UNIVERSIDADE FEDERAL DO PARÁ

INSTITUTO DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

MINERAÇÃO DE DADOS EDUCACIONAIS APLICADA À BUSCA DE PERFIS DE ALUNOS EM CASOS DE EVASÃO OU RETENÇÃO: UMA ABORDAGEM ATRAVÉS DE REDES BAYESIANAS

DIEGO DA COSTA DO COUTO

DM 35/2017

UFPA / ITEC / PPGEE

Campus Universitário do Guamá
Belém-Pará-Brasil

2017

UNIVERSIDADE FEDERAL DO PARÁ

INSTITUTO DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

DIEGO DA COSTA DO COUTO

MINERAÇÃO DE DADOS EDUCACIONAIS APLICADA À BUSCA DE PERFIS DE ALUNOS EM CASOS DE EVASÃO OU RETENÇÃO: UMA ABORDAGEM ATRAVÉS DE REDES BAYESIANAS

DM 35/2017

UFPA / ITEC / PPGEE

Campus Universitário do Guamá
Belém-Pará-Brasil

2017

UNIVERSIDADE FEDERAL DO PARÁ

INSTITUTO DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

DIEGO DA COSTA DO COUTO

MINERAÇÃO DE DADOS EDUCACIONAIS APLICADA À BUSCA DE PERFIS DE ALUNOS EM CASOS DE EVASÃO OU RETENÇÃO: UMA ABORDAGEM ATRAVÉS DE REDES BAYESIANAS

Dissertação submetida à Banca Examinadora do Programa de Pós-Graduação em Engenharia Elétrica da UFPA para obtenção do Grau de Mestre em Engenharia Elétrica na área de Computação Aplicada.

Orientador: Prof. Dr. Ádamo Lima de Santana

UFPA / ITEC / PPGEE

Campus Universitário do Guamá

Belém-Pará-Brasil

2017

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

DO COUTO, DIEGO DA COSTA
MINERAÇÃO DE DADOS EDUCACIONAIS APLICADA À BUSCA DE PERFIS DE ALUNOS EM CASOS DE EVASÃO OU RETENÇÃO: UMA ABORDAGEM ATRAVÉS DE REDES BAYESIANAS / DIEGO DA COSTA DO COUTO. - 2017.
76 f. : il.

Dissertação (Mestrado) - Programa de Pós-graduação em Engenharia Elétrica (PPGEE), Instituto de Tecnologia, Universidade Federal do Pará, Belém, 2017.
Orientação: Prof. Dr. ÁDAMO LIMA DE SANTANA

1. DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS. 2. REDES BAYESIANAS. 3. CLASSIFICAÇÃO. I. DE SANTANA, ÁDAMO LIMA, orient. II. Título

CDD 006.312

DIEGO DA COSTA DO COUTO

MINERAÇÃO DE DADOS EDUCACIONAIS APLICADA À BUSCA DE PERFIS DE ALUNOS EM CASOS DE EVASÃO OU RETENÇÃO: UMA ABORDAGEM ATRAVÉS DE REDES BAYESIANAS

Dissertação submetida à Banca Examinadora do Programa de Pós-Graduação em Engenharia Elétrica da UFPA para obtenção do Grau de Mestre em Engenharia Elétrica na área de Computação Aplicada.

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil, 12 de setembro de 2017:

Prof. Dr. Ádamo Lima de Santana
Orientador

Prof. Dr. Francisco Carlos Bentes Frey Muller
(Avaliador Externo ao Programa – FCT/UFPA)

Prof. Dr. Marcelino Silva da Silva
(Avaliador Interno – PPGEE/UFPA)

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2017

Agradecimentos

Agradeço a Deus pela vida, por iluminar a minha mente, guiar meus passos e guardar-me do mal.

Sou grato até o último dia de vida aos meus pais, pelo amor, esforço e dedicação. O maior tesouro que me deram foi a educação, assim recursos financeiros escassos não me impediram de sonhar e conquistar muitas mudanças em nossas vidas.

Agradeço, em especial, à minha mãe que incontáveis vezes demonstrou na prática o valor da amizade e lealdade. As simples e profundas palavras de Drummond, eternizadas no poema Para Sempre, exprimem o sentimento que tenho cultivado ao longo da vida em relação à minha mãe.

À minha irmã, pessoa pela qual eu tenho muito respeito e admiração.

Ao meu orientador Prof. Dr. Ádamo Santana pela paciência e grande contribuição ao desenvolvimento desta pesquisa.

Aos colegas do CTIC que contribuíram para este trabalho, em especial Prof. Dr. Eloi Favero, Gustavo Lobato, Ernani Sales e Marco Aurélio Capela.

À toda a equipe do PPGEE, especialmente ao professor Pelaes e a Socorro, servidores competentes que sempre prestaram todo apoio a mim e aos demais alunos.

Finalmente, sou grato a Universidade Federal do Pará, instituição que proporcionou-me excelentes oportunidades, desempenhando um papel fundamental na minha formação profissional.

Sumário

Lista de ilustrações

Lista de tabelas

Resumo

Este trabalho investiga os perfis de alunos de cursos da graduação da Universidade Federal do Pará propensos a dois problemas enfrentados em diversas universidades brasileiras denominados evasão e retenção. Estas problemáticas estimularam o estudo de metodologias que detectassem padrões que suscitam a extrapolação ou o fim prematuro dos estudos. A ferramenta elegida a este fim, a Rede Bayesiana é poderosa ao propiciar raciocínio sobre incertezas, especialmente em diagnósticos de causas e efeitos tendo como pressuposto o relacionamento das variáveis e suas probabilidades de ocorrências conjuntas e marginais. Outro aspecto inerente a estrutura das Redes Bayesianas diz respeito à compreensibilidade da representação e dos resultados, os quais geram subsídios voltados a especialistas e usuários inseridos no domínio. Considerando tais colocações, essas potencialidades da metodologia em questão fortaleceram a sua aplicação nesta pesquisa. Dessa forma, registros acadêmicos contendo dezenas de milhares de amostras oriundas de alunos imersos em ambientes de ensino presencial pertencentes aos alunos de graduação ingressantes na Universidade Federal do Pará até o ano de 2016 foram submetidos ao processo de Descoberta de Conhecimento em Base de Dados, especificamente na etapa de Mineração de Dados os padrões desejados foram extraídos valendo-se da tarefa de classificação. Em adição, realizou-se na etapa de Mineração de Dados várias análises de desempenhos da Rede Bayesiana junto a outros algoritmos clássicos do aprendizado supervisionado, e aquela revelou a sua grande acurácia e eficiência, ressaindo dentre as melhores soluções encontradas, isto posto o seu uso foi certificado sobre a base de dados selecionada. Em três estudos de casos avaliados, os resultados indicaram a qualidade do classificador baseado em Redes Bayesianas que apresentou acurácia superior a 82%, condição que legitima a sua utilidade no domínio pesquisado. Assim, os resultados atingidos foram satisfatórios e apontaram fortes influências de algumas variáveis à propensão da evasão ou retenção.

Palavras-chaves: Descoberta de Conhecimento em Base de Dados. Redes Bayesianas. Classificação.

Abstract

This work investigates the profiles of undergraduate students at the University of Federal University of Pará prone to two problems faced in several universities evasion and retention. These problems stimulated the study of methodologies that detect patterns that lead to extrapolation or the premature end of the studies. The tool chosen for this purpose, the Bayesian Network is powerful in providing reasoning about uncertainties, especially in causes and effects diagnoses. Assumption of the relationship of the variables and their probability of occurrence and marginal. Another aspect inherent in the structure of Bayesian Networks is the comprehensibility of representation and results, which generate specialists and users entered into the domain. Considering such placements, these potential of the methodology in question strengthened its application in this research. So, academic records containing tens of thousands of samples from students immersed in presential teaching environments belonging to undergraduate students at the Federal University of Pará until the year 2016 were submitted to the of Knowledge Discovery in the Database, specifically in Data Mining the desired patterns were extracted using the classification task. In addition, several performance analyzes were performed during Data Mining stage The Bayesian Network together with other classic algorithms of supervised learning, and which revealed its great accuracy and efficiency, rising from the best solutions found, its use has been certified on the selected database. In three Study of Case, the results shows classifier's quality based on Bayesian Networks, which presented an accuracy of more than 82%, a condition that its usefulness in the researched domain. Thus, the results achieved were satisfactory and strong influences of some variables on the propensity of evasion or retention.

Keywords: Knowledge Discovery in Database. Bayesian Network. Classification.

1 Introdução

O ingresso no ensino superior representa à maioria dos estudantes brasileiros a concretização de um sonho. Toda a trajetória de um jovem do ensino médio até a matrícula na universidade envolve: dedicação aos estudos, investimento financeiro, concorrência e, em alguns casos, até reprovações. Ao ingressar na faculdade, o discente se depara com outra etapa desafiadora: cumprir todas as exigências curriculares do curso escolhido dentro de um prazo estabelecido. Neste sentido, a permanência do aluno até a sua formatura relaciona-se com um conjunto de fatores internos e externos ao ambiente acadêmico os quais são determinantes para o sucesso ou fracasso durante a graduação.

Neste novo desafio repleto de incertezas, muitos estudantes falham e abandonam os seus estudos, em muitos casos não se reconhecem os motivos que desencadeiam esta decisão, entretanto alguns autores consideram que isto não ocorre de maneira repentina. As respostas não são óbvias porém existem os meios que conduzem a elas, pois há evidências capazes de compô-las. O histórico escolar representa, em geral, o registro de dados considerados relevantes referentes ao desempenho acadêmico do discente nas diversas atividades curriculares. Então seria bastante sensato empregá-lo como fonte de provas necessárias para a solução do problema da evasão. Vale ressaltar que nem todos os dados tem relevância neste processo de análise, o que adiciona dificuldade aos fins pretendidos, em alguns casos outros dados subjacentes tornam-se imprescindíveis.

1.1 Justificativa e problema de pesquisa

A educação superior está em ascensão no Brasil. O Censo da Educação Superior revelou que até o ano de 2013 existiam cerca de 32.049 cursos de graduação em todo o país, distribuídos entre os graus bacharelado, licenciatura e tecnológico nas modalidades de ensino presencial e a distância (INEP, 2014b). Em 2014, houve um acréscimo de aproximadamente $2,5\%$ (32.878) no número dos cursos ofertados em 2.368 Instituições de Ensino Superior (IES). De acordo com a Tabela 1 as IES pertencentes à categoria privada apresentaram os maiores números em todos os anos, sendo que, em 2013, foram responsáveis por oferta de $66,1\%$ dos cursos. Entre os anos de 2010 e 2013 a evolução do quantitativo de cursos de graduação, de modo geral, corresponde percentualmente a $8,6\%$ (INEP, 2014b).

O gráfico da Figura 1.1 mostra que as faculdades representam $84,3\%$ das IES, contudo concentram apenas $29,2\%$ do total de matrículas. Por outro lado, são as universidades, que apesar de representarem um pequeno percentual das IES, $8,2\%$, concentram $53,4\%$ das

Tabela 1 – Evolução do Número de Cursos de Graduação, por Categoria Administrativa – Brasil – 2010-2013

Ano	Total	Categoria Administrativa				
		Total Pública	Federal	Estadual	Municipal	Privada
2010	29.507	9.245	5.326	3.286	633	20.262
2011	30.420	9.833	5.691	3.359	783	20.587
2012	31.866	10.905	5.978	3.679	1.248	20.961
2013	32.049	10.850	5.968	3.656	1.226	21.199

Fonte: (INEP, 2014b)

matrículas (INEP, 2014b; INEP, 2014a), sendo os cursos ofertados na modalidade presencial frequente em 90% das universidades. Vale destacar que entre 2003 e 2014, a matrícula na educação superior registrou aumento de 96,5% (INEP, 2014a). Essas constatações corroboram os avanços em termos quantitativos da educação superior no país nas iniciativas privada e estatal. Contudo, ressalta-se que gestores devem continuamente avaliar se quantidade se converteu em qualidade, ao estudante, à instituição e à sociedade.

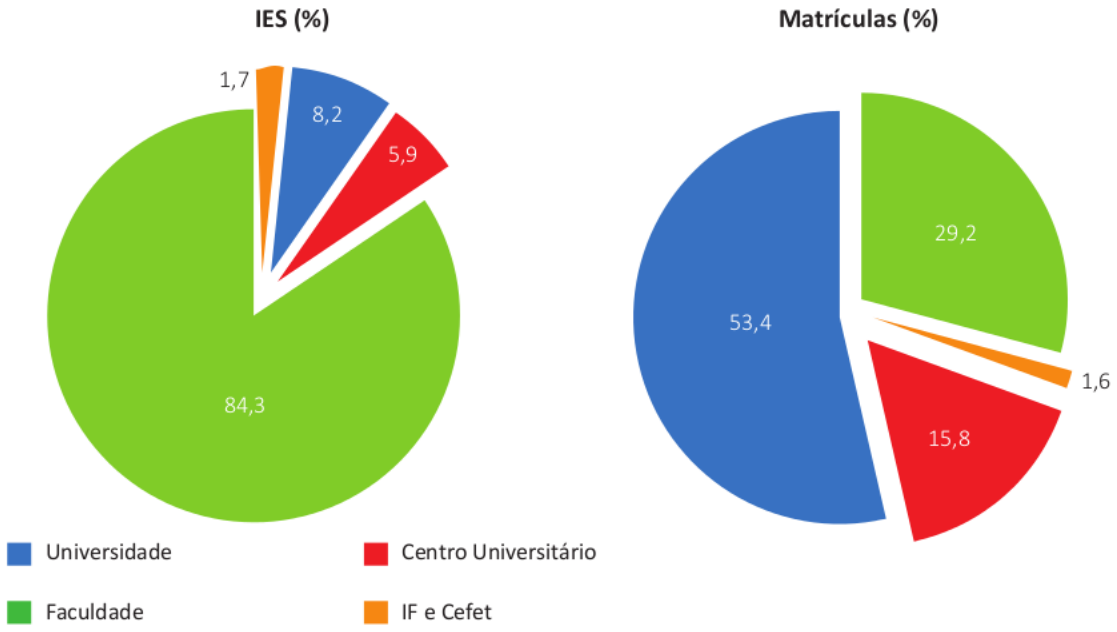

Figura 1.1 – Percentual do Número de Instituições de Educação Superior e Percentual do Número de Matrículas por Organização Acadêmica – Brasil – 2013

Fonte: (INEP, 2014b)

Os levantamentos realizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), formatados no Censo da Educação Superior, também apontam a existência de um descompasso entre os números de matrícula, ingressantes, cursos e concluintes. A Tabela 2 apresenta a evolução do número de concluintes de cursos de graduação, segundo a organização acadêmica. Constata-se queda no número de concluintes em todas as organizações acadêmicas, representando uma redução de 6,4% para as faculdades, 6,8% em

relação centros universitários, $4,4\%$ tendo em conta universidades e $26,0\%$ considerando os IFs e Cefets (INEP, 2014b). Estas informações denotam um importante diagnóstico: o aumento na quantidade de vagas não está impactando diretamente na permanência do aluno no ambiente acadêmico até a sua formatura. O resultado perceptível da evasão são as vagas ociosas ou remanescentes, as quais se destinam a outros processos de seleção. Acerca disso observou-se que apenas 17% dessas vagas foram ocupadas em 2014, embora a rede federal possua o maior percentual de preenchimento ($24,4\%$), mais de 86 mil dessas vagas não foram preenchidas nesses estabelecimentos de ensino (INEP, 2014a).

Tabela 2 – Evolução do Número de Concluintes de Cursos de Graduação, segundo a Organização Acadêmica – Brasil – 2010-2013

Organização Acadêmica	2010	2011	2012	2013
Brasil	973.839	1.016.713	1.050.413	991.010
Universidade	506.234	522.928	545.454	521.685
Centro Universitário	155.114	152.683	173.579	161.780
Faculdade	307.021	328.750	318.650	298.126
IF e Cefet	5.470	12.352	12.730	9.419

Fonte: (INEP, 2014b)

Neste contexto maior, a Universidade Federal do Pará (UFPA) está como uma Instituição Federal de Ensino Superior (IFES), *multicampi* que foi criada pela Lei nº 3.191, de 2 julho de 1957 (GOVERNO FEDERAL, 1957). Atualmente é considerada uma das maiores IFES do segmento do ensino, pesquisa e extensão em âmbito nacional. Os números institucionais demostram essa importância da instituição no cenário regional e brasileiro, por exemplo, no ano de 2014 havia aproximadamente 40.189 alunos matriculados, além de 4.598 diplomados e 9.631 ingressantes em 551 cursos no mesmo período (UFPA, 2015b).

Nos últimos 15 anos, a Universidade Federal do Pará e a sociedade conquistaram grandiosos instrumentos que a conduzem ao atingimento das metas institucionais (UFPA, 2015a), dentre os quais se destacam: criação de novos cursos de graduação; programa de Recuperação da Infraestrutura Física; crescimento e fortalecimento da Pós-Graduação *stricto sensu* e *lato sensu*; adesão ao Programa de Apoio a Planos de Reestruturação e expansão das Universidades Federais – REUNI; aprovação do Regulamento do Ensino de Graduação na UFPA; e expansão dos grupos de projetos de pesquisa e pesquisadores.

Contudo, o planejamento de ações que visem à excelência da UFPA, nos seus mais diversos aspectos, não é tarefa fácil (UFPA, 2015a) e dentro da problemática estudada esta instituição não está isenta. Algumas das razões que dificultam este processo são pontuadas no Plano de Desenvolvimento Institucional (PDI): situação geográfica e características peculiares da região na qual se situa; atuação *multicampi* em um Estado de grandes dimensões; processo ainda embrionário de uma cultura de planejamento; e ausência de modelos avançados de gestão e de uma cultura de avaliação e *feedback* (UFPA, 2015a).

No PDI, a UFPA possui, dentre vários outros, o objetivo de "Formar cidadãos capazes de transformar a realidade local" mensurado através do número de titulados na graduação e *stricto sensu*. Neste sentido, infere-se que as metas estipuladas serão atingidas se a instituição antes compreender as causas da evasão no ambiente acadêmico e, posteriormente, aplicar o conhecimento adquirido para implementar políticas que possam atenuar os seus efeitos. O último valor calculado (71,33) (UFPA, 2015b) do indicador Taxa de Sucesso na Graduação (TSG) que avalia o número de diplomados em relação ao total de ingressantes (MEC, 2004) revela que aproximadamente 30% dos estudantes não conseguem obter o diploma de graduação.

O Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais (REUNI) instituído pelo Decreto no 6.096, de 24 de abril de 2007 (GOVERNO FEDERAL, 2007) torna explícita em suas diretrizes gerais (MEC, 2007) a preocupação do Governo Federal acerca do problema da evasão, sob a asserção "*os índices de evasão de estudantes nos cursos de graduação atingem, em alguns casos, níveis alarmantes*". Dentre os focos do programa, destinados à resolver a problemática do insucesso na graduação, salienta-se a criação de condições para a ampliação do acesso e permanência na educação superior no nível de graduação, bem como a elevação gradual da taxa de conclusão média.

Outra medida adotada pelo Governo Federal brasileiro foi a Lei Nº 12.089 de novembro de 2009 (GOVERNO FEDERAL, 2009), a qual proíbe que uma pessoa ocupe duas vagas simultaneamente no curso de graduação em instituições públicas de ensino superior. Pode-se inferir que esta lei visa minimizar os casos nos quais discentes, por desinteresse ou vários outros motivos, abandonem um dos cursos ou demorem mais que o tempo normal para concluírem os estudos. Ambas as situações mencionadas não promovem benefícios para a sociedade e são absolutamente opostas às premissas do REUNI.

Os esforços do Ministério da Educação (MEC) face a problemática resultou na elaboração de um trabalho coletivo conduzido pela Comissão Especial de Estudos sobre Evasão para analisar um conjunto significativo de dados sobre o desempenho das universidades públicas brasileiras relativo aos índices de diplomação, retenção e evasão dos estudantes de seus cursos de graduação (MEC, 1997). O documento citado além de apresentar a evasão e sua complexidade, segmenta o seu significado em três vertentes, a saber:

1. *Evasão de curso*: quando o estudante desliga-se do curso superior em situações diversas tais como: abandono (deixa de matricular-se), desistência (oficial), transferência ou reopção (mudança de curso), exclusão por norma institucional;

2. *Evasão da instituição*: quando o estudante desliga-se da instituição na qual está matriculado;

3. *Evasão do sistema*: quanto o estudante abandona de forma definitiva ou temporária o ensino superior.

Segundo Silva Filho et al. (2007) a evasão estudantil no ensino superior, de modo geral, causa desperdícios de ordem social, acadêmica e econômica. Os reflexos deste entrave no setor público de ensino se manifesta quando os recursos são aplicados sem o devido retorno à sociedade. Enquanto que no ramo privado, os empresários perdem receitas e aumentam os gastos em manutenção. Em ambos os casos, a evasão implica em ociosidade de professores, funcionários, equipamento e espaço físico. A desistência do estudante tem consequências diretas no seu cotidiano, visto que este não consegue obter a qualificação necessária para atuar na área pretendida e, em outros casos, não retorna à IES em busca de novas oportunidades.

Quando os indivíduos que abandonaram o ensino superior são questionados acerca da motivação atrelada à evasão, algumas respostas são dadas, dentre as quais podemos destacar: i) dificuldades sócio-econômicas; ii) expectativas frustradas em relação ao curso; iii) rendimento acadêmico insuficiente; iv) problemas de relacionamento com colegas de classe ou funcionários da universidade; v) pressão exercida pela família para escolha de outro curso. Segundo o ponto de vista de (TINTO; TINTO, 1975, 1987 apud ANDRIOLA, 2009, p. 2) os argumentos supracitados estão totalmente válidos pois as causas da evasão emanam da falta de integração com o ambiente acadêmico e social da instituição, além disso os autores Barroso e Falcão (apud MANHÃES et al., 2011, p. 151) classificaram estes impedimentos sob três grupos, a saber: **vocacional**, **econômica** e **institucional**.

Alguns fatores associados à interrupção dos estudos por parte dos universitários são conhecidos (e alguns já foram elucidados nessa discussão) por gestores e estudantes, contudo há certas dúvidas que impedem a compreensão deste fenômeno em sua totalidade e, em geral, este entendimento é inerente à realidade de cada universidade. O principal questionamento a se fazer diz respeito a *identificação* de quais percalços tornam os estudantes vulneráveis a desistir do curso, o seu grau de *impacto* e *como* eles se configuram dentro do ambiente acadêmico.

Contudo, o tema ainda não é debatido como um problema real e que repercute no ensino superior. Neste sentido, Lobo (2011) argumenta que "*praticamente não existem estudos e politicas específicas sobre a Evasão no Ensino Superior*". Na concepção da autora esta ação deveria ser uma política do governo voltada a qualidade acadêmica e a aplicação responsável dos recursos, sejam eles públicos ou privados. Silva Filho et al. (2007) ratifica esta argumentação ao afirmar que "*são raríssimas as IES brasileiras que possuem um programa institucional profissionalizado de combate à evasão, com planejamento de ações, acompanhamento de resultados e coleta de experiências bem sucedidas*".

Afinal como deve-se avaliar a evasão? A resposta para este questionamento é fornecida por (SILVA FILHO et al., 2007):

> A evasão pode ser medida em uma instituição de ensino superior, em um curso, em uma área de conhecimento, em um período de oferta de cursos e em qualquer outro universo, desde que tenhamos acesso a dados e informações pertinentes. Em princípio, pode-se estudar a evasão no âmbito de uma IES, ou em

> um sistema, ou seja, um conjunto de instituições. (SILVA FILHO et al., 2007, pg. 644)

Diante disso, percebe-se que as IES, no atual cenário, possuem sistemas informatizados que apoiam a decisão dos gestores por meio do armazenamento de dados e geração de relatórios. Todavia, as informações consolidadas por estes sistemas são insuficientes para a extração de conhecimento referente ao domínio, com efeito, imprescindíveis para se criar vantagens competitivas para que a IES possa alcançar suas metas ou ainda detectar precocemente fatores negativos desencadeadores de fenômenos como a evasão ou a retenção dentro do ambiente acadêmico.

O desenvolvimento de pesquisas na utilização de ferramentas que auxiliem a extração de informações relevantes estão sendo realizadas por várias organizações (GOLDSCHMIDT; PASSOS, 2005). Estas ferramentas em conjunto com profissionais especializados são responsáveis por maximizar os casos de sucessos nas decisões gerenciais e de *marketing* nos mais variados domínios de aplicação (CARVALHO, 2001; HAN; KAMBER, 2001). A ferramenta encontrada para apresentar uma solução parcial para o problema discutido é denominada Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases* – KDD), juntamente com a técnica Mineração de Dados, cuja finalidade de ambas é a extração de informações escondidas em grandes bases de dados históricas (FAYYAD et al., 1996).

Não há dúvidas de que a evasão e a retenção são problemas sérios e com poucos estudos nacionais voltados para diagnosticar as causas e seus reflexos nas IES. Infere-se dos argumentos supracitados de Silva Filho et al. (2007) e (LOBO, 2011) que há condições favoráveis para mensurar a evasão na UFPA. Assim, busca-se nesta pesquisa aplicar e analisar o desempenho das técnicas de KDD sobre a base de dados do SIGAA, buscando revelar o perfil dos estudantes em situação de *evasão de curso* ou retenção e como estes atores se distinguem em relação aos diplomados.

1.2 Estado da arte de *Educational Data Mining* (EDM)

O sítio do Journal of Educational Data Mining (2016) define Mineração de Dados Educacionais (*Educational Data Mining* – EDM) como uma disciplina emergente cujo objetivo está no desenvolvimento de métodos para explorar os dados provenientes de cenários educacionais e essas metodologias são utilizadas para compreender os alunos nos seus ambientes de aprendizagem. Sachin e Vijay (2012) argumentam que há um aumento considerável no interesse por pesquisas valendo-se de EDM. Nesta perspectiva, Romero e Ventura (2010) elaboraram um trabalho sobre o estado da arte da Mineração de Dados Educacionais, no qual são discutidos 235 publicações mais relevantes até o ano de 2009, eles verificaram crescimento exponencial no número de publicações ao longo dos últimos anos, destacando o aparecimento de jornais e edições de livros especializados no assunto. Outras pesquisas importantes do tipo *survey* acerca

da temática e aplicações no domínio de EDM podem ser consultadas em (BAKER; YACEF, 2009) e Penã-Ayala (2013b), Penã-Ayala (2013a).

Baker e Yacef; Castro et al. (apud MANHÃES et al., 2012) garantem que o emprego de algoritmos de Mineração de Dados sobre dados educacionais consiste em um campo de investigação ainda não consolidado e necessita de investigações complementares. O artigo "Mineração de Dados Educacionais: Oportunidades para o Brasil"publicado em 2011 por Baker, Isotani e Carvalho (2011), evidencia que a área de EDM vem crescendo rapidamente em países da Europa e nos EUA, contudo pesquisas nacionais ainda estão em seu estágio inicial, visto que há poucas publicações de autoria ou participação brasileira.

Destas constatações, deduz-se que a comunidade científica está cada vez mais interessada neste campo emergente de pesquisa, o que o torna tendência promissora de investimentos e investigações no âmbito científico e acadêmico (BAKER, 2009). Esta argumentação corrobora a relevância do tema, já que neste trabalho busca-se explorar o potencial da área de pesquisa no domínio da educação superior, especificamente para apresentar uma solução parcial ao problema da evasão no domínio dos cursos de graduação da Universidade Federal do Pará.

