

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ  
DEPARTAMENTO ACADÊMICO DE INFORMÁTICA  
TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

JEAN CARLOS RODRIGUES  
RAFAEL FERNANDES SIQUEIRA

**MINERAÇÃO DE DADOS DO DESEMPENHO ACADÊMICO NA  
EDUCAÇÃO A DISTÂNCIA**

TRABALHO DE CONCLUSÃO DE CURSO

PONTA GROSSA  
2015

**JEAN CARLOS RODRIGUES  
RAFAEL FERNANDES SIQUEIRA**

**MINERAÇÃO DE DADOS DO DESEMPENHO ACADÊMICO NA  
EDUCAÇÃO A DISTÂNCIA**

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Tecnólogo em Análise e Desenvolvimento de Sistemas, do Departamento Acadêmico de Informática da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Esp. Marcos Vinicius Fidelis

**PONTA GROSSA  
2015**



Ministério da Educação  
**Universidade Tecnológica Federal do Paraná**  
Campus Ponta Grossa  
Diretoria de Graduação e Educação Profissional  
Departamento Acadêmico de Informática  
Tecnologia em Análise e Desenvolvimento de Sistemas



---

## **TERMO DE APROVAÇÃO**

Mineração de Dados do Desempenho Acadêmico na Educação a Distância

por

**JEAN CARLOS RODRIGUES**  
**RAFAEL FERNANDES SIQUEIRA**

Este Trabalho de Conclusão de Curso (TCC) foi apresentado em 02 de junho de 2015 como requisito parcial para a obtenção do título de Tecnólogo em Análise de Desenvolvimento de Sistemas. Os candidatos foram arguidos pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

---

Marcos Vinicius Fidelis  
Prof.(a) Orientador(a)

---

Helyane Bronoski Borges  
Membro titular

---

Geraldo Ranthum  
Membro titular

- O Termo de Aprovação assinado encontra-se na Coordenação do Curso -

Dedico este trabalho à minha família pelo apoio e confiança. À meu colega Rafael, pela paciência e parceria ao longo desse trabalho. Ao nosso orientador Prof. Marcos Vinicius Fidelis, pelos valorosos aconselhamentos e orientações. Em especial à minha namorada, Solange Maria César, pelo apoio e compreensão em todos os momentos.

*Jean Carlos Rodrigues*

Dedico este trabalho à minha família. Para nosso Orientador Prof. Marcos Vinicius Fidelis, pela sua dedicação, atenção e incentivo. À minha namorada, Dilmara Rodrigues, pela sua compreensão e apoio integral nos momentos em que estive ausente, você foi e sempre será minha inspiração. É para vocês que dedico essa conquista.

*Rafael Fernandes Siqueira*

## **AGRADECIMENTOS**

Primeiramente agradecemos a Deus que permitiu que tudo isso acontecesse, ao longo de nossas vidas, e não somente nesses anos como universitários, por renovar a cada momento a nossa força e disposição e pelo discernimento concedido ao longo dessa jornada.

Agradecemos ao nosso orientador Prof. Marcos Vinicius Fidelis, pela sabedoria e dedicação com que nos guiou nesta trajetória.

À todos os colegas e professores da UTFPR, que estiveram junto conosco em algum momento no curso.

À equipe e a Coordenação do NUTEAD, pelo apoio e por nos fornecer os dados para a realização do estudo.

Gostaríamos também de deixar registrado nosso reconhecimento à nossas famílias, pois o apoio e compreensão foi de vital importância para vencermos esse desafio.

Enfim, a todos os que por algum motivo contribuíram para a realização desta pesquisa.

RODRIGUES, Jean Carlos; SIQUEIRA, Rafael Fernandes. **Mineração de Dados do Desempenho Acadêmico na Educação a Distância**. 2015. 89 p. Trabalho de Conclusão de Curso de Tecnologia em Análise e Desenvolvimento de Sistemas - Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2015.

## RESUMO

A análise da eficiência dos recursos empregados no ensino a distância ainda é um desafio para as instituições de educação no Brasil. O objetivo deste trabalho é realizar um estudo sobre a aplicação de técnicas de Mineração de Dados em uma base de dados real do Ambiente Virtual de Aprendizagem utilizada no Núcleo de Tecnologia e Educação Aberta e a Distância da Universidade Estadual de Ponta Grossa. Para tanto, é descrita a aplicação do processo de Descoberta de Conhecimento em Banco de Dados nesta base, utilizando dados de três disciplinas do curso Bacharelado em Administração Pública, realizado através do programa Universidade Aberta do Brasil. Será realizada a análise das árvores de decisão obtidas com a aplicação do algoritmo de classificação J48 e comparada a eficiência desse algoritmo com outros algoritmos de classificação.

**Palavras-chaves:** Mineração de Dados. DCBD. Ambiente Virtual de Aprendizagem.

RODRIGUES, Jean Carlos; SIQUEIRA, Rafael Fernandes. **Academic Performance's Data Mining in Distance Education** . 2015. 89 p. Trabalho de Conclusão de Curso de Tecnologia em Análise e Desenvolvimento de Sistemas - Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2015.

#### **ABSTRACT**

The efficiency analysis of the resources used in online education still a challenge for education institutions in Brasil. The main objective of this paper is to study the application of Data Mining techniques on the Virtual Learning Environment's database used by Núcleo de Tecnologia e Educação Aberta e a Distância of the Universidade Estadual de Ponta Grossa. To accomplish the objectives we describe the application of the Knowledge Discover in Databases process on three disciplines's data of the Bacharelado em Administração Pública course offered through Universidade Aberta do Brasil program. The decision trees generated by the J48 classification algorithm will be discuss, and his efficiency will be compared to other classification algorithms.

**Keywords:** Data Mining. KDD. Virtual Learning Environment.

## LISTA DE FIGURAS

Figura 1	–	Estrutura e funcionamento da UAB .....	17
Figura 2	–	Sistema de Apoio a Decisão .....	22
Figura 3	–	Etapas do processo de DCBD .....	25
Figura 4	–	Principais áreas relacionadas com a EDM .....	31
Figura 5	–	Experimenter .....	40
Figura 6	–	Arquivo iris.arff .....	41
Figura 7	–	Distribuição de alunos por Polo .....	48
Figura 8	–	Resultado do algoritmo <i>Ranker</i> nos dados da disciplina de RI .....	50
Figura 9	–	Resultado do algoritmo <i>Ranker</i> nos dados da disciplina de GA .....	52
Figura 10	–	Resultado do <i>Ranker</i> utilizando <i>GainRatioAttributeEval</i> .....	53
Figura 11	–	Avaliação do algoritmo <i>Ranker</i> para a base de GQ .....	54
Figura 12	–	Árvores de decisão geradas pelo algoritmo J48 .....	58
Figura 13	–	Árvore de decisão para a disciplina GA .....	59
Figura 14	–	Árvore de decisão para a base GQ-exame .....	61
Figura 15	–	Árvore de decisão para a base RI-nota .....	61
Figura 16	–	Árvores de decisão das bases GA-nota e GQ-nota .....	62



## LISTA DE TABELAS

Tabela 1	–	Valores dos atributos da base Gestão Ambiental .....	53
Tabela 2	–	Atributos da disciplina de GQ selecionados pelo algoritmo Ranker .....	55
Tabela 3	–	Comparação entre os atributos das disciplinas .....	55
Tabela 4	–	Resultados da primeira execução do algoritmo J48 .....	57
Tabela 5	–	Resultados do algoritmo J48 com diferentes parâmetros.....	59
Tabela 6	–	Porcentagem de acerto dos algoritmos .....	60
Tabela 7	–	Comparação do Kappa Statistic da classificação das bases .....	60

## LISTA DE QUADROS

Quadro 1	– Educação a Distância na UEPG .....	16
Quadro 2	– Áreas de conhecimento encontradas em DCBD e sua utilização .....	24
Quadro 3	– Tabela de concordância dos valores do Kappa Statistic .....	37
Quadro 4	– Tabelas do banco de dados selecionadas para estudo.....	43
Quadro 5	– Categorias de cursos criadas dentro do AVA .....	44
Quadro 6	– Subcategorias de cursos cadastradas no AVA.....	44
Quadro 7	– Quadro com os cursos e o total de logs de acesso por curso .....	45
Quadro 8	– Disciplinas do curso de Bacharelado em Administração Pública.....	46
Quadro 9	– Ações realizadas pelos usuários .....	47
Quadro 10	– Resultados do algoritmo <i>Ranker</i> agrupados pelo tipo da ação .....	50
Quadro 11	– Valores dos atributos relacionados a realização dos questionários .....	51
Quadro 12	– Valores dos atributos relacionados a realização dos trabalhos.....	52

## SUMÁRIO

<b>1 INTRODUÇÃO</b> .....	12
1.1 OBJETIVOS .....	13
1.2 JUSTIFICATIVA .....	13
1.3 ESTRUTURA DO TRABALHO .....	14
<b>2 REFERENCIAL TEÓRICO</b> .....	15
2.1 EDUCAÇÃO A DISTÂNCIA .....	15
2.1.1 Educação a Distância na Universidade Estadual de Ponta Grossa .....	16
2.1.2 Universidade Aberta do Brasil .....	17
2.1.3 Programa Nacional de Formação em Administração Pública .....	18
2.2 AMBIENTE VIRTUAL DE APRENDIZAGEM .....	18
2.2.1 Sistema de Gerenciamentos de Curso .....	19
2.2.1.1 Moodle .....	20
2.2.1.2 Outros SGC .....	20
2.3 SISTEMAS DE APOIO A DECISÃO .....	21
2.3.1 Sistemas de Apoio a Decisão em Educação a Distância .....	23
2.4 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS .....	24
2.4.1 Seleção .....	25
2.4.2 Pré-processamento .....	25
2.4.3 Transformação dos Dados .....	26
2.4.4 Mineração de Dados .....	26
2.4.5 Interpretação e Avaliação .....	26
2.5 MINERAÇÃO DE DADOS .....	27
2.5.1 Tarefas da Mineração de Dados .....	27
2.5.1.1 Tarefas Descritivas .....	27
2.5.1.2 Tarefas de Amostragem .....	28
2.5.1.3 Tarefas de Prognóstico .....	28
2.5.2 Técnicas da Mineração de dados .....	28
2.5.2.1 Descoberta de Regras de Associação .....	29
2.5.2.2 Árvores de Decisão .....	29
2.5.2.3 Raciocínio Baseado em Casos .....	29
2.5.2.4 Algoritmos Genéticos .....	30
2.5.2.5 Redes Neurais Artificiais .....	30
2.5.3 Mineração de Dados Educacionais .....	30
2.6 TRABALHOS CORRELATOS .....	31
<b>3 METODOLOGIA</b> .....	34
3.1 APLICAÇÃO DO PROCESSO DE DCBD .....	34

3.1.1 Seleção .....	34
3.1.2 Pré-Processamento .....	35
3.1.3 Transformação .....	35
3.1.4 Mineração de Dados .....	35
3.1.5 Interpretação e Avaliação .....	36
3.2 FERRAMENTAS UTILIZADAS .....	37
3.2.1 PostgreSQL .....	37
3.2.1.1 Arquivo <i>dump</i> .....	37
3.2.2 EXCEL .....	38
3.2.3 Weka .....	38
3.2.3.1 Algoritmos .....	39
3.2.3.2 Experimenter .....	39
3.2.3.3 Arquivo ARFF .....	41
<b>4 RESULTADOS .....</b>	<b>43</b>
4.1 SELEÇÃO .....	43
4.2 PRÉ-PROCESSAMENTO .....	46
4.2.1 Limpeza .....	46
4.2.2 Seleção de atributos .....	48
4.2.2.1 Seleção de atributos da disciplina de Relações Internacionais .....	49
4.2.2.2 Seleção de atributos da disciplina de Gestão Ambiental .....	52
4.2.2.3 Seleção de atributos da disciplina de Gestão da Qualidade .....	54
4.3 TRANSFORMAÇÃO .....	56
4.4 MINERAÇÃO DOS DADOS .....	56
4.4.1 Primeiro experimento .....	57
4.4.2 Segundo experimento .....	60
4.5 AVALIAÇÃO DOS RESULTADOS .....	62
<b>5 CONCLUSÃO .....</b>	<b>65</b>
<b>APÊNDICE A – SQL UTILIZADAS PARA EXTRAÇÃO DE DADOS .....</b>	<b>70</b>
<b>APÊNDICE B – DESCRIÇÃO DOS DADOS .....</b>	<b>75</b>
<b>APÊNDICE C – TERMO DE CESSÃO DE DADOS DO AVA .....</b>	<b>88</b>

## 1 INTRODUÇÃO

A Educação à Distância (EAD) é um recurso importante a ser explorado pelas instituições de ensino no país, pois aumenta as oportunidades para que muitas pessoas tenham acesso à educação de nível superior. Essa modalidade de ensino, permite a separação temporal/espacial entre o aluno e o professor. Portanto, prover informações que auxiliem a condução das atividades inerentes ao sistema educacional, na EAD, é um grande diferencial para professores, tutores e instituições de ensino.

Para que a interação entre aluno e professor aconteça, utiliza-se a internet e uma plataforma de suporte. No Brasil, estas plataformas são chamadas de Ambientes Virtuais de Aprendizagem (AVA). Segundo Peters (2003), um AVA é um sistema informatizado ou *software* formado pela integração de diversos recursos que facilitam a interação professor-aprendiz. Nesse sentido, existem diversos *softwares* disponíveis no mercado de forma gratuita ou não, que permitem que todas as interações entre professores e alunos aconteçam nesse ambiente e por isso são fundamentais nessa modalidade de ensino.

Sabendo que no cenário atual, o AVA é a principal ferramenta utilizada na EAD, onde se dá a comunicação, aprendizagem e avaliação do aluno, procura-se neste trabalho encontrar um padrão de comportamento e possíveis relações existentes entre o desempenho do acadêmico ao final de um período letivo e a sua participação no curso por meio da interação e utilização do AVA.

Utilizando o processo de Descoberta de Conhecimento em Bases de Dados (DCBD), pode-se encontrar informações de interesse que ainda não foram observadas, estas sendo de difícil detecção por métodos tradicionais de análise, pois estes em geral tratam apenas de informações explícitas, sendo assim, as informações encontradas pela aplicação do processo de DCBD são potencialmente úteis para a auxiliar a tomada de decisão.

Frawley, Shapiro e Matheus (1992) definem a DCBD como a extração não trivial de informações implícitas, previamente desconhecidas e potencialmente úteis dos dados. Esse processo geralmente consiste em numa série de passos que englobam a transformação dos dados brutos, o pré-processamento, a Mineração de Dados (*Data Mining*) e o pós-processamento dos resultados obtidos.

Tan, Steinbach e Kumar (2009), explicam que a Mineração de Dados(MD) é uma parte do processo da descoberta de conhecimento, e é utilizada em grandes depósitos de dados, com o intuito de descobrir padrões úteis que poderiam, de outra forma, permanecer ignorados. Dessa maneira, utilizando o processo de DCBD na base de dados do ambiente virtual, pode-se encontrar informações que auxiliem o processo de tomada de decisão, através das correlações que podem ser encontradas entre o uso do ambiente virtual e o desempenho do aluno.

No entanto, a principal preocupação é que os resultados obtidos com esse processo sejam úteis e compreensíveis, pois apesar das diversas ferramentas disponíveis para a análise desses dados, os resultados ainda necessitam de análise humana, para que possam auxiliar no processo de tomada de decisão.

## 1.1 OBJETIVOS

O objetivo geral do trabalho é identificar possíveis relações existentes entre o desempenho dos acadêmicos de um Curso de Ensino a Distância da Universidade Estadual de Ponta Grossa (UEPG) e a utilização do AVA, disponibilizado para os alunos pelo Núcleo de Tecnologia de Educação Aberta e a Distância (NUTEAD).

Os objetivos específicos do trabalho são:

- Elaborar o referencial teórico para desenvolvimento do trabalho.
- Aplicar as etapas do processo de DCDB nos dados do AVA utilizado pelo NUTEAD.
- Avaliar as relações descobertas entre o desempenho acadêmico e a utilização do ambiente virtual.

## 1.2 JUSTIFICATIVA

A EAD no ensino superior, por ser um método ainda em difusão no Brasil, possui campo de estudo ainda pouco explorado. Dessa maneira, fornecer informações relevantes sobre a eficácia das tecnologias empregadas para o aprendizado do aluno em cursos a distância, é um grande desafio, que pode ajudar a manter o gerenciamento das atividades inerentes ao sistema educacional e consequentemente garantir a expansão e qualidade dessa modalidade de ensino.

Para Moran (2009) a utilização da modalidade no país, em instituições de nível superior, governamentais e privadas, encontra-se numa fase de mudanças rápidas e intensa consolidação pedagógica. Ganhou forte apoio do governo federal brasileiro com a criação da Universidade Aberta do Brasil (UAB) e de uma política reguladora consolidada pelo Ministério da Educação e Cultura (MEC), com decretos e portarias que definem o que é válido ou não na EAD, objetivando o ensino e a aprendizagem de qualidade.

Assim a EAD torna-se referência para uma grande mudança no ensino superior. Segundo Iaralham (2009), a EAD é um processo educativo, sistemático e organizado que exige uma comunicação em via dupla e que resulta na interação dos meios tecnológicos de informação e comunicação na aprendizagem. Portanto, a utilização do AVA, deve garantir o *feedback* necessário para que as instituições se aproximem e atendam os anseios de seus alunos e as necessidades da sociedade.

Nesse sentido, este trabalho pretende aplicar os conceitos da DCDB no conjunto de informações disponíveis da base de um AVA, utilizando ferramentas de mineração de dados, com a possibilidade de descobrir padrões e relações, identificar possíveis tendências de comportamento dos alunos, que podem responder a questão da existência de uma relação entre a utilização do AVA pelo aluno e o seu desempenho.

### 1.3 ESTRUTURA DO TRABALHO

Este documento está dividido em 5 capítulos. No primeiro capítulo apresenta-se a contextualização do tema, a proposta de avaliar os dados de um AVA através da DCBD, o objetivo geral e os objetivos específicos e por fim os benefícios que podem ser alcançados com a realização deste trabalho.

No segundo capítulo é apresentado o referencial teórico. Este capítulo mostra o contexto e a popularização da Educação a Distância, apresenta o conceito de Ambientes Virtuais de Aprendizagem, uma revisão sobre Sistemas de Apoio a Decisão, a contextualização da Descoberta de Conhecimento em Bases de Dados, assim como da Mineração de Dados e por fim suas etapas que representam a base para a metodologia deste trabalho.

O terceiro capítulo descreve a metodologia utilizada para a realização do trabalho. Este capítulo apresenta os métodos utilizados em cada uma das etapas do processo de DCBD e também as ferramentas necessárias para se atingir os objetivos propostos para este trabalho.

