UNIVERSIDADE FEDERAL DA BAHIA

DISSERTAÇÃO DE MESTRADO

Uma Análise Probabilística da Retenção na Universidade Federal da Bahia: um Estudo de Caso no Curso de Ciência da Computação

Marcelo Silva Santos

Programa de Pós-Graduação em Ciência da Computação

Salvador
18 de novembro de 2015

PGCOMP-Msc-2015

MARCELO SILVA SANTOS

UMA ANÁLISE PROBABILÍSTICA DA RETENÇÃO NA UNIVERSIDADE FEDERAL DA BAHIA: UM ESTUDO DE CASO NO CURSO DE CIÊNCIA DA COMPUTAÇÃO

Esta Dissertação de Mestrado foi apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientadora: Profa. Dra. Daniela Barreiro Claro
Co-orientadora: Profa. Dra. Verônica Maria Cadena Lima

Salvador
18 de novembro de 2015

Ficha catalográfica.

SEU NOME EM CITACOES

Uma Análise Probabilística da Retenção na Universidade Federal da Bahia: um Estudo de Caso no Curso de Ciência da Computação/ Marcelo Silva Santos– Salvador, 18 de novembro de 2015.

109p.: il.

Orientadora: Profa. Dra. Daniela Barreiro Claro.
Co-orientadora: Profa. Dra. Verônica Maria Cadena Lima.
TIPO DE TRABALHO– UNIVERSIDADE FEDERAL DA BAHIA, INSTITUTO DE MATEMÁTICA, 18 de novembro de 2015.

TOPICOS PARA FICHA CATALOGRAFICA.
I. NOME DO SEU ORIENTADOR EM CITACOES. II. NOME DO SEU CO-ORIENTADOR EM CITACOES.
III. UNIVERSIDADE FEDERAL DA BAHIA. INSTITUTO DE MATEMÁTICA. IV. Título.

NUMERO CDD

AGRADECIMENTOS

Uma das lições de vida que aprendi, é que conquistas pessoais estão diretamente ligadas as pessoas que estão ao seu redor. Essa conexão é mais intensa quando chegamos a pensar, "se não fosse por cicrano e beltrano, talvez eu nem estivesse onde estou".

Ao finalizar um mestrado, não tem como deixar de fazer uma retrospectiva na vida e reconhecer as pessoas que fizeram com que eu chegasse a ter a oportunidade de entrar em um mestrado, quiça concluir.

A toda conquista obtida em minha vida, primeiramente devo agradecer a mainha, uma mulher guerreira que me ensinou a lutar para conseguir os meus objetivos e que sempre fez de tudo para me ver feliz. Obrigado por tudo minha mãe, eu tenho muito orgulho da senhora.

Outras pessoas importantes que contribuíram para eu estar em um mestrado foram minha tia Silvana e meu tio Valfredo que me acolheram em momentos complicados da minha vida e me trataram como um filho. Serei grato a eles pelo resto da vida.

Ao professor José Craveiro, pelo incentivo dado para entrar na vida acadêmica durante a graduação. Foi a partir daqui que vislumbrei o início de um mestrado.

Do início ao fim do mestrado, existiram pessoas que fizeram toda a diferença para que eu chegasse a conclusão deste trabalho.

Agradeço principalmente à orientadora e amiga Daniela Claro, por sua dedicação e empenho, não só para o desenvolvimento do trabalho, mas também para o meu crescimento pessoal e profissional, sem sua orientação seria quase impossível este trabalho ter sido desenvolvido. Obrigado por me mostrar como é ser um excelente orientador, os seus ensinamentos com certeza serão passados a diante. Agradeço também a minha coorientadora Verônica Lima por me socorrer ao surgirem dúvidas a respeito dos conceitos de probabilidade e por colaborar do início ao fim no desenvolvimento do trabalho.

A minha namorada Vanessa Sales, agradeço por estar sempre ao meu lado nesta jornada, me acalmando, aconselhando e me incentivando a concluir este trabalho. Você fez toda a diferença neste processo, te amo.

Também, agradeço a amiga Mydiã Freitas por ouvir os desabafos sobre o andamento da pesquisa do mestrado e por torcer para o sucesso do mesmo.

Por fim, mas não menos importante, agradeço aos meus amigos Albert, Jovane, Flávio, Juca, Tia Cris e Tio Rui por estarem torcendo pelo meu sucesso mesmo estando longe. São amizades assim que levaremos a vida toda.

“No fim tudo dá certo, e se não deu certo é porque ainda não chegou ao fim.”

—FERNANDO SABINO

RESUMO

O crescimento nas universidades brasileiras vem permitindo que vários alunos tenham acesso ao ensino superior. Esse crescimento foi possível devido ao projeto de restruturação nas universidades federais (REUNI). No entanto, grande parte dos estudantes que se matriculam em um curso excedem o tempo regular para obtenção do grau. Ao ultrapassar esse tempo, tanto a instituição quanto o aluno acumulam prejuízos financeiros e pedagógicos.

Diante deste contexto, universidades e institutos federais começaram a desenvolver pesquisas cujo o principal objetivo constitui em analisar a retenção desses alunos buscando métodos, técnicas ou metodologias que auxiliem na redução da retenção.

É nesse cenário que este trabalho se insere, com o objetivo principal de analisar a retenção dos alunos do curso de Ciência da Computação da Universidade Federal da Bahia no intuito de proporcionar informações que possam auxiliar os gestores na criação de políticas para a redução da retenção. Para isso, foram utilizadas redes bayesianas no intuito de obter resultados probabilísticos a respeito da aprovação/reprovação dos alunos em disciplinas cursadas em um determinado semestre, em certa tentativa.

Para obter os resultados esperados, foram feitos dois experimentos: i) definição de uma rede bayesiana manualmente baseada no fluxo de disciplinas da grade curricular do curso a fim de identificar o desempenho do aluno no fluxo das disciplinas da grade curricular de acordo com a sua aprovação ou reprovação em disciplinas cursadas anteriormente, e ii) definição de uma classificador bayesiano através do algoritmo *Naive Bayes* no intuito de identificar perfis de alunos que tendem a ultrapassar o tempo regular de conclusão (retenção final), bem como a identificação do impacto da retenção em um dado semestre na retenção final do aluno.

A partir dos resultados probabilísticos obtidos, foi possível analisar os resultados e obter as seguintes conclusões: i) ser aprovado nos pré-requisitos aumenta a probabilidade de aprovação da disciplina seguinte; ii) disciplinas iniciais tem um grande impacto no resultado do aluno em disciplinas posteriores; iii) a reprovação em disciplinas básicas gera uma retenção nos semestres posteriores que dificilmente poderá ser reparada; iv) todas as disciplinas do primeiro semestre devem ser priorizadas para que o aluno não fique retido no fim do curso; v) a probabilidade da retenção final dado a retenção em qualquer um dos semestres do curso fica em torno de 93%.

Assim, através dos resultados probabilísticos e as análises realizadas este trabalho propõe algumas políticas de retenção como: possíveis alterações na grade curricular, políticas de acompanhamento dos alunos durante e no final do semestre, assistência na escolha de disciplinas para determinado perfis de alunos. Além disso, essas informações tem o principal objetivo de proporcionar aos gestores subsídios para a criação de políticas eficazes para a redução da retenção.

Palavras-chave: Mineração de Dados, Probabilidades, Redes Bayesianas, Retenção.

ABSTRACT

The growth in Brazilian universities has allowed several students have access to higher education, this growth was possible due to restructuring project in the federal universities. However, most of the students who enroll in a course exceeds the regular time for the degree. To exceed this time, both the institution and the student accumulate financial and educational losses.

Given this context, universities and federal institutes began developing research whose main objective is to analyze the retention of these students seeking methods, techniques or methodologies to assist in reducing retention.

It is in this scenario that this work is inserted, with the main objective of analyzing the retention of students of Computer Science of the Federal University of Bahia in order to provide information that may assist managers in creating policies to reduce retention. For this, we used Bayesian networks in order to obtain probabilistic results about the pass/fail of students in courses taken in a given semester, a certain try.

For the expected results, two experiments were made: i) definition of a Bayesian network manually based on the course curriculum subjects flow to identify the student's behavior in the flow of the subjects of the curriculum according to their approval or failure in courses taken before, and ii) building a Bayesian classifier through the algorithm *Naive Bayes* in order to identify profiles of students who tend to exceed the regular time of completion (final retention) as well as the identification of impact of retention in a given semester in the final retention of the student.

From the probabilistic results obtained, it was possible to analyze the results and get the following conclusions: i) to pass the prerequisites increases the likelihood of approval of the following discipline; ii) initial disciplines have a major impact on the results of the student in later disciplines; iii) failure in basic disciplines generates a retention in later semesters that can hardly be repaired; iv) all disciplines of the first semester should be prioritized so that the student does not get trapped at the end of the course; v) the likelihood of final retention given retention in any of the semesters of the course is around 93%.

Thus, through the probabilistic results and analyzes this paper proposes some retention policies such as possible changes in the curriculum, monitoring policies for students during and at the end of the semester, assistance in choosing subjects for a given student profile. In addition, this information has the primary objective to provide managers subsidies for the creation of effective policies to reduce retention.

Keywords: Data Mining, Probabilitys, Bayes Network, Retention.

SUMÁRIO

LISTA DE ABREVIATURAS

AUC *Universidade Federal de Pernambuco*

BI Bacharelado Interdisciplinar

DAG Gráfico Acíclico Dirigido

EDM *Education Data Mining*

PROUFBA Programa Pense, Pesquisa e Inove a UFBA

RB Rede Bayesiana

REUNI Restruturação e Expansão das Universidades Federais

TPC Tabela de Probabilidade Condicional

UFBA Universidade Federal da Bahia

UFPE Universidade Federal de Pernambuco

LISTA DE FIGURAS

LISTA DE TABELAS

Capítulo 1

.

INTRODUÇÃO

Nas universidades brasileiras têm havido um aumento no número de vagas, principalmente por causa dos procedimentos do REUNI (Restruturação e Expansão das Universidades Federais) (BRASIL,). No projeto e implementação do REUNI na Universidade Federal da Bahia (UFBA) foi previsto a matrícula de aproximadamente 32.000 novos estudantes devido a criação dos novos cursos na Universidade. Deste total, 10.500 alunos estariam inscritos em cursos noturnos, sendo 2.107 alunos matriculados em cursos noturnos de BI (Bacharelado Interdisciplinar) . Dentre as principais instruções do REUNI, é possível destacar: um aumento no número de ingressos, especialmente para os cursos noturnos, redução no número de evasões e ocupação de vagas ociosas. De acordo com (FILHO et al., 2010), na implementação do REUNI, a taxa de conclusão esperada foi estimada em torno de 90% dos ingressos. No entanto, a taxa de conclusão nos cursos de graduação ainda está muito abaixo do desejado. Dentre tantos fatores que contribuem com a taxa de conclusão baixa, é possível destacar (MANHÃES et al., 2012): i) a evasão dos cursos e ii) a não conclusão dos cursos no período regular, ou seja a retenção.

A não conclusão do curso no período regular é reflexo principalmente, da retenção propiciada por determinados componentes curriculares no desempenho do alunado, prejudicando a sua semestralização. A não semestralização causa prejuízos à Universidade, principalmente em cursos em implantação, que devem oferecer obrigatoriamente os componentes curriculares do semestre em questão, alocando docentes em disciplinas para uma proporção muito aquém do almejado pelo REUNI (FILHO et al., 2010) de 18:1 (18 discentes para 1 docente). Além disso, a retenção no ensino superior tem efeitos negativos do ponto de vista da universidade, família, sociedade e pedagógico, o que evidencia a necessidade de pesquisas na área.

Do ponto de vista da Universidade, os alunos ao ultrapassarem o tempo regular para a conclusão do curso, gera uma necessidade de destinar mais recursos públicos para custeá-los. Caso não houvessem alunos retidos, o dobro de recurso destinado ao custeio dos alunos poderia ser economizado. Já para a família, o discente que despende mais que o tempo regular do curso para concluir acaba gerando custos superiores do que o planejado pela família. Do ponto de vista social, quanto mais tempo o aluno despende para

conclusão do seu curso, maior será o tempo necessário para que possa atuar profissionalmente tendo impacto direto na economia do país. Por fim, do ponto de vista pedagógico, a repetência em disciplinas, que leva a retenção tem alguns efeitos: a desmotivação e a diminuição da auto-estima que interferem no processo de aprendizagem e o aumento da probabilidade de reprovações futuras.

A utilização de técnicas de mineração de dados e/ou modelos probabilísticos em dados educacionais tem sido conhecido como EDM (*Education Data Mining*). No Brasil ainda são poucos os trabalhos publicados nesta área de pesquisa. Um dos trabalhos pioneiros no uso de mineração de dados na educação (BRANDÃO; RAMOS; TRÓCCOLI, 2003) analisou os dados do programa nacional de informática na educação. Autores em (BAKER; ISOTANI; CARVALHO, 2011) apresentam alguns métodos e aplicações da EDM e a visão sobre o potencial benefício que a EDM pode trazer ao sistema educacional brasileiro.

Nesse contexto, este trabalho ocorreu em duas etapas: i) utilização de técnicas de mineração de dados na análise da retenção nos cursos de graduação da Universidade Federal da Bahia (UFBA) desenvolvido através do projeto "Análise da Retenção do Alunado da UFBA via Desempenho Acadêmico" no Programa Pense, Pesquisa e Inove a UFBA (PROUFBA) e ii) uma análise probabilística da retenção especificamente no curso de Ciência da Computação que foi o foco dessa dissertação.

A primeira etapa, desenvolvido no projeto PROUFBA, focou na utilização de regras de associação para identificar disciplinas que cursadas em um mesmo semestre levam o aluno a reprovação implicando em uma retenção em um semestre específico. Além disso, foi possível identificar os resultados obtidos pelos alunos classificados como retidos e não retidos em uma disciplina quando cursada em um determinado semestre. Por fim, também foram obtidos resultados dos cursos que mais retêm na UFBA e das disciplinas, por semestre, que mais retêm. Essa implicação de retenção em um semestre específico foi possível dada a definição de uma heurística de retenção no PROUFBA; **para cada disciplina recomendada para um estudante em um determinado semestre, se o estudante foi reprovado ou não se inscreveu em pelo menos um dos pré-requisitos necessários para cursar uma das disciplinas do seu semestre atual, o aluno é caracterizado como retido**. Os resultados obtidos no projeto PROUFBA foram validados através de artigos publicados em conferências específicas da área e todos os resultados podem ser vistos no relatório final produzido no final do projeto (CLARO et al., 2014).

Com a primeira etapa, foi possível entender o contexto da retenção nos cursos da UFBA, possibilitando uma melhor visão de possíveis técnicas e analises que pudessem contribuir ainda mais para os resultados a respeito da retenção. Tecnicamente, o projeto PROUFBA contribuiu com uma metodologia de análise dos resultados que consiste em analisar os alunos quando cursam a disciplina em um determinado semestre, em dados parcialmente pré-processados para serem utilizados na segunda etapa do trabalho, com uma definição heurística da retenção que foi essencial na análise dos resultados na segunda etapa do trabalho, bem como na percepção de questões que pudessem ser respondidas com modelos probabilísticos que colaborasse com uma análise da retenção.

A partir dos resultados obtidos na primeira etapa da pesquisa, percebeu-se que in-

formações probabilísticas poderiam melhorar a compreensão sobre a retenção dos alunos, visto que no PROUFBA foram obtidos apenas resultados descritivos, ou seja, os resultados sempre são verdadeiros ou falsos. A diferença entre os resultados descritivos e probabilísticos podem ser vistos diante da informação a respeito do resultado do aluno retido em um determinado semestre ao cursar uma determinada disciplina. No modelo descritivo, a informação pode ser descrita da seguinte forma: os alunos retidos em um semestre Y aprovam na disciplina X, já uma informação probabilística pode informar que esses mesmos alunos retidos no semestre Y aprovam em 51% dos casos e reprovam em 49%. Nesses casos, a informação probabilística apresenta resultados mais precisos que a informação descritiva, visto que no exemplo dado, existem 49% de alunos que reprovam na disciplina X, porém a informação descritiva afirma que todos os alunos retidos aprovam nessa disciplina.

Diante dos efeitos negativos da retenção no ensino superior aqui apresentados e dos resultados obtidos no projeto PROUFBA o presente trabalho propõe realizar uma análise probabilística do curso de Ciência da Computação através da utilização de Redes Bayesianas.

Nesta etapa, realizou-se dois experimentos, o primeiro definiu uma rede bayesiana baseado na grade curricular do curso a fim de analisar probabilisticamente o fluxo do aluno através dos resultados obtidos nos componentes curriculares cursados em um determinado semestre, em certa tentativa. Uma certa tentativa refere-se a quantidade de tentativas que o aluno se matriculou na disciplina que foi reprovado. No segundo a utilização do algoritmo *naive bayes* para predizer a probabilidade de retenção final do aluno a partir de um conjunto de resultados obtidos pelos alunos ao cursar componentes curriculares em um determinado semestre, em uma certa tentativa.

1.1 OBJETIVOS

Como objetivo geral da pesquisa, procurou-se analisar a retenção causada pela reprovação em componentes curriculares dos alunos através de redes bayesianas para que os resultados obtidos possam ser utilizados pelos colegiados para intervir e criar políticas para evitar a retenção.

Especificamente, este trabalho procurou analisar a retenção em duas perspectivas:

- Analisar o fluxo acadêmico dos alunos através dos resultados obtidos ao cursar disciplinas em um determinado semestre em uma certa tentativa.
  - Identificar disciplinas com maiores probabilidades de implicar em retenção por semestre, bem como os resultados dos alunos retidos em disciplinas que retêm em semestres específicos.
  - Calcular a probabilidade da aprovação/reprovação do aluno em disciplinas do semestre posterior dado um resultado obtido em disciplinas de semestre anterior.
- Predizer a probabilidade do aluno concluir o curso no tempo regular baseado nas disciplinas cursadas em determinado semestre, em uma certa tentativa e na retenção

do semestre.

1.2 ORGANIZAÇÃO DO TRABALHO

O restante deste trabalho está estruturado como segue. No Capítulo 2 é fornecida uma revisão da literatura a respeito das definições de retenção, bem como os modelos probabilísticos e técnicas de mineração de dados aplicadas no problema da retenção. No Capítulo 3 são apresentados os conceitos básicos da probabilidade para uma melhor compreensão das redes bayesianas. O Capítulo 4 apresenta a definição de uma rede bayesiana, sua modelagem e algoritmos utilizados na aprendizagem dos parâmetros e nas inferências probabilísticas. Em seguida, no Capítulo 5 é proposto um modelo de redes bayesianas para analisar o comportamento do aluno na grade curricular e outro para predizer probabilisticamente a retenção final dos alunos a partir de um conjunto de variáveis. No Capítulo 6 são apresentados os experimentos realizados para obter a solução proposta no Capítulo 5. O Capítulo 7 fornece os resultados desta pesquisa, bem como suas respectivas análises. Por fim, no Capítulo 8 são apresentadas as conclusões deste trabalho e seus trabalhos futuros.

PARTE I

FUNDAMENTAÇÃO TEÓRICA

Capítulo 2

.

ANÁLISE DA RETENÇÃO NO ENSINO SUPERIOR

Este capítulo, aborda as principais pesquisas já realizadas na literatura para a identificação e previsão da retenção em instituições de ensino superior.

Na primeira parte desta revisão procurou-se entender o conceito do termo retenção usado internacionalmente e nacionalmente no intuito de definir o conceito de retenção abordada nesta pesquisa e identificar as diferenças entre os conceitos nacionais e internacionais.

Após a definição dos conceitos utilizados na retenção, alguns trabalhos internacionais que utilizam mineração de dados e/ou modelos probabilísticos são listados, porém, estes não foram detalhados diante do sentido diferente do termo retenção entre os trabalhos internacionais e nacionais.

2.1 RETENÇÃO NO ENSINO SUPERIOR

No intuito de compreender os estudos que abordam a retenção no ensino superior de instituições nacionais e internacionais foi considerado essencial revisar as definições e os conceitos apresentados pelos autores a fim de estabelecer uma definição mais adequada ao propósito deste estudo.

Para isso, utilizou-se palavras chaves em repositório de artigos científicos como: Elsevier, IEEE Xplore, Science Direct, Google Scholar e Springer para encontrar artigos relacionados. As palavras chaves utilizadas foram: "student retention", "student attrition", "student progression", "student retention". Além disso, buscou-se trabalhos nacionais através do Google Scholar e em eventos relevantes da área com as chaves de busca: "retenção", "mineração de dados" & "retenção", "permanência prolongada", "redes bayesianas" & "retenção".

Os autores (LENNING et al., 1980) *apud* (PEREIRA, 2013) desenvolveram um estudo denominado "*Retention and Attrition: evidence for action and research*" visando esclarecer os diversos conceitos de retenção (*retention*) e de desgaste ou abandono (*attrition*). Segundo os autores, a retenção (*retention*) pode ser definida como aquela que ocorre quando os alunos completam, continuam ou retomam seus estudos, ou seja, quando

a universidade conseguiu reter o aluno até ele concluir o seu objetivo, enquanto o desgaste (*attrition*) ocorre quando os estudantes já não estão matriculados em faculdade ou universidade.

Autores como (PEREIRA, 2013) definem a retenção como a rematrícula progressiva na faculdade, de um período para o próximo, garantindo que o aluno não evada do curso.

O autor (SEIDMAN, 2005) define *retetion* como a realização de objetivos acadêmicos e/ou pessoais do estudante, independentemente do tempo que o aluno leva para concluí-lo.

Dessa forma, pode-se concluir que nos estudos internacionais, o termo retenção tem uma conotação positiva, geralmente referindo-se à permanência do estudante na universidade até o alcance de seu objetivo, independente do tempo necessário. Esse termo é semelhante ao termo evasão utilizado no Brasil.

No Brasil o termo retenção tem sido utilizado em alguns estudos, porém, diferente dos estudos internacionais, a conotação do termo retenção é predominantemente negativa. Na pesquisa bibliográfica realizada por (PEREIRA, 2013) mais a pesquisa realizada neste trabalho foram encontrados 19 trabalhos que apontam qual definição de retenção foi adotada. Também foi visto que alguns trabalhos adotaram o termo "permanência prolongada" como sinônimo de retenção no intuito de diferenciar do significado internacional que tem um sentido positivo. Diante disto, esta revisão da literatura a respeito da retenção em trabalhos nacionais estendeu o trabalho de (PEREIRA, 2013) adicionando os seguintes trabalhos não abordados pelo autor ou mais atuais encontrados através das buscas já citadas: (NORONHA; CARVALHO; SANTOS, 2001), (DIAS; LINS, 2010), (SOARES; FUNDÃO, 2006), (RIOS; SANTOS; LIMA, 2003), (VIEIRA, 2014), (GENEVOIS; LYRA; LIMA, 2008). A Tabela 2.1 apresenta o conjunto de definições e conceitos de retenção adotados em estudos brasileiros.

A maior parte dos trabalhos apresentados na Tabela 2.1 refere-se a retenção no sentido do aluno que permanece prolongadamente na instituição após o tempo necessário para concluir o seu curso. No entanto, mesmo com a definição adotada pelo MEC sobre retenção no Brasil, as pesquisas não se adequam a este conceito, definindo um conceito de retenção baseado no contexto e nas variáveis do seu estudo.

Diante do relatório publicado em (EVASÃO, 1997) e no estudo realizado por (VASCONCELOS; SILVA, 2012), o aluno só é caracterizado como retido ao ultrapassar o prazo máximo de integralização do curso e continuar matriculado. Porém, esse método de avaliação não permite que a instituição tenha um diagnóstico dos níveis de retenção em tempo hábil para adotar ações preventivas e corretivas, comprometendo o potencial de intervenção por parte de cada colegiado.

Diante destas limitações, (NEY, 2010) *apud* (PEREIRA, 2013) adaptou a definição de retenção onde considerou como retido o aluno que permaneceu matriculado no curso por prazo superior à soma do tempo previsto na matriz curricular mais o número de períodos letivos disponíveis para o trancamento. No entanto, a inclusão de períodos de trancamento no prazo para definir o aluno como retido implica em aceitar um período de ociosidade da vaga. Nesse sentido, a autora (POLYDORO, 2000) fez um estudo sobre o trancamento dos alunos e constatou que mais de 90% dos estudantes que trancam a matrícula indicaram ao sair que pretendiam concluir sua formação, porém apenas 10% destes alunos retornaram aos cursos. Diante disto a constatação do aluno retido somada

Tabela 2.1 Definição e Conceitos de Retenção ou Permanência Prolongada adotados em estudos nacionais

Conceito de Retenção	Definição	Referência
Tempo de permanência > prazo máximo de integralização	Condição do aluno que apesar de esgotado o prazo máximo de integralização curricular fixado pelo Conselho Federal de Educação ainda não concluiu o curso, mantendo-se matriculado na universidade.	(EVASÃO, 1997)
	Condição do aluno que após o período máximo de integralização curricular ainda se mantêm matriculado no curso.	(VASCONCELOS; SILVA, 2012)
Tempo de permanência > prazo máximo de integralização	Ultrapassagem ou superação do tempo de permanência no curso para além daquele previsto para a sua integralização curricular.	(SANTOS; NASCIMENTO; RIOS, 2000)
	Situação de prolongamento de curso em que o tempo de titulação é maior que o preestabelecido.	(NORONHA; CARVALHO; SANTOS, 2001)
	Permanência prolongada em um curso que ocorre quando o aluno completa o curso em um tempo maior do que aquele planejado pelo currículo.	(CORRÊA; NORONHA; MIURA, 2004)
	Condição do aluno que leva um tempo maior para completar o curso do que o planejado no currículo ou projeto pedagógico.	(CISLAGHI et al., 2008), (VASCONCELOS; SILVA, 2012), (DIAS; LINS, 2010)
	Condição do aluno que inicia um curso, mas não consegue terminar no tempo projetado.	(DIAS; CERQUEIRA; LINS, 2009)
	Situação do aluno que permanece matriculado no curso mesmo após o tempo suficiente para concluí-lo. O tempo suficiente é a soma do tempo previsto na matriz curricular do curso mais o número de períodos letivos disponíveis para trancamento.	(NEY, 2010)
	Condição do aluno que não conclui o curso dentro da duração normal ou que faz trancamento de matrícula, mesmo que tenha ingressado há pouco tempo no curso.	(CAMPELLO; LINS, 2008)
	Condição do aluno que não concluiu no tempo médio de conclusão do curso.	(SOARES; FUNDÃO, 2006)
	Não definido, mas deixa claro no texto a condição do aluno que permanence no curso após o tempo de conclusão.	(RIOS; SANTOS; LIMA, 2003), (VIEIRA, 2014), (VASCONCELOS; SILVA, 2012), (GENEVOIS; LYRA; LIMA, 2008)
Quando há reprovação em disciplinas	Condição do estudante que em função da não obtenção do conceito mínimo de aprovação nas avaliações escolares é reprovado.	(LAUTERT; ROLIM; LODER, 2011)
	Condição do aluno que reprovar por nota ou falta em uma ou mais disciplinas ou reprovar na disciplina essencial.	(RISSI; MARCONDES, 2011)
Aluno que está matriculado	Condição do aluno regularmente matriculado no curso de origem quando da realização do estudo.	(SANTOS, 1999)

aos possíveis períodos de trancamento contrapõem à definição proposta por (NEY, 2010).

Os autores (CAMPELLO; LINS, 2008) caracterizam o trancamento como elemento causador da retenção, visto que os alunos que trancam o semestre ou disciplinas não conseguem concluir o curso no prazo previsto. Outras abordagens similares adotadas por (SANTOS; NASCIMENTO; RIOS, 2000), (LAUTERT; ROLIM; LODER, 2011), consideram como elemento causador da retenção a reprovação em disciplinas e por isso os estudos são desenvolvidos em torno deste aspecto.

A definição do prolongamento do aluno no curso após o tempo previsto para concluir o seu curso foi predominante nas pesquisas no tocante da retenção como pôde ser visto na Tabela 2.1.

Neste trabalho, foi utilizado como causas de retenção as reprovações dos alunos em disciplinas durante o andamento do curso. Diante disto, a análise foi realizada utilizando os alunos vinculados a grade curricular mais atual para que os resultados obtidos fossem os mais representativos possíveis, o que implica na redução do número de informações possíveis a serem analisadas dado a diferença entre os componentes curriculares das grades curriculares mais antigas. Além disto, alguns cursos analisados no PROUFBA ainda não tinham um tempo mínimo suficiente para que fosse possível um aluno completar o curso.

Nesse sentido, as definições de retenção abordadas até aqui afetam dois pontos importantes nesta pesquisa: uma maior redução no número de informações, dado que essas definições precisam ter um tempo mínimo para que os alunos possam concluir o curso, e o ponto mais crítico seria inviabilizar a análise de cursos novos que não apresentam tempo suficiente para que alunos os alunos possam completar o curso.

Tabela 2.2 Mineração de Dados ou Modelos Probabilísticos aplicados a retenção

Autores	Conceito de Retenção	Técnicas	Tipos de Dados	Conjunto de Dados	Resultados
(Dias, et, al.2010)	Ultrapassa o tempo regular de conclusão	Cadeias de Markov	quantitativos	Período: 2000 a 2005 240 alunos	Média de retenção, 3 semestre; 67% prob. de formar 10% prob. de jubilar 23% prob. de evadir
(Campelo, et, al. 2008)	Ultrapassa o tempo regular de conclusão	Clustering	socioeconômicos e aproveitamento do aluno	Período: 2000 a 2006 280 alunos	Grupos: Excelente, Bons, regulares, fracos, péssimos e desinteressados
(Silva, 2013)	Ultrapassa o tempo regular de conclusão	regressão logística, redes neurais e árvores de decisão	socioeconômicos e aproveitamento do aluno	Período: 1998 a 2011 85467 alunos	possível estratégia de combate ao problema via aconselhamento dos alunos
(Genevois, et al. 2008)	Ultrapassa o tempo regular de conclusão	Estatística Descritiva	dados pessoais e notas do vestibular	Não informado	Conhecimento básico do aluno é o critério mais relevante

Diante disto, foi necessária realizar adaptações nestas definições para o trabalho proposto. No PROUFBA e no primeiro experimento desenvolvido neste trabalho a seguinte definição de retenção foi conceituada e utilizada: **para cada disciplina recomendada para um estudante em um determinado semestre, se o estudante foi reprovado em pelo menos um dos pré-requisitos de uma disciplina recomendada que ele não se inscreveu, este aluno é caracterizado como retido**.

No segundo experimento, no intuito de analisar a retenção final do aluno de acordo com resultados obtidos em disciplinas e/ou na retenção dos alunos em cada semestre, utilizou-se o conceito de retenção mais utilizada pelos autores até aqui citados: **o aluno retido é aquele que ultrapassa o tempo de conclusão estipulado pelo curso**.

Em relação a estas definições, deve-se observar que a avaliação da retenção pode ser realizada ao longo de todo o período da graduação, permitindo intervenções institucionais ainda nos primeiros semestres dos cursos, além de compreender os diversos elementos envolvidos no processo de retenção.

2.2 MINERAÇÃO DE DADOS E REDES PROBABILÍSTICAS NA RETENÇÃO

Na revisão bibliográfica realizada para encontrar trabalhos que utilizam técnicas de mineração de dados e/ou modelos probabilísticos na retenção dos alunos de graduação, pôde ser visto que existem alguns trabalhos internacionais com o objetivo de predizer e intensificar a retenção dos alunos nas universidade, como: (NANDESHWAR; MENZIES; NELSON, 2011), (ZHANG et al., 2010),(PITTMAN, 2008), (YADAV; BHARADWAJ; PAL, 2012), (YU et al., 2010).

Diante da diferença no sentido de *retention* e retenção nos trabalhos internacionais e nacionais, apenas são detalhadas as pesquisas nacionais relacionadas ao conceito de retenção que utilizam a mineração de dados e/ou modelos probabilísticos para analisar e predizer a retenção que se assemelha a definição e o sentido da retenção proposta neste trabalho. De forma concisa a Tabela 2.2 apresenta os trabalhos encontrados.

O trabalho de (DIAS; LINS, 2010) que fez uso do modelo de cadeia de Markov, caso particular de processo estocástico com tempo discreto que apresenta a propriedade de que os seus estados anteriores são irrelevantes para a predição dos estados seguintes, desde que o estado atual seja conhecido (GOMES; WANKE, 2008). Os autores desenvolveram um modelo a partir das probabilidades dos estados criados para representar os alunos no curso (matriculado, ausente, trancamento, desvinculado, formado e jubilado) e utilizaram os

dados do curso de Engenharia de Produção para os ingressantes de 2000 a 2005. Segundo os autores, os resultados obtidos pelo modelo refletem com os dados reais de evasão e retenção do curso estudado (DIAS; LINS, 2010).

Um estudo realizado por (CAMPELLO; LINS, 2008) utiliza a mineração de dados para analisar a evasão e retenção no curso de engenharia de produção entre os anos 2000 a 2006. Foram utilizados dados sócio-econômico-culturais e os dados de desempenho acadêmico do aluno. Com isso os autores apresentaram que o curso tem cerca de 48,6% de alunos com problemas de evasão ou retenção dentre os 280 alunos ingressantes do curso. Com a utilização da técnica de *clustering* da mineração de dados, os autores dividiram os alunos em seis grupos de alunos evadidos ou retidos, a saber: excelentes, bons, regulares, fracos, péssimos e desinteressados. Para cada grupo foi realizada uma análise detalhada com características de cada grupo, além de algumas características gerais. A partir da análise realizada, foi possível formular estratégias no combate a evasão e retenção nas engenharias da Universidade Federal de Pernambuco (UFPE) .

No trabalho realizado por (GENEVOIS; LYRA; LIMA, 2008) apud (SILVA, 2013), utilizou-se a estatística descritiva para encontrar relações entre as notas do vestibular, sexo e semestre de entrada dentre outros fatores de evasão/retenção nos cursos de engenharia da UFPE. Para cada variável foi realizada uma estratificação em razão do percentual de formandos por curso. O autor chegou à conclusão de que os cursos com menores notas para aprovação possuíam maiores taxas de evasão, embora que nos cursos de alta demanda como as engenharias eletrônica e de produção, menos da metade dos alunos ingressantes se formava. Para a retenção, também foi concluído pelo autor que o conhecimento básico do aluno é um critério relevante.

A pesquisa desenvolvida por (SILVA, 2013) utiliza técnicas de mineração de dados para identificar e estimar riscos de retenção ou evasão dos alunos a partir de informações sócio-econômico-culturais e do seu aproveitamento acadêmico. Os autores utilizaram regressão logística, redes neurais e árvores de decisão. Para cada técnica os autores dividiram o conjunto de dados em partes para serem utilizados, a primeira parte contendo apenas dados sócio-econômico-culturais, na segunda, apenas dados acadêmicos e uma terceira com todos os dados. Para cada experimento, os autores apresentaram os resultados obtidos. Após os resultados obtidos pelos autores, eles analisam uma possível estratégia de combate ao problema via aconselhamento dos alunos que poderia produzir uma redução de custos e com isso um aumento na economia da universidade.

A análise da retenção no Brasil ainda é baixa, poucos trabalhos foram apresentados desde a primeira proposta de definição de retenção no Brasil em 1997 (EVASÃO, 1997). Mais especificamente, na análise da retenção através de técnicas de mineração de dados, de acordo com os conhecimentos adquiridos, observou-se que dois trabalhos tem destaque nacional: (CAMPELLO; LINS, 2008), (SILVA, 2013).

A respeito de modelos probabilísticos para a análise da retenção, considerando o estado da arte deste trabalho, não se tem conhecimento de pesquisas que utilizem esta abordagem. Diferente dos trabalhos aqui apresentados, o estudo de caso apresentado procura entender o fluxo acadêmico do aluno a partir dos resultados obtidos em disciplinas em um determinado semestre em uma certa tentativa.

Capítulo 3

.

PROBABILIDADE

A teoria da probabilidade teve sua origem em jogos de azar. No entanto, esta associação com o jogo contribuiu muito pouco para o crescimento da teoria da probabilidade como uma disciplina matemática (ROHATGI; SALEH, 2011). A primeira tentativa com rigor matemático é creditada a Laplace (LAPLACE, 1812). Laplace deu a definição clássica da probabilidade: se os eventos elementares são igualmente prováveis, a probabilidade de um evento A é igual ao quociente entre os números de casos favoráveis ao evento A e o número de casos possíveis.

Em 1933, a probabilidade em eventos aleatórios foram definidos (KOLMOGOROV, 1950) e representados por conjuntos. Assim a probabilidade é apenas uma medida normalizada nestes conjuntos.

O objetivo da teoria da probabilidade é proporcionar métodos para descrever e analisar experiências com resultados aleatórios. Esses experimentos tem o objetivo de observar os "eventos" ou "magnitudes aleatórios", bem como o cálculo da probabilidade de que tais acontecimentos ocorram, ou os valores esperados de tais magnitudes (BAUER, 1996).