Segundo Romero e Ventura (2007) a Mineração de Dados pode ser aplicada em dados oriundos de dois cenários de ensino distintos – tradicional ou presencial e a Educação à Distância (EaD). Em ambos, notam-se diferenças quanto a origem das amostras e os objetivos em aplicações das técnicas de descoberta de conhecimento em dados. O ensino presencial ou tradicional representa a modalidade de ensino mais difundida mundialmente, a sua peculiaridade reside no contato e a constante interação estabelecida entre educadores e estudantes por intermédio de aulas ocorridas em espaços físicos, como salas de aula, auditórios etc (ROMERO; VENTURA, 2007). Neste trabalho avaliamos indicadores acadêmicos registrados em histórico dos discentes os quais, em quase sua totalidade, estão inseridos em ambientes tradicionais de ensino e matriculados em disciplinas com cargas horária práticas (laboratório) ou teóricas.

1.3 Relevância do tema

A meta principal deste trabalho é a criação de um diagnóstico acerca dos fatores motivadores à evasão e retenção dos alunos da graduação da Universidade Federal do Pará em seus diversos *campi*. Aplica-se neste trabalho uma metodologia nova conhecida por *Educational Data Mining* (EDM), este fato impulsiona a realização de novas pesquisas e concede à base de dados corporativa mais que as funcionalidades primárias de armazenamento, recuperação de dados e síntese de relatórios, todavia, esta se torna uma fonte de extração de conhecimento até então inexplorada.

1.4 Trabalhos relacionados

Minaei-Bidgoli et al. (2003) desenvolveram uma pesquisa na *Michigan State University* (MSU), Estados Unidos. Coletaram-se 261 registros de *logs* do sistema *Learning Computer-Assisted Personalized Approach* (LON-CAPA) visando predizer o desempenho dos estudantes nas notas finais da disciplina de Física ofertada nesta plataforma *on-line*. Os algoritmos experimentados incluem *Quadratic Bayesian*, *1-nearest neighbor (1-NN)*, *k-nearest neighbor (k-NN)*, *Parzen-window*, Redes Neurais e Árvore de Decisão. Os autores aferiram as acurácias individuais e comparam à metodologia chamada combinação de múltiplos classificadores (*Combination of Multiple Classifiers* – CMC) com o auxílio de Algoritmos Genético (AG) no processo de otimização. A composição dos algoritmos demonstrou melhores resultados se comparado ao desempenho individualizado.

Os autores (KOTSIANTIS et al., 2003) examinaram o uso de técnicas de apredizagem de máquina (*Machine Learning* – ML) na predição da performance de alunos na disciplina "Introdução à Informática" de um curso da modalidade EAD da instituição de ensino grega *Helenic Open University*. Avaliaram-se 354 instâncias com atributos demográficos, frequência nas atividades e notas em avaliações. Os autores estudaram comparativamente 5 algoritmos das seguintes técnicas de ML: Árvore de Decisão (*Decision Tree* – DT), Redes Bayesianas (*Bayesian Nets*), *Perceptron-based Learning*, Aprendizado Baseado em Instâncias (*Instance-Based Learning*) e Aprendizado Baseado em Regras (*Rule-Based Learning*). Os pesquisadores concluíram que, estatisticamente, não havia diferença entre os classificadores, todavia o algoritmo *Naive Bayes* requereu menor tempo de processamento para construção do modelo e obteve melhor acurácia. Estas características deste classificador o posicionam como uma possível escolha na modelagem e desenvolvimento de ferramentas voltadas à EDM.

Al-Radaideh, Al-Shawakfa e Al-Najjar (2006) coletaram dados através de questionários referentes a estudantes de graduação dos cursos de Ciência da Computação e Tecnologia da *Yarmouk University*, Jordânia. Entre os atributos selecionados estão o gênero, idade, departamento, etc. Os autores testaram as técnicas árvore de decisão (algoritmos ID3 e C4.5) e o algoritmo *Naive* Bayes para classificação de quatro possíveis conceitos na disciplina Programação I (linguagem C++). As acurácias de todos os classificadores apresentados no trabalho não foram suficientemente altos. A melhor taxa de acerto (38.4615%) foi obtida pela árvore de decisão ID3. Segundo os pesquisadores tanto amostras quanto os atributos não foram relevantes à geração dos modelos.

Hämäläinen e Vinni (2006) originaram um trabalho no Departamento de Ciência da Computação da *University of Joensuu*, Finlândia. Os autores coletaram registros das notas finais e resultados de exercícios por dois anos letivos, concernentes a duas disciplinas de programação. Os dados de 125 alunos são referentes a primeira disciplina (*Prog. 1*) enquanto que 88 estudantes constituem a outra (*Prog. 2*). Experimentaram-se os seguintes algoritmos classificadores: regressão linear (*Linear Regression*), *Naive Bayes*, redes bayesianas (*Bayesian networks*)

e *Bayesian mutinets*. Os pesquisadores verificaram que o classificador *Naive Bayes* foi melhor em prever os alunos potencialmente propensos a desistir das disciplinas. Todos os métodos conseguiram acurácias superiores a 80%.

Superby, Vandamme e Meskens (2006) avaliaram o desempenho acadêmico de 533 estudantes pertecentes a três universidades da comunidade francesa na Bélgica (*French Community of Belgium*). A investigação destinava-se a identificar no primeiro ano de curso a propensão do aluno abandonar os estudos, sob três categorias de risco: baixo (*low risk*), médio (*medium risk*) e alto (*high risk*). Os dados foram coletados por meio de questionários aplicados entre os anos de 2003 e 2004. Os classificadores aplicados na pesquisa foram DT, Redes Neurais Artificiais (*Neural Network*) e *Random Forest*. Os autores argumentaram que os resultados não foram satisfatórios em decorrência da heterogeneidade na fonte das amostras, isto é, existem grandes disparidades entre as três instituições belgas.

O artigo (NGHE; JANECEK; HADDAWY, 2007) comparou as técnicas árvore de decisão e redes bayesianas buscando predizer o desempenho acadêmico de estudantes da graduação e da pós-graduação de duas universidades bastantes diferentes quanto a sua estrutura: *Can Tho University* (CTU), uma grande universidade do Vietnã; e a *Asian Institute of Technology* (AIT), um pequeno instituto internacional de pós-graduação da Tailândia. Coletaram-se 20.492 registros dos discentes que ingressaram entre 1995 a 2002 na universidade vietnamita. Para o instituo tailandês, foram coletados 936 registros dos ingressantes nos anos de 2003 a 2005. Os autores investigaram a utilização de três *softwares* destinados à mineração de dados, a saber: Weka, Orange e Yale. A primeira ferramenta citada se mostrou mais apropriada em três requisitos: maior número de algoritmos disponíveis, suporta grandes *data sets* e por deter mais funcionalidades destinada s ao pré-processamento. O trabalho supracitado apresentou 4 classes, e para cada uma delas obteve precisão superior 70%, sendo mais perceptível ao se valer de apenas duas classes, onde se atingiu $94,03\%$ para a base proveniente da CTU. Os dois estudos de caso apresentados revelaram a superioridade quanto à acurácia da árvore de decisão se comparada à redes bayesianas.

Os autores de (CORTEZ; SILVA, 2008) obtiveram dados no período letivo de 2005 e 2006 de escolas públicas de Portugal. Os atributos constituíram-se de registros coletados de relatórios emitidos pelo sistema escolar e questionários com perguntas sobre aspectos sociais, demográficos e emocionais dos estudantes. A finalidade dos autores era predizer o desempenho escolar nas disciplinas básicas de Matemática e Português. Os autores trabalharam com três metodologias: classificação binária, em cinco níveis e regressão. Os algoritmos testados foram Árvore de Decisão, *Random Forest*, RNA e *Support Vector Machine*. Os resultados atingidos foram satisfatórios, a exemplo, para o teste com classes binárias a DT conseguiu a melhor taxa de acerto ($93,0\%$). Os autores descobriram ainda importantes regras as quais associam as notas aos demais atributos.

Romero et al. (2008) estudaram o emprego de diferentes técnicas de ML voltadas à aná-

lise do desempenho de estudantes associado à algumas atividades efetuadas em um dos 7 cursos ofertados na plataforma *web* Moodle (MOODLE, 2016). Os dados são pertencentes a 438 alunos da *Cordoba University*, Espanha. Os autores desenvolveram uma ferramenta integrada ao Moodle cujo objetivo é auxiliar tutores no *feedback* concernente ao aprendizagem do estudante inserido neste ambiente. Eles propuseram alguns cenários com dados numéricos ou categóricos. Os algoritmos CART e C.45, incluídos na técnica de árvore de decisão, apresentaram os melhores resultados se aplicados sobre *data set* cujos atributos estão no formato discretizado.

Os estudos conduzidos por Dekker, Pechenizkiy e Vleeshouwers (2009) analisaram no departamento de Engenharia Elétrica da *Eindhoven University*, Holanda, os dados referentes a 648 discentes entre os anos de 2000 a 2009. O objetivo dos autores consistia em aplicar técnicas de mineração de dados para identificar alunos desistentes ou reprovações ainda no primeiro ano do curso de Engenharia Elétrica. O *data set* foi composto por dados pré-universitários, tais como: notas nas disciplinas Matemática e Ciências e o curso feito anteriormente. Após o ingresso dos alunos foram registrados as notas de três exames. Os autores testaram os algoritmos Árvore de Decisão, *Bayesian Nets* e Regras de Associação, todos implementados no *software* Weka. De acordo com o estudo o classificador baseado em árvore de decisão apresentou melhor desempenho em relação aos demais.

A pesquisa elaborada por Delen (2010) coletou dados relativos às características acadêmicas, financeiras e demográficas de 16066 estudantes ingressantes nos anos entre 2004 e 2008. O objetivo do trabalho consistia em aplicar EDM para analisar a propensão do discente cursar o segundo ano de graduação após o ingresso. O número de variáveis para estudo consistiram em 39 e um atributo que armazenava a classe (binária). Foram aplicados quatro métodos de classificação: Redes Neurais Artificiais (*Artificial Neural Network* – ANN), Árvores de Decisão (*Decision Tree* – DT), *Support Vector Machines* (SVM) e *Logistic Regression*. Comparam-se estes classificadores a três outras técnicas denominadas *ensembles*[1]: *Bagging* (*Random Forest*), *Busting* (*Boosted Tree*) e *Information Fusion*. Os resultados apontaram o melhor desempenho de SVM ($87,23\%$), seguida de DT ($87,16\%$), ANN (86.45%) e *Logistic Regression* (86.12%). Contudo, no cenário evidenciado, a base de dados estava desbalanceada, isto é, havia proporção de 80% de uma classe e 20% da outra. Assim, para atenuar os efeitos do fenômeno conhecido por *overfit*, que em geral se revela nesse tipo de configuração, os dados foram distribuídos uniformemente para ambas as classes. Os resultados, para o estudo citado, novamente ratificaram a precisão global do *Support Vector Machine* (81.18%). De acordo com o estudo, a acurácia das três técnicas *ensembles* não foram, individualmente, superiores aos classificadores elencados.

Kovačić (2010) inspecionou o uso de árvore de decisão para predizer o sucesso de 450 estudantes do curso de Sistema de Informação da *Open Polytechnic*, localizada na Nova Zelândia. As variáveis coletadas foram sócio-demográficas e relacionadas ao ambiente de ensino. O pesquisador aplicou quatro algoritmos chamados: CHAID, *exhaustive* CHAID, QUEST e

[1] Optou-se por manter em inglês este e alguns outros termos empregados na literatura.

CART. Conclui-se que todos os métodos aplicados não obtiveram acurácias satisfatoriamente altas, dessa forma a melhor performance foi conseguida pela variante CART ($60,5\%$). Kovačić (2010) explica que a entrada de certos dados (gênero, idade, etnia, escola de origem e profissão[2]) no processo de inscrição não contém informações suficientes à classificação, isto provavelmente está atrelado a falta de qualidade das técnicas aplicadas.

Alkhasawneh e Hobson (2011) construíram dois modelos distintos de Redes Neurais Artificiais (RNA) usando *feed-foward backpropagation*. A primeira configuração destinava-se a prever se o aluno prosseguiria no curso. O outro modelo aplicava-se a classificação em uma das três categorias (*at-risk*, *intermediate* e *advanced*). Os autores coletaram 338 amostras pertencentes a estudantes das áreas de Ciência, Tecnologia, Engenharia e Matemática da *Virginia Commonwealth University*, localizada em Richmond, Estados Unidos. No primeiro cenário, a melhor taxa de acerto foi $68,9\%$ (Engenharia), enquanto que o segundo apresentou acurácia em torno de $70,1\%$.

Manhães et al. (2011) avaliaram a aplicação de 10 algoritmos classificadores à base de dados coletada do sistema acadêmico da Universidade Federal do Rio de Janeiro – UFRJ, a qual contém informações sobre os alunos de graduação que ingressaram no curso de Engenharia Civil da Escola Politécnica no período de 1994 a 2005. Foram analisadas 887 amostras, divididas em duas classes, destas 543 é composta por informações de alunos que concluíram o curso e 344 registros de alunos que não concluíram. Identificaram-se cinco disciplinas mais cursadas relativas ao primeiro semestre do curso. Realizaram-se três experimentos com o objetivo de comparar o desempenho dos algoritmos de mineração de dados. Os resultados alcançados mostraram que a acurácia varia em média em torno de 75 a 80%, além desta métrica a taxa de erro dos classificadores foi considerada relevante, assim, a previsão incorreta de discentes com risco de evasão é considerada erro grave do classificador.

Nos trabalhos (PANDEY; PAL, 2011a; PANDEY; PAL, 2011b) aplicaram-se o algoritmo *Naive Bayes* para criar um modelo probabilístico destinado a inferência do desempenho acadêmico em cinco classes. A base foi composta por dados de 300 alunos de cinco cursos da instituição de ensino chamada *Dr. R. M. L. Awadh University*, na Índia. Os atributos coletados referem-se à dados demográficos, acadêmicos e sócio-econômico, os quais foram obtidos por meio de questionários e do próprio sistema acadêmico. As variáveis potencialmente influentes ou preditores, com as suas probabilidades, na classificação foram: nota no ensino secundário ($0,8642$), logradouro ($0,7862$) e proficiência do aluno em um ou mais idiomas ($0,7225$).

As investigações em EDM se consolidaram ainda mais no país em 2012, na ocasião, MANHÃES et al. elaboraram um estudo de caso para avaliar a evasão em 155 cursos de graduação ofertados por 28 unidades da UFRJ. Para a pesquisa em discussão, foram selecionados dados acadêmicos dos discentes que ingressaram nos dois semestres letivos dos anos de 2003 e 2004. Dois aspectos merecem destaque neste trabalho: o primeiro fato refere-se ao total de amostras

[2] *work status*

disponíveis; e, por conseguinte, o número de classes tratadas, já que muitos trabalhos na literatura tratam apenas duas categorias. A acurácia dos classificadores e a interpretabilidade dos seus resultados foram dois requisitos considerados na escolha do método apropriado para solucionar a problemática. Neste contexto, o classificador *Naive Bayes* foi elegido pois configurou-se aos propósitos da pesquisa, pois conseguiu atingir precisão global superior a 80%. As contribuições dessa pesquisa também foram publicadas em (MANHÃES et al., 2014b). Nos trabalhos (MANHÃES et al., 2014a; MANHÃES et al., 2015), analisaram-se somente cursos pertencentes as áreas de ciência, tecnologia, engenharia e matemática.

Cheewaprakobkit (2013) pesquisou a respeito dos fatores relativos ao desempenho acadêmico. Foram empregados 1600 registros com 22 atributos, coletados no período de 2001 a 2011, de alunos participantes de um programa internacional desenvolvido por uma universidade particular da Tailândia. Os pesquisadores aplicaram o uso das técnicas Árvore de Decisão e Redes Neurais através da validação cruzada com 10 conjuntos (*cross-validation with 10 folds*). A comparação experimental entre os modelos construídos indicou alta taxa de classificação correta para DT (85,188%), a qual é ligeiramente maior que a RNA por 1,313%. Dentro dos artigos elencados, o trabalho discutido possuiu uma das melhores acurácias, isto revela o valor da relação existente entre seleção dos dados, pré-processamento e o uso da técnica adequada ao domínio de aplicação.

Kabakchieva (2013) examinou o uso de alguns algoritmos disponíveis na ferramenta Weka com a finalidade de classificar corretamente o desempenho de 10330 estudantes da *University of National and World Economy* (UNWE), na Bulgária. Dados acadêmicos e pré-universitários dos alunos foram utilizados na predição de 5 notas (classes). As técnicas adotadas pela autora foram árvore de decisão (*J48*), classificadores *bayesianos* (*Naive Bayes* e *BayesNet*), *Nearest Neighbour* e aprendizagem por regras (*One Rip* e *JRip*). Os resultados conseguidos mostraram que a técnica de árvore de decisão obteve maior precisão global, seguida pelo método *JRip* e *Nearest Neighbour*. Em todos os testes os classificadores *bayesianos* não obtiveram acurácias satisfatórias (abaixo de 70%), isto implica em alta taxa de erro e pouca confiabilidade na classificação.

Recentemente, em meados de 2014, (MANHÃES et al., 2014c) propuseram uma arquitetura em três camadas (dados, aplicação e apresentação) denominada por EDM WAVE a qual utiliza alguns classificadores implementados na API da ferramenta WEKA. No trabalho citado, aplica-se cada algoritmo aos dados acadêmicos dos discentes da graduação e avalia-se a predição destes pertencerem a uma das três classes estabelecidas no estudo. Segundo os autores esta arquitetura foi integrada ao sistema legado que suporta as bases de registros acadêmicos da UFRJ. Contudo, é importante destacar que não se apresentou o funcionamento de um *software* desenvolvido a partir da arquitetura, com efeito esta lacuna não invalida os excelentes resultados alcançados pelos autores, visto que pelo nosso entendimento os objetivos da tese de doutorado foram atingidos.

As investigações desenvolvidas por Mayilvaganan e Kalpanadevi (2014) comparam algumas técnicas de classificação aplicadas para previsão da capacidade de aprendizagem, categorizada em 4 tipos, de 197 estudantes pertencentes aos cursos das áreas de Artes, Ciência da Computação, Comércio e Engenharia de Sofware da *PSG College of Arts and Science College* localizada em Coimbatore, Índia. Alguns atributos selecionados são: idade, sexo, especialidade, horas dedicadas ao estudo, acesso à recursos como internet e biblioteca, as notas obtidas nas atividades desenvolvidas, entre outros. Os autores analisaram os algoritmos: árvore de decisão; *Aggregating One-Dependence Estimators* (AODE); *Naive Bayes*; e *Multi Label K-Nearest Neighbor*. Segundo eles o método *Multi Label K-Nearest Neighbor* apresentou melhor acurácia se confrontado às outras técnicas elencadas.

Os trabalhos relacionados ao projeto estão dispostos na Tabela 3.

Tabela 3 – Trabalhos relacionados elencados para a execução do projeto

Trabalho	**Origem**	**Número de algoritmos testados**	**Número de Amostras**	**Análise em disciplinas**	**Análise em cursos**
(MINAEI-BIDGOLI et al., 2003)	Estados Unidos	6	261	X	
(KOTSIANTIS et al., 2003)	Grécia	5	354	X	
(AL-RADAIDEH; AL-SHAWAKFA; AL-NAJJAR, 2006)	Jordânia	3	Não informado	X	
(HäMäLäINEN; VINNI, 2006)	Finlândia	4	213	X	
(SUPERBY; VANDAMME; MESKENS, 2006)	Bélgica	5	533		X
(NGHE; JANECEK; HADDAWY, 2007)	Vietnã	3	**20492**		X
(ROMERO et al., 2008)	Espanha	25	438		X
(DEKKER; PECHENIZKIY; VLEESHOUWERS, 2009)	Holanda	8	648		X
(DELEN, 2010)	Estados Unidos	5	16066		X
(KOVAČIĆ, 2010)	Nova Zelândia	4	450		X
(ALKHASAWNEH; HOBSON, 2011)	Estados Unidos	1	338		X
(MANHÃES et al., 2011)	Brasil	10	887		X
(MANHÃES et al., 2012)	Brasil	6	14024		X
(MANHÃES et al., 2014b)	Brasil	1	14237		X
(MANHÃES et al., 2014a)	Brasil	5	14237		X
(MANHÃES et al., 2015)	Brasil	5	1359		X

Trabalho	**Origem**	**Número de algoritmos testados**	**Número de Amostras**	**Análise em disciplinas**	**Análise em cursos**
(PANDEY; PAL, 2011a; PANDEY; PAL, 2011b)	Índia	1	300		X
(CHEEWAPRAKOBKIT, 2013)	Tailândia	2	1600		X
(KABAKCHIEVA, 2013)	Bulgária	6	10330		X
(MAYILVAGANAN; KALPANADEVI, 2014)	Índia	4	197		X

Portanto, o trabalho proposto compartilha algumas características com aqueles citados e discutidos anteriormente. Contudo, possui alguns pontos que merecem relevância dentro da pesquisa:

a) Supera os trabalhos relacionados em número de amostras, pois a maioria dos trabalhos usam *data sets* com poucas centenas até alguns milhares;

b) Consegue aplicar grande volume de amostras a uma quantidade considerável de classificadores;

c) Excede os trabalhos relacionados em número de cursos;

d) Os trabalhos correlatos não dispunham de métricas capazes de relacionar o desempenho acadêmico dos alunos àqueles formados em determinado período. Nesta dissertação, apresentam-se atributos que correlacionam, por probabilidades, a performance acadêmica do aluno em relação aos formados;

e) Investigação em uma grande universidade pública federal com abrangência na região Norte e Nordeste;

f) Fortalecimento do campo de EDM, uma vez que esta área é nova e não há muitos estudos nacionais;

g) Diagnóstico inédito dentro da Universidade Federal do Pará acerca da evasão valendo-se da Mineração de Dados.

1.5 Contribuições

A contribuição para UFPA consiste na criação de um diagnóstico acerca da evasão nos *campi* da capital e do interior. Esta ação favorecerá, a longo prazo, na adoção e vinculação do processo de EDM junto ao SIGAA como um procedimento auxiliar no diagnóstico e avaliação acerca da evasão. A consolidação da proposta deste trabalho baseia-se em apresentar uma resposta inicial sobre a problemática analisada no contexto local, à vista disso espera-se que estudos posteriores sugiram a modelagem e desenvolvimento de *softwares*, ou aprimoramento

daqueles existentes, com foco nos requisitos: seleção de atributos, transformação de dados, visualização, mineração de dados e representação do conhecimento.

A temática é complexa, além disso poucos pesquisadores empenham grandes esforços na investigação do fenômeno. Os trabalhos existentes em geral aplicam métodos tradicionais como: revisões de relatórios, estudo teórico ou restringem-se à analise estatística dos dados. Assim, no Brasil nota-se ainda poucas pesquisas no campo emergente de EDM, desta forma busca-se novos resultados propícios de revelar um panorama da problemática em uma IFES de grande porte, com destaque na região Norte e Nordeste e importante papel no ensino, pesquisa e extensão do país.

1.6 Objetivos

O objetivo geral desta dissertação visa analisar a evasão e a retenção de estudantes na Universidade Federal do Pará utilizando Redes Bayesianas, visto que é uma estrutura de dados eficiente na representação quantitativa (probabilidades) e qualitativa (gráfica) sobre a modelagem de cenários de incertezas dentro de um domínio de aplicação. Diante disso, a UFPA suporta uma vasta base de dados do Sistema Integrado de Gestão Acadêmica (SIGAA) que recebe anualmente milhares de estudantes, todavia a instituição não dispõe de mecanismos modernos e eficazes, os quais consigam identificar o perfil de alunos propensos ao abandono ou à prescrição de acordo com o Regimento da Graduação. Assim, os gestores poderão incorporar este novo conceito aos seus instrumentos de avaliação institucional, a fim de implementar políticas, embasadas em diagnóstico de combate efetivo à problemática enfrentada.

A realização do objetivo geral está associada com os objetivos específicos enumerados a seguir:

a) Introduzir a temática da evasão e retenção no âmbito de outras IES;

b) Entender a modelagem da base de dados corporativa para identificar quais os dados estavam disponíveis e são mais relevantes à pesquisa;

c) Projetar um banco de dados relacional de dados auxiliar que contemple os atributos elegidos à pesquisa;

d) Aplicar à base de dados ferramentas de Mineração de Dados para extração de padrões, buscando obter conhecimento útil acerca da evasão e consequentemente apoiar decisões no âmbito acadêmico;

1.7 Metodologia

Os aspectos metodológicos desta pesquisa estão relacionados aos seguintes itens:

a) Estudo preliminar dos conhecimentos envolvidos: técnicas de Mineração de Dados, noções de probabilidade e estatística, modelos de dados do banco e, em alguns casos, técnicas avançadas em SQL;

b) Formular um cenário, para não generalizar e consequentemente gerar resultados equivocados;

c) Solicitar os dados relevantes para a pesquisa (sem identificação de alunos ou dados pessoais) ao setor acadêmico da instituição;

d) Selecionar os dados obtidos com foco na sua representatividade para a pesquisa, removendo inconsistências; e

e) Aplicar as técnicas de Mineração de Dados e extrair os conhecimentos importantes para apresentar repostas ao problema identificado.

1.8 Organização da Dissertação

Este trabalho está organizado da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica relativa à inteligência computacional e classificação. O Capítulo 3 trata dos conceitos de probabilidades e redes bayesianas bem como suas particularidades. Por sua vez, no Capítulo 4 é apresentada a modelagem da aplicação proposta. Em seguida, Capítulo 5 evidenciam-se os resultados de três estudos de casos propostos nesta pesquisa. Por fim, no Capítulo 6 são debatidas as conclusões, contribuição, dificuldades e propostas de trabalhos futuros.

2 Fundamentação Teórica

2.1 Considerações Iniciais

Neste capítulo são tratados questões relativas à inteligência artificial, aprendizado de máquina supervisionado e as fases da mineração de dados dentro do processo de descoberta de conhecimento. Além disso, são apresentadas as métricas de qualidade de um classificador juntamente com alguns conceitos fundamentais da classificação.

2.2 Inteligência Artificial

Em 1950 Alan Turing (TURING, 1950) propôs o teste que ganhou o próprio nome e foi projetado a fim de garantir uma definição operacional e adequada de inteligência (RUSSEL; NORVING, 2010). A proposta consistia em submeter o computador a um teste baseado na impossibilidade de distingui-lo de um ser humano. Uma pessoa (interrogador) formula algumas perguntas e um computador será aprovado caso aquele não consiga diferençar se as respostas foram informadas por um ser humano ou não. Russel e Norving (2010) evidenciam quatro principais capacidades indispensáveis à aprovação no teste:

1. Processamento da linguagem natural (*natural language processing*);
2. Representação de conhecimento (*knowledge representation*);
3. Raciocínio automatizado (*automated reasoning*);
4. Aprendizado de máquina (*machine learning*).

Acerca desta discussão Rich e Knight (1993) asseguram que:

> Inteligência Artificial (IA) é o estudo de como fazer os computadores realizarem coisas que, no momento, as pessoas fazem melhor". É claro que esta definição é um tanto quanto efêmera por causa de sua referência ao estado atual da ciência da computação. E ela não consegue incluir áreas de impacto potencialmente grande, a saber, problemas que não podem presentemente ser solucionados muito bem nem pelos computadores nem pelas pessoas. (RICH; KNIGHT, 1993, p.3)

A lacuna da qual (RICH; KNIGHT, 1993) afirmam refere-se aos problemas que tanto seres humanos quanto computadores não resolvem, ou se solucionam, não o fazem da melhor

forma, embora os autores ratificam os propósitos quanto à inteligência referente à máquina introduzidos por Turing. Os autores Russel e Norving (2010) chamam a esta capacidade de realizar tarefas de maneira eficiente de *racionalidade*. Russel e Norving (2004) defendem a existência de uma enorme variedade de áreas de uso geral, como aprendizado e percepção, além de tarefas específicas como jogos de xadrez, demonstração de teoremas matemáticos, criação de poesia e diagnóstico de doenças.