O quarto capítulo apresenta os resultados obtidos e avalia as possibilidades de análise das relações encontradas.

Por fim, o quinto capítulo contém as considerações finais e os trabalhos futuros.

## 2 REFERENCIAL TEÓRICO

Neste capítulo é descrito o referencial teórico utilizado neste trabalho. A seção 2.1 apresenta o conceito de Educação a Distância. Já à seção 2.2 é feita a revisão de literatura sobre os Sistemas de Gerenciamento de Cursos e é apresentado o ambiente virtual de aprendizagem que será utilizado nesse trabalho. Na seção seguinte, 2.3, apresenta-se o conceito de Sistemas de Apoio a Decisão que está relacionada com a proposta desse trabalho. À seção 2.4 mostra o conceito do processo de Descoberta de Conhecimento em base de dados. E finalmente, a seção 2.5 apresenta o processo de Mineração de dados e suas principais tarefas.

### 2.1 EDUCAÇÃO A DISTÂNCIA

A EAD é uma modalidade de ensino onde alunos e professores encontram-se em tempo e espaço distintos e permite estudos independentes, onde o controle do aprendizado é realizado mais intensamente pelo aluno, assim podendo-se definir que “Educação a distância é o processo de ensino-aprendizagem, mediado por tecnologias, onde professores e alunos estão separados espacial e/ou temporalmente” (MORAN, 2009, p. 1).

Nesse sentido a EAD utiliza principalmente tecnologias de transmissão de dados das telecomunicações e de recursos da informática, para que a conexão entre professor e aluno que estão separados espacialmente e/ou temporalmente aconteça.

O MEC, por meio do Decreto 5622, de 19 de dezembro de 2005, art. 1º caracteriza EAD como:

[...] modalidade educacional na qual a mediação didático-pedagógica nos processos de ensino e aprendizagem ocorre com a utilização de meios e tecnologias de informação e comunicação, com estudantes e professores desenvolvendo atividades educativas em lugares ou tempos diversos.(MEC, 2007, p. 5).

No Brasil, a EAD é praticada desde o início do século XX e tinha como principal objetivo a universalização do acesso à educação, e sua expansão pode ser dividida em 3 etapas. A primeira etapa caracteriza-se pelos cursos por correspondência. A segunda etapa deu enfoque para o uso de programas radiofônicos e televisivos (década de 70), após com áudios e vídeos (década de 80), como por exemplo, o telecurso. E enfim a terceira etapa iniciou-se com a transmissão via satélite (década de 90) e atualmente a tecnologia está totalmente integrada e caracteriza pelo uso dos mais diversos recursos tecnológicos tendo como principais o computador e a internet.(RODRIGUES; SCHIDMIT, 2010, p.7)

De acordo com o Censo da Educação Superior de 2013, divulgado pelo MEC juntamente com o INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), os cursos de graduação a distância registraram crescimento de 15,7% entre 2012 e 2013. Nos últimos 11 anos, a oferta de cursos na modalidade cresceu 24 vezes. No ano de 2013 foram registrados o total de 1.258 cursos de ensino a distância, destes 240 eram bacharelados, 592 licenciaturas e 426 tecnológicos. O volume de matrículas também aumentou e os cursos de li-



cenciatura registraram 451.193 inscrições (39%); cursos de bacharelado e tecnológicos somaram 361.202 (31%) e 341.177 (30%), respectivamente, totalizando 1.153.572 matrículas no ensino superior na modalidade a distância (INEP, 2014).

A EAD portanto, é modalidade de ensino que mais cresce devido a grandes diferenciais oferecidos, sendo um dos maiores atrativos a flexibilidade de horários, onde o aluno não fica preso a horários estipulados pela instituição e pode estudar dentro da sua possibilidade. A Evolução da tecnológica e seu barateamento, juntamente com o apoio do governo federal, possibilitaram impulsionar o crescimento desta modalidade de ensino.

### 2.1.1 Educação a Distância na Universidade Estadual de Ponta Grossa

A UEPG iniciou a sua trajetória na EAD no ano de 2000, ofertando o Curso Normal Superior com Mídias Interativas, um programa especial de formação em nível superior para atender os profissionais da rede pública de ensino que teriam dificuldades em frequentar cursos presenciais. Contou com o apoio do Governo do Paraná, e graduou mais de 3.300 professores num período de 5 (cinco) anos utilizando-se de mídias interativas.

Em 2002, a UEPG criou o NUTEAD, uma estrutura administrativa e pedagógica destinada a incentivar e apoiar o desenvolvimento de cursos e programas de educação a distância na Instituição.

Em 2004, a UEPG foi credenciada junto ao MEC, para ministrar cursos sequenciais, de extensão, de graduação e pós-graduação *lato sensu* na modalidade de educação a distância.

Atualmente, a UEPG/NUTEAD oferta, orienta, desenvolve e avalia cursos e programas em Educação a Distância de relevância nacional, desenvolvidos em parceria com o Ministério da Educação.

No Quadro 1, é apresentada uma síntese do histórico da EAD na UEPG desde a criação do NUTEAD.

2000	Criação, pela UEPG, do primeiro curso de educação a distância – Curso Normal Superior com Mídias Interativas.
2002	é criado, na UEPG, o Núcleo de Tecnologia e Educação Aberta e a Distância- NUTEAD.
2004	o MEC credencia UEPG para ministrar cursos superiores de EaD – de extensão, seqüenciais, graduação e pós – graduação <i>lato sensu</i> (especialização). Nessa ocasião, além do Normal Superior, a UEPG passa a ofertar outros cursos de especialização e seqüenciais na modalidade a distância.
2004	a UEPG passa a integrar a Rede Nacional De Formação Continuada de Professores das Redes Públicas de Ensino e cria o CEFORTEC um dos centros de Alfabetização e Linguagem dessa Rede.
2008	iniciam-se na UEPG os cursos do Programa Pró- Licenciatura - MEC/FNDE.

**Quadro 1 – Educação a Distância na UEPG**

Fonte: Rodrigues e Schidmit (2010)

### 2.1.2 Universidade Aberta do Brasil

O Sistema UAB, é um programa do MEC, criado em 2005, que tem como base o aprimoramento da EAD, seu objetivo é estimular a articulação e integração de um sistema nacional de educação superior através das instituições já existentes, dessa maneira possibilita levar ensino superior público aos mais diversos municípios do Brasil atingindo assim um maior número de cidadãos. Para isso, o sistema tem como base, fortes parcerias entre as esferas federais, estaduais e municipais do governo (MIRANDA, 2008).

O programa disponibiliza cursos de licenciatura, bacharelado e formação continuada. Para realizar a oferta de cursos a distância, cada município deve montar um pólo presencial, com laboratórios de informática, ambientes administrativos, bibliotecas, ambientes acadêmicos, condições de acessibilidade. Nessa infra-estrutura, que inclui ainda o apoio de tutores, fica à disposição dos alunos. Já a elaboração dos cursos é de responsabilidade das instituições públicas de ensino superior de todo país, que desenvolvem material didático e pedagógico (MEC, 2007)

Na Figura 1 são exemplificadas as interações que acontecem entre instituições e polos que oferecem cursos da UAB.

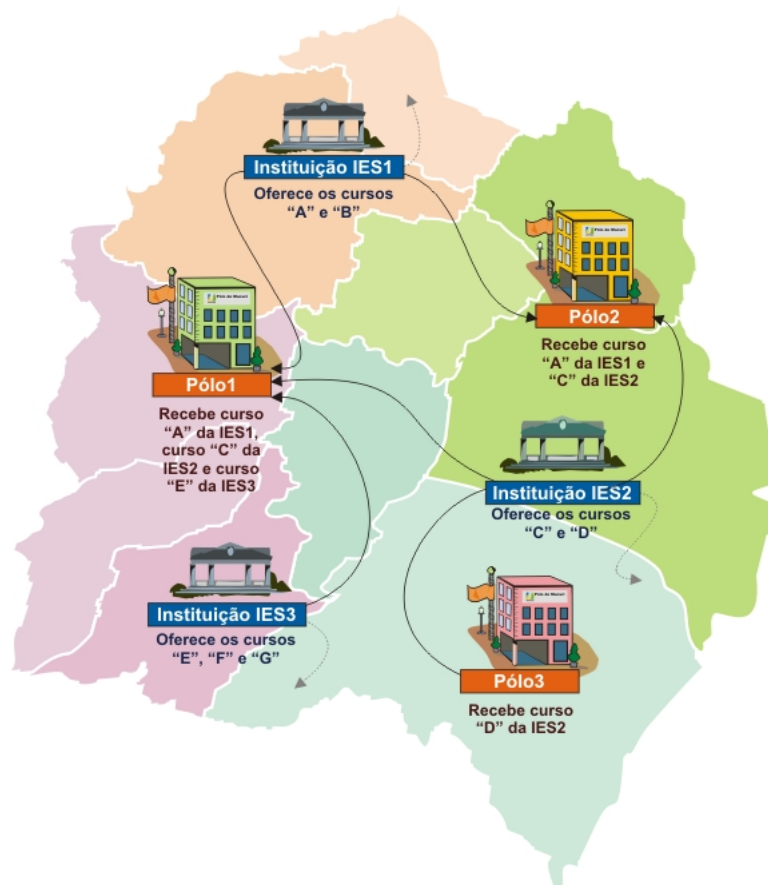


Figura 1 – Estrutura e funcionamento da UAB

Fonte: Miranda (2008)

### 2.1.3 Programa Nacional de Formação em Administração Pública

Em 2009 o MEC, visando a ampliação da UAB, implementou o Programa Nacional de Formação em Administração Pública (PNAP). Segundo Demarco (2013, p. 5), o programa tem por objetivo:

[...] propiciar aos estudantes, gestores públicos, uma tomada de consciência sobre as atuais políticas de governo, a partir do desenvolvimento das capacidades necessárias para conhecer o contexto socioeconômico, cultural e político que conformam o campo da gestão pública no Brasil.

O PNAP surgiu como continuidade ao curso de Administração a distância e foi construído coletivamente de maneira colaborativa em conjunto com a Escola Nacional de Administração Pública (ENAP), Ministério da Saúde, Conselho Federal de Administração (CFA), Secretaria de Educação a Distância (SEED/MEC) e por mais de 20 instituições públicas de ensino superior (IPES), vinculadas à UAB (DEMARCO; VIEIRA, 2014).

O programa caracteriza-se pela reafirmação do caráter estratégico da UAB em prol do desenvolvimento científico e da inovação tecnológica visando o crescimento sustentável do Brasil. E também é uma resposta a necessidade da qualificação de gestores públicos em todos os níveis governamentais. Ao adotar a modalidade EAD manteve um padrão nacional de grade curricular e materiais didáticos visando a criação de um perfil nacional do Administrador Público que compreenda as especificidades das esferas municipais, estaduais e Federal (DEMARCO; VIEIRA, 2014).

O PNAP engloba o curso de Bacharelado em Administração Pública e os cursos de Especialização (*Lato Sensu*) em Gestão Pública, Gestão Pública Municipal e Gestão em Saúde (PNAP, 2014).

## 2.2 AMBIENTE VIRTUAL DE APRENDIZAGEM

Um Ambiente Virtual de Aprendizagem, é o “espaço virtual” planejado especialmente para abrigar cursos, geralmente na modalidade a distância ou semi presencial. O objetivo de um AVA é permitir ao aluno o acompanhamento organizado e sistematizado da rotina de estudo além da possibilidade da recuperação de conteúdo já estudado.

Os ambientes virtuais foram pensados observando vários softwares que através do uso da internet, agregavam e conectavam pessoas, a grande maioria deles voltada ao entretenimento. Então surgem os softwares semelhantes, mas agora focados na no ensino e aprendizagem pela internet que possibilitam organizar e controlar as atividades, interações, cursos através de diversas ferramentas como fóruns, conferências, bate-papos, documentos de texto, imagens, vídeos e arquivo de áudio. Para (ALMEIDA, 2003), ambientes virtuais são:

[...] sistemas computacionais disponíveis na internet, destinados ao suporte de atividades mediadas pelas tecnologias de informação e comunicação. Permitem integrar múltiplas mídias, linguagens e recursos, apresentar informações de

maneira organizada, desenvolver interações entre pessoas e objetos de conhecimento, elaborar e socializar produções tendo em vista atingir determinados objetivos.

Possuem ferramentas disponíveis como *e-mails*, fóruns, conferências, bate-papos, arquivos de textos, imagens, vídeos, *wikis*, *blogs*, dentre outros que potencializam o poder de educação através da comunicação e da integração de mídias. O *hiperlink* é um recurso que possibilita o aumento do raio de conhecimento a ser desenvolvido pelos alunos, é utilizado dentro do próprio AVA possibilitando a navegação na plataforma e em seu conteúdo por exemplo entre textos indicados ou entre discussões em fóruns diferentes, como também de dentro para fora em casos de documentos e arquivos que estão disponíveis na web.

Os AVA portanto são facilitadores para a Educação à Distância, pois permitem uma interação assíncrona e síncrona entre alunos, professores e tutores através do uso da internet e das diversas ferramentas neles disponíveis. Nessa perspectiva existem ambientes estruturados e desenvolvidos com o objetivo de apoiar o processo de ensino e aprendizagem via rede; são os chamados Sistemas de Gerenciamento de Curso, que são apresentados a seguir.

### 2.2.1 Sistema de Gerenciamentos de Curso

Um Sistema de Gerenciamento de Cursos (SGC) também chamado de Sistema de Gestão da Aprendizagem(SGA)(do inglês, *Learning Management System, LMS*) é um AVA que disponibiliza uma série de recursos, síncronos e assíncronos, que dão suporte ao processo de aprendizagem, permitindo seu planejamento, implementação e avaliação.

Têm como objetivo primordial apoiar o processo de aprendizado, proporcionando e facilitando diversas formas de interação, como as que ocorrem entre os próprios alunos, entre os alunos e o conteúdo, e entre os alunos e o professor. Seguem algumas das características básicas dos SGC, segundo Gomes et al. (2009):

- Possuem recursos interativos;
- Permitem o controle das atividades e monitoramento de todas as interações e acessos dos alunos;
- Permitem a gestão de conteúdo, como a criação de cursos e a customização das informações de maneira que os usuários encontram facilmente o que precisam;
- Possuem um controle de usuários e sistemas de permissões de acesso ao conteúdo;

Portanto os SGC, além de agregarem as características dos AVA – e por isso muitas vezes serem identificados simplesmente como tal – trazem algumas perspectivas de controle e de estruturação de cursos como: gerenciamento de integrantes (alunos, tutores, professores e administradores do sistema), relatórios de acesso e de atividades, recursos para promover a interação e para a submissão de atividades, além de possibilitar a publicação e armazenamento

de conteúdos. Como exemplos de ambientes virtuais que podem ser caracterizados como SGC pode citar-se: o Moodle (detalhado na seção 2.2.1.1), o TelEduc , o AulaNet , o e-ProInfo , o WebCT, o Blackboard, entre outros.

### 2.2.1.1 Moodle

É um ambiente virtual muito utilizado pelas Instituições de Ensino no Brasil, oferece aos professores a possibilidade de criar e conduzir cursos, e ofertá-los no apoio a atividades do ensino presencial ou semipresencial e é peça chave no ensino a Distância. Possui várias ferramentas que auxiliam na administração do curso e forte embasamento na Pedagogia Construcionista Social (colaboração, atividades, reflexão crítica, etc.). Iniciado no ano de 1990, pelo australiano Martim Dougiamas, que buscava facilitar o uso da Internet como instrumento facilitador do Ensino a Distância (PEREIRA; CHAVES, 2007).

O SGC Moodle é um software livre, ou seja, qualquer pessoa têm acesso ao código fonte do software, podendo ser alterado, ampliado e modificado conforme a necessidade. Pode ser instalado em plataformas que consigam executar a linguagem php tais como Unix, Linux, Windows, MAC OS. Como base de dados pode utilizar Oracle, Interbase , MySQL, PostgreSQL, Access ou ODBC.

Moodle é o único sistema de fonte aberta atualmente disponível que pode competir com os grandes sistemas comerciais. Segundo o site oficial da ferramenta Moodle (2015), na documentação da ferramenta, é definida como:

O Moodle é um Sistema de Gerenciamento de Cursos - um pacote de software de código livre usando princípios pedagógicos, para ajudar educadores na criação de comunidades de aprendizado on-line. A palavra Moodle vem do acróstico “Modular Object Oriented Dynamic Learning Environment”, traduzindo significa objeto orientado para aprendizagem em ambiente dinâmico (virtual). O verbo to moodle traduz a ação de navegar sem pretensões, adequando a elaboração de outras tarefas simultaneamente.

A grande vantagem Moodle é o fato de permitir aos professores/tutores e aos alunos, de forma bastante simples e prática ensinar ou estudar um curso online, devido ao fato de possuir um conjunto de ferramentas que podem ser selecionadas pelo professor ou administrador do sistema de acordo com seus objetivos pedagógicos e necessidades de seu público-alvo e assim modelar um curso ou disciplina utilizando recursos como fóruns, diários, chats, lições, questionários, textos, wiki, tarefas, glossários e vídeos além de inserir e remover blocos, com informações sobre participantes, últimas notícias, calendários, entre outros.

### 2.2.1.2 Outros SGC

Existem diversos SGC disponíveis atualmente, todos possuem semelhanças entre as plataformas, mas apresentam cada um suas especificidades. Dois dos mais relevantes no mercado são: o TelEduc e o Blackboard.

O TelEduc é um ambiente para criação, participação e administração de cursos na Web desenvolvido pela Universidade Estadual de Campinas (UNICAMP). Foi concebido tendo como alvo o processo de formação dos professores para informática educativa, baseado na metodologia de formação contextualizada, desenvolvida por pesquisadores no Núcleo de Informática Aplicada à Educação da Unicamp (NIED) em parceria com o Instituto de computação (IC) (TELEDUC, 2015).

É um software livre de SGC inteiramente nacional e gratuito de fácil instalação, desenvolvido de forma participativa, todas as ferramentas foram planejadas e desenvolvidas conforme necessidades relatadas por seus usuários e pessoas não especialistas em computação. A flexibilidade deve-se a um conjunto enxuto de funcionalidades. Com isso, essas características que o diferenciam de outros ambientes para educação a distancia disponível no mercado (RAMOS, 2006).

O Blackboard é um SGC baseado na Internet, tem como objetivo facilitar a criação e gestão de cursos online proporcionando recursos multidirecionais para a construção da experiência educativa. Entre as 20 melhores instituições de ensino do mundo, 90% destas utilizam o Blackboard. Foi Criado em 1997, é considerado líder em plataformas educacionais para o aprendizado à distância e apoio ao ensino presencial devido a sua facilidade de uso, flexibilidade pedagógica, amplitude de funções e características intuitivas, propiciando maior autonomia no desenvolvimento, gerenciamento e oferecimento de conteúdos on-line (BLACKBOARD, 2015)..