3.1 MODELO PROBABILÍSTICO

Um modelo probabilístico é composto por três essenciais elementos, definidos como (ASIMOW; MAXWELL, 2010):

- **experimento**: é qualquer atividade ou processo cujo resultado está sujeito à incerteza. Assim, experimentos que podem ser de interesse incluem: jogar uma moeda uma ou várias vezes; determinar o tempo de deslocamento de casa para o trabalho; obtenção de tipos de sangue a partir de um grupo de indivíduos.

- **espaço amostral**: conjunto de todos os resultados possíveis de um experimento, representado por S. Por exemplo, se o experimento é jogar uma moeda para cima, o espaço amostral é cara/coroa; se o experimento é lançar um dado, o espaço amostral é 1, 2, 3, 4, 5, 6.

  - **eventos**: subconjunto do espaço amostral S, ou seja, uma coleção de resultados possíveis. Cada possível resultado é denominado de *ponto* ou *elemento* de S. Por exemplo, ao lançar um dado, pode-se ter eventos como (1, 2, 3, 4, 5, 6), a ocorrência de uma face par (2,4,6) ou ímpar (1,2,3).

- **probabilidades** probabilidade de um elemento ocorrer ao longo de uma escala de 0-100. Estes elementos devem satisfazer a propriedade em que a soma das probabilidades deve ser igual a 1 para cada evento.

3.2 DEFINIÇÃO DA PROBABILIDADE

Dentre os elementos apresentados de um modelo probabilístico, as probabilidades podem ser definidas de duas formas: frequência relativa e clássica. Além disto, estas probabilidades devem satisfazer os axiomas, ou seja, propriedades mínimas que devem ser satisfeitas pela probabilidade de qualquer evento (KOLMOGOROV, 1950).

Dado um espaço amostral S, onde para cada evento A de S assume que a probabilidade de A, formalmente descrita como P(A), deve satisfazer os seguintes axiomas:

1. Axioma 1: $0 \leq P(A) \leq 1$, ou seja, a probabilidade de A é um número entre 0 e 1;

2. Axioma 2: P(S) = 1, a probabilidade de que algum evento elementar em todo o espaço amostral irá ocorrer é 1. Mais especificamente, não há eventos elementares fora do espaço amostral.

3. Axioma 3: para qualquer conjunto de n eventos mutuamente exclusivos definido no mesmo espaço de amostras, a probabilidade de pelo menos um desses eventos ocorridos, é a soma das suas respectivas probabilidades, formalmente:

$$P(A_1 \cup A_2 \cup ...A_n) = P(A_1) + P(A_2) + ...P(A_n) \tag{3.1}$$

A definição da frequência relativa de uma probabilidade é formalmente apresentada da seguinte forma: dado um experimento que é executado n vezes, se um evento A ocorrer n_a vezes, então P(A), é definido como:

$$P(A) = \lim_{n \to \infty} \frac{n_a}{n} \tag{3.2}$$

A definição clássica, utilizada em experimentos onde o espaço amostral é finito, é descrito da seguinte forma: dado um experimento que tem um espaço amostral S, a probabilidade de um evento A ocorrer, P(A), é dada pela razão entre a ocorrência de A em relação a todos os outros eventos ocorridos, formalmente:

$$P(A) = \frac{n_A}{n(S)} \tag{3.3}$$

Em um exemplo simples ao contexto estudado, no curso de ciência da computação existem inúmeros alunos cursando disciplinas em semestres distintos. De acordo com a Tabela 3.1, deseja-se calcular a probabilidade do aluno cursar a disciplina 1 (*Disc1*).

Tabela 3.1 Quantidade de alunos cursando disciplinas em um determinado semestre

Semestre/Disciplina	Disc1	Disc2	Disc3	Disc4	Total
1° Semestre	35	53	52	46	186
2° Semestre	84	72	42	20	218
3° Semestre	120	85	32	9	246
Total	239	210	126	75	650

De acordo com a Tabela 3.1 existem 650 alunos que cursam disciplinas em semestres distintos, no qual 239 cursam a *Disc1*, 210 *Disc2*, 126 *Disc3* e 75 *Disc4*. A probabilidade do aluno cursar *Disc1* é:

$$P(Disc1) = \frac{239}{650} = 0,367 \tag{3.4}$$

Logo, pode-se afirmar que a probabilidade do aluno cursar *Disc1* é 36,7%, *Disc2* é 32,3%, *Disc3* é 19,3% e a probabilidade de *Disc4* pode ser calculada como P($Disc4$) = 1 - $P(Disc1 \cup Disc2 \cup Disc3) = 11,7\%$, onde $P(Disc1 \cup Disc2 \cup Disc3)$ é o complemento de $P(Disc4)$. Esta propriedade e as outras foram definidas a partir dos axiomas da probabilidade que são descritos a seguir.

Dado um espaço amostral S e os eventos A e B. Tem-se:

(P1) $P(A) = 1 - P(A^c)$, onde $P(A^c)$ é o complemento de $P(A)$, ou seja, todos os resultados possíveis que não satisfazem A;

(P2) $P(B) = P(B \cap A) + P(B \cap A^c)$;

(P3) Se $A \subset B$ então $P(A) \leq P(B)$;

(P4) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;

Caso seja necessário calcular a probabilidade do aluno cursar *Disc1*, dado que ele esteja no seu segundo semestre, faz-se necessário a utilização das probabilidades condicionais, ou seja, as probabilidades condicionadas a eventos anteriores.

3.3 PROBABILIDADE CONDICIONAL

Dado que se tenha o conhecimento da probabilidade de um evento X ter ocorrido e deseja-se calcular a probabilidade de um evento Y ocorrer, condicionado ao evento X, utiliza-se a probabilidade condicional. Uma maneira simples de pensar sobre probabilidades condicionais dado Y é imaginar que o universo de eventos S encolheu para Y. De acordo com (KORB; NICHOLSON, 2010), a notação utilizada para representar a probabilidade condicional é a seguinte:

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} \tag{3.5}$$

A partir da definição da probabilidade condicional, pode-se calcular a probabilidade do aluno cursar a disciplina 1 (*Disc1*), dado que ele esteja no seu segundo semestre. De acordo com a Tabela 3.1, 239 alunos cursam *Disc1*, no qual 35 cursam no primeiro semestre, 84 no segundo e 120 no terceiro. Logo, a probabilidade pode ser calculada da seguinte forma:

$$P(Disc1|2Semestre) = \frac{(84/650)}{218/650} = \frac{0,129}{0,335} = 0,385 \tag{3.6}$$

Em uma probabilidade condicional $P(X|Y)$, se esta probabilidade for diferente de $P(X)$, implica que o evento X é condicionalmente dependente do evento Y, pois a sua probabilidade é alterada dada a ocorrência do evento Y. Nos casos em que a probabilidade $P(X|Y) = P(X)$, defini-se que os eventos X e Y são independentes, pois a ocorrência de um evento não altera a probabilidade da ocorrência do outro. A relação de variáveis dependentes (condicional) e independentes (não-condicional) serão abordadas com mais ênfase no Capítulo 5.

3.4 TEOREMA DE BAYES

Teorema de Bayes é o fundamento da inferência bayesiana no qual, mostra a relação entre uma probabilidade condicional e sua inversa. Este teorema foi publicado após a morte de Thomas Bayes (1702-1761) em sua obra (BAYES; PRICE, 1763), e também é conhecido como Regra de Bayes ou Leis de Bayes. O teorema de bayes expressa a probabilidade condicional ou probabilidade *a posteriori*, de um evento Y dado X, observando a probabilidade *a priori* de Y, e a probabilidade condicional de X dado Y. Formalmente, é descrito pela seguinte fórmula:

$$P(Y|X) = \frac{P(X|Y) * P(Y))}{P(X)} \tag{3.7}$$

Uma analogia ao teorema de bayes são exames solicitados nos diagnósticos médicos para confirmar ou não uma doença, que pode ser feita através da seguinte comparação: o que o especialista pensa antes da realização do exame é a probabilidade *a priori* e o que os especialistas pensam depois é a probabilidade *a posteriori*.

Os autores (RUSSELL; NORVIG, 1995) utilizam um exemplo de um caso médico para exemplificar o teorema de bayes: "um médico sabe que a meningite causa torcicolo em 50% dos casos. Porém o médico também conhece algumas probabilidades não-condicionais que dizem que um caso de meningite atinge 1/50000 pessoas, e a probabilidade de alguém ter torcicolo é de 1/20. Deseja-se calcular, utilizando o teorema de Bayes, a probabilidade de um paciente que esteja com torcicolo ter meningite".

Considerando T como a probabilidade não-condicional de um paciente ter torcicolo, tem-se que P(T) = 1/20. Considerando M como a probabilidade não-condicional de um paciente ter meningite tem-se que P(M) = 1/50000. A probabilidade condicional de que um paciente que tenha meningite apresente torsicolo se dá por P(T—M) = 0,5.

Aplicando o teorema de bayes:

$$P(M|T) = \frac{P(T|M) * P(M))}{P(T)} = \frac{P(0,5) * P(1/50000))}{P(1/20)} = 0,0002 \qquad (3.8)$$

Ou seja, é esperado que apenas 1 em 5000 pacientes com torcicolo tenha meningite. Note que mesmo tendo torcicolo uma alta probabilidade nos casos de meningite (0,5), a probabilidade de um paciente ter meningite continua pequena, devido ao fato de a probabilidade não-condicional de torcicolo ser muito maior que a probabilidade de meningite (RUSSELL; NORVIG, 1995).

Uma argumentação válida surge do fato de que o médico poderia também possuir a probabilidade não-condicional $P(M|T)$, a partir de amostras de seu universo de pacientes, da mesma forma que $P(T|M)$, evitando o cálculo realizado anteriormente. Porém, imagine que um surto de meningite aflija o universo de pacientes do médico em questão, aumentando o valor de $P(M)$. Caso $P(M|T)$ tenha sido calculado estatisticamente a partir de observações em seus pacientes, o médico não terá nenhuma ideia de como este valor será atualizado (visto que $P(M)$ aumentou). Entretanto, caso tenha realizado o cálculo de $P(M|T)$ em relação aos outros três valores (como demonstrado) o médico verificará que $P(M|T)$ crescerá proporcionalmente em relação a $P(M)$ (RUSSELL; NORVIG, 1995).

As probabilidades aqui apresentadas, obtidas através das inferências baseadas no Teorema de Bayes, é um dos tipos de informações que podem ser obtidas quando não se tem certeza da informação, ou seja quando as informações disponíveis não apresentam todos os dados, variáveis, aspectos necessários para que se obtenha um diagnóstico perfeito, ou seja, verdadeira ou falso.

3.5 RACIOCÍNIO SOB INCERTEZA

Normalmente as informações reais são incertas ou incompletas, dificultando em alguns casos a utilização de modelos lógicos que admitem apenas valores verdadeiros ou falsos. Nesses casos, é preferível a utilização de teorias que atuem com a incerteza ou a imprecisão das informações.

De acordo com (KLIR; FOLGER, 1988), incerteza origina-se de alguma deficiência da informação. A informação pode estar incompleta, ser vaga, imprecisa ou contraditória. A Tabela 3.2 apresenta aspectos que podem causar a incerteza na informação:

Analisando a Tabela 3.2, pode-se concluir que quando a informação é perfeita, a lógica pode ser utilizada para resolver o problema, porém geralmente as informações reais não são precisas. Segundo (NASSAR, 2003), para dar tratamento a essas incertezas é recomendada a utilização de teorias como fuzzy sets, probabilidades, teoria da evidência, entre outras.

De acordo com (CHARNIAK, 1991), a principal vantagem do raciocínio probabilístico sobre raciocínio lógico (true or false) é o fato de que agentes podem tomar decisões racionais mesmo quando não existe informação suficiente para se provar que uma ação funcionará.

Este capítulo foi descrito com um intuito de apresentar os conceitos básicos de probabilidade, bem como o Teorema de Bayes, teorema base para a existência das redes

Tabela 3.2 Aspectos que podem causar a incerteza na informação

Tipo de Informação	Que horas o ônibus passa por aqui?
Perfeita	O ônibus passa às 08h e 15mn.
Imprecisa	O ônibus passa entre 08h e 09h.
Incerta	Eu acho que o ônibus passa às 08h.
Vaga	O ônibus passa lá pelas 08h.
Probabilista	É provável que o ônibus passe às 08h.
Possibilista	É possível que o ônibus passe às 08h.
Inconsistente	Fulano disse que passa às 08h,
	mas Cicrano disse que passa às 10h.
Incompleta	Eu não sei que horas o ônibus passa,
	mas geralmente os ônibus passam aqui às 08h.
Ignorância	Não faço a menor ideia do horário do ônibus.

bayesianas. Com isso, espera-se que se tenha uma melhor compreensão dos conceitos que serão abordados no capítulo seguinte sobre as redes bayesianas.

Capítulo

4

.

REDES BAYESIANAS

O termo Rede Bayesiana (RB) surgiu na década de 80, especificamente nas áreas de inteligência artificial e pesquisa operacional para denominar um tipo específico de modelo probabilístico que representa relações de dependência entre um conjunto de variáveis aleatórias (PEARL, 1988).

Segundo (SOBERANIS, 2010), essas redes são apropriadas para modelar processos causais com incerteza e oferecem potencial de modelar rupturas de suprimento de forma efetiva. As probabilidades desta rede podem ser obtidas com fatos do passado para estimar a condição futura ou ainda com opiniões subjetivas de especialistas.

Uma rede Bayesiana pode ser compreendida imaginando situações onde existem efeitos de causalidades e o entendimento sobre as causalidades é incompleto, então é necessário descrever a situação probabilística (CHARNIAK, 1991). Assim, as redes Bayesianas são mais frequentemente construídas utilizando as noções de relação de causa e efeito (MADSEN et al., 2005).

Esses modelos possuem como componentes uma estrutura qualitativa, a qual representa as dependências entre os nós (arcos), e quantitativa, que avalia, em termos probabilísticos, essas dependências.

Segundo (RUSSELL; NORVIG, 1995) uma rede Bayesiana consiste em:

1. Um conjunto de variáveis $x = x_1, x_2, ...x_n$ e um conjunto de arcos entre as variáveis;

2. Cada variável x_i possui um conjunto finito e limitado de estados mutualmente exclusivos ;

3. Os nós e os arcos foram um grafo orientado e acíclico;

4. Para cada nó A que possui como pais $y_1, y_2, ...y_n$ existe uma tabela de probabilidade condicional (TPC) , no qual cada linha nesta tabela contêm a probabilidade condicional para cada caso condicional dos nós pais, ou seja $P(x|y_1, y_2, ...y_n)$.

Levando em consideração a definição 1, pode-se definir a rede Bayesiana como modelo gráfico probabilístico que representa um conjunto de variáveis $x = x_1, x_2, ...x_n$ e suas dependências probabilísticas. A estrutura gráfica da RB permite que a relação probabilística possa ser representada por um grande número de variáveis. A estrutura da rede S é um gráfico acíclico dirigido (DAG) , onde os nós representam as variáveis e os arcos representam as dependências condicionais entre as variáveis. Quando uma variável está condicionada a outra (existe um arco de uma variável para outra, $A \rightarrow B$), diz-se que A é pai de B, pois a probabilidade de B está condicionada a probabilidade do seu pai A. Quando uma variável não recebe nenhum arco diz-se que esta variável é independente, ou seja, a probabilidade de sua ocorrência independe das outras variáveis. Para o espaço de probabilidade (S, P), a distribuição da probabilidade local P, é igual ao produto das suas distribuições condicionais de todos os nós do grafo condicionados as variáveis que correspondem aos seus nós pai (HECKERMAN, 1998). Formalmente, para um gráfico acíclico de N variáveis aleatórias x, a junção distribuição da probabilidade conjunta é dada pela fórmula:

$$p(x) = \prod_{i=1}^{N} p(x_i|pais(x_i)) \tag{4.1}$$

onde pais(x_i) indica as variáveis que são pais do nó i e $x = x_1, x_2, ...x_n$. No caso em que o nó i não tem pai, a probabilidade associada com a variável x_i é reduzida a probabilidade não-condicional tal que $p(x_i|pais(x_i)) = p(x_i)$.

Por exemplo, a Figura 4.1, apresenta uma rede Bayesiana com os nós A, B e C e suas probabilidades condicionais e não-condicionais, onde B tem A como nó Pai e C tem B. Logo, a probabilidade de A tem influência na probabilidade de B e C é influenciado diretamente por B e indiretamente por A. Como descrito na Seção 3.2, $P(X^c)$ é o complemento de $P(X)$, ou seja, todos os resultados possíveis que não satisfazem X, sendo x uma variável da rede. Para calcular a probabilidade de cada variável, pode-se utilizar a fórmula da junção da distribuição como um produto das distribuições condicionadas de cada variável (exemplo adaptado do (SOBERANIS, 2010)):

$$p(x_1, ...x_N) = p(x_N|x_1, ..., x_{N-1})...p(x_2|x_1)p(x1) \tag{4.2}$$

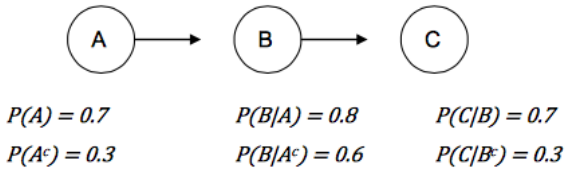

Figura 4.1 Exemplo I de uma rede bayesiana
(SOBERANIS, 2010)

Como ilustrado na Figura 4.1, para cada nó da rede é calculada a probabilidade condicionada ao nó pai correspondente. Os arcos representam a influência de A em B e

de A e B em C. Baseado nas probabilidades associadas de cada nó e repetindo a aplicação da distribuição da conjunção de probabilidades, pode-se calcular a probabilidade de B e C dadas as probabilidades de seus pais, aplicando o teorema de Bayes:

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c) = (0,8)(0,7) + (0,6)(0,3) = 0,74$$
$$P(C) = P(C|B)P(B) + P(C|B^c)P(B^c) = (0,7)(0,74) + (0,3)(0,26) = 0,596 \quad (4.3)$$

As probabilidades de B e C foram calculadas a partir das informações dos pais passadas pela rede considerando-se as probabilidades dos nós pais utilizando o teorema de Bayes. Pode-se agora reescrever a rede bayesiana com as seguintes probabilidades:

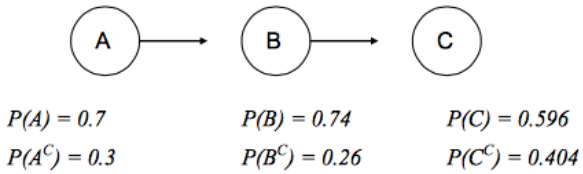

Figura 4.2 Exemplo II de uma rede bayesiana
(SOBERANIS, 2010)

4.1 MODELAGEM DE UMA REDE BAYESIANA

Os autores (LUCAS; GAAG; ABU-HANNA, 2004) definem a modelagem de uma rede Bayesiana em quatro etapas:

1. Seleção de variáveis relevantes: a primeira fase da construção de uma RB é identificar as variáveis importantes que serão utilizadas, juntamente com os valores que podem adotar. A seleção das variáveis relevantes geralmente é baseada em entrevistas com especialistas, descrições do domínio, e uma extensa análise do objetivo da rede em construção. Neste trabalho, foram selecionadas as disciplinas que são pré-requisitos para alguma disciplina em semestre posterior que tenha no mínimo 20 alunos que a cursam em um determinado semestre em certa tentativa.

2. Identificação do relacionamento entre as variáveis: uma vez que as variáveis a serem incluídas na rede foram decididas, as relações de dependência e independência entre elas devem ser analisadas e expressas em uma estrutura gráfica. Geralmente a noção de causalidade é empregada como um princípio orientador: "O que poderia causar este efeito?" e "Quais manifestações esta causa tem?" As relações então são criadas tendo uma relação de causalidade para dirigir os arcos entre as variáveis. A noção de causalidade muitas vezes parece corresponder a maneira de pensar sobre os processos fisiológicos em seu domínio (GAAG; HELSPER, 2002). O relacionamento entre as variáveis também podem ser definidos através de algoritmos de aprendizagem de estruturas que utilizam um conjunto de dados como base do conhecimento para identificar as melhores relações entre os nós. Neste trabalho, a relação entre

as variáveis baseou-se no sistema de pré-requisitos da grade curricular, onde uma variável X depende de uma variável Y, se Y é um pré-requisito para que se possar cursar X.

3. Identificação das probabilidades qualitativas e restrições lógicas: identificar o tipo de distribuição das probabilidades requeridas para a construção da rede. A restrição lógica objetiva limita o universo de probabilidades que devem ser avaliadas. Geralmente, esta etapa consiste em mapear uma base de dados. Neste trabalho, utilizaram-se todos os alunos do curso de Ciência da Computação que foram admitidos entre 2004.1 à 2013.2 vinculados às grades curriculares 2007.2 ou 2008.1.

4. Avaliação das probabilidades: calcular as distribuições de probabilidades condicionais e não-condicionais de cada variável da RB. Essas probabilidades podem ser obtidas a partir de um especialista no domínio. Porém, este levantamento pode ser uma tarefa difícil. Alternativamente, as probabilidades podem ser calculadas a partir de algoritmos de aprendizagem de parâmetros quando já se tem a estrutura da RB definida: Estimação Bayesiana (MICHALSKI; CARBONELL; MITCHELL, 2013) e Estimação via Máxima verossimilhança (MICHALSKI; CARBONELL; MITCHELL, 2013). O método de estimação via Máxima verossimilhança foi utilizado neste trabalho e será detalhado na Seção 4.2.1.

4.2 APRENDIZAGEM EM REDES BAYESIANAS

A modelagem da estrutura de uma RB a partir do conhecimento de um especialista no domínio pode ser difícil e demorado. Além disto, calcular as probabilidades condicionais de uma RB, cujo o número de variáveis seja maior do que dez, pode se tornar uma tarefa complicada com apenas a experiência do especialista.

A aprendizagem de RBs consiste em induzir, a partir de uma amostra de dados, as distribuições de probabilidades simples e condicionais e/ou identificar as relações de interdependência entre as variáveis de um domínio de dados, que se constitui na população de interesse. Este processo de aprendizagem indutiva pode ser de duas formas: aprendizagem da estrutura, quando definida a estrutura da RB a partir de um conjunto de dados, e aprendizagem dos parâmetros numéricos, quando já foi definida a estrutura da RB e precisa-se calcular as probabilidades das variáveis.

4.2.1 Aprendizagem de Parâmetros

A aprendizagem de parâmetros consiste em estimar as probabilidades condicionais de cada variável da rede. Existem dois algoritmos para realizar este aprendizado quando não existem dados nulos na base de dados: estimadores bayesianos e estimadores de máxima verossimilhança (MICHALSKI; CARBONELL; MITCHELL, 2013). Neste trabalho, a aprendizagem de parâmetros foi realizada através dos estimadores de máxima verossimilhança. Com isso, apenas esse será detalhado.

De acordo com (MICHALSKI; CARBONELL; MITCHELL, 2013), a estimação por Máxima Verossimilhança não considera nenhum conhecimento *a priori*, ou seja, suas

Tabela 4.1 Amostra de dados das variáveis mata01dis1_s1 e mata02dis2_s2

Variáveis/Eventos	APR	REP	NC
mata01dis1_s1	288	270	23
mata02dis2_s2	73	89	419

Tabela 4.2 Frequência Relativa da variável mata02dis2_s2 condicionada a mata01dis1_s1

		mata01dis1_s1			
		APR	REP	NC	Total
mata02dis2_s2	APR	0	73	0	73
	REP	0	89	0	89
	NC	288	108	23	419
Total		288	270	23	581

estimativas baseiam-se na frequência relativa e na contagem de eventos das variáveis da base de dados. Como cada variável possui seus estados e possíveis pais dentro da rede bayesiana, a probabilidade estimada é o cálculo da frequência relativa do nó em função de seu pai.

Para um melhor entendimento deste modelo, a estimação por máxima verossimilhança é utilizada para estimar as probabilidades da rede apresentada na Figura 4.3.

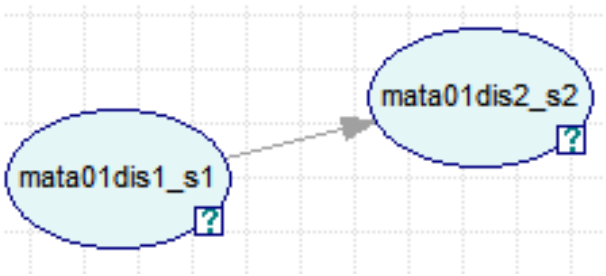

Figura 4.3 Exemplo de uma rede bayesiana com duas variáveis

Para realizar essas estimativas é necessária a utilização da base de dados para identificar a frequência relativa de cada variável. Uma amostra de dados reais das duas variáveis da RB é apresentada na Tabela 4.1.

Onde, APR é aprovado, REP é reprovado e NC Não Cursou a disciplina.

A partir do conjunto de dados apresentado na Tabela 4.1 é criada uma Tabela para identificar a frequência de cada evento da variável condicionada as variáveis que são seus respectivos pais. Para este exemplo, a Tabela 4.2 apresenta as seguintes frequências:

Para calcular as probabilidades de cada estado de uma variável utiliza-se a Tabela 4.2. A estimativa da probabilidade condicional de cada estado x de uma variável Y se da pela razão entre as ocorrências de um estado x de uma variável Y condicionada a um estado z da variável pai H e a ocorrência do estado Z da variável pai H. Formalmente, tem-se:

Tabela 4.3 Tabela da Probabilidade Condicional TPC de mata01dis2_s2

		mata01dis1_s1		
		APR	REP	NC
mata02dis2_s2	APR	0	0,2703	0
	REP	0	0,3296	0
	NC	1	0,4	1

$$P(Y = x|H = z) = \frac{OcorrenciaXeH}{OcorrenciadeH} \tag{4.4}$$

Desta forma, para calcular a probabilidade do aluno ser aprovado em mata02dis2_s2 dado que ele tenha sido reprovado em mata01dis1_s1 utiliza-se a quantidade de alunos que foram aprovados em mata02dis2_s2 e reprovados em mata01dis1_s1 (73) dividido pela quantidade de alunos que foram reprovados em mata01dis1_s1 (270). Assim, a probabilidade é de $73/270 = 0{,}2703$.

No caso do exemplo apresentado, a tabela da probabilidade condicional (TPC) da variável mata02dis2_s2 é apresentado na Tabela 4.3

4.2.2 Aprendizagem da Estrutura

A aprendizagem da estrutura de uma rede bayesiana busca a melhor disposição de dependências e independências entre as variáveis que mais reflita o mundo real a partir de um determinado conjunto de dados.

Segundo (JR, 1997), a estimação de estrutura de uma rede bayesiana, também conhecida na literatura como aprendizado de estrutura, pode ser dividida em duas partes: a primeira baseada em uma busca heurística e a segunda baseada no conceito de independência condicional dos atributos da rede. Assim, algoritmos são requeridos para ambos os tipos de estimação.

De acordo com (ARA-SOUZA, 2010) os algoritmos de busca heurística pesquisam a melhor estrutura com base na busca de uma pontuação adequada, assim, começam com uma rede sem arcos e, gradativamente, adicionam arcos ligando variável a variável, analisando um determinado *score* em cada passagem. Por fim, indica como sendo a melhor estrutura aquela com o máximo score obtido. Uma desvantagem desse tipo de algoritmo é que ele depende diretamente da ordenação inicial das variáveis. Dentre estes algoritmos existem o algoritmo PC e o K2 (RUSSELL; NORVIG, 1995).

Nesta pesquisa, a aprendizagem na estrutura da rede foi utilizada no experimento II baseado no conceito de independência condicional dos atributos da rede através do algoritmo *Naive Bayes* que é descrito na seção seguinte.

4.2.3 Classificador Bayesiano Naive Bayes

Os classificadores bayesianos baseiam-se no Teorema de Bayes. Esses classificadores são facilmente aplicáveis em grandes volumes de dados. A respeito do seu desempenho, em diversos casos, como por exemplo, classificação de textos, os classificadores bayesianos

podem ser mais eficazes que árvores de decisão e redes neurais (HAN; KAMBER; PEI, 2011).

Para predizer as probabilidades de classificação, os classificadores aplicam o Teorema de Bayes para descobrir a probabilidade condicional *a posteriori* $P(B|A)$ de um determinado evento B ocorrer dado a ocorrência do evento A. Com a aplicação do Teorema de Bayes é possível descobrir a probabilidade *a posteriori* a partir do conhecimento de estimativas das probabilidades de ocorrência de A e B (probabilidades *apriori*) e do conhecimento da probabilidade condicional $P(A|B)$

O classificador bayesiano *Naive Bayes* parte da hipótese que todos as variáveis são independentes entre si e todos eles são dependentes da variável classificadora. A representação deste classificador é apresentada na Figura 4.4

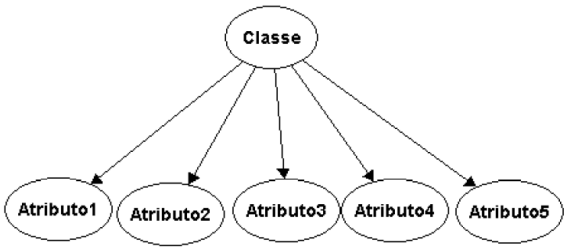

Figura 4.4 Estrutura do Naive Bayes com 5 atributos e uma variável classe (KARCHER, 2009)

Sob a hipótese de independência condicional entre as variáveis dada a classe, é obtida a distribuição conjunta de probabilidades do classificador *Naive Bayes* diante da seguinte fórmula:

$$P(A_1, ..., A_n, C) = P(C) \times \prod_{i=1}^{n} P(A_i|C) \tag{4.5}$$

Dado que um classificador bayesiano tenha apenas atributos discretos e a classe C assumindo valores $0, 1$, a probabilidade de classificar um novo caso, $A_1 = a_1, ..., A_n = a_n$, onde C=1 é calculado como segue:

$$P(C = 1|A_1 = a_1, ..., A_n = a_n) = \frac{P(C = 1) \times P(A_1 = a_1, ..., A_n = a_n|C = 1))}{P(A_1 = a_1, ..., A_n = a_n)} \tag{4.6}$$

Para os casos em que a variável C=0, calcula-se da seguinte forma:

$$P(C = 1|A_1 = a_1, ..., A_n = a_n) = \frac{P(C = 0) \times P(A_1 = a_1, ..., A_n = a_n|C = 0))}{P(A_1 = a_1, ..., A_n = a_n)} \tag{4.7}$$

Após o cálculo dos casos para os dois possíveis resultados (C=0 e C=1), o resultado classificado será o que ocorre com maior probabilidade.

Conhecido por sua simplicidade e eficiência diante da sua estrutura fixa, sua suposição de independência é problemática, visto que esta hipótese raramente se verifica no mundo real. Porém, os classificadores *Naive Bayes* têm apresentado um bom desempenho em um grande número de aplicações, especialmente naquelas em que as variáveis preditoras não são fortemente correlacionadas (CHENG; GREINER, 2001).

4.3 INFERÊNCIA BAYESIANA

Após a definição da estrutura da rede bayesiana e a aprendizagem das probabilidades condicionais e não-condicionais das variáveis, uma das tarefas mais importantes consiste em obter estimativas de probabilidades dos eventos relacionados, à medida que novas informações ou evidências sejam conhecidas. Em redes bayesianas, as inferências podem ser obtidas de três formas:

- causal: das causas para os efeitos, por exemplo, qual a probabilidade de aprovação do aluno em uma disciplina do terceiro semestre dado a aprovação em uma disciplina do primeiro.
- Diagnóstico: dos efeitos para a causas, por exemplo, dado que o aluno não tenha cursado uma disciplina no terceiro semestre, qual a probabilidade de ter sido aprovado na disciplina do primeiro semestre?
- Intercausais: dada a evidência de uma causa, obtêm-se a probabilidade de outra causa, por exemplo, uma disciplina Z do segundo semestre tem X e Y como pré-requisito. Qual a probabilidade de aprovação em X, dado que o aluno tenha sido aprovado em Y?

Estas possíveis inferências são apresentadas na Figura 4.5.

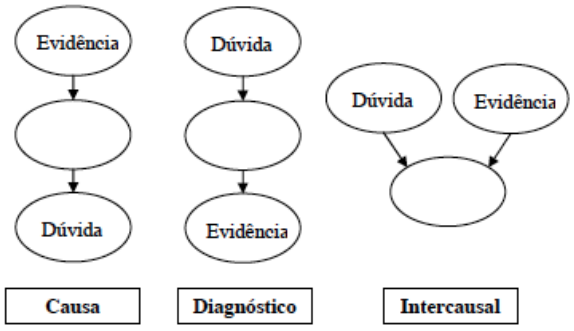

Figura 4.5 Possíveis inferências em redes bayesianas

Existem dois tipos de algoritmos que executam inferência probabilística: aproximado e exatos. O primeiro é baseado em métodos de simulação para inferir a probabilidade com uma precisão aproximada, porém consegue obter os resultados rapidamente. O segundo obtêm as inferências exatas, mas necessita de um esforço computacional elevado quando a rede tem um grande número de variáveis. Dos algoritmos que trazem soluções exatas,

destacam-se o algoritmo de *clustering* (RENOOIJ; GAAG, 2002) e elimição de variáveis (RUSSELL; NORVIG, 1995). Nesta pesquisa, optou-se por utilizar algoritmos exatos para obter resultados mais precisos. Dos algoritmos exatos, o algoritmo de *clustering* foi utilizado nesta pesquisa por ser o mais utilizado em inferências probabilísticas exatas na literatura.

O algoritmo de *clustering* consiste em agrupar as variáveis que tem filhos em comum de forma que esta se torne uma rede simplesmente conexa. Uma rede simplesmente conexa é aquele em que existe no máximo um caminho não direcionado, não importando a orientação dos arcos, entre um par de nós. Em contrapartida, uma rede com múltiplas conexões, possuem pares de nós se conectando por mais de uma combinação de arcos. A Figura 4.6 mostra a conversão de uma rede com múltiplas conexões em uma rede simplificada utilizando o algoritmo de *clustering*.

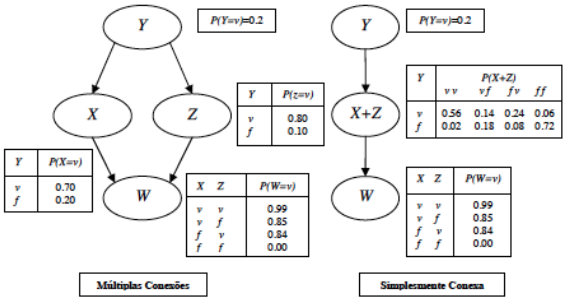

Figura 4.6 Agrupamento de variáveis com o algoritmo de *clustering*

O nó gerado pelo agrupamento dos outros dois nós X e Z representa uma única variável com quatro estados que são combinações dos estados dos nós originais, obtendo assim os seguintes estados: *vv*, *vf*, *fv* e *ff*. O calculo da Tabela de probabilidade condicional do nó gerado (X+Z) se faz pela combinação das probabilidades dos nós originais. Por exemplo, para o caso do novo estado *vv* que representam os estados verdadeiro para X e verdadeiro para Z, multiplica-se $P(X = v|Y = v)$ por $P(Z = v|Y = v)$, ou seja, 0,7 por 0,8. Com isso, é obtido 0,56 como a probabilidade de X e Z serem verdadeiros, dado que Y é verdadeiro. O pseudocódigo do algoritmo é apresentado na Figura 4.7.

0. Definir numeração $\alpha: U \leftrightarrow \{1,2, \ldots, |U|\}$ em função inversa da ordem da eliminação, tal que para $u, v \in U$, se *u* for eliminado após *v*:
   a. Então $\alpha(u) < \alpha(v)$;
1. Seja *v*, a variável de C de maior numeração, tal que as *w*, com numeração menor do que *v*, isto é, $u \in C$ e $\alpha(u) < \alpha(v)$. Se v existir:
   a. Então o índice de C é $\alpha(v)$
   b. Senão é 1.

   Após obter todos os índices, numerar os cliques por ordem crescente de índice, iniciando com 1;
2. Considerar o clique de número 1, isto é, C_1, como a raiz da árvore;
3. Ligar C_k a algum clique C_j, *j*<*k*>1, já na árvore, que contenha $S_k = C_k \cap \cup^{k-1}_{i=1} C_i$. Se existir mais de um, ligar ao de menor índice.

Figura 4.7 Algoritmo *clustering*
(CAMARINHA, 2011)

Este capítulo apresentou conceitos das redes bayesianas que foram utilizados para definição da solução proposta, bem como para a construção dos experimentos e a obtenção dos resultados, apresentados nos capítulos seguintes.

