013.

APPIO, A. *Classificando Dados de Evasão do Curso de Licenciatura em Matemática da Unioeste - Campus Cascavel no Período de 2003-2010*. Dissertação (Monografia) — Universidade Estadual do Oeste do Paraná, 2012.

BAGGI, C. A. dos S.; LOPES, D. A. *Evasão e Avaliação Institucional no Ensino Superior: Uma Discussão Bibliográfica*. 2011. Disponível em: ⟨http://www.scielo.br/pdf/aval/v16n2/a07v16n2.pdf⟩. Acesso em: 22 abr 2020.

BRASIL. *Diplomação, Retenção e Evasão nos Cursos de Graduação em Instituições de Ensino Superior Públicas*. 1997. Disponível em: ⟨http://www.dominiopublico.gov.br/download/texto/me002240.pdf⟩. Acesso em: 08 jun 2020.

CASTRO, L. P. V. de. *Evasão escolar no ensino superior: um estudo nos cursos de licenciatura da Universidade Estadual do Oeste do Paraná UNIOESTE - campus Cascavel*. Dissertação (Dissertação de Mestrado) — Universidade Estadual do Oeste do Paraná, 2013.

COSTA, E. et al. Mineração de dados educacionais: Conceitos, técnicas, ferramentas e aplicações. In: *Jornada de Atualização em Informática na Educação*. [S.l.: s.n.], 2012.

COUTO, D. da C. do; SANTANA, L. de. Mineração de dados educacionais aplicada à identificação de variáveis associadas à evasão e retenção. In: *II Congresso sobre Tecnologia na Educação*. [S.l.: s.n.], 2017. p. 333–344.

DAMASCENO, M. Introdução à mineração de dados utilizando o weka. In: *V Congresso de Pesquisa e Inovação da Rede Norte Nordeste de Educação Tecnológica – CONNEPI*. [S.l.: s.n.], 2010.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, 1996.

GAIOSO, N. P. de L. *O fenômeno da evasão escolar na educação superior no Brasil*. Dissertação (Dissertação) — Universidade Católica de Brasília, 2005.

GOLDSHMIDT, R.; PASSOS, E.; BEZERRA, E. *Data Mining: conceitos, técnicas, algoritmos, orientações e aplicações*. 2. ed. [S.l.]: Elsevier, 2015.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: concepts and techniques*. [S.l.]: Elsevier, 2012.

HOED, R. M. *Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de computação*. Dissertação (Dissertação) — Universidade de Brasília, 2016.

LANES, M.; ALCâNTARA, C. Predição de alunos com risco de evasão: estudo de caso usando mineração de dados. In: *XXIX Simpósio Brasileiro de Informática na Educação (SBIE 2018)*. [S.l.: s.n.], 2018.

Oliveira Júnior, J. G. de; NORONHA, R. V.; KAESTNER, C. A. A. Método de seleção de atributos aplicados na previsão da evasão de cursos de graduação. *Revista de Informática Aplicada*, v. 13, n. 2, p. 54–67, 2017.

PAZ, F. J.; CAZELLA, S. C. Identificando o perfil de evasão de alunos de graduação através da mineração de dados educacionais: um estudo de caso de uma universidade comunitária. In: *Workshops do Congresso Brasileiro de Informática na Educação*. [S.l.: s.n.], 2017. p. 624–633.

POLYDORO, S. A. J. *O trancamento de matrícula na trajetória acadêmica no universitário: condições de saída e de retorno à instituição*. Dissertação (Tese) — Universidade Estadual de Campinas, 2000.

QUINOT, D. A. *Investigação de Características Socioeconômicas Relevantes para o desempenho das Escolas na Prova Brasil*. Dissertação (Monografia) — Universidade Estadual do Oeste do Paraná, 2018.

ROMERO, C.; VENTURA, S. Educational data mining: A review of the state-of-the-art. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews*, 2010.

SILVA FILHO, R. L. L. E. et al. *A Evasão No Ensino Superior Brasileiro*. 2007. Disponível em: ⟨http://www.scielo.br/pdf/cp/v37n132/a0737132.pdf⟩. Acesso em: 18 mar 2020.

SOUZA, L. F. D. de. *Evasão do curso de Licenciatura em Matemática (Noturno) da Universidade de Brasília*. Dissertação (Monografia) — Universidade de Brasília, 2016.

SOUZA, S. L. de. *Evasão no ensino superior: um estudo utilizando a mineração de dados como ferramenta de gestão do conhecimento em um banco de dados referente à graduação de engenharia*. Dissertação (Dissertação) — Universidade Federal do Rio de Janeiro, 2008.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introdução ao Data Mining*. Rio de Janeiro: Ciência Moderna, 2009.

WAIKATO, U. O. *WEKA*. 2010. Disponível em: ⟨https://ai.waikato.ac.nz/weka⟩. Acesso em: 08 jun 2020.

WITTEN, H. I.; FRANK, E. *Data mining: practical machine learning tools and techniques*. [S.l.]: Elsevier, 2005.