A plataforma Blackboard é um programa pago, considerado também um sistema de gestão acadêmica, que além dos recursos didáticos, oferece ferramentas administrativas para o gerenciamento de cursos, como a visualização e o controle de todas as informações referentes a um curso - matrícula, notas, frequência dos alunos, etc; e também permite a integração de diferentes setores de uma instituição de ensino, como áreas administrativas, financeiras e acadêmicas, sob um só sistema (BLACKBOARD, 2015).

### 2.3 SISTEMAS DE APOIO A DECISÃO

Os Sistemas de Apoio a Decisão (SAD) são sistemas que não produzem apenas informações utilizadas para monitorar e controlar a organização, como nos Sistemas de Informações Gerenciais (SIG), mas que através de informações e modelos especializados são utilizados para apoiar o processo de tomada de decisão em níveis estratégicos.

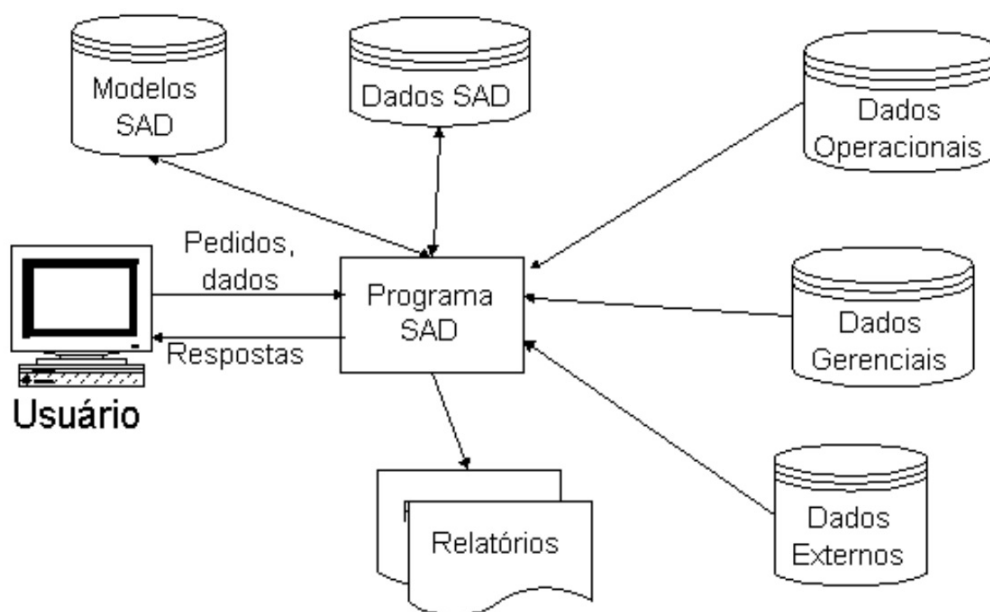
Um SIG é definido como um “processo de transformação de dados em informações que são utilizadas na estrutura decisória da empresa, proporcionando, ainda, a sustentação administrativa para otimizar os resultados esperados” (OLIVEIRA, 2005, p. 273). Os SIG são os mais antigos aplicativos de apoio a tomada de decisão, eles coletam, validam, executam operações, transformam, armazenam e apresentam todas essas informações através da geração de relatórios. Tais relatórios são apresentados da melhor maneira possível e com o nível de detalhe do conteúdo adequado ao usuário, tanto na forma textual, em planilhas ou visual através de gráficos.

Portanto um SIG está condicionado no auxílio para responder a questões como:

- Qual o número médio de vendas nos últimos 3 meses?
- Quais os produtos mais rentáveis da empresa?

Os SAD são “sistemas que tratam de assuntos específicos, estatísticas, projeções e comparações de dados referentes ao desempenho da empresa, estabelecendo parâmetros para novas ações dentro do negócio da empresa” (POLLONI, 2001, p. 32). Portanto, um SAD foca em partes específicas de um problema e que se alteram com rapidez, fornecendo respostas interativas para questões não rotineiras, utilizando ferramentas de análise e modelagem sofisticada através de simulações, regressão e modelos interativos .

Tais sistemas captam informações de diversas bases de dados, como mostrado na Figura 2, onde nota-se a abrangência dos dados que são utilizados para a criação de elementos confiáveis para a auxiliar na tomada de decisão.



**Figura 2 – Sistema de Apoio a Decisão**

**Fonte: Marques e Carvalho (2013)**

Um SAD “é qualquer sistema que forneça informações qualificadas (sintetizadas e estatísticas) baseadas em um ou mais sistemas de informação integrados” (POLLONI, 2001, p. 54), por este motivo, estes sistemas possuem mais poder analítico do que outros sistemas, pois são construídos com uma variedade de modelos para favorecer a análise de dados, ou por muitas vezes condensam grandes quantidades de dados, dentro de um formulário, onde podem ser analisados pelos tomadores de decisão.

Dessa maneira, um SAD também fornece relatórios, mas permite que o usuário além de perguntas rotineiras, elabore novas perguntas ao sistema, intervindo diretamente como os dados serão apresentados. Algumas questões que um SAD ajuda a responder para uma empresa:

- Se subíssemos os preços em 10% em quanto aumentaria o lucro?
- Quanto custaria a mais fabricar nosso produto se o custo relativo a salários dos funcionários subisse 15%?
- Devemos vender diretamente ao mercado varejista, continuar a vender através dos distribuidores, ou ambos?

Uma vez que o objetivo é dar suporte às decisões a serem tomadas, o SAD deve simular situações, analisar alternativas e propor soluções adaptadas ao estilo cognitivo do usuário, tornando possível o alinhamento de estratégias. O processo de tomada de decisão se torna mais confiável e acontece através da interação constante do usuário com o SAD.

### 2.3.1 Sistemas de Apoio a Decisão em Educação a Distância

Segundo Lopes (2003), um dos grandes desafios da EAD é a dificuldade de se realizar a avaliação e o acompanhamento do aprendizado do aluno que é ocasionada por diversos motivos, sendo um deles, a falta de contato presencial entre professores e alunos. Em seu trabalho, é apresentada uma proposta de acompanhamento do aprendizado capturando e analisando informações do aluno armazenadas nos AVA.

Santos (1999 apud Lopes, 2003) enumera algumas atitudes e comportamentos de alunos em cursos a distância via internet, que podem ser identificados, para estudar a atividade e desempenho do aluno com o ambiente:

- Que caminhos foram percorridos sobre os conteúdos disponibilizados pelo professor?
- Qual o grau de utilização e pesquisa de fontes suplementares fornecidos pelo professor?
- Quais foram às contribuições e em que grau, na realização de tarefas cooperativas?
- Com que frequência os alunos contataram o professor por e-mail?
- Os alunos entram em contato com o professor somente no período próximo às avaliações?
- Quais foram suas assiduidades e graus de participação em bate-papos, videoconferência, lista e fóruns de discussão?
- Os trabalhos desenvolvidos pelos aprendizes mostraram boa utilização dos recursos educacionais disponíveis no curso?

Segundo Nunes (2007) avaliar e acompanhar o aprendizado de um aluno em um curso a distância não são tarefas triviais, pois envolvem, além de teorias pedagógicas, questões tecnológicas, como autenticação e rastreamento do aluno na ferramenta e apoio à tomada de decisão, mediante situações problemáticas na dinâmica de ensino-aprendizagem.



A potencialidade de uma ferramenta de apoio à decisão voltada a EAD, está na possibilidade do acompanhamento das disciplinas e turmas através de informações relevantes ao coordenador de curso, que analisando essas informações, será capaz de detectar prováveis situações em que a tomada de decisão, como intervenções pessoais com turmas, professores ou alunos, se faça necessária para fins de adequação, no que diz respeito ao cumprimento dos objetivos propostos.

Neste caso, considerando somente as interações do aluno com o ambiente de ensino, informações sobre o comportamento do estudante em relação ao curso, o interesse, a participação e o desempenho podem agora serem analisados pela ótica computacional.

#### 2.4 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

A Descoberta de Conhecimento em Banco de Dados (DCBD), do inglês, *Knowledge Discovery in Databases (KDD)*, é o processo no qual busca-se obter informações úteis em bases de dados, que são de difícil detecção por métodos tradicionais de análise, pois estes tratam apenas as informações explícitas. O Quadro 2 mostra as diferentes áreas do conhecimentos que fazem parte do processo de DCBD.

<b>Principais áreas de conhecimento do DCBD</b>	
Aprendizado de Máquina	Utilização de estratégias de aprendizado de máquina e modelos cognitivos para aquisição automática de conhecimento
Bases de Dados	Utilização de técnicas e pesquisas com o objeto de melhorar e aprimorar a exploração das características dos dados.
Estatística e Matemática	Utilização de modelos matemáticos ou estatísticos para criação e identificação de padrões e regras entre os dados
Sistemas Especialistas	Programas de computadores de inteligência artificial, ou seja, soluções criadas em linguagem de máquina para resolver problemas do mundo real.
Visualização de Dados	Descoberta da informação, ou seja, análise do resultado final que pode ser demonstrado em forma de gráficos, figuras e ícones.

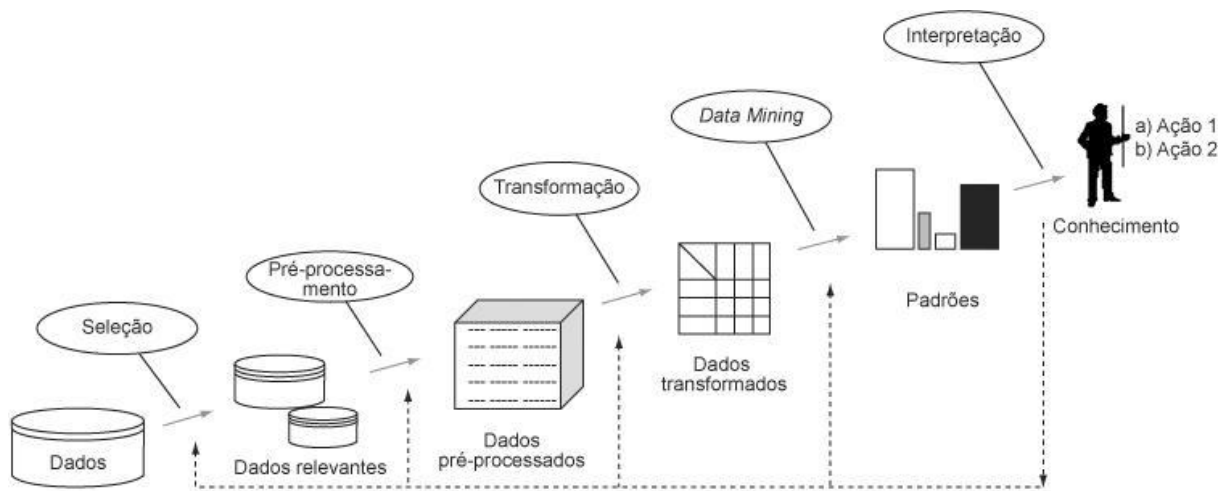
**Quadro 2 – Áreas de conhecimento encontradas em DCBD e sua utilização**

**Fonte: (LEMOS, 2003)**

Fayyad et al. (1996) define a DCBD como um processo não trivial de identificação de padrões em dados. Estes padrões devem ser válidos, novos, potencialmente úteis e compreensíveis. Sendo assim, não trivial significa que deve-se utilizar uma técnica de busca ou de inferência para a descoberta de informações; novos, potencialmente úteis e compreensíveis, mostra a necessidade que estas informações sejam inéditas e que possam ser relevantes para o usuário ou sistema e devem ser apresentadas de maneira clara e legível.

Essas etapas são iterativas, pois o resultado de cada uma é dependente dos resultados das que a precedem. E interativo, pois os profissionais envolvidos no processo de DCBD podem tomar decisões e intervir, controlando fluxo das atividades. O processo de DCBD compreende

todo o ciclo que o dado extraído percorre até se tornar conhecimento ou informação, a Figura 3 ilustra esse processo.



**Figura 3 – Etapas do processo de DCBD**

Fonte: Fayyad et al. (1996)

#### 2.4.1 Seleção

A fase de seleção de dados é a primeira fase do processo, possui grande complexidade e participação na qualidade do resultado final, uma vez que nesta fase é escolhido o conjunto de dados que contém todas as possíveis variáveis, características ou atributos, e os registros que farão parte da análise, estes podem vir de uma série de fontes diferentes (*data warehouses*, planilhas, sistemas legados) e possuir os mais diversos formatos.

#### 2.4.2 Pré-processamento

O Pré-processamento dos dados é uma parte crucial no processo de DCBD, pois a qualidade dos dados vai determinar a eficiência dos algoritmos de mineração. Aqui são realizadas tarefas que eliminam dados redundantes e inconsistentes, recuperam dados incompletos e avaliam possíveis dados discrepantes ao conjunto, estes devem ser analisados por um especialista do domínio, pois em grande parte dos casos apenas alguém que realmente entende do assunto é capaz de dizer se um dado inconsistente é um valor atípico a série ou um erro de digitação.

Nesta fase também são utilizados métodos de redução ou transformação para diminuir o número de variáveis envolvidas no processo, visando com isto melhorar o desempenho do algoritmo de análise.

A identificação de dados inapropriados dentro do conjunto selecionado é problemática, e isto dificulta a automatização desta fase. Definir um dado como “ruim” dentro do conjunto depende da estrutura do mesmo e também de que aplicação é dada a ele (DUNKEL et al., 1997).

### 2.4.3 Transformação dos Dados

A Transformação do Dados é a fase do DCBD que antecede a fase de Mineração de Dados. Depois de selecionados, limpos e pré-processados, os dados necessitam ser armazenados e formatados adequadamente para que os algoritmos possam ser aplicados.

Nesta etapa também é possível obter dados faltantes através da transformação ou combinação de outros, são os chamados “dados derivados”. Um exemplo de um dado que pode ser calculado a partir de outro é a idade de um indivíduo, que pode ser encontrada a partir de sua data de nascimento.

### 2.4.4 Mineração de Dados

A etapa de Mineração de Dados, do inglês *Data Mining*, recebe o maior destaque e ênfase dentro deste processo.

Segundo Berry e Linoff (1997), MD é a exploração e análise, de forma automática ou semi-automática, de grandes bases de dados com objetivo de descobrir padrões e regras. O objetivo do processo de mineração é fornecer as corporações informações melhores estratégias nas mais diversas áreas que sejam de seu interesse.

Tan, Steinbach e Kumar (2009) explicam que a MD é uma etapa da DCBD, com o intuito de descobrir padrões úteis e recentes que poderiam, de outra forma, permanecer ignorados. Combinando métodos tradicionais de análise de dados com algoritmos sofisticados, as informações descobertas permitem analisar novos tipos de dados e também tipos antigos de novas maneiras.

Dessa maneira, as técnicas e ferramentas utilizadas na etapa de mineração de dados, são as que causam a difusão de um conceito errado sobre mineração de dados, que definem essas técnicas como sistemas que podem automaticamente minerar todos os conceitos valiosos que estão escondidos dentro de um grande banco de dados sem a intervenção ou direcionamento humano.

Portanto Mineração de Dados é um processo altamente cooperativo entre homens e máquinas, que visa a exploração de grandes repositórios de dados, com o objetivo de extrair conhecimento através de diversas técnicas confiáveis e válidas pela sua expressividade estatística. Este passo é descrito mais detalhadamente na seção 2.5.

### 2.4.5 Interpretação e Avaliação

A última a etapa do processo de DCBD é a de Interpretação e Avaliação dos resultados. Esta etapa consiste em avaliar o conhecimento extraído das bases de dados, identificar padrões e interpretá-los e o resultado deve ser compreensível para os tomadores de decisão, os quais são responsáveis por validar o conhecimento adquirido. Dessa forma, observa-se a importância do trabalho em conjunto dos profissionais envolvidos no DCBD e o especialista no domínio, a fim

de que os resultados do processo de descoberta de conhecimento sejam cada vez mais relevantes e alcancem a confiabilidade desejada.

## 2.5 MINERAÇÃO DE DADOS

A mineração de dados é a principal etapa do DCBD, consiste na aplicação de algoritmos capazes de lidar com escalabilidade e alta dimensionalidade em grandes bases de dados, com a finalidade extrair padrões de comportamento delas. Segundo Fayyad et al. (1996), o processo de mineração possui dois objetivos principais: predição, onde a ideia é prever o comportamento futuro de algumas variáveis da base de dados estudada, e descrição, onde o objetivo é identificar padrões que representam a distribuição dos itens de tal forma que esses padrões sejam passíveis de interpretação.

### 2.5.1 Tarefas da Mineração de Dados

De acordo com o que se pretende, podem ser realizadas várias tarefas de mineração de dados. Entende-se como tarefa, a especificação da intenção do que se busca nos dados, que tipo de padrões ou regularidades se busca encontrar. De modo geral, podem-se agrupar as tarefas de Data Mining quanto ao objetivo pretendido em 3 grandes grupos: Tarefas Descritivas, Tarefas de Amostragem e Tarefas de Prognóstico.

#### 2.5.1.1 Tarefas Descritivas

Busca encontrar associações e relações, caracterizando e descrevendo o modelo e também encontrar informações que sejam relevantes de difícil visualização.

- **Classificação:** consiste em categorizar os dados em classes que possuam alguma similaridade em alguma característica e após a criação desse modelo aplicar a dados não classificados a fim de categorizá-los também.
- **Associações:** Tem como objetivo identificar fatos que ocorrem em conjunto ou de forma condicionada e que possuem relacionamentos entre seus itens.
- **Agrupamento:** semelhante a tarefa de classificação, mas as classes são definidas durante a tarefa de acordo com os atributos ditos direcionadores da categorização e os grupos são formados conforme similaridade entre esses atributos.
- **Descrição:** consiste na descrição textual de um conjunto de particularidades, observadas com frequência para um determinado evento. Utilizada geralmente para traçar perfis comportamentais como descrever os perfis dos clientes que tem comportamentos de compra similares.

- **Detecção de Sequências:** Tem por objetivo estabelecer estabelecer relações temporais entre fatos. Um exemplo seria detectar que em 33% das compras de notebooks em até um mês os compradores voltam para comprar um mouse.
- **Segmentação:** é a subdivisão do conjunto de dados em conjunto menores de acordo com alguma distinção. Por exemplo, segmentar consumidores por região e sexo antes de buscar possíveis associações, assim seria possível encontrar diferenças nos hábitos de compras nestas regiões entre homens e mulheres.

#### 2.5.1.2 Tarefas de Amostragem

Esta tarefa tem como objetivo encontrar comportamentos que fogem demais a normalidade da situação, garantindo assim a confiabilidade da amostragem e seus resultados.