PARTE II

CONTRIBUIÇÕES DA DISSERTAÇÃO

Capítulo

5

.

SOLUÇÃO PROPOSTA

Avaliar a retenção dos alunos em um curso superior é crucial para entender o que está levando os alunos a não concluírem o curso no tempo previsto. Com os resultados obtidos no PROUFBA, percebeu-se que uma análise mais detalhada com a utilização de técnicas probabilísticas poderia contribuir para uma melhor compreensão deste problema.

Com o intuito de analisar probabilisticamente a retenção no curso de Ciência da Computação, este trabalho concentrou-se em dois problemas para serem avaliados:

1. analisar o fluxo acadêmico do aluno, que consiste em verificar probabilidades dos resultados em disciplinas quando cursam em um determinado semestre em uma certa tentativa, bem como a probabilidade dos seus resultados quando está retido em um semestre;

2. predizer a retenção final (não conclusão no tempo previsto para integralização do currículo) considerando se o aluno foi aprovado ou reprovado em uma disciplina, cursada em determinado semestre em uma certa tentativa.

Para cada problema foram definidas algumas questões específicas a serem respondidas no intuito de refinar os resultados esperados:

1. Problema 1

   (a) Quais as disciplinas com maiores probabilidades de reter e não reter os alunos em cada semestre?

   (b) Quais as probabilidades de aprovação/reprovação dos alunos retidos?

   (c) Desempenho dos alunos reprovados/aprovados em disciplinas básicas do curso ao cursar disciplinas de semestres posteriores

      i. Qual o resultado dos alunos ao cursar novamente as disciplinas em que foram reprovados?

ii. Qual o resultado do aluno em disciplina de semestre posterior dado a aprovação ou reprovação no pré-requisito direto da disciplina?

iii. Qual o resultado do aluno em disciplina de semestre posterior dado a aprovação ou reprovação no pré-requisito indireto da disciplina?

2. Problema II

   (a) Qual a probabilidade de retenção final diante do resultado em disciplinas?

   (b) Qual a probabilidade de retenção final quando o aluno fica retido em determinado semestre?

Como solução proposta para o primeiro problema, definiu-se a estrutura de uma rede bayesiana manualmente com base na grade curricular para criar os nós e as dependências entre as variáveis de acordo com os pré-requisitos das disciplinas. Para o segundo problema, utiliza-se um classificador bayesiano para predizer a retenção final do aluno a partir de resultados obtidos em componentes curriculares. Para isto, primeiramente foi necessário coletar os dados e pré-processá-los para serem utilizados nas redes definidas.

5.1 CONJUNTO DE DADOS

No conjunto de dados utilizados há três entidades: alunos, disciplinas e cursos. Um aluno é identificado por um id, uma data de nascimento, a cidadania e a nacionalidade. Uma disciplina é identificada por um código (por exemplo MATA02) seguido por um nome (por exemplo, Cálculo A) e a quantidade de horas em um dado semestre. Um curso é identificado por um nome (ou seja, Ciência da Computação - CC), seguido por um código.

Um estudante pode ou não se matricular em uma ou mais disciplinas a cada semestre. O aluno pode também sair do curso por diferentes razões ou ser removido dele. Também é registrada a razão pelo qual o estudante trancou uma disciplina. Durante a inscrição em um curso, o aluno recebe um currículo recomendado. Ele também terá um coeficiente de rendimento e o total de horas de trabalho relativo as disciplinas que obteve aprovação até o semestre mais recente. Para completar o curso, o aluno deve ser aprovado em um conjunto de disciplinas em um tempo definido (pois há um tempo mínimo, tempo certo ou normal e tempo máximo) e completar as atividades complementares. Algumas disciplinas são obrigatórias, outras são optativas, e devem ser escolhidas pelos alunos.

Finalmente, um estudante escolhe em que semestre ele irá cursar as disciplinas do curso. No caso de um semestre regular, o aluno pode ter resultados diferentes nas disciplinas, tais como: ser aprovado em uma disciplina, ser reprovado, reprovado por falta, cursado a disciplina em outra instituição, cursado a disciplina em curso diferente. Se o aluno é aprovado, a carga horária da disciplina é contabilizada na carga horária total do aluno, caso contrário, a carga horária não é contabilizada.

A fim de analisar o conjunto de dados dos estudantes, é importante definir a granularidade que se deve usar para analisar o problema de retenção.

5.1.1 Granularidade

Diante da base de dados existente, foi realizada uma análise das granularidades possíveis e informações extraídas de cada uma delas. Com isso, chegou-se a conclusão que a melhor granularidade é (estudante, semestre, disciplina), ou seja, a disciplina que um estudante está matriculado em um determinado semestre. Especificamente, esta análise reforça a hipótese de que a retenção de estudantes pode ser causada pela combinação de disciplinas em cada semestre.

Foram analisados dois outros pontos de vista de granularidade: (estudante,semestre) e (estudante, disciplina). Na granularidade (estudante, semestre), é possível agregar ao conjunto de dados a informação de quantos alunos estão fazendo determinada disciplina, em determinado semestre e o desempenho de determinado aluno em relação aos seus colegas. Na segunda, uma das potenciais variáveis é a carga horária obtida em cada semestre, o número de disciplinas intensivas cursadas, etc.

No projeto PROUFBA foram disponibilizados os dados dos alunos dos cursos superiores da UFBA no formato de planilha *Excel*. Diante disto, foi realizado um estudo nos dados disponíveis e o desenvolvimento de um modelo relacional de banco de dados para organizar os dados e inseri-los em um banco de dados. Este modelo pode ser visto na Figura 5.1.

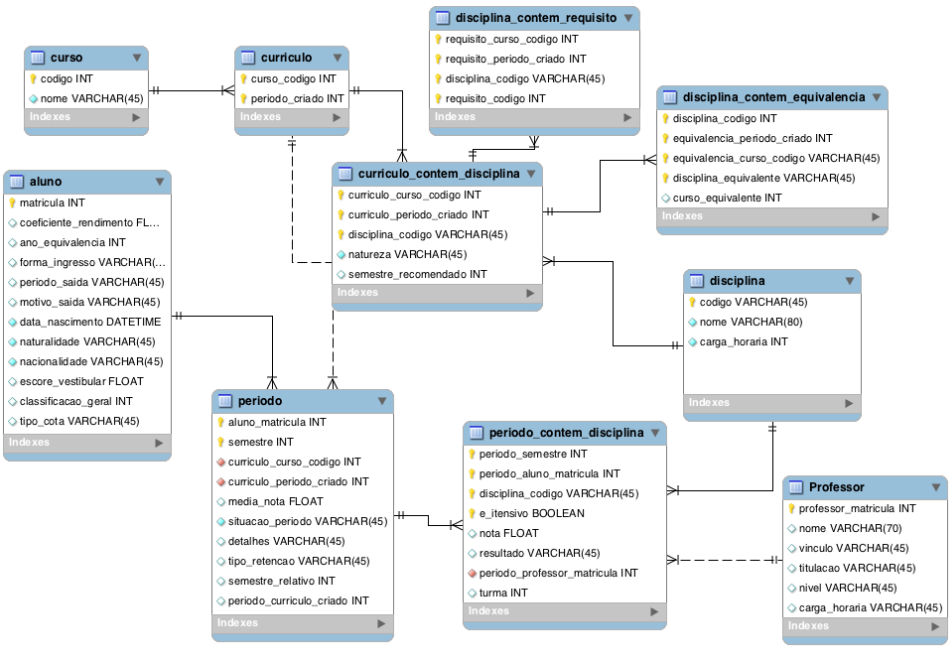

Figura 5.1 Modelo relacional do banco de dados

Após a definição do modelo relacional, foi desenvolvido um *script* para ler os dados na planilha *Excel* e inseri-los no banco de dados criado a partir do modelo desenvolvido.

No PROUFBA, foram realizadas algumas transformações nos dados para aplicação das técnicas de regras de associação: **Semestre Relativo** e **Refinamento da Repetição em disciplina** que serão descritos na seção seguinte.

A partir dos dados já pré-processados do PROUFBA, foi realizada mais uma transformação nestes dados adicionando o **Semestre de Inscrição** do aluno, ou seja, a identificação de quando o aluno se inscreveu na disciplina apenas para o curso de Ciência da Computação, estudo de caso desta pesquisa. Essa transformação também é descrita na seção seguinte com suas respectivas justificativas.

O conjunto de dados utilizados nesta pesquisa inclui todos os alunos do curso Ciência da Computação que foram admitidos entre 2004.1 à 2013.2. Apenas os estudantes pertencentes a uma mesma grade curricular foram considerados, portanto, este conjunto de dados foi limitado a estudantes vinculados a grade curricular de 2007.2 e 2008.1. Isso foi necessário para que os resultados refletidos sejam a respeito das disciplinas que os alunos cursam atualmente (detalhes já foram descritos na Seção 2.1)

5.1.2 Transformação

Como a escolha da granularidade foi a mais fina (estudante, semestre, disciplina), realizou-se três transformações do curso que são descritas nesta seção.

Semestre Relativo. A partir do semestre em que o aluno está inscrito no curso, definiu-se cada novo semestre como primeiro, segundo, terceiro, etc. semestre, o que pode ser entendido como o semestre em relação ao ano de inscrição do aluno (por exemplo, 2009.1, 2009.2, 2010.1, etc.). Com essa transformação pode-se comparar os alunos que se inscreveram em anos diferentes, mas estão enfrentando um fluxo semelhante de disciplinas durante o curso.

Refinamento da Repetição em disciplina. Nessa transformação, refina-se cada disciplina em várias disciplinas com base no número de matrículas realizadas pelo aluno. Por exemplo, os alunos podem estar matriculados em uma disciplina pela sua segunda, terceira ou quarta vez. A partir disso, é possível distinguir as probabilidades associadas da mesma disciplina para diferentes quantidades de inscrição.

Semestre de Inscrição. De acordo com a grade curricular que está inserido, o aluno deveria se matricular em um conjunto de disciplinas por semestre. Entretanto, os alunos não se matriculam em todos as disciplinas recomendadas por vários motivos, tais como: o não cumprimento dos pré-requisitos, choque de horário entre disciplinas, um curto período de tempo para conciliar a disciplina com atividades extra classe, etc. Assim, é importante identificar o semestre que o estudante se matricula em uma disciplina específica.

A partir destas transformações cada variável foi composta pela disciplina, mais o semestre que o aluno se inscreveu e uma certa tentativa, um valor referente ao número de vezes que o aluno está matriculado em determinada disciplina. Por exemplo: *mata01dis1_s1*, são os alunos que se inscreveram em MATA01 (Geometria Analítica) na primeira tentativa no primeiro semestre. E a variável *retencao_s2* apresenta os alunos retidos no segundo semestre, *retencao_s3* no terceiro semestre, etc. Uma amostra do conjunto de dados é apresentado na Figura 5.2.

Na Figura 5.2, a sigla *APR* significa aluno aprovado, *REP* reprovado, *NC* não cursou nas variáveis a respeito de disciplinas, t ou t informa se o aluno ficou retido (t) ou não (f) no segundo, terceiro e quarto semestre e NC da variável de retenção significa que o

mata01dis1_s1	mata02dis1_s1	mata37dis1_s1	mata38dis1_s1	mata39dis1_s1	retencao_s2	retencao_s3	retencao_s4
REP	REP	REP	REP	REP	t	NC	NC
REP	REP	APR	APR	APR	t	t	NC
APR	REP	APR	APR	APR	t	t	f
REP	REP	APR	APR	APR	t	t	t
REP	REP	APR	APR	APR	t	t	t
REP	REP	APR	APR	NC	t	t	t
APR	APR	APR	APR	APR	f	NC	NC
APR	APR	REP	APR	APR	t	t	NC
APR	APR	APR	APR	APR	f	f	f
APR	REP	REP	APR	APR	NC	NC	NC
APR	APR	APR	APR	APR	f	t	t
APR	APR	APR	REP	APR	f	t	NC
REP	APR	APR	APR	APR	t	NC	NC
REP	REP	REP	APR	REP	t	t	t
APR	APR	APR	REP	APR	f	t	NC
APR	REP	APR	APR	APR	t	t	NC
APR	REP	APR	APR	APR	t	t	t
REP	REP	APR	APR	APR	t	NC	NC
REP	REP	REP	REP	REP	t	NC	NC
APR	REP	APR	REP	APR	t	t	t
REP	REP	REP	APR	APR	t	t	t
REP	APR	APR	APR	APR	t	NC	NC
REP	REP	APR	APR	APR	t	t	t
REP	APR	APR	APR	APR	t	t	NC
APR	APR	APR	APR	APR	f	f	f
REP	APR	APR	APR	APR	t	t	NC
REP	REP	APR	APR	APR	t	t	t
REP	REP	APR	APR	APR	t	t	t

Figura 5.2 Parte dos dados após a transformação no pré-processamento

aluno não está mais cursando disciplinas no curso.

Após o pré-processamento através das transformações listadas, o conjunto de dados foi composto por 530 variáveis a respeito de 581 alunos cursando disciplinas em um determinado semestre em uma certa tentativa.

5.2 SOLUÇÃO PROPOSTA PARA O PROBLEMA 1

Diante dos dois problemas destacados nesta pesquisa, o primeiro deles concentra-se na análise probabilística do fluxo acadêmico e os resultados dos alunos ao estarem classificados como retidos ou não retidos. Nesse sentido, buscou-se identificar quais métodos probabilísticos poderiam contribuir com uma solução proposta para este problema.

Diante disto, duas técnicas foram destacadas como possíveis para a solução proposta: cadeias de markov (GOMES; WANKE, 2008) e redes bayesianas (KORB; NICHOLSON, 2010). Primeiramente, as cadeias de markov foram vistas como possível técnica a se utilizar, porém, a modelagem da cadeia se mostrou muito complexa e limitada, visto a necessidade de criar para cada fluxo acadêmico do aluno uma cadeia de markov diferente, bem como a impossibilidade de utilizar três variáveis distintas (APROVADO, REPROVADO e NÃO CURSOU) na modelagem de uma cadeia. De forma contrária, as redes bayesianas são um modelo probabilístico flexível possibilitando a definição da estrutura da rede por um especialista de acordo com o domínio estudado, bem como obter informações probabilísticas das variáveis contidas da rede modelada.

Com as redes bayesianas, foi possível criar uma estrutura que refletisse o fluxo acadêmico das disciplinas e seus pré-requisitos de acordo com a grade curricular específica do curso e a partir desta estrutura obter informações probabilísticas dos resultados obtidos pelos alunos ao cursar componentes curriculares em um dado semestre em uma certa tentativa. Com isso, é possível responder as questões mais específicas destacadas no primeiro problema desta pesquisa. Além disso, outros pontos importantes da utilização de redes bayesianas neste projeto são listadas a seguir:

- Representação quantitativa e qualitativa de um problema específico através de softwares específicos no auxílio da modelagem da rede. Neste trabalho, esta vantagem possibilitou representar a grade curricular do curso através da rede (representação qualitativa) e as probabilidades dos resultados dos alunos ao decorrer do curso (representação quantitativa).

- As redes bayesianas permitem associar diversos tipos de conhecimento numa única ferramenta de projeção. Neste trabalho, foi possível implementar o conceito de pré-requisito de uma grade curricular, bem como o conceito de retenção definido no PROUFBA.

- Permite inferência probabilística do efeito para causa e da causa para o efeito. Com isso, foi possível identificar a probabilidade de aprovação do aluno em uma disciplina recomendada em semestre posterior ao primeiro dado que tenha sido aprovado/reprovado em uma disciplina recomendada primeiro semestre, bem como a probabilidade de aprovação do aluno em uma disciplina recomendada no primeiro semestre dado que não tenha cursado uma disciplina recomendada em um semestre posterior ao primeiro.

Como solução proposta para responder as questões destacadas no primeiro problema, foi definida uma rede bayesiana onde cada variável representa alunos que cursam uma disciplina em um determinado semestre em uma certa tentativa e estas variáveis dependem uma das outras de acordo os seus pré-requisitos. Além disto, as variáveis de retenção de cada semestre definidas de acordo com a heurística do PROUFBA, devem ser dependentes das variáveis que implicam na retenção daquele semestre, ou seja, as variáveis que representam as disciplinas do semestre anterior que são pré-requisitos de disciplinas do semestre da retenção.

Após a definição da rede, os parâmetros de cada variável são calculados (probabilidades condicionais e incondicionais) a partir dos dados dos alunos já pré-processados através do algoritmo de estimação de máxima verossimilhança (MICHALSKI; CARBONELL; MITCHELL, 2013).

Depois da definição da estrutura da rede e a aprendizagem dos parâmetros das variáveis, é possível realizar as inferências probabilísticas através do algoritmo de *clustering* (RENOOIJ; GAAG, 2002) para obter informações a respeito das questões definidas no primeiro problema.

5.3 SOLUÇÃO PROPOSTA PARA O PROBLEMA 2

O segundo problema destacado concentra-se em predizer probabilisticamente a retenção final do aluno a partir dos resultados obtidos pelos alunos em disciplinas cursadas em um determinado semestre em uma certa tentativa e a partir da retenção dos alunos em um determinado semestre.

Como solução proposta deste problema, buscou-se técnicas que resultassem em predições probabilísticas, excluindo a possibilidade de utilizar técnicas de classificação comuns na mineração de dados: árvores de decisão, k-vizinhos mais próximos, etc, que trazem resultados lógicos (*true* ou *false*). Diante disso as técnicas regressão logística (DEGROOT,

1975) e *Naive Bayes* (HAN; KAMBER; PEI, 2011) apresentaram-se viáveis para responder as questões deste problema, visto que as duas técnicas predizem probabilisticamente o resultado de uma variável X, dado os valores de outras variáveis. Como o *Naive Bayes* é um classificador bayesiano, ou seja, utiliza o Teorema de Bayes como base para predizer a variável de retenção, este algoritmo foi utilizado como solução proposta do segundo problema. Além disto, o desempenho e a precisão são uma das características importantes deste algoritmo que o viabiliza para utilização nesta solução proposta.

Assim, como solução proposta para responder as questões destacadas no segundo problema utiliza-se as variáveis já transformadas no pré-processamento mais a variável de retenção final no algoritmo *Naive Bayes*. Com isso, gera-se uma rede bayesiana, onde todas as variáveis são dependentes da variável retenção final (característica do *Naive Bayes*) possibilitando a realização de inferências probabilísticas que respondem as questões do segundo problema.

Inicialmente, para criar a variável retenção final, foi analisada na base de dados a quantidade de semestres cursados pelos alunos para que eles tivessem um motivo de saída (Graduado, Cumpriu Grade Curricular, Esperando colação de grau, etc.) vinculado a sua matrícula, porém foi observado que alguns alunos, mesmo cumprindo os componentes da grade curricular no tempo regular para integralização do curso, não possuíam um motivo de saída vinculado ao mesmo. Diante disso, buscou-se uma alternativa para a retenção final que se aproximasse o mais possível do conceito de conclusão no tempo regular do curso e que também incluísse esses alunos.

Uma alternativa viável foi utilizar a disciplina Projeto Final de Curso II (MATA97) (última disciplina obrigatória do curso) para classificar o aluno como retido ou não. Dessa forma, se o aluno cursa o Projeto Final de Curso II no semestre recomendado (último semestre do tempo regular do curso) e é aprovado, o aluno é classificado como não retido. Caso tenha reprovado ou não cursou, significa que o aluno irá cursar mais um semestre do curso, ultrapassando o tempo regular sendo então classificado como retido.

Depois disto, uma alteração ainda foi necessária no conceito de retenção visto que os alunos utilizados neste experimento são vinculados as duas grades curriculares (2007.2 e 2008.1), onde na grade curricular de 2007.2 a disciplina Projeto Final de Curso II é recomendada no nono período. Já na grade curricular 2008.1 a disciplina é recomendada no décimo período. Diante disso, foi necessário ajustar o conceito de retenção adicionando mais um semestre no tempo regular do curso para poder contemplar os alunos das duas grades curriculares. Assim, o aluno retido é aquele que não consegue obter aprovação na disciplina Projeto Final de Curso II no semestre recomendado ou no semestre posterior ao recomendado.

Dois experimentos foram realizados com o intuito de validar as soluções propostas.

Capítulo 6

.

EXPERIMENTOS

Neste capítulo será apresentado detalhadamente como foram desenvolvidos os experimentos das duas soluções propostas. Para a primeira solução proposta definiu-se uma rede bayesiana baseado na grade curricular do curso, no segundo aplicou-se o algoritmo *Naive Bayes*.

6.1 EXPERIMENTO I

O software Genie (DRUZDZEL, 1999) foi utilizado para auxílio na definição da rede bayesiana, dado a sua fácil utilização através de uma interface gráfica intuitiva e por ser um software com código fonte aberto. Por ser código aberto, o software beneficia toda a sociedade permitindo a cooperação e compartilhamento dos recursos do software entre todos. Além disso, essa foi uma das poucas ferramentas que não impõe limites no número de nós da rede, tão pouco no conjunto de dados utilizados para aprendizado da mesma. Por fim, essa ferramenta facilitou as inferências probabilísticas na rede bayesiana definida através dos algoritmos de inferência já implementados.

Inicialmente a rede bayesiana foi definida utilizando as disciplinas após a transformação dos dados, exceto as disciplinas optativas. As disciplinas optativas foram excluídas pois não fazem parte do fluxo acadêmico obrigatório que os alunos devem percorrer. Desta forma, os resultados a respeito dessas disciplinas não iriam contribuir com o objetivo da pesquisa, porém, sabe-se a importância e a necessidade de cumprimento destas disciplinas para a obtenção do grau.

A rede definida com as variáveis após a transformação dos dados dificultou a realização de inferências visto que o algoritmo de inferência não conseguiu calcular as probabilidades das inferências diante da alta quantidade de variáveis na rede. Diante disto, houve a necessidade de criar critérios de exclusão para selecionar as variáveis mais importantes para o estudo. Com isso, foram adotados os seguintes critérios de exclusão: a) variáveis de disciplinas devem ter pelo menos 20 alunos inscritos, quantidade julgada mínima para que a probabilidade apresentada seja satisfatória e b) uma disciplina deve ser pré-requisito para alguma disciplina recomendada em semestre posterior, característica necessária para

se criar o fluxo das disciplinas. Nesse trabalho, não foi considerado os casos que ocorre quebra de pré-requisito, visto que as quebras de pré-requisitos ocorre apenas em casos excepcionais.

Após a remoção das variáveis consideradas menos importantes para a inferência, a rede Bayesiana foi definida com 105 nós e 253 arcos entre eles.

Cada nó da rede representa estudantes que se inscreveram em uma disciplina recomendada em um determinado semestre e esta inscrição ocorreu na 1ª, 2ª, ..., ou enésima vez. Dessa forma, foi possível identificar probabilidades de casos específicos que ocorrem durante o curso relativo a um dado semestre em uma certa tentativa, diferente de probabilidades a respeito da disciplina isolada, sem o refinamento de quando ocorreu a matricula nem em que vez. Assim, cada variável transformada no pré-processamento, como: mata01dis1_s1, mata40dis1_s2, mata38dis1_s1, etc, é representado por um nó da rede. Em cada nó definiu-se os seguintes estados ou eventos: aprovado (APR), reprovado (REP) ou não se inscreveu na disciplina (NC).

A definição dos arcos entre as variáveis partiu do conhecimento obtido no PROUFBA a respeito da grade curricular do curso e da definição da heurística de retenção. Nesta rede, os arcos retratam o conhecimento qualitativo dos pré-requisitos entre as disciplinas que compõem o fluxo acadêmico que o aluno deve seguir. Além disso, na heurística de retenção definida no PROUFBA, em cada semestre existem disciplinas que retêm o aluno em um semestre específico. Essas disciplinas são os pré-requisitos necessários para cursar todas as disciplinas no semestre em questão. Diante disso, para cada nó que representa os alunos retidos e não retidos em um determinado semestre, existem arcos das disciplinas que retêm naquele semestre para a variável de retenção em questão.

No domínio estudado, cada disciplina pode ou não ter um pré-requisito. Este pré-requisito é a condição necessária para que o aluno possa cursar a disciplina. Dessa forma, caso a disciplina tenha pré-requisito, a matrícula em uma disciplina Y, que tem como pré-requisito X, apenas irá ocorrer dado que o aluno tenha sido aprovado na disciplina X. Assim o resultado na disciplina Y está condicionado a um resultado ocorrido na disciplina X. A Figura 6.1 apresenta um caso real do estudo de caso realizado.

Figura 6.1 mata07dis1_s2 depende da variável mata01dis1_s1

A variável MATA01 (Geometria Analítica) é pré-requisito de MATA07 (Álgebra Linear A). A variável mata01dis1_s1 contém os alunos que cursaram MATA01 na primeira tentativa no primeiro semestre e a variável mata07dis1_s2 contém os alunos que cursaram MATA07 na primeira tentativa no segundo semestre.

Existem casos em que os alunos reprovam em uma disciplina X. Consequentemente o aluno irá se matricular novamente na disciplina X em uma segunda tentativa. Logo, cursar a disciplina X novamente em uma segunda tentativa está condicionada a um evento ocorrido na disciplina X quando cursada na primeira tentativa. A Figura 6.2 apresenta

este caso a respeito do domínio estudado.

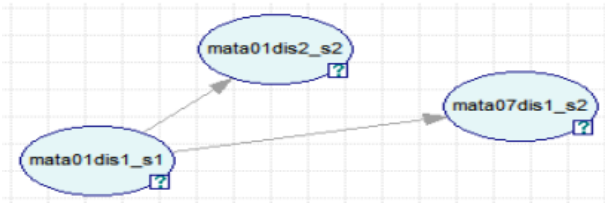

Figura 6.2 A variável mata07dis1_s2 depende de mata01dis1_s1 e mata01dis2_s2 depende de mata01dis1_s1

A variável mata01dis2_s2 contém os alunos que cursam novamente a disciplina MATA01 (Geometria Analítica) na sua segunda tentativa em seu segundo semestre, no qual os resultados em mata01dis2_s2 está condicionado aos eventos (APR, REP, NC) ocorridos em mata01dis1_s1.

Nos dois casos apresentados, alunos escolhem o semestre que irá cursar as disciplinas, por exemplo: o aluno aprovou na disciplina X no primeiro semestre, mas cursou a disciplina Y no terceiro semestre, mas deveria ter cursado no segundo. Logo, a variável dos alunos que cursam Y no terceiro semestre na primeira tentativa também é dependente de X que contêm os alunos que cursam-na no primeiro semestre. O outro caso são dos alunos repetentes. Caso o aluno reprove na disciplina X no primeiro semestre na primeira tentativa e apenas curse a disciplina X no terceiro semestre na segunda tentativa, existe uma dependência da variável X cursada no terceiro semestre na segunda tentativa para a variável dos alunos que cursam X no primeiro semestre na primeira tentativa. Um caso real é apresentado na Figura 6.3

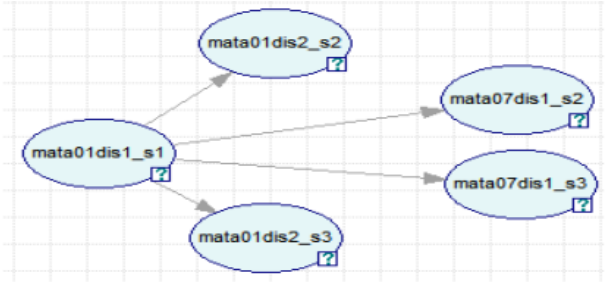

Figura 6.3 Dependência de mata01dis2_s2 e mata01dis2_s3 a variável mata01dis1_s1

A variável mata01dis2_s3 contém os alunos que se inscreveram na disciplina MATA01 (Geometria Analítica) pela segunda vez apenas no terceiro semestre e mata07dis1_s3 contêm os alunos que cursam MATA07 (Álgebra Linear A) pela primeira vez no terceiro semestre.

Quando uma variável Y, que contém os alunos que cursam uma disciplina em um determinado semestre e tem X como pré-requisito, não é possível que Y seja dependente de X nos casos em que X é cursada por alunos em um semestre igual ou superior ao semestre em que Y está sendo cursada. Isso acontece pois não existe possibilidade do

aluno cursar Y em um semestre menor ou igual ao semestre em que ele está cursando X, pois ele precisa ter cursado X primeiro devido ao pré-requisito para poder cursar Y. Os casos com pré-requisitos foram ignorados, visto que no curso de Ciência da Computação essa prática ocorre apenas em casos excepcionais. A Figura 6.4 apresenta um caso real do domínio estudado.

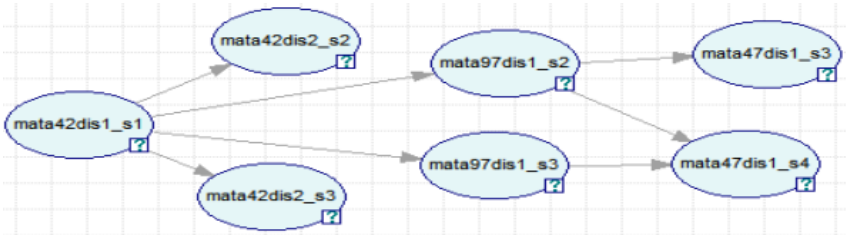

Figura 6.4 Dependências entre as variáveis

A disciplina MATA97 (Matemática Discreta II) é pré-requisito de MATA47. A variável mata97dis1_s2 contém os alunos que cursam a disciplina MATA97 no segundo semestre na primeira tentativa, os que cursam na primeira tentativa no terceiro semestre é representado na variável mata97dis1_s3. A variável mata47dis1_s3 contém os alunos que cursam a disciplina MATA47 (Lógica para Programação) no terceiro semestre na primeira tentativa e é dependente, ou seja, é pré-requisito da variável mata97dis1_s2, a ocorrência de eventos na variável mata47dis1_s3 depende de eventos que ocorrem na variável que é dependente (mata97dis1_s2). Percebe-se que a variável mata47dis1_s3 não é dependente da variável mata97dis1_s3, pois não existe possibilidade do aluno cursar MATA97 no terceiro semestre e ao mesmo tempo cursar MATA47 no terceiro semestre. Diferente dos casos em que os alunos cursam MATA47 no quarto semestre, pois podem ter cursado MATA97 no segundo ou terceiro semestre, logo a variável mata47dis1_s4 é dependente das variáveis mata97dis1_s2 e mata97dis1_s3.

Por fim, cada variável retenção de um determinado semestre é dependente das disciplinas que retêm naquele semestre, pois o aluno só é classificado como retido dado a eventos que ocorram nos pré-requisitos. Logo, para cada disciplina que retém em um determinado semestre, cria-se um arco do nó que caracteriza o pré-requisito para a variável retenção do semestre. Um exemplo é apresentado na Figura 6.5.

No sétimo semestre, os pré-requisitos para que os alunos possam cursar todas as disciplinas do sétimo semestre são: MATA07 (Álgebra Linear A), MATA95 (Complementos de Cálculo) e MATA57 (Laboratório de Programação I). Logo, as variáveis que dependem de *retencao_s7* são todas aquelas que representam os alunos que cursam as disciplinas MATA07, MATA95 e MATA57 antes do sétimo semestre. Se o aluno cursa um destes pré-requisitos no sétimo semestre ele já está classificado como retido, pois chegou no semestre e não cumpriu os pré-requisitos necessários.

Como descrito na Seção 4.2.1, dado que a estrutura de uma rede bayesiana foi definida por um especialista, é necessário estimar a distribuição de probabilidade conjunta da rede, ou seja, as probabilidades de cada variável dado todas as dependências entre as variáveis. Assim, a probabilidade deve ser estimada de acordo com as frequências observadas em um

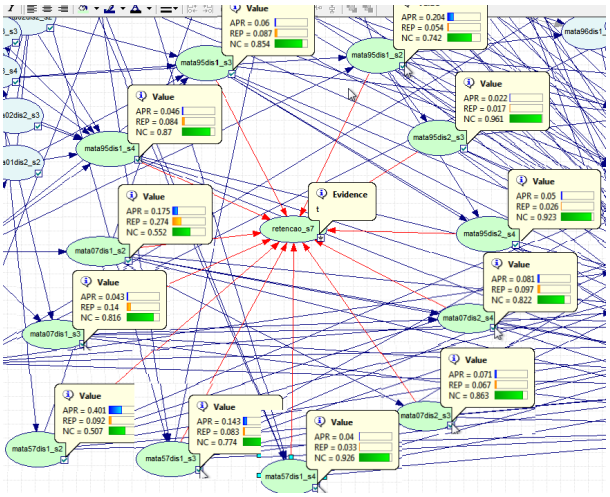

Figura 6.5 Dependências das disciplinas que são pré-requisitos de disciplinas recomendadas no sétimo semestre

conjunto de dados. Neste trabalho, o algoritmo de estimação utilizado foi o da máxima verossimilhança (MICHALSKI; CARBONELL; MITCHELL, 2013) já descrito na Seção 4.2.1.

Finalmente, após a definição da rede e estimação da probabilidade dos eventos de cada variável é possível realizar as inferências probabilísticas para responder as questões definidas através do algoritmo de inferência *clustering* (LAURITZEN; SPIEGELHALTER, 1988) já descrito na Seção 4.3. A Figura 6.6 apresenta maior parte da rede definida para que se tenha uma noção geral de como foi construída. As variáveis utilizadas neste experimento são apresentadas no Anexo C.

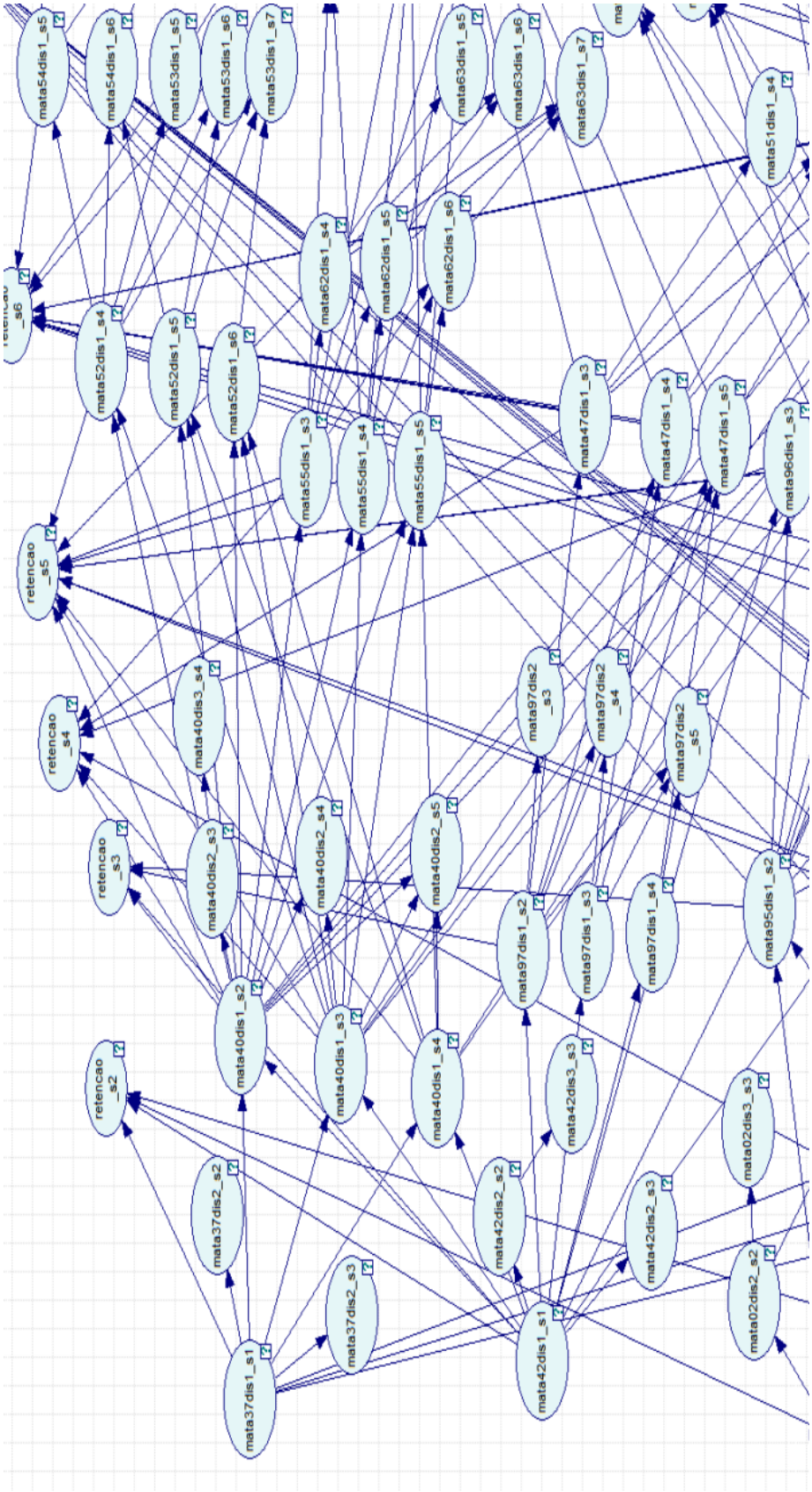

Figura 6.6 Parte da rede bayesiana definida manualmente

6.2 EXPERIMENTO II

O classificador bayesiano *naive bayes* foi utilizado no intuito de predizer probabilisticamente alunos retidos a partir dos resultados obtidos nas disciplinas durante o curso e/ou através da retenção por semestre definida pela heurística de retenção do PROUFBA. Dessa forma, a partir de um conjunto de resultados em disciplinas cursadas em semestres distintos em suas determinadas tentativas é possível informar a probabilidade da retenção final do aluno.