Isto posto, sustensa-se a ideia que Inteligência Artificial (IA) é uma ciência relativamente nova e com numerosas aplicabilidades com o intuito de subsidiar a resolução de atividades intelectuais com utilização de mecanismos automatizados. Neste trabalho será empregado esse potencial da IA visando alcançar os objetivos estipulados por intermédio do Aprendizado de Máquina (*Machine Learning*), da Descoberta de Conhecimento em Base de Dados (*Knwoledge Discovery Database*) e da Mineração de Dados (*Data Mining*) os quais são ramos advindos da Ciência da Computação.

2.3 Hierarquia do Aprendizado de Máquina

Rezende (2005, p 90) define indução como "a forma de inferência lógica que permite obter conclusões genéricas sobre um conjunto particular de exemplos". O raciocínio se consolida da parte para o todo, isto é, um conceito específico é generalizado a partir de exemplos apresentados ao processo de inferência indutiva. Destaca-se que a seleção dessas instâncias é crucial para que as hipóteses construídas preservem a verdade e ofereça resultados de qualidade. O aprendizado indutivo está subdivido em duas categorias: **supervisionado** e **não-supervisionado**. Na primeira, é informado ao algoritmo de aprendizado (indutor), um conjunto de exemplos ou instâncias para os quais se conhece o rótulo da classe. Na segunda vertente, por sua vez, o indutor analisa os exemplos e tenta agrupá-los por similaridade entre os seus atributos, formando os *clusters* ou agrupamentos.

O aprendizado supervisionado se caracteriza pela presença de um algoritmo de indução ou classificador capaz de determinar corretamente a classe ou conceito de novos exemplos ainda não rotulados, os quais formam a base de testes. Rezende (2005, p 91, grifo do autor) faz uma importante distinção de conceitos ao declarar que "para rótulos de classe discretos, esse problema é conhecido como **classificação** e para os valores contínuos como **regressão**". A Figura 2.1 mostra esquematicamente a hierarquia do aprendizado indutivo, os elementos na cor azul são de interesse a este trabalho.

2.4 Aprendizado supervisionado

Segundo Rich e Knight (1993, p. 526, grifo do autor) "*classificação* é o processo de atribuir, a uma determinada informação recebida, o nome de uma classe à qual ela pertence". Este

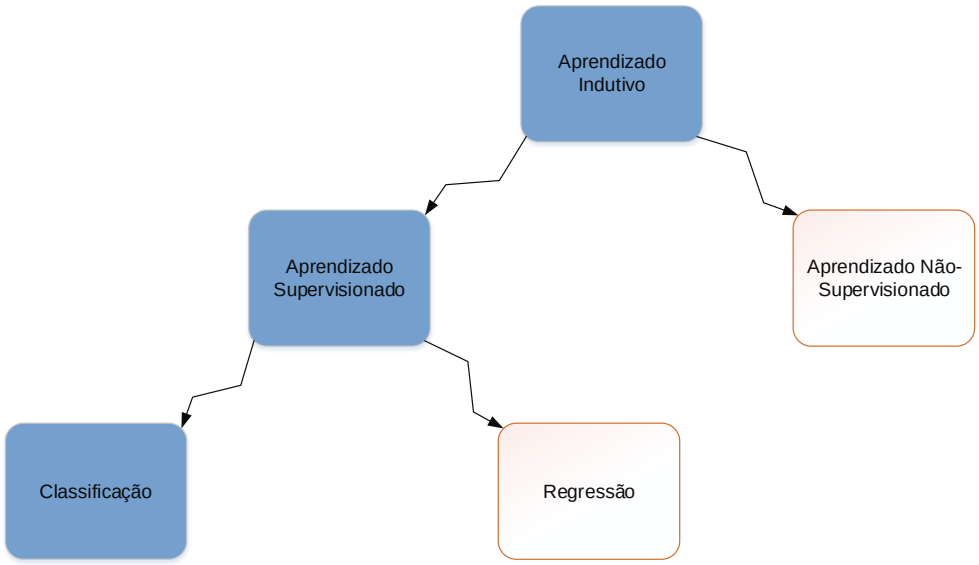

Figura 2.1 – Hierarquia do aprendizado indutivo

Fonte: (REZENDE, 2005)

processo é compreendido por duas etapas: **aprendizagem** (*learning*) e a **classificação** (*classification*) (HAN; KAMBER, 2006). Na primeira etapa, exemplificado na Figura 2.2, um algoritmo de classificação constrói um classificador a partir de um conjunto de treinamento oriundo das tuplas[1] de uma base de dados, assim são criadas as regras de classificação, as quais representam os conhecimentos extraídos dessa base. A segunda etapa se efetiva após a construção do modelo ou das regras, deste modo novos exemplos submetidos ao classificador são rotulados sob uma das possíveis classes disponíveis, a sistemática está apresentada na Figura 2.3.

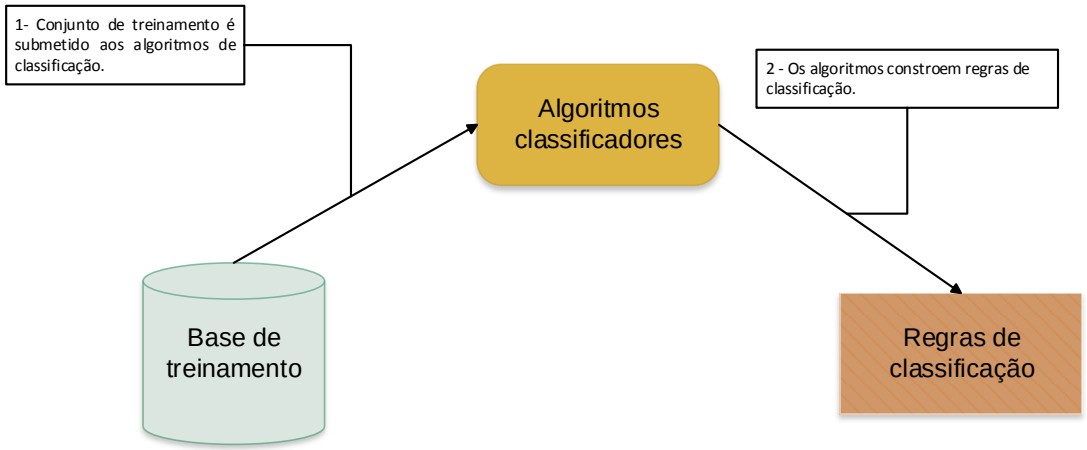

Figura 2.2 – Etapa de aprendizado

Fonte: Adaptado de (HAN; KAMBER, 2006)

[1]Outras terminologias podem ser utilizadas, tais como: amostras (*samples*), exemplos (*examples*), instâncias (*instances*) e objetos (*objects*). Assim, doravante, qualquer uma delas será empregada ao longo deste trabalho

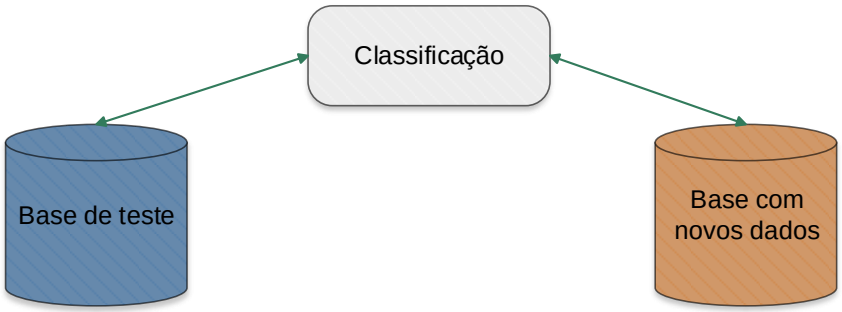

Figura 2.3 – Etapa de classificação

Fonte: Adaptado de (HAN; KAMBER, 2006)

A representação de uma tupla X é um vetor de atributos n-dimensional, tal que $X = (x_1, x_2, \ldots, x_n)$ e uma instância corresponde a um atributo conhecido por rótulo de classe (HAN; KAMBER, 2006). Russel e Norving (2010) formalizam também o aprendizado supervisionado como um conjunto de N pares ordenados $(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)$, onde x_i é um exemplo da base de dados e y_i corresponde a uma função $f(x)$ desconhecida. A inferência indutiva consiste em obter uma função h, para a coleção de exemplos, que se aproxime de f . A função h é chamada de hipótese (*hypothesis*). A descoberta de h embasa-se em um processo de busca dentro do espaço de hipóteses H.

2.5 Descoberta de Conhecimento em Base de Dados – *Knowledge Discovery Database*

Fayyad (1997) evidencia a distinção entre dois termos tratados como sinônimos, pois KDD consiste em um processo completo voltado para descobrir informações potencialmente úteis sobre uma determinada base de dados, enquanto Mineração de Dados (*data mining*) refere-se a uma etapa particular dentro deste processo. Segundo Frawley, Piatetsky-Shapiro e Matheus (1992, p. 58, tradução nossa) o termo *Knowledge Discovery in Databases (KDD)* remete a um "processo não-trivial de extração de informações ímplicitas, previamente desconhecidas e potencialmente úteis a partir dos dados". Mineração de Dados (*Data Mining*), por sua vez, é uma das etapas do KDD (FAYYAD et al., 1996) no qual há extração de padrões por meio do uso de algoritmos específicos, esta etapa pode ser apoiada por basicamente quatro áreas transversais: Apredizado de Máquina, banco de dados, estatística e visualização (FERRO; LEE, 2001).

Assim, de acordo com o contexto explicitado, ao longo deste trabalho, utilizar-se-á a mesma abordagem conceitual de (FAYYAD, 1997), ou seja, ressaltando a distinção entre KDD como processo e *data mining* como uma parte deste, embora alguns autores (GROTH, 2000) (SACHIN; VIJAY, 2012) considerem as definições como sinônimas. A Seção 2.6 e Seção 2.7 expõem maiores informações acerca do KDD e das tarefas da Mineração de Dados, o

intento é promover uma abordagem completa o suficiente para evidenciar o assunto e a motivação das escolhas definidas neste trabalho.

2.6 Etapas do Processo de KDD

As fases do processo de KDD são: seleção, pré-processamento, transformação, transformação, *data mining* e interpretação. Dentro de cada uma delas há a execução de uma ou mais atividades com o propósito de extrair conhecimento. A Figura 2.4 mostra as fases que formam a sistemática de *Knowledge Discovery in Database*, bem como o resultado após o término de cada uma delas.

A primeira fase, a seleção, os dados ou um subconjunto de variáveis são selecionados para os quais a descoberta de conhecimento será aplicada. Frawley, Piatetsky-Shapiro e Matheus (1992) recomendam, antes do início desta etapa, o entendimento a respeito do domínio de aplicação e das expectativas relativas ao usuário final. Os dados considerados relevantes ao problema investigado são a saída desta etapa conforme ilustra a Figura 2.4.

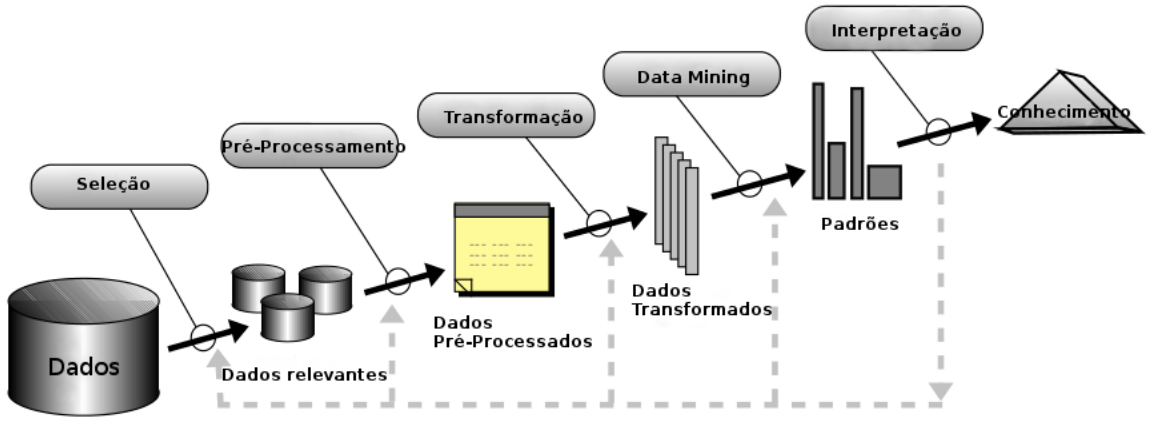

Figura 2.4 – Etapas do processo de *Knowledge Discovery in Database*

Fonte: adaptado de (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, p. 41)

No pré-processamento, os dados selecionados e tidos como indispensáveis ao domínio, em muitos casos, estão ruidosos ou inconsistentes. Os dados ruidosos são aqueles errados (*outliers*) ou que contam com valores considerados divergentes do padrão esperado (GOLDSCHMIDT; PASSOS, 2005). A inconsistência, ocorre quando os dados contêm alguma incoerência semântica, por exemplo, valores nulos em campos cuja restrição seja apenas valores numéricos. Consta como objetivo central desta etapa a limpeza dos dados, a fim de torná-los de qualidade, consequentemente, o modelo de conhecimento abstraído no processo de KDD também será otimizado. Goldschmidt e Passos (2005, p. 37) argumentam que a fase de limpeza de dados

"envolve uma verificação da consistência das informações, a correção de possíveis erros e o preenchimento ou a eliminação de valores desconhecidos e redundantes".

A fase de transformação se desenvolve na projeção e redução dos dados, isto é, são escolhidas formas de representar os dados dependendo dos objetivos da aplicação (FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1992). Aplicam-se métodos de transformações para reduzir o número de variáveis. Em outros casos, faz-se necessário a conversão dos dados quanto ao tipo ou representação, para se tornarem adequados à manipulação e processamento em determinadas ferramentas computacionais.

A etapa de *data mining*, como já discutido anteriormente, é a aplicação de algoritmos específicos com a finalidade de extrair os conhecimentos provenientes da base de dados. Nesta fase é indispensável a escolha da tarefa de mineração de dados, tais como: classificação, regressão, sumarização, agrupamento (*clustering*) etc. A escolha da tarefa tem por base os objetivos da aplicação desenvolvida, ou seja, com vistas no que se deseja descobrir acerca dos dados, além da representação final deste conhecimento. As tarefas de KDD são apresentadas com detalhes na Seção 2.7.

O produto da fase de *data mining* são padrões descobertos a partir dos dados de treinamento, conforme nota-se pela Figura 2.4. Contudo, a interpretação destes padrões é crucial à consolidação do conhecimento que será descoberto. Desta forma, na fase de interpretação, como a própria denominação sugere, todos os padrões são efetivamente analisados, através de uma representação particular, por especialistas do negócio, com o auxílio de *softwares* ou ferramentas matemáticas e estatísticas. Depreende-se da Figura 2.4 que durante a interpretação os requisitos e os objetivos da aplicação de KDD, elencados na fase de seleção dos dados, são validados em relação as características analisadas. Caso os padrões interpretados sejam irrelevantes, novas iterações são realizadas nas etapas anteriores.

Assim, com base em toda a sistemática apresentada e da argumentação supracitada, Fayyad, Piatetsky-Shapiro e Smyth (1996, p. 42, tradução nossa) defendem que o "processo de KDD é interativo e iterativo". O termo *interativo* remete a ideia de que os especialistas de negócio e analistas constantemente tomam decisões para validarem os resultados quanto aos requisitos do campo de aplicação. Quanto ao conceito *iterativo*, por sua vez, entende-se que as etapas podem ser repetidamente efetuadas para atestar a validade e qualidade do modelo em relação aos resultados.

2.7 Tarefas de KDD

Os dois objetivos primários da Mineração de Dados tendem a ser a predição e a descrição (FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1992). A primeira, utiliza algumas variáveis ou campos na base de dados para prever valores de outras variáveis de interesse. A segunda, por sua vez, concentra-se na busca de um modelo essencialmente interpretável para o

homem, a partir dos dados utilizados.

Os macro objetivos da aplicação de KDD tem sua origem na análise inicial sobre a base de dados e, como já argumentado, a intervenção de analistas e especialistas do negócio faz-se indissociável da definição dos objetivos que nortearão todo o desenvolvimento da aplicação (GOLDSCHMIDT; PASSOS, 2005). As tarefas de KDD são responsáveis por promover, de forma operacional, o alcance dos macro objetivos vislumbrados nos requisitos da aplicação, isto é, por meio delas são empregados algoritmos capazes de gerar padrões interpretáveis dentro de um contexto e finalmente, consolidados em conhecimento potencialmente útil. Fayyad, Piatetsky-Shapiro e Smyth (1996) destacam as principais tarefas de KDD, as quais estão sucintamente enumeradas a seguir.

a) **Classificação** (*classification*): esta tarefa consiste em mapear ou classificar um conjunto de dados em apenas uma de algumas classes predefinidas. Esta tarefa foi apresentada em detalhes de acordo com a necessidade deste trabalho na Seção 2.4.

b) **Regressão** (*regression*): caracteriza-se por mapear ou realizar predição de dados em uma variável numérica e real. As principais aplicabilidades desta tarefa são: predição da soma da biomassa presente em uma floresta; estimar a probabilidade de um paciente sobreviver, baseada em um conjunto de diagnósticos passados; e previsão de séries temporais (FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1992).

c) **Clusterização** (*clustering*) ou **Agrupamentos**: é uma modalidade de classificação não-supervisionada, na qual os dados que possuem certa similaridade são agrupados em*clusters* (JAIN; MURTY; FLYNN, 1999). Cabe destacar que em *clustering*, ao contrário da classificação, os rótulos de classe não são previamente conhecidos, portanto o usuário deve identificar, de maneira subjetiva, os grupos formados e nomeá-los de acordo com as características encontradas.

d) **Sumarização** (*summarization*): traduz-se na aplicação de métodos para encontrar uma descrição compacta para um conjunto de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Goldschmidt e Passos (2005) afirma que é comum aplicar a tarefa de sumarização aos agrupamentos oriundos da clusterização, a fim de gerar uma análise descritiva para os *clusters* encontrados.

e) **Modelo de dependências** (*dependency modeling*): consiste em encontrar dependências significativas entre as variáveis do problema, a dependência existe em dois níveis: estrutural e quantitativo (FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1992).

f) **Detecção de mudanças e desvios** (*change and deviation detection*): objetivo principal desta tarefa de KDD está na descoberta de mudanças significativas nos padrões esperados para os dados no domínio da aplicação. Goldschmidt e Passos (2005) acreditam que esta tarefa pode ser aplicada em cenários onde se pretende detectar

características que divirjam do perfil normal de compras em cartão de créditos de usuários, o que pode representar fraudes ou uso indevido por terceiros.

2.8 Conceitos e métricas usadas em classificação

Nesta seção evidenciam-se alguns conceitos e métricas amplamente aplicados em classificação, as quais também serão usados neste trabalho. Antes da apresentação de algumas terminologias importantes é fundamental entender a distinção entre os conceitos de tuplas positivas e negativas (HAN; KAMBER; PEI, 2012). As primeiras referem-se à classe de interesse, enquanto que as últimas dizem respeito aos demais rótulos. As informações aqui apresentadas foram coletadas da literatura referente a Mineração de Dados (GOLDSCHMIDT; PASSOS, 2005) (WITTEN; FRANK, 2005) (HAN; KAMBER; PEI, 2012) e estão enumeradas a seguir.

a) **Verdadeiros Positivos (*True Positive* – TP)**: as tuplas positivas corretamente rotuladas pelo classificador. Esta quantidade será chamada por TP.

b) **Verdadeiros Negativos (*True Negative* – TN)**: as tuplas negativas corretamente rotuladas pelo classificador. Esta quantidade será chamada por TN.

c) **Falsos Positivos (*False Positive* – FP)**: representam as tuplas negativas que são incorretamente classificadas como positivas. Esta quantidade será chamada por FP.

d) **Falsos Negativos (*True Negative* – FN)**: representam as tuplas positivas que são incorretamente classificadas como negativas. Esta quantidade será chamada por FN.

e) **Taxa de Verdadeiro Positivo (*True Positive Rate*)**: conhecida por **sensitividade** (*sensitivity*), taxa de reconhecimento ou *recall* denota a proporção de tuplas positivas que são corretamente identificadas. Dado P, a quantidade de tuplas positivas, a métrica avaliada é dada pela Equação 2.1.

$$TPR = \frac{TP}{P} \tag{2.1}$$

f) **Taxa de Verdadeiro Negativo (*True Negative Rate*)**: denominada por **especificidade** (*specifity*), isto é, a parcela de tuplas negativas corretamente identificadas. Considere N, o número de tuplas negativas, desta forma a taxa de verdadeiro negativo (TNR) é calculada pela Equação 2.3.

$$TNR = \frac{TN}{N} \tag{2.2}$$

g) **Taxa de Falso Positivo (*False Positive Rate*)**: representa a parcela de tuplas negativas incorretamente identificadas, expressa na Equação 2.3.

$$FPR = 1 - TNR \tag{2.3}$$

h) **Precisão**: mensura o percentual de exatidão do classificador. Considere a precisão P, evidenciada na Equação 2.4.

$$P = \frac{TP}{TP + FP} \tag{2.4}$$

Outra importante métrica capaz de avaliar o desempenho de classificadores e também amplamente discutida na literatura (HAN; KAMBER; PEI, 2012) é chamada de *F-score*, quantificada pela Equação 2.5.

$$F - score = \frac{2 \times P \times recall}{P \times recall} \tag{2.5}$$

i) **Taxa de Erro**: retrata a taxa de classificação incorreta. A Equação 2.6 é a convenção matemática para esta métrica.

$$Err(h) = \frac{1}{n} \sum_{i=1}^{n} \|y_i \neq h(i)\| \tag{2.6}$$

Tal que,

- o operador $\|y_i \neq h(i)\|$ retorna 1 se a expressão for verdadeira e 0 caso contrário;
- n é o número de exemplos;
- y_i é a classe real associada ao *i-ésimo* exemplo;
- $h(i)$ é a classe indicada (hipótese) pelo classificador para o *i-ésimo* exemplo

j) **Acurácia**: representa a capacidade do classificador mapear corretamente uma instância para uma dada classe. A Equação 2.7 expressa matematicamente a acurácia de h em função de $Err(h)$.

$$Acc(h) = 1 - Err(h) \tag{2.7}$$

k) **Matriz de Confusão**: tabela que relaciona, por classe, os números de classificações corretas e incorretas. A Tabela 4 mostra k classes distintas $\{C_1, C_2, C_3, \ldots, C_k\}$. O elemento $M(C_i, C_j)$ significa o número de exemplos pertencentes à C_i e que foram classificados em C_j. Todas as quantidades da diagonal principal correspondem denotam os acertos do indutor, pois $i = j$. A matriz de confusão, em classificação binária, recebe outro formato de acordo com a Tabela 5.

l) **Matriz de Custo**: o custo $Cost(C_i, C_j)$ é um número que corresponde a uma penalidade aplicada quando o classificador comete erro ao categorizar exemplos de C_i em C_j. A convenção matemática define:

Tabela 4 – Matriz de confusão para um problema com k classes

Classes	Predita C_1	Predita C_2	...	Predita C_k
Verdadeira C_1	$M(C_1,C_1)$	$M(C_1,C_2)$	...	$M(C_1,C_k)$
Verdadeira C_2	$M(C_2,C_1)$	$M(C_2,C_2)$	...	$M(C_2,C_k)$
...	...	...	...	...
Verdadeira C_k	$M(C_k,C_1)$	$M(C_k,C_2)$	...	$M(C_k,C_k)$

Fonte: (GOLDSCHMIDT; PASSOS, 2005, p. 69)

Tabela 5 – Matriz de confusão para um problema com 2 classes

Classes	Predita C_+	Predita C_-
Verdadeira C_+	Verdadeiros Positivos	Falsos Negativos
Verdadeira C_-	Falsos Positivos	Verdadeiros Negativos

Fonte: (GOLDSCHMIDT; PASSOS, 2005, p. 69)

- $Cost(C_i,C_j) = 0$, se $i = j$
- $Cost(C_i,C_j) > 0$, se $i \neq j$

A taxa de erro pode ser escrita em função do custo e dos elementos da matriz de confusão (Equação 2.8).

$$Err(h) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} M(C_i,C_j)Cost(C_i,C_j) \tag{2.8}$$

m) **Estatística Kappa**: método estatístico para avaliar o nível de concordância ou reprodutibilidade para dois conjuntos de dados (FLEISS, 1973), é calculado pela Equação 2.9.

$$k = \frac{I_o - I_e}{1 - I_e} \tag{2.9}$$

Onde,

- I_o taxa de aceitação relativa;
- I_e taxa hipotética de aceitação;
- Se a concordância é completa entre dois conjuntos de dados, $k = 1$.

n) ***Overfiting***: este fenômeno ocorre quando o classificador se ajusta em excesso ao conjunto de treinamento. Isto ocorre porque este conjunto pode não ser bastante representativo, desta maneira o classificador pode ter bom desempenho no conjunto de treinamento, em detrimento do conjunto de teste. A técnica capaz de atenuar os efeitos dessa característica é descrita na Seção 2.9.

o) ***Underfiting***: ocorre quando o classificador ajusta-se muito pouco ao conjunto de treinamento.

2.9 Validação Cruzada Estratificada em K Conjuntos (*Stratified K-Fold Cross-Validation*)

A técnica de validação cruzada (*cross-validation*) é um método estatístico empregado para comparar o desempenho de algoritmos classificadores, valendo-se da partição dos dados em base de treinamento (aprendizado) e testes destinados à estimar a precisão e confiabilidade do modelo construído pelo indutor (KOHAVI, 1995). Há muitas abordagens descritas na literatura, tais como: validação cruzada regular (*regular cross-validation*), *leave-one-out cross-validation* e validação cruzada com k conjuntos estratificada (*stratified k-fold cross-validation*) (KOHAVI, 1995) (WITTEN; FRANK, 2005). O estudo comparativo de Kohavi (1995) apontam que o uso da técnica *stratified k-Fold cross-validation* apresenta melhores resultados, quando a variável k assume o valor igual a 10, desta forma utilizar-se-á esta mesma abordagem nos experimentos, dada sua relevância em pesquisas (DIETTERICH, 1998).

O método *K-Fold Cross-Validation* segmenta a base de dados (*dataset*) em k subconjuntos mutuamente exclusivos de tamanhos aproximadamente iguais. Os $k-1$ subconjuntos são utilizados para treinamento, enquanto apenas um deles é usado para validação do modelo. O processo é repetido k vezes até que todos os *folds* tenham sidos aplicados nas fases de treinamento e testes. A finalidade desta metodologia consiste em avaliar o desempenho médio e a precisão dos algoritmos classificadores. O método *Stratified K-fold Cross-Validation* acrescenta a característica de possuir os subconjuntos com aproximadamente a mesma proporção dos rótulos existentes na base original (KOHAVI, 1995).

A Figura 2.5 ilustra a técnica de validação cruzada para $k = 3$, nota-se a cada momento uma partição, o elemento mais escuro, submetida ao teste, enquanto o restante do *dataset* é aplicado a classificação.