- **Detecção de Desvios:** aqui devem ser detectados os dados ditos desarmônicos que não obedecem o comportamento geral do modelo de dados, eles podem ser tratados ou simplesmente descartados antes de iniciar o processo de mineração.
- **Análise de Desvios:** similar a detecção de desvios, mas nesse caso, a medida de comparação que vai definir se um dado é desarmônico em relação ao modelo estudado já é um padrão estabelecido anteriormente. Por exemplo, analisar fraude de um cartão de crédito, se em um determinado mês, a fatura aumenta muito em relação ao padrão de consumo do usuário, incluindo localidade, valores e tipos de itens das compras realizadas.

#### 2.5.1.3 Tarefas de Prognóstico

Essa tarefa tem como objetivo estimar valores desconhecidos ou comportamentos futuros ou inferir um determinado valor utilizando resultados ou informações obtidas em análises descritivas.

- **Estimação:** estimar valores desconhecidos com bases em valores conhecidos. Estimar o salário e o número de filhos de um individuo utilizando a analisando seus gastos e sua idade.
- **Predição:** predizer um valor futuro baseando se em valores já conhecidos. Predizer o salário de um individuo daqui a alguns anos, baseado em sua formação escolar, seu emprego e ramo profissional atual,

#### 2.5.2 Técnicas da Mineração de dados

Harrison (1998) afirma que não há uma técnica que resolva todos os problemas de mineração de dados. Diferentes métodos servem para diferentes propósitos, cada método oferece suas vantagens e suas desvantagens. A familiaridade com as técnicas é necessária para facilitar

a escolha de uma delas de acordo com os problemas apresentados. Nesta seção são descritas algumas técnicas de mineração de dados comumente usadas.

#### 2.5.2.1 Descoberta de Regras de Associação

A técnica de descoberta de regras de associação estabelece uma correlação estatística entre certos itens de dados em um conjunto de dados. Uma regra de associação tem a forma geral  $X_1 \dots X_n \Rightarrow Y [C,S]$ , onde  $X_1, \dots, X_n$  são itens que prevêm a ocorrência de  $Y$  com um grau de confiança  $C$  e com um suporte mínimo de  $S$  e “ $\Rightarrow$ ” denota um operador de conjunção (AND).

Um exemplo desta regra pode ser que 90% dos clientes que comprem leite, também comprem pão; o percentual de 90% é chamado a confiança da regra. O suporte da regra leite  $\Rightarrow$  pão é o número de ocorrências deste conjunto de itens na mesma transação.

A técnica de descoberta de regras de associação é apropriada à tarefa de associação. Como exemplos de algoritmos que implementam regras de associação tem-se: Apriori, AprioriTid, AprioriHybrid, AIS, SETM e DHP.

#### 2.5.2.2 Árvores de Decisão

Uma árvore de decisão é uma árvore onde cada nó não terminal representa um teste ou decisão sobre o item de dado considerado (GOEBEL; GRUENWALD, 1999). O objetivo principal é separar as classes; tuplas de classes diferentes tendem a ser alocadas em subconjuntos diferentes, cada um descrito por regra simples em um ou mais itens de dados. Essas regras podem ser expressas como declarações lógicas, em uma linguagem como SQL, de modo que possam ser aplicadas diretamente a novas tuplas. Uma das vantagens principais das árvores de decisão é o fato de que o modelo é bem explicável, uma vez que tem a forma de regras explícitas Harrison (1998).

A técnica de árvore de decisão, em geral, é apropriada às seguintes tarefas: classificação e regressão. Alguns exemplos de algoritmos de árvore de decisão são: CART, CHAID, C5.0, Quest, ID-3, SLIQ e SPRINT.

#### 2.5.2.3 Raciocínio Baseado em Casos

Também conhecido como MBR (*Memory-Based Reasoning* – raciocínio baseado em memória), o raciocínio baseado em casos tem base no método do vizinho mais próximo. O MBR procura os vizinhos mais próximos nos exemplos conhecidos e combina seus valores para atribuir valores de classificação ou de previsão Harrison (1998). Portanto o MBR tenta solucionar um dado problema fazendo uso direto de experiências e soluções passadas. A distância dos vizinhos dá uma medida da exatidão dos resultados.

Na aplicação do MBR, segundo Berry e Linoff (1997), existem quatro passos importantes: 1) escolher o conjunto de dados de treinamento; 2) determinar a função de distância; 3) escolher o número de vizinhos mais próximos; e 4) determinar a função de combinação.

A técnica de raciocínio baseado em casos é apropriada às seguintes tarefas: classificação e segmentação. Os seguintes algoritmos implementam a técnica de raciocínio baseado em casos: BIRCH, CLARANS e CLIQUE.

#### 2.5.2.4 Algoritmos Genéticos

Os algoritmos genéticos são métodos generalizados de busca e otimização que simulam os processos naturais de evolução. Um algoritmo genético é um procedimento iterativo para evoluir uma população de organismos e é usado em mineração de dados para formular hipóteses sobre dependências entre variáveis, na forma de algum formalismo interno (GOEBEL; GRUENWALD, 1999).

Os algoritmos genéticos usam os operadores de seleção, cruzamento e mutação para desenvolver sucessivas gerações de soluções. Com a evolução do algoritmo, somente as soluções com maior poder de previsão sobrevivem, até os organismos convergirem em uma solução ideal (HARRISON, 1998).

A técnica de algoritmos genéticos é apropriada às tarefas de classificação e segmentação. Exemplos de algoritmos genéticos: Algoritmo Genético Simples Genitor e CHC, Algoritmo de Hillis, GA-Nuggets, GA-PVMINER.

#### 2.5.2.5 Redes Neurais Artificiais

Uma das principais vantagens das redes neurais é sua variedade de aplicação, mas os seus dados de entrada são difíceis de serem formados e os modelos produzidos por elas são difíceis de entender (HARRISON, 1998).

As redes neurais são uma classe especial de sistemas modelados seguindo analogia com o funcionamento do cérebro humano e são formadas de neurônios artificiais conectados de maneira similar aos neurônios do cérebro humano. Como no cérebro humano, a intensidade de interconexões dos neurônios pode alterar (ou ser alterada por algoritmo de aprendizagem) em resposta a um estímulo ou uma saída obtida que permite a rede aprender (GOEBEL; GRUENWALD, 1999).

A técnica de redes neurais é apropriada às seguintes tarefas: classificação, estimativa e segmentação. Exemplos de redes neurais: Perceptron, Rede MLP, Redes de Kohonen, Rede Hopfield, Rede BAM, Redes ART, Rede IAC, Rede LVQ, Rede Counterpropagation, Rede RBF, Rede PNN, Rede Time Delay, Neocognitron, Rede BSB

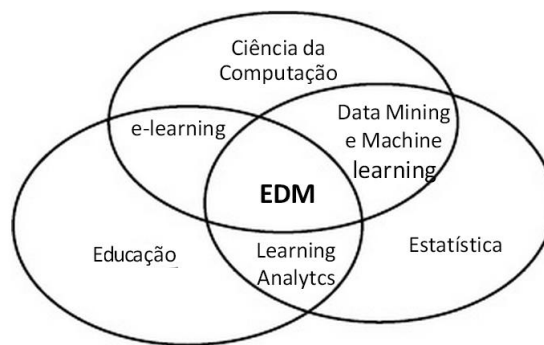
#### 2.5.3 Mineração de Dados Educacionais

Dentro do contexto da Mineração de Dados, encontra-se a área de Mineração de Dados Educacionais (do inglês Educational Data Mining, EDM), que é uma área recente de pesquisa e com enorme potencial para melhorar a qualidade do ensino, influenciando diretamente no processo de aprendizagem.

Para Romero e Ventura (2013), os esforços da EDM estão focados, principalmente, em três áreas de pesquisa:

- I- Desenvolver ferramentas e técnicas computacionais.
- II- Definir quais perguntas se deve fazer aos dados, ou seja, como obter as respostas mineando os dados educacionais.
- III- Determinar quem são os interessados que poderiam se beneficiar de mecanismos de relatórios aprimorados, possíveis através das técnicas de EDM.

Ainda segundo Romero e Ventura (2013), EDM é a combinação de 3 principais áreas do conhecimento: Computação, Educação e Estatística. A interseção dessas áreas ainda fornece três subáreas: E-learning, Data Mining e Machine Learning e a Learning Analytics que também se relacionam diretamente com a EDM (Figura 4).



**Figura 4 – Principais áreas relacionadas com a EDM**

**Fonte: Romero e Ventura (2013)**

O surgimento da Mineração de Dados Educacionais acompanhou a grande ampliação da disponibilidade de dados sobre as atividades associadas com a Educação, principalmente aquelas mediadas por AVA. Assim a EDM trata do desenvolvimento de métodos e técnicas para explorar dados originados em ambientes educacionais, como predição, agrupamento, mineração de relações, descoberta com modelos e tratamento de dados para apoio à decisão, com o propósito de contribuir para o entendimento de como é o processo de aprendizagem nesses ambientes (RIGO et al., 2014).

Portanto a EDM está orientada ao uso de técnicas de mineração de dados em ambientes educacionais, tendo em vista a quantidade de dados gerados nas instituições de ensino e por sistemas. Conhecer e saber como utilizar técnicas de mineração de dados, leva a melhor compreensão das relações existentes e potencialmente úteis dentro das bases dos AVA por exemplo.

## 2.6 TRABALHOS CORRELATOS

Na literatura técnica há trabalhos que utilizaram a DCBD e a Mineração de Dados afim de identificar padrões de desempenho no âmbito educacional. A seguir, são apresentados alguns



trabalhos que fazem uso de métodos de mineração de dados e técnicas de classificação, assim como outros que se correlacionam com o tema, tendo em vista suas características em comum.

- Detoni, Araujo e Cechinel (2014) buscaram identificar através da EDM prever com antecedência risco de reprovação de um estudante, utilizando a contagem de interações do ambiente virtual Moodle da Universidade Federal de Pelotas. Na etapa de coleta, foram extraídos os identificadores dos usuários envolvidos no processo, professores, tutores e alunos, e a data e horário das interações, utilizando somente o número de interações dos usuários com o sistema. Os alunos foram rotulados como aprovado e reprovado e foram gerados atributos derivados do número de interações: interações por semana, média de interações por semana, número de semanas sem interação e a razão de interações com professores e tutores e o fator de empenho, que é a razão entre as interações do alunos e a média de interações da turma.

A pesquisa, portanto, não englobou o meio utilizado (fórum, atividade, chat, etc.) mas sim a interação de modo geral com a plataforma, não abordando dessa maneira o recurso utilizado. Os modelos apresentados para prever a situação do aluno foram focados em 3 diferentes casos: utilizando o número absoluto de interações por estudante, adicionando os atributos derivados e balanceando o conjunto de exemplos com atributos derivados. Para avaliar foi utilizado a ferramenta *Waikato Environment for Knowledge Analysis*(WEKA), e os algoritmos Rede Bayesiana, Rede Neural (Perceptron de Múltiplas Camadas), J48 e Floresta Aleatória como classificadores. Os resultados mostraram que o modelo utilizando somente o número de interações, durante a primeira semana, são próximas a zero, depois da quinta semana, novos dados acrescentam muito pouco.

O classificador Redes Bayesianas obteve o melhor resultado. Já utilizando atributos derivados, estes apresentaram um comportamento bastante semelhante ao anterior, e novamente o classificador Redes Bayesianas foi o mais adequado. Já utilizando o balanceamento e os atributos derivados, houve um considerável aumento da predição com acurácia média de 80 % ao longo das sete semanas. A comparação apresentada, mostra que em geral, Redes Bayesianas foram o melhor classificador para o modelo de predição proposto.

- Junior, Noronha e Kaestener (2014) questionaram a existência de alguma correlação entre o empréstimo de livros em biblioteca com o abandono do curso, e utilizam o conceito da mineração de dados educacionais e séries temporais para responder tal questionamento. A técnica de Mineração de dados, utilizada nessa pesquisa foi a classificação, e os classificadores investigados nesta pesquisa foram: árvore de decisão, vizinho mais próximo, redes bayesianas, redes neurais e máquina de vetores de suporte. Na extração foi utilizada uma série temporal, com informações do vínculo do aluno com o curso. Nos experimentos realizados nesta pesquisa foram utilizados os seguintes algoritmos de classificação: J48 (árvore de decisão), Naive Bayes (redes bayesianas), SVM, usando a implementação SMO - *Sequential Minimal Optimization*, Multilayer Perceptron (redes neurais) e IBk (vi-

zinho mais próximo). As bases utilizadas na pesquisa foram a do sistema acadêmico da UTFPR (Universidade Tecnologia Federal do Paraná) e do sistema integrado de bibliotecas Pergamun, já a série temporal foi criada com informações dos alunos ao final de cada ano e semestre como, idade, curso, coeficiente, notas relativas ao Enem, situação do aluno e do Pergamun a quantidade de empréstimos na Biblioteca. A ferramenta de mineração foi o Weka e houve dois tipos de experimentos, utilizando o atributo de coeficiente, e o segundo não. O primeiro experimento, o Atributo coeficiente foi discretizado, para que pudesse ser minerado, o resultado do J48. verifica-se que o atributo coeficiente, é mais significativo que o número de empréstimos, desta forma não foi possível estabelecer uma correlação da evasão com o número de empréstimo de livros. O segundo experimento, não utiliza o atributo do coeficiente, assim foi encontrada a correlação que 80,6 % dos alunos com mais de 18 anos que emprestaram livros na biblioteca, não desistiram do curso, representando 70 % da amostra. Isto é um indicio que o empréstimo e livros pode estar relacionado com a permanência do aluno no curso. Ambos os experimentos permitiram que se descobrisse uma informação não investigada: 92,6% dos alunos com idade entre 17 e 18 anos desistiram do curso.

- Conti (2011) explorou técnicas de mineração de dados para análise dos prazos de entregas de atividades no AVA Moodle do Instituto de Farroupilha – Campus de São Vicente do sul. Para o estudo, foram considerados os períodos que as atividades permaneceram abertas, o curso proveniente da tarefa e o período que a postagem foi realizada. O processo foi baseado nas etapas do DCBD utilizando a ferramenta de mineração Weka. Como resultado, observou-se a incidência de postagens próximas ao termino do prazo de postagem, quando esse era superior a 15 dias na graduação e nos cursos de pós-graduação identificou-se que os tempos para envio de tarefas são maiores que os oferecidos na graduação e a maioria das postagens é realizada no final do prazo, desta maneira, é proposto a diminuição dos prazos das atividades afim de melhorar o acompanhamento por parte de professores e tutores dos seus alunos. O trabalho também propõe a integração do DCBD automaticamente no ambiente virtual Moodle, tendo como base os algoritmo que apresentaram resultados satisfatórios, sendo eles o EM e J.48 através de agrupamento e classificação.
- Santos, Camargo e Camargo (2012) apresentaram um estudo de caso sobre a utilização de técnicas de avaliações formativas a partir de um Ambiente Virtual de Aprendizagem. Resultados preliminares mostraram que a abordagem sugerida constitui-se em uma alternativa viável para um correto diagnóstico da evolução dos alunos e identificação dos alunos com maior nível de dificuldade na disciplina. A partir de dados coletados do ambiente foram criados modelos preditivos que permitem a identificação da tendência à reprovação com acurácia superior a 72%.

### 3 METODOLOGIA

Este capítulo descreve os métodos e as ferramentas utilizadas para alcançar os objetivos propostos neste trabalho. A metodologia aplicada para a análise dos dados da base do AVA; apresentado na seção 3.2; é baseado nas etapas do processo de Descoberta de Conhecimento em Base de Dados propostas por Fayyad et al. (1996), descritos na seção 2.4, e os resultados da aplicação desta metodologia serão expostos no capítulo 4.

As ferramentas utilizadas, apresentadas na seção 3.3, foram o Weka, que é uma ferramenta de mineração de dados de código aberto; o PostgreSQL, que é um sistema gerenciador de banco de dados objeto-relacional, também de código aberto; e o programa de planilhas eletrônicas Microsoft Excel.

#### 3.1 APLICAÇÃO DO PROCESSO DE DCBD

A partir da base de dados selecionada para o Estudo de Caso, foram seguidas todas as etapas do processo de DCBD, conforme a exposto na seção 2.4.

Ocorreram momentos onde foi necessário o retorno à etapas anteriores, devido ao fato da DCBD ser um processo interativo e iterativo. Sua interatividade, permitiu intervir e controlar o curso das atividades. Já sua iteratividade, caracterizada por uma sequência finita de operações onde o resultado de cada uma das etapas foi dependente dos resultados das que a precederam, permitiu a possibilidade de repetições integrais ou parciais das etapas ou do processo como um todo.

##### 3.1.1 Seleção

A fase de seleção tem grande impacto na qualidade do resultado final do processo de DCBD. Conforme o procedimento descrito na seção 2.4.1, para a seleção do conjunto de dados inicial, optou-se por estudar a fonte de dados, buscando compreender os relacionamentos entre as tabelas; a estrutura dos logs disponíveis no AVA; como e quais dados estavam armazenados. Isso permitiu a compreensão do problema e dos dados disponíveis.

Para realizar a análise dos cursos, disciplinas e alunos a serem selecionados para o estudo, foram realizadas várias pesquisas no banco de dados, utilizando-se a ferramenta PgAdmin III, a fim de responder algumas questões iniciais:

- Quais cursos ofertados e disciplinas possuíam o maior número de alunos?
- Quais cursos e disciplinas possuíam o maior número de informações na tabela *ava\_log*?
- Quais disciplinas possuíam o maior número de alunos em comum entre si?
- Quais disciplinas possuíam dados referentes ao desempenho dos alunos?

Identificadas tais informações, foram selecionados os conjuntos mais relevantes ao estudo. Após, foram encontradas as interações dos alunos selecionados com o sistema. Essas interações foram classificadas por tipo de interação e módulo do sistema no qual foi realizada. Ao final desse processo foi gerada uma única base de dados, que continha os registros e todos os atributos que fizeram parte da análise. A próxima etapa deste processo, foi a aplicação do pré-processamento nesta base de dados.

### 3.1.2 Pré-Processamento

Nesta etapa realizou-se a análise do conjunto de dados selecionado, afim de garantir a eficiência do processo de mineração. O trabalho proposto na seção 2.4.2, primeiramente buscou encontrar os atributos desnecessários que anteriormente foram selecionados.

Após a remoção destes atributos, efetuou-se uma verificação sobre a integridade do conjunto alvo, procurando encontrar possíveis atributos ou registros com informações incompletas ou redundantes, por exemplo: foram pesquisadas instâncias com dados faltantes em alguma das disciplinas selecionadas; disciplinas que não tinham o número suficiente de alunos etc. Após terminada o processo de limpeza, alcançou-se a redução da dimensão da base de dados.