Para este experimento, é necessário um conjunto de variáveis e uma variável classe. Esse conjunto de variáveis é utilizado para evidenciar os seus possíveis eventos e com isso passar as informações necessárias para que se possa predizer probabilisticamente os eventos da variável classe. Neste trabalho, a variável classe é a retenção final (conclusão do curso após o tempo regular) e as demais variáveis são as mesmas utilizadas no experimento anterior, variáveis que representam os alunos que estão cursando uma disciplina em um determinado semestre em uma certa tentativa.

Dado o conceito de retenção final utilizado neste experimento, o número de alunos utilizados foi reduzido comparado ao último experimento, pois apenas é possível analisar os alunos que ingressaram entre 2004.1 e 2009.1, visto que apenas existem informações na base de dados dos alunos até o período 2013.2. Para verificar a probabilidade do aluno cursar a disciplina Projeto Final de Curso II é necessário que exista uma quantidade mínima de semestres entre o período de entrada e o último período da base de dados que viabilize o aluno cursar a disciplina, por exemplo: o aluno que ingressa em 2005.2 tem até 2011.2 para cursar o Projeto Final de Curso II, já os alunos que ingressaram em 2012.1 não podem ser analisados pois não existe a possibilidade de identificar se esse aluno cursou ou não a disciplina Projeto Final de Curso II no semestre recomendado (2016.2) porque não existe essa informação na base de dados.

Dessa forma, para o experimento com o *naive bayes*, foram utilizados os alunos vinculados as grades curriculares 2007.2 ou 2008.1 e que ingressaram entre 2004.1 e 2009.1 somando 230 alunos. Os que ingressaram em 2009.1 devem cursar o Projeto Final de Curso II em 2013.2, último período com informações a respeito dos alunos na base de dados.

Como ferramenta para aplicação do algoritmo *naive bayes*, realização de inferências e validação do classificador utilizou-se o software Genie (DRUZDZEL, 1999). O classificador bayesiano construído através do algoritmo *naive bayes* é ilustrado na Figura 6.7.

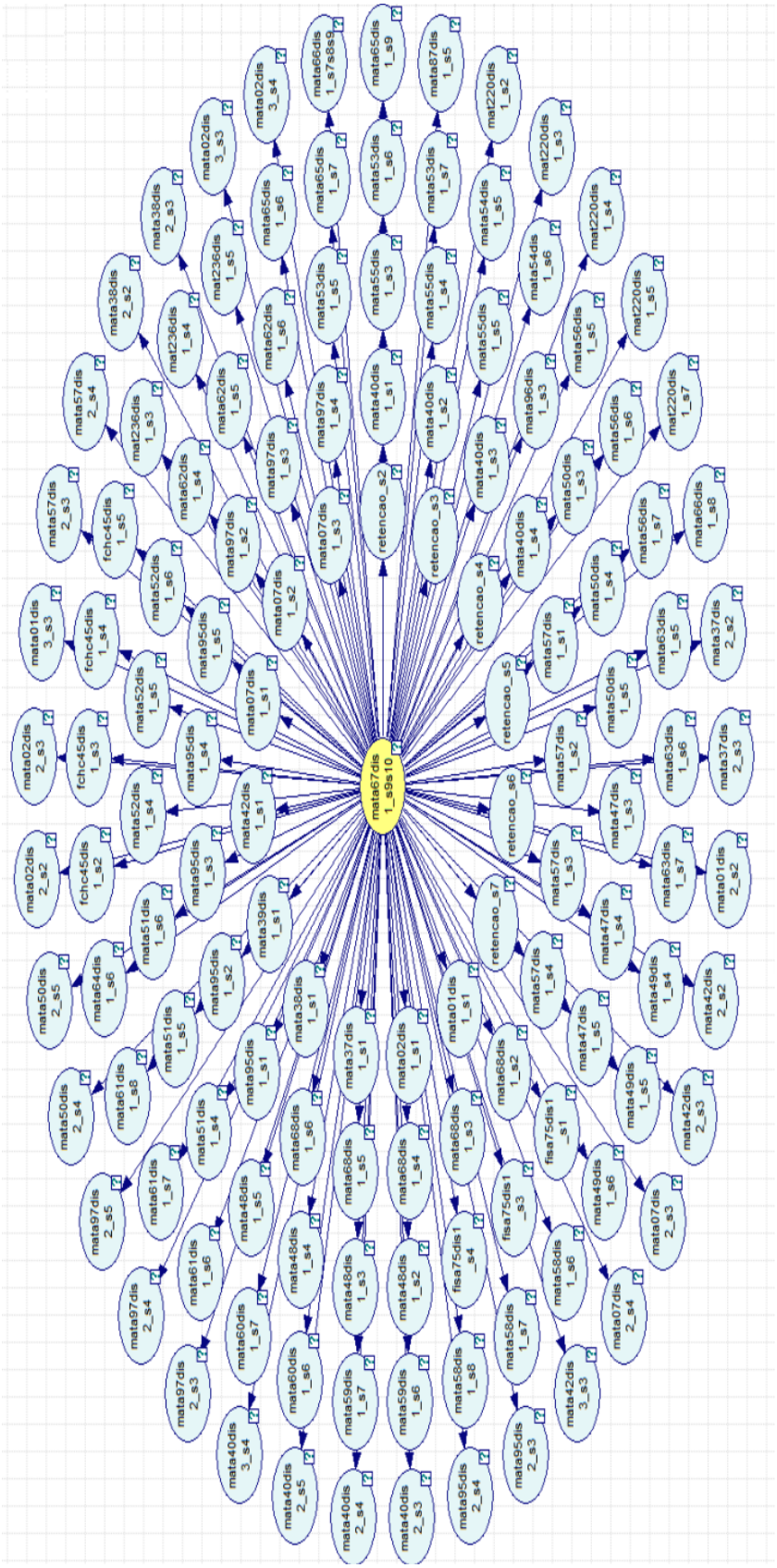

Figura 6.7 Estrutura do classificador bayesiano *Naive Bayes*

6.2.1 Avaliação do Classificador Bayesiano

Quando a estrutura da rede bayesiana é definida a partir de um conjunto de dados para classificar uma variável, é necessário fazer uma avaliação das classificações probabilísticas realizadas pela rede no intuito de avaliar os resultados que estão sendo obtidos com ela. Nesse sentido, a rede bayesiana do experimento II que foi construída através do algoritmo *Naive Bayes* foi avaliada através de algumas medidas de avaliação de classificadores a fim de verificar se a predição realizada por ele pode ser utilizada como fonte de informação para o mundo real.

Uma maneira natural de apresentar as estatísticas para a avaliação de um modelo de classificação é por meio de uma tabulação cruzada entre a classe prevista pelo modelo e a classe real dos exemplos. Essa tabulação é conhecida como tabela de contingência (também chamada de matriz de confusão) (PRATI; BATISTA; MONARD, 2008)

Primeiramente, foi analisada a matriz de confusão obtida com o naive bayes para as classes possíveis: aprovado, reprovado ou não cursou. Como pode ser visto na Figura 6.8, o classificador mostra que dos 27 casos em que o aluno cursou MATA57 (Laboratório de Programação I) e foi aprovado, o classificador prediz corretamente 23 casos, uma probabilidade de acerto de 85,1%. Porém para os casos em que os alunos cursam mas são reprovados, o classificador tem uma probabilidade de acerto muito baixa 11%, inviabilizando a predição de casos em que os alunos são reprovados. Essa baixa probabilidade pode ter sido causado devido ao pequeno número de alunos que são reprovados. Por fim, o classificador prediz corretamente com probabilidade de 88,8% os casos em que os alunos não cursam, visto que dos 170 casos, 151 casos foram previstos corretamente.

	APR	NC	REP
APR	23	2	2
NC	14	151	5
REP	5	3	1

Figura 6.8 Matriz de Confusão do *Naive Bayes*

O grau de acerto de um classificador consiste na probabilidade de predizer corretamente uma classe, parâmetro que se define como *sensibilidade*, quantificado como a razão entre a predição positivo verdadeiro para a classe e o total de casos (verdadeiros positivos e falsos negativos) classificados. No caso estudado a sensibilidade para classificar o aluno que cursou MATA57 (Laboratório de Programação I) e foi aprovado é calculada como a razão dos alunos previstos como aprovados e foram aprovados entre todos os aprovados (23/(23+14+5)) = 54,7%. Associado a este parâmetro, existe outro que serve de contra-prova: a *especificidade*, definida como a probabilidade de predizer negativamente uma classe que, de fato, não é a classe, ou seja, a razão entre os casos previstos como não sendo a classe pelo total de todos os casos previstos para esta classe (positivos e negativos). Para os alunos aprovados em MATA57 (Laboratório de Programação I) a especificidade é dada pelos casos em que os alunos foram aprovados, mas previstos como não aprovados sobre todos os não aprovados (160/160+19) = 89,3%.

A validade de uma predição está na capacidade de um classificador detectar o maior número possível de acertos (resultados positivos verdadeiros) e minimizar os erros (falsos

resultados positivos). Em outras palavras, maximizando a sensibilidade e minimizando as falsas predições positivas. Isso é convenientemente avaliado pela curva ROC (PRATI; BATISTA; MONARD, 2008), registrando-se todos os valores de sensibilidade (a proporção de acertos verdadeiros) no eixo y, contra os valores correspondentes à proporção de falsos acertos (calculados como 1 - especificidade), no eixo x.

Além das avaliações qualitativas dos modelos de classificação obtidas a partir da curva ROC, é possível consolidar a interpretação dos valores da matriz de contingência calculando a área sob a curva (*Area Under Curve* (AUC)) (HANLEY; MCNEIL, 1982) , que é numericamente igual à probabilidade de, dado dois exemplos de classes distintas, o exemplo positivo ser ordenado antes do exemplo negativo. Dessa maneira, as curvas ROC de bons classificadores possuem AUC tendendo a 1. Por outro lado, valores para a AUC próximos de 0.5 são obtidos por classificadores randômicos.

A respeito da análise da área ROC. O ponto (0,0) representa a estratégia de nunca classificar um exemplo como positivo. Modelos que correspondem a esse ponto não apresentam nenhum falso positivo, mas também não conseguem classificar nenhum verdadeiro positivo. A estratégia inversa, de sempre classificar um novo exemplo como positivo, é representada pelo ponto (100,100). O ponto (0,100), representa o modelo perfeito, i.e., todos os exemplos positivos e negativos são corretamente classificados. O ponto (100,0) representa o modelo que sempre faz predições erradas. A linha diagonal indica uma classificação aleatória, ou seja, um sistema que aleatoriamente seleciona saídas como positivas ou negativas, como jogar uma moeda para cima e esperar cara ou coroa (AUC = 50%). Logo, a curva no mínimo deve estar acima da linha diagonal.

Como pode ser visto na Figura 6.9 que apresenta a curva ROC a respeito da predição dos aprovados em MATA57 (Laboratório de Programação I), a curva está mais próxima ao canto inferior esquerdo caracterizando-o como o um modelo "conservativo", onde fazem uma classificação positiva somente se têm grande segurança na classificação e cometem poucos erros falsos positivos. Além disso, percebe-se que a linha da curva até um certo ponto está bem próxima ao ponto (0,100) que representa o modelo perfeito. A AUC tende a 100% classificando-o como um bom modelo de acordo com a curva apresentada.

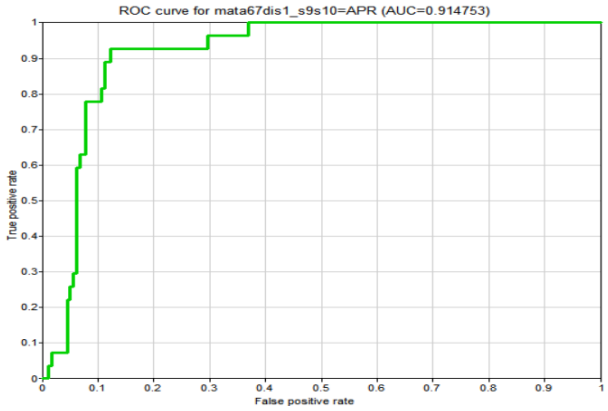

Figura 6.9 Curva ROC predição dos aprovados em MATA57 no semestre recomendado ou posterior

Diferente do bom modelo de predição dos estudantes aprovados em MATA57 (Laboratório de Programação I), a curva ROC sobre a predição de alunos reprovados visto na Figura 6.10 apresenta uma AUC próximo a 50% (predição aleatória) e uma curva muito distante do ponto (0,100). Isso nos mostra que a predição de um aluno reprovado pelo classificador bayesiano dado um conjunto de disciplinas não apresenta uma predição satisfatória. Porém, como neste trabalho analisa-se a retenção do aluno, o problema está em classificá-lo como "não cursou" e ele tenha cursado.

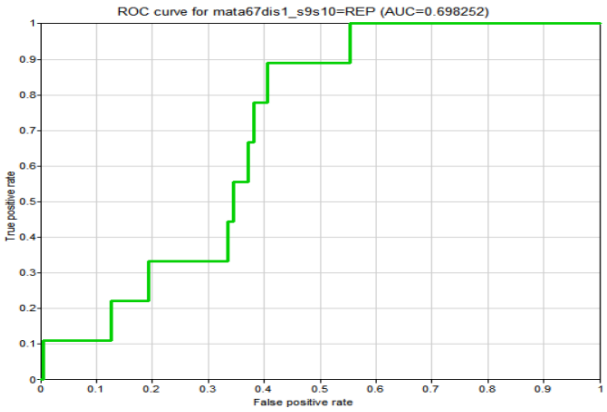

Figura 6.10 Curva ROC predição dos reprovados em MATA57 no semestre recomendado ou posterior

Já para predição dos alunos que não cursam a disciplina MATA57 (Laboratório de Programação I), a curva ROC apresentada na Figura 6.11 mostra uma modelo perfeito até o ponto (0,0.5), depois deste ponto se distancia um pouco, mas sempre perto do eixo (0,1). Esse foi o modelo com melhor predição, visto que mais se aproximou do modelo perfeito e sua AUC foi de 94,1%.

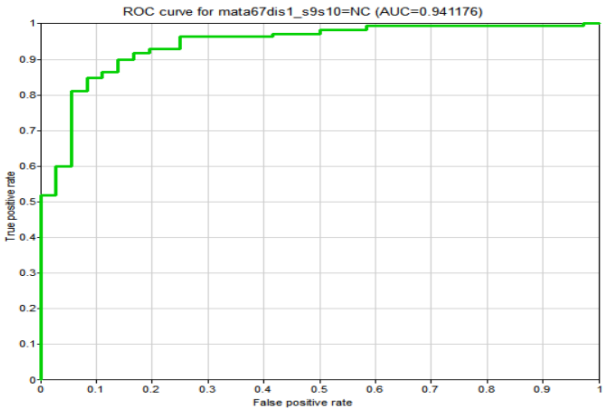

Figura 6.11 Curva ROC predição dos que não cursaram MATA57 no semestre recomendado ou posterior

Uma vez avaliado o experimento, o próximo capítulo apresenta e discute os resultados obtidos.

Capítulo 7

.

RESULTADOS

Os resultados desta pesquisa se concentraram em responder as questões específicas dos dois problemas abordados na solução proposta: i) analisar o fluxo acadêmico do aluno, verificando probabilidades dos resultados em disciplinas quando o aluno cursa em um determinado semestre em uma certa tentativa, bem como seus resultados quando esta retido ou não em um semestre; ii) predizer a retenção final (não conclusão no tempo regular) a partir de resultados obtidos em disciplinas quando o aluno cursa em um determinado semestre em uma certa tentativa. Para cada solução proposta dos problemas citados foram realizados experimentos para obter resultados no intuito de responder questões específicas dos problemas.

Para o experimento I as seguintes questões procuraram ser respondidas:

1. Quais as disciplinas com maiores probabilidades de reter e não reter os alunos em cada semestre?

2. Quais as probabilidades de aprovação/reprovação dos alunos retidos?

3. Desempenho dos alunos reprovados/aprovados em disciplinas básicas do curso ao cursar disciplinas de semestres posteriores

   (a) Qual o resultado dos alunos ao cursar novamente as disciplinas que foram reprovados?

   (b) Qual o resultado do aluno em disciplina de semestre posterior dado a aprovação ou reprovação no pré-requisito direto da disciplina?

   (c) Qual o resultado do aluno em disciplina de semestre posterior dado a aprovação ou reprovação no pré-requisito indireto da disciplina?

As disciplinas básicas do curso são as disciplinas do primeiro e segundo semestre que dão a base inicial para o andamento do aluno no curso. Sobre essas disciplinas, focou-se

na análise a respeito das disciplinas do primeiro semestre visto que são as disciplinas com maiores probabilidades de reprovação e onde obrigatoriamente todos os alunos a cursam.

Com estes resultados, foi possível verificar as probabilidades de aprovação e reprovação do aluno em disciplinas básicas e em disciplinas recomendadas em semestres posteriores de acordo com o resultado obtido em uma ou mais disciplinas básicas. Além disso, foi realizado uma análise da repetência dos alunos em disciplinas básicas, verificando se ao reprovar, o aluno cursa a disciplina novamente em um determinado semestre e qual a probabilidade de aprovação. Também a probabilidade do aluno retido ser aprovado ou reprovado em disciplinas que retêm os alunos. Por fim, foi observado quais as probabilidades de aprovação dos alunos nas disciplinas básicas dado que ele não tenha cursado uma disciplina do quinto semestre.

Além disso, espera-se que resultados a respeito destas questões, possam contribuir com uma análise de como os alunos do curso estão progredindo no fluxo da grade curricular e quais os fatores que fazem com que eles não consigam cursar determinadas disciplinas no semestre em que foi recomendado.

Com a utilização do experimento II espera-se responder as seguintes questões:

1. Qual a probabilidade de retenção final diante do resultado em disciplinas?

2. Qual a probabilidade de retenção final diante de uma retenção em um determinado semestre?

Com os resultados do experimento II, foi possível identificar o impacto na retenção final do aluno diante um resultado obtido em disciplinas do primeiro ao quarto semestre, sendo possível destacar disciplinas que contribuem mais ou menos na conclusão do curso no tempo regular. Além disso foi possível verificar a probabilidade de retenção final diante de uma retenção em um dos semestres do curso.

A partir destes resultados, realizou-se uma análise detalhada dos resultados no intuito de contribuir com informações relevantes sobre a retenção no curso para que os órgãos competentes possam intervir e possivelmente reduzir a retenção no curso.

7.1 RESULTADOS DO EXPERIMENTO I

Nesta seção serão apresentados os resultados obtidos para cada questão que devem ser respondidas pelo primeiro experimento. A primeira questão concentra-se em buscar responder: **Quais as disciplinas com maiores probabilidades de reter e não reter os alunos em cada semestre e quais as probabilidades de aprovação/reprovação dos alunos retidos?**.

A primeira análise concentra-se em verificar a probabilidade dos resultados dos alunos retidos nas disciplinas que podem reter o aluno em um determinado semestre. Assim, a disciplina com maior probabilidade de não reter o aluno é aquele onde maior parte dos alunos retidos conseguiram obter aprovação. De forma contrária a disciplina com maior probabilidade de reter o aluno é aquela onde maior parte dos alunos retidos foram reprovados ou não cursaram as possíveis disciplinas que retêm naquele semestre.

A rede bayesiana definida manualmente foi utilizada para identificar o resultado dos alunos retidos em cada semestre nas disciplinas que causaram a sua retenção. Para realizar esta inferência a variável retencao_sn (n é um determinado semestre) foi evidenciada como *true* para analisar apenas os alunos que foram classificados como retidos no semestre. No intuito de um melhor entendimento, os nós que foram utilizados estão na cor verde e os arcos estão com a cor vermelha nas Figuras apresentadas a seguir.

Para os alunos retidos no segundo semestre a Figura 7.1 apresenta probabilidade do aluno ser aprovado, reprovado ou não ter cursado as disciplinas que levaram ele a ser classificado como retido.

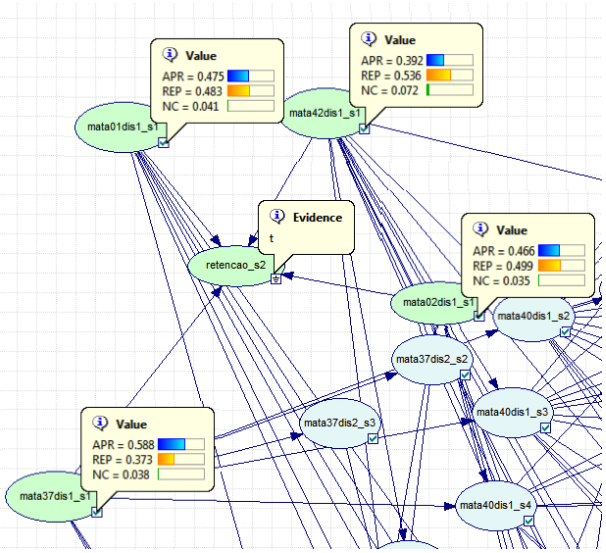

Figura 7.1 Probabilidade de aprovação/reprovação dos alunos retidos no 2° semestre

Cada variável apresenta a probabilidade do aluno ser aprovado, reprovado ou não cursar a disciplina de acordo com evidências na rede. Como pode ser visto na Figura 7.1, dado que o aluno seja retido no segundo semestre (variável retencao_s2 é evidenciada como t = true) a variável mata01dis1_s1 informa que a probabilidade de aprovação é de 47,5%, a de reprovação é de 48,3% e do aluno não cursar é de 4,1%.

Dos alunos retidos a disciplina que mais tem probabilidade de implicar em uma não retenção no segundo semestre é MATA37 (Introdução a Lógica de Programação), no qual a probabilidade de aprovação é 58,8% e 37,3% de reprovação . No entanto a disciplina MATA42 (Matemática Discreta I) é a que tem uma maior probabilidade de reprovação 53,6%, ou seja, a disciplina com maior probabilidade de reter o aluno no segundo semestre.

Para os alunos retidos no terceiro semestre a Figura 7.2 apresenta probabilidade do aluno ser aprovado, reprovado ou não ter cursado as disciplinas que levaram ele a ser classificado como retido.

Dos alunos retidos no terceiro semestre, a disciplina que mais tem probabilidade de reter o aluno é a disciplina MATA95 (Complementos de Cálculo) com probabilidade de 75,7% do aluno não cursar e 4,1% do aluno ser reprovado. A disciplina com maior probabilidade de implicar em não retenção no terceiro semestre é a disciplina MATA42

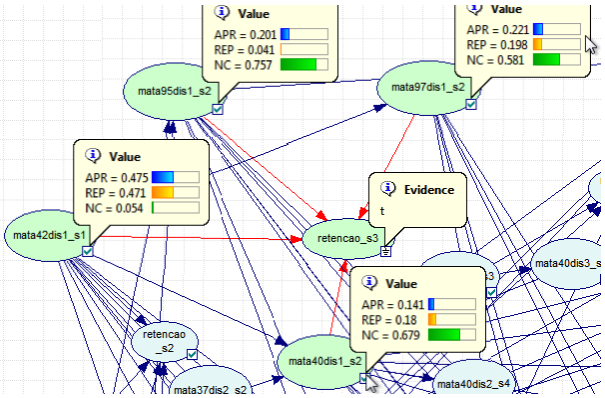

Figura 7.2 Probabilidade de aprovação/reprovação dos alunos retidos no 3° semestre

(Matemática Discreta I), dado que a probabilidade de aprovação é de 47,5%.

Para os alunos retidos no quarto semestre a Figura 7.3 apresenta probabilidade do aluno ser aprovado, reprovado ou não ter cursado as disciplinas que levaram ele a ser classificado como retido.

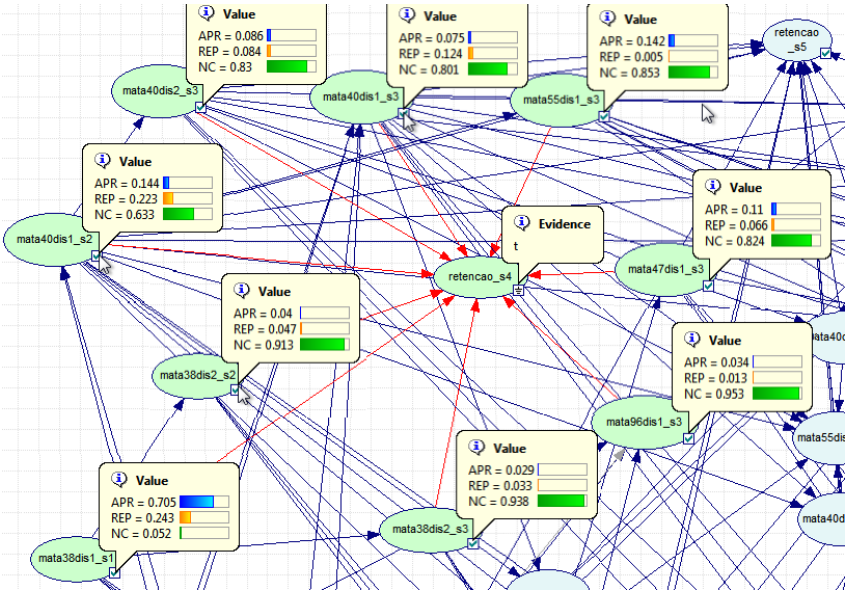

Figura 7.3 Probabilidade de aprovação/reprovação dos alunos retidos no 4° semestre

A disciplina com maior probabilidade de implicar em uma retenção no quarto semestre é a disciplina MATA96 (Estatística A), dado que a probabilidade do aluno não cursar a disciplina MATA96 (Estatística A) recomendada no terceiro semestre é de 95,3% e 1,3% de ser reprovado. A disciplina com maior probabilidade de não reter o aluno no quarto semestre é a disciplina MATA38 (Projeto de Circuitos Lógicos), dado que a probabilidade de aprovação é de 70,5% ao cursa-la no primeiro semestre, de 4% ao cursa-la pela segunda tentativa no segundo semestre e 2,9% ao cursar pela segunda tentativa no terceiro semestre. Assim, a probabilidade da disciplina MATA38 (Projeto de Circuitos Lógicos)

não reter o aluno no quarto semestre é de 77,4% dada pela soma das probabilidades de aprovação quando cursadas em um determinado semestre em uma certa tentativa (77,4% = 70,5% + 4% + 2,9%).

Para os alunos retidos no quinto semestre a Figura 7.4 apresenta probabilidade do aluno ser aprovado, reprovado ou não ter cursado as disciplinas que levaram ele a ser classificado como retido.

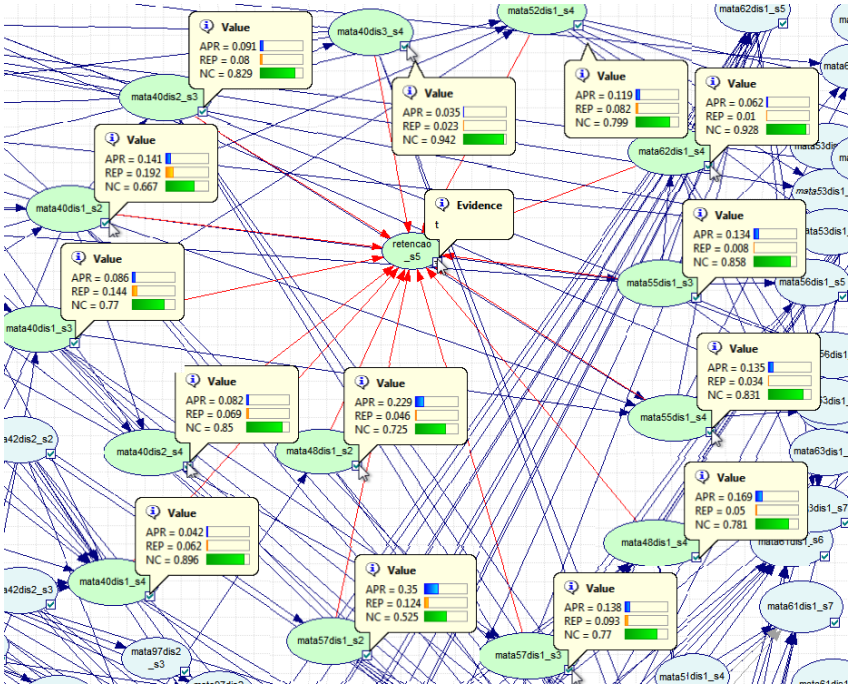

Figura 7.4 Probabilidade de aprovação/reprovação dos alunos retidos no 5° semestre

Como pode ser visto na Figura 7.4 a disciplina MATA62 (Engenharia de Software I) é a que mais implica em uma retenção no quinto semestre, dado que a probabilidade do aluno não cursar a disciplina no quarto semestre é de 92,8% e 1% de ser reprovado. A disciplina MATA40 (Estrutura de Dados e Algoritmos I) é a disciplina com maior probabilidade de não reter no quinto semestre, pois 47,7% dos alunos que foram retidos no quinto semestre, aprovaram na disciplina MATA40. Os 47,7% dos alunos aprovados em MATA40 (Estrutura de Dados e Algoritmos I) estão espalhados entre as variáveis que informam quando o aluno cursou MATA40 em um determinado semestre diante de uma certa tentativa, são elas: mata40dis1_s2, mata40dis1_s3, mata40dis1_s4, mata40dis2_s3, mata40dis2_s4, mata40dis3_s4, no qual respectivamente 14,1%, 8,6%, 4,2%, 9,1%, 8,2% e 3,5% foram aprovados. A soma destas probabilidades é possível, dado que é impossível um aluno ser aprovado em uma disciplina mais de uma vez. Neste caso, é impossível um aluno ser aprovado em MATA40 (Estrutura de Dados e Algoritmos I) no segundo semestre na primeira tentativa (mata40dis1_s2) e também aprovado em MATA40 no terceiro semestre na primeira tentativa (mata40dis1_s3).

Para os alunos retidos no sexto semestre a Figura 7.5 apresenta probabilidade do aluno ser aprovado, reprovado ou não ter cursado as disciplinas que levaram ele a ser

classificado como retido.

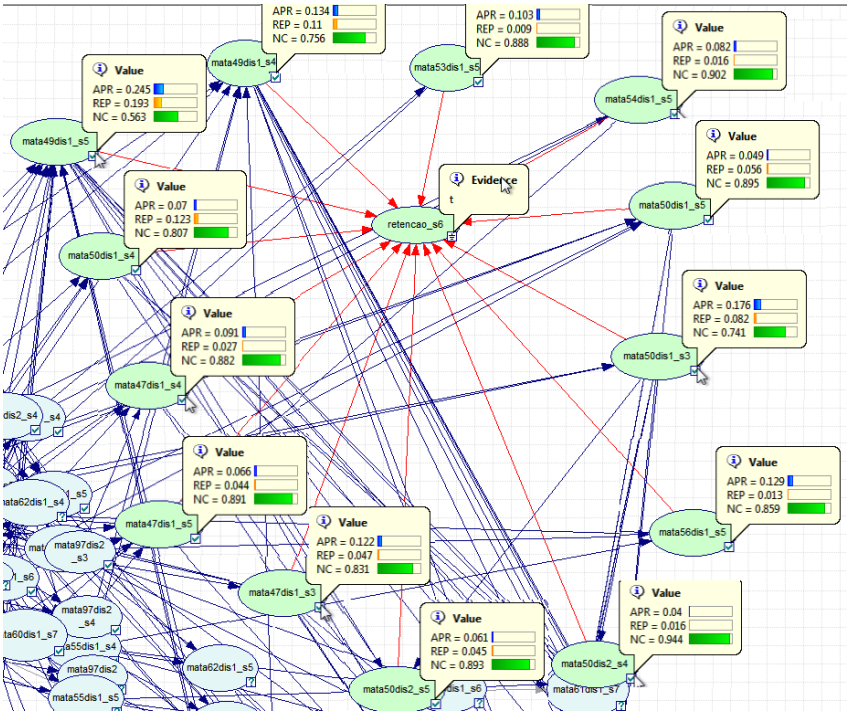

Figura 7.5 Probabilidade de aprovação/reprovação dos alunos retidos no 6° semestre

No sexto semestre, a disciplina que é pré-requisito de uma das disciplinas recomendadas no sexto semestre com maior probabilidade de reter é MATA54 (Estrutura de Dados e Algoritmos II), dado que a probabilidade de aprovação é de 8,2% e a probabilidade do aluno não cursar a disciplina é de 90,2% e 1,6% de reprovação. A disciplina MATA50 (Linguagens Formais e Autômatos), é a que menos tem probabilidade de reter no sexto semestre, dado que dos alunos retidos, 43% conseguem obter aprovação antes do sexto semestre. Os 43% dos alunos aprovados em MATA50 (Linguagens Formais e Autômatos) estão espalhados entre as variáveis que informam quando o aluno cursou MATA50 (Linguagens Formais e Autômatos) em um determinado semestre diante de uma certa tentativa, são elas: mata50dis1_s3, mata50dis1_s4, mata50dis1_s5, mata50dis2_s4, mata40dis2_s5 no qual respectivamente 17,6%, 7%, 4,9%, 4%, 6,1% foram aprovados.

Para os alunos retidos no sétimo semestre a Figura 7.6 apresenta probabilidade do aluno ser aprovado, reprovado ou não ter cursado as disciplinas que levaram ele a ser classificado como retido.

Diante dos resultados apresentados na Figura 7.6 a disciplina MATA07 (Álgebra Linear A) é a que mais tem probabilidade de reter o aluno no sétimo semestre, dado que dos retidos no sétimo semestre, a probabilidade de aprovação em MATA07 (Álgebra Linear A) é de 37% (soma das probabilidades de aprovação das variáveis: mata07dis1_s2, mata07dis1_s3, mata07dis2_s3, mata07dis2_s4). Já a disciplina MATA57 (Laboratório de Programação I) é a que menos tem probabilidade de reter os alunos dado que a probabilidade de aprovação é de 58,4% (soma das probabilidades de aprovação das variáveis:

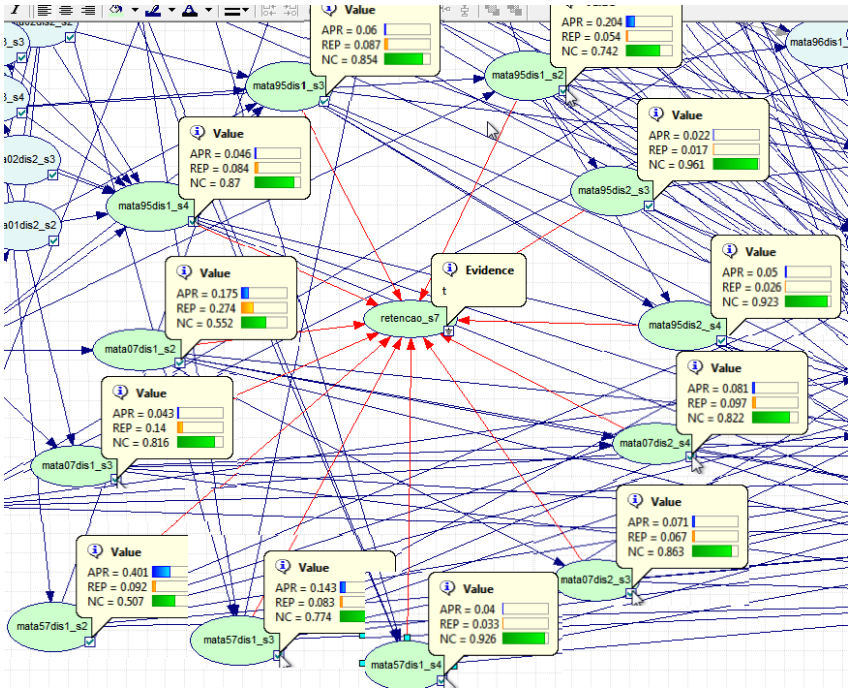

Figura 7.6 Probabilidade de aprovação/reprovação dos alunos retidos no 7° semestre

mata57dis1_s2, mata57dis1_s3, mata57dis1_s4).

A Tabela 7.1 resume as disciplinas e suas respectivas probabilidades que implicam numa maior probabilidade de reter ou não reter o aluno em um determinado semestre. É importante ressaltar que as probabilidades dos resultados dos alunos em outras disciplinas são apresentadas nas Figuras e podem ser avaliadas. Neste trabalho apresentam-se apenas as que mais contribuem.