Kohavi (1995) faz uma descrição mais formal acerca do processo, uma vez que afirma que a base de dados $\mathcal{D}$ é dividida em k subconjuntos mutuamente exclusivos (*folds*) $D_1, D_2 \ldots D_k$ de tamanhos aproximadamente iguais. O indutor I é treinado e testado k vezes; cada vez $t \in \{1,2,\ldots,k\}$, sendo treinado em $\mathcal{D} \setminus \mathcal{D}_t$ e validado no segmento $\mathcal{D}_t$. Seja $\mathcal{D}(i)$ o conjunto de teste com instâncias $x_i = \langle v_i, y_i \rangle$, a técnica de validação cruzada estima a acurácia do classificador de acordo com a Equação 2.10, tal que δ representa a função custo e $\langle v_i, y_i \rangle$ é uma amostra pertecente ao conjunto $\mathcal{V}$ para os possíveis rótulos y_i que pertencem a $\mathcal{Y}$.

$$a_{cc} = \frac{1}{n} \sum_{\langle v_i, y_i \rangle} \delta(I(\mathcal{D} \setminus \mathcal{D}(i), v_i), y_i) \tag{2.10}$$

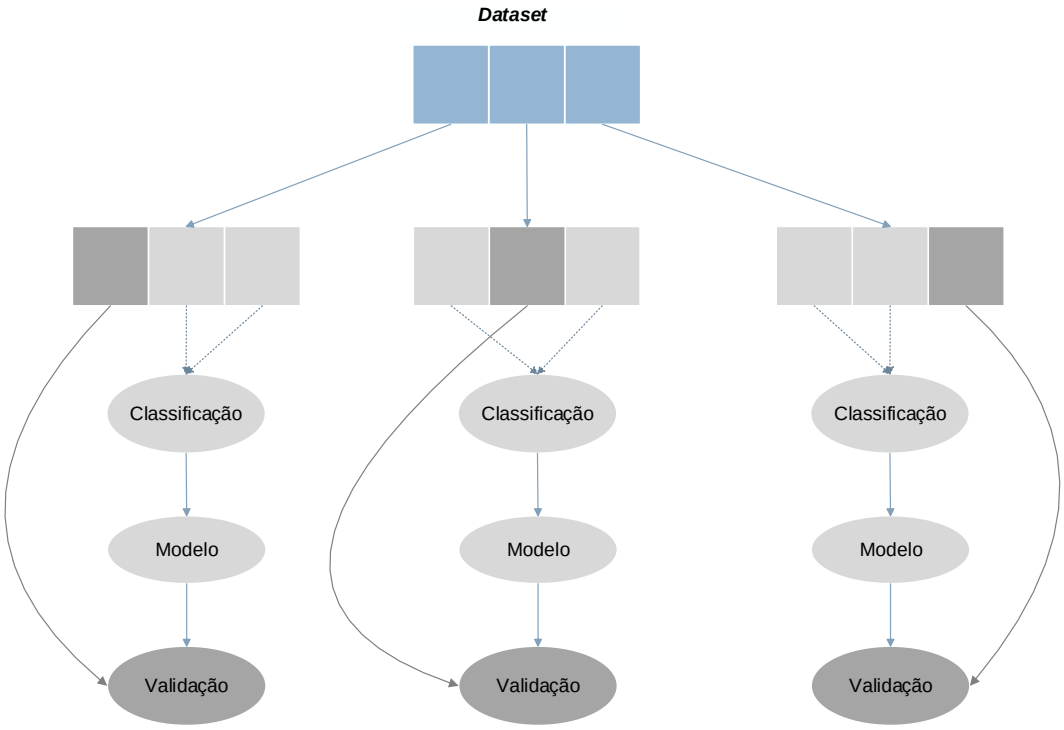

Figura 2.5 – Validação cruzada para 3 *folds*

Fonte: Da pesquisa

2.10 Considerações Finais

Neste capítulo evidenciaram-se os conceitos e fundamentos da mineração de dados, especificamente na sua etapa de classificação. Outro fator importante diz respeito as métricas de desempenho, indispensáveis na determinação da qualidade e precisão dos algoritmos. Todavia, a ocorrência do *overfiting* pode ocultar o ajuste em excesso do classificador à base de treinamento, assim a solução deste problema se dá pelo método de validação cruzada estratificada em k conjuntos.

3 Conceitos sobre Probabilidade e Redes Bayesianas

3.1 Considerações Iniciais

Neste capítulo são apresentados alguns conceitos básicos da teoria da probabilidade, cuja base teórica pode ser encontrada em obras específicas como (SOONG, 1986; SOONG, 2004) e (CLARKE; DISNEY, 1985). A abordagem voltada à Inteligência Computacional está publicada em trabalhos clássicos dentro da Ciência da Computação, tais como (HAN; KAMBER; PEI, 2012) e (RUSSEL; NORVING, 2010). O teorema de Bayes e da probabilidade condicional são as bases do desenvolvimento de toda a formulação das redes Bayesianas e o raciocínio sobre incertezas.

Após essa discussão apresentam-se os conceitos teóricos centrais do estudo das Redes Bayesianas, além disso, exemplos práticos são oferecidos pretendendo-se aprimorar a compreensão acerca desse assunto. Diante disso, esclarecem-se os seguintes pontos: componentes da rede e sua estrutura; definições básicas; representação gráfica e matemática; o processo de inferência e; a aprendizagem de estrutura da rede.

O estudo das redes Bayesianas se mostrou atrativo aos objetivos desta pesquisa em razão de propiciar: uma representação gráfica da qual é possível a extração de conhecimento inseridas em domínio de aplicação; e o relacionamento de variáveis visando detectar causas e efeitos atrelados à incerteza. Os fundamentos supracitados são julgados indispensáveis e favorecem a compreensão do restante do trabalho, em particular, o capítulo que expõe os resultados alcançados nesta pesquisa, onde são apresentadas as redes Bayesianas usadas em estudos de caso.

3.2 Propriedades da probabilidade

Soong (1986) afirma que três axiomas são a base de postulados para se deduzir todas as propriedades de uma função de probabilidade, a seguir eles estão evidenciados:

Axioma 1 $0 \leq P(A) \leq 1$

Axioma 2 $P(S) = 1$, *tal que S corresponde ao espaço amostral.*

Axioma 3 *Se A_1, A_2, $\ldots A_n$ é uma coleção numerável de eventos mutuamente exclusivos, isto é $A_i \cap A_j = \varnothing$, então:*

$$P(A_1 \cup A_2 \cup \ldots) = P(\sum_j A_j) = \sum_j P(A_j) \tag{3.1}$$

3.3 Probabilidade Condicional

Define-se a probabilidade condicional $P(A|B)$ como sendo a probabilidade da ocorrência do evento A tendo a evidência que B ocorreu, em outras palavras também conhecida por *probabilidade A dado B*. Matematicamente, o conceito está expressado na Equação 3.2

$$P(A|B) = \frac{P(A \cup B)}{P(B)} \quad P(B) \neq 0 \tag{3.2}$$

A probabilidade condicional demanda a existência de um espaço amostral restrito, deste modo, B substitui S e a probabilidade condicional $P(A|B)$ é conseguida como sendo a probabilidade de A em relação ao novo espaço amostral. Percebe-se pelo axioma 2 que

$$P(S|B) = 1$$

Ou ainda,

$$P(S|B) = \frac{P(SB)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

Pelo axioma 3, se $A_1, A_2, \ldots A_n$ são mutuamente exclusivos, em decorrência disso A_1B, A_2B, $\ldots A_nB$ também são. Assim,

$$\begin{aligned} P[(A_1 \cup A_2 \cup \ldots A_n)|B] &= \frac{P[(A_1 \cup A_2 \cup \ldots A_n)|B]}{P(B)} \\ &= \frac{P[A_1B \cup A_2B \cup \ldots A_n]}{P(B)} \\ &= \frac{P(A_1B)}{P(B)} + \frac{P(A_2B)}{P(B)} + \ldots \frac{P(A_nB)}{P(B)} \\ &= P(A_1|B) + P(A_1|B) + \ldots + P(A_n|B) \end{aligned}$$

3.4 Independência Estatística

Conceitualmente, os eventos A e B são *estatisticamente* independentes se ambos não afetam a ocorrência ou não-ocorrência do outro, para isso as probabilidades associadas a cada

um dele deve atender à Equação 3.3. A respeito disso, vale destacar que as probabilidades dos eventos individuais, em geral, não dizem respeito ao comportamento em conjunto.

$$P(A \cap B) = P(A)P(B) \tag{3.3}$$

3.5 Teorema da Probabilidade Total

Sejam os eventos $B_1, B_2, B_3, \ldots, B_n$, mutuamente exclusivos, tal que $S = B_1 \cup B_2 \cup B_3 \cup B_n$. O diagrama de Venn-Euler mostrado na Figura 3.1 evidencia que um evento A pode ser representado pela união de eventos mutuamente excludentes $AB_1, AB_2, \ldots AB_n$. Dessa observação e do teorema da probabilidade condicional (Equação 3.2) deriva-se a Equação 3.4.

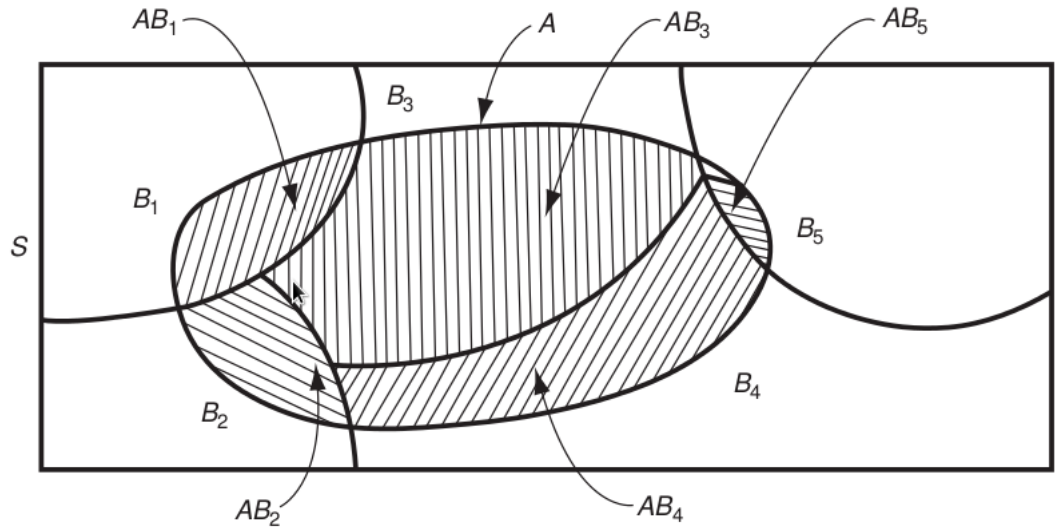

Figura 3.1 – Diagrama de Venn-Euler aplicado a um espaço amostral particionado em eventos mutuamente exclusivos

Fonte: (SOONG, 2004) p. 23.

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots \\ &= P(A|B_n)P(B_n) \\ &= \sum_{j=1}^{n} P(A|B_j)P(B_j) \end{aligned} \tag{3.4}$$

3.6 Teorema de Bayes

Considerando os eventos arbitrários A e B, tal que $P(A) \neq 0$ e $P(B) \neq 0$, tem-se a Equação 3.5.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{3.5}$$

O teorema de Bayes representa uma conjugação do teorema de probabilidade condicional e da fórmula de probabilidades totais (Seção 3.5) conforme está disposto na Equação 3.6, permitido-se calcular a probabilidade a *posteriori* $P(B|A)$ em termos da informação a *priori* $P(B)$ e $P(A|B)$.

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n}[P(A|B_j)P(B_j)]} \tag{3.6}$$

3.7 Redes Bayesianas

Esta seção introduz uma estrutura de dados proposta por Judea Pearl em 1986 (PEARL, 1986), chamada Redes Bayesianas (RB); conhecida também por Redes de crença, Rede probabilística, Rede Causal e mapa de conhecimento (RUSSEL; NORVING, 2010)(HAN; KAMBER; PEI, 2012). Essa estrutura refere-se à representação gráfica de variáveis e suas relações causais sob um determinado cenário de aplicação.

Ko e Kim (2014) defendem a importância da representação do conhecimento em Inteligência Computacional (Seção 2.2), pois isso determina o quão fácil as características presentes na base de dados podem ser compreendidas. A argumentação dos autores, dentro do contexto estudado, sobreleva a relevância da Rede Probabilística, uma vez que Hlel, Jamoussi e Hamadou (2016) julgam a RB como um dos melhores modelos de interpretação do conhecimento baseado em formalismos teóricos.

As RBs têm sido empregadas na resolução de problemas em diversos campos de estudo. Borunda et al. (2016) destacam aplicações em ramos como diagnósticos médico, busca heurística, questões ambientais, gestões de bacias hidrográficas, sensores virtuais, entre outras. Essa gama de trabalhos confirma a utilidade e flexibilidade das redes de crenças na modelagem de cenários complexos em uma notação simples, não obstante objetiva. Isto posto, na pesquisa apresentada nesta dissertação explora-se esse potencial para o desenvolvimento de modelos voltados a descoberta de padrões alusivos à desistência ou retenção em cursos de graduação.

3.8 Estrutura das Redes Bayesianas

A rede de crenças é um grafo direcionado, conectado e acíclico, conhecido na literatura por *Directed Acyclic Graph* (DAG), onde cada nó é uma variável aleatória. Dessa forma os conceitos enumerados anteriormente acerca de teoria dos grafos significam, respectivamente: estruturas onde há a direção dos arcos; todos os nós estão conectados na rede e; finalmente, partindo-se de um nó arbitrário e seguindo por todos os arcos conectados não há retorno.

Os arcos referem-se às dependências ou influências, se houver um vínculo de um nó X até outro Y, denomina-se que X é pai de Y. A cada nó X_i há uma distribuição de probabilidade condicional $P(X_i|Pais(X_i))$ que quantifica o efeito dos pais sobre o nó (RUSSEL; NORVING, 2010). Segundo Russel e Norving (2010), a topologia da rede ou o conjunto de nós e vínculos (arcos) especifica os relacionamentos de independência condicional e a sua semântica gráfica sugere que as causas são pais dos efeitos.

Outro elemento importante às redes probabilísticas é a tabela de probabilidade condicional (TPC), que denota uma matriz que contém a probabilidade condicional de cada valor de nó à combinação de possíveis valores dos nós pai. A soma de cada linha deve ser igual a 1, uma vez que as entradas representam combinações exaustivas dos possíveis eventos em cada variável.

A RB de um domínio hipotético mostrada na Figura 3.2, esquematicamente exemplifica os conceitos debatidos nesta seção. Observa-se a presença de cinco nós *booleanos*, sendo que A e B são pais de C, enquanto D e E, filhos de C, são nós folhas. Cada linha das TPCs, exceto aos nós A e B (variáveis-raiz), apresentam a probabilidade combinada exaustivamente para todas as possibilidades dos nós pais. Diante disso, por exemplo, nota-se que $P(C|A,B)$ ou $P(C = \text{“}V\text{”} \mid A = \text{“}V\text{”}, B = \text{“}F\text{”})$ vale 0,25.

As topologias das redes e as tabelas de probabilidades são elementos cruciais para o raciocínio sobre incertezas, isto é, na descoberta de padrões dos quais se conhece pouco acerca do domínio de aplicação, no que tange as relações de causas e efeitos. Assim pode-se responder muitas questões acerca do domínio somente por meio de inferências.

3.9 Inferência em Redes Bayesianas

Segundo (KORB; NICHOLSON, 2010) o termo inferência ou atualização de crença (*belief updating*) contextualizado em sistemas aplicados a redes de bayesianas, refere-se ao cálculo da distribuição de probabilidade *posteriori*, atualizada por toda a estrutura da rede, dado um conjunto de variáveis aleatórias de evidências.

O teorema de Bayes, apresentado na Equação 3.5, é a base de todos os sistemas modernos de inferência probabilística (RUSSEL; NORVING, 2010). A probabilidade *posteriori* é calculada a partir da generalização desse teorema (Equação 3.6), onde o termo $\frac{1}{\sum_{j=1}^{n}[P(A|B_j)P(B_j)]}$

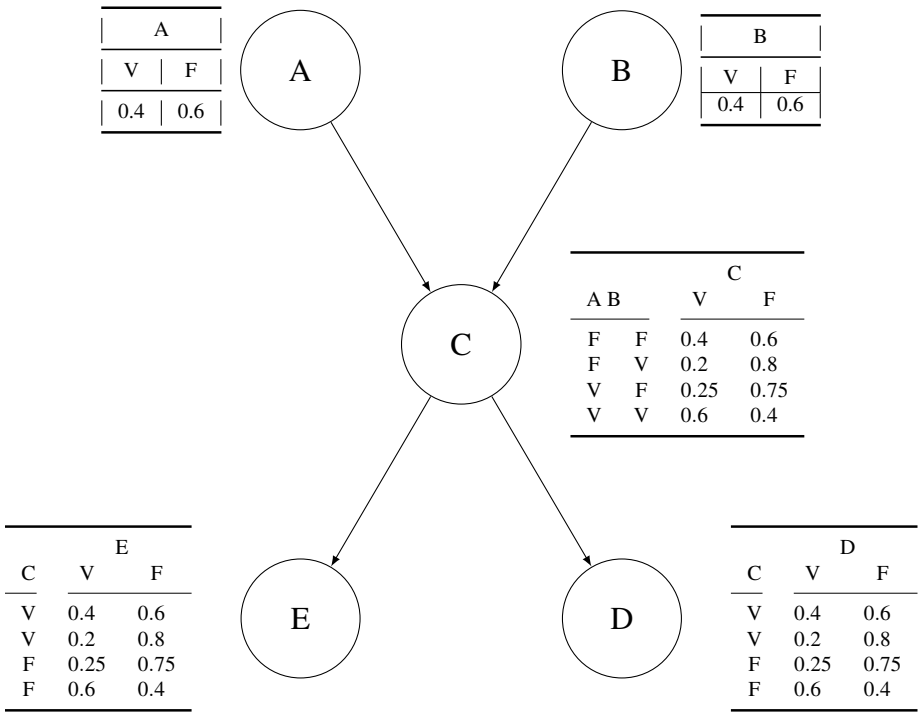

Figura 3.2 – Rede Bayesiana com tabelas de probabilidade condicional dos nós A, B, C, D e E

denota uma constante de normalização α, necessária para tornar a soma das entradas $P(A|B_j)$ igual a 1. Dessa discussão resulta que a probabilidade a *posteriori* é conseguida pela Equação 3.7, tal que j representa os possíveis estados da variável B.

$$P(B_j|A) = \alpha P(A|B_j)P(B_j) \tag{3.7}$$

Retomando à topologia disposta na Figura 3.2, é permitida a formalização do processo de inferência. Por conseguinte, considera-se saber a saída do sistema, dadas as consultas $P(D,\overline{C},E,A,B)$. A resposta da inferência se obtém a partir do teorema de Bayes e o seu detalhamento matemático está mostrado na Equação 3.8.

$$\begin{aligned} P(D,\overline{C},E,A,B) &= P(D|\overline{C})P(\overline{C},E,A,B) \\ &= P(D|\overline{C})P(E|\overline{C})P(\overline{C},A,B) \\ &= P(D|\overline{C})P(E|\overline{C})P(\overline{C}|A,B)P(A)P(B) \end{aligned} \tag{3.8}$$

A generalização do exemplo anterior permite escrever a Equação 3.9, a qual denota a forma geral da inferência, onde $pais(X_i)$ representam todos os valores em $Pais(X_i)$ que aparece em $x_1,\ldots,x_n$ (RUSSEL; NORVING, 2010).

$$P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i|pais(X_i)) \tag{3.9}$$

3.10 Aprendizagem da estrutura

O processo de aprendizagem da estrutura de redes de crenças pode ser concebido em duas modalidades distintas. A primeira forma, recorre ao emprego de algoritmos que aprendem a partir de uma base de dados o grafo que melhor representa a distribuição de probabilidade do conjunto de dados (SARDINHA; PAES; ZAVERUCHA, 2009); enquanto a outra possibilidade, vale-se de interações com usuários especialistas no domínio de aplicação. De acordo com Santana (2008), contudo, verifica-se grande consumo de tempo na execução desta última, visto que é mais factível a combinação das duas estratégias dirigida à validação de modelos.

Santana (2008) assegura a grande relevância do estudo de técnicas dedicadas ao aprendizado de estrutura pelo fato de o tamanho do espaço de busca de possíveis estruturas aumentar exponencialmente junto com o número de variáveis do modelo. Em virtude disso, as topologias de certas redes tornam-se demasiadamente complexas, assim dificultam a interpretabilidade oriunda dos seus elementos estruturais.

Sardinha, Paes e Zaverucha (2009) destacam a existência de três categorias de algoritmos de aprendizado de estrutura, a saber: baseado em pontuação (*score-based*); baseados em restrições (*constraint-based*); e híbridos, agregadores das particularidades de ambos os anteriores. Os Algoritmos *score-based* adotam uma função de avaliação para qualificar a rede bayesiana e uma heurística para contornar situações nas quais o problema de buscas das redes, em todo o espaço, torna-se intratável. A abordagem *constraint-based*, por sua vez, faz uso de diversos testes de independência condicional como por exemplo, algum teste de hipótese ou *score* oriundo da teoria da informação (SARDINHA; PAES; ZAVERUCHA, 2009).

Considerando as colocações debatidas anteriormente, neste trabalho será empregado um método clássico de aprendizado de estrutura baseado em pontuação, denominado K2 (COOPER; HERSKOVITS, 1992). Esta opção se justifica, segundo as asserções de Ko e Kim (2014), como sendo um dos melhores e mais eficientes métodos de aprendizado, além da facilidade de implementação (YANG; CHANG, 2002). O K2 percorre todo o espaço de busca valendo-se da métrica de pontuação dada por Equação 3.10.

$$P(B_s|X) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \tag{3.10}$$

Onde:

- X é a base de dados com n observações;
- B_s a dimensão de estrutura;
- ri é a quantidade de valores que a variável X_i pode assumir;

- N_{ijk} é o número de observações na base X, tal que X_i é configurado com o valor k e seus pais com o valor j.

3.11 Considerações Finais

Neste capítulo foram apresentados os principais fundamentos de probabilidade e suas implicações no entendimento das Redes Bayesianas. Finalmente, esta estrutura foi compreendida sob o enfoque de sua representação gráfica indissociável das TPCs, o que proporciona o processo de inferência, metodologia basilar na criação de agentes capazes de raciocinar sobre incertezas, isto é, aqueles que geram respostas a partir de cenários nos quais se conhece pouco acerca do relacionamento entre variáveis.

4 Modelagem da Aplicação de Mineração de Dados Educacionais destinada à busca de perfis de alunos em casos de evasão ou retenção

4.1 Considerações Iniciais

Neste capítulo são apresentados pontos relativos às especificidades do projeto do sistema acadêmico da UFPA juntamente à seleção dos dados e do processo de migração destes a uma nova base. Em adição, foram debatidas os procedimentos, ferramentas e metodologias voltadas ao pré-processamento, transformação e Mineração de Dados do conjunto de amostras.

4.2 Base de dados do SIGAA

O Sistema Integrado de Gestão de Atividades Acadêmicas (SIGAA) é parte dos Sistemas Institucionais Integrados de Gestão (SIG) e informatiza os procedimentos da área acadêmica através de módulos como: graduação, pós-graduação (*stricto* e *lato sensu*), ensino técnico, entre outros. O SIGAA foi adquirido pela Universidade Federal do Pará, por meio de um contrato firmado com a Universidade Federal do Rio Grande do Norte (UFRN); além disso outras IFES e Institutos Federais (IF) também adquiriram o produto a fim de promover integração entre sistemas, processar dados e oferecer serviços da área fim através de interface *web*.

Destaca-se que os dados empregados nesta pesquisa, apresentados na Seção 4.3, são advindos do módulo de graduação e, antes de estarem prontos ao *Data Mining*, passaram por certas etapas como: migração de dados, limpeza, pré-processamento e transformação. As fases descritas são discutidas com detalhes nas sessões 4.4, 4.3 e 4.5.

4.3 Dados selecionados

Os dados selecionados à pesquisa são registros acadêmicos oriundos do SIGAA, referentes a 157.298 discentes dos cursos de graduação da Universidade Federal do Pará, ingressantes até o ano de 2016. Desta quantidade, as amostras inconsistentes[1] foram removidas do

[1]Não estavam de acordo com restrições ou semanticamente errados (por exemplo, idade menor que 10 anos)

conjunto de dados finais, permanecendo 98.698 linhas. A Tabela 7 mostra os 31 atributos elegidos, seus respectivos significados, os tipos dos dados e a indicação se a variável foi transformada da base de dados original.

A forma de ingresso (*forma_ingresso*) pode assumir sete valores possíveis: Processo Seletivo (PS), Plataforma Freire (PARFOR), Mobilidade Externa (MOBEX), Mobilidade Interna (MOBIN), Exame Nacional do Ensino Médio (ENEM) e outras seleções (OUTRAS) diferentes destas citadas que foram ou ainda são aplicadas como critério de ingresso aos cursos de graduação da UFPA.

As variáveis de 13 a 19 representam os indicadores de rendimento acadêmico acumulado, estabelecidos no regimento da UFRN. Esses índices mensuram o desempenho dos alunos da graduação e nos cálculos consideram-se dados do histórico acadêmico, tais como: quantidades de reprovações, aprovações, trancamentos, cargas horárias acumuladas e esperadas para integralização do curso, entre outros. A Universidade Federal do Pará, oficialmente (UFPA, 2013), implementa somente o Coeficiente de Regimento Geral (CRG), também denominado Índice de Rendimento Acadêmico (IRA). Contudo, neste trabalho avaliaram-se todos os indicadores implementados no SIGAA. As fórmulas para calculá-los e os seus respectivos significados estão dispostos no Anexo A o qual foi extraído do Regimento da Graduação da Universidade Federal do Rio Grande do Norte (UFRN, 2013).

Os atributos de 20 a 22 denotam a probabilidade de um discente formado nos últimos cinco anos possuir um dos índices acadêmicos igual ou superior aos demais alunos pertencentes ao mesmo curso e matriz curricular. Foram usados os indicadores MC, IRA e IEA, uma vez que estes, em suas definições matemáticas e conceituais, aferem a eficiência do aluno durante o seu percurso acadêmico.

A média das notas obtidas pelo estudante em cada disciplina, em um período letivo, é convertida em conceito, definido segundo a escala apresentada na Tabela 6. As variáveis indexadas de 23 a 30 referem-se ao percentual de um determinado conceito de acordo com o período de avaliação, seja este geral (acumulado por todo o curso) ou para o primeiro ano cursado. Por exemplo, a variável perc_ins_primeiro_ano denota o percentual de conceitos do tipo INS obtido pelo discente no primeiro ano de graduação.

Quanto aos demais atributos, considerados intuitivos, podem ser consultados nas descrições dispostas na Tabela 7.

Tabela 6 – Correspondência entre a média das notas e o conceito

Conceito	Intervalo da média
Insuficiente (INS)	[0-4,99]
Regular (REG)	[5-6,99]
Bom (BOM)	[7-8,99]
Excelente (EXC)	[9-10]

Finalmente, o atributo 31 representa a classe a qual o discente pertence, cujos possíveis valores são: “Formado”, “Evadido” e “Retido”. Os alunos considerados na classe “Formado” são aqueles que conseguiram integralizar a carga horária prevista pelo curso. Por sua vez, o rótulo “Evadido” remete-se aos alunos que, por decisão própria ou processo de prescrição previsto em regimento da instituição, abandonaram a graduação. Os estudantes com matrículas ativas no SIGAA, porém que ultrapassaram um ano do prazo de conclusão estabelecido no currículo do curso foram classificados como “Retido”. Existem na base de dados 65.758 (66,63%) amostras referentes a classe dos alunos formados; 25.581 (25,92%) dos registros, pertencem aqueles que desistiram dos estudos; e por fim, os alunos em retenção são menos representativos 7.359 (7,46%).