Foi realizada também a padronização dos atributos e outras adequações necessárias nos registros que levaram a melhorar a qualidade do conjunto de dados, melhorando assim a acurácia e eficiência dos processos de mineração subsequente.

Nesta etapa, para realizar a seleção de atributos, foi utilizado o algoritmo *Ranker* em conjunto com dois avaliadores de atributos, *InfoGainAttributeEval* e *GainRatioAttributeEval*, que avaliam o valor de cada atributo da base de dados para classificação. Ao fim da avaliação de atributos por um dos avaliadores, o algoritmo *Ranker* lista os atributos por ordem de valor.

### 3.1.3 Transformação

Na etapa de Transformação foram executadas atividades expostas na seção 2.4.3, que englobaram ações que buscaram adequar os dados para melhor interpretação pelos algoritmos de mineração de dados.

Esta etapa é necessária pois dados selecionados e pré-processados poderiam ainda apresentar formato inapropriado para um determinado algoritmo, como por exemplo, determinados algoritmos trabalham somente com valores numéricos e outros somente valores categóricos.

Por fim, foi realizada a criação de atributos derivados utilizando ou combinando atributos existentes, afim de gerar um conjunto de novos atributos que facilitaram a interpretação na etapa seguinte.

### 3.1.4 Mineração de Dados

Nesta etapa buscou-se escolher o método de mineração para se atingir os objetivos propostos. Com base nas tarefas descritas na seção 2.5.1, avaliou-se a aplicabilidade de cada

uma, conforme os objetivos e a solução que era desejada ser encontrada.

Optou-se pela aplicação da tarefa de classificação, que através de exemplos formados por um conjunto de atributos previsores e um atributo meta, que são pertencentes a classes conhecidas, busca-se encontrar correlações entre esses atributos. Desta forma utiliza-se as relações descobertas a partir dos exemplos para prever a classificação dos atributos metas que são desconhecidos. Na abordagem desse trabalho, os alunos foram classificados conforme seu desempenho acadêmico.

Foi utilizada a ferramenta Weka, descrita na seção 3.3.3. Aqui, foram aplicadas diversas técnicas de mineração com a ferramenta Explorer, e após foram geradas comparações entre as técnicas utilizadas e seus resultados, com a ferramenta *Experimenter*, ambas presentes no Weka.

Por fim, os resultados obtidos nesse processo, foram as respostas encontradas aos objetivos anteriormente propostos. Como o processo é interativo, foram feitas diversas tentativas para correção, até que a análise estivesse adequada.

Nesta etapa, para execução da tarefa de classificação, foi utilizado o algoritmo J48. Para avaliar o desempenho da classificação com o J48, foram utilizados os algoritmos Multilayer Perceptron, NaiveBayes, ZeroR e OneR. Estes algoritmos são apresentados na seção 3.3.3.1.

### 3.1.5 Interpretação e Avaliação

Após o término da etapa de MD, os resultados, as regras e as informações encontradas foram avaliadas conforme exposto na seção 2.4.5, afim da correta interpretação e para então apresentar uma leitura do conhecimento descoberto.

Nesta etapa também avaliou-se o processo como um todo, o que por muitas vezes implicou na reexecução de qualquer uma das fases anteriores caso o resultado obtido não fosse satisfatório ou as informações descobertas não tivessem apresentado consistência.

Os resultados da aplicação dos algoritmos são apresentados em formato texto pela ferramenta WEKA. Neste trabalho são utilizados para comparação os valores *Kappa Statistic* (Estatística Kappa), *TP Rate - True Prediction Rate* e *FP Rate - False Prediction Rate*.

O *TP Rate* representa a taxa de predição verdadeira, ou seja, a taxa de acerto do algoritmo para uma determinada classe, já o *FP Rate* representa a taxa de predição falsa, ou taxa de erro.

Segundo o site da FMUP (2015), a Estatística Kappa é uma medida de concordância usada em escalas nominais que nos fornece uma ideia do quanto as observações se afastam daquelas esperadas, fruto do acaso, indicando-nos assim o quão legítimas as interpretações são.

No quadro 3 são apresentados os valores de concordância conforme o valor da Estatística Kappa. Essa tabela pode ser usada para basear a interpretação dos resultados.

Valor de kappa	Concordância
0	Pobre
0 – 0,20	Ligeira
0,21 – 0,40	Considerável
0,41 – 0,60	Moderada
0,61 – 0,80	Substancial
0,81 – 1	Excelente

**Quadro 3 – Tabela de concordância dos valores do Kappa Statistic**

Fonte: FMUP (2015)

## 3.2 FERRAMENTAS UTILIZADAS

Para o desenvolvimento deste trabalho verificou-se a necessidade de duas classes de ferramentas: uma com a capacidade de tratamento, manipulação e extração e informações em bancos de dados, o PostgreSQL, descrito na seção 3.2.1; e outra para a aplicação de técnicas de mineração de dados, de forma parametrizada, o Weka, apresentado na seção 3.2.2, que possui uma série de algoritmos para as tarefas de mineração e tem a capacidade de expressar os resultados obtidos e permite a interpretação dos mesmos.

### 3.2.1 PostgreSQL

O PostgreSQL é um Sistema de Gerenciamento de Banco de Dados (SGBD) objeto-relacional de código aberto, tem mais de 15 anos de desenvolvimento, extremamente robusto, confiável e flexível. Roda em todos os grandes sistemas operacionais. É considerado objeto relacional por possuir algumas características da orientação a objetos, como herança e tipos personalizados (POSTGRESQL, 2015).

Suas principais características são: o tamanho ilimitado, possibilidade de criação de tabelas com capacidade de até 32 TB, tuplas com capacidade de até 1.6 TB e atributos com limite de 1GB de capacidade de armazenamento. não possui restrição quanto a quantidade de tuplas que podem ser armazenadas em uma tabela e podendo ser criadas em um única tabela até 1600 colunas (dependendo do tipo de dado armazenado).

Possui também o pgAdmin, software que disponibiliza uma interface gráfica para a administração do PostgreSQL com diversos recursos.

Optou-se por manter esta ferramenta para a manipulação e extração de informações da bases de dados desse trabalho, pois além de ser um software livre e um dos melhores e mais bem documentados SGBD disponíveis, o banco de dados do AVA utilizado nesse trabalho era proveniente de um ambiente PostgreSQL.

#### 3.2.1.1 Arquivo *dump*

Os arquivos em formato *dump* são comumente gerados pelas ferramentas de backup dos SGDB, nesses arquivos são armazenados toda a estrutura de tabelas, usuários existentes no

banco de dados, assim como todos os dados gravados no banco de dados. A recuperação ou restauração de um banco de dados pode ser feita a partir de um arquivo *dump* desse banco.

Para a utilização nesse trabalho, foi restaurado o arquivo *dump* disponibilizado pelo NUTEAD do banco de dados do AVA, para criação da base e restauração dos dados.

### 3.2.2 EXCEL

O programa Excel é um programa escrito e produzido pela empresa Microsoft baseado em planilhas eletrônicas. O sistema é utilizado para cálculos, estatísticas, gráficos, relatórios, formulários e entre outros requisitos das rotinas empresariais, econômicas, administrativas e domésticas.

Após a base de dados alvo ter sido selecionada, os dados foram exportadas para o formato .CSV (*Comma Separated Values* - Valores Separados por Vírgulas). Nesta planilha eletrônica, tais dados foram manipulados nas etapas iniciais do processo de DCBD, afim de melhorar visualização e controle do processo e também pela facilidade e rapidez de se trabalhar com planilhas eletrônicas.

Nesta ferramenta, foi instalado também um complemento, chamado *Excel 2007 & 2010 addin* (*Disponível em <http://blog.tomaskafka.com/node/625>*) que auxilia na transformação de arquivos .CSV, em arquivos .ARFF (Attribute-Relation File Format) o qual é utilizado pela Weka para aplicar a mineração de dados.

### 3.2.3 Weka

Existem disponíveis no mercado vários sistemas proprietários e não proprietários capazes de executar técnicas de mineração de dados. Para este trabalho optou-se pela ferramenta Weka, que é um software gratuito, usa a licença GNU (GNU/GPL). Foi desenvolvida pela universidade de Waikato na Nova Zelândia. Está implementado na linguagem Java e contém uma GUI (*Graphical User Interface*, em português Interface Gráfica do Usuário) que permite interagir com arquivos de dados e produzir resultados visuais (WEKA, 2015).

Esta ferramenta possui diversos métodos de classificação, associação e clusterização, além de ser uma ferramenta customizável e expansível, pois permite a remoção ou inclusão de novos métodos de forma bastante simples. Suporta apenas a manipulação de arquivos do tipo ARFF, baseado em ASCII, utilizados para definir atributos e seus valores. Permite também a apresentação gráfica de dados em forma de histogramas, possui modelos de gráficos para a montagem de redes neurais e a possibilidade de visualização de resultados em árvores de decisão (GOLDSCHMIDT; PASSOS, 2005).

A equipe de desenvolvimento lança periodicamente correções e *releases* da ferramenta e mantém um grupo de discussão sobre o software. A maioria das funções se originou de teses e pesquisas da Universidade de Waikato. Uma limitação está no fato que o volume de dados a serem manipulados fica limitado a quantidade de memória principal da máquina, sendo então a

escalabilidade um ponto negativo dessa ferramenta (SILVA, 2010).

### 3.2.3.1 Algoritmos

A escolha dos algoritmos disponíveis na ferramenta Weka, foi realizada de modo a exercitar as principais técnicas relacionadas com a tarefa de Classificação.

O algoritmo J48 tem a finalidade de gerar uma árvore de decisão baseada em um conjunto de dados de treinamento, sendo este modelo usado para classificar as instâncias no conjunto de teste. Um dos aspectos para a grande utilização do algoritmo J48 é que o mesmo se mostra adequado para os procedimentos, envolvendo as variáveis (dados) qualitativas contínuas e discretas presentes nas bases de dados (WITTEN; FRANK, 2005).

É considerado o algoritmo que apresenta o melhor resultado na montagem de árvores de decisão, a partir de um conjunto de dados de treinamento. Para a montagem da árvore, o algoritmo J48 utiliza a abordagem de dividir-para-conquistar, onde um problema complexo é decomposto em subproblemas mais simples, aplicando recursivamente a mesma estratégia a cada subproblema, dividindo o espaço definido pelos atributos em subespaços, associando-se a eles uma classe (WITTEN; FRANK, 2005).

A fim de comparar os resultados obtidos com o J48, foram utilizados os seguintes algoritmos encontrados no Weka:

- **MultilayerPerceptron**, baseado na técnica de redes neurais artificiais, é capaz de classificar dados cujos atributos sejam numéricos (com classe discreta), apresenta excelente capacidade de separação de dados, mas pela natureza da técnica é computacionalmente custoso.
- **NaiveBayes**: utiliza a classificação bayesiana, que é uma classificação estatística baseada em probabilidades e tem como objetivo prever a classe mais provável, isto é, calcular a probabilidade que uma amostra desconhecida pertença a cada uma das classes possíveis.
- **OneR**: é baseado em regras de decisão e gera apenas uma regra na execução do algoritmo. Utiliza o atributo do mínimo-erro para a predição (IREP - *Incremental Reduced Error Pruning*, técnica de simplificação de árvores que melhoram erros em conjuntos de dados com ruídos).
- **ZeroR**: este classificador prevê a classe mais frequente para atributos categóricos e a média para atributos numéricos. Tem como objetivo servir de base para avaliação de outros classificadores.

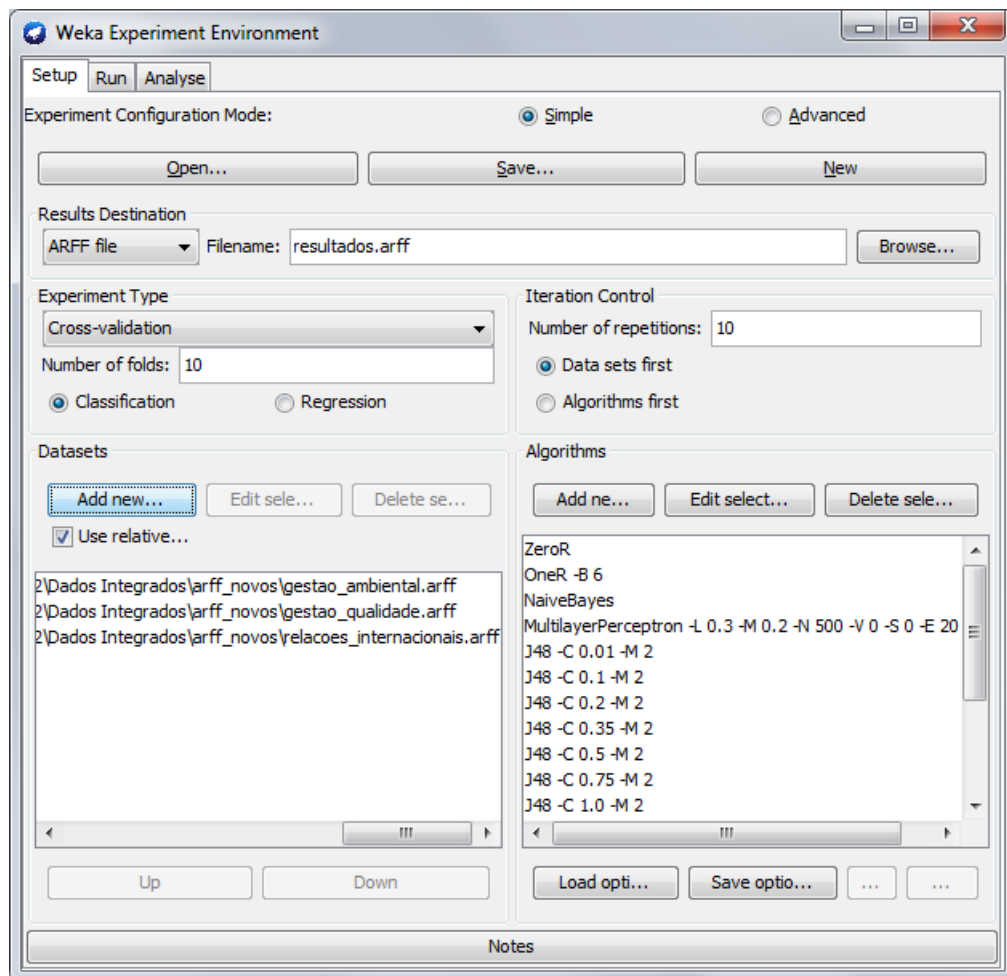
### 3.2.3.2 Experimentar

Para a aplicação dos algoritmos de mineração de dados nas bases, foram utilizadas duas ferramentas existentes no Weka, a ferramenta Explorer, onde foram analisadas as bases



de dados, a seleção de atributos e as execuções dos algoritmos de classificação. Já a ferramenta Experimenter foi utilizada para execução dos testes automatizados das execuções dos algoritmos, para gerar os resultados de comparação entre algoritmos distintos e configurações diferentes dos algoritmos.

A ferramenta Experimenter, apresentada na Figura 5, permite criar vários experimento e configurar várias bases de dados diferentes e vários algoritmos para serem aplicados à essas bases. Ao final da configuração, pode-se executar o experimento. Está ferramenta possui três abas: *setup*, *run* e *analyse*.



**Figura 5 – Experimenter**

**Fonte: Autoria própria**

Na aba *Setup*, é feita a configuração do experimento, podendo ser adicionada as bases de dados, configurados os algoritmos a serem executados e o arquivo onde serão gravados os resultados do experimento.

No campo *Experiment Type*, (tipo do experimento), é configurada o tipo de validação que será utilizado nos algoritmos, e o tipo do experimento, classificação ou regressão.

Nos experimentos executados, foram utilizadas as opções de validação cruzada (*Cross-validation*) com número de *Folds* de 10. O método de validação cruzada, separa o arquivo de

dados em  $k$  partes, a cada iteração é utilizado uma parte do conjunto para teste e o treinamento do algoritmo em  $k-1$  partes, ao final da execução são combinados os resultados das  $k$  iterações.

Em textitRun são executados os experimentos configurados na aba *Setup*, mostrando o tempo de execução de cada experimento e os erros de execução encontrados.

Na aba *Analyse* é feita a análise dos resultados dos experimentos, permite carregar os resultados do experimento atual, assim como resultados de experimentos salvos. Pode ser configurado para mostrar a análise em vários tipos de formatação: Latex, texto simples, etc. Além disso, permite a personalização dos algoritmos, bases e variáveis utilizadas para comparação. Utilizando-se desta ferramenta, foram aplicados os algoritmos selecionados às bases de dados configuradas, depois salvando todos os resultados em um arquivo ARFF.

### 3.2.3.3 Arquivo ARFF

Para aplicar a mineração de dados com a ferramenta Weka, os dados devem estar organizados no formato ARFF. Neste formato estão presentes informações como: domínio do atributo do atributo, os valores que os atributos podem representar e um atributo classe (DAMASCENO, 2010).

O arquivo ARFF é dividido em duas partes, como observa-se no exemplo de arquivo arff apresentado na Figura 6, a primeira contém uma lista de todos os atributos, onde se deve definir o tipo do atributo e/ou os valores que ele pode representar. Os valores devem estar entre chaves ( $\{\}$ ) e separados por vírgulas. A segunda é composta pelas instâncias presentes nos dados, os atributos de cada instância devem ser separados por vírgula, e aqueles que não contêm valor, o valor deve ser representados pelo caractere '?' (DAMASCENO, 2010).

```
@RELATION iris

@ATTRIBUTE sepallength REAL
@ATTRIBUTE sepalwidth REAL
@ATTRIBUTE petallength REAL
@ATTRIBUTE petalwidth REAL
@ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
```

Figura 6 – Arquivo iris.arff

Fonte: Exemplo de arquivo ARFF - Weka (2015)

Para a aplicação dos algoritmos de MD pelo Weka nas bases de dados, os arquivos precisaram ser convertidos para arquivos com o formato ARFF. Essa conversão foi realizada por meio da ferramenta citada na seção 3.2.2.

## 4 RESULTADOS

Neste capítulo são apresentados os resultados da aplicação do processo de DCBD proposta por Fayyad et al. (1996), descrevendo a aplicação de cada etapa do processo nos dados disponíveis.

Os dados utilizados para estudo no presente trabalho, são provenientes da base de dados do sistema Moodle, utilizado como AVA pelo NUTEAD. O termo de cessão dos dados está disponível no Apêndice C. Os dados foram disponibilizados através de um arquivo *dump*, portanto para análise e consulta, foi necessária a importação deste *dump* em um banco de dados Postgres.