Após os resultados descritos a respeito da primeira questão, são apresentados os resultados sobre o **Desempenho dos alunos reprovados/aprovados em disciplinas básicas do curso ao cursar disciplinas de semestres posteriores**, que são compostas por três questões específicas. Nesse sentido são descritos os resultados da questão

Tabela 7.1 Disciplinas com maiores probabilidade de reter ou não reter o aluno em um semestre

Semestre	Disciplina com maiores probabilidade de reter	Disciplina com maiores probabilidade de não reter
2	MATA42 - 60,8%	MATA37 - 58,8%
2	MATA02 - 53,4%	MATA01 - 47,5%
3	MATA95 - 79,8%	MATA42 - 47,5%
3	MATA40 - 78,7%	MATA97 - 22,1%
4	MATA96 - 96,6%	MATA38 - 77,4%
5	MATA62 - 93,8%	MATA40 - 47,7%
6	MATA54 - 91,8%	MATA50 - 43%
7	MATA07 - 63%	MATA57 - 58,4%

2(a): **Qual o resultado dos alunos ao cursar novamente as disciplinas que foram reprovados?**

No primeiro semestre do curso as disciplinas recomendadas e são pré-requisitos em semestres posteriores são: MATA42 (Matemática Discreta I), MATA02 (Cálculo A), MATA01 (Geometria Analítica), MATA37 (Introdução a Lógica de Programação) e MATA38 (Projeto de Circuitos Elétricos).

É importante deixar claro que a soma das probabilidades dos eventos de cada variável somam 1, porém os resultados apresentam apenas as probabilidades de dois eventos: aprovado e reprovado, porém a probabilidade do evento não cursou pode ser calculado como 1 - (probabilidade de aprovação + probabilidade de reprovação) (já descrito na Seção 3.2).

As variáveis da rede utilizadas para obter estes resultados são ilustradas na Figura 7.7, onde os nós foram destacados com a cor verde e os arcos com a cor vermelha.

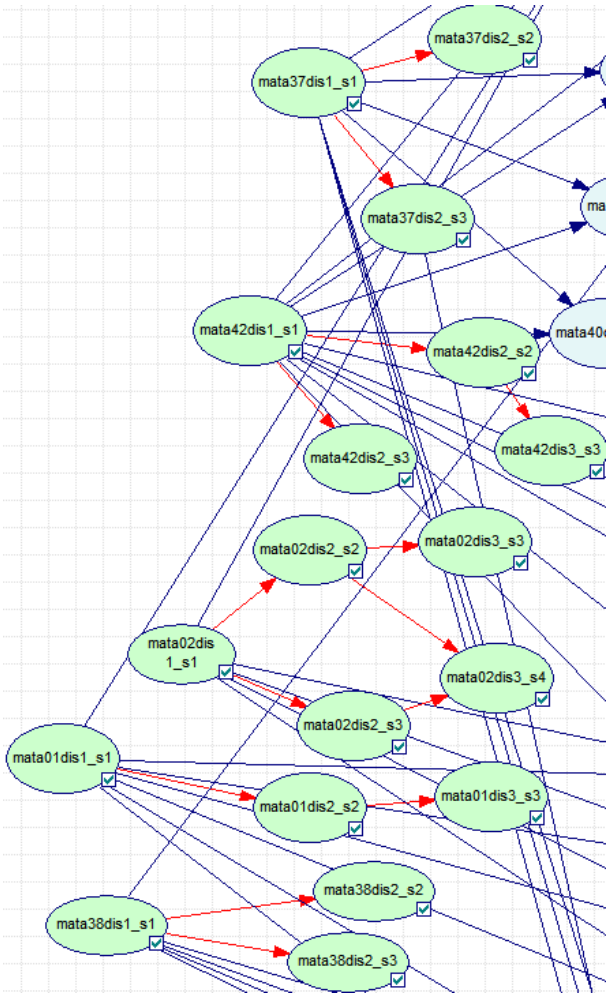

Figura 7.7 Variáveis envolvidas nas inferências

Diante disto as seguintes inferências probabilísticas respondem a seguinte questão: **Qual o resultado dos alunos ao cursar novamente as disciplinas iniciais que**

foram reprovados?.

- Dado a reprovação na disciplina MATA42 (Matemática Discreta I) (48,8%), 62,43% dos alunos cursam novamente a disciplina, no qual 48,8% no segundo semestre e 13,63% no terceiro. Para aqueles que cursaram no segundo semestre, 27,46% foram aprovados e 21,47% reprovados. Para os que cursaram no terceiro, 6,33% foram aprovados e 7,39% reprovados. Dos que foram reprovados novamente quando cursaram no segundo semestre na segunda tentativa, 53% cursou no semestre seguinte, no qual 42% foram aprovados e 11% reprovados. A Figura 7.8 apresenta um gráfico a respeito dessa inferência com probabilidades aproximadas.

- Dado a reprovação na disciplina MATA02 (Cálculo A) (46,5%), 69,24% dos alunos cursam novamente a disciplina, onde 59,99% cursam novamente no segundo semestre e 9,25% no terceiro. Para aqueles que cursaram novamente no segundo semestre, 27,03% foram aprovados e 32,96% reprovados. Para os que cursaram no terceiro, 2,22% foram aprovados e 7,3% reprovados. Dos que foram reprovados na segunda tentativa no segundo semestre, 28,08% foram aprovados na terceira tentativa no terceiro semestre e 29,21% reprovados. Para aqueles reprovados na segunda tentativa no terceiro semestre, 23,8% foram aprovados e 28,6% reprovados na terceira tentativa no quarto semestre.

- Dado a reprovação na disciplina MATA01 (Geometria Analítica) (44,9%), 53,63% dos alunos cursam a disciplina na segunda tentativa no segundo semestre, no qual 26,05% dos alunos são aprovados e 27,58% reprovados. Para aqueles reprovados, 18,05% são aprovados e 30,55% reprovados quando cursam a disciplina na terceira tentativa no terceiro semestre, os restantes desistiram do curso ou cursou em semestre posterior ao terceiro.

- Dado a reprovação na disciplina MATA37 (Introdução a Lógica de Programação) (35,5%), a probabilidade do aluno cursar a disciplina novamente no segundo semestre é de 50,5%, no qual a probabilidade de aprovação é 32,5% e 18% de reprovação. Para os que cursam novamente apenas no terceiro semestre, a probabilidade de aprovação é de 6,3% e 5,8% de reprovação.

- Dado a reprovação na disciplina MATA38 (Projeto de Circuitos Lógicos) (22,2%), a probabilidade de cursar novamente no segundo semestre é de 27,3%, no qual 11,7% é a probabilidade de aprovação e 15,6% de reprovação. Para os que cursam novamente apenas no terceiro semestre a probabilidade de aprovação é de 7% e 10,9% de reprovação.

Na Figura 7.8 as siglas 1° Sem. - 1° Tent. identifica os alunos que cursam a disciplina MATA42 (Matemática Discreta I) no primeiro semestre, na primeira tentativa. As outras siglas como 2° Sem. - 2° Tent., 3° Sem. - 2° Tent., seguem a mesma linha de raciocínio, alunos cursando MATA42 (Matemática Discreta I) no segundo semestre na segunda tentativa e alunos cursando MATA42 (Matemática Discreta I) no terceiro semestre na segunda tentativa. As setas indicam qual caminho possível do aluno. Ao ser

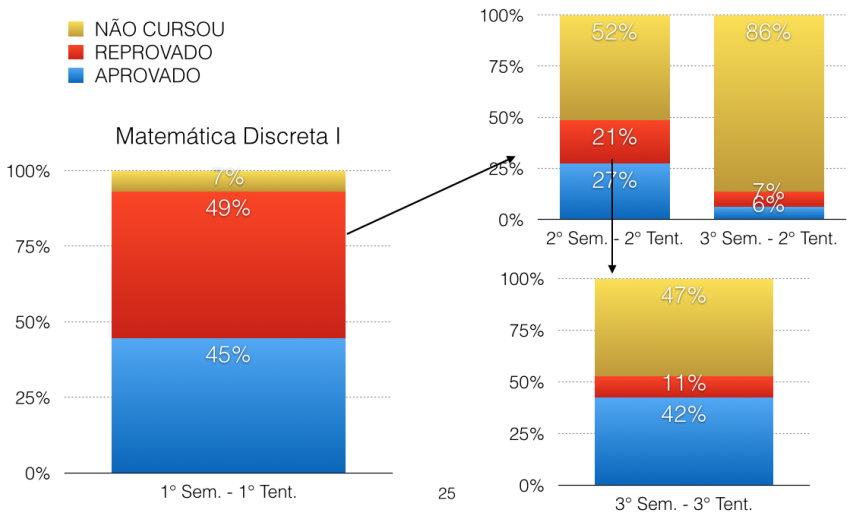

Figura 7.8 Resultado do aluno ao ser reprovado em MATA42

reprovado na primeira tentativa no primeiro semestre, o aluno pode cursar MATA42 no segundo ou terceiro semestre na sua segunda tentativa. Para os reprovados na segunda tentativa no segundo semestre, esses podem cursar MATA42 no terceiro semestre em sua terceira tentativa.

Como já informado, a probabilidade de não cursar não é apresentada nas inferências, porém pode ser caculada como 1 - (probabilidade de aprovação + probabilidade de reprovação). No exemplo apresentado na Figura 7.8, a probabilidade do aluno não cursar MATA42 (Matemática Discreta I) no segundo semestre na sua segunda tentativa é de $1 - (0.21 + 0.27) = 0.52$, ou seja, dos alunos que reprovaram em MATA42 (Matemática Discreta I) ao cursar pela primeira vez no seu primeiro semestre, 52% não cursaram a disciplina novamente no segundo semestre.

Ainda sobre o comportamento dos alunos reprovados/aprovados em disciplinas básicas do curso, os resultados a questão 2(b): **Qual o resultado do aluno em disciplina de semestre posterior dado a aprovação ou reprovação no pré-requisito direto da disciplina?**, são apresentados.

Nos casos em que os alunos são aprovados nas disciplinas do primeiro semestre, é importante observar se os alunos cursam a disciplina no semestre posterior e se são aprovados ou reprovados. Além disso, deseja-se comparar os alunos aprovados e reprovados nas disciplinas básicas, identificando se ao cursar a disciplina do semestre posterior em um determinado semestre o aluno que foi aprovado na primeira tentativa tem probabilidade maior de ser aprovado na disciplina posterior do que o aluno que foi reprovado na primeira tentativa. Com isso, é importante analisar as implicações que a reprovação e aprovação nestas disciplinas implica no fluxo acadêmico do aluno.

Para obter estes resultados, para cada disciplina do primeiro semestre foi evidenciado o seu estado como aprovado e reprovado, ou seja, seleciona-se apenas os alunos aprovados em um caso e os reprovados em outro caso. Para cada evidência foi realizada uma inferência probabilística a respeito dos resultados das variáveis que são dependentes das

variáveis evidenciadas, ou seja, as disciplinas do segundo semestre que tem disciplinas do primeiro semestre como pré-requisito.

A disciplina MATA97 (Matemática Discreta II), recomendada no segundo semestre, tem MATA42 (Matemática Discreta I) como seu pré-requisito direto.

A respeito destas disciplinas foi realizada as seguintes inferências:

- Dado a **reprovação** na disciplina MATA42 (Matemática Discreta I) (48,88%) na primeira tentativa, 22,53% cursam MATA97 (Matemática Discreta II) no terceiro semestre, no qual 8,45% são aprovados e 14,08% reprovados. Para aqueles que cursam apenas no quarto semestre, 4,92% são aprovados e 9,85% reprovados.

- Dado a **aprovação** na disciplina MATA42 (Matemática Discreta I) (44,57%) na primeira tentativa, 82,8% cursam MATA97 (Matemática Discreta II) no segundo semestre, no qual 45,1% são aprovados e 37,1% reprovados. Para aqueles que cursam apenas no terceiro semestre, 1,7% são reprovados.

A figura 7.9 apresenta essas inferências de forma gráfica.

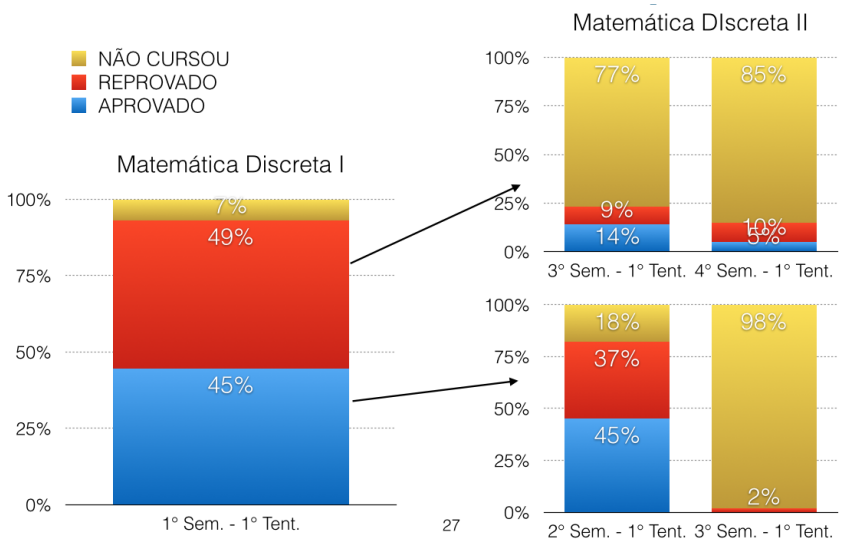

Figura 7.9 Resultado do aluno ao ser reprovado em MATA42

Sobre a disciplina MATA95 (Complementos de Cálculo), recomendada no segundo semestre, tem MATA01 (Geometria Analítica) e MATA02 (Cálculo A) como seus pré-requisitos direto. A respeito destas disciplinas, obtêm-se as seguintes inferências:

- Dado a **reprovação** em MATA02 (Cálculo A) e MATA01 (Geometria Analítica) na primeira tentativa (20,87%), 6,73% cursam no terceiro semestre, 6,72% no quarto e 4,65% no quinto. Dos que cursam no terceiro semestre, 4,66% são aprovados e 2,07% reprovados, já os que cursam no quarto, 3,62% são aprovados e 3,1% reprovados, por fim, para os que cursam no quinto semestre 1,03% são aprovados e 3,62% reprovados.

- Dado a **aprovação** em MATA01 (Geometria Analítica) e MATA02 (Cálculo A) na primeira tentativa (25,41%), 64,1% cursam MATA95 (Complementos de Cálculo) no segundo semestre, no qual 64,09% são aprovados e 14,54% reprovados, já os que cursam no terceiro semestre, 1,81% são aprovados ou reprovados.

Sobre a disciplina MATA07 (Álgebra Linear), recomendada no segundo semestre, tem MATA01 (Geometria Analítica) como pré-requisito direto. A respeito destas disciplinas, obtêm-se a seguinte inferência:

- Dado a **reprovação** na disciplina MATA01 (Geometria Analítica) na primeira tentativa (44,92%), 18,38% cursam MATA07 (Álgebra Linear A) no terceiro semestre, no qual 3,44% são aprovados e 14,94% são reprovados.

- Dado a **aprovação** na disciplina MATA01 (Geometria Analítica) na primeira tentativa (51,92%), 78,2% cursam MATA07 (Álgebra Linear A) no segundo semestre, no qual 41,3% são aprovados e 36,9% são reprovados. A probabilidade de cursar MATA07 (Álgebra Linear A) no terceiro semestre é de apenas 4%, no qual 2% são aprovados e 2% reprovados.

Sobre a disciplina MATA40 (Estrutura de Dados e Algoritmos I), recomendada no segundo semestre, tem MATA37 (Introdução à Lógica de Programação) e MATA42 (Matemática Discreta I) como seus pré-requisitos direto. A respeito destas disciplinas, obtêm-se as seguintes inferências:

- Dado a **reprovação** na disciplina MATA37 (Introdução a Lógica de Programação) e MATA42 (Matemática Discreta I), a probabilidade do aluno cursar MATA40 (Estrutura de Dados e Algoritmos I) no terceiro semestre é de 11,8%, no qual a probabilidade de aprovação é de 2,5% e 9,3% de reprovação. Já para o quarto semestre, a probabilidade de aprovação é de 3,7% e 6,8% de reprovação.

- Dado a **aprovação** nas disciplinas MATA42 (Matemática Discreta I) e MATA37 (Introdução a Lógica de Programação) (26,84%) na primeira tentativa, 80,34% cursam MATA40 (Estrutura de Dados e Algoritmos I) no segundo semestre, 4,91% cursa no terceiro e 5,4% no quarto, os restantes cursam em semestres posteriores ou desistiram do curso. A probabilidade do aluno ser aprovado ou reprovado quando cursa no segundo é 40,17%, para os que cursam no terceiro, a probabilidade de aprovação é de 3,12% e 1,78% de reprovação, por fim, aqueles que cursam no quarto tem uma probabilidade de aprovação de 2,13% e 3,27% de reprovação.

Sobre a disciplina MATA57 (Laboratório de Programação I), recomendada no segundo semestre, tem MATA37 (Introdução à Lógica de Programação) como seu pré-requisito direto. A respeito destas disciplinas, obtêm-se as seguintes inferências:

- Dado a reprovação na disciplina MATA37 (Introdução a Lógica de Programação) na primeira tentativa no primeiro semestre (35,5%), a probabilidade do aluno cursar MATA57 (Laboratório de Programação I) no terceiro semestre é de 26,9%, no qual

a probabilidade de aprovação é de 14.7% e 12,2% de reprovação. Já para os que cursam a disciplina MATA57 (Laboratório de Programação I) no quarto semestre, a probabilidade de aprovação é de 4,5% e 6% de reprovação.

- Dado a aprovação na disciplina MATA37 (Introdução a Lógica de Programação) na primeira tentativa no primeiro semestre (60,2%), a probabilidade do aluno cursar MATA57 (Laboratório de Programação I) no segundo semestre é de 76%, no qual a probabilidade de aprovação é de 58.2% e 17,8% de reprovação. Já para os que cursam a disciplina MATA57 (Laboratório de Programação I) no terceiro semestre, a probabilidade de aprovação é de 7.2% e 2.4% de reprovação.

Finalmente, sobre o comportamento sobre o comportamento dos alunos reprovados/aprovados em disciplinas básicas do curso, os resultados da terceira e última questão: **Qual o resultado do aluno em disciplina de semestre posterior dado a aprovação ou reprovação no pré-requisito indireto da disciplina?**, são apresentados.

Para isso, procurou-se observar qual o impacto da reprovação/aprovação em uma disciplina básica que é pré-requisito indireto a uma disciplina de semestre posterior, ou seja, qual a probabilidade do aluno que foi reprovado ou aprovado em uma disciplina do primeiro ou segundo semestre cursar uma disciplina de dois, três, etc. semestres recomendados posteriores a ela?

Uma disciplina X é pré-requisito direto de uma disciplina Y, dado que em um semestre posterior recomendado para cursar X, Y só possa ser cursado se houve aprovação na disciplina X, logo X $\rightarrow$ Y. Uma disciplina Z tem X como pré-requisito indireto dado que Z tem como pré-requisito Y, como Y depende de X (X $\rightarrow$ Y) e Z depende de Y (Y $\rightarrow$ Z), logo Z depende de X, X $\rightarrow$ Z.

Para obter estes resultados, foi selecionada as disciplinas do quinto semestre como variável de observação e as disciplinas do primeiro semestre como evidências no intuito identificar as probabilidades do aluno cursar a disciplina do quinto semestre em um dado semestre, bem como o seu resultado dado que tenha sido aprovado ou reprovado nas disciplinas do primeiro semestre que são pré-requisitos indiretos a ela. Desta forma é possível analisar todos os fluxos possíveis dos alunos até cursar as disciplinas do quinto semestre. Foi selecionado as disciplinas do quinto semestre visto que este é o último semestre que irá conter disciplinas que serão pré-requisitos em semestre posterior, dado que nenhuma disciplina do sexto semestre é pré-requisito para uma disciplina de semestre posterior. A Figura 7.10 apresenta os possíveis fluxos de disciplinas que o aluno obrigatoriamente deve seguir.

A disciplina MATA37 (Introdução a Lógica de Programação) e MATA42 (Matemática Discreta I), recomendada no primeiro semestre, é pré-requisito de MATA40 (Estruturas de Dados e Algoritmos I), recomendada no segundo semestre. MATA40 é pré-requisito de MATA52 (Análise e Projeto de Algoritmos), recomendada no quarto semestre, MATA52 é pré-requisito de MATA54 (Estruturas de Dados e Algoritmos II) recomendada no quinto semestre.

De acordo com a rede bayesiana a probabilidade do aluno cursar MATA54 (Estrutura de Dados e Algoritmos II) no quinto semestre é de apenas 9,8%, no qual a probabilidade

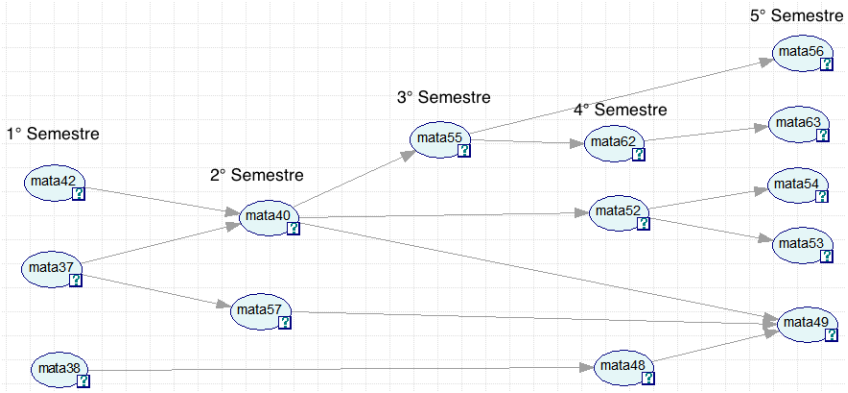

Figura 7.10 Fluxo Acadêmico do 1° ao 5° semestre

de aprovação é de 9% e 0,08% de reprovação. Já para o sexto semestre a probabilidade é de 4,7% de aprovação e 2% de reprovação.

A respeito dos alunos que aprovam em MATA37 (Introdução a Lógica de Programação) e MATA40 (Estrutura de Dados e Algoritmos I) as probabilidades de cursar MATA54 (Estrutura de Dados e Algoritmos II) bem como os seus resultados são ilustrados na Figura 7.11.

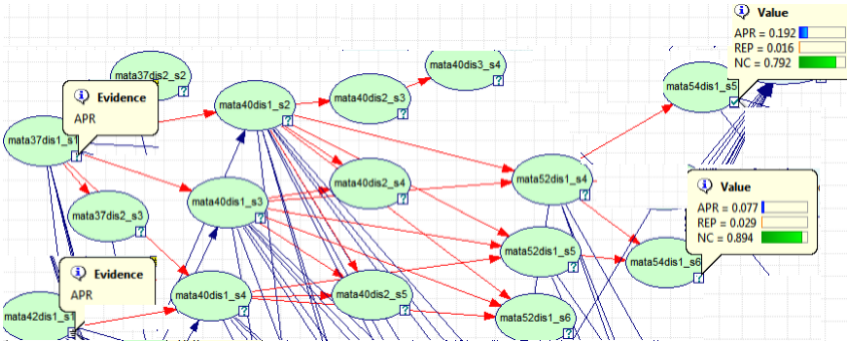

Figura 7.11 Probabilidade de cursar MATA54 dado a **aprovação** em MATA37 e MATA42

A Figura 7.11 apresenta o caminho possível, baseado nos requisitos indiretos da disciplina MATA54 (Estrutura de Dados e Algoritmos II) para que o aluno possa cursar a disciplina MATA54 evidenciando apenas os alunos aprovados nas disciplinas MATA37 (Introdução a Lógica de Programação) e MATA42 (Matemática Discreta I).

Como pode ser visto nas inferências apresentadas na Figura 7.11, dado a aprovação em MATA37 (Introdução a Lógica de Programação) e MATA42 (Matemática Discreta I) a probabilidade do aluno cursar MATA54 (Estrutura de Dados e Algoritmos II) no semestre recomendado é de 20,8%, no qual a probabilidade de aprovação é de 19,2% e 1,6% de reprovação. A probabilidade de cursar no sexto semestre é menor para os alunos aprovados, onde a probabilidade de aprovação é de 7,7% e 2,9% de reprovação.

É importante observar que dada as evidências para os aprovados em MATA37 (Introdução a Lógica de Programação) e MATA42 (Matemática Discreta I), as informações

são a respeito apenas dos aprovados nestas duas disciplinas. Segunda a inferência realizada, sabe-se que 20,8% dos alunos cursam no quinto semestre e 10,6% no sexto, logo o restante dos alunos ou cursaram em semestre posterior ao sexto (não existe a variável mata54dis1_s7 pois não contêm o mínimo de 20 alunos cursando, dada as restrições impostas na rede, ver Seção 6.1) ou desistiram do curso.

A respeito dos alunos que reprovam em MATA37 (Introdução a Lógica de Programação) e MATA40 (Estrutura de Dados e Algoritmos I) as probabilidades de cursar MATA54 (Estrutura de Dados e Algoritmos II) bem como os seus resultados são ilustrados na Figura 7.12.

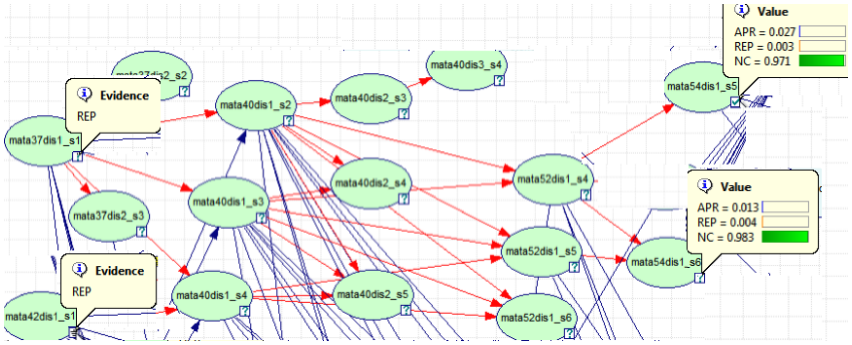

Figura 7.12 Probabilidade de cursar MATA54 dado a **reprovação** em MATA37 e MATA42

Para os que reprovam nas disciplinas MATA37 (Introdução a Lógica de Programação) e MATA40 (Estrutura de Dados e Algoritmos I) no primeiro semestre, a probabilidade de cursar no quinto semestre e ser aprovado é de 2,7% e 0,03 de ser reprovado. Logo, percebe-se que mesmo tendo a possibilidade de se recuperar no semestre posterior e conseguir cursar a disciplina no semestre recomendado de acordo com os pré-requisitos deste fluxo de disciplinas, os alunos não conseguem ou não fazem isso. No sexto semestre a probabilidade do aluno cursar e ser aprovado é de 1,3% e 0,04% de ser reprovado. O restante dos reprovados, ou cursaram em semestres posteriores, ou evadiram do curso.

Sobre a disciplina MATA49 (Programação de Software Básico), recomendada no quinto semestre, para o aluno cursar esta disciplina ele deve ter cumprir os seguintes pré-requisitos: MATA40 (Estrutura de Dados e Algoritmos I), MATA57 (Laboratório de Programação I) recomendadas no segundo semestre e MATA48 (Arquitetura de Computadores) recomendada no quarto semestre. A disciplina MATA40 tem como pré-requsito as disciplinas MATA37 (Introdução a Lógica de Programação) e MATA42 (Matemática Discreta I).

Diante deste números de pré-requisitos, para cursar a disciplina MATA49 (Programação de Software Básico) o aluno deve cumprir os seguintes pré-requisitos indiretos: MATA37 (Introdução a Lógica de Programação), MATA38 (Projeto de Circuitos Lógicos) e MATA42 (Matemática Discreta I) recomendadas no primeiro semestre, MATA40 (Estrutura de Dados e Algoritmos I) e MATA57 (Laboratório de Programação I), recomendadas no segundo semestre e seu pré-requisito direto MATA48 recomendada no quarto semestre.

A probabilidade do aluno cursar MATA49 (Programação de Software Básico) no quarto semestre é de 14,1%, no qual a probabilidade de aprovação é de 8% e 6,1% de

reprovação. Já no quinto semestre a probabilidade de aprovação é de 23% e 22,7% de reprovação. Por fim, há uma probabilidade de 45,6% de cursar a disciplina MATA49 (Programação de Software Básico) no sexto semestre, no qual a probabilidade de aprovação é de 23,1% e 22,5% de reprovação.

Dado a evidência de ter sido aprovado ou reprovado nas disciplinas básicas do primeiro semestre que são pré-requisitos indiretos de MATA49 (Programação de Software Básico), destaca-se as implicações que uma aprovação ou reprovação nestas disciplinas implicam em cursar a disciplina MATA49. Os resultados a respeito dos alunos aprovados são apresentados na Figura 7.13.

A respeito dos alunos que aprovam em MATA37 (Introdução a Lógica de Programação), MATA42 (Matemática Discreta I) e MATA38 (Projeto de Circuitos Lógicos) as probabilidades de cursar MATA54 (Estrutura de Dados e Algoritmos II) bem como os seus resultados são ilustrados na Figura 7.13.

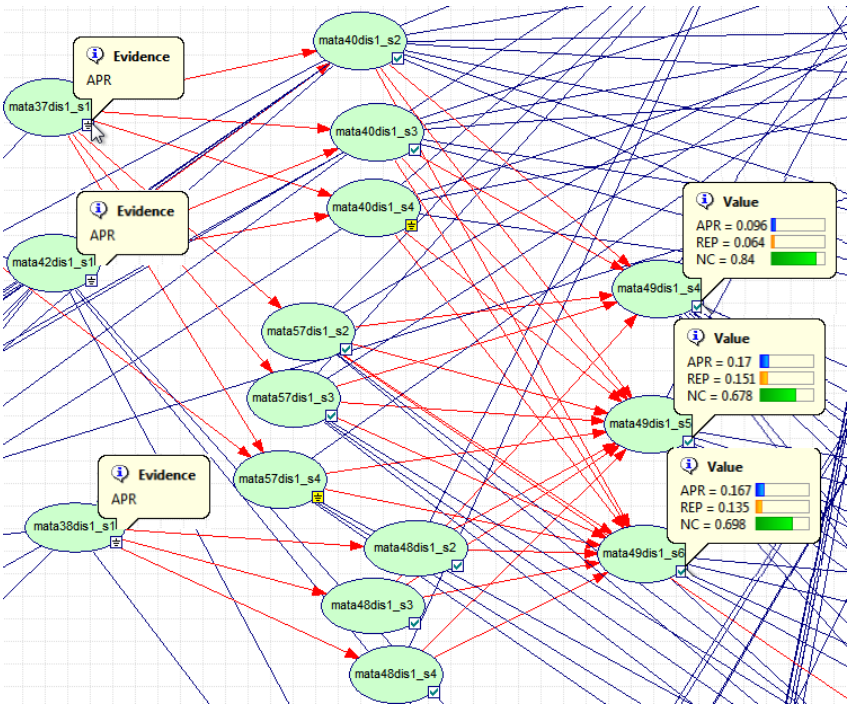

Figura 7.13 Probabilidade de cursar MATA49 dado a **aprovação** em MATA37, MATA42 e MATA38

Dado a aprovação em MATA37 (Introdução a Lógica de Programação), MATA42 (Matemática Discreta I) e MATA38 (Projeto de Circuitos Lógicos), a probabilidade de cursar MATA49 (Programação de Software Básico) no quarto semestre é de 16%, no qual a probabilidade de aprovação é de 9,6% e 6,4% de reprovação. Para o quinto semestre, a probabilidade do aluno cursar é de 32,1%, no qual a aprovação é de 17% e 15,1% de reprovação. Por fim, os que cursam no sexto semestre 16,7% aprovam e 13,5% reprovam. O restante dos alunos cursaram em semestres posteriores ou evadiram do curso.

Como pode ser visto nos resultados das inferências, alguns alunos cursam a disciplina MATA49 (Programação de Software Básico) em um semestre anterior ao recomendado,

isto é possível dado que no terceiro semestre já é possível ter cumprindo todos os pré-requisitos de MATA49 (Programação de Software Básico). Além disso, ao cursar em qualquer um dos semestre a probabilidade de aprovação sempre é maior do que a de reprovação, porém não existe uma grande diferença.

A respeito dos alunos que reprovam em MATA37 (Introdução a Lógica de Programação), MATA42 (Matemática Discreta I) e MATA38 (Projeto de Circuitos Lógicos) as probabilidades de cursar MATA49 (Programação de Software Básico) bem como os seus resultados são ilustrados na Figura 7.14.

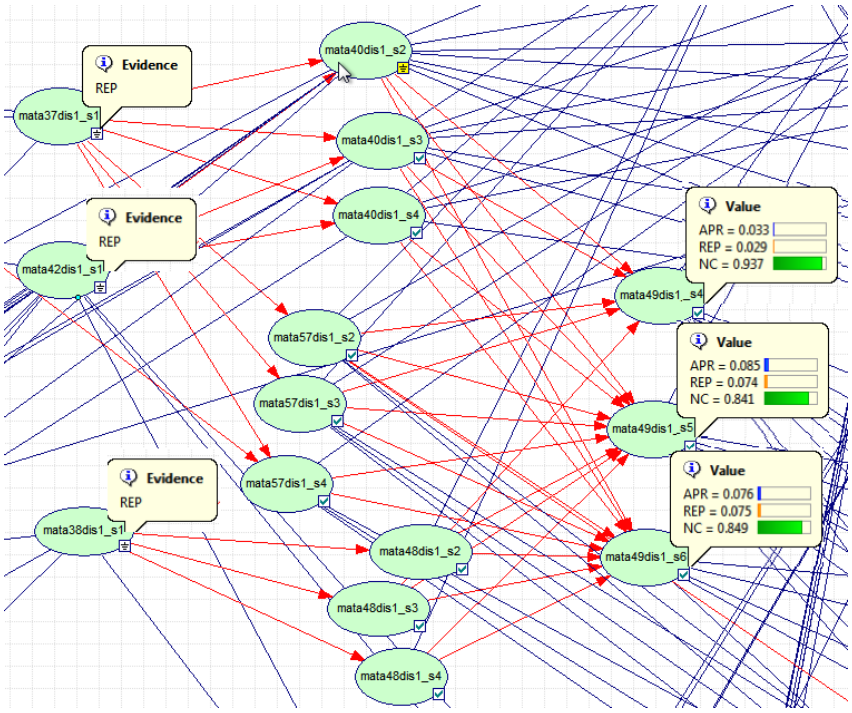

Figura 7.14 Probabilidade de cursar MATA49 dado a **reprovação** em MATA37, MATA42 e MATA38

Diante dos pré-requisitos de MATA49 (Programação de Software Básico), mesmo com a reprovação nas disciplinas básicas existe a possibilidade do aluno cursar MATA49 em qualquer um dos semestres. Como pode ser visto na Figura 7.14, a probabilidade de cursar no quarto semestre é de 6,2%, no quinto 15,9% e no sexto 15,1%. Da mesma forma que os aprovados nas disciplinas básicas, a probabilidade de aprovação sempre é maior que a de reprovação, porém a diferença é pequena entre elas.

A disciplina MATA63 (Engenharia de Software II) recomendada no quinto semestre, tem como pré-requisito MATA62 (Engenharia de Software I), recomendada no quarto semestre, que tem como pré-requisito MATA55 (Programação Orientada a Objetos), recomendada no terceiro semestre, que tem MATA40 (Estrutura de Dados e Algoritmos I), recomendada no segundo semestre, que tem como pré-requisito MATA37 (Introdução à lógica de programação) e MATA42 (Matemática Discreta I) recomendadas no primeiro semestre.

A probabilidade do aluno ser aprovado em MATA63 (Engenharia de Software II) ao cursar no quinto semestre, é de 6,5% e 0,07% de ser reprovado. Para os que cursam no sexto semestre a probabilidade de aprovação é de 3,4% e 0,05% de ser reprovado. Por fim, a probabilidade do aluno cursar no sétimo semestre é de 5,1%, no qual a probabilidade de aprovação é de 3,3% e 1,8% de reprovação.

A respeito dos alunos que aprovam em MATA37 (Introdução a Lógica de Programação) e MATA42 (Matemática Discreta I) as probabilidades de cursar MATA63 (Engenharia de Software II) bem como os seus resultados são ilustrados na Figura 7.15.