Tabela 7 – Variáveis selecionadas à pesquisa

Número	Atributo	Descrição	Tipo	Processado
1	sexo	sexo que o discente pertence	"M"ou "F"	
2	idade	idade que o aluno ingressou no curso	[14,70]	
3	interior	informa se o discente estuda no campus capital ou do interior	true ou false	X
4	turno	turno no qual o discente estuda, sendo os possíveis valores: diurno (D), integral (I) e noturno (N)	{"D", "I", "N"}	X
5	forma_ingresso	forma de seleção pela qual o discente ingressou na universidade	{"PS", "PARFOR", "OUTRAS", "MOBEX", "PSE", "MOBIN", "ENEM"}	X
6	numero_trancamento	Número de vezes que o discente trancou a matrícula	inteiro	X
7	numero_vinculos	Número de vínculos até o ingresso no curso atual	inteiro	X
8	perc_ch_pratico	Percentual de carga horária dedicada às aulas práticas	[0,100]	
9	perc_ch_estagio	Percentual de carga horária dedicada ao estágio	[0,100]	
10	perc_ch_teorico	Percentual de carga horária dedicada às aulas da modalidade presencial	[0,100]	
11	sem_ordem	O percentual das disciplinas cursadas fora da ordem proposta pelo currículo do discente	[0,100]	X
12	primeiro_semestre_ocorr	Informa qual o semestre que o discente cursou pela primeira vez uma disciplina fora de ordem	inteiro	X
13	mc	Média de Conclusão	[0,10]	X
14	crg	Coeficiente de Rendimento Geral	[0,10]	X
15	mcn	Média de Conclusão Normalizada	reais maiores ou iguais a 0.	X
16	iech	Indíce de Eficiência em Carga Horária	reais maiores ou iguais a 0.	X
17	iepl	Índice de Eficiência em Períodos Letivos	reais maiores ou iguais a 0.	X
18	iea	Índice de Eficiência Acadêmica	reais maiores ou iguais a 0.	X
19	iean	Índice de Eficiência Acadêmica Normalizada	reais maiores ou iguais a 0.	X

Número	Atributo	Descrição	Tipo	Processado
20-22	probabilidade_{índice}[2]	Refere-se a probabilidade de um discente formado nos últimos 5 anos possuir o índice acadêmico maior ou igual ao aluno avaliado (*Teste z*)	[0,100]	X
23-30	perc_conceito_{tipo}[3]_{avaliação}[4]	Refere-se ao percentual de um conceito conseguido pelo discente dentro do período avaliado	[0,100]	X
31	status	rótulos usados na classificação	{"FORMADO", "EVADIDO", "RETIDO"}	X

4.4 Migração de dados

A migração de dados foi realizada em 4 estágios, devido sua complexidade de implementação e processamento. O primeiro passo expressa-se em fazer uma cópia[5] das tabelas do banco de dados do SIGAA e restaurá-lo num outro servidor Linux. Essa escolha referente ao sistema operacional baseado em Unix se deve ao fato deste disponibilizar ferramentas nativas as quais permitiram automatizar muitas rotinas através de *shell script*. Após esse procedimento, foram criadas algumas tabelas e funções indicadas a suportar fases de seleção, pré-processamento e transformação dos dados empregadas nesta pesquisa. A sistemática descrita está apresentada na Figura 4.1 e esta é imprescindível para outros estágios.

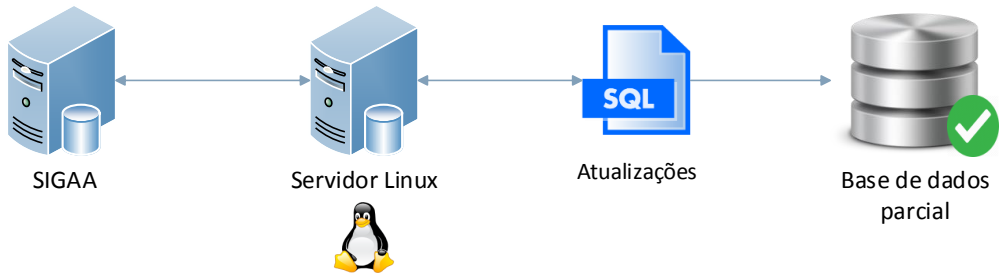

Figura 4.1 – Criação da infraestrutura da base de dados

No segundo estágio, desenvolveu-se uma rotina em *Bash Script* (NEWHAM; ROSENBLATT, 1998) que iniciou alguns processos concorrentes e paralelos os quais interagiram com um conjunto de funções (*functions*), escritas em linguagem SQL. Cada processo manipulou um segmento da base de dados, proporcional ao número de discentes da graduação registrados no SIGAA. A estratégia multiprocessada e concorrente adotada reduziu consideravelmente o tempo de processamento de dados, o procedimento descrito está apresentado na Figura 4.2.

A terceira etapa funciona essencialmente como a anterior, porém o conjunto de funções desenvolvidas em SQL referem-se ao cálculo de índices acadêmicos (veja Seção 4.3). É impor-

[2] índice = MC, IRA e IEA
[3] tipo = INS, REG, BOM e EXC
[4] avaliação = Geral e Primeiro semestre
[5] *Dump* é uma expressão de língua inglesa empregada nesse sentido

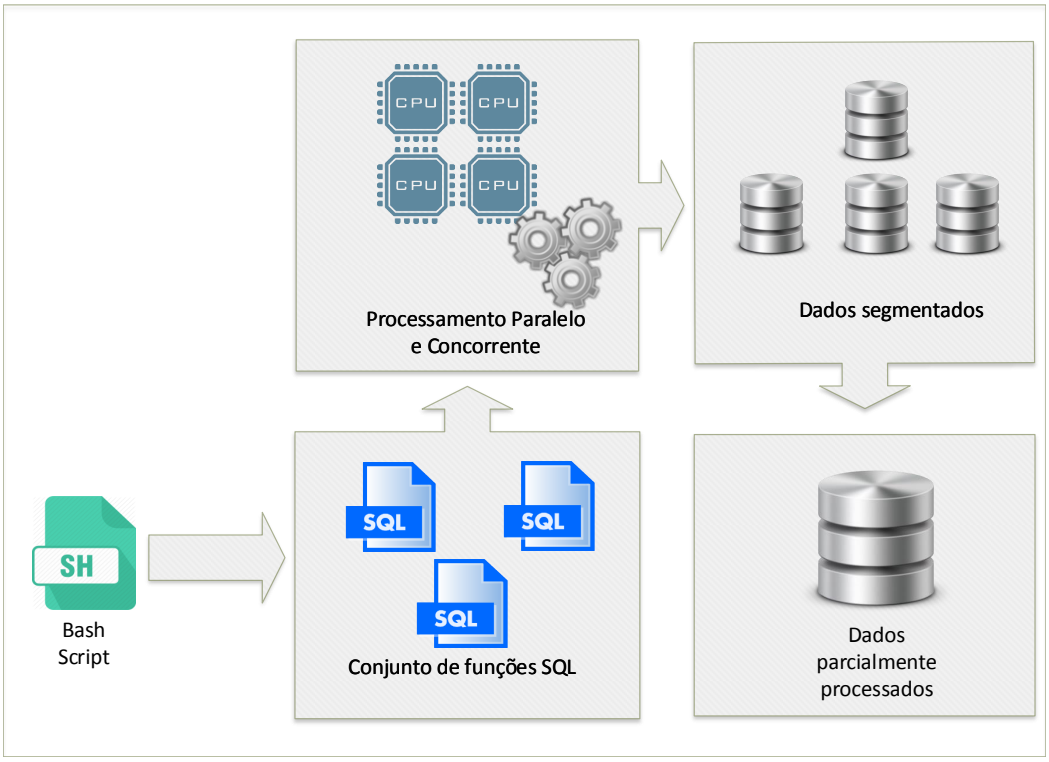

Figura 4.2 – Migração dos dados de discentes da base do SIGAA ao novo servidor

tante frisar que um ambiente de programação paralela[6] e concorrente[7] manipula essas funções. Assim, denota-se da Figura 4.3 que a base de dados não foi segmentada, isto é, cada processo executa o seu código sobre os dados completos, contudo recebe a sua função (código SQL) de cálculo e salva o resultado em tabelas específicas.

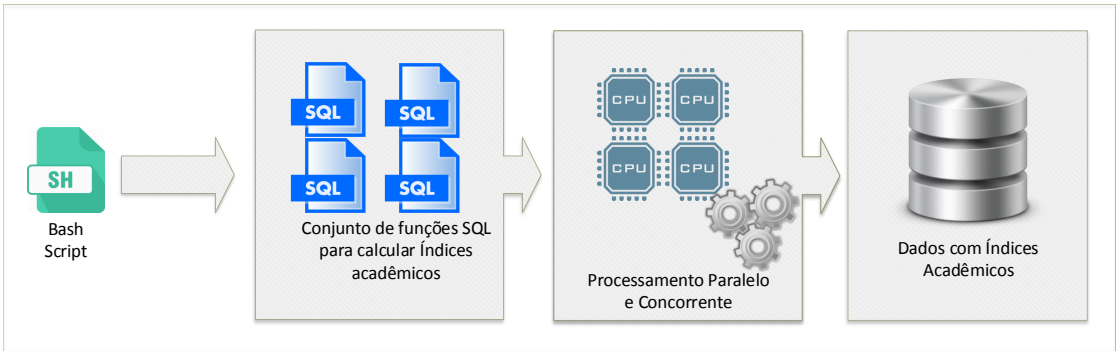

Figura 4.3 – Cálculo dos índices acadêmicos

Por fim, a quarta fase da migração diz respeito aos cálculos das probabilidades dos índices acadêmicos (veja Apêndice A). Optou-se pela programação sequencial, visto que a solução apresentou baixo tempo de processamento, ressaltam-se que as soluções testadas – uma na linguagem de programação Java e outra extensão do banco de dados – se mostraram ineficientes

[6] mais de um processador recebeu alguns processos

[7] os processos na mesma unidade de processamento foram alocados de acordo com a política de gerência de processo

à carga de dados, uma vez que demoraram várias horas para realizar os cálculos ou revelaram erros de representação matemática (*overflow* ou *underflow*). O mecanismo desenvolvido traduz-se em submeter um *Bash Script* que conecta-se ao banco de dados e exporta um arquivo em formato *Comma-Separated Values* (CSV) (SHAFRANOVICH, 2005) o qual contém o identificador do aluno, o valor do índice acadêmico, média e desvio padrão amostral do índice acadêmico dos alunos formados nos últimos 5 anos. Posteriormente, esse arquivo é automaticamente submetido ao *software* GNU Octave (GNU OCTAVE, 2016), projetado para realizar vários cálculos matemáticos de maneira eficiente. Os valores são salvos em outro arquivo CSV, sendo persistido em lote ao servidor de banco de dados. O processo se repete até que todos os índices sejam computados. Toda a dinâmica apresentada é mostrada na Figura 4.4.

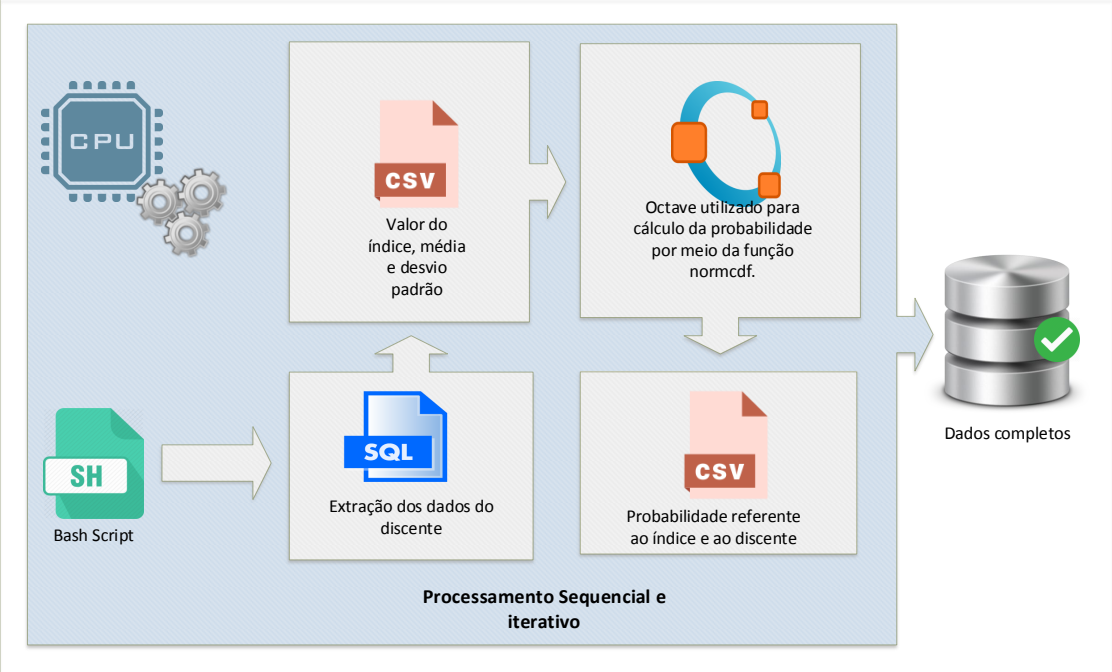

Figura 4.4 – Interface do cálculo das probabilidades entre banco de dados, *Bash Script*, arquivos CSV e a ferramenta matemática Octave.

4.5 Pré-processamento e Tranformação

Os atributos selecionados (Tabela 7) foram inicialmente pré-processados durante uma das fases apresentadas na Seção 4.4. A Figura 4.5 mostra os passos restantes até a obtenção do *dataset* usado na mineração de dados. Inicialmente, o arquivo CSV resultante da etapa ilustrada pela Figura 4.4, foi convertido em formato ARFF (*Attribute-Relation File Format*), o qual representa um arquivo de texto específico da ferramenta Weka (WITTEN; FRANK; HALL, 2016), desenvolvido pelo Departamento de Ciência da Computação da *University of Waikato*.

Os classificadores empregados neste trabalho, em suas implementações, não manipulam variáveis contínuas, visto que estas necessitam estar discretizadas segundo alguma notação.

Dessa forma, utilizou-se o filtro do Weka voltado à conversão de valores numéricos em uma notação categórica e significativa aos algoritmos estudados. A estratégia adotada na discretização foi a divisão em intervalos arbitrários com distribuição igual de frequência, esta abordagem tem a mesma finalidade da técnica *Cross-Validation*, isto é, evitar *overfiting*.

O conjunto de dados construído anteriormente não apresentava uma notação intuitiva, dessa maneira esta foi ajustada com o auxílio de um *script* de processamento de texto o qual emprega algumas expressões regulares compatíveis com o editor Sed (*Stream Editor*) e AWK. O produto do passo anterior, conforme denota a Figura 4.5 representa o ARFF utilizado no processo de extração de conhecimento, cujo cabeçalho é apresentado pelo Código A.1 no Apêndice A.

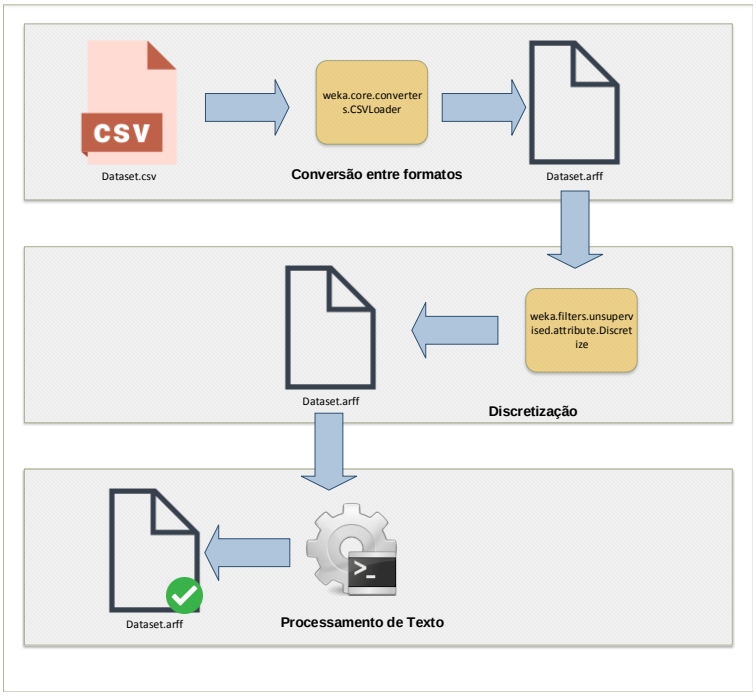

Figura 4.5 – Interface do cálculo das probabilidades entre banco de dados, *Bash Script*, arquivos CSV e a ferramenta matemática Octave.

4.6 Data Mining

Durante a etapa de *Data Mining*, foram testados algoritmos classificadores, a partir disso analisou-se a precisão global (acurácia) de cada um deles, para finalmente selecionar aquele com taxa de acerto aceitável. Considerou-se ainda à seleção do algoritmo dois critérios: a representação dos resultados e o quanto esta informação pode ser interpretada por especialistas e usuários inseridos no domínio.

Quanto a estas finalidades, a Rede Bayesiana se mostra uma importante ferramenta, pelos seguintes aspectos: representação gráfica da relação entre estados; a rede expressa o conhecimento especialista acerca do domínio; e os resultados numéricos (probabilidades) podem

ser visualizados através de gráficos. Nesse sentido, os indutores foram testados quanto às métricas de desempenho com intuito de certificar a confiabilidade do algoritmo bayesiano.

A estratégia utilizada para segmentar a base de dados em conjuntos de treinamento e testes, destinados a estimar precisão e confiabilidade do modelo construído pelo classificador, foi a validação cruzada com k conjuntos estratificada (*stratified k-fold cross-validation*), por ser uma das mais empregadas em mineração de dados (HAN; KAMBER; PEI, 2012); detalhes do funcionamento desta técnica podem ser consultados na Seção 2.9.

Os algoritmos de aprendizado supervisionado (REZENDE, 2005) empregados nesta pesquisa estão disponíveis no *software* Weka. Os classificadores estão divididos de acordo com as seguintes abordagens: árvores de decisão, probabilísticos, baseados em instâncias, baseados em funções e redes neurais artificiais. A Tabela 8 apresenta todos os métodos experimentados, as respectivas abordagens de construção do modelo e a configuração dos parâmetros de execução, sendo estes *default* da ferramenta.

A ferramenta Weka tem algumas características que legitimam a sua utilização nesta pesquisa: i) código-fonte aberto (*open source*), permitindo integração e customização do *software* de acordo com as necessidades do desenvolvedor; ii) isenta de pagamento de licenças; iii) flexibilidade, visto que ela tem sido aplicada nos últimos anos em numerosos domínios; iv) possui componentes voltados ao pré-processamento, mineração de dados e visualização dos resultados; v) possibilita a interação do usuário por meio de interfaces gráficas, extensões codificadas a partir da API (*Application Programming Interface*) em Java ou em *batch* através de *scripts*.

Tabela 8 – Parâmetros configurados nos algoritmos classificadores

Algoritmos	**Parâmetros**
Naive Bayes	Não se aplica
Redes Bayesianas (*Bayesian Network*)	Algoritmo de construção da rede: K2; Máximo número de pais em cada nós 5
K-Nearest Neighbor (KNN)	K=1
Support Vector Machine (SVM)	Função kernel gaussiana: $\exp(-\gamma * \lvert u-v\rvert^2)$; $C = 1$; $\gamma = 1/k$, seja k o número de instâncias.
Multilayer Perceptron	Tipo *Backpropagation*; Função de ativação sigmóide; Número de épocas = 500; Taxa de aprendizado = 0.3; *Momentum Rate* = 0.2.
C4.5	Mínimo de instâncias por folha = 2; Limite de confiança para *prunning* = 25%
Random Tree	Número de iterações = 100; Profundidade máxima da árvore ilimitada; Mínimo de instâncias por folha = 1.
Random Forest	Número de iterações = 100; Profundidade máxima da árvore ilimitada; Mínimo de instâncias por folha = 1.
Classification And Regression Trees (CART)	Número máximo de instâncias em nós terminais = 2

4.7 Considerações Finais

Neste capítulo foram discutidos os procedimentos e ferramentas relacionadas à modelagem da aplicação de mineração de dados educacionais proposta nesta pesquisa. Apresentaram-se os atributos selecionados e os métodos de transformação e processamento de dados, além disso os algoritmos indutores (Tabela 8) foram adotados em todas as análises como critério de comparação e garantia de confiabilidade dos modelos baseados em Redes Bayesianas. O *data set* alcançado após a execução dos mecanismos apresentados foi aplicado em três estudos de casos os quais são discutidos em suas especificidades no Capítulo 5.

5 Resultados

5.1 Considerações Iniciais

Neste capítulo são discutidos os resultados advindos de três Estudos de Caso, detalhados nas seções 5.3 a 5.5. Os resultados, referentes à inferência bayesiana (Seção 3.9), debatidos neste capítulo são variações das probabilidades da variável aleatória antes e após um conjunto de evidências, esta abordagem permite visualizar a tendência em relação ao crescimento e decrescimento de um conjunto de configurações e as classes de alunos.

5.2 Processo de Inferência

O processo de inferência bayesiana utilizada ao longo deste capítulo é apresentado nesta seção. Matematicamente, denotam-se as variáveis aleatórias por X_i e os eventos x_{ji}, sendo o intervalo j, conseguido da discretização dos dados em X_i. Constata-se ainda que a classe de alunos foi representada especialmente por C, onde os valores possíveis (eventos) foram expressados por c_k, onde k varia de 1 até a quantidade de classes t (t=3). Dadas as condições supramencionadas, combinaram-se as evidências factíveis sobre x_{ji} e verificaram-se as saídas de C_k. A Equação 5.1 demonstra o cálculo das variações de C_k, dada probabilidade inicial $P(C_0k)$ e a inferência $P(C_k, x_{j_1 1}, \ldots, x_{j_1 i}, x_{j_n 1}, \ldots, x_{n_1 i})$, a combinação de n variáveis.

$$\Delta P(C_k) = P(C_k, x_{j_1 1}, \ldots, x_{j_1 i}, x_{j_n 1}, \ldots, x_{n_1 i}) - P(C_0 k) \tag{5.1}$$

Por consequência da Equação 5.1, quando $\Delta P(C_k) > 0$, significa um acréscimo na probabilidade da ocorrência de C_k, em outras palavras, aumenta a tendência de o discente pertencer à classe k. Pelo contrário, $\Delta P(C_k) < 0$ remete um decréscimo nas chances de o aluno se tornar do rótulo avaliado. Ressalva-se a inexistência de probabilidade negativa (Axioma 1), conquanto neste trabalho a semântica de variação negativa ou positiva está intrinsecamente correlacionada à diminuição e aumento, respectivamente.

5.3 Estudo de Caso I

O primeiro estudo de caso evidencia os resultados de duas análises. A primeira trata de verificar o impacto do índice acadêmico IEA, por retratar a mensuração do quão eficiente é um aluno na graduação. A outra projeção dos dados diz respeito à descobrir padrões relativos

aos conceitos BOM e INS, essa opção baseia-se na hipótese de que as notas particularmente no primeiro ano de graduação fornecem um prognóstico de como será o desempenho de um aluno ao longo do curso.

5.3.1 Análise de desempenho dos algoritmos

A Tabela 9 apresenta os 9 algoritmos e as métricas usadas: tempos para treinar e testar modelo, acurácia, coeficiente Kappa e raiz do erro médio quadrático (*Root Mean Square Error* – RMSE). Os resultados mostram que a melhor solução foi conseguida através do indutor *Random Forest* cuja acurácia superou 87%, não obstante o algoritmo de Redes Bayesianas revelou precisão global próxima de 86% e tempos aceitáveis para construção e testes do modelo, além disso este algoritmo obteve valor de estatística Kappa igual a 0,6961, considerado um nível substancial de concordância interobservador. A Rede Bayesiana ainda demonstrou baixo RMSE (0.2658) quando comparada as outras soluções. Destaca-se que aplicações nas quais o tempo de processamento é considerado requisito crucial ao domínio, soluções como Multilayer Perceptron e SVM são consideradas inviáveis, embora apresentem boas taxas de acerto.

Tabela 9 – Métricas de desempenho geral dos classificadores

Algoritmos	**Tempo para treino (s)**	**Tempo para teste (s)**	**Acurácia (%)**	**Kappa**	**RMSE**
Naive Bayes	0.31	2.14	78.7736	0.5688	0.353
Bayesian Network	6.77	1.26	85.865	0.6961	0.2658
KNN	0.29	1153.27	83.8041	0.6483	0.3103
SVM	1418.69	1288.21	86.6938	0.6999	0.2978
Multilayer Perceptron	4739.78	3.56	86.2054	0.7048	0.2762
C4.5	1.62	1.44	86.2449	0.6984	0.2775
Random Tree	0.55	1.48	80.5021	0.5924	0.3542
Random Forest	12.31	8.24	87.1102	0.7118	0.2529
CART	236.39	0.63	86.5104	0.7023	0.2646

Diferentemente do método Naive Bayes, por exemplo, que serve como um classificador natural, a rede Bayesiana necessita ter uma boa precisão para ser aplicado ao domínio, e os testes comprovaram a sua eficiência quando comparada às técnicas clássicas. A escolha pela rede Bayesiana é satisfatória aos objetivos desta pesquisa, porquanto, neste experimento: não penalizou tempo de construção e testes do modelo; demonstrou taxa de acerto adequada se confrontada as demais; e agrega conhecimento especialista sobre o domínio em representação gráfica. Diante do exposto, apresentam-se na Subseção 5.3.2, a geração da RB e o conhecimento extraído da própria topologia da rede, por intermédio da inferência probabilística.

5.3.2 Análise da evasão e retenção via Redes Bayesianas

Foram selecionados os 14 atributos mais relevantes, dispostos na Figura 5.13, além da classe (status), de acordo com ganho de informação (HAN; KAMBER; PEI, 2012). Após a redução no número de variáveis, aferiu-se novamente a acurácia do algoritmo Bayesian Network,

apresentando precisão de 83,5%, ratificando a sua robustez. O algoritmo K2 foi empregado para construção da topologia da rede, atingindo-se maior precisão global com o parâmetro de número esperado de pais por nó definido a 5.

A Figura 5.1 mostra a rede Bayesiana resultante, as cores de fundo dos nodos estão diferentes com propósito de agrupar os tipos das variáveis conforme seus significados no domínio de aplicação. Os elementos em laranja, verde e azul são respectivamente, índices acadêmicos, probabilidades de o discente possuir o valor índice acadêmico menor ou igual aos formados, e os percentuais avaliados para determinado conceito (notas). O *status* está em fundo amarelo.