Todas as consultas a esta base de dados, foram realizadas por meio de SQL queries, utilizando-se a ferramenta pgAdmin III. Algumas das principais consultas SQL utilizadas, são apresentadas no Apêndice A. Para ter uma maior flexibilidade na manipulação e compreensão dos dados, as consultas com resultados de maior relevância foram exportadas em arquivos do tipo CSV, para serem trabalhadas em planilhas do programa Microsoft Excel.

Para uma melhor compreensão do cenário que este conjunto de dados representa, os dados foram representados em gráficos, apresentados no Apêndice B. Nas próximas seções estão descritos a extração e processamento dos dados seguindo as etapas do DCBD.

### 4.1 SELEÇÃO

Com a análise do banco de dados do AVA, foi verificada a existência de 228 tabelas de dados, destas, 208 são as tabelas padrão utilizadas pelo sistema Moodle, as 21 restantes armazenam informações personalizadas do NUTEAD. Além destas, foi criada uma tabela adicional para armazenar os dados relativos ao desempenho acadêmico dos alunos.

Dessas tabelas identificou-se quatro tabelas de maior relevância, apresentadas no Quadro 4, que contém os dados necessários sobre os cursos, disciplinas, alunos e *logs* de acesso.

Nome	Descrição
<i>ava_course_categories</i>	Contém o cadastro dos cursos e categorias de curso, é estruturada de tal maneira que permite o aninhamento de categorias para criação da estrutura dos cursos.
<i>ava_course</i>	Contém o cadastro das disciplinas, toda tupla está ligada a uma tupla da tabela <i>ava_course_categories</i> , ficando atrelada a um curso cadastrado. Possui informações sobre a disciplina, data de início, término e de cadastro.
<i>ava_user</i>	Contém o cadastro interno dos usuários do sistema Moodle, com informações pessoais de cada usuário, como por ex: nome, email, id, idacademico, polo, data do último acesso.
<i>ava_log</i>	Contém os registros das ações realizadas pelos usuários dentro do sistema, com informações como: data e horário, id do usuário, disciplina acessada, url acessada, tipo da ação realizada.

**Quadro 4 – Tabelas do banco de dados selecionadas para estudo**  
**Fonte: Autoria própria**

Cada tupla da tabela *ava\_course\_categories* contém o nome de uma categoria e o identificador de uma categoria pai, que é gravado na coluna *parent*, permitindo o aninhamento de categorias para a criação da estrutura dos cursos.

No Quadro 5, são listadas as categorias em que o identificador da categoria pai é nulo, isto é, que representam no AVA os tipos de cursos ofertados, todas as outras categorias cadastradas na tabela tem uma dessas como categoria pai.

Identificador	Categoria
1	EXTENSAO
2	GRADUACAO
6	POS-GRADUACAO

**Quadro 5 – Categorias de cursos criadas dentro do AVA**

**Fonte: Autoria própria**

No Quadro 6, são apresentadas as categorias que vem abaixo das categorias pai na estrutura criada dentro do sistema, denominadas aqui como subcategorias.

Identificador	Subcategoria
4	PROLICEN
5	UAB
22	INSTITUCIONAL
66	PARFOR
69	UAB
71	INSTITUCIONAL
72	AGENTES UNIVERSITÁRIOS UEPG (FAU)
125	OFERTA ESPECIAL
157	EDUCAÇÃO A DISTÂNCIA (PAFC)
308	EDUCAÇÃO A DISTÂNCIA (PAFC 2013)
314	PLANO INOVADOR DE CAPACITAÇÃO (PIC)

**Quadro 6 – Subcategorias de cursos cadastradas no AVA**

**Fonte: Autoria própria**

A tabela *ava\_log* contém o registro de todas as ações realizadas dentro do sistema pelos usuários, com esses registros é possível ter informações detalhadas do comportamento dos usuários, horários de login, quantidade de acessos, permanência dos usuários no site, quantidade de acessos ao fórum, questões realizadas no fórum, etc.

A tabela *ava\_course* contém os registros das disciplinas dos cursos, cada tupla desta tabela contém necessariamente um identificador de uma categoria cadastrada na tabela *ava\_course\_categories*, atrelando dessa maneira as disciplinas aos cursos disponíveis.

A maneira mais simples de encontrar os alunos vinculados às disciplinas é a partir da tabela *ava\_role\_assignments*, onde são armazenados os registros das disciplinas que os alunos estão cursando no momento, porém, como ao final de cada semestre os alunos passam a cursar novas disciplinas, as anteriores acabam por ser excluídas da tabela, impossibilitando a consulta através desse método às disciplinas que foram cursadas pelos alunos nos semestres anteriores.

Para encontrar os alunos que participaram de um curso, primeiramente foi necessário agrupar os *logs* dos alunos por disciplina, pois, na modelagem do banco de dados do Moodle, o aluno não é associado diretamente ao curso, e sim à disciplina. Tendo encontrado as disciplinas que os alunos cursaram, esses alunos foram agrupados pelos cursos disponíveis na tabela *ava\_course\_categories*.

No Quadro 7 são apresentados os cursos de graduação e pós-graduação e o número total de logs de acessos de cada curso, que foram encontrados com o processo descrito anteriormente. Os dois cursos com mais *logs* de acessos são: Bacharelado em Administração Pública e Licenciatura em Geografia.

<b>Tipo</b>	<b>Categoria</b>	<b>Nome</b>	<b>Total de logs</b>
Graduação	Oferta Especial	Licenciatura em Pedagogia	744083
Graduação	Prolicen	Licenciatura em Geografia	3012
Graduação	Prolicen	Licenciatura em História	1302
Graduação	Prolicen	Licenciatura em Letras Portugues/Espanhol	1460
Graduação	UAB	Bacharelado em Administração Pública	787200
Graduação	UAB	Licenciatura em Educação Física	614242
Graduação	UAB	Licenciatura em Geografia	783577
Graduação	UAB	Licenciatura em História	353154
Graduação	UAB	Licenciatura em Letras Portugues/Espanhol	398549
Graduação	UAB	Licenciatura em Matemática	342147
Graduação	UAB	Licenciatura em Pedagogia	251373
Pós-Graduação	Institucional	Especialização em Gestão Ambiental	16
Pós-Graduação	Institucional	Especialização em Gestão Pública (GPR)	113058
Pós-Graduação	Institucional	Mestrado em Odontologia	18
Pós-Graduação	UAB	Especialização em Gestão em Saúde (E1) - Plano de Recuperação de Estudos	23619
Pós-Graduação	UAB	Especialização em Educação Física Escolar	159027
Pós-Graduação	UAB	Especialização em Educação Matemática	188022
Pós-Graduação	UAB	Especialização em Gestão Educacional	59450
Pós-Graduação	UAB	Especialização em Gestão Pública (2013/2)	458
Pós-Graduação	UAB	Especialização em Gestão Pública Municipal (2013/2)	542
Pós-Graduação	UAB	Especialização em Gestão em Saúde (2013/2)	140584
Pós-Graduação	UAB	Especialização em História, Arte e Cultura (EHAC) (2013/2)	170536

**Quadro 7 – Quadro com os cursos e o total de logs de acesso por curso**

**Fonte: Autoria própria**

Dentre os cursos encontrados e suas informações, foi selecionado o curso de Bacharelado em Administração Pública, pois além de ter a maior quantidade de logs entre os cursos

disponíveis (787200 *logs* de acesso), é ofertado por várias instituições de ensino no Brasil, por meio do programa UAB-PNAP, e o mesmo havia sido finalizado recentemente, no segundo semestre de 2013, por esse motivo foram utilizadas as disciplinas do 8º semestre, pois além de serem recentes, eram as que mais possuíam dados de acesso.

No Quadro 8 são apresentadas as disciplinas selecionadas para estudo ofertadas no 8º semestre. Das outras disciplinas ofertadas neste semestre, foram excluídas do estudo as seguintes: Orientação de Trabalho de Conclusão de Curso, Estágio Supervisionado e Empreendedorismo Governamental, as duas primeiras por não serem teóricas e a última por ser uma disciplina não obrigatória.

<b>Disciplina</b>	<b>Alunos</b>	<b>Logs</b>
Comércio Internacional	271	13687
Gestão Ambiental e Sustentabilidade	319	52297
Gestão da Qualidade no Setor Público	313	108593
Marketing Governamental	277	16877
Políticas Públicas e Sociedade	311	43569
Relações Internacionais	297	50259

**Quadro 8 – Disciplinas do curso de Bacharelado em Administração Pública**

**Fonte: Autoria própria**

Com as disciplinas selecionadas, foi realizada a extração de dados dos logs, agrupando todos os dados dos acessos dos alunos nas disciplinas. Foi realizado o pré-processamento dos dados, afim de selecionar atributos e instancias relevantes para o trabalho. Este processo foi descrito na próxima seção.

## 4.2 PRÉ-PROCESSAMENTO

Com a base de dados gerada na etapa de seleção, apresentada na seção 4.1, na etapa de pré-processamento foram removidos os dados que não agregavam informações úteis para a análise, assim como dados redundantes, inconsistentes ou discrepantes no conjunto e também a recuperação de dados incompletos quando possível.

### 4.2.1 Limpeza

Como a quantidade de *logs* é grande, seria inviável trabalhar com cada *log* de ação realizado no sistema, para realizar o estudo foi necessário agregar todos os logs, gerando um atributo para representar o número de acessos e um atributo para cada ação contida nesses logs.

Para gerar o dado quantificador de acesso, foi feita a extração de todos os logs das interações realizadas pelos alunos, o atributo número de acesso foi gerado a partir do processamento desses logs, como a análise foi realizada por disciplina e o usuário pode ter logs no sistema sem necessariamente acessar os recursos das disciplinas estudadas, foi preciso agrupar esses dados por sessão realizada pelo usuário, considerando como uma nova sessão toda vez

que o usuário realiza login no sistema. Após o agrupamento, foram quantificadas as sessões onde o aluno acessou algum recurso de cada disciplina, com esses valores foi gerado para cada estudante o atributo *acessos*.

Cada log de acesso é gerado quando o aluno realiza alguma ação dentro do AVA, desde o login até a leitura de conteúdos, interações no fórum, envio de trabalhos, etc. Ao todo, nas pesquisas realizadas nos logs de acesso, foram encontradas 34 tipos de ações diferentes realizadas pelos alunos. No Quadro 9 são listadas as ações e uma breve descrição de cada uma delas.

<b>Ação</b>	<b>Descrição</b>
assignment_upload	Envio de tarefa
assignment_view	Visualização de tarefa
assignment_view_all	Visualização da lista de tarefas
course_user_report	Visualização das notas da disciplina
course_view	Visualização da disciplina
discussion_mark_read	Marcação de leitura em uma discussão no fórum
forum_add_discussion	Inclusão de discussão no fórum
forum_add_post	Inclusão de postagem no fórum
forum_delete_discussion	Exclusão de discussão no fórum
forum_delete_post	Exclusão de postagem no fórum
forum_search	Busca realizada no fórum
forum_subscribe	Inscrição em um fórum
forum_update_post	Atualização de postagem no fórum
forum_user_report	Visualização dos detalhes do usuário
forum_view_discussion	Visualização de discussão no fórum
forum_view_forum	Visualização da página do fórum
forum_view_forums	Visualização dos fóruns disponíveis
quiz_attempt	Tentativa de realização do questionário
quiz_close_attempt	Fechamento da realização do questionário
quiz_continue_attemp	Continuação da realização do questionário
quiz_review	Revisão do questionário
quiz_view	Visualização do questionário
quiz_view_all	Visualização de todos os questionários
resource_view	Visualização de recurso
resource_view_all	Visualização de todos os recursos disponíveis
survey_submit	Submissão da pesquisa
survey_view_all	Visualização de todas as pesquisas disponíveis
survey_view_form	Visualização da pesquisa
survey_view_graph	Visualização dos resultados da pesquisa
upload_upload	Envio de arquivo
user_change_password	Alteração de senha do usuário
user_update	Atualização de cadastro de usuário
user_view	Visualização de usuário
user_view_all	Visualização da lista de usuários

**Quadro 9 – Ações realizadas pelos usuários**

**Fonte: Autoria própria**



Para representar cada ação, foi gerado um atributo contendo o número de vezes que foi realizada, sendo que, para os alunos que não realizaram alguma ação específica o atributo foi zerado. Como cada ação é realizada dentro de um módulo do sistema, foi adicionado o nome do módulo ao atributo.

Para realizar esse estudo, foi decidido que as disciplinas deveriam conter os mesmos alunos. Das seis disciplinas selecionadas, foram excluídas duas, Comércio Internacional e Marketing Governamental, que possuíam um menor número de *logs*, assim como um menor número de alunos em comum com as outras quatro disciplinas restantes. A disciplina Políticas Públicas, também foi excluída na sequencia, apesar de possuir os mesmos alunos em comum com as outras disciplinas, nessa disciplina, os alunos tiveram um agrupamento por turma e polo diferente do padrão encontrado nas outras 3. Após a remoção dessas três disciplinas e dos alunos que não constavam nelas, foram geradas três bases de dados no formato de arquivo ARFF, uma por disciplina, para permitir o estudo em separado.

#### 4.2.2 Seleção de atributos

As bases geradas com os dados de cada disciplina possuem 237 instâncias, cada instância representa os dados de um estudante e contém 42 atributos, 5 nominais (*polo*, *sexo*, *grau\_graduacao*, *exame*, *status*) e outros 37 atributos numéricos, que representam os quantificadores de acessos, diferentes ações realizadas e a média do aluno na disciplina.

Os atributos nominais *polo*, *sexo* e *grau\_graduacao* possuem os mesmos valores para as três bases de dados geradas, para os outros atributos os valores são distintos.

O atributo *polo* indica o polo no qual o aluno está matriculado, no conjunto de dados existem 10 polos distintos: Cerro Azul, Congonhinhas, Faxinal, Ibaiti, Ipiranga, Jacarezinho, Jaguariaíva, Jaú, Palmeira e Ponta Grossa. O polo com o maior número de estudantes é o de Ponta Grossa, com 65 estudantes, e o com menor número é o de Congonhinhas com 11 estudantes. A Figura 7 apresenta a distribuição de alunos entre os polos.

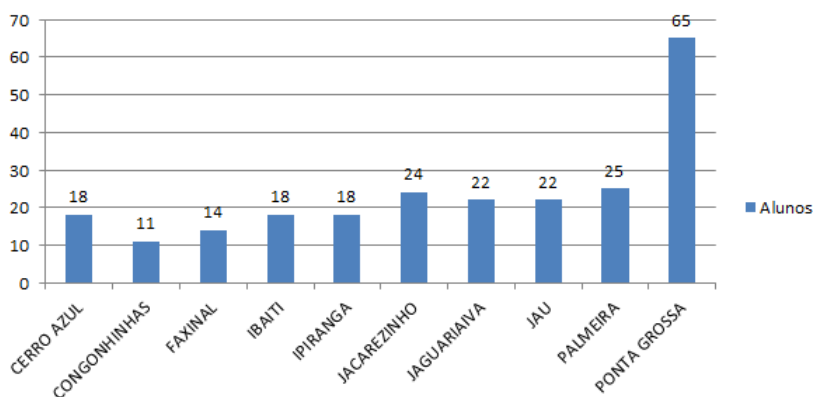


Figura 7 – Distribuição de alunos por Polo

Fonte: Autoria própria

O atributo *grau\_graduacao* representa o nível máximo de graduação que o aluno possui,

para esse atributo existem somente dois valores distintos na base de dados: 2 (indica que o aluno possui o Ensino Médio completo), 3 (indica que o aluno já possui algum curso de Graduação completo).

O atributo *sexo* representa o sexo do aluno, na base, 133 alunos são do sexo feminino e 104 alunos são do sexo masculino. Na base de dados o atributo pode conter os valores F e M.

O atributo *exame* representa se o aluno ficou em exame na disciplina, todos os alunos que possuem notas inferiores à 7.0 e superiores que 3.0, devem realizar o exame, que consiste em uma avaliação final para recuperação de nota. Os alunos que, com a nota do exame ficarem com notas inferiores à 5.0 são reprovados, os que obterem nota igual ou acima de 5.0 são aprovados. O atributo pode conter o valor 1 (realizou exame) e 0 (não realizou exame).

O atributo *status* representa o status final do aluno na disciplina, os alunos reprovados são representados na base pelo valor R e os aprovados pelo valor A. Esse será inicialmente o atributo utilizado como classe nos testes realizados.

O atributo *media* foi mantido na base de dados para geração de novos atributos derivados, porém, por ser um atributo óbvio de predição de classes, não foi utilizado para classificação dos alunos.

O atributo idade mostra que os estudantes tinham idades entre 21 e 62 anos no período o qual foram disponibilizadas as disciplinas no AVA. A média de idade é de aproximadamente 36 anos. Este atributo foi criado a partir da data de nascimento, esse processo foi descrito na etapa de transformação.

Como o estudo foi realizado com três disciplinas distintas, a seleção de atributos foi feita separadamente para cada disciplina, pois cada uma contém informações específicas, e revelam diferentes padrões de comportamento dos alunos em relação ao AVA naquela disciplina.

#### 4.2.2.1 Seleção de atributos da disciplina de Relações Internacionais

Como não havia um especialista no AVA para auxílio na seleção de atributos, foi preciso utilizar o algoritmo Ranker, apresentado na seção 3.1.2, para essa tarefa.

Utilizando a base de dados da disciplina Relações Internacionais(RI), aplicou-se o algoritmo citado. Na Figura 8 é apresentado o resultado do algoritmo Ranker aplicado na base de dados dessa disciplina.

O atributo *acessos* foi considerado pelo algoritmo o atributo de maior relevância. Para esta disciplina esse atributo tem o valor mínimo 1 e máximo de 96, e sua média é de 26,5 acessos. Analisando o atributo na aba Preprocess do Weka, notou-se que, 134 alunos realizaram mais de 20 acessos na disciplina, sendo que desses, nenhum foi reprovado. Isso indica que os alunos que realizaram mais acessos, ou seja, tiveram mais interesse pela disciplina, conseguiram ser aprovados com mais facilidade e os 10 alunos reprovados dessa disciplina estão entre os estudantes que realizaram 20 ou menos acessos. Com essa informação pode-se dizer que o número de acessos representa um fator importante para a análise do comportamento dos alunos.