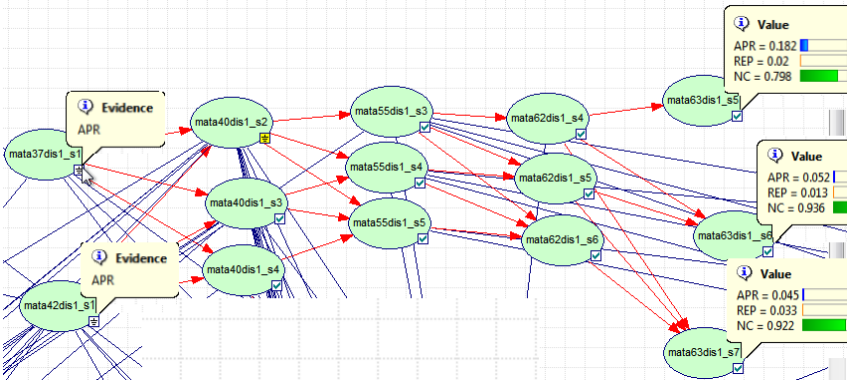

Figura 7.15 Probabilidade de cursar MATA63 dado a **aprovação** em MATA37 e MATA42

Dado a aprovação nas disciplinas MATA42 (Matemática Discreta I) e MATA37 (Introdução a Lógica de Programação) na primeira tentativa, há uma probabilidade do aluno cursar (Engenharia de Software II) no quinto semestre de 18,2%, onde todos são aprovados, já para o sexto semestre a probabilidade do aluno cursar é de 8,8%, no qual 6,5% são aprovados e 2,3% reprovados. Por fim, a probabilidade do aluno cursar MATA63 (Engenharia de Software II) no sétimo semestre dado sua aprovação em MATA42 (Matemática Discreta I) e MATA37 é de 13,4%, no qual a probabilidade de aprovação é de 7,5% e de reprovação 5,9%.

A respeito dos alunos que reprovam em MATA37 (Introdução a Lógica de Programação) e MATA42 (Matemática Discreta I) as probabilidades de cursar MATA63 (Engenharia de Software II) bem como os seus resultados são ilustrados na Figura 7.15.

Dado a reprovação nas disciplinas MATA42 (Matemática Discreta I) e MATA37 (Introdução a Lógica de Programação) na primeira tentativa, há probabilidade de do aluno cursar MATA63 (Engenharia de Software II) no quinto semestre é quase nula, já para cursar no sexto semestre a probabilidade é de 0,9%, onde todos são aprovados. Por fim, a probabilidade de cursar no sétimo semestre é de 3,5%.O restante dos alunos cursaram em semestres posteriores ou evadiram do curso.

A disciplina MATA56 (Paradigmas de Linguagem de Programação) recomendada no quinto semestre, tem a disciplina MATA55 (Programação Orientada a Objetos), recomendada no terceiro semestre, como seu pré-requisito direto, onde a disciplina MATA55 tem MATA40 (Estrutura de Dados e Algoritmos I), recomendada no segundo semestre, como seu pré-requisito e MATA40 (Estrutura de Dados e Algoritmos I) tem MATA37 (In-

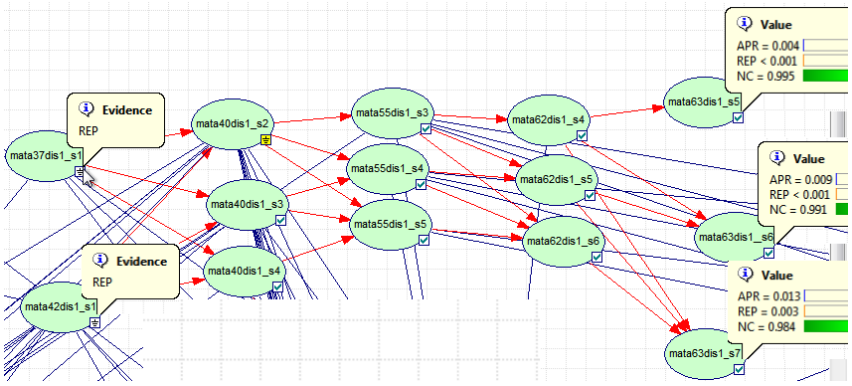

Figura 7.16 Probabilidade de Cursar MATA63 dado a **reprovação** em MATA37 e MATA42

trodução a Lógica de Programação) e MATA42 (Matemática Discreta I), recomendadas no primeiro semestre como seu pré-requisito.

A probabilidade dos alunos cursarem a disciplina MATA55 (Programação Orientada a Objetos) no quinto semestre, é de 8,17%, no qual a probabilidade de aprovação é de 8,1% e 0,07% de reprovação. No sexto semestre a probabilidade do aluno cursar é de 8,1%, no qual a de aprovação é de 5,6% e 2,5% de reprovação. Por fim, a probabilidade de ser aprovado no sétimo semestre é de 2,9% e 3,1% de ser reprovado.

Dado que o aluno tenha sido aprovado nas disciplinas MATA42 (Matemática Discreta I) e MATA37 (Introdução a Lógica de Programação), pré-requisitos indiretos a MATA56 (Paradigmas de Linguagem de Programação), a probabilidade de cursar MATA56, bem como a probabilidade dos seus resultados são apresentados na Figura 7.17. Para os que foram reprovados em MATA42 (Matemática Discreta I) e MATA37 (Introdução a Lógica de Programação) as probabilidades são apresentadas na Figura 7.18.

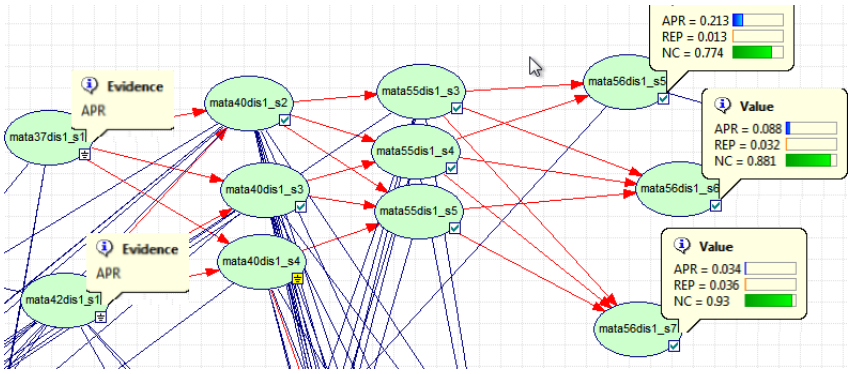

Figura 7.17 Probabilidade de cursar MATA56 dado a **aprovação** em MATA37 e MATA42

Como ilustrado na Figura 7.17 parte dos alunos aprovados em MATA37 (Introdução a Lógica de Programação) e MATA42 (Matemática Discreta I) a probabilidade de cursar a disciplina MATA56 (Paradigmas de Linguagem de Programação) no quinto semestre é de 22,6%, no qual 21,3% conseguem obter aprovação e 1,3% reprovam. Os que cursam no sexto a probabilidade de aprovação é de 8,8% e 3,2% de reprovação. Para os que só

cursam no sétimo semestre a probabilidade de aprovação é de 3,4% e 3,6% de reprovação. O restante dos alunos que não aparecem nas probabilidades das variáveis relativas a MATA56 (Paradigmas de Linguagem de Programação) cursaram em semestres posteriores ou desistiram do curso.

Dado que o aluno tenha sido reprovado nas disciplinas MATA42 (Matemática Discreta I) e MATA37 (Introdução a Lógica de Programação), pré-requisitos indiretos a MATA56 (Paradigmas de Linguagem de Programação), a probabilidade de cursar MATA56, bem como a probabilidade dos seus resultados são apresentados na Figura 7.18.

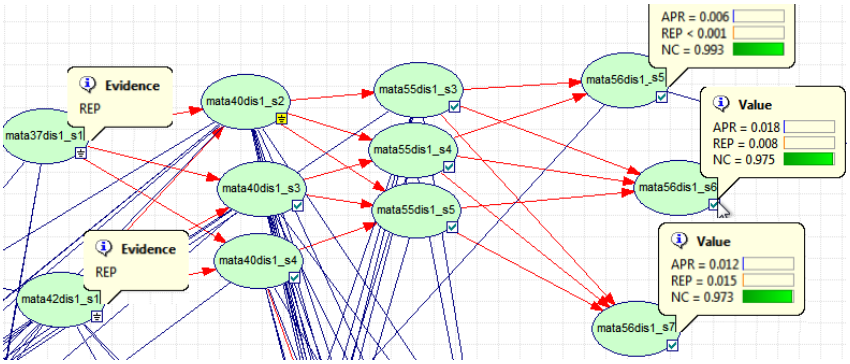

Figura 7.18 Probabilidade de cursar MATA56 dado a **reprovação** em MATA37 e MATA42

Como pode ser visto na Figura 7.18, a probabilidade do aluno cursar MATA56 (Paradigmas de Linguagem de Programação) no quinto semestre é de apenas 0,06%. Já no sexto semestre existe uma probabilidade de 1,88%, no qual 1,8% aprovam e 0,08% reprovam. Por fim, os que cursam no sétimo semestre a probabilidade de aprovação é de 1,2% e reprovação de 1,5%. Mesmo a diferença sendo baixa, maior parte dos alunos tendem a cursar MATA56 (Paradigmas de Linguagem de Programação) no sétimo semestre ao ser reprovado nas duas disciplinas básicas, porém a probabilidade de reprovação é maior do que a de aprovação.

A disciplina MATA53 (Teoria dos Grafos) é recomendada no quinto semestre. Tem como pré-requisito MATA62 (Engenharia de Software I), recomendada no quarto semestre, que tem como pré-requisito MATA55 (Programação Orientada a Objetos), recomendada no terceiro semestre, que tem MATA40 (Estrutura de Dados e Algoritmos I), recomendada no segundo semestre, que tem como pré-requisito MATA37 (Introdução à lógica de programação) e MATA42 (Matemática Discreta I) recomendadas no primeiro semestre.

A probabilidade dos alunos cursarem a disciplina MATA53 no quinto semestre, é de 8,68%, no qual a probabilidade de aprovação é de 8,6% e 0,08% de reprovação. No sexto semestre a probabilidade do aluno cursar é de 5,7%, no qual a de aprovação é de 4,3% e 1,4% de reprovação. Por fim, a probabilidade de ser aprovado no sétimo semestre é de 4,4% e 2,6% de ser reprovado.

Dado a aprovação nas duas disciplinas básicas MATA37 (Introdução a Lógica de Programação) e MATA42 (Matemática Discreta I), a probabilidade do aluno cursar MATA53 em um dado semestre é apresentado na Figura 7.19.

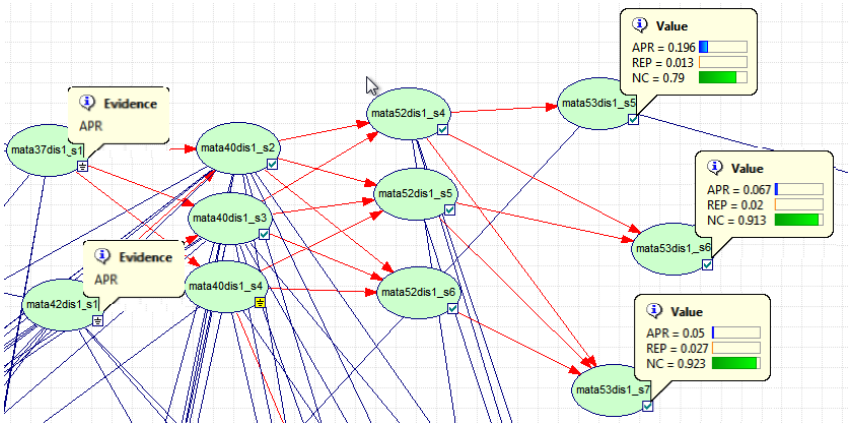

Figura 7.19 Probabilidade de cursar MATA53 dado a **aprovação** em MATA37 e MATA42

De acordo com os resultados apresentados na Figura 7.19 maior parte dos alunos a serem aprovados nas disciplinas MATA37 (Introdução a Lógica de Programação) e MATA42 (Matemática Discreta I) no primeiro semestre cursam a disciplina MATA53 no semestre recomendado (20,9%), onde apenas 1,3% reprovam na disciplina. No sexto semestre a probabilidade de cursar é de 6,9%, no qual a probabilidade de aprovação chega a 6,7% e 2% de reprovação. Por fim, menor parte dos alunos cursam a disciplina no sétimo semestre (7,7%), onde a probabilidade de aprovação é de 5% e 2,7% de reprovação.

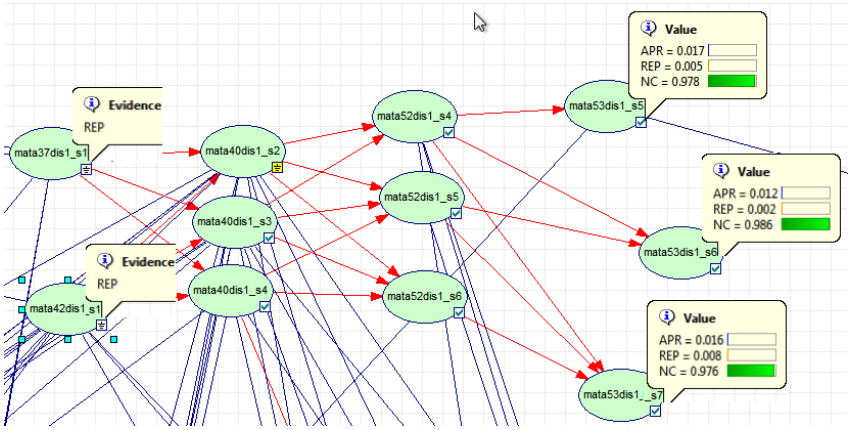

Figura 7.20 Probabilidade de cursar MATA53 dado a **reprovação** em MATA37 e MATA42

Como apresentado na Figura 7.20, a probabilidade do aluno cursar MATA53 no semestre recomendado dado que tenha sido reprovado nas disciplinas MATA37 (Introdução a Lógica de Programação) e MATA42 (Matemática Discreta I) é baixo, mesmo existindo a possibilidade do aluno se recuperar e cursar MATA53 no quinto semestre.

No entanto, dos alunos que são reprovados e cursam MATA53 em algum dos semestres possíveis da rede (quinto, sexto ou sétimo), maior parte dos alunos cursam a disciplina no semestre recomendado (1,75%). No sexto semestre a probabilidade do aluno cursar foi de 1,4% e no sétimo 1,68%. No sétimo semestre a probabilidade de aprovação é de

1,6% e 0,08% de reprovação.

7.2 RESULTADOS DO EXPERIMENTO II

Nesta seção serão apresentados os resultados obtidos para cada questão que deve ser respondida com o segundo experimento. Inicialmente buscou-se responder a seguinte questão 1: **Quais as disciplinas com maiores probabilidades de reter e não reter os alunos em cada semestre?**.

Para isso, utilizou-se a rede bayesiana gerada no experimento II (ver Seção 6.2) evidenciando os estados aprovados e reprovados das variáveis referentes aos alunos cursando disciplinas pela primeira vez recomendadas do primeiro ao quarto semestre. Para cada evidência foi analisada a probabilidade do aluno ser aprovado, reprovado ou não cursar a disciplina MATA67 (Projeto Final de Curso II) (variável utilizada para predizer a retenção final do aluno, ver Seção 6.2).

É importante deixar claro que diferente de um dos resultados apresentados na Seção 7.1, os resultados aqui apresentados é relativo a retenção final do aluno, ou seja se irá concluir o seu curso no tempo regular ou não.

Como pode ser visto na Figura 7.21, são apresentadas as probabilidades dos alunos aprovar, reprovar ou não cursar a disciplina MATA67 (Projeto Final de Curso II) no nono ou décimo semestre. A probabilidade do aluno estar retido no curso é a probabilidade dele não cursar mais a probabilidade de ser reprovado, logo a probabilidade do aluno estar retido é de 86,8%.

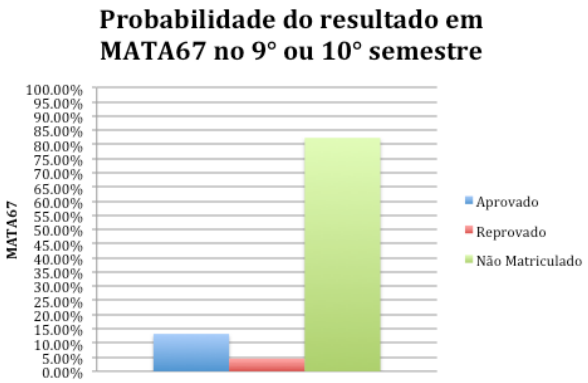

Figura 7.21 Probabilidade da retenção final (probabilidade de reprovado + probabilidade de não matriculado)

Em cursos de graduação, inúmeras vezes os alunos acabam se dedicando mais a uma disciplina do que outras por vários motivos: não conseguem conciliar as disciplinas, julgam uma disciplina como "mais importante" ou "menos importante", tem mais facilidade no aprendizado em determinadas disciplinas, entre outros. Assim, ao decorrer do curso o aluno realiza inúmeras escolhas de quando cursar a disciplina e quanto irá se dedicar a ela.

Desta forma, a identificação dos resultados em disciplinas que mais contribuem para que o aluno consiga não ficar retido no curso irá ajuda-lo a escolher quais disciplinas

cursar em determinados semestre que contribuem mais para uma não retenção no curso.

Diante disto, buscou-se identificar disciplinas que cursadas em um determinado semestre e em uma determinada tentativa aumentam ou reduzem a probabilidade de retenção.

A respeito das disciplinas recomendadas no primeiro semestre e cursadas na primeira tentativa, os alunos a serem aprovado na disciplina MATA42 (Matemática Discreta I) a probabilidade do aluno ser aprovado em MATA67 (Projeto Final de Curso II) é de 26%, a aprovação nas disciplinas MATA01 (Geometria Analítica), MATA02 (Cálculo A) e MATA37 (Introdução a Lógica de Programação) levam a probabilidade de aproximadamente 20% do aluno não ser retido. A aprovação nas disciplinas MATA38 (Projeto de Circuitos Lógicos) e MATA39 foram as que menos contribuíram para uma não retenção, no qual a probabilidade de não retenção é de 15,3%. A Figura 7.22 apresenta estas probabilidades.

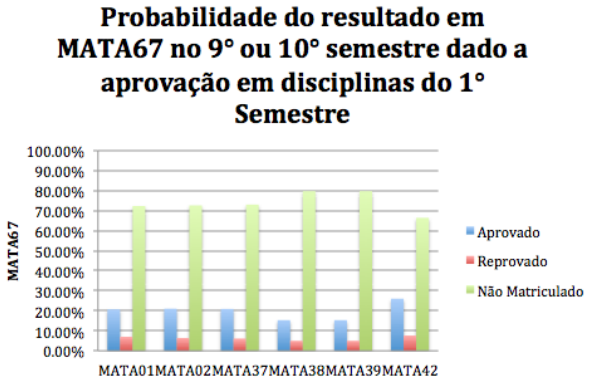

Figura 7.22 Probabilidade da retenção final dado a **aprovação** em disciplinas do primeiro semestre

A reprovação em uma disciplina contribui muito para retenção dos alunos, visto que no curso de Ciência da Computação existem vários pré-requisitos que se não atendidos comprometem diretamente a probabilidade do aluno ser retido. No entanto, o curso estudado apresenta altas probabilidades de reprovação. Nesse sentido, procurou-se identificar as disciplinas que menos contribuem para a retenção dado que o aluno tenha sido reprovado.

Para as disciplinas recomendadas no primeiro semestre na primeira tentativa, a Figura 7.23 mostra que a reprovação em qualquer uma delas leva a quase 100% a probabilidade de retenção. É importante ressaltar que esta probabilidade também foi contribuída pelos alunos que evadiram do curso e estes casos ocorrem na maior parte das vezes quando os alunos estão cursando as disciplinas iniciais.

Para os alunos que cursam disciplinas no segundo semestre na primeira tentativa, logo eles aprovados na disciplina do primeiro semestre, a Figura 7.24 apresenta a probabilidade do aluno obter aprovação, reprovação ou não se matricular no projeto final do curso. A disciplinas que mais contribuíram para não retenção dado que tenha sido aprovado foi MATA97 (Matemática Discreta II), no qual a probabilidade de obter aprovação em MATA97 é de 44,1% e MATA40 (Estrutura de Dados e Algoritmos I), no qual a probabilidade de aprovação em MATA97 (Matemática Discreta II) foi de 40,4%.

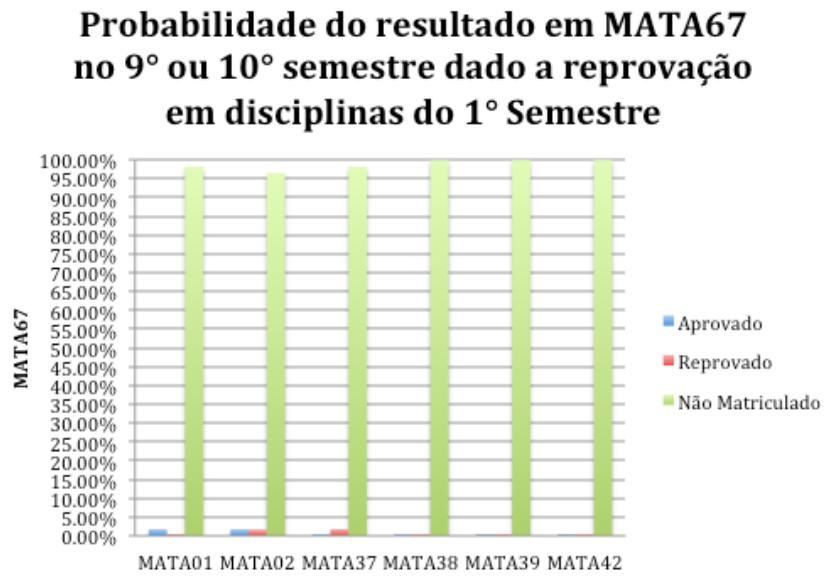

Figura 7.23 Probabilidade da retenção final dado a **reprovação** em disciplinas do primeiro semestre

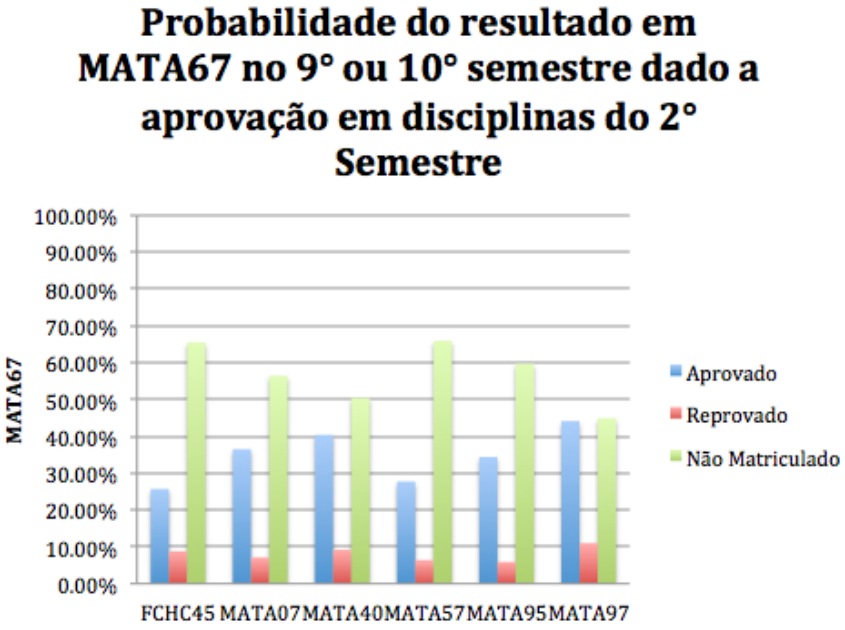

Figura 7.24 Probabilidade da retenção final dado a **aprovação** em disciplinas do segundo semestre

Como já esperado, a probabilidade do aluno ficar retido cursando disciplinas do segundo semestre no semestre recomendado tendem a ser menor se comparada aos alunos que estão cursando disciplinas do primeiro semestre, pois o aluno que cursa uma determinada disciplina no segundo semestre, já cumpriu o pré-requisito do semestre anterior, ou seja, necessita cumprir um número menor de disciplinas para poder cursar a disciplina MATA97 (Matemática Discreta II).

A respeito da reprovação em uma das disciplinas do segundo semestre. De acordo com a Figura 7.25 a disciplina que menos contribui com a retenção dada a sua reprovação foi a disciplina FCHC45 (Metodologia e Expressão Técnico-Científico), no qual a probabilidade de cursar MATA67 (Projeto Final de Curso II) e ser aprovado é de 11,9%. Para os alunos que reprovam em MATA95 (Complementos de Cálculo) na primeira tentativa no segundo semestre, a probabilidade de cursar MATA67 (Projeto Final de Curso II) é de 8,3%, porém a probabilidade de reprovação é de 13,6%.

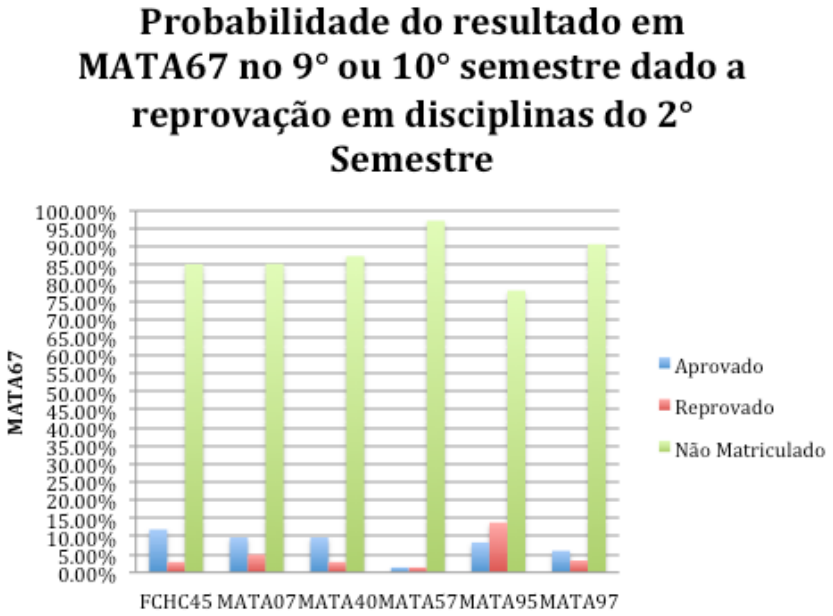

Figura 7.25 Probabilidade da retenção final dado a **reprovação** em disciplinas do segundo semestre

Para os alunos que cursam uma das disciplinas no terceiro semestre na primeira tentativa, a aprovação na disciplina MATA96 (Estatística A) é a que mais contribui para uma não retenção, pois a probabilidade de ser aprovado em MATA67 (Projeto Final de Curso II) é de 56,2%. É importante observar que a disciplina MATA40 (Estrutura de Dados e Algoritmos I) recomendada no segundo semestre contribui mais para uma não retenção do que a disciplina MATA50 (Linguagens Formais e Autômatos). A Figura 7.26 apresenta estas probabilidades.

Das disciplinas recomendadas no terceiro semestre. De acordo com a Figura 7.27 a reprovação na disciplina MATA47 (Lógica de Programação) é a que menos contribui para a retenção dos alunos, onde mesmo reprovado em MATA47 (Lógica de Programação) a

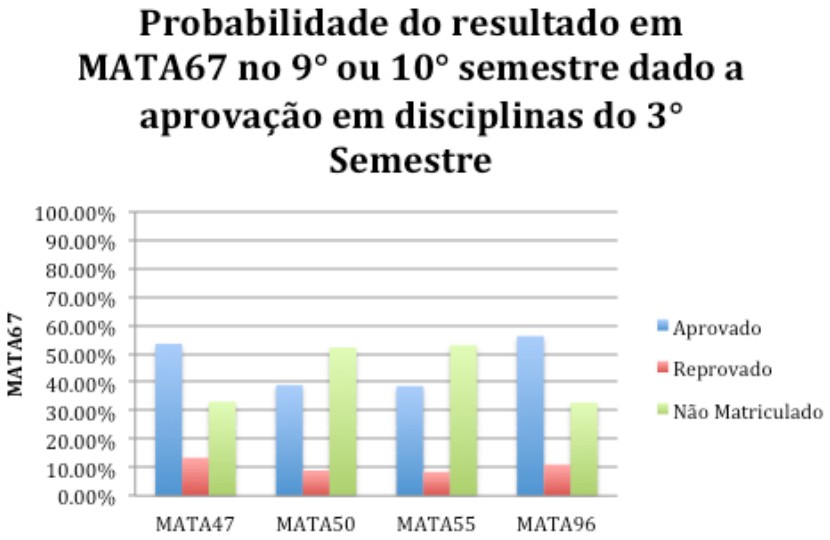

Figura 7.26 Probabilidade da retenção final dado a **aprovação** em disciplinas do terceiro semestre

probabilidade de cursar MATA67 (Projeto Final de Curso II) e ser aprovado é de 20,2%. Interessante que a disciplina MATA50 (Linguagens Formais e Autômatos) é a que menos contribui para uma não retenção ao ser aprovado, porém quando o aluno é reprovado em MATA50 (Linguagens Formais e Autômatos) a probabilidade dele ser aprovado em MATA67 (Projeto Final de Curso II) é a menor dentre as possíveis reprovações em disciplinas do terceiro semestre, apenas 6% .

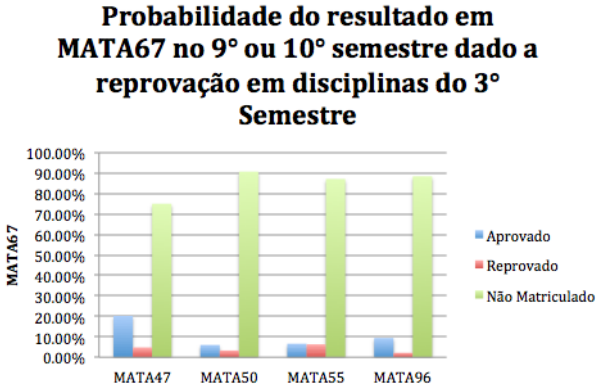

Figura 7.27 Probabilidade da retenção final dado a **reprovação** em disciplinas do terceiro semestre

Por fim, das disciplinas recomendadas no quarto semestre a aprovação na disciplina MATA51 (Teoria da Computação) quando cursada no semestre recomendado na primeira tentativa é a que mais contribui para uma não retenção, onde a probabilidade do aluno não ficar retido é de 55,6%. É interessante observar que a aprovação na disciplina MATA96 (Estatística A) cursada no terceiro semestre na primeira tentativa contribui

mais para uma não retenção comparada a disciplina MATA51 (Teoria da Computação). A disciplina MATA68 (Computador, Ética e Sociedade) é a que menos contribui, onde os aprovados tem a probabilidade de 29,2% de não ser retido. A Figura 7.28 apresenta estas probabilidades.

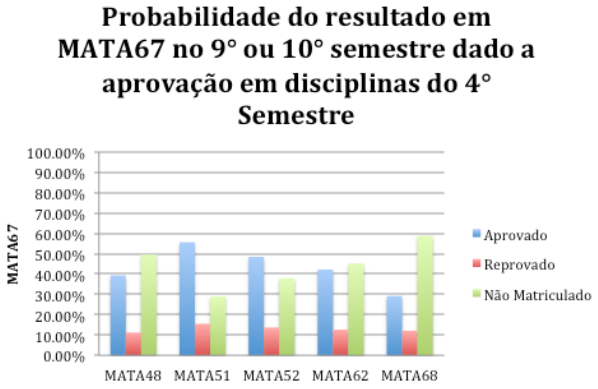

Figura 7.28 Probabilidade da retenção final dado a **aprovação** em disciplinas do quarto semestre

Para as disciplinas recomendadas no quarto semestre, a reprovação em uma delas implica em uma alta probabilidade de retenção. A reprovação nas disciplinas MATA51 (Teoria da Computação), MATA62 (Engenharia de Software I) e MATA68 (Computador, Ética e Sociedade) foram as que menos contribuíram com a retenção, onde a aprovação em qualquer uma delas levou a probabilidade de aprovação em MATA67 (Projeto Final de Curso II) de 11%. A Figura 7.29 apresenta estas probabilidades.

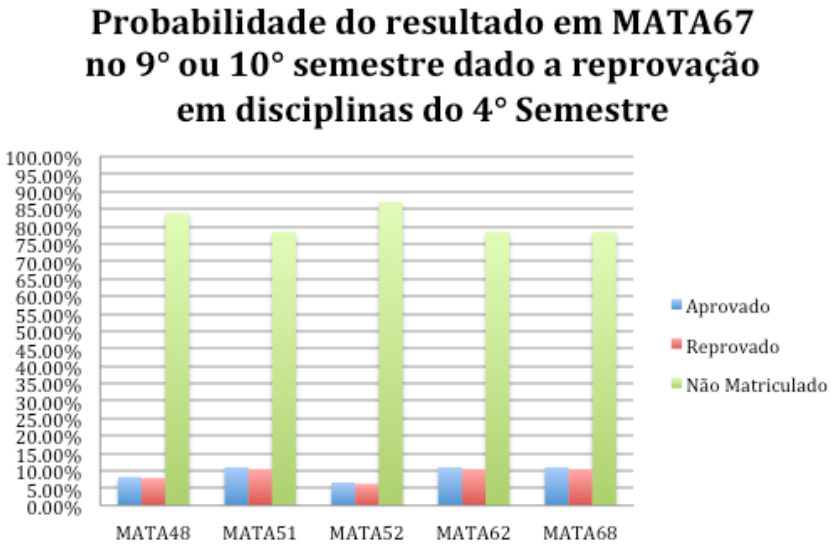

Figura 7.29 Probabilidade da retenção final dado a **reprovação** em disciplinas do quarto semestre

Após a obtenção da probabilidade de retenção final dado o resultado em disciplinas do

1° ao 4° semestre, buscou-se responder a questão 2: **Qual a probabilidade de retenção final diante de uma retenção em um determinado semestre?**.

A partir do possibilidade de classificar um aluno retido ou não em um determinado semestre através da heurística de retenção definida no PROUFBA, buscou-se identificar a probabilidade da retenção final do aluno retido ou não retido do 1° ao 4° semestre.

Para obter este resultado, cada variável de retençãon (n relativo a um determinado semestre) foi evidenciada como *true*, alunos retidos e *false*, alunos não retidos. Para cada evidência foi verificada a probabilidade do aluno concluir o curso no tempo regular, ou seja, ser aprovado na disciplina Projeto Final de Curso II. A Figura 7.30 apresenta estas probabilidades.

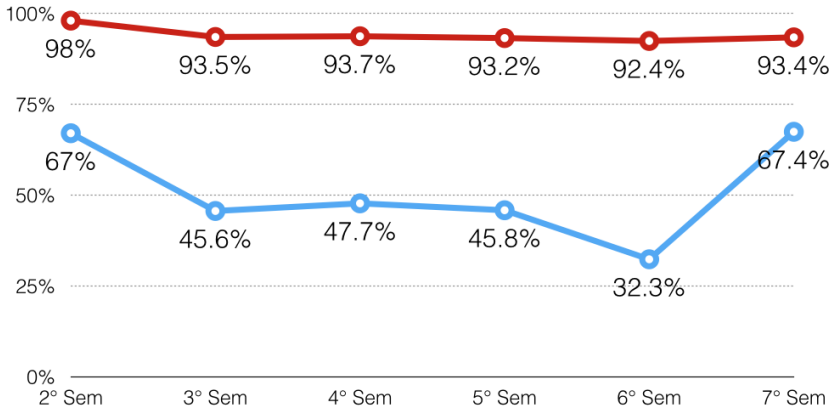

Figura 7.30 Probabilidade da retenção final para os alunos retidos e não retidos em cada semestre

Como pode ser visto na Figura 7.30 a maior probabilidade de retenção final é para os alunos retidos no segundo semestre 98%, para os alunos que chegam no segundo semestre não retidos, a probabilidade de retenção final é de 67%. O sexto semestre apresenta as menores probabilidade de retenção final para os alunos retidos e não retidos, para os retidos a probabilidade é de 92,4% e para os não retidos a probabilidade é de 32,3%.

Ao decorrer dos semestres as probabilidades de retenção final para os alunos não retidos vão diminuindo, isto também acontece para os alunos retidos, porém com uma redução não muito significativa na probabilidade.

Após os resultados aqui apresentados, a seção seguinte apresenta uma análise dos resultados obtidos através dos dois experimentos.

7.3 ANÁLISE DOS RESULTADOS

Os resultados apresentados neste trabalho propuseram uma melhor compreensão sobre o comportamento dos alunos no curso de Ciência da Computação através das redes bayesianas. A partir dos dois experimentos utilizados foi possível responder questões importantes a respeito do comportamento do aluno durante o curso com o foco na retenção.