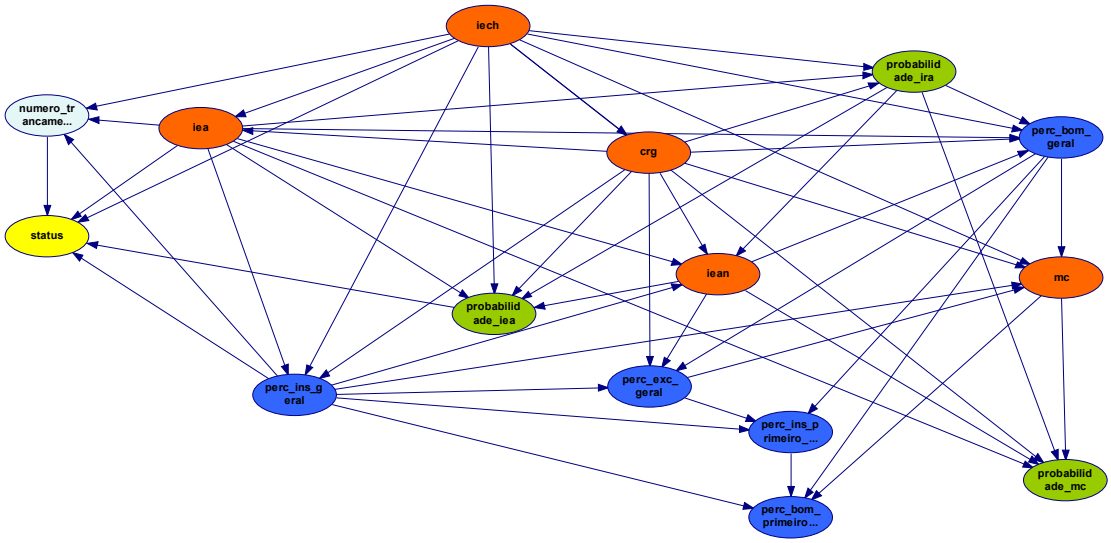

Figura 5.1 – Rede Bayesiana construída para analisar a evasão e retenção no âmbito educacional

A partir da rede apresentada na Figura 5.1, percebe-se que o desempenho estudante depende diretamente do número de trancamentos, do Índice de Eficiência Acadêmico (IEA), da probabilidade em relação a este índice, e do percentual de reprovações durante todo o curso. Com efeito, essas relações fazem sentido, uma vez que o aumento do número de trancamentos e baixo IEA implica, consequentemente, em uma elevação na probabilidade do índice (probabilidade_iea); isto é, o discente fica abaixo da expectativa de conclusão dos estudos em tempo hábil, ou ainda de concluir do curso. Vale destacar que, a partir da topologia apresentada, outras interpretações são possíveis e válidas no contexto pesquisado.

Para a inferência Bayesiana, escolheram-se os atributos IEA (iea) e número de trancamentos (numero_trancamento) visando quantificar o efeito inverso entre o índice e a interrupção da matrícula. Os estados, de todas as variáveis, foram conseguidos por intermédio da discretização com distribuição uniforme de frequência, sendo que *numero_trancamento* e *iea* tiveram, respectivamente, três e quatro intervalos para conversão de um domínio contínuo em discreto, essas quantidades foram determinadas após análise da frequência das amostras em determinado intervalo.

As Figuras 5.2 a 5.4 representam os gráficos das variações de probabilidade, compara-

tivamente por classes de discentes, número de trancamento e evolução do IEA.

Depreende-se do gráfico da Figura 5.2 que alunos sem trancamento possuem mais chances de graduar-se à medida que o IEA aumenta. O intervalo do índice de eficiência acadêmica entre 3,905 e 6,435, revela um importante diagnóstico – nesta faixa acentuam-se as possibilidades de diplomação (7,6%), atingindo o valor máximo (28,2%) com o índice variando de 7,975 até 10. Por sua vez, o aluno que conseguiu IEA entre 0 e 3,95, incrementa em aproximadamente 11,5% as possibilidades de abandonar os estudos. Portanto, o cenário desejável ocorre a um IEA superior a 7,975, pois neste caso há acréscimo de quase 30% de o aluno formar-se e há redução de 25,9% para evasão e 6,1% a retenção.

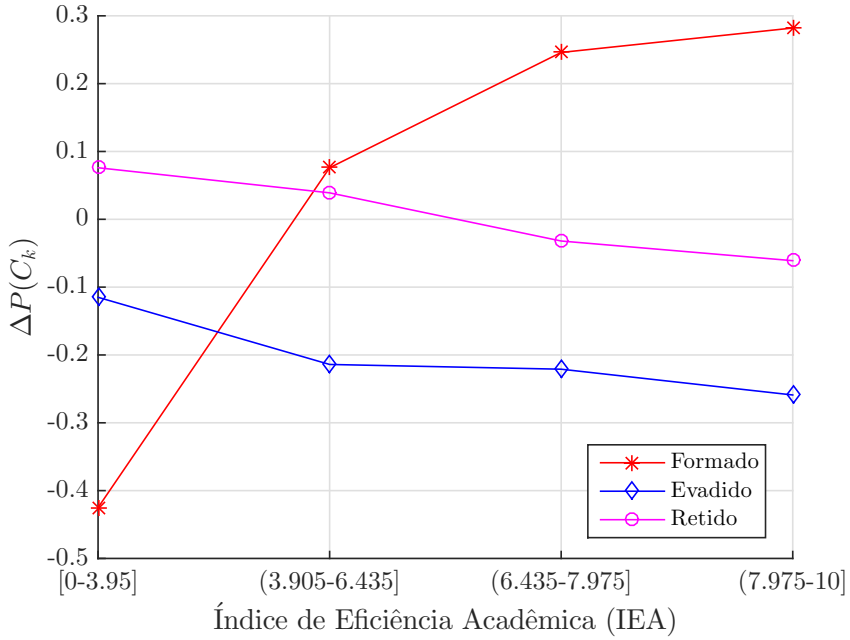

Figura 5.2 – Variação da probabilidade de o discente, sem trancamento de matrícula, pertencer a uma das classes de acordo com a variação do Índice de Eficiência Acadêmica

Conforme ilustra o gráfico da Figura 5.3, alunos com exatamente 1 trancamento têm suas possibilidades de formatura significativamente reduzidas (-57,2%), caso possuam IEA variando de 0 a 3,95, após este valor acentuam-se as chances de diplomação (em torno de 9,2%), contudo sem grande expressividade se comparado ao caso anterior, isto é, sem trancamento de matrícula. No segundo intervalo (representado por (3.905-6.435]) visualiza-se o primeiro decréscimo (-11.5) da classe de alunos evadidos, atingindo o valor mínimo onde as chances de formatura são máximas, ou seja, com variação do IEA de 7,975 até 10.

Após 1 trancamento, há um aumento 64,5% para evasão, caso o aluno possua até 3,95% de índice de eficiência acadêmica e atenua-se, ao passo que este indicador cresce, até 22,8%, de acordo com a Figura 5.4. A variação da probabilidade à formatura não atingiu em nenhum momento valor positivo, isto é, na prática a interrupção da matrícula por mais de uma vez diminuem drasticamente as chances de o aluno conseguir a diplomação, elevando consideravelmente as chances da desistência.

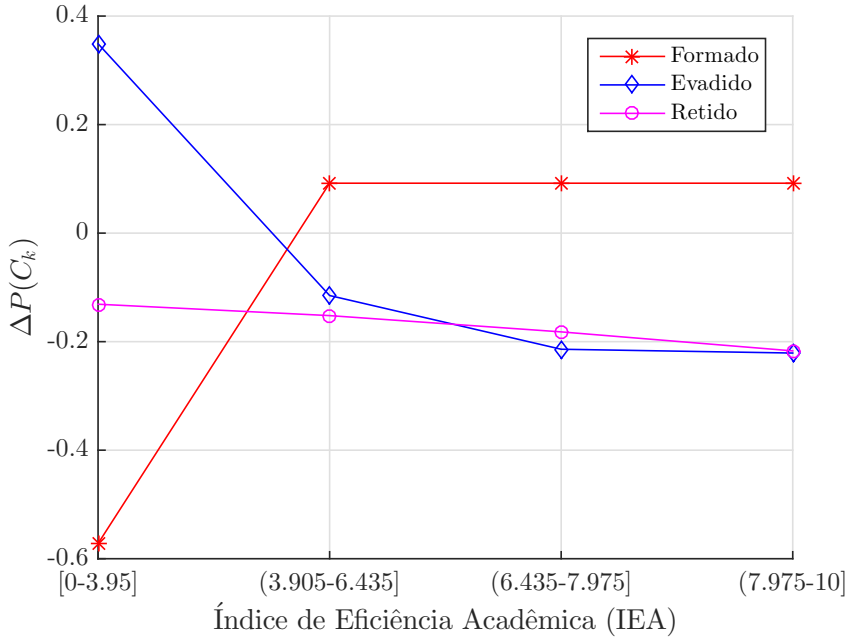

Figura 5.3 – Variação da probabilidade de o discente, com 1 trancamento de matrícula, pertencer a uma das classes de acordo com a variação do Índice de Eficiência Acadêmica

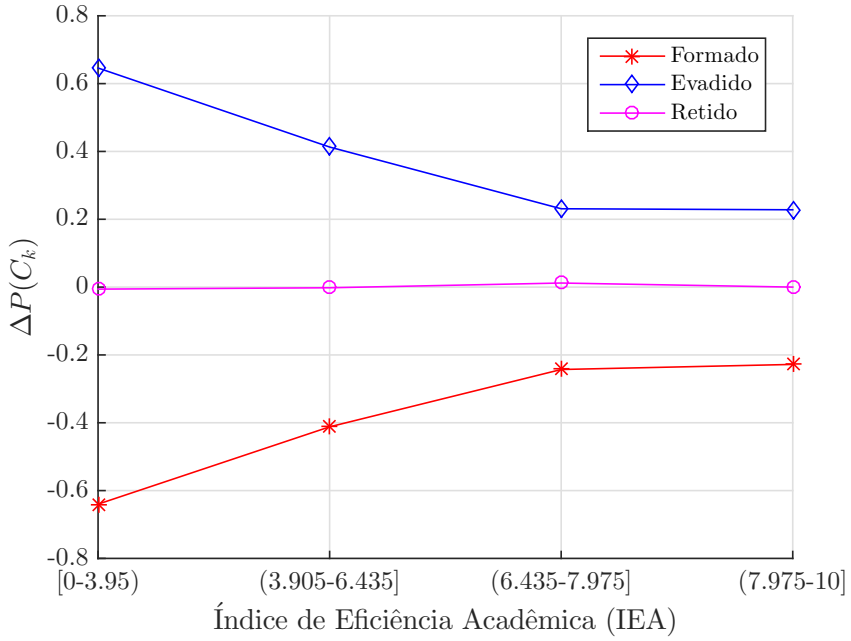

Figura 5.4 – Variação da probabilidade de o discente, com número de trancamentos de matrícula superior a 1, pertencer a uma das classes de acordo com a variação do Índice de Eficiência Acadêmica

As Figuras 5.5 a 5.8 representam os gráficos das variações das probabilidades para classe de alunos, consoante ao percentual de conceitos do tipo BOM e INS no decurso do primeiro ano de graduação. O eixo das abscissas (horizontal) estão as categorias discretizadas da variável *perc_conceito_bom*, enquanto no eixo ordenado (vertical) estão os valores probabilísticos. Salienta-se que os gráficos evidenciam os resultados da inferência de um *status* dadas as evidências das possíveis taxas de reprovação e conceitos do tipo BOM.

O gráfico da Figura 5.5 aponta um acréscimo superior a 20% nas chances de o discente graduar-se, caso haja aprovação em todas as disciplinas matriculadas no decorrer do primeiro ano de estudo, sem realização de trancamento de matrícula. Tais considerações apontam maior variação de probabilidade (22,5%), caso o discente supere 51,985% de conceitos BOM em relação as outras notas obtidas. Convém salientar, o aproveitamento total no decurso do primeiro ano e a não interrupção da matrícula acrescentam as chances de o discente concluir os seus estudos. Ainda pertencente a essa situação, examina-se que o trancamento de matrícula, as chances de o discente concluir os estudos são variações negativas, ou seja, diminuem à proporção que o aluno obtém menos conceitos BOM, amplificando as possibilidades de evasão, as quais são maiores (16,3%) com até 20,02% de notas desse tipo.

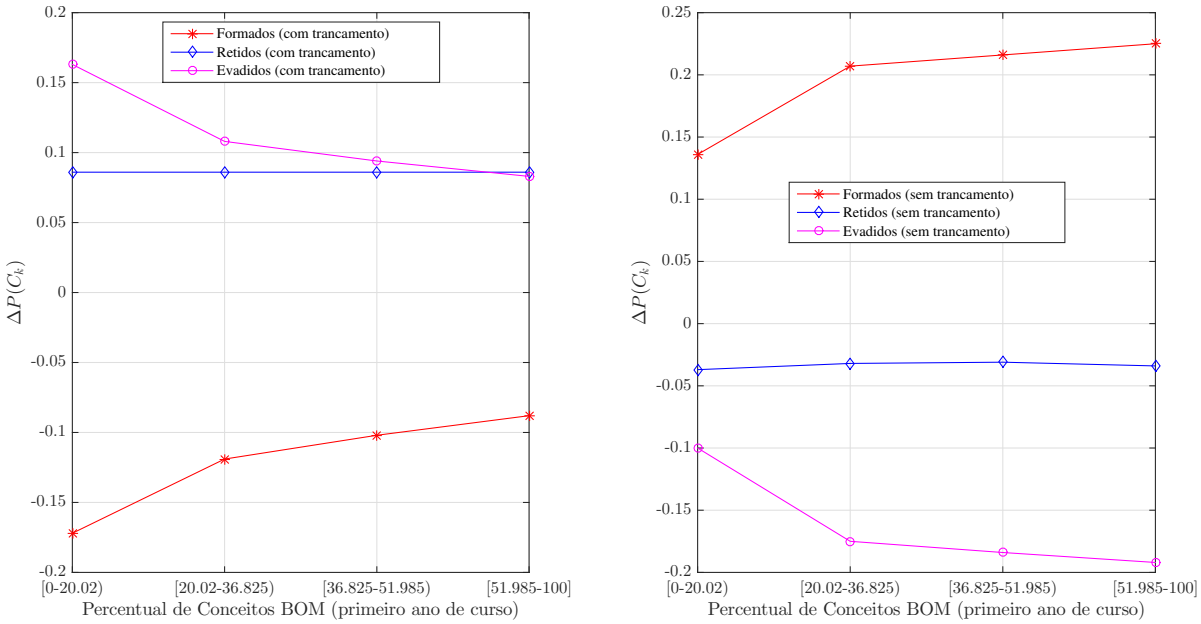

Figura 5.5 – Variação da probabilidade de o discente, sem reprovações no primeiro ano de curso, pertencer a uma determinada classe (por situação da matrícula)

Ao percentual de INS de até 13,8% (Figura 5.6), atenta-se que o aumento de 16% nas chances de formatura se configuram em uma faixa acima de 51% no percentual de notas do tipo BOM. Ao levar em consideração esta perspectiva, é importante situar que sob essas mesmas condições, todavia sem reprovações, a alteração probabilística supera o valor citado (22,5%). Um outro aspecto refere-se à retenção, a qual não demonstra muita variação (aproximadamente 3% de acréscimo), porém é maior se confrontado ao cenário sem reprovações.

A partir do gráfico da Figura 5.7, infere-se que reprovações entre 13,8% e 38,97% acentuam a probabilidade de evasão (maior variação de 33%) a menores taxas de conceitos BOM, ao passo que reduz-se à medida que este percentual cresce. Neste cenário, as possibilidades de formatura sofrem decréscimos (todas negativas) em situações de trancamento e possuem a maior variação próxima de 4,9% (sem trancamento de matrícula). Um exemplo claro disso também pode ser visualizado na Figura 5.8, onde há a relação entre a taxa de conceitos BOM e a variação da probabilidade para os casos de reprovações que superam 38,97%. Nesse sentido,

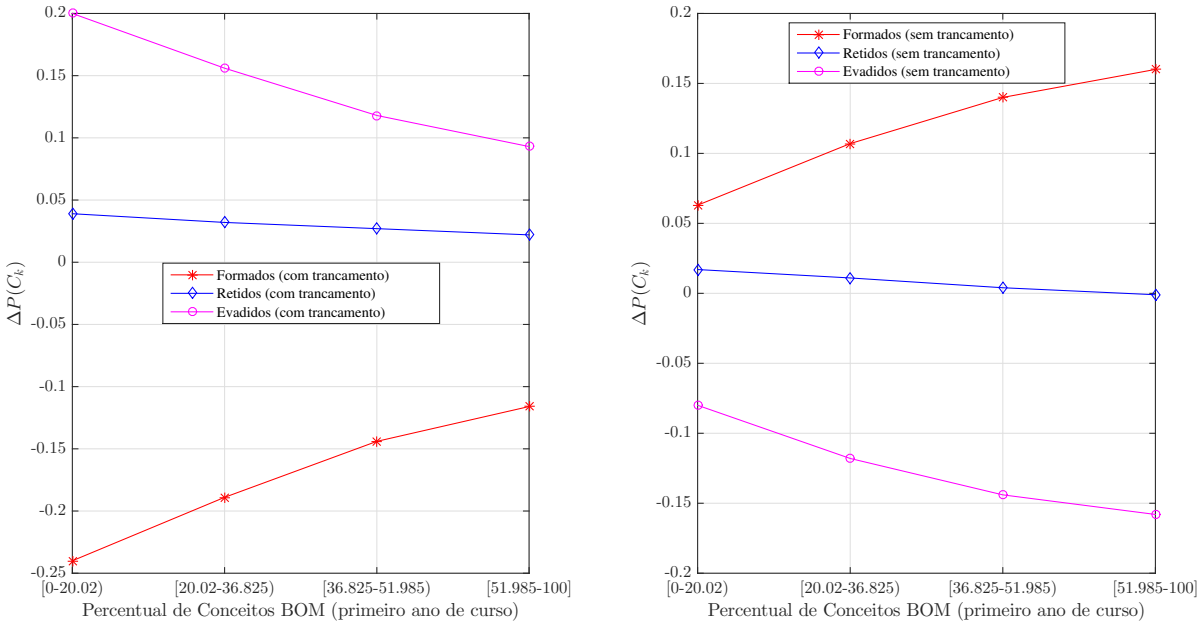

Figura 5.6 – Variação da probabilidade de o discente, 13,8% de conceitos insuficientes no primeiro ano de curso, pertencer a um determinado *status* (por situação da matrícula)

as chances de evasão e retenção podem ser detectadas ainda no primeiro ano de estudo e estão atreladas a disparidades entre boas notas e reprovações, somando-se a isso a interrupção da matrícula em componentes curriculares (trancamento).

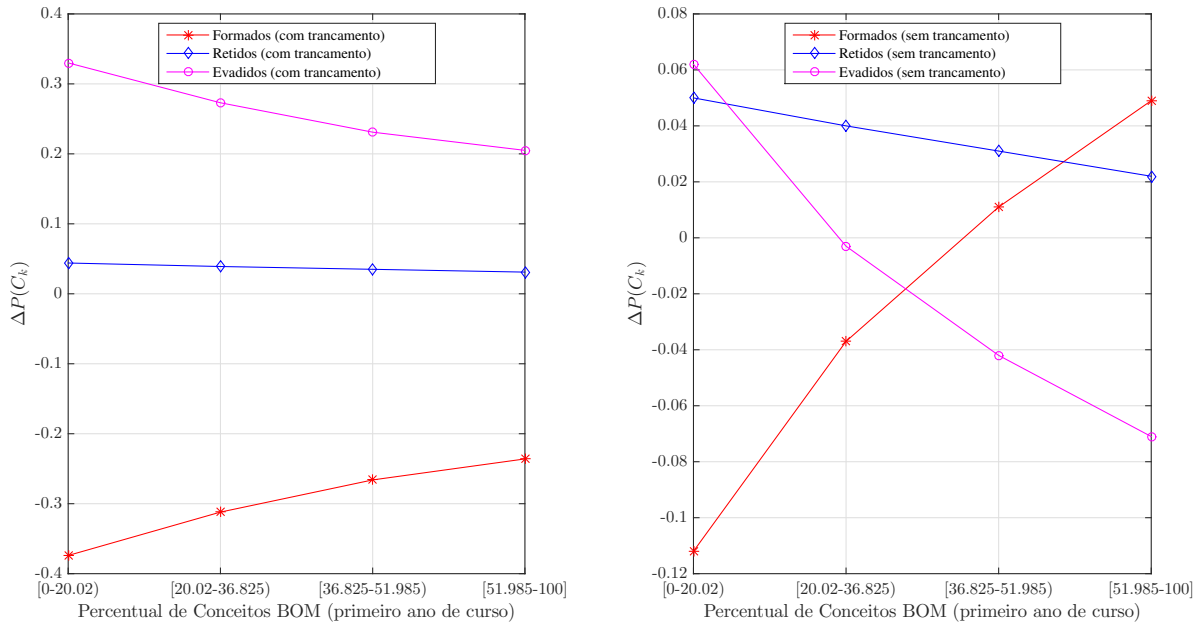

Figura 5.7 – Variação da probabilidade de o discente pertencer a um *status*, com percentual de conceitos insuficientes no primeiro ano do curso entre 13,8% e 38,97% (por situação da matrícula)

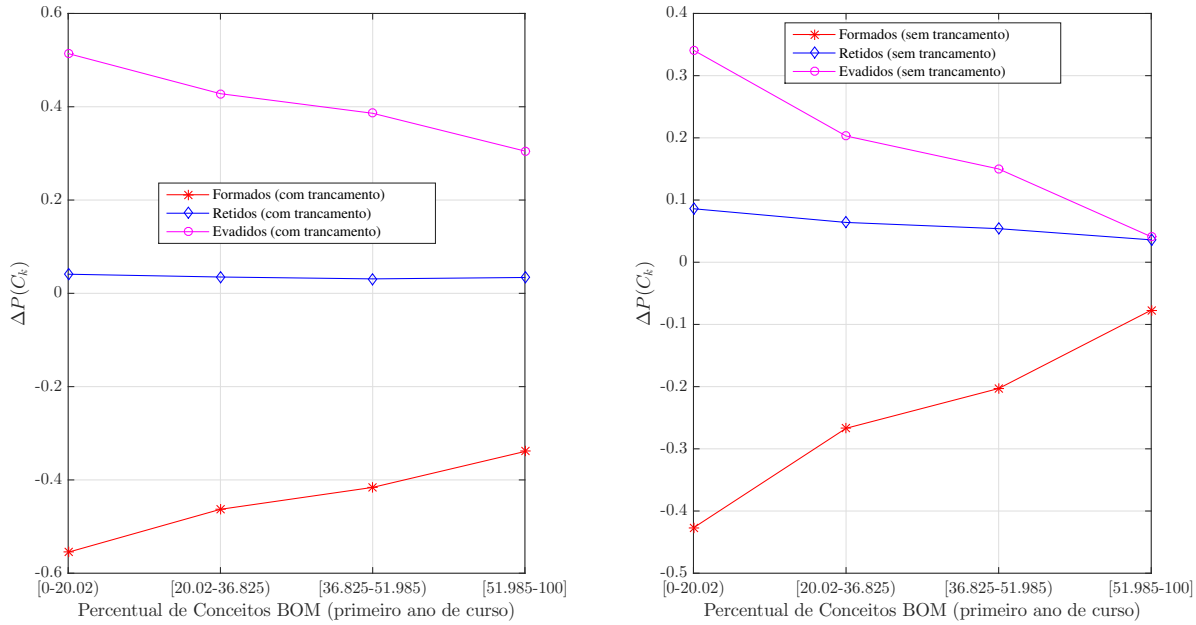

Figura 5.8 – Variação da probabilidade de o discente pertencer a um *status*, com percentual de conceitos insuficientes no primeiro ano do curso superior a 38,97% (por situação da matrícula)

5.4 Estudo de Caso II

Neste estudo de caso analise-se o nexo entre atributos *idade*, *turno*, indicadores acadêmicos (*crg* e *mc*), ingresso (*forma_ingresso*) no curso e a situação do aluno (*status*) em referência ao prosseguimento dos estudos. Para esta finalidade emprega-se a mesma base de dados, todavia reduzindo-se o número de atributos na composição da rede, já que muitos nós dificultam a interpretação dos resultados.

5.4.1 Avaliação da Rede Bayesiana

A rede conseguida, incluindo-se 9 atributos, com auxílio do algoritmo K2 é mostrada na Figura 5.9. Depreende-se do grafo a dependência entre os atributos *turno*, *interior* e *mc* da variável *idade*. Outro fator observado diz respeito a associação da forma de ingresso (*forma_ingresso*), número de trancamentos e número de vínculos do *campus* onde o discente está matriculado. Estas relações são contundentes, alunos mais novos estão vinculados a turnos da manhã ou da tarde, enquanto que aqueles com mais idade e, em geral, desempenham atividades laborativas tendem a estudarem em horários noturnos. Em razão dessa realidade e ainda inspecionando a topologia da rede, o turno interfere em questões relacionadas ao trancamento, consequentemente, isto afeta o rendimento geral e a situação (*status*) na universidade.

As Figuras 5.10 a 5.12 apresentam os gráficos das variações de probabilidades à inferência bayesiana segundo o *campus* e as principais formas de ingresso existentes durante a história da universidade. Esta opção é uma parte do conhecimento passível de extração da rede, fato que releva a entrada na instituição e seus mais variados *campi*.

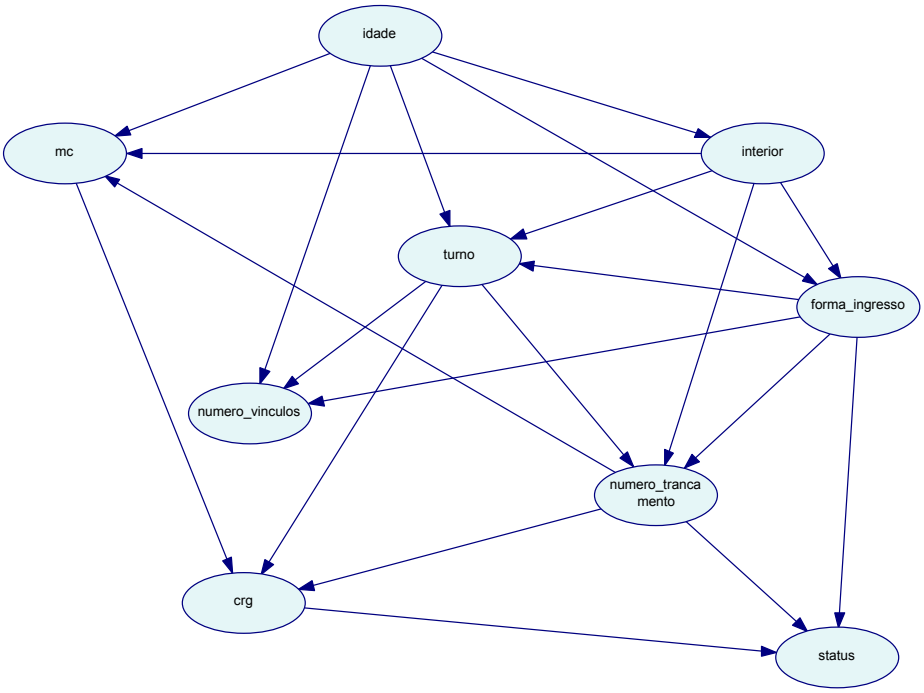

Figura 5.9 – Rede Bayesiana construída a partir de atributos *idade*, *turno*, *numero_trancamento*, *interior*, *numero_vinculos*, indicadores acadêmicos e forma de ingresso

Graficamente observa-se por meio da Figura 5.10 que os alunos da capital ingressantes pelos antigos processos seletivos (PS) possuíam um aumento de 8,4% de chances à formatura, porém vale destacar que esta modalidade de ingresso na instituição foi extinta, dessa forma, nos próximos anos não mais haverá alunos com matrículas ativas que ingressaram na IES por intermédio do PS. Outro ponto relevante refere-se ao Plano Nacional de Formação dos Professores da Educação Básica (PARFOR), os alunos dos *campi* do interior possuem 5,4% a mais de probabilidade em formarem-se, quanto aos discentes do capital há redução (-3,4%) neste número, o qual representa o segundo menor.

A maior redução probabilística se configura, aos discentes da capital, ingressantes pelo Exame Nacional do Ensino Médio (ENEM), aproximadamente 25,8%, todavia esta realidade deve ser acompanhada continuamente pelos gestores porque o ENEM está implantado há poucos anos na UFPA e com mais dados históricos será possível otimizar a análise acerca desta seleção. Cabe avaliar que, para todas as formas de ingresso, os discentes dos *campi* do interior são mais propensos à formatura.