Dos atributos relevantes selecionados pelo algoritmo Ranker, 11 atributos representa-

vam ações dos usuários dentro da disciplina, esses atributos, excluindo-se o atributo *course\_view*, que representa o número de visualizações da página da disciplina, podem ser agrupados em três grupos: ações relacionadas ao fórum da disciplina, ações relacionadas com a realização dos questionários da disciplina, e ações relacionadas com a realização dos trabalhos da disciplina.

```

=== Attribute Selection on all input data ===

Search Method:
    Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 41 status):
    Information Gain Ranking Filter

Ranked attributes:
0.1183      6 acessos
0.107568   25 course_view
0.077113   23 assignment_view
0.068262   20 quiz_close_attempt
0.068262   12 quiz_attempt
0.05835    14 upload_upload
0.05835    40 assignment_upload
0.055269   15 forum_add_discussion
0.054533    5 exame
0.049449   26 forum_view_discussion
0.049448   22 quiz_continue_attemp
0.046      10 forum_view_forum
0.044138   18 forum_add_post
0.031854    1 polo
0.00048     2 sexo
0.000444    4 grau_graduacao

```

**Figura 8 – Resultado do algoritmo *Ranker* nos dados da disciplina de RI**

**Fonte: Autoria própria**

No Quadro 10 observa-se que os atributos que foram considerados de mais peso pelo algoritmo são os do grupo dos questionários, seguido pelo grupo ações dos trabalhos e por fim, dos do fórum.

Grupo ações questionários		Grupo ações trabalhos		Grupo ações fórum	
Ação	Peso Ranker	Ação	Peso Ranker	Ação	Peso Ranker
quiz_close_attempt	0,068	assignment_view	0,077	forum_add_discussion	0,055
quiz_attempt	0,068	assignment_upload	0,058	forum_view_discussion	0,049
quiz_continue_attemp	0,049	upload_upload	0,058	forum_view_forum	0,046
				forum_add_post	0,044

**Quadro 10 – Resultados do algoritmo *Ranker* agrupados pelo tipo da ação**

**Fonte: Autoria própria**

Para essa disciplina, com base nos resultados obtidos com o algoritmo Ranker, nota-se que o que mais influência na classificação dos estudantes são os questionários realizados, representados pelos atributos *quiz\_close\_attemp*, *quiz\_attempt* e *quiz\_continue\_attempt*. A seguir são descritos os atributos que representam as ações relacionadas com a realização dos questionários da disciplina.

O atributo que teve o maior peso atribuído é *quiz\_close\_attempt*, como mostrado no Quadro 11, esse atributo representa o número de vezes que o usuário finalizou um questionário na disciplina. O atributo *quiz\_attempt* representa o número de vezes que o estudante iniciou a realização de um questionário, e o atributo *quiz\_continue\_attempt* representa o número de vezes que o estudante continuou um questionário iniciado anteriormente.

Atributo	Min	Max	Média
<i>quiz_close_attempt</i>	0	11	5,56
<i>quiz_attempt</i>	0	9	5,54
<i>quiz_continue_attempt</i>	0	35	8,71

**Quadro 11 – Valores dos atributos relacionados a realização dos questionários**

Ambos os atributos *quiz\_close\_attempt* e *quiz\_attempt* possuem valores parecidos, pois quando o estudante abre um questionário pra realização ele também realiza a conclusão deste questionário, somente o valor máximo está discrepante, isso pode indicar que o professor reabriu algum questionário para os alunos responderem. A maioria dos alunos desta disciplina (203 alunos) realizaram entre 3 e 7 questionários (*quiz\_attempt*), sendo a maioria (199) dos que concluíram os questionários (*quiz\_close\_attempt*) também estão nessa faixa entre 3 e 7 questionários. A maioria dos reprovados (5) teve entre 0 e 1 questionário finalizado.

Entre os 4 estudantes que não finalizaram nenhum questionário, um foi aprovado na disciplina, analisando os outros atributos da instância que representa esse estudante (instância 23), tem-se que o aluno realizou poucos acessos na disciplina (7 acessos), e não realizou nenhum envio de trabalhos. O valor (*status* aprovado) pode ter sido lançado erroneamente pelo professor, ou o aluno pode ter entregue algum trabalho ou questionário para o professor fora do AVA. Como a análise dos alunos está sendo realizada somente por meio dos atributos presentes nos logs do ambiente, essa instância foi considerada como ruído por apresentar valores discrepantes, e foi excluída da base de dados para a próxima etapa do processo.

Dos atributos relacionados aos trabalhos realizados pelos estudantes da disciplina, assim como os atributos *quiz\_attempt* e *quiz\_continue\_attempt* do grupo de questionários, os atributos *assignment\_upload* e *upload\_upload* tiveram o mesmo peso atribuído pelo algoritmo Ranker, a ação *assignment\_upload* representa o acesso à página de upload de um trabalho pelo estudante, já a ação *upload\_upload* acontece quando o usuário efetivamente faz o envio do arquivo do trabalho, já o atributo *assignment\_view* representa o número de acessos às páginas dos trabalhos disponibilizados na disciplina.

No Quadro 12 são apresentados os valores mínimo e máximo, e média dos atributos relacionados à execução dos trabalhos da disciplina. Dos 10 alunos reprovados nessa disciplina, 7 não realizaram o envio de nenhum arquivo, os outros 3 reprovados realizaram um envio de arquivo. Outros 28 estudantes não realizaram o envio de nenhum trabalho, porém, todos eles (exceto o aluno representado pela instância 23, já discutido anteriormente) finalizaram pelo menos um questionário da disciplina.

Esse comportamento reforça o resultado do algoritmo Ranker, pois mostra que, mesmo o aluno não enviando nenhum trabalho, porém realizar ao menos um questionário, ainda pode ser aprovado na disciplina. Dos alunos que visualizaram as páginas dos trabalhos, a maioria (125 alunos) teve entre 15 e 38 visualizações dos trabalhos, todos os alunos que acessaram mais de 15 vezes as páginas dos trabalhos foram aprovados na disciplina.

Atributo	Min	Max	Média
upload_upload	0	10	1,85
assignment_upload	0	10	1,83
assignment_view	0	100	20,22

**Quadro 12 – Valores dos atributos relacionados a realização dos trabalhos**

Fonte: Autoria própria

#### 4.2.2.2 Seleção de atributos da disciplina de Gestão Ambiental

A disciplina de Gestão Ambiental(GA) possui os mesmos atributos que a disciplina anterior, porém essa disciplina tem 12 reprovados, contra 10 alunos na disciplina RI. Aplicando o algoritmo Ranker na base de dados dessa disciplina foi gerado o resultado mostrado na Figura 9.

```

=== Attribute Selection on all input data ===

Search Method:
  Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 41 status):
  Information Gain Ranking Filter

Ranked attributes:
0.04582  18 forum_add_post
0.0457   14 upload_upload
0.04388  40 assignment_upload
0.03926   1 polo
0.03466  12 quiz_attempt
0.03237  15 forum_add_discussion
0.03227  20 quiz_close_attempt
0.01452   5 exame
0.01126   4 grau_graduacao
0.0059    2 sexo

```

**Figura 9 – Resultado do algoritmo Ranker nos dados da disciplina de GA**

Fonte: Autoria própria

Observa-se que o algoritmo Ranker selecionou menos atributos em comparação com a base de dados de RI, e que os pesos dos atributos selecionados são menores. Porém nessa disciplina o atributo *forum\_add\_post*, relacionado ao fórum, foi considerado o mais relevante pelo algoritmo. Esse atributo representa o número de postagens realizadas no fórum pelos alunos, a maioria dos alunos dessa disciplina (135 alunos) realizaram somente 2 postagens no fórum,

com o mínimo de postagens sendo 0 postagens e o máximo 9, com média de postagens de 1,916. Como o atributo *forum\_add\_post* não possui nenhuma particularidade em seus valores que possa ser relevante, foi executado novamente o algoritmo Ranker na base de dados, utilizando dessa vez o filtro avaliador de atributos *GainRatioAttributeEval*, ao invés do filtro padrão *InfoGainAttributeEval*, gerando o resultado apresentado na Figura 10.

```

=== Attribute Selection on all input data ===

Search Method:
    Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 41 status):
    Gain Ratio feature evaluator

Ranked attributes:
0.1279    12 quiz_attempt
0.11163   20 quiz_close_attempt
0.10584   14 upload_upload
0.09843   40 assignment_upload
0.07755   15 forum_add_discussion
0.05808   18 forum_add_post
0.02286    5 exam
0.01934    4 grau_graduacao
0.01259    1 polo
0.00596    2 sexo

```

**Figura 10 – Resultado do Ranker utilizando *GainRatioAttributeEval***

**Fonte: Autoria própria**

Na segunda execução chegou-se a um resultado mais parecido com o da disciplina de Relações Internacionais, porém com um menor número de atributos. Com o filtro *GainRatioAttributeEval* foram obtidos atributos com peso maior na seleção de atributos, comparado com o filtro *InfoGainAttributeEval*. Os atributos seguem a mesma ordem de relevância apresentado na disciplina anterior, com os atributos relacionados aos questionários com mais peso, seguido dos atributos dos trabalhos e por fim dois atributos referentes à utilização do fórum.

Na Tabela 1 são apresentados os valores dos atributos relacionados às ações realizadas pelos alunos no ambiente. Na média, os alunos na disciplina de GA fizeram menos questionários, e realizaram menos envios de trabalhos, do que na disciplina RI.

**Tabela 1 – Valores dos atributos da base Gestão Ambiental**

Atributo	Min	Max	Média
quiz_attempt	0	7	4,37
quiz_close_attempt	0	7	4,34
upload_upload	0	4	1,19
assignment_upload	0	5	1,16
forum_add_discussion	0	4	0,99
forum_add_post	0	9	1,92

**Fonte: Autoria própria**

Analisando os valores das finalizações dos questionários, tem-se que, a maioria dos estudantes (220 alunos) finalizaram entre 3 e 6 questionários na disciplina. Entre os 3 alunos que

não realizaram nenhuma finalização de questionário (*quiz\_close\_attempt*), dois foram aprovados e um reprovado. Dos dois aprovados, um deles não havia realizado nenhum envio de tarefas (instância 232), e teve um número pequeno de acessos (4 acessos), foi realizada a exclusão dessa instância da base de dados, seguindo o mesmo critério aplicado na disciplina anterior.

Para o envio de trabalhos, contrastando com a disciplina RI, onde os alunos que realizaram pelo menos dois envios foram aprovados, nessa disciplina não existem um padrão aparente, sendo que existem alunos reprovados que realizaram um, dois ou quatro envios, sendo o máximo cinco envios na disciplina. A maioria (165 estudantes) realizou somente um envio de arquivo.

#### 4.2.2.3 Seleção de atributos da disciplina de Gestão da Qualidade

A disciplina de Gestão da Qualidade (GQ) é a última disciplina selecionada a ser estudada, ela possui, assim como a disciplina Gestão Ambiental, 12 alunos reprovados, e 225 alunos aprovados.

Para essa disciplina, o resultado do algoritmo, mostrado na Figura 11, mantém o padrão de atributos selecionados nas outras duas disciplinas, porém, invertendo-se a ordem das ações dos trabalhos (*assignment\_upload*, *upload\_upload*), com as ações dos questionários (*quiz\_view*, *quiz\_close\_attempt*, *quiz\_view\_all*).

```
=== Attribute Selection on all input data ===

Search Method:
  Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 41 status):
  Information Gain Ranking Filter

Ranked attributes:
0.114208  40 assignment_upload
0.114208  14 upload_upload
0.093508  23 assignment_view
0.072066   6 acessos
0.05662   25 course_view
0.044138  24 quiz_view
0.037343  20 quiz_close_attempt
0.032114   1 polo
0.021568   5 exame
0.018408  32 survey_view_form
0.018408  31 quiz_view_all
0.011263   4 grau_graduacao
0.000581   2 sexo
```

Figura 11 – Avaliação do algoritmo *Ranker* para a base de GQ

Fonte: Autoria própria

Na seleção de atributos dessa disciplina, tem-se um atributo que não foi selecionado nas outras duas disciplinas, o atributo *survey\_view\_form*, que representa o número de visualizações

do formulário de pesquisa de opinião da disciplina. A pesquisa de opinião (do inglês *survey*), é uma ferramenta disponibilizada aos professores para criação de questionários, a fim de obter *feedback* dos estudantes sobre o ensino, disciplina, recursos ou outro assunto relevante que os professores desejem obter opiniões dos alunos e usuários.

Observando os valores mínimo, máximo e média dos atributos selecionados, apresentados na Tabela 2, nota-se que, enquanto nas disciplinas RI e GA, tem-se uma média aproximada de realização de questionários de 5,56 e 4,34, e envio de trabalhos de 1,85 e 1,19 respectivamente, na disciplina de GQ a média de envio de trabalhos é de 17,45 e realização de questionários é de 0,31, esses valores mostram que, além dessa disciplina ter uma quantidade maior de interações dos alunos em relação às outras duas, sua principal forma de avaliação está na aplicação de trabalhos, e não nos questionários aplicados.

**Tabela 2 – Atributos da disciplina de GQ selecionados pelo algoritmo Ranker**

Atributo	Min	Max	Média
assignment_upload	0	48	17,42
upload_upload	0	48	17,45
assignment_view	2	703	138,14
course_view	3	334	68,24
quiz_view	0	22	0,97
quiz_close_attempt	0	2	0,31
survey_view_form	0	2	0,10
quiz_view_all	0	4	0,03

**Fonte: Autoria própria**

A maioria (176 alunos) realizaram entre 12 e 24 envios de arquivos, sendo que os alunos reprovados estão entre os que realizaram menos de 20 envios. Entre os alunos que não realizaram nenhum envio de trabalho, 4 alunos foram aprovados, porém três deles têm uma finalização de questionário, o único que não realizou nem envio de trabalhos nem finalização de questionários foi o da instância 232, mesma instância com ruído da disciplina de GA, devido a isso, excluímos também da base de GQ essa instância.

Na comparação entre as disciplinas, a partir dos resultados apresentados na Tabela 3, observa-se que, enquanto nas disciplinas de GA e RI existe uma maior quantidade de questionários realizados em relação aos envios de trabalhos, na disciplina de GQ observa-se o comportamento inverso.

**Tabela 3 – Comparação entre os atributos das disciplinas**

	RI	GA	GQ
Média Acessos	26,58	28,92	40,93
Média Envios Trabalhos	1,86	1,19	17,52
Média Questionários Realizados	5,59	4,36	0,31
Alunos com exame	21	38	72
Alunos reprovados	10	12	12
Alunos aprovados com exame	15	33	65

**Fonte: Autoria própria**

Na disciplina de GQ observa-se uma média de acessos maior em relação às outras



duas. A disciplina de GQ possui uma maior média de acessos em relação às outras duas. Nessa disciplina também encontra-se o maior número de alunos que precisaram realizar o exame final.

Os atributos das três bases de dados que não foram selecionados pelo algoritmo Ranker não foram excluídos das bases, pois têm influência nos resultados dos algoritmos de mineração de dados, aumentando as taxas de acerto desses algoritmos.

A etapa de seleção de atributos com a utilização do algoritmo Ranker, foi realizada a fim de obter-se uma melhor compreensão das bases de dados das disciplinas e de seus atributos.

Na etapa seguinte é mostrada como foi realizada a transformação de atributos, para realizar o ajuste das bases de dados aos algoritmos que foram aplicados na etapa de Mineração de Dados.

### 4.3 TRANSFORMAÇÃO

Na etapa de transformação foi gerado o atributo *categoria\_nota*, para identificar possíveis padrões dos alunos que tiram notas boas e ruins, e para comparar os resultados com a classificação da classe *status* do aluno.

Para a geração do novo atributo foi utilizado o atributo *media* presentes nas bases de dados. Foram criadas 3 categorias diferentes para representar as notas ruins, regulares e boas. Para os alunos que tiveram médias finais inferiores à 5,0, foi atribuído o valor RUIM para o atributo *categoria\_nota*, os alunos que tiveram notas entre 5,0 e 7,0 exclusive tiveram o valor REGULAR atribuído, e, por fim, os alunos que tiraram notas iguais ou superiores que 7,0 tiveram o valor BOA atribuído.

Também foi gerado o atributo *idade*, a partir da data de nascimento dos alunos, para isso foi considerado o ano que os alunos estavam cursando as matérias selecionadas (2013), assim subtraindo-se o ano de nascimento do ano 2013, obteve-se as idades dos alunos.

Nesta etapa, foram transformadas as bases de dados que estavam sendo manipuladas em arquivos do formato .CSV para arquivos .ARFF, formato utilizado pela ferramenta de mineração WEKA, através da ferramenta EXCEL. Tais ferramentas foram descritas na seção 3.3.

### 4.4 MINERAÇÃO DOS DADOS

Nesta etapa foram realizados dois experimentos com as bases de dados das disciplinas, utilizando-se os algoritmos de classificação J48, NaiveBayes e MultilayerPerceptron, presentes no Weka. O primeiro experimento foi a aplicação dos algoritmos para classificação da classe *status*. No segundo experimento foi realizada a classificação da classe obtida a partir da discretização do atributo *media*, realizada na etapa anterior.

Ao final foi feita a comparação dos resultados dos dois experimentos, e a comparação do desempenho dos algoritmos utilizados.

#### 4.4.1 Primeiro experimento

O primeiro experimento utiliza o atributo *status* para classificação, que representa a aprovação ou reprovação do aluno na disciplina. Executando primeiramente o algoritmo J48 com suas configurações padrão, com o parâmetro de poda da árvore setado em 0,25 (poda moderada), analisando os resultados mostrados na tabela 4, nota-se que o algoritmo obteve os mesmos números de instâncias classificadas corretamente e incorretamente para todas as três disciplinas, classificando 226 corretamente e 10 incorretamente.

Na classificação de aprovados tem-se uma alta taxa de acerto, porém, a classificação dos reprovados teve uma taxa de acerto menor, isso se deve ao fato de que, na base existem poucas instâncias de alunos reprovados, aproximadamente 5% do número de instâncias. A disciplina que apresentou melhor valor de *Kappa statistic* foi a de GQ, conseguindo classificar 50% dos reprovados corretamente.

**Tabela 4 – Resultados da primeira execução do algoritmo J48**

Disciplina	Kappa statistic	Aprovados		Reprovados	
		TP Rate	FP Rate	TP Rate	FP Rate
RI	0,48	0,98	0,50	0,50	0,02
GA	0,42	0,99	0,67	0,34	0,01
GQ	0,52	0,98	0,50	0,50	0,02

**Fonte: Autoria própria**

O atributo de envio de trabalhos (*upload\_upload*), foi considerado pelo algoritmo Ranker como um dos mais relevantes, como mostrado na seção 4.2.2.3, isso se confirma na árvore de decisão da disciplina, onde aparece como a raiz da árvore.

Comparando as árvores de decisão geradas para as três disciplinas, apresentadas na Figura 12, constata-se que, para a disciplina GQ, se um aluno tiver mais que 5 envios de trabalhos ele consegue ser aprovado. Já para os reprovados, a regra "se envio de trabalhos for menor ou igual a 5, e finalização de questionário igual a 0 então reprovado", se aplicou a 5 instâncias das 12 de reprovados presentes na base.

Para as disciplinas RI e GA, as árvores geradas são maiores que a da disciplina GQ, evidenciado pelo valor do número de folhas (Number of Leaves), nas duas árvores são utilizados um número maior de atributos do que na de GQ, nota-se que nas duas disciplinas o atributo *exame* foi utilizado pelo algoritmo.