A análise dos resultados realizada neste trabalho, propõe discutir os resultados obtidos em cada questão respondida pelos experimentos I e II. Diante disto, inicialmente é

apresentada uma discussão sobre o resultado obtidos nas disciplinas pelos alunos retidos nos semestres, possibilitando destacar as disciplinas que mais contribuem com a retenção e não retenção por semestre.

Apenas no segundo semestre a causa da retenção está mais vinculada a reprovação do aluno, ou seja, os alunos cursam a disciplina mas não conseguem obter aprovação. No segundo semestre a disciplina com maior probabilidade de implicar em retenção foi a disciplina MATA42 (Matemática Discreta I), dado que 53,6% reprovam na disciplina ao cursar na primeira tentativa. A disciplina que menos implica em retenção no segundo semestre é a disciplina MATA37 (Introdução a Lógica de Programação), no qual a probabilidade de aprovação é de 58,8%, ou seja a disciplina que os alunos retidos mais tem sido aprovado.

Do terceiro ao sétimo semestre a causa da retenção está mais vinculada ao aluno não cursar o pré-requisito necessário, ou seja, os alunos não conseguem cursar a disciplina que é pré-requisito no semestre atual dado as suas reprovações em semestres anteriores que o impossibilitaram de cursar as disciplinas posteriores.

No terceiro semestre, a disciplina MATA95 (Complementos de Cálculo) tem a maior probabilidade de implicar em retenção, dado que 75,7% dos alunos não cursaram a disciplina, ou seja, chega no terceiro semestre mas ainda não cursou MATA95 (Complementos de Cálculo) recomendada no segundo semestre. É importante observar que MATA95 (Complementos de Cálculo) tem como pré-requisito MATA01 (Geometria Analítica) e MATA02 (Cálculo A) que tem altas probabilidades de reprovação, logo a probabilidade dos alunos não se matricularem em MATA95 (Complementos de Cálculo) no segundo semestre está diretamente relacionada a probabilidade de reprovação nestas duas disciplinas iniciais. A disciplina MATA42 (Matemática Discreta I) tem uma probabilidade de 47,5% de não reter o aluno, sendo classificada como a disciplina que menos retém.

No quarto semestre, a disciplina MATA97 (Matemática Discreta II) tem a maior probabilidade de implicar em retenção, dado que 95,3% dos alunos não cursaram a disciplina. MATA42 (Matemática Discreta I) e MATA95 (Complementos de Cálculo) são os pré-requisitos de MATA97 (Matemática Discreta II), ou seja, MATA42 foi a disciplina com maior probabilidade de reter no segundo semestre, e MATA95 (Complementos de Cálculo) no terceiro. Já a disciplina que menos contribuiu com a retenção dos alunos foi a disciplina MATA38 (Projeto de Circuitos Lógicos), no qual a probabilidade de não reter o aluno no quarto semestre foi de 77,4%. Semelhante ao caso do terceiro semestre, a disciplina MATA38 (Projeto de Circuitos Lógicos) é recomendada no primeiro semestre, assim o aluno tem do primeiro ao terceiro semestre para obter aprovação nesta disciplina e chegar no quarto semestre não retido por causa desta disciplina.

Do segundo ao quarto semestre, percebe-se que as disciplinas com maiores probabilidades que implicam em retenção são disciplinas de cunho matemático. Além disso, com a alta probabilidade de reprovação nas disciplinas do primeiro semestre, principalmente de cunho matemático, os alunos ficam impossibilitados de cursar as disciplinas posteriores, emperrando o seu fluxo acadêmico e contribuindo para a não conclusão no tempo correto.

A respeito do fluxo acadêmico, sabe-se que segui-lo da forma que é recomendado contribui para que o aluno conclua o seu curso no tempo correto, porém inúmeros fatores levam os alunos a não seguir este fluxo, o que acaba prejudicando a conclusão do curso

no tempo regular. Diante disto, intensifica-se a importância de analisar este fluxo para que se crie medidas para facilitar o fluxo do aluno nas disciplinas.

Nesse sentido uma discussão sobre as questões respondidas pelo experimento I a respeito dos resultados dos alunos ao cursar novamente as disciplinas iniciais em que foram reprovados são apresentadas.

A partir das probabilidades de reprovação das disciplinas do primeiro semestre, percebe-se que os alunos tem uma grande dificuldade em obter aprovação quando cursam não só na primeira tentativa mais também em suas tentativas posteriores, principalmente nas disciplinas de cunho matemático.

Na grade curricular do curso de Ciência da Computação, essas disciplinas são pré-requisito em disciplinas posteriores, logo a reprovação nestas disciplinas e a repetição da reprovação prejudica e intensifica a retenção do aluno.

O aluno quando não cumpre o pré-requisito precisa cursar a disciplina novamente e obter aprovação para estar apto a cursar a disciplina que requer o requisito. O caminho menos prejudicial nestes casos é o aluno cursar a disciplina novamente logo no semestre posterior ao que foi reprovado, obter aprovação e posteriormente cursar a disciplina que a tem como pré-requisito, porém, pelas inferências probabilísticas analisadas, na maior parte dos casos este caminho não é o que mais ocorre, pois grande parte dos alunos são reprovados novamente na disciplina ou não a cursa no semestre posterior ao que foi reprovado, implicando em uma retenção difícil de ser reparada.

Das disciplinas analisadas a respeito desta questão, MATA42 (Matemática Discreta I) é a disciplina que mais reprova ao cursar pela primeira vez no primeiro semestre (48,8%), porém ao cursar novamente em semestres posteriores, maior parte dos alunos são aprovados, mas a probabilidade de reprovação ainda é muito alta. No entanto, nas disciplinas MATA02 (Cálculo A) e MATA01 (Geometria Analítica), a probabilidade de reprovação sempre é mais alta do que a de aprovação, quando cursadas na segunda tentativa ou terceira. A reprovação continuada na disciplina irá implicar em uma retenção que dificilmente será revertida.

As disciplinas específicas de computação MATA37 (Introdução a Lógica de Programação) e MATA38 (Projeto de Circuitos Lógicos) apresentam menores probabilidades de reprovação comparadas as de cunho matemático, porém os alunos que reprovam na disciplina MATA38 (Projeto de Circuitos Lógicos) tem maior probabilidade de reprovar novamente do que de aprovar, diferente da disciplina MATA37 (Introdução a Lógica de Programação) que os alunos ao cursar novamente tem maior probabilidade de ser aprovado.

Com esta alta probabilidade de reprovação nas disciplinas do primeiro semestre do curso, grande parte dos alunos não poderão cursar disciplinas no segundo semestre, implicando em uma retenção já no segundo semestre e possivelmente numa retenção maior caso o aluno reprove novamente na disciplina.

Ainda no intuito de entender o fluxo do aluno são realizadas algumas considerações a respeito dos resultados obtidos sobre qual semestre o aluno irá cursar as disciplinas que tem pré-requisitos direto a disciplina que o reprovou ou aprovou e qual será a probabilidade de ser aprovado ou reprovado.

Há uma probabilidade maior dos alunos cursarem MATA97 (Matemática Discreta

II) no terceiro semestre (22,53%) comparado ao quarto semestre (14,77%) para aqueles que foram reprovados em MATA42 (Matemática Discreta I) no primeiro semestre, logo, maior parte dos alunos reprovados em MATA42 tendem obter aprovação em MATA42 no segundo semestre em sua segunda tentativa, visto que maior parte deles cursam MATA97 (Matemática Discreta II) terceiro semestre, porém os alunos continuam com dificuldades em ser aprovado em MATA97 (Matemática Discreta II), pois a probabilidade de reprovação sempre é maior do que a probabilidade de aprovação, independente do semestre em que o aluno cursa MATA97. Desta forma, percebe-se que a reprovação na disciplina MATA42 (Matemática Discreta I) implica em uma probabilidade maior de reprovação em disciplinas com pré-requisito direto a ela.

Observou-se também que dos alunos que cursam MATA97 (Matemática Discreta II) no segundo semestre dado que tenham sido aprovados em MATA42 (Matemática Discreta I) no primeiro semestre, maior parte dos alunos seguem o fluxo de disciplinas recomendadas pela grade curricular. Além disso, a probabilidade de aprovação é maior que a de reprovação, porém a probabilidade de reprovação ainda é alta (37,06%).

Comparando os aprovado dos reprovados em MATA42 (Matemática Discreta I) no primeiro semestre, os alunos que aprovam tendem a aprovar em MATA97 disciplina que requer o pré-requisito no semestre posterior, porém a probabilidade de reprovação ainda é alta. Já os que reprovam na disciplina MATA42 quando cursam MATA97 (Matemática Discreta II) a probabilidade de reprovação é maior do que a de aprovação.

A probabilidade do aluno ser reprovado em MATA02 (Cálculo A) e MATA01 (Geometria Analítica) na primeira tentativa é de 20,87%. Dentre estes reprovados, apenas 18,1% chegam a cursar MATA95 (Complementos de Cálculo) no terceiro, quarto ou quinto semestre, para os que chegam a cursar em algum dos semestres a probabilidade de aprovação é maior do que a de reprovação, exceto para os que cursam no quinto semestre onde a probabilidade de reprovação é de 3,62% e 1,03% de aprovação.

Para os alunos aprovados em MATA02 (Cálculo A) e MATA01 (Geometria Analítica), 78,6% segue o fluxo acadêmico e cursa a disciplina MATA95 (Complementos de Cálculo) no segundo semestre, onde a probabilidade de aprovação (64,09%) é bem maior do que a de reprovação (14,54%) .

Para os alunos reprovados em MATA01 (Geometria Analítica) no primeiro semestre, percebe-se que menos da metade dos alunos cursam a disciplina MATA07 (Álgebra Linear A) no terceiro semestre. Daqueles que conseguiram obter aprovação em MATA01 (Geometria Analítica) no segundo semestre na segunda tentativa e cursou MATA07 (Álgebra Linear A) no terceiro, maior parte deles foram reprovados, ou seja, os alunos continuam com dificuldade em conseguir obter aprovação na disciplina posterior a ela. Para os aprovados em MATA01 (Geometria Analítica) no primeiro semestre, maior parte dos alunos tendem a cursar MATA07 (Álgebra Linear A) no semestre posterior, porém a probabilidade de reprovação ainda é alta.

Por ser uma disciplina de cunho matemático em um curso de Computação, é normal, não ideal, que os alunos tenham dificuldades em obter aprovação, apenas 26,84% são aprovados em MATA42 (Matemática Discreta I) e MATA37 (Introdução a Lógica de Programação) no primeiro semestre, logo, percebe-se que maior parte dos alunos já estarão retidos no segundo semestre, implicando posteriormente em uma retenção continua,

já que a disciplina MATA40 (Estrutura de Dados e Algoritmos I), que tem as duas disciplinas como pré-requisito, é pré-requisito de outras disciplinas em semestres posteriores. Dado a aprovação nestas duas disciplinas, maior parte deles (80,34%) cursam MATA40 (Estrutura de Dados e Algoritmos I) logo no segundo semestre, porém a probabilidade de obter aprovação é de 50%, assim, metade destes alunos estarão retidos no terceiro semestre, já que MATA40 é pré-requisito para uma disciplina no terceiro semestre. Para aqueles que mesmo aptos a cursar MATA40 (Estrutura de Dados e Algoritmos I) no segundo semestre, cursam no terceiro ou quarto, a probabilidade de reprovação é maior.

Nos casos em que os alunos são reprovados em MATA37 (Introdução a Lógica de Programação) no primeiro semestre, apenas 26,9% destes conseguem se recuperar da reprovação em MATA37 (Introdução a Lógica de Programação) e cursar a disciplina MATA57 (Laboratório de Programação I) no terceiro semestre, onde a probabilidade de aprovação é maior do que a de reprovação, porém a probabilidade de reprovação é quase a mesma da de aprovação ao cursar a disciplina no terceiro semestre. Aqueles que cursam MATA37 (Introdução a Lógica de Programação) apenas no quarto semestre a probabilidade de reprovação é maior do que a de aprovação. Para os que são aprovados em MATA37 maior parte cursa a disciplina MATA57 (Laboratório de Programação I) no semestre seguinte e a probabilidade de aprovação é alta comparada a de reprovação, porém 24% dos alunos não cursam a disciplina MATA57 (Laboratório de Programação I) no segundo semestre, mesmo tendo cumprindo o pré-requisito.

A respeito do impacto das disciplinas iniciais em disciplinas do quinto semestre, inicialmente foi constatado que a probabilidade de cursar uma das disciplina no quinto semestre se mostrou extremamente baixa, o que indica que são poucos os casos que os alunos chegam a cursar uma destas disciplinas no semestre recomendado ou em um ou dois semestres a diante. No entanto, as probabilidades de aprovação se mostraram sempre superiores a de reprovação quando o aluno chega a cursar uma das disciplinas do quinto.

Os resultados a respeito do impacto do resultado da disciplina do primeiro semestre em uma disciplina no quinto semestre que a tem como pré-requisito indireto possibilitou verificar que os alunos que são reprovados nas disciplinas do primeiro semestre dificilmente chegam a cursar a disciplina do quinto no semestre recomendado, sexto ou sétimo semestre, indicando que os alunos a serem reprovados nas disciplinas do primeiro semestre tendem a evadir ou cursar as disciplinas do quinto semestre em semestres muito mais superiores.

Sobre as disciplinas do quinto semestre a disciplina que tem maior probabilidade do aluno cursar no semestre recomendado é a disciplina MATA49 (Programação de Software Básico) 45,7%, uma probabilidade muito superior as outras disciplinas do quinto semestre que em média tem uma probabilidade de 8% do aluno cursar. Esta alta probabilidade em cursar MATA49 (Programação de Software Básico) no semestre recomendado pode estar vinculada ao fato dela ter disciplinas do segundo semestre como pré-requisitos direto a ela. O ponto importante nesta alta probabilidade é que esta disciplina é importante no fluxo de disciplinas dado que é pré-requisito para três disciplinas no semestre posterior.

Além disso, foi visto que quando uma disciplina do quinto semestre tem disciplinas do quarto semestre como pré-requisitos direto a ela a probabilidade do aluno cursa-la é menor.

Em todos os casos em que uma disciplina recomendada no quinto semestre tem uma disciplina do quarto semestre como pré-requisito direto, verifica-se que maior parte dos alunos não cursam a disciplina no quinto semestre, pois não conseguem chegar a cursar a disciplina do quarto semestre que é pré-requisito da disciplina recomendada no quinto. Nos casos em que o aluno chega a cursar a disciplina no quarto semestre, dificilmente ele é reprovado e não cursa a disciplina no quinto.

Os alunos que não cursam a disciplina no quinto tem suas probabilidades de reprovação nas disciplinas iniciais muito altas, em alguns casos chega a ser maior do que a de aprovação, ou seja, a maior parte dos alunos que não cursam a disciplina do quinto já inicia o curso sendo reprovado nas disciplinas o que dificulta a possibilidade de cursa-la no semestre recomendado.

De forma geral sobre os resultados obtidos no experimento I algumas considerações devem ser ressaltadas:

- ser aprovado nos pré-requisitos aumenta a probabilidade de aprovação da disciplina seguinte na maior parte dos casos;
- ser reprovado nos pré-requisitos aumenta a probabilidade de reprovação da disciplina seguinte na maior parte dos casos;
- as disciplinas iniciais tem um grande impacto no resultado do aluno em disciplinas posteriores;
- a reprovação em disciplinas básicas gera uma retenção nos semestres posteriores que dificilmente poderá ser reparada;
- a probabilidade de retenção de um pré-requisito em um determinado semestre tende a aumentar ao decorrer dos semestres;
- a probabilidade de reprovação nas disciplinas tende a diminuir em semestres mais avançados, como o quinto semestre;
- as disciplinas de cunho matemático tem maiores probabilidades de reprovação comparado a outras disciplinas. Além disso, os alunos que aprovam nessas disciplinas, tendem a ser aprovado nas disciplinas de cunho específico da computação;
- a disciplina MATA40 (Estrutura de Dados e Algoritmos I), recomendada no segundo semestre, e tem como pré-requisito MATA42 (Matemática Discreta I) e MATA37 (Introdução a Lógica de Programação), recomendadas no primeiro semestre, é pré-requisito para disciplina do terceiro (MATA55 - Programação Orientada a Objetos), quarto (MATA52) e quinto (MATA49 - Programação de Software Básico) semestre. Assim, o aluno que não cursa MATA40 (Estrutura de Dados e Algoritmos I) no semestre recomendado, está retido até o quinto semestre. Diante das altas probabilidade de reprovação em MATA42 (Matemática Discreta I) e MATA40 (Estrutura de Dados e Algoritmos I), percebe-se que esses casos acontecem constantemente. A partir destas probabilidades, sugere-se que a disciplina MATA40 (Estrutura de Dados e Algoritmos I) não tenha como pré-requisito MATA42 (Matemática Discreta

I) que é de cunho matemático e tem uma grande probabilidade de reprovação a fim de flexibilizar o caminho percorrido pelo aluno na grade curricular e reduzir a retenção;

- a probabilidade de retenção está mais vinculada a reprovação dos alunos nos pré-requisitos apenas no segundo semestre, ou seja, o aluno cursa os requisitos no semestre anterior, mas não obtém aprovação. Nos semestres posteriores, a retenção está mais vinculada ao aluno não cursar o pré-requisito, ou seja, os alunos estão retidos em um determinado semestre porque não cursaram as disciplinas anteriores que são pré-requisitos no semestre em que está retido.

- ao reprovarem em disciplinas básicas, a probabilidade de reprovação novamente na disciplina ainda é alta em grande parte dos casos.

Sobre os resultados do experimento II que focou na predição da retenção final do aluno e não no fluxo acadêmico como no experimento I alguns pontos importantes devem ser discutidos.

No curso de Ciência da Computação a probabilidade do aluno não concluir o curso no tempo regular é muito alta (86,8%). As probabilidades a respeito da retenção dado o resultado em disciplinas podem ajudar na compreensão da retenção no curso, bem como ajudar os alunos a identificar disciplinas que o seu resultado são mais importantes em um determinado semestre em certa tentativa.

Além disso, a predição da retenção em semestres iniciais podem ajudar o colegiado a intervir em casos em que a probabilidade de retenção para um determinado perfil de aluno é muito alta, mesmo que a intervenção não possa garantir que o aluno conclua o curso no semestre recomendado, reduzir ao máximo o tempo para que possa obter a conclusão do curso.

Quando o aluno reprova em qualquer uma das disciplinas recomendadas no primeiro semestre a probabilidade dele não concluir o curso no tempo correto é de quase 100%.

Das disciplinas recomendadas no segundo semestre a disciplina FCHC45 (Metodologia e Expressão Técnico-Científico) foi a que menos impactou numa retenção quando o aluno é reprovado ao cursa-la pela primeira vez no segundo semestre. Isto acontece visto que esta disciplina apenas é pré-requisito para MATA66 (Projeto Final de Curso I) no oitavo semestre. Como já descrito a importância da disciplina MATA40 (Estrutura de Dados e Algoritmos I) nesta seção, os resultados apresentados deste experimento intensificam essa constatação ao mostrar que das disciplinas do segundo semestre ela é a disciplina que mais contribui para uma não retenção quando se obtêm aprovação no semestre recomendado.

No terceiro semestre a disciplina que ao ser aprovado no semestre recomendado mais contribuiu para a não retenção é MATA96 (Estatística A). É interessante observar que MATA96 é a última disciplina de cunho matemático que tem como pré-requisito disciplinas de cunho matemático em semestres anteriores. Assim, o aluno que chega no terceiro semestre cumprindo as disciplinas de cunho matemático aumenta a sua probabilidade de não retenção. Além disso, a reprovação na disciplina MATA50 (Linguagens Formais e Autômatos) é a que menos contribui com a retenção comparada a reprovação nas outras disciplinas do terceiro.

Diante destes resultados foi possível identificar as disciplinas em que o aluno pode priorizar para que aumente a sua probabilidade de não retenção e possíveis disciplinas que mesmo sendo reprovados não irá implicar em uma alta probabilidade de retenção.

Por fim, a predição da retenção final do aluno a partir da retenção em um determinado semestre tem o objetivo de ajudar o colegiado a intervir nos casos em que os alunos estão retidos em determinados semestres e a acompanhar os que não estão.

Sobre esses resultados, observou-se que a probabilidade de retenção final é sempre alta quando o aluno chega retido em qualquer um dos semestres. Isto pode estar correlacionado com o número de pré-requisitos da grade curricular, onde dificilmente o aluno consegui se recuperar quando chega a estar retido em um determinado semestre. De forma contrária a probabilidade de retenção final do aluno tende a reduzir ao decorrer dos semestres ao ser classificado como não retido.

De forma geral sobre os resultados obtidos no experimento II algumas considerações devem ser ressaltadas:

- todas as disciplinas do primeiro semestre devem ser priorizadas para que o aluno não fique retido no fim do curso;
- no segundo semestre, a única disciplina que não contribui significativamente com a probabilidade de não retenção final no curso é a disciplina FCHC45 (Metodologia e Expressão Técnico-Científico).
- no terceiro semestre, as disciplinas que mais devem ser priorizadas são MATA47 (Lógica de Programação) e MATA96 (Estatística A). A disciplina MATA47 (Lógica de Programação) é a que menos deve se priorizar.
- no quarto semestre, as disciplinas que mais contribuem para que o aluno não fique retido no fim do curso são MATA51 (Teoria da Computação) e MATA52 (Análise e Projeto de Algoritmos.
- a probabilidade da retenção final dado a retenção em qualquer um dos semestres do curso fica em torno de 93%. Assim, o aluno que fica retido em qualquer um dos semestres tende a não concluir o curso no período regular.
- a probabilidade da retenção final dado a não retenção em um dos semestres reduz significativamente quando o aluno está no terceiro, quarto, quinto ou sexto semestre não retido.

Com as discussões aqui apresentadas sobre os resultados obtidos, espera-se que sejam utilizadas para uma melhor compreensão do comportamento do aluno diante a grade curricular, bem como sua retenção para que seja possível criar políticas de retenção a fim de reduzir as taxas de retenção dos alunos.

Capítulo

8

.

CONCLUSÕES

A análise da retenção é uma tarefa complexa que ainda está crescendo no Brasil e precisa ser explorada em várias áreas da ciência para que se possa ter cada vez mais informações que contribuam para a criação de políticas de retenção eficazes nas instituições.

O projeto PROUFBA apresentou resultados importantes na retenção dos alunos e teve contribuição significativa para o desenvolvimento deste trabalho. A metodologia utilizada neste trabalho, em analisar o aluno que cursa uma disciplina em um determinado semestre em uma certa tentativa no curso de Ciência da Computação baseou-se na metodologia utilizada no PROUFBA que analisou o aluno inscrito em uma disciplina em um determinado semestre. Assim, foi possível utilizar parte dos dados já pré-processados para compor o conjunto de dados desta pesquisa. Além disso, foi a partir dos resultados obtidos que percebeu-se a necessidade de elaborar questões a serem respondidas probabilisticamente a respeito do fluxo acadêmico do aluno de acordo com os pré-requisitos entre as disciplinas. Por fim, só foi possível realizar algumas inferências probabilísticas a respeito da retenção em cada semestre do curso diante da heurística de retenção definida anteriormente.

Este trabalho teve o objetivo de analisar probabilisticamente a retenção dos alunos do curso de Ciência da Computação a partir do fluxo acadêmico do aluno ao cursar disciplinas em determinados semestres em certas tentativas e predizer probabilisticamente a retenção final dos alunos dado alguns resultados em disciplinas. Para isso, foi proposto dois experimentos: i) definição de uma rede bayesiana manualmente baseado na grade curricular dos alunos e ii) a utilização do algoritmo Naive Bayes para criar um classificador bayesiano. A partir destes experimentos foi possível inferir uma série de probabilidades que compõem os resultados deste trabalho e uma análise dos resultados obtidos.

Os resultados apresentados neste trabalho têm o objetivo de proporcionar uma melhor compreensão do comportamento dos alunos no curso de Ciência da Computação da UFBA, especificamente na retenção para que o colegiado ou orgãos competentes possam criar políticas de retenção para acompanhar e intervir durante o andamento do curso. Algumas políticas podem ser: i) acompanhamento do desempenho do aluno durante e

no fim de cada semestre; ii) orientação sobre disciplinas com maiores probabilidades de reprovação; iii) orientação sobre um melhor conjunto de disciplinas para o aluno se matricular a depender do seu desempenho em disciplinas anteriores; iv) cursos de reforço para disciplinas com altas probabilidades de reprovação; v) orientação sobre altas probabilidades de reprovação em uma disciplina de acordo com resultados anteriores obtidos pelo aluno.

Além disso, com esses resultados pretende-se colaborar com o colegiado nos ajustes da grade curricular, uma sugestão foi feita na seção 7.3, visando possíveis modificações que procurem contribuir com redução da retenção.

Com os resultados da predição da retenção final do aluno a partir de resultados em disciplinas do 1° ao 4° semestre pretende-se proporcionar subsídios ao colegiado para intervir em casos em que a probabilidade de retenção final para um determinado perfil de aluno é muito alta, mesmo que a intervenção não possa garantir que o aluno conclua o curso no tempo regular, que ela possa reduzir ao máximo o tempo para se obter a conclusão do curso. Essa intervenção têm o sentindo de acompanhar o andamento do aluno procurando motivá-lo e identificar maneiras para que ele reduza o número de semestres necessários para concluir o curso que podem ser: i) cursar uma disciplina de férias e ii) selecionar disciplinas para que o aluno curse em determinados semestres.

Nesse sentindo as principais contribuições deste trabalho são descritas:

- Identificação de disciplinas com maiores probabilidades de reter e não reter em cada semestre, bem como a probabilidade de aprovação dos alunos retidos em cada semestre;

- Análise do impacto de um resultado em uma disciplina do primeiro semestre em disciplinas do semestre posterior que a tem como pré-requisito direto ou indireto;

- Reconhecimento de resultados em disciplinas que aumentam ou reduzem a probabilidade de retenção final do aluno;

- Mensuração da probabilidade de retenção final do aluno diante da retenção ou não retenção em um determinado semestre;

- Análise dos resultados obtidos com propósito de auxiliar o curso em uma melhor compreensão do andamento do aluno no curso para que seja possível criar políticas de retenção no intuito de reduzir as altas probabilidades de retenção.

Por fim, seguem as publicações dos artigos em simpósios de renome no intuito de validar os resultados que foram obtidos, bem como contribuir para a pesquisa científica da área. Estas publicações são listadas a seguir:

- Silva, Carlos VA; Santos, Marcelo S; Silva, Marcos; Claro, Daniela B; Lima, Veronica MC; Ribeiro, Silvana; Telles, Ana R; Lopes, Denivaldo. *Mining Retention Rules from Student Transcripts: A Case Study of the Information Systems programme at a Federal University. Simpósio Brasileiro de Informática na Educação,*

Novembro 25-29, São Paulo, Brazil, 2013. Anais do Simpósio Brasileiro de Informática na Educação, volume 24, número 1 (*Qualis B2*) (SILVA et al., 2013). Neste trabalho foi realizado uma análise sobre a retenção dos alunos no curso de Sistemas de Informação da UFBA. Este trabalho teve por objetivo utilizar regras de associação para identificar se existe correlação entre disciplinas que contribuem para a retenção dos alunos, ou seja, disciplinas que recomendadas em um determinado semestre implicam na reprovação em ambas, bem como disciplinas que alunos retidos tendem a ser aprovado ou reprovado. Com essas informações, foi possível sugerir algumas modificações na grade curricular a fim de reduzir o índice de retenção dos alunos.

- Santos, Marcelo S; Santana, Liz C; Pereira, Quemuel L; Silva, Marcos; Claro, Daniela B e Lima, Veronica MC e Vieira, Vaninha e Ribeiro, Silvana; Telles, Ana R; Lopes, Denivaldo. *Mining Retention Rules from Student Transcripts: A Case Study of the programs at a Federal University. Simpósio Brasileiro de Informática na Educação, Novembro 03-06, Mato Grosso do Sul, Brazil, 2014. Anais do Simpósio Brasileiro de Informática na Educação, volume 25, número 1* (*Qualis B2*) (SANTOS et al., 2014). Este trabalho utilizou as regras de associação para identificar regras frequentes e infrequentes de todos os cursos da UFBA. Com a utilização das regras infrequentes foi possível identificar relações entre disciplinas que não ocorriam frequentemente, porém eram correlações importantes no contexto da retenção avaliada pelos especialistas da área. Além disto, criou-se um panorama de disciplinas e cursos que mais retêm por semestre a partir da heurística de retenção definida, obtendo as seguintes informações: cursos que mais retêm na UFBA, semestres que têm maiores taxas de retenção e disciplinas que mais retêm por semestre.

- Santos, Marcelo S e Claro, Daniela B e Lima, Veronica MC. *A probabilistic analysis of student retention in a Federal University: A Case Study of a Computer Science Program. Simpósio Brasileiro de Informática na Educação, Outubro 26-30, Maceió, Brazil, 2015. Anais do Simpósio Brasileiro de Informática na Educação* (*Qualis B2*) (SILVA; CLARO; LIMA, 2015). Neste trabalho utilizou-se redes bayesianas para analisar probabilidades no fluxo acadêmico dos alunos no intuito de predizer se o aluno vai aprovar, reprovar ou não cursar uma disciplinas em semestres posteriores a partir de resultados obtidos em disciplinas básicas do curso, bem como a predição de resultados em disciplinas que o aluno reprova e cursa novamente. Parte dos resultados desta dissertação foram publicadas neste artigo.

j

8.1 TRABALHOS FUTUROS

No trabalho desenvolvido, algumas variáveis e informações foram reduzidas ou removidas devido a não fazer parte dos objetivos do trabalho ou a limitações de algoritmos, porém, não deixam de ser informações relevantes que devem ser avaliadas. Nesse sentido algumas hipóteses são apresentadas:

- A reprovação em disciplinas sem pré-requisito tem impacto no resultado em outras disciplinas? Para responder essa questão, pode ser utilizado regras de associação para encontrar regras onde a reprovação em disciplina que não é pré-requisito implicam na reprovação em outras disciplinas.

- Como fazer uma análise com todas as variáveis a respeito de alunos cursando disciplinas em um determinado semestre em certa tentativa? Nesta pesquisa, utilizou-se algoritmos de aprendizado de parâmetros e inferências probabilísticas exatos nas redes bayesianas, ou seja, as probabilidades calculas apresentam resultados exatos, o que demanda de um custo computacional grande. Para utilizar todas as variáveis, talvez seja possível com a utilização de algoritmos aproximados que tem um custo computacional menor, mas trazem resultados próximos ao esperado, como: *Forward Sampling, Likelihood Weighting e Gibbs Sampling.*

Além disso, foi visto nos resultados altas probabilidades a respeito dos alunos não cursarem uma disciplina em um dado semestre em certa tentativa. A partir dos dados utilizados, é impossível identificar os motivos que fizeram o aluno a não cursar a disciplina. Nesse sentido, seria interessante aplicar um *survey* para os alunos a respeito dos motivos que fizeram o aluno a não cursar uma disciplina. Assim, nos casos em que exista uma alta probabilidade do aluno não cursar uma certa disciplina e seus respectivos motivos, o colegiado terá mais informações para poder criar políticas para que esses casos não ocorram.

Também, como um projeto futuro que estende o realizado no PROUFBA sugere-se destacar quais combinações de locais e horários tem prejudicado os discentes na sua semestralização. Especificamente, pretende-se responder aos seguintes questionamentos: (i) quais horários há um maior índice de reprovações por componente curricular cursado, diferenciando cursos noturnos e diurnos? (ii) quais combinações de horários e locais interferem no desempenho acadêmico do alunado? Há uma combinação de horários e locais independente de componentes curriculares que prejudicam o desempenho do alunado? (iii) Há necessidade de transporte intercampi para minimizar as taxas de reprovações? (iv) Há diferenciações em relação aos discentes dos BIs?

Por fim, pretende-se utilizar dados sócio-econômicos e dados do desempenho dos alunos durante o ensino médio no intuito de criar um sistema bayesiano para predição probabilística da retenção em Universidades e Institutos Federais utilizando vários modelos probabilísticos (Naive Bayes, Tree-Argumented-Naive Bayes, General Bayes).

A ideia é desenvolver um sistema onde os colegiados possam utilizar para analisar a probabilidade de um aluno reter com a utilização de vários algoritmos, onde apenas o algoritmo com melhor classificação (baseado nas medidas de avaliação) será apresentado. Assim, ao entrar no curso a ferramenta irá informar uma probabilidade de retenção do aluno a partir do seu perfil e essa probabilidade será atualizada diante dos resultados obtidos pelos alunos durante o curso. Com isso, os colegiados podem estar intervindo em possíveis retenções desde a entrada do aluno no curso.

Nesse sentido, pretende-se estender este trabalho com o desenvolvimento da ferramenta aqui descrita em Institutos Federais, não só nos cursos superiores, mas também

em cursos subsequentes e integrados que vem encontrando problemas no tocante da retenção.

Apêndice

A

.