O gráfico ilustrado na Figura 5.11 evidencia a propensão à evasão para discentes da capital ingressantes pelas mobilidades interna (MOBIN) e externa (MOBEX), cujas variações em probabilidades são respectivamente 18,1% e 21,1%. Ambas as formas de entrada à universidade destinam-se a alunos externos (MOBEX) ou internos (MOBIN) à instituição, os quais pretendem trocar de cursos, diante desse contexto é válido argumentar que tais estudantes já desistiram de pelo menos uma graduação e ficam mais vulneráveis a ocorrência deste fato novamente, por motivos que devem ser minunciosamente analisados dentro do contexto acadêmico. Fato similar

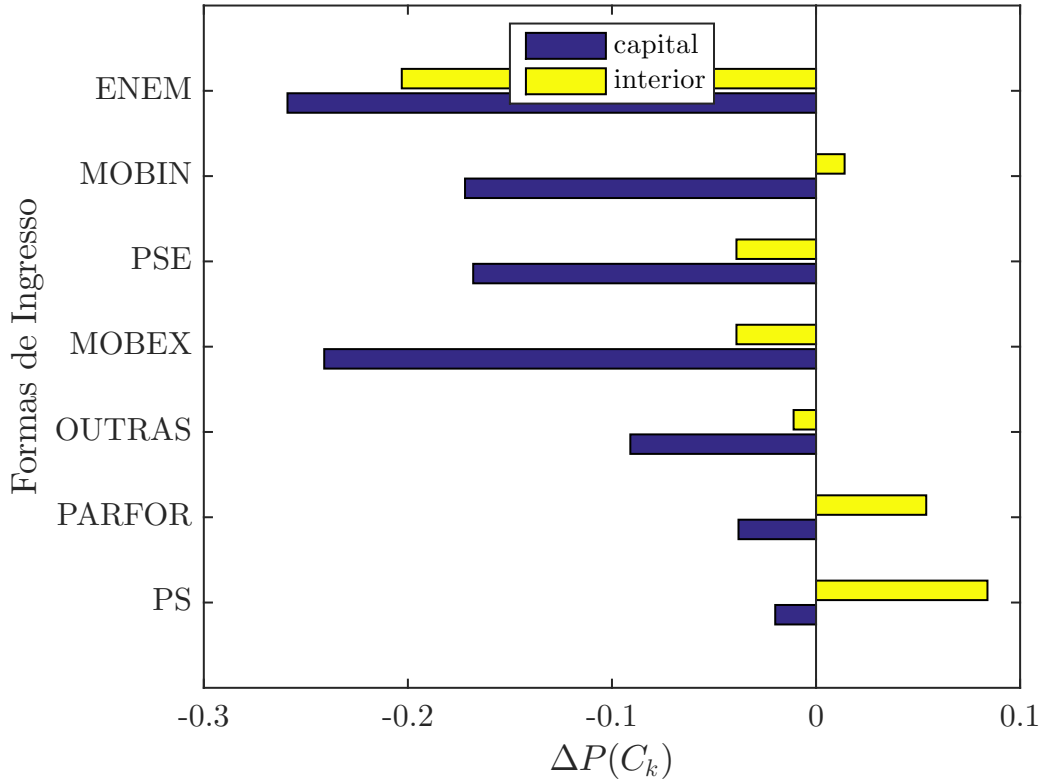

Figura 5.10 – Variação da probabilidade concernente à formatura de acordo com o *campus* e forma de ingresso

ocorre com alunos do Processo Seletivo Especial (PSE), todavia com menor frequência (6,5%). Acrescente-se, ainda, que os resultados ratificam a discussão anterior acerca do PARFOR, ou seja, os estudantes, independente do município, tornam-se menos propensos à abandonarem os estudos (redução de mais de 20%). A retenção, por sua vez, é um problema mais evidente, a partir do gráfico da Figura 5.12, aos ingressantes da capital pelo ENEM (32,9%), acompanhado do PSE (10,3%) e PARFOR (23,9%).

5.5 Estudo de Caso III

Neste estudo de caso, expõe-se a conexão das problemáticas da evasão e retenção com os cursos de ciências exatas cujo número de reprovações são elevados. Em virtude disso, novos atributos foram empregues nos experimentos, além daqueles organizados na Tabela 7, totalizando 37 variáveis em 8.839 amostras.

A Tabela 10 mostra o novo conjunto de dados, os atributos 31 a 33 expressam os totais de reprovações nas disciplinas de cálculo, a saber I, II e III respectivamente. O fato motivador à seleção dessas variáveis vincula-se a conjuntura de que muitos alunos desistem ou excedem o período ideal para conclusão do curso por não alcançarem êxito nessas disciplinas iniciais e básicas da formação de muitas carreiras acadêmicas. Vale destacar que este estudo é expansível a outras áreas de conhecimento, decorrente da necessidade de gestores de faculdades ou institutos.

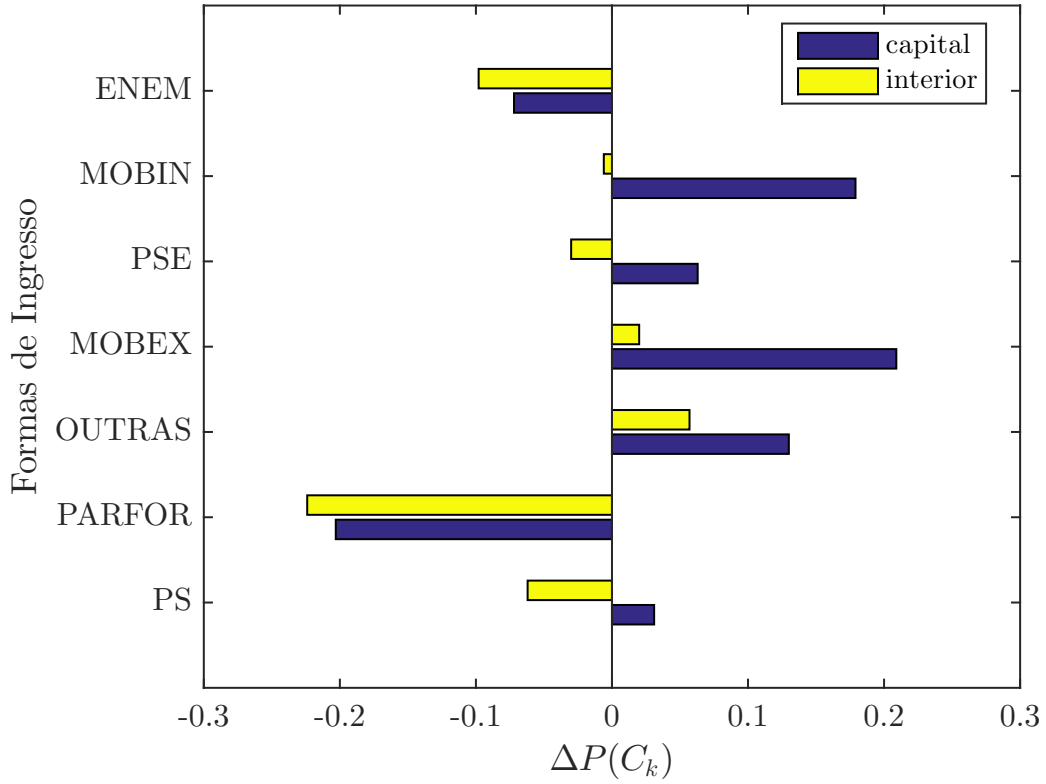

Figura 5.11 – Variação da probabilidade concernente à evasão de acordo com o *campus* e forma de ingresso

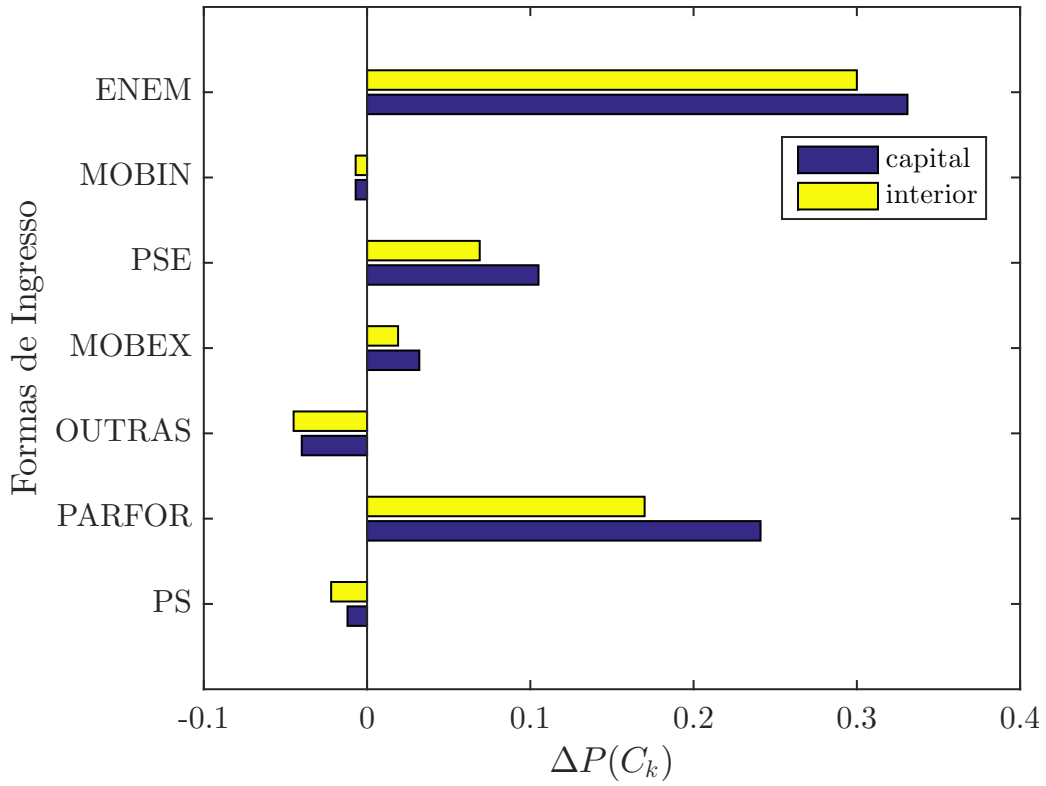

Figura 5.12 – Variação da probabilidade concernente à retenção de acordo com o *campus* e forma de ingresso

Os atributos 34 e 35 denotam a quantidade de reprovações em disciplinas com maior e menor índice de reprovações no curso avaliado, respectivamente. Esta variáveis, em tese,

indicariam o desempenho do aluno em componentes curriculares de acordo com o seu grau de complexidade. No decurso do pré-processamento foi imposta uma restrição na qual julgaram-se imprescindíveis as disciplinas que tiveram mais de 300 reprovações. O atributo 36 retrata a área de conhecimento onde o curso está inserido, para os casos analisados são permitidas engenharias, matemática e estatística e outras menos específicas.

Tabela 10 – Variáveis selecionadas à pesquisa

Número	Atributo	Descrição	Domínio
31	repro_calculoi	Número de reprovações em cálculo I	inteiros
32	reprov_calculoii	Número de reprovações em cálculo II	inteiros
33	reprov_calculoiii	Número de reprovações em cálculo III	inteiros
34	reprov_disciplina_facil	Número de reprovações na disciplina com menor índice de reprovações no curso	inteiros
35	reprov_disciplina_dificil	Número de reprovações na disciplina com maior índice de reprovações no curso	inteiros
36	area_conhecimento	Área do curso de exatas	"MATEMATICA_ESTATISTICA", "ENGENHARIAS" e "OUTRAS"

5.5.1 Análise de desempenho dos algoritmos

O desempenho dos algoritmos foram avaliados, similarmente ao que foi efetuado na Subseção 5.3.1. Os resultados apontam maior precisão do classificador SVM (85,1567%), todavia repetidamente a Rede Bayesiana revelou precisão global admissível (83,5728%) ao domínio, embora seja inferior aquela do Estudo de Caso I (85,865%). Em face disso os números dispostos na Tabela 12 seguem análises correlatas às debatidas na Subseção 5.3.1, o que deve ser evidenciado é a capacidade e validação das redes bayesianas ao domínio, frente aos demais algoritmos clássicos empregados na literatura.

Tabela 12 – Métricas de desempenho geral dos classificadores para o novo *dataset*

Algoritmos	Tempo para treino (s)	Tempo para teste (s)	Acurácia (%)	Kappa
Naive Bayes	0.41	0.54	75.5628	0.4954
Bayesian Network	2.04	0.53	83.5728	0.619
KNN	0.4	18.49	80.5295	0.5582
SVM	15.51	13.63	85.1567	0.6314
C4.5	0.84	0.45	82.8827	0.5916
Multilayer Perceptron	1273.54	1.1	83.7764	0.6269
Random Tree	0.69	0.39	76.0493	0.4644
Random Forest	2.13	0.89	84.6476	0.6155
CART	39.75	0.23	83.9348	0.6082

5.5.2 Análise da evasão e retenção via Redes Bayesianas

A topologia da rede, mostrada na Figura 5.13, foi conseguida por meio do método K2, com o número máximo de pais igual a 3 e a esta configuração o algoritmo *Bayesian Network* atingiu precisão global superior a 82%. Foram selecionadas 11 variáveis, tendo como objetivo a simplificação da rede, particularmente voltada ao entendimento do domínio estudado. Além das variáveis específicas a este Estudo de Caso, organizadas na Tabela 10 e indicadas em fundo de cor amarela na rede; inseriram-se o *crg*, *probabilidade_ira*, *idade* e *numero_trancamento*. Essa predileção se deu pelo motivo de empenhar-se em conseguir uma combinação de efeitos de indicadores acadêmicos individuais e comparativos; e atributos concernentes a performance nas disciplinas de cálculo e daquelas associadas a um grau de dificuldade para aprovação.

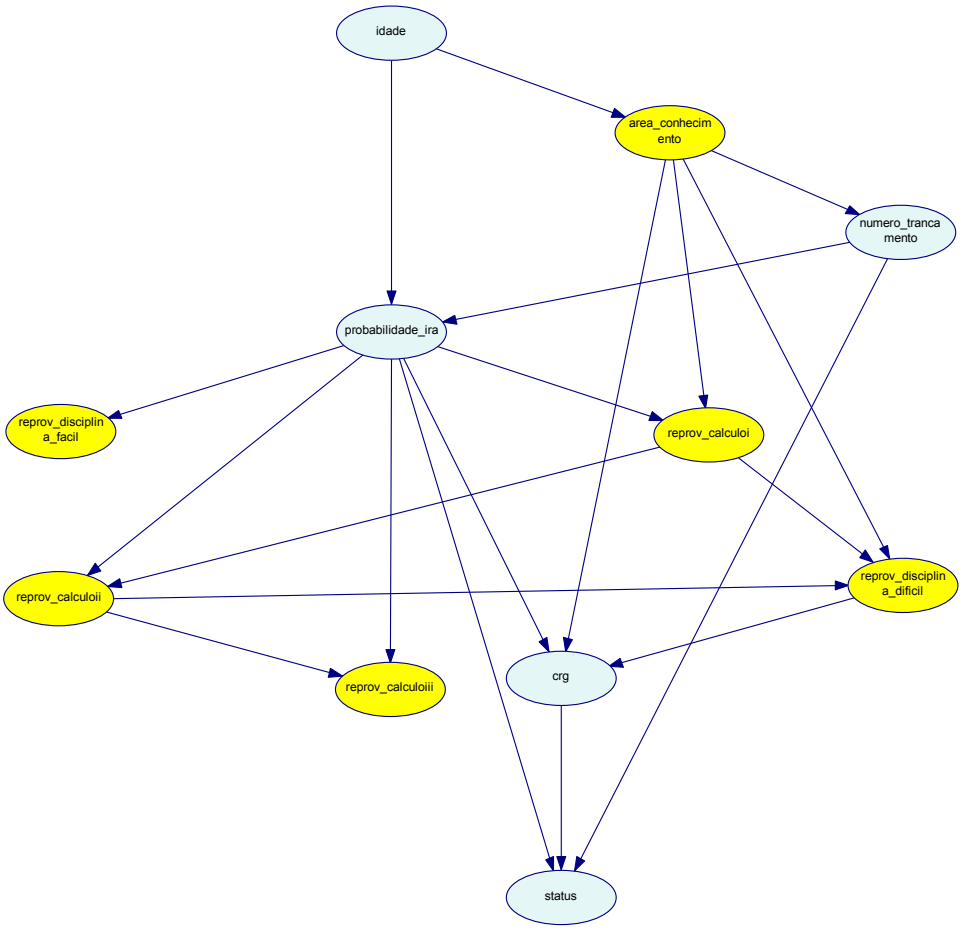

Figura 5.13 – Rede Bayesiana construída pelo método K2 tendo como base os alunos de Ciências Exatas

A topologia da rede (Figura 5.13) oportuniza a interpretação acerca da ligação entre atributos e domínio de aplicação, assim nota-se, por exemplo, que a faixa etária pode ser influente na escolha da área de conhecimento na qual o estudante ingressa na universidade. Observa-se ainda a influência da disciplina de Cálculo I nas subsequentes II e III, ratificando o que se vivencia na prática em muitos cursos de ciências exatas, por vezes discentes sem êxitos em componentes curriculares básicas à graduação pretendida tornam-se incapazes de prosseguir

nos estudos. O exame da rede em seus nós e arestas (ou ramos) concede outras deduções que, dependendo do conhecimento do especialista, podem ter fundamento no contexto estudado.

Para a inferência bayesiana foram testados todos os atributos relativos à reprovação em disciplinas juntamente às áreas de conhecimento, a saber: Matemática e Estatística; Engenharias e; outras, englobando física, química e graduações que possuem componentes de cálculos. Os resultados conseguidos são similares quanto à interpretação, possuindo obviamente pequenas variações numéricas, dessa forma não serão debatidos todos os gráficos nesta subseção. Por conseguinte, a reprovação é debatida sob o enfoque de Cálculo I.

As Figuras 5.14 a 5.16 ilustram os gráficos, de acordo com a área de conhecimento das ciências exatas, das variações das probabilidades sob o número de reprovações em Cálculo I. De modo geral, infere-se que o crescimento das chances de evasão e retenção crescem na mesma proporção do acréscimo da quantidade de reprovações em Cálculo I.

A Figura 5.14 possui o gráfico da variação de probabilidade após a inferência, dadas condições de reprovações em Cálculo I e ingresso nas áreas de Matemática ou Estatística. Nota-se que após duas reprovações, tanto as chances de evasão quanto retenção tornam-se acima de 10%. Por sua vez, sob as mesmas condições, a possibilidade de formatura tende fortemente a reduzir-se (-19,1%).

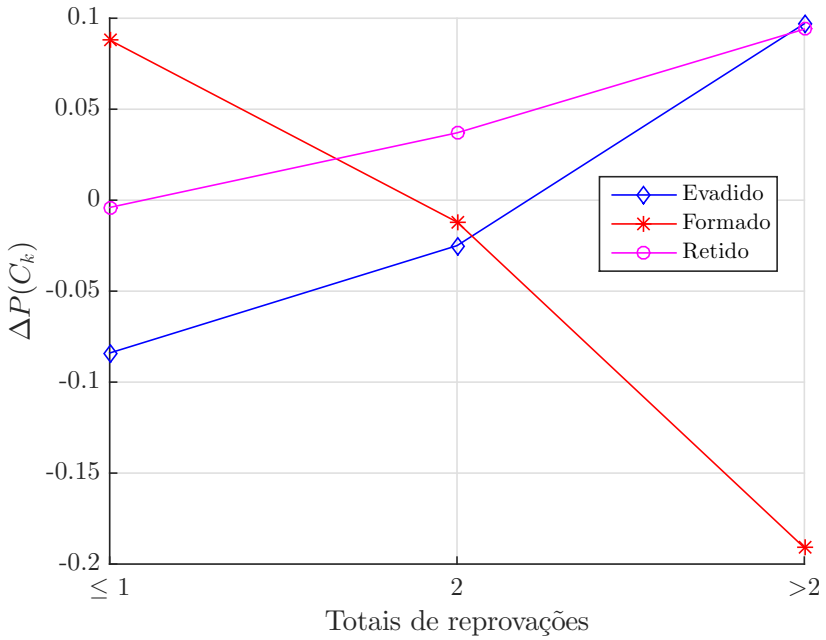

Figura 5.14 – Variação de probabilidade concernente a reprovações na disciplina de Cálculo I, dado que o aluno e de Matemática ou Estatística

O comportamento gráfico aos alunos das engenharias é similar ao apresentado anteriormente, porém a evasão torna-se consideravelmente mais provável (24,7%) de acontecer se comparada ao caso de Matemática e Estatística, já a retenção é pouco visualizada (6%) para estudantes das engenharias. A condição de formatura apresenta, por consequência, o maior decréscimo, aproximadamente 30,1%. Conforme a Figura 5.16 à proporção que eleva a quantidade

de reprovações, as demais áreas das ciências exatas demonstram crescimento máximo de 20,5% à evasão e decrescimento próximo a 25,5% na possibilidade de conclusão da graduação.

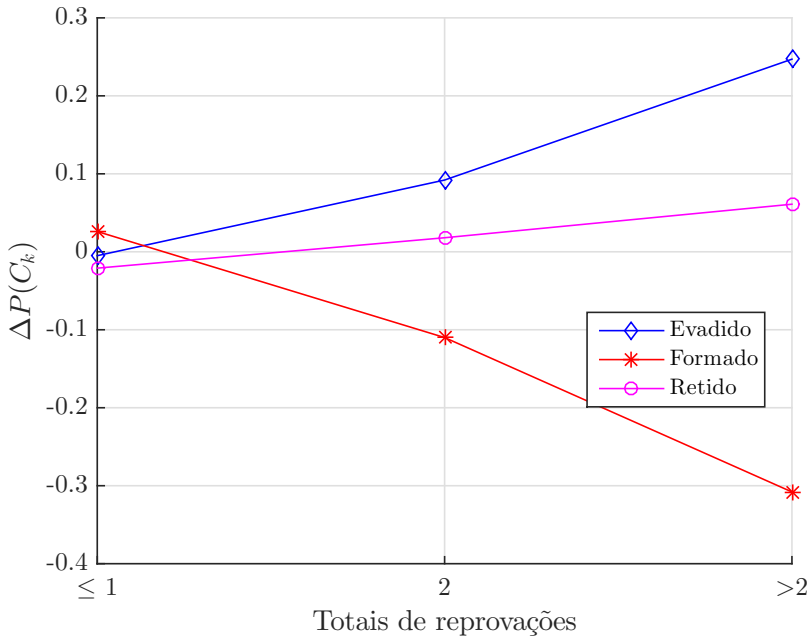

Figura 5.15 – Variação de probabilidade concernente a reprovações na disciplina de Cálculo I, dado que o aluno e estudante de Engenharia

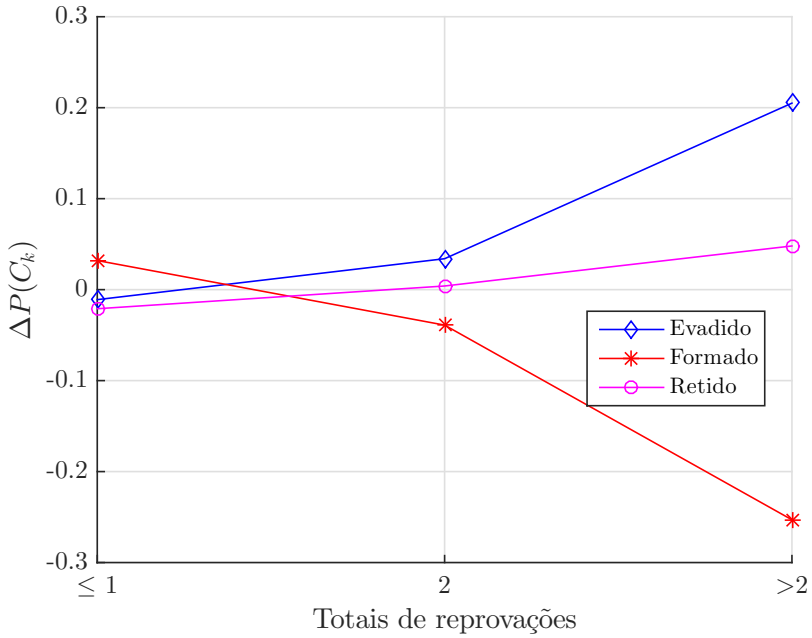

Figura 5.16 – Variação de probabilidade concernente a reprovações na disciplina de Cálculo I, dado que o aluno e estudante de outras areas de ciencias exatas

5.6 Considerações Finais

Este capítulo foi essencial à exposição de três estudos de caso nos quais são abordados os padrões extraídos após o processo de inferência bayesiana, além disso, a eficiência do método

bayesiano foi atestada em comparação aos algoritmos clássicos do aprendizado supervisionado. Em suma, os resultados foram suficientes aos desígnios deste trabalho, pois foi permitido compreender o relacionamento entre a problemática e as variáveis, sob a premissa do uso de uma metodologia objetiva, eficiente e de fácil entendimento ao usuário inserido no domínio.

6 Conclusões

Esta pesquisa proporcionou o entendimento a respeito das problemáticas da evasão e retenção. Os referenciais teóricos correlatos desenvolvidos na academia nos últimos anos foram imprescindíveis às buscas de métodos e ferramentas hábeis à avaliação dos percalços ligados à formação de alunos da graduação do ensino presencial da Universidade Federal do Pará. A metodologia empregada aos propósitos estabelecidos foi o processo de *Knowledge Discovery in Database*, tendo um grande volume de dados acadêmicos registrado nas bases de dados acadêmicas da instituição, dessa forma pôde-se depreender novos padrões relativos aos fatores dificultadores à conclusão dos estudos.

Em adição, primou-se pela capacidade de expressar esses padrões de maneira concisa em uma estrutura onde fosse possível também aplicar e, concomitantemente, extrair conhecimento de especialistas. Diante do exposto, priorizou-se a escolha da Rede Bayesiana em razão de atender a essas exigências, além disso é permitida uma representação hierárquica entre variáveis e atribuição de probabilidades a determinado evento por meio de evidências em relação a algo que se sabe (ou pretende descobrir) sobre o domínio.

A contribuição mais significativa deste trabalho ao domínio de aplicação consistiu na utilidade dos padrões obtidos, importantes ao processo de tomada de decisão por parte de gestores e pesquisadores em educação, sobretudo aqueles empenhados no entendimento na evasão e retenção. Em adição, a aplicação de rede bayesiana em cenários educacionais, especificamente no caso da UFPA, evidenciou a flexibilidade e robustez dessa metodologia, fato que permitirá uma gama de possibilidades voltadas ao aprimoramento e expansão deste estudo.

Assim, os resultados conseguidos em três estudos de caso, foram bastantes satisfatórios e as metas inicialmente atingidas. Em todos os casos, a Rede Bayesiana apresentou acurácia superior a 80%, evidenciando a sua robustez à capacidade de resolução de problemas em diversos domínios. Outro ponto que cabe relevância refere-se as topologias aprendidas pelo algoritmo K2, as quais mostraram-se contundentes quanto à semântica das observações vivenciadas no cotidiano de muitos alunos inclinados ao abandono do curso ou à extrapolação do tempo ideal da diplomação.

Contudo, é válido destacar as dificuldades existentes à aquisição e seleção de atributos, por alguns motivos: certos dados, em especial, sócio-econômicos quando registrados, não o são de forma centralizada, em geral, estão dispersos em variadas bases de dados; mecanismo de coleta sem validação ou migração provenientes de sistemas legados, acabam por vezes, tornando alguns dados inconsistentes; e o próprio questionamento discutido na área de Mineração

de Dados Educacionais acerca de quais variáveis ou características merecem relevância para estudo. Embora os resultados supramencionados tenham sidos de grande valor à academia, é basilar a sua expansão através de novas pesquisas, dentro de outros contextos, a fim de consolidar respostas a novos questionamentos, certamente, ainda presentes relativamente a evasão e retenção.