DESCRIÇÃO VARIÁVEIS UTILIZADAS NOS EXPERIMENTOS I E II

Tabela A.1: Descrição das variáveis utilizadas no Experimento I e II

Variáveis	Descrição da Variável
fisa75dis1_s3	Alunos que cursam FISA75 pela primeira vez no terceiro semestre
fisa75dis1_s4	Alunos que cursam FISA75 pela primeira vez no quarto semestre
mata01dis1_s1	Alunos que cursam MATA01 pela primeira vez no primeiro semestre
mata01dis2_s2	Alunos que cursam MATA01 pela segunda vez no segundo semestre
mata01dis3_s3	Alunos que cursam MATA01 pela terceira vez no terceiro semestre
mata02dis1_s1	Alunos que cursam MATA02 pela primeira vez no primeiro semestre
mata02dis2_s2	Alunos que cursam MATA02 pela segunda vez no segundo semestre
mata02dis2_s3	Alunos que cursam MATA02 pela segunda vez no terceiro semestre
mata02dis3_s3	Alunos que cursam MATA02 pela terceira vez no terceiro semestre
mata02dis3_s4	Alunos que cursam MATA02 pela terceira vez no quarto semestre
mata07dis1_s3	Alunos que cursam MATA07 pela primeira vez no terceiro semestre
mata07dis2_s3	Alunos que cursam MATA07 pela segunda vez no terceiro semestre
mata07dis2_s4	Alunos que cursam MATA07 pela segunda vez no quarto semestre
mata37dis1_s1	Alunos que cursam MATA47 pela primeira vez no primeiro semestre
mata37dis2_s2	Alunos que cursam MATA07 pela segunda vez no segundo semestre
mata38dis1_s1	Alunos que cursam MATA38 pela primeira vez no primeiro semestre
mata38dis2_s2	Alunos que cursam MATA38 pela segunda vez no segundo semestre
mata38dis2_s3	Alunos que cursam MATA38 pela segunda vez no terceiro semestre
mata40dis1_s2	Alunos que cursam MATA40 pela primeira vez no segundo semestre
mata40dis1_s3	Alunos que cursam MATA40 pela primeira vez no terceiro semestre
mata40dis1_s4	Alunos que cursam MATA40 pela primeira vez no quarto semestre
mata40dis2_s3	Alunos que cursam MATA40 pela segunda vez no terceiro semestre

mata40dis2_s4	Alunos que cursam MATA40 pela segunda vez no quarto semestre
mata40dis2_s5	Alunos que cursam MATA40 pela segunda vez no quinto semestre
mata40dis3_s4	Alunos que cursam MATA40 pela terceira vez no quarto semestre
mata42dis1_s1	Alunos que cursam MATA42 pela primeira vez no primeiro semestre
mata42dis2_s2	Alunos que cursam MATA42 pela segunda vez no segundo semestre
mata42dis2_s3	Alunos que cursam MATA42 pela segunda vez no segundo semestre
mata42dis3_s3	Alunos que cursam MATA42 pela terceira vez no terceiro semestre
mata47dis1_s3	Alunos que cursam MATA47 pela primeira vez no terceiro semestre
mata47dis1_s4	Alunos que cursam MATA47 pela primeira vez no quarto semestre
mata47dis1_s5	Alunos que cursam MATA47 pela primeira vez no quinto semestre
mata48dis1_s2	Alunos que cursam MATA48 pela primeira vez no segundo semestre
mata48dis1_s3	Alunos que cursam MATA48 pela primeira vez no terceiro semestre
mata48dis1_s4	Alunos que cursam MATA48 pela primeira vez no quarto semestre
mata48dis1_s5	Alunos que cursam MATA48 pela primeira vez no quinto semestre
mata49dis1_s4	Alunos que cursam MATA49 pela primeira vez no quarto semestre
mata49dis1_s5	Alunos que cursam MATA49 pela primeira vez no quinto semestre
mata49dis1_s6	Alunos que cursam MATA49 pela primeira vez no sexto semestre
mata50dis1_s3	Alunos que cursam MATA50 pela primeira vez no terceiro semestre
mata50dis1_s4	Alunos que cursam MATA50 pela primeira vez no quarto semestre
mata50dis1_s5	Alunos que cursam MATA50 pela primeira vez no quinto semestre
mata50dis2_s4	Alunos que cursam MATA50 pela segunda vez no quarto semestre
mata50dis2_s5	Alunos que cursam MATA50 pela segunda vez no quinto semestre
mata51dis1_s4	Alunos que cursam MATA51 pela primeira vez no quarto semestre
mata51dis1_s5	Alunos que cursam MATA51 pela primeira vez no quinto semestre
mata51dis1_s6	Alunos que cursam MATA51 pela primeira vez no sexto semestre
mata52dis1_s5	Alunos que cursam MATA52 pela primeira vez no quinto semestre
mata52dis1_s6	Alunos que cursam MATA52 pela primeira vez no sexo semestre
mata53dis1_s6	Alunos que cursam MATA53 pela primeira vez no sexto semestre
mata53dis1_s7	Alunos que cursam MATA53 pela primeira vez no sétimo semestre
mata54dis1_s5	Alunos que cursam MATA54 pela primeira vez no quinto semestre
mata54dis1_s6	Alunos que cursam MATA54 pela primeira vez no sexto semestre
mata55dis1_s3	Alunos que cursam MATA55 pela primeira vez no terceiro semestre
mata55dis1_s4	Alunos que cursam MATA55 pela primeira vez no quarto semestre
mata55dis1_s5	Alunos que cursam MATA55 pela primeira vez no quinto semestre
mata56dis1_s5	Alunos que cursam MATA56 pela primeira vez no quinto semestre
mata56dis1_s6	Alunos que cursam MATA56 pela primeira vez no sexto semestre
mata56dis1_s7	Alunos que cursam MATA56 pela primeira vez no sétimo semestre
mata57dis1_s2	Alunos que cursam MATA57 pela primeira vez no segundo semestre
mata57dis1_s3	Alunos que cursam MATA57 pela primeira vez no terceiro semestre
mata57dis1_s4	Alunos que cursam MATA57 pela primeira vez no quarto semestre
mata58dis1_s6	Alunos que cursam MATA58 pela primeira vez no sexto semestre
mata58dis1_s7	Alunos que cursam MATA58 pela primeira vez no sétimo semestre

mata58dis1_s8	Alunos que cursam MATA58 pela primeira vez no oitavo semestre
mata59dis1_s6	Alunos que cursam MATA59 pela primeira vez no sexto semestre
mata59dis1_s7	Alunos que cursam MATA59 pela primeira vez no sétimo semestre
mata60dis1_s6	Alunos que cursam MATA60 pela primeira vez no sexto semestre
mata60dis1_s7	Alunos que cursam MATA60 pela primeira vez no sétimo semestre
mata61dis1_s6	Alunos que cursam MATA61 pela primeira vez no sexto semestre
mata61dis1_s7	Alunos que cursam MATA61 pela primeira vez no sétimo semestre
mata61dis1_s8	Alunos que cursam MATA61 pela primeira vez no oitavo semestre
mata62dis1_s4	Alunos que cursam MATA62 pela primeira vez no quarto semestre
mata62dis1_s5	Alunos que cursam MATA62 pela primeira vez no quinto semestre
mata62dis1_s6	Alunos que cursam MATA62 pela primeira vez no sexto semestre
mata63dis1_s6	Alunos que cursam MATA63 pela primeira vez no sexto semestre
mata63dis1_s7	Alunos que cursam MATA63 pela primeira vez no sétimo semestre
mata64dis1_s6	Alunos que cursam MATA64 pela primeira vez no sexto semestre
mata65dis1_s6	Alunos que cursam MATA65 pela primeira vez no sexto semestre
mata65dis1_s7	Alunos que cursam MATA65 pela primeira vez no sétimo semestre
mata65dis1_s9	Alunos que cursam MATA65 pela primeira vez no nono semestre
mata95dis1_s2	Alunos que cursam MATA95 pela primeira vez no segundo semestre
mata95dis1_s3	Alunos que cursam MATA95 pela primeira vez no terceiro semestre
mata95dis1_s4	Alunos que cursam MATA95 pela primeira vez no quarto semestre
mata95dis1_s5	Alunos que cursam MATA95 pela primeira vez no quinto semestre
mata95dis2_s3	Alunos que cursam MATA95 pela segunda vez no terceiro semestre
mata96dis1_s3	Alunos que cursam MATA96 pela primeira vez no terceiro semestre
mata97dis1_s2	Alunos que cursam MATA97 pela primeira vez no segundo semestre
mata97dis1_s3	Alunos que cursam MATA97 pela primeira vez no terceiro semestre
mata97dis1_s4	Alunos que cursam MATA97 pela primeira vez no quarto semestre
mata97dis2_s3	Alunos que cursam MATA97 pela segunda vez no terceiro semestre
mata97dis2_s4	Alunos que cursam MATA97 pela segunda vez no quarto semestre
mata97dis2_s5	Alunos que cursam MATA97 pela segunda vez no quinto semestre
retencao_s2	Alunos retidos no segundo semestre
retencao_s3	Alunos retidos no terceiro semestre
retencao_s4	Alunos retidos no quarto semestre
retencao_s5	Alunos retidos no quinto semestre
retencao_s6	Alunos retidos no sexto semestre
retencao_s7	Alunos retidos no sétimo semestre
mata67dis1_s9s10	Alunos que cursam MATA67 no nono ou décimo semestre

Anexo

A

GRADE CURRICULAR 2007.2

Curso: 112140 Currículo: 2007-2 Turno: Diurno Duração em anos: Mínima 4,5 Média 5,5 Máxima 7
Ciência da Computação

Área: Matemática, Ciências Físicas e Tecnologia Titulação: Bacharel em Ciência da Computação

Habilitação: Bacharelado

Base Legal: AUTORIZAÇÃO: RESOLUÇÃO CONSUNI/UFBA Nº 04 DE 22.01.1971. PARECER CFE Nº 417/80, APROVADO EM 09.04.1980.
RECONHECIMENTO: DECRETO Nº 82027 DE 24.07.78. PARECER CFE Nº 1910/78

1º SEMESTRE	Crédito / Semestre 0	Horas / Semana 25	Horas / Semestre 425

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
MATA01 GEOMETRIA ANALÍTICA	68	0	OB		
MATA02 CÁLCULO A	102	0	OB		
MATA37 INTRODUÇÃO À LÓGICA DE PROGRAMAÇÃO	68	0	OB		
MATA38 PROJETO DE CIRCUITOS LÓGICOS	68	0	OB		
MATA39 SEMINÁRIOS DE INTRODUÇÃO AO CURSO	51	0	OB		
MATA42 MATEMÁTICA DISCRETA I	68	0	OB		

2º SEMESTRE	Crédito / Semestre 0	Horas / Semana 25	Horas / Semestre 425

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
FCHC45 METODOLOGIA E EXPRESSÃO TÉCNICO-CIENTÍF	68	0	OB		
MATA07 ÁLGEBRA LINEAR A	68	0	OB	01	MATA01
MATA40 ESTRUTURAS DE DADOS E ALGORITMOS I	68	0	OB	01	MATA37 MATA42
MATA57 LABORATÓRIO DE PROGRAMAÇÃO I	51	0	OB	01	MATA37
MATA95 COMPLEMENTOS DE CÁLCULO	102	0	OB	01	MATA01 MATA02
MATA97 MATEMÁTICA DISCRETA II	68	0	OB	01	MATA42

3º SEMESTRE	Crédito / Semestre 0	Horas / Semana 27	Horas / Semestre 459

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
FISA75 ELEMENTOS DO ELETROMAGNETISMO E DE CIR	102	0	OB	01	MATA95
MATA47 LÓGICA PARA COMPUTAÇÃO	68	0	OB	01	MATA97
MATA50 LINGUAGENS FORMAIS E AUTÔMATOS	68	0	OB	01	MATA42
MATA55 PROGRAMAÇÃO ORIENTADA A OBJETOS	68	0	OB	01	MATA40
MATA96 ESTATÍSTICA A	102	0	OB	01	MATA42 MATA95
OPT051 OPTATIVA 051	51	0	OP		

4º SEMESTRE	Crédito / Semestre 0	Horas / Semana 19	Horas / Semestre 323

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
MATA48 ARQUITETURA DE COMPUTADORES	68	0	OB	01	MATA38
MATA51 TEORIA DA COMPUTAÇÃO	68	0	OB	01	MATA47
MATA52 ANÁLISE E PROJETO DE ALGORITMOS	68	0	OB	01	MATA40
MATA62 ENGENHARIA DE SOFTWARE I	68	0	OB	01	MATA55
MATA68 COMPUTADOR, ÉTICA E SOCIEDADE	51	0	OB		

5º SEMESTRE	Crédito / Semestre 0	Horas / Semana 23	Horas / Semestre 391

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
MATA49 PROGRAMAÇÃO DE SOFTWARE BÁSICO	68	0	OB	01	MATA40 MATA48 MATA57
MATA53 TEORIA DOS GRAFOS	68	0	OB	01	MATA52
MATA54 ESTRUTURAS DE DADOS E ALGORITMOS II	68	0	OB	01	MATA52
MATA56 PARADIGMAS DE LINGUAGENS DE PROGRAMAÇ	68	0	OB	01	MATA55
MATA63 ENGENHARIA DE SOFTWARE II	68	0	OB	01	MATA62
OPT051 OPTATIVA 051	51	0	OP		

6º SEMESTRE	Crédito / Semestre 0	Horas / Semana 23	Horas / Semestre 391

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
MATA58 SISTEMAS OPERACIONAIS	68	0	OB	01	MATA49
MATA59 REDES DE COMPUTADORES I	68	0	OB	01	MATA49
MATA60 BANCO DE DADOS	68	0	OB	01	MATA54
MATA61 COMPILADORES	68	0	OB	01	MATA49 MATA50
MATA64 INTELIGÊNCIA ARTIFICIAL	68	0	OB	01	MATA47 MATA53 MATA56
OPT051 OPTATIVA 051	51	0	OP		

7º SEMESTRE	Crédito / Semestre 0	Horas / Semana 24	Horas / Semestre 408

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
MATA65 COMPUTAÇÃO GRÁFICA	68	0	OB	01	MATA07 MATA57 MATA95
OPT068 OPTATIVA 068	68	0	OP		
OPT068 OPTATIVA 068	68	0	OP		

7º SEMESTRE	Crédito / Semestre 0	Horas / Semana 24	Horas / Semestre 408

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
OPT068 OPTATIVA 068	68	0	OP		
OPT068 OPTATIVA 068	68	0	OP		
OPT068 OPTATIVA 068	68	0	OP		

8º SEMESTRE	Crédito / Semestre 0	Horas / Semana 21	Horas / Semestre 357

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
MATA66 PROJETO FINAL DE CURSO I	51	0	OB	01	FCHC45
OPT051 OPTATIVA 051	51	0	OP		
OPT051 OPTATIVA 051	51	0	OP		
OPT068 OPTATIVA 068	68	0	OP		
OPT068 OPTATIVA 068	68	0	OP		
OPT068 OPTATIVA 068	68	0	OP		

9º SEMESTRE	Crédito / Semestre 0	Horas / Semana 23	Horas / Semestre 391

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
MATA67 PROJETO FINAL DE CURSO II	136	0	OB	01	MATA66
OPT051 OPTATIVA 051	51	0	OP		
OPT051 OPTATIVA 051	51	0	OP		
OPT051 OPTATIVA 051	51	0	OP		
OPT051 OPTATIVA 051	51	0	OP		
OPT051 OPTATIVA 051	51	0	OP		

OPTATIVAS

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
ADM001 INTRODUCAO À ADMINISTRACAO	68	0	OP		
ADM011 PESQ OPERACIONAL	60	4	OP	01	ADM001 MATA96
ADM100 ADMINISTRACAO CONTABIL I	60	4	OP	01	FCC001
ADM171 ELEMENTOS E ANALISES DE CUSTOS	68	0	OP	01	ADM001
ADM241 INTRODUCAO AO MARKETING	51	0	OP		
ADM243 GERÊNCIA CONTEMPORÂNEA	68	0	OP		
ECO001 FUNDAMENTOS DE ECONOMIA	51	0	OP		
EDC001 EDUCAÇÃO ABERTA, CONTINUADA E À DISTÂNC	68	0	OP		
ENG229 APLICAÇÕES INDUSTRIAIS DA COMPUTAÇÃO	68	0	OP		
ENG336 ELETRÔNICA DIGITAL	68	0	OP	01	MATA38
ENG646 AUTOMACAO DE SISTEMAS	51	3	OP	01	MATA48
ENG648 CONTROLE E AUTOMACAO DE PROCESSOS	51	3	OP	01	MATA07 MATA95
FCC001 CONTABILIDADE GERAL I	68	0	OP		
FCH162 PSICOLOGIA DAS RELACOES HUMANAS	68	0	OP		
FISA76 OSCILAÇÕES E ONDAS ELETROMAGNÉTICAS	102	0	OP	01	FISA75
LET358 INGLES INSTRUMENTAL III N-100	51	0	OP		
LET359 INGLES INSTRUMENTAL IV N-100	51	0	OP	01	LET358
LETE46 LIBRAS-LÍNGUA BRASILEIRA DE SINAIS	34	0	OP		
MAT174 CALCULO NUMÉRICO I	68	0	OP	01	MATA07 MATA37 MATA95
MAT220 EMPREENDEDORES EM INFORMATICA	68	0	OP		
MATA04 CÁLCULO C	102	0	OP		
MATA05 CALCULO D	102	0	OP	01	MATA95
MATA41 INFORMÁTICA NA EDUCAÇÃO	68	0	OP		
MATA69 MODELAGEM E SIMULAÇÃO DE SISTEMAS	68	0	OP	01	MATA07 MATA96
MATA71 ANÁLISE NUMÉRICA	68	0	OP	01	MAT174 MATA07 MATA95
MATA72 TÓPICOS EM ARQUITETURA DE COMPUTADORES	51	0	OP	01	MATA48
MATA73 LABORATÓRIO DE CIRCUITOS DIGITAIS	51	0	OP	01	MATA38
MATA74 TÓPICOS EM COMPUTAÇÃO E ALGORITMOS	51	0	OP	01	MATA50 MATA51 MATA52
MATA75 SEMÂNTICA DE LINGUAGEM DE PROGRAMAÇÃO	68	0	OP	01	MATA47 MATA56
MATA76 LINGUAGENS PARA APLICAÇÃO COMERCIAL	51	0	OP	01	MATA40
MATA77 PROGRAMAÇÃO FUNCIONAL	68	0	OP	01	MATA40 MATA97
MATA79 TÓPICOS EM PROGRAMAÇÃO	51	0	OP	01	MATA54 MATA56
MATA80 LABORATÓRIO DE PROGRAMAÇÃO II	51	0	OP	01	MATA52
MATA81 LABORATÓRIO DE SISTEMAS OPERACIONAIS	51	0	OP	01	MATA58
MATA82 SISTEMAS DE TEMPO REAL	68	0	OP	01	MATA58

Anexo

B

GRADE CURRICULAR 2008.1

Curso: 112140 Currículo: 2008-1 Turno: Diurno Duração em anos: Mínima 5 Média 6 Máxima 7
Ciência da Computação
Área: Matemática, Ciências Físicas e Tecnologia Titulação: Bacharel em Ciência da Computação
Habilitação: Bacharelado
Base Legal: AUTORIZAÇÃO: RESOLUÇÃO CONSUNI/UFBA Nº 04 DE 22.01.1971. PARECER CFE Nº 417/80, APROVADO EM 09.04.1980. RECONHECIMENTO: DECRETO Nº 82027 DE 24.07.78. PARECER CFE Nº 1910/78. RENOVAÇÃO DE RECONHECIMENTO: PORTARIA Nº 291 DE 02.02.2011

1º SEMESTRE	Crédito / Semestre 0	Horas / Semana 25	Horas / Semestre 425

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
MATA01 GEOMETRIA ANALÍTICA	68	0	OB		
MATA02 CÁLCULO A	102	0	OB		
MATA37 INTRODUÇÃO À LÓGICA DE PROGRAMAÇÃO	68	0	OB		
MATA38 PROJETO DE CIRCUITOS LÓGICOS	68	0	OB		
MATA39 SEMINÁRIOS DE INTRODUÇÃO AO CURSO	51	0	OB		
MATA42 MATEMÁTICA DISCRETA I	68	0	OB		

2º SEMESTRE	Crédito / Semestre 0	Horas / Semana 25	Horas / Semestre 425

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
FCHC45 METODOLOGIA E EXPRESSÃO TÉCNICO-CIENTÍF	68	0	OB		
MATA07 ÁLGEBRA LINEAR A	68	0	OB	01	MATA01
MATA40 ESTRUTURAS DE DADOS E ALGORITMOS I	68	0	OB	01	MATA37 MATA42
MATA57 LABORATÓRIO DE PROGRAMAÇÃO I	51	0	OB	01	MATA37
MATA95 COMPLEMENTOS DE CÁLCULO	102	0	OB	01	MATA01 MATA02
MATA97 MATEMÁTICA DISCRETA II	68	0	OB	01	MATA42

3º SEMESTRE	Crédito / Semestre 0	Horas / Semana 27	Horas / Semestre 459

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
FISA75 ELEMENTOS DO ELETROMAGNETISMO E DE CIR	102	0	OB	01	MATA95
MATA47 LÓGICA PARA COMPUTAÇÃO	68	0	OB	01	MATA97
MATA50 LINGUAGENS FORMAIS E AUTÔMATOS	68	0	OB	01	MATA42
MATA55 PROGRAMAÇÃO ORIENTADA A OBJETOS	68	0	OB	01	MATA40
MATA96 ESTATÍSTICA A	102	0	OB	01	MATA42 MATA95
OPT051 OPTATIVA 051	51	0	OP		

4º SEMESTRE	Crédito / Semestre 0	Horas / Semana 19	Horas / Semestre 323

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
MATA48 ARQUITETURA DE COMPUTADORES	68	0	OB	01	MATA38
MATA51 TEORIA DA COMPUTAÇÃO	68	0	OB	01	MATA47
MATA52 ANÁLISE E PROJETO DE ALGORITMOS	68	0	OB	01	MATA40
MATA62 ENGENHARIA DE SOFTWARE I	68	0	OB	01	MATA55
MATA68 COMPUTADOR, ÉTICA E SOCIEDADE	51	0	OB		

5º SEMESTRE	Crédito / Semestre 0	Horas / Semana 23	Horas / Semestre 391

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
MATA49 PROGRAMAÇÃO DE SOFTWARE BÁSICO	68	0	OB	01	MATA40 MATA48 MATA57
MATA53 TEORIA DOS GRAFOS	68	0	OB	01	MATA52
MATA54 ESTRUTURAS DE DADOS E ALGORITMOS II	68	0	OB	01	MATA52
MATA56 PARADIGMAS DE LINGUAGENS DE PROGRAMAÇ	68	0	OB	01	MATA55
MATA63 ENGENHARIA DE SOFTWARE II	68	0	OB	01	MATA62
OPT051 OPTATIVA 051	51	0	OP		

6º SEMESTRE	Crédito / Semestre 0	Horas / Semana 23	Horas / Semestre 391

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
MATA58 SISTEMAS OPERACIONAIS	68	0	OB	01	MATA49
MATA59 REDES DE COMPUTADORES I	68	0	OB	01	MATA49
MATA60 BANCO DE DADOS	68	0	OB	01	MATA54
MATA61 COMPILADORES	68	0	OB	01	MATA49 MATA50
MATA64 INTELIGÊNCIA ARTIFICIAL	68	0	OB	01	MATA47 MATA53 MATA56
OPT051 OPTATIVA 051	51	0	OP		

7º SEMESTRE	Crédito / Semestre 0	Horas / Semana 20	Horas / Semestre 340

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
MATA65 COMPUTAÇÃO GRÁFICA	68	0	OB	01	MATA07 MATA57 MATA95
OPT068 OPTATIVA 068	68	0	OP		
OPT068 OPTATIVA 068	68	0	OP		

7º SEMESTRE	Crédito / Semestre 0	Horas / Semana 20	Horas / Semestre 340

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
OPT068 OPTATIVA 068	68	0	OP		
OPT068 OPTATIVA 068	68	0	OP		

8º SEMESTRE	Crédito / Semestre 0	Horas / Semana 20	Horas / Semestre 340

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
OPT068 OPTATIVA 068	68	0	OP		
OPT068 OPTATIVA 068	68	0	OP		
OPT068 OPTATIVA 068	68	0	OP		
OPT068 OPTATIVA 068	68	0	OP		
OPT068 OPTATIVA 068	68	0	OP		

9º SEMESTRE	Crédito / Semestre 0	Horas / Semana 23	Horas / Semestre 391

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
MATA66 PROJETO FINAL DE CURSO I	51	0	OB	01	FCHC45
OPT068 OPTATIVA 068	68	0	OP		
OPT068 OPTATIVA 068	68	0	OP		
OPT068 OPTATIVA 068	68	0	OP		
OPT068 OPTATIVA 068	68	0	OP		
OPT068 OPTATIVA 068	68	0	OP		

10º SEMESTRE	Crédito / Semestre 0	Horas / Semana 16	Horas / Semestre 272

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
MATA67 PROJETO FINAL DE CURSO II	136	0	OB	01	MATA66
OPT068 OPTATIVA 068	68	0	OP		
OPT068 OPTATIVA 068	68	0	OP		

OPTATIVAS

Disciplina	C.H.	CR	Nat.	Gr	Pré Requisito
ADM001 INTRODUCAO À ADMINISTRACAO	68	0	OP		
ADM011 PESQUISA OPERACIONAL	68	0	OP	01	ADM001 MATA96
ADM170 ADMINISTRACAO CONTABIL I	68	0	OP		
ADM171 ELEMENTOS E ANALISES DE CUSTOS	68	0	OP	01	ADM001
ADM241 INTRODUCAO AO MARKETING	51	0	OP		
ADM243 GERÊNCIA CONTEMPORÂNEA	68	0	OP		
ECO001 FUNDAMENTOS DE ECONOMIA	51	0	OP		
EDC001 EDUCAÇÃO ABERTA, CONTINUADA E À DISTÂNC	68	0	OP		
ENG229 APLICAÇÕES INDUSTRIAIS DA COMPUTAÇÃO	68	0	OP		
ENG336 ELETRÔNICA DIGITAL	68	0	OP	01	MATA38
ENG646 AUTOMACAO DE SISTEMAS	51	3	OP	01	MATA48
ENG648 CONTROLE E AUTOMACAO DE PROCESSOS	51	3	OP	01	MATA07 MATA95
FCC001 CONTABILIDADE GERAL I	68	0	OP		
FISA76 OSCILAÇÕES E ONDAS ELETROMAGNÉTICAS	102	0	OP	01	FISA75
IPSA39 PSICOLOGIA DAS RELACOES HUMANAS	68	0	OP		
LET358 INGLES INSTRUMENTAL III N-100	51	0	OP		
LET359 INGLES INSTRUMENTAL IV N-100	51	0	OP	01	LET358
LETE46 LIBRAS-LÍNGUA BRASILEIRA DE SINAIS	34	0	OP		
MAT174 CALCULO NUMÉRICO I	68	0	OP	01	MATA07 MATA37 MATA95
MAT220 EMPREENDEDORES EM INFORMATICA	68	0	OP		
MATA04 CÁLCULO C	102	0	OP		
MATA05 CALCULO D	102	0	OP	01	MATA95
MATA41 INFORMÁTICA NA EDUCAÇÃO	68	0	OP		
MATA69 MODELAGEM E SIMULAÇÃO DE SISTEMAS	68	0	OP	01	MATA07 MATA96
MATA71 ANÁLISE NUMÉRICA	68	0	OP	01	MAT174 MATA07 MATA95
MATA72 TÓPICOS EM ARQUITETURA DE COMPUTADORES	51	0	OP	01	MATA48
MATA73 LABORATÓRIO DE CIRCUITOS DIGITAIS	51	0	OP	01	MATA38
MATA74 TÓPICOS EM COMPUTAÇÃO E ALGORITMOS	51	0	OP	01	MATA50 MATA51 MATA52
MATA75 SEMÂNTICA DE LINGUAGEM DE PROGRAMAÇÃO	68	0	OP	01	MATA47 MATA56
MATA76 LINGUAGENS PARA APLICAÇÃO COMERCIAL	51	0	OP	01	MATA40
MATA77 PROGRAMAÇÃO FUNCIONAL	68	0	OP	01	MATA40 MATA97
MATA79 TÓPICOS EM PROGRAMAÇÃO	51	0	OP	01	MATA54 MATA56

REFERÊNCIAS BIBLIOGRÁFICAS

ARA-SOUZA, A. L. Redes bayesianas: Uma introdução aplicada a credit scoring. *Simpósio Nacional de Probabilidade e Estatística*, 2010.

ASIMOW, L. A.; MAXWELL, M. M. *Probability and statistics with applications: A problem solving text*. [S.l.]: Actex Publications, 2010.

BAKER, R. S. J. de; ISOTANI, S.; CARVALHO, A. M. J. B. de. Mineraçao de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, v. 19, n. 2, 2011.

BAUER, H. *Probability theory, volume 23 of de Gruyter Studies in Mathematics*. [S.l.]: Walter de Gruyter & Co., Berlin, 1996.

BAYES, M.; PRICE, M. An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfrs. *Philosophical Transactions (1683-1775)*, JSTOR, p. 370–418, 1763.

BRANDÃO, M. d. F. R.; RAMOS, C. R. dos S.; TRÓCCOLI, B. T. Análise de agrupamento de escolas e núcleos de tecnologia educacional: mineração na base de dados de avaliação do programa nacional de informática na educação. In: *Anais do Simpósio Brasileiro de Informática na Educação*. [S.l.: s.n.], 2003. v. 1, n. 1, p. 366–374.

BRASIL. *Institui o Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais - REUNI*. ⟨http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2007/decreto/d6096.htm⟩, year=2007, note = Disponível. Acessado em 20 de Fevereiro de 2013.

CAMARINHA, M. M. d. O. *Auditoria na Banca Utilizando Redes Bayesianas*. Dissertação (Mestrado), 2011.

CAMPELLO, A. d. V. C.; LINS, L. N. Metodologia de análise e tratamento da evasão e retenção em cursos de graduação instituições federais de ensino superior. *Anais XXVIII Encontro Nacional de Engenharia de Produção, Rio de Janeiro*, 2008.

CHARNIAK, E. Bayesian networks without tears. *AI magazine*, v. 12, n. 4, p. 50, 1991.

CHENG, J.; GREINER, R. Learning bayesian belief network classifiers: Algorithms and system. In: *Advances in Artificial Intelligence*. [S.l.]: Springer, 2001. p. 141–151.

CISLAGHI, R. et al. Um modelo de sistema de gestão do conhecimento em um framework para a promoção da permanência discente no ensino de graduação. Florianópolis, SC, 2008.

CLARO, D. B. et al. Análise da retenção do alunado da ufba via desempenho acadêmico. 2014. Disponível em: https://repositorio.ufba.br/ri/handle/ri/15760. Acessado em 15 de Setembro de 2014.

CORRÊA, A. C.; NORONHA, A. B.; MIURA, I. K. Avaliação da evasão e permanência prolongada em um curso de graduação em administração de uma universidade pública. *Anais do Semead-Seminários de Administração, São Paulo, SP, Brasil*, v. 7, 2004.

DEGROOT, M. *Probability and statistics*. [S.l.]: Addison-Wesley Pub. Co., 1975. (Addison-Wesley series in behavioral science).

DIAS, A. F. M.; CERQUEIRA, G. S.; LINS, L. N. Fatores determinantes da retenção estudantil em um curso de graduação em engenharia de produção. In: *Congresso Brasileiro de Educação em Engenharia*. [S.l.: s.n.], 2009. v. 37.

DIAS, A. F. M.; LINS, L. N. Cadeias de markov para análise da evasão/retenção em cursos de graduação. *XVIII CONIC*, 2010.

DRUZDZEL, M. J. Smile: Structural modeling, inference, and learning engine and genie: a development environment for graphical decision-theoretic models. In: *AAAI/IAAI*. [S.l.: s.n.], 1999. p. 902–903.

EVASÃO, C. E. de Estudo de. *Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas*. 1997. ⟨http://www.udesc.br/arquivos/id_submenu/102/diplomacao.pdf⟩, note = Acessado em 20 de Fevereiro de 2013.

FILHO, N. d. et al. *Memorial da Universidade Nova*. [S.l.], 2010.

GAAG, L. C. van der; HELSPER, E. M. Experiences with modelling issues in building probabilistic networks. In: *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*. [S.l.]: Springer, 2002. p. 21–26.

GENEVOIS, B. B.; LYRA, P. R.; LIMA, E. S. de. Evasão e retenção nos cursos do centro de tecnologia da universidade federal de pernambuco. *Congresso Brasileiro de Educação de Engenharia*, 2008.

GOMES, A. V. P.; WANKE, P. Modelagem da gestão de estoques de peças de reposição através de cadeias de markov. *Gestão & Produção*, SciELO Brasil, v. 15, n. 1, p. 57–72, 2008.

HAN, J.; KAMBER, M.; PEI, J. *Data mining: concepts and techniques: concepts and techniques*. [S.l.]: Elsevier, 2011.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, v. 143, n. 1, p. 29–36, 1982.

HECKERMAN, D. *A tutorial on learning with Bayesian networks*. [S.l.]: Springer, 1998.

JR, R. H. Propagação de evidências em redes bayesianas: diagnóstico sobre doenças pulmonares. *Brasília: CIC/UNB*, 1997.

KARCHER, C. *Redes Bayesianas aplicadas à análise do risco de crédito*. Tese (Doutorado) — Universidade de São Paulo, 2009.

KLIR, G. J.; FOLGER, T. A. Fuzzy sets, uncertainty, and information. Prentice Hall, 1988.

KOLMOGOROV, A. N. Foundations of the theory of probability. Chelsea Publishing Co., 1950.

KORB, K. B.; NICHOLSON, A. E. *Bayesian artificial intelligence*. [S.l.]: CRC press, 2010.

LAPLACE, P. S. marquis de. *Théorie analytique des probabilités*. [S.l.]: Mme Ve Courcier, 1812.

LAURITZEN, S. L.; SPIEGELHALTER, D. J. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 157–224, 1988.

LAUTERT, L. V.; ROLIM, M.; LODER, L. L. Investigando processos de retenção no âmbito de um curso de engenharia elétrica. In: *Anais do Congresso Brasileiro de Educação em Engenharia*. [S.l.: s.n.], 2011.

LENNING, O. et al. Retention and attrition: Evidence for action and research. ERIC, 1980.

LUCAS, P. J.; GAAG, L. C. van der; ABU-HANNA, A. Bayesian networks in biomedicine and health-care. *Artificial intelligence in medicine*, Elsevier, v. 30, n. 3, p. 201–214, 2004.

MADSEN, A. L. et al. The hugin tool for probabilistic graphical models. *International Journal on Artificial Intelligence Tools*, World Scientific, v. 14, n. 03, p. 507–543, 2005.

MANHÃES, L. M. et al. Identificação dos fatores que influenciam a evasão em cursos de graduação através de sistemas baseados em mineração de dados: Uma abordagem quantitativa. *VIII Simpósio Brasileiro de Sistemas de Informação (SBSI 2012)*, 2012.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. *Machine learning: An artificial intelligence approach*. [S.l.]: Springer Science & Business Media, 2013.

NANDESHWAR, A.; MENZIES, T.; NELSON, A. Learning patterns of university student retention. *Expert Systems with Applications*, Elsevier, v. 38, n. 12, p. 14984–14996, 2011.

NASSAR, S. M. Tratamento de incerteza: Sistemas especialistas probabilísticos. v. 1, 2003. Disponível em: ¡http://www.inf.ufsc.br/ silvia/disciplinas/sep/MaterialDidatico.pdf¿. Acessado em 20 de agosto de 2014.

NEY, O. A. D. S. *Sistemas de informação acadêmica para o controle da evasão*. Dissertação (Mestrado) — Universidade Federal da Paraíba, 2010.

NORONHA, B.; CARVALHO, B. M.; SANTOS, F. F. F. Perfil dos alunos evadidos da faculdade de economia, administração e contabilidade campus ribeirão preto e avaliação do tempo de titulação dos alunos atualmente matriculados. *Documento de Trabalho. NUPES–Núcleo de Pesquisa sobre Ensino Superior, Universidade de São Paulo, São Paulo, SP, Brasil*, 2001.

PEARL, J. Probabilistic reasoning in intelligent systems: Networks of plausible inference (1992) morgan kaufmann. *San Mateo, CA, USA*, 1988.

PEREIRA, A. S. *Retenção Discente nos Cursos de Graduação Presencial da Universidade Federal do Espírito Santo*. Dissertação (Mestrado) — Universidade Federal do Espírito Santos, 2013.

PITTMAN, K. *Comparison of data mining techniques used to predict student retention*. [S.l.]: ProQuest, 2008.

POLYDORO, S. A. J. O trancamento de matrícula na trajetória acadêmica do universitário: condições de saída e de retorno à instituição. Biblioteca Digital da Unicamp, 2000.

PRATI, R.; BATISTA, G.; MONARD, M. Curvas roc para avaliação de classificadores. *Revista IEEE América Latina*, v. 6, n. 2, p. 215–222, 2008.

RENOOIJ, S.; GAAG, L. C. van der. From qualitative to quantitative probabilistic networks. In: *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*. [S.l.: s.n.], 2002. p. 422–429.

RIOS, J. R. T.; SANTOS, A. P. dos; LIMA, L. B. de. Evasão e retenção na escola de minas da ufop: a perspectiva dos colegiados de cursos. *Congresso Brasileiro de Educação de Engenharia*, 2003.

RISSI, M.; MARCONDES, M. Estudo sobre a reprovação e retenção nos cursos de graduação: 2009. *UEL. Londrina, PR, Brasil*, 2011.

ROHATGI, V. K.; SALEH, A. M. E. *An introduction to probability and statistics*. [S.l.]: John Wiley & Sons, 2011.

RUSSELL, S.; NORVIG, P. Artificial intelligence: a modern approach. 1995.

SANTOS, A. P. D. Diagnóstico do fluxo de estudantes nos cursos de graduação da ufop. retenção, diplomação e evasão. *Avaliação*, Unicamp, v. 4, n. 4, p. 55–66, 1999.

SANTOS, A. P. dos; NASCIMENTO, C.; RIOS, J. R. T. Estudo da evasão e da retenção nos cursos de engenharia da escola de minas da universidade federal de ouro preto. 2000.

SANTOS, M. S. et al. Mining retention rules from student transcripts: A case study of the programs at a federal university. In: *Anais do Simpósio Brasileiro de Informática na Educação*. [S.l.: s.n.], 2014. v. 25, n. 1, p. 762–771.

SEIDMAN, A. *College student retention: Formula for student success*. [S.l.]: Greenwood Publishing Group, 2005.

SILVA, C. V. et al. Mining retention rules from student transcripts: A case study of the information systems programme at a federal university. In: *Anais do Simpósio Brasileiro de Informática na Educação*. [S.l.: s.n.], 2013. v. 24, n. 1.

SILVA, H. R. *Uma abordagem de mineração de dados para a prevenção da evasão/retenção na UFPE*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2013. Não disponível eletronicamente.

SILVA, M.; CLARO, D. B.; LIMA, V. A probabilistic analysis of student retention in a federal university: A case study of a computer science program. In: *Anais do Simpósio Brasileiro de Informática na Educação*. [S.l.: s.n.], 2015. v. 26, n. 1, p. 1245.

SOARES, I. S.; FUNDÃO, C. U.-I. do. Evasão, retenção e orientação acadêmica: Ufrj–engenharia de produção–estudo de caso. In: *Anais do XXXIV Congresso Brasileiro de Ensino de Engenharia-COBENGE. Passo Fundo: Ed. Universidade de Passo Fundo*. [S.l.: s.n.], 2006.

SOBERANIS, I. E. D. An extended bayesian network approach for analyzing supply chain disruptions. University of Iowa, 2010.

VASCONCELOS, A. L. F. de S.; SILVA, M. N. D. Uma investigação sobre os fatores contribuintes na retenção dos alunos no curso de ciências contábeis em uma ifes: um desafio à gestão universitária. *Registro Contábil*, v. 2, n. 3, p. 21–34, 2012.

VIEIRA, E. T. Índices de retenção na universidade de brasília: abordagem do ponto de vista do financiamento. 2014.

YADAV, S. K.; BHARADWAJ, B.; PAL, S. Mining education data to predict student's retention: A comparative study. *arXiv preprint arXiv:1203.2987*, 2012.

YU, C. H. et al. A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science*, v. 8, n. 2010, p. 307–325, 2010.

ZHANG, Y. et al. Use data mining to improve student retention in higher education - a case study. In: *ICEIS (1)*. [S.l.: s.n.], 2010. p. 190–197.