A Universidade pode empregar a metodologia e os resultados conseguidos nesta dissertação para elaboração de políticas de combate à evasão ou retenção, visando principalmente a implementação de mecanismos que monitorem continuamente alunos em condições adversas à formatura. Assim, a instituição poderá oferecer suporte aos discentes com potenciais problemas a formação acadêmica durante o período de graduação.

6.1 Trabalhos Futuros

Para trabalhos futuros, planeja-se o desenvolvimento de um sistema *web* que suporte a criação de redes bayesianas por diversos métodos de aprendizado de estruturas e inferência em tempo real com interface de exportação de padrões extraídos. Outros requisitos indispensáveis ao projeto estão listados a seguir:

- Integração com o sistema acadêmico da UFPA;
- Criação de uma interface orientada à estudos de casos, na qual seja possível o gestor optar pelas relações entre variáveis, classes e projeções (curso, disciplina, *campus* entre outras possibilidades);
- Desenvolvimento de uma arquitetura comum às instituições que utilizam o mesmo sistema da UFPA, desse modo visando a interação com outras IES a partir de poucas ou nenhuma modificação em código-fonte;
- Sistema escalável em relação ao volume de dados, ou seja, compute poucas dezenas até centenas de milhares de amostras com tempo de processamento admissível.

6.2 Publicações

Os resultados desta pesquisa foram publicados nos artigos:

- Do Couto, D.C, Santana, A.L. Mineração de Dados Educacionais Aplicada à Identificação de Variáveis Associadas à Evasão e Retenção. Aceito para apresentação no Congresso sobre Tecnologias na Educação (Ctrl + E 2017), Mamanguape, Brasil, 18 a 20 de maio, 2017.

- Do Couto, D.C, Santana, A.L. Análise dos problemas da Evasão e Retenção: Uma abordagem através de Mineração de Dados Educacionais. Aceito para apresentação no XIII Encontro Anual de Computação - EnAComp 2017, Catalão, Goiás, Brasil, 24 a 26 de maio, 2017.

Referências

AL-RADAIDEH, Q. A.; AL-SHAWAKFA, E. M.; AL-NAJJAR, M. I. Mining Student Data Using Decision Trees. *The 2006 International Arab Conference on Information Technology (ACIT'2006)*, 2006. Jordan, 2006.

ALKHASAWNEH, R.; HOBSON, R. Modeling Student Retention in Science and Engineering Disciplines Using Neural Networks. *IEEE Global Engineering Education Conference (EDUCON)*, 2011. p. 660–663, 2011.

ANDRIOLA, W. Fatores associados à evasão discente na Universidade Federal do Ceará (UFC) de acordo com as opiniões de docentes e de coordenadores de cursos. 2009. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, v. 7, n. 4, 2009.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de Dados Educacionais: Oportunidades para o Brasil. In: . [S.l.: s.n.], 2011. v. 19, n. 3–13.

BAKER, R. S. Data Mining for Education. *International Encyclopedia of Education*, 2009. Elsevier, v. 3, 2009.

BAKER, R. S.; YACEF, K. The State of Educational Data Mining in 2009: A Review and Future Visions. *JEDM - Journal of Educational Data Mining*, 2009. v. 1, n. 1, p. 3–17, out. 2009.

BARROSO, M. F.; FALCÃO, E. B. M. Evasão Universitária: O Caso do Instituto de Física da UFRJ. 2004. IX Encontro Nacional de Pesquisa em Ensino de Física, 2004.

BORUNDA, M. et al. Bayesian networks in renewable energy systems: A bibliographical survey. *Renewable and Sustainable Energy Reviews*, 2016. v. 62, p. 32 – 45, 2016. ISSN 1364-0321.

CARVALHO, L. A. V. *Data Mining: A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração*. 2. ed. [S.l.]: Érica, 2001.

CASTRO, F. et al. Applying Data Mining Techniques to e-Learning Problems, Studies in Computational Intelligence (SCI). *Springer-Verlag*, 2007. p. 183–221, 2007.

CHEEWAPRAKOBKIT, P. Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program. In: *Proceedings of the International MultiConference of Engineers and Computer Scientists*. Hong Kong: [s.n.], 2013. v. 1.

CLARKE, A. B.; DISNEY, R. L. *Probability and random processes : a first course with applications*. New York, Chichester, Brisbane: John Wiley, 1985. (Wiley Series in Probability and Mathematical Statistics). ISBN 0-471-08535-9.

COOPER, G. F.; HERSKOVITS, E. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Mach. Learn.*, 1992. Kluwer Academic Publishers, Hingham, MA, USA, v. 9, n. 4, p. 309–347, out. 1992. ISSN 0885-6125.

CORTEZ, P.; SILVA, A. Using data mining to predict secondary school student performance. In: *Proceedings of 5th Annual Future Business Technology Conference*. Porto, Portugal: [s.n.], 2008. p. 5–12.

DEKKER, G.; PECHENIZKIY, M.; VLEESHOUWERS, J. Predicting Students Drop Out: A Case Study. In: *International Conference on Educational Data Mining*. Spain: [s.n.], 2009. p. 41–50.

DELEN, D. A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 2010. Elsevier, p. 498–506, 2010.

DIETTERICH, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.*, 1998. MIT Press, Cambridge, MA, USA, v. 10, n. 7, p. 1895–1923, out. 1998. ISSN 0899-7667.

FAYYAD, U. Data mining and knowledge discovery in databases: implications for scientific databases. In: *Scientific and Statistical Database Management, 1997. Proceedings., Ninth International Conference on*. [S.l.: s.n.], 1997. p. 2–11.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 1996. v. 17, p. 37–54, 1996.

FAYYAD, U. M. et al. *Advances In Knowledge Discovery And Data Mining*. [S.l.]: American Association for Artificial Intelligence, 1996. ISBN 02625600976.

FERRO, M.; LEE, H. D. O Processo de KDD – Knowledge Discovery in Database para Aplicações na Medicina. *SEMINC*, 2001. p. 57–62, 2001.

FLEISS, J. *Statistical methods for rates and proportions Rates and proportions*. [S.l.]: Wiley, 1973.

FRAWLEY, W.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. Knowledge Discovery in Databases: An Overview. *AI Magazine*, 1992. p. 57–70, 1992.

GNU Octave – Scientific Programming Language. 2016. Disponível em: <https://www.gnu-.org/software/octave/doc/interpreter/>.

GOLDSCHMIDT, R.; PASSOS, E. *Data Mining: Um Guia Prático*. [S.l.]: Editora Campus, 2005.

GROTH, R. *Data mining: building competitive advantage*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2000. ISBN 0-13-086271-1.

HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. [S.l.]: Morgan Kaufmann, 2001.

HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. 2. ed. [S.l.]: Morgan Kaufmann, 2006.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3. ed. [S.l.]: Morgan Kaufmann, 2012.

HLEL, E.; JAMOUSSI, S.; HAMADOU, A. B. Bayesian network for discovering the interests of authors. In: *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*. [S.l.: s.n.], 2016. p. 1–6.

HäMäLäINEN, W.; VINNI, M. Comparison of machine learning methods for intelligent tutoring systems. In: IKEDA, M.; ASHLEY, K. D.; CHAN, T.-W. (Ed.). *Intelligent Tutoring Systems*. [S.l.]: Springer, 2006. (Lecture Notes in Computer Science, v. 4053), p. 525–534. ISBN 3-540-35159-0.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. *Censo da Educação Superior 2014 - Notas Estatísticas*. [S.l.], 2014. Disponível em: <http://download.inep.gov.br/educacao_superior/censo_superior/documentos/2015-/notas_sobre_o_censo_da_educacao_superior_2014.pdf>. Acesso em: 16/04/2016.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. *Resumo Técnico Censo da Educação Superior 2013*. [S.l.], 2014. Disponível em: <http://download.inep.gov.br/download/superior/censo/2013-/resumo_tecnico_censo_educacao_superior_2013.pdf>. Acesso em: 16/04/2016.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. *Data Clustering: A Review*. 1999.

JOURNAL OF EDUCATIONAL DATA MINING. *Journal of Educational Data Mining*. [S.l.], 2016. Disponível em: <http://www.educationaldatamining.org/JEDM>. Acesso em: 26/01/2016.

KABAKCHIEVA, D. Predicting student performance by using data mining methods for classification. *CYBERNETICS AND INFORMATION TECHNOLOGIES*, 2013. Sofia, v. 13, n. 1, 2013.

KO, S.; KIM, D.-W. An efficient node ordering method using the conditional frequency for the k2 algorithm. *Pattern Recogn. Lett.*, 2014. Elsevier Science Inc., New York, NY, USA, v. 40, p. 80–87, abr. 2014. ISSN 0167-8655.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence (IJCAI)*, 1995. p. 1137–1145, 1995.

KORB, K. B.; NICHOLSON, A. E. *Bayesian Artificial Intelligence, Second Edition*. 2nd. ed. Boca Raton, FL, USA: CRC Press, Inc., 2010. ISBN 1439815917, 9781439815915.

KOTSIANTIS, S. B. et al. Predicting students performance in distance learning using machine learning techniques. 2003. Recent Advances in Mechanics and Related Works, p. 297–305, 2003.

KOVAČIĆ, Z. J. Early prediction of student success: Mining students enrolment data. *Proceedings of Informing Science & IT Education Conference (InSITE)*, 2010. 2010.

LOBO, M. B. de C. M. *PANORAMA DA EVASÃO NO ENSINO SUPERIOR BRASILEIRO: ASPECTOS GERAIS DAS CAUSAS E SOLUÇÕES*. [S.l.]: Instituto Lobo para Desenvolvimento da Educação, da Ciência e da Tecnologia, 2011.

MANHÃES, L. M. B. et al. Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. In: ANAIS DO XXII SBIE, 22., 2011, Aracaju. Aracaju: SBIE, 2011.

MANHÃES, L. M. B. et al. Identificação dos Fatores que Influenciam a Evasão em Cursos de Graduação Através de Sistemas Baseados em Mineração de Dados: Uma Abordagem Quantitativa. In: *Simpósio Brasileiro de Sistemas de Informação*. [S.l.: s.n.], 2012. p. 468–479.

MANHÃES, L. M. B. et al. Investigating Withdraw of STEM Courses in a Brazilian University with EDM. In: . [S.l.: s.n.], 2014. p. 1–8.

MANHÃES, L. M. B. et al. The Impact of High Dropout Rates in a Large Public Brazilian University. *CSEDU – 6th International Conference on Computer Supported Education*, 2014. p. 126–129, 2014.

MANHÃES, L. M. B. et al. WAVE: an Architecture for Predicting Dropout in Undergraduate Courses using EDM. In: SYMPOSIUM OF APPLIED COMPUTING (SAC 2014), 2014, Gyeongju, Korea. [S.l.]: SAC, 2014.

MANHÃES, L. M. B. et al. Towards Automatic Prediction of Student Performance in STEM Undergraduate Degree Programs. In: SYMPOSIUM OF APPLIED COMPUTING (SAC 2015), 2015, Salamanca, Spain. [S.l.]: SAC, 2015.

MAYILVAGANAN, M.; KALPANADEVI, D. Comparison of classification techniques for predicting the performance of students academic environment. In: *Communication and Network Technologies (ICCNT), 2014 International Conference on*. [S.l.: s.n.], 2014. p. 113–118.

MINAEI-BIDGOLI, B. et al. Predicting student performance: an application of data mining methods with an educational web-based system. In: *Frontiers in Education, 2003. FIE 2003 33rd Annual*. [S.l.: s.n.], 2003. v. 1, p. T2A–13. ISSN 0190-5848.

MINISTÉRIO DA EDUCAÇÃO. *Reestruturação e Expansão das Universidades Federais: Diretrizes Gerais*. [S.l.], 2007. Disponível em: <http://portal.mec.gov.br/sesu/arquivos/pdf-/diretrizesreuni.pdf>. Acesso em: 15/04/2016.

MINISTÉRIO DA EDUCAÇÃO E CULTURA. *Diplomação, Retenção e Evasão nos Cursos de Graduação em Instituições de Ensino Superior Públicas.* [S.l.], 1997. Disponível em: <http://www.udesc.br/arquivos/id_submenu/102/diplomacao.pdf>. Acesso em: 15/04/2016.

MINISTÉRIO DA EDUCAÇÃO E CULTURA. *ORIENTAÇÕES PARA O CáLCULO DOS INDICADORES DE GESTãO*. [S.l.], 2004. Disponível em: <http://portal.mec.gov.br/setec-/arquivos/pdf/indicadores.pdf>. Acesso em: 15/04/2016.

MOODLE. *Moodle – Open-source learning plataform*. [S.l.], 2016. Disponível em: <https://moodle.org/?lang=pt_br>. Acesso em: 22/04/2016.

NEWHAM, C.; ROSENBLATT, B. *Learning the Bash Shell*. 2nd. ed. Sebastopol, CA, USA: O'Reilly & Associates, Inc., 1998. ISBN 1565923472.

NGHE, N. T.; JANECEK, P.; HADDAWY, P. A comparative analysis of techniques for predicting academic performance. In: *Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE '07. 37th Annual*. [S.l.: s.n.], 2007. p. T2G–7–T2G–12. ISSN 0190-5848.

PANDEY, U. K.; PAL, S. Data mining : A prediction of performer or underperformer using classification. *International Journal of Computer Science and Information Technologies (IJCSIT)*, 2011. v. 2, p. 686–690, 2011. ISSN 0975-9646.

PANDEY, U. K.; PAL, S. Data mining : A prediction of performer or underperformer using classification. *CoRR*, 2011. abs/1104.4163, 2011. Disponível em: <http://arxiv.org/abs/1104-.4163>.

PEARL, J. Fusion, propagation, and structuring in belief networks. *Artif. Intell.*, 1986. Elsevier Science Publishers Ltd., Essex, UK, v. 29, n. 3, p. 241–288, set. 1986. ISSN 0004-3702.

PENÃ-AYALA, A. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, 2013. Elsevier, 2013.

PENÃ-AYALA, A. Educational data mining: applications and trends, studies in computational intelligence. 2013. Springer Verlang, 2013.

PRESIDÊNCIA DA REPÚBLICA. *LEI Nº 3.191, DE 2 DE JULHO DE 1957*. [S.l.], 1957. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/1950-1969/L3191.htm>. Acesso em: 14/04/2016.

PRESIDÊNCIA DA REPÚBLICA. *DECRETO Nº 6.096, DE 24 DE ABRIL DE 2007*. [S.l.], 2007. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2007-2010/2007/decreto-/d6096.htm>. Acesso em: 15/04/2016.

PRESIDÊNCIA DA REPÚBLICA. *LEI Nº 12.089 DE 11 DE NOVEMBRO DE 2009*. [S.l.], 2009. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2007-2010/2009/Lei-/L12089.htm>. Acesso em: 15/04/2016.

REZENDE, S. O. *Sistemas Inteligentes: Fundamentos e Aplicações*. [S.l.]: Manole, 2005.

RICH, E.; KNIGHT, K. *Inteligência Artificial*. [S.l.]: Makron Books, 1993.

ROMERO, C.; VENTURA, S. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 2007. v. 1, n. 33, p. 135–146, 2007.

ROMERO, C.; VENTURA, S. Educational Data Mining: A Review of the State of the Art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 2010. v. 40, n. 6, p. 601–618, 2010. ISSN 1094-6977.

ROMERO, C. et al. Data mining algorithms to classify students. In: BAKER, R. S. J. de; BARNES, T.; BECK, J. E. (Ed.). *EDM*. [S.l.]: www.educationaldatamining.org, 2008. p. 8–17.

RUSSEL, J. S.; NORVING, P. *Inteligência Artificial*. [S.l.]: Elsevier, 2004.

RUSSEL, J. S.; NORVING, P. *Artificial Intelligence: A Modern Approach*. 3. ed. [S.l.]: Pearson, 2010.

SACHIN, R.; VIJAY, M. A Survey and Future Vision of Data Mining in Educational Field. In: *Advanced Computing Communication Technologies (ACCT), 2012 Second International Conference on*. [S.l.: s.n.], 2012. p. 96–100.

SANTANA Ádamo Lima de. *Estratégias para a melhoria da modelagem e interpretabilidade de redes bayesianas*. Doutorado em Engenharia ElÃ©trica, 2008.

SARDINHA, R. de L.; PAES, A.; ZAVERUCHA, G. Aprendizado Local da Estrutura de Redes Bayesianas a partir de Dados Incompletos - Bayes Ball Structure Learning (BBSL). 2009. 2009.

SHAFRANOVICH, Y. *Common Format and MIME Type for Comma-Separated Values (CSV) Files*. [S.l.], 2005. Disponível em: <https://tools.ietf.org/html/rfc4180>.

SILVA FILHO, R. L. L. et al. A evasão no ensino superior brasileiro. 2007. Fundação Carlos Chagas, v. 37, n. 132, p. 641–659, 2007.

SOONG, T. T. *Probabilistic modeling and analysis in science and engineering*. 1. ed. [S.l.]: LTC, 1986.

SOONG, T. T. *Fundamentals of Probability and Statistics for Engineers*. 1. ed. [S.l.]: Wiley-Interscience, 2004. Paperback. ISBN 0470868147.

SUPERBY, J. F.; VANDAMME, J.-P.; MESKENS, N. Determination of factors influencing the achievement of the first-year university students using data mining methods. In: *Proceedings International Conference Intelligent Tutoring System of the Workshop on Educational Data Mining*. Taiwan: [s.n.], 2006. p. 1–8.

TINTO, V. Dropout from higher education: a theoretical synthesis of recent research. 1975. Review of Educational Research, p. 89–125, 1975.

TINTO, V. *Leaving college: rethinking the causes of student attrition*. [S.l.]: University of Chigago Press, 1987.

TURING, A. Computing machinery and intelligence. *Mind*, 1950. p. 433–460, 1950.

UNIVERSIDADE FEDERAL DO PARÁ. *Plano de Desenvolvimento Institucional 2011–2015*. [S.l.], 2015. Disponível em: <https://www.portal.ufpa.br/docs/pdi_aprovado_final.pdf>. Acesso em: 14/04/2016.

UNIVERSIDADE FEDERAL DO PARÁ. *UFPA em Números*. [S.l.], 2015. Disponível em: <http://www.ufpanumeros.ufpa.br/index.php/br/ensino>. Acesso em: 14/04/2016.

UNIVERSIDADE FEDERAL DO PARá. *Resolução Nº 4399 de 14 de maio de 2013 – Regulamento do Ensino de Graduação da Universidade Federal do Pará*. [S.l.], 2013. Disponível em: <http://www.proeg.ufpa.br/view/inicio/downloads.php?idDoc=263>. Acesso em: 08/01/2017.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE. *Resolução Nº 171/2013-CONSEPE – Regulamento dos Cursos Regulares de Graduação da Universidade Federal do Rio Grande do Norte*. [S.l.], 2013. Disponível em: <http://www.sistemas.ufrn.br/download-/sigaa/public/regulamento_dos_cursos_de_graduacao.pdf>. Acesso em: 08/01/2017.

WITTEN, I. H.; FRANK, E. *Data Mining Pratical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann, 2005.

WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. 4th. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2016. ISBN 0123748569, 9780123748560.

YANG, S.; CHANG, K.-C. Comparison of score metrics for bayesian network learning. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 2002. v. 32, n. 3, p. 419–428, May 2002. ISSN 1083-4427.

Anexos

ANEXO A – Cálculo dos Indicadores de Rendimento Acadêmico Acumulado

ANEXO II DO REGULAMENTO DOS CURSOS REGULARES DE GRADUAÇÃO DA UFRN

ANEXO II
CÁLCULO DOS INDICADORES DE RENDIMENTO ACADÊMICO ACUMULADO

A **Média de Conclusão (MC)** é a média ponderada do rendimento acadêmico final nos componentes curriculares em que o estudante conseguiu êxito ao longo do curso, obtida pela seguinte fórmula:

$$MC = \frac{\sum_{i=1}^{Nx} n_i \times c_i}{\sum_{i=1}^{Nx} c_i}$$

São contabilizados os $\mathbf{N_x}$ componentes curriculares concluídos com êxito após o início do curso, sendo $\mathbf{n_i}$ a nota (rendimento acadêmico) final obtida no **i**-ésimo componente curricular e $\mathbf{c_i}$ a carga horária discente do **i**-ésimo componente curricular. São excluídos do cálculo os componentes curriculares trancados, cancelados, reprovados, aproveitados, incorporados e dispensados e os componentes curriculares cujo rendimento acadêmico não é expresso de forma numérica.

A **Média de Conclusão Normalizada (MCN)** é a MC do estudante normalizada em relação à média (μ) e desvio padrão amostral (σ) das MC dos concluintes do mesmo curso, obtida pela seguinte fórmula:

$$MCN = 500 + 100 * \left(\frac{MC - \mu}{\sigma} \right)$$

Nessa fórmula, **MC** é a Média de Conclusão do estudante para o qual está sendo calculada a **MCN**. A média (μ) e desvio padrão amostral (σ) são calculados pelas seguintes fórmulas:

$$\mu = \frac{1}{M} \sum_{i=1}^{M} MC_i \qquad \sigma = \sqrt{\frac{1}{M-1} \sum_{i=1}^{M} (MC_i - \mu)^2}$$

São contabilizados os **M** estudantes que concluíram o mesmo curso nos últimos 5 (cinco) anos, sendo $\mathbf{MC_i}$ a Média de Conclusão final obtida pelo **i**-ésimo concluinte. São excluídos do cálculo os estudantes que não concluíram com êxito o curso por qualquer motivo bem como aqueles que fizeram apenas apostilamento de habilitação ou certificação de ênfase.

Para os cursos com mais de um turno ou mais de uma habilitação ou ênfase, a média e desvio padrão amostral são os mesmos para todos os estudantes das diferentes matrizes curriculares.

A média e desvio padrão são calculados para os cursos que têm estudantes concluintes há pelo menos 5 (cinco) anos ou em número superior a 100 (cem). Caso contrário, utilizam-se os valores médios do centro acadêmico do curso ou, caso impossível, do centro com maior similaridade.

O **Índice de Eficiência em Carga Horária (IECH)** é o percentual da carga horária utilizada pelo estudante que se converteu em aprovação, obtido pela seguinte fórmula:

$$IECH = \frac{\sum_{i=1}^{Np} c_i}{\sum_{i=1}^{Nm} c_i}$$

São contabilizados no numerador os $\mathbf{N_p}$ componentes curriculares nos quais o estudante obteve aprovação ou integralizou após o início do curso, incluindo-se os componentes incorporados depois do início do curso e excluindo-se os componentes aproveitados, cursados antes do início do curso, e os dispensados.

São contabilizados no denominador os $\mathbf{N_m}$ componentes curriculares nos quais o estudante teve a matrícula efetuada após o início do curso, incluindo-se os componentes incorporados após o início do curso e os trancamentos, reprovações e cancelamentos de matrícula e excluindo-se os componentes curriculares aproveitados, cursados antes do início do curso, e os dispensados.

$\mathbf{c_i}$ é a carga horária discente do **i**-ésimo componente curricular.

O **Índice de Eficiência em Períodos Letivos (IEPL)** é a divisão da carga horária acumulada pela carga horária esperada, obtida pela seguinte fórmula:

$$IEPL = \frac{\sum_{i=1}^{Na} c_i}{P \times \frac{CHM}{DP}}$$

São contabilizados no numerador todos os $\mathbf{N_a}$ componentes curriculares nos quais o estudante acumulou carga horária após o início do curso, incluindo-se os componentes curriculares incorporados após o início do curso e excluindo-se os componentes curriculares aproveitados, cursados antes do início do curso, e os dispensados.

$\mathbf{c_i}$ é a carga horária discente do **i**-ésimo componente curricular.

P é o número de períodos já cursados pelo estudante, excluindo-se os períodos letivos nos quais o programa foi suspenso e aqueles durante os quais o estudante esteve realizando mobilidade acadêmica em outra instituição, não incluindo também os períodos letivos contados no perfil inicial.

CHM e **DP** são a carga horária mínima e a duração padrão, respectivamente, para integralização da estrutura curricular do estudante.

O **Índice de Eficiência Acadêmica (IEA)** é o produto da MC pelo IECH e pelo IEPL, conforme a seguinte fórmula:

$$IEA = MC \times IECH \times IEPL$$

O **Índice de Eficiência Acadêmica Normalizado (IEAN)** é o produto da MCN pelo IECH e pelo IEPL, conforme a seguinte fórmula:

$$IEAN = MCN \times IECH \times IEPL$$

APÊNDICE A – Cabeçalho do arquivo ARFF utilizado

```
@attribute sexo {F,M}
@attribute idade {[14-19],(19-21],(21-26],(26-100]}
@attribute interior {f,t}
@attribute turno {I,D,N}
@attribute forma_ingresso {PS,PARFOR,OUTRAS,MOBEX,PSE,MOBIN,ENEM}
@attribute numero_trancamento {0,1,>1}
@attribute numero_vinculos {0,>=1}
@attribute perc_ch_pratico {0,(0-27.615],>27.615}
@attribute perc_ch_teorico {[0-50.275],>50.275}
@attribute perc_ch_estagio {0,(0-100]}
@attribute sem_ordem {[0-23.57],(23.57-67.43],(67.43-100]}
@attribute primeiro_semestre_ocorr {1,2,>2}
@attribute mc {[0-6.935],(6.935-7.645],(7.645-8.355],>8.355}
@attribute crg {[0-5.055],(5.055-6.975],(6.975-8.065],>8.065}
@attribute mcn {[0-426.385],(426.385-501.375],(501.375-568.865],>568.865}
@attribute iech {[0-0.665],(0.665-0.915],(0.915-0.995],>0.995}
@attribute iepl {[0-0.825],(0.825-0.965],(0.965-1.065],>1.065}
@attribute iea {[0-3.905],(3.905-6.435],(6.435-7.975],(7.975-10]}
@attribute iean {[0-242.115],(242.115-407.555],(407.555-532.455],>532.455}
@attribute probabilidade_mc
    {[0-32.135],(32.135-65.965],(65.965-90.505],(90.505-100]}
@attribute probabilidade_ira
    {[0-30.245],(30.245-64.435],(64.435-96.645],(96.645-100]}
@attribute probabilidade_iea
    {[0-22.985],(22.985-59.585],(59.585-94.755],(94.755-100]}
@attribute perc_ins_geral {0,(0-7.665],(7.665-29.665],(29.665-100]}
@attribute perc_reg_geral
    {[0-4.985],(4.985-13.615],(13.615-24.355],(24.355-100]}
@attribute perc_bom_geral
    {[0-21.115],(21.115-33.745],(33.745-44.145],(44.145-100]}
@attribute perc_exc_geral
    {[0-9.825],(9.825-22.355],(22.355-39.365],(39.365-100]}
@attribute perc_ins_primeiro_ano {0,(0-13.8],(13.8-38.97],(38.97-100]}
@attribute perc_reg_primeiro_ano {0,(0-16.655],(16.655-31.83],(31.83-100]}
@attribute perc_bom_primeiro_ano
    {[0-20.02],(20.02-36.825],(36.825-51.985],(51.985-100]}
@attribute perc_exc_primeiro_ano
    {[0-6.905],(6.905-18.98],(18.98-35.885],(35.885-100]}
```

```
@attribute status {FORMADO, RETIDO, EVADIDO}

@data
```

Código A.1 – Parte do arquivo ARFF empregado na etapa de *Data Mining*