Na disciplina RI, nota-se dois comportamentos distintos entre os alunos, os que realizaram exame e os que não o fizeram. Entre os estudantes que não realizaram o exame final, os que tiveram mais de 4 visualizações dos trabalhos foram aprovados (204 instâncias), entre os estudantes que visualizaram 4 ou menos vezes os trabalhos, os que fizeram algum questionário também foram aprovados (7 instâncias), os que não fizeram nenhum questionário foram reprovados (4 instâncias), esta regra é semelhante à regra encontrada na disciplina GQ, porém, em RI é considerada a variável de visualização dos trabalho *assignment\_view*, e na árvore de GQ

foi considerada a variável de envio (*upload\_upload*). Entre os alunos que fizeram exame, tem-se que, os que tiveram mais de 15 acessos na disciplina conseguiram aprovação (11 instâncias),

<pre> exame = 0   assignment_view &lt;= 4     quiz_close_attempt &lt;= 1: R (4.0)     quiz_close_attempt &gt; 1: A (7.0)   assignment_view &gt; 4: A (204.0) exame = 1   acessos &lt;= 15     quiz_continue_attemp &lt;= 3: A (2.0)     quiz_continue_attemp &gt; 3       sexo = M: R (3.0)       sexo = F         acessos &lt;= 10: R (3.0)         acessos &gt; 10: A (2.0)           acessos &gt; 15: A (11.0)  Number of Leaves :      8 Size of the tree :     15         </pre>	<pre> upload_upload &lt;= 0   quiz_attempt &lt;= 2: R (4.0)   quiz_attempt &gt; 2     course_user_report &lt;= 17: A (14.0)     course_user_report &gt; 17: R (2.0) upload_upload &gt; 0   exame = 0: A (184.0/1.0)   exame = 1     survey_view_form &lt;= 0         acessos &lt;= 21           acessos &lt;= 20: A (4.0)           acessos &gt; 20: R (3.0)           acessos &gt; 21: A (21.0)           survey_view_form &gt; 0             assignment_view_all &lt;= 1: A (2.0)             assignment_view_all &gt; 1: R (2.0)  Number of Leaves :      9 Size of the tree :     17         </pre>
---	---

(a) Árvore de RI

(b) Árvore de GA

```

upload_upload <= 5
| quiz_close_attempt <= 0: R (5.0)
| quiz_close_attempt > 0
| | sexo = M: R (2.0)
| | sexo = F: A (5.0/1.0)
upload_upload > 5: A (224.0/4.0)

Number of Leaves :      4

Size of the tree :      7
        
```

(c) Árvore de GQ

Figura 12 – Árvores de decisão geradas pelo algoritmo J48

Fonte: Autoria própria

Na árvore de decisão da disciplina GA, assim como na de RI, existem duas ramificações aparentes, os alunos que realizaram o envio de trabalhos e os que não enviaram. Entre os que não enviaram trabalhos, os estudantes que fizeram dois ou menos questionários foram reprovados (4 instâncias), os que fizeram mais que dois questionários e tiveram 17 ou menos visualizações da página com as notas da disciplina foram aprovados(14 instâncias).

Dos alunos que realizaram o envio de arquivos, os que não ficaram para exame foram aprovados (184 instâncias), e entre os que ficaram para exame e tiveram mais que 21 acessos foram aprovados (21 instâncias).

Na segunda execução do algoritmo J48, Tabela 5, foram alterados os parâmetros confidenceFactor (grau de confiança) e minNumObj(número mínimo de instâncias por folha da árvore), para o primeiro parâmetro, foram testados os valores 0,1 (poda alta), 0,25 (padrão -

poda moderada) e 0,5 (poda baixa), para o segundo, os valores 2 (padrão), 3 e 4, combinando-se os valores dos parâmetros para cada execução.

**Tabela 5 – Resultados do algoritmo J48 com diferentes parâmetros**

Teste	ConfidenceFactor	minNumObj	RI		GA		GQ	
			Taxa de acerto	Kappa	Taxa de acerto	Kappa	Taxa de acerto	Kappa
1	0,10	2	95,34	0,40	96,61	0,487	95,339	0,45
2	0,25	2	95,76	0,48	95,76	0,42	95,76	0,52
3	0,50	2	95,34	0,40	94,92	0,37	95,76	0,52
4	0,10	3	96,19	0,45	96,92	0	95,34	0,40
5	0,25	3	96,61	0,58	95,76	0,42	96,19	0,59
6	0,50	3	95,76	0,42	95,76	0,42	96,19	0,59

Fonte: Autoria própria

Para ambas as disciplinas obteve-se melhores resultados alterando-se os valores padrão dos atributos do algoritmo, aumentando-se o valor da variável de objetos mínimos por folha no caso das disciplinas RI e GQ e diminuindo o valor do grau de confiança para a disciplina GA.

Das árvores de decisão geradas na segunda execução, a que teve mais alterações em relação as árvores apresentadas para a primeira execução, foi a árvore da disciplina GA, apresentada na Figura 13. Para essa execução foi gerada uma árvore menor, com somente 4 folhas, sendo que a primeira possui 9 regras. Na segunda execução, para essa disciplina, o algoritmo realizou a classificação correta de todos os alunos aprovados, porém, entre os 12 reprovados, conseguiu a classificação correta somente em 4 alunos.

```

upload_upload <= 0
|   quiz_attempt <= 2: R (4.0)
|   quiz_attempt > 2
|   |   course_user_report <= 17: A (14.0)
|   |   course_user_report > 17: R (2.0)
upload_upload > 0: A (216.0/6.0)

Number of Leaves :      4

Size of the tree :      7

```

**Figura 13 – Árvore de decisão para a disciplina GA**

Fonte: Autoria própria

Para a obtenção de um parâmetro de comparação, a fim de demonstrar a eficácia do algoritmo J48 na execução dos experimentos, foi realizada a classificação utilizando os algoritmos NaiveBayes, MultilayerPerceptron e ZeroR, e comparou-se os resultados destes com os resultados apresentados pelo algoritmo J48, para isso foi utilizada a ferramenta Experimenter presente no Weka, o resultado desse experimento é mostrado na Tabela 6.

Conforme os resultados do experimento, nota-se entre os algoritmos testados que, em média, o J48 com a configuração dos parâmetros *confidenceFactor* com valor 0,25 e *minNunObj* com valor 3 foi o algoritmo com melhor taxa de acerto, tendo um desempenho menor do que a configuração padrão somente para a disciplina GQ. O algoritmo MultilayerPerceptron obteve valores semelhantes aos do J48, inclusive tendo desempenho melhor para a disciplina de RI. O

algoritmo NaiveBayes teve o menor desempenho entre os algoritmos testados, ficando abaixo do desempenho do algoritmo ZeroR.

**Tabela 6 – Porcentagem de acerto dos algoritmos**

Base	(1)	(2)	(3)	(6)	(8)	(10)
RI	95.76	94.95	94.95	95.37	83.11	95.59
GA	94.93	96.00	94.80	95.13	79.88	94.58
GQ	94.93	95.25	96.45	96.28	80.78	95.68
Média	95.20	95.41	95.40	95.59	81.26	95.28

**Fonte: Autoria própria**

- (1) rules.ZeroR
- (2) trees.J48 '-C 0.1 -M 2'
- (3) trees.J48 '-C 0.25 -M 2' - Configuração padrão
- (6) trees.J48 '-C 0.25 -M 3'
- (8) bayes.NaiveBayes
- (10) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a' - Configuração padrão

#### 4.4.2 Segundo experimento

No segundo experimento, foram comparados os resultados da classificação dos alunos por status, com outras duas bases geradas para cada disciplina. Na primeira base foi utilizado o atributo *categoria\_nota*, gerado na etapa de transformação, mostrada na seção 4.3, como classe. Na segunda base foi utilizado o atributo nominal *exame*. O objetivo desse experimento é realizar uma comparação entre os métodos de classificação por status, nota e exame.

Utilizando a ferramenta Experimenter, foram executados os algoritmos J48, Naive-Bayes e MultilayerPerceptron e OneR. Comparando os resultados da classificação, apresentados na Tabela 7, nota-se que os algoritmos tiveram melhores resultados de classificação nas bases que utilizam o atributo *categoria\_nota* como classe. A classificação pelo atributo *exame* não apresentou resultados satisfatórios, pois, somente na disciplina GQ foi possível a classificação por essa classe.

**Tabela 7 – Comparação do Kappa Statistic da classificação das bases**

Base	(1)	(2)	(3)	(4)	(5)
RI-status	0.00	0.22	0.41	0.25	0.33
GA-status	-0.01	0.17	0.39	0.28	0.25
GQ-status	0.44	0.19	0.42	0.46	0.52
RI-nota	0.67	0.17	0.66	0.63	0.68
GA-nota	0.71	0.32	0.71	0.66	0.66
GQ-nota	0.69	0.55	0.70	0.69	0.70
RI-exame	-0.01	0.11	0.07	0.00	0.00
GA-exame	0.18	-0.02	0.20	0.00	-0.01
GQ-exame	0.96	0.94	0.90	0.96	0.96

**Fonte: Autoria própria**

- (1) rules.OneR '-B 6'
- (2) bayes.NaiveBayes ''
- (3) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a'
- (4) trees.J48 '-C 0.2 -M 2'
- (5) trees.J48 '-C 0.25 -M 3'

Na árvore de decisão gerada para a base GQ-exame, apresentada na Figura 14, observa-se que os alunos que finalizaram algum questionário realizaram exame na disciplina. Na árvore com a classificação de aprovados e reprovados para essa disciplina, apresentada no primeiro experimento, mostra-se que, os alunos que realizaram 5 ou menos envios de trabalhos e não realizaram nenhum questionário eram reprovados, analisando as duas árvores em conjunto, pode-se concluir que os alunos somente precisaram realizar o questionário de exame da disciplina.

A classificação da categoria da nota apresentou resultados melhores que a classificação do *status* do aluno, mostrando que é possível realizar a classificação dos alunos tanto por sua aprovação ou reprovação, quanto pela categoria da nota. Os resultados apresentados mostram também que, ambos os algoritmos J48 e MultilayerPerceptron tiveram resultados satisfatórios para as bases de dados testadas, ambos obtendo valores de Kappa Statistic acima de 0,6 para as bases com a classe *categoria\_nota*, e maiores que 0,4 nos melhores resultados para as bases com a classe *status*.

```

quiz_close_attempt <= 0: 0 (165.0/2.0)
quiz_close_attempt > 0: 1 (72.0/2.0)

Number of Leaves :      2
Size of the tree :      3

```

**Figura 14 – Árvore de decisão para a base GQ-exame**  
**Fonte: Autoria própria**

No primeiro experimento foi identificado que, a realização de questionários na disciplina de RI, foi mais importante para a aprovação. Na Figura 15, observa-se que o mesmo comportamento para os alunos que conseguiram nota boa na disciplina. Esse comportamento, tanto na classificação do status aprovado ou reprovado, quanto na categoria nota do aluno, mostra que realmente a realização dos questionários foi o principal meio de avaliação utilizado nessa disciplina.

```

exame = 0
| quiz_attempt <= 3
| | forum_view_discussion <= 0: RUIM (4.0)
| | forum_view_discussion > 0: BOA (4.0)
| quiz_attempt > 3: BOA (208.0)
exame = 1
| quiz_review <= 3
| | course_view <= 23
| | | quiz_continue_attemp <= 3: REGULAR (2.0)
| | | quiz_continue_attemp > 3: RUIM (6.0/1.0)
| | course_view > 23: REGULAR (11.0/1.0)
| quiz_review > 3: RUIM (2.0/1.0)

Number of Leaves :      7
Size of the tree :     13

```

**Figura 15 – Árvore de decisão para a base RI-nota**  
**Fonte: Autoria própria**

Nas árvores geradas para as disciplinas GA e GQ, apresentadas na Figura 16, na classificação da categoria da nota, o envio de trabalhos é o principal meio de avaliação para essas disciplinas, sendo que, os alunos que realizaram, no mínimo um envio de trabalho na disciplina de GA, e os alunos que tiveram no mínimo de 7 envios na disciplina de GQ, conseguiram uma nota classificada como BOA.

Para a disciplina de GA, observa-se que, entre os alunos que realizaram exame, os que tiveram nota RUIM ou REGULAR, tiveram números baixos para as ações: duas ou menos visualizações de lista de usuários (*user\_view\_all*) e quatro ou menos visualizações dos fóruns (*forum\_view\_forums*), já os alunos que realizaram exame e tiveram nota BOA, tiveram mais visualizações da lista de usuários (mais que duas) e postagens no fórum (mais que quatro). Com base nesses números, pode-se concluir que, os alunos com mais interesse e interações na disciplina e no fórum, mesmo precisando realizar exame, conseguiram uma nota boa na disciplina.

```

exame = 0
| upload_upload <= 0
| | sexo = M: BOA (6.0)
| | sexo = F
| | | forum_add_post <= 1: RUIM (5.0)
| | | forum_add_post > 1: BOA (4.0/1.0)
| upload_upload > 0: BOA (184.0/1.0)
exame = 1
| user_view_all <= 2
| | forum_add_post <= 4
| | | forum_view_forums <= 4
| | | | idade <= 22: RUIM (2.0)
| | | | idade > 22: REGULAR (30.0/4.0)
| | | forum_view_forums > 4: RUIM (2.0/1.0)
| | | forum_add_post > 4: BOA (2.0)
| | user_view_all > 2: BOA (2.0)

Number of Leaves :    9

Size of the tree :   17

```

(a) GA-nota

```

exame = 0
| upload_upload <= 6
| | sexo = M: BOA (2.0)
| | sexo = F: RUIM (4.0)
| upload_upload > 6: BOA (159.0)
exame = 1
| course_user_report <= 12: REGULAR (69.0/21.0)
| course_user_report > 12: BOA (3.0)

Number of Leaves :    5

Size of the tree :    9

```

(b) GQ-nota

Figura 16 – Árvores de decisão das bases GA-nota e GQ-nota

Para as três disciplinas, nas árvores de decisão, o atributo *exame* foi utilizado como raiz da árvore. Observa-se um comportamento distinto entre os alunos que realizaram e os que não realizaram exame nas disciplinas.

#### 4.5 AVALIAÇÃO DOS RESULTADOS

Esta etapa foi realizada ao término de cada uma das etapas do processo da DCBD, afim interpretar o produto final de cada uma delas, validando ou não o resultado encontrado, acarretando por muitas vezes na repetição da etapa avaliada, da etapa anterior ou do processo como um todo. A ênfase da avaliação dos resultados se dá principalmente na etapa da Mineração de Dados, onde o pós-processamento dos dados, deve estar de encontro para os objetivos inicialmente propostos.

Observa-se que com base nos resultados obtidos durante a etapa de mineração de dados, em ambos os experimentos realizados, que existe uma relação entre a utilização do AVA e o desempenho do aluno. Dessa maneira o objetivo inicialmente proposto para este estudo, depois de diversas repetições do processo, foi alcançado.

Em relação ao primeiro experimento, embora o comportamento em cada árvore de decisão gerada seja específico de cada disciplina, alguns atributos utilizados pelos algoritmos de classificação para compor a regra de classificação, extraídos destas árvores, são repetidos em ambas.

Nas disciplinas de RI,GA e GQ, a interação com as ferramentas e recursos utilizados para a avaliação foram determinantes para a aprovação dos alunos. Interpretando a árvore de decisão apresentada para a disciplina de GQ, a regra mais relevante mostra que:

- Alunos que realizaram mais de 5 envios de trabalho (*upload\_upload*) foram aprovados – 220 alunos tiveram seu desempenho (Aprovação) classificado corretamente somente com esta regra.

Na disciplina de GA, a regra mais expressiva é:

- Os alunos que realizaram pelo menos um envio de trabalho (*upload\_upload*) foram aprovados sem Exame - 183 alunos tiveram seu desempenho (Aprovação) classificado corretamente somente com esta regra.

Na disciplina de RI, a regra extraída é:

- Dentre os alunos que não ficaram para exame, aqueles que interagiram mais de 4 vezes com a visualização dos trabalhos foram aprovados – 204 alunos tiveram seu desempenho (Aprovação) classificado corretamente somente com esta regra.

Para prever a aprovação, essas árvores tiveram no mínimo 98% de acerto, e em média uma Estatística Kappa de 0.47, mostrando que estas regras são capazes de classificar novos dados com concordância moderada em relação ao resultado esperado.

Com base nos resultados, observa-se que o comportamento mais relevante do aluno para a classificação como aprovado ou reprovado, é a interação com as ferramentas de avaliação, questionários e envio de trabalhos, sendo que interações com os demais recursos e ferramentas utilizadas, como o fórum, visualização de recursos, e até mesmo a quantidade de acessos, no escopo das disciplinas estudadas, não influenciam diretamente na classificação no desempenho do aluno quanto a aprovação ou reprovação.

Para a classificação da classe *status*, entre as disciplinas estudadas, o algoritmo J48 obteve melhor desempenho entre os algoritmos utilizados para comparação, pois, além de ter médias de acerto melhor para as 3 disciplinas estudadas, ainda permitiu a interpretação e visualização das regras, através da apresentação da árvore de decisão gerada.



No segundo experimento, observou-se que, dentro do escopo utilizado para o estudo, a classificação utilizando o atributo *categoria\_nota* demonstra ser mais eficaz quando comparado os valores da Estatística Kappa das execuções dos algoritmos. Seguida pela classificação utilizando o atributo *status* e *exame* respectivamente.

Já classificação quanto a realização ou não de exame com base na interação com a plataforma, teve um valor de Estatística Kappa com um nível de concordância pobre com a classificação esperada em duas disciplinas, GA e RI, com qualquer um dos algoritmos. Como os testes consistiam na validação dos resultados nas três disciplinas estudadas, mesmo que na disciplina de GQ a Estatística Kappa apresentada tem nível de concordância excelente com a classificação proposta, essa informação não foi considerada como estável e consistente, por ser um comportamento discrepante em relação aos resultados obtidos nas outras duas disciplinas do estudo.

Em ambos os experimentos realizados, observa-se que os algoritmos de árvores de decisão (J48) e redes neurais (MultilayerPerceptron) tiveram um desempenho semelhante entre si no escopo deste trabalho.

Por tanto, o fator determinante para a classificação do desempenho de um aluno (aprovação, reprovação e nota), no escopo das disciplinas estudadas, é a interação do estudante com as ferramentas de avaliação das disciplinas. Sendo que a relevância das demais interações em outras ferramentas do AVA, ainda é baixa ou inexistente. Desta maneira, esta informação pode ser alarmante, pois sugere:

- A falta de diversidade dos recursos utilizados para avaliar os alunos nestas disciplinas,
- Baixo incentivo aos alunos para acessar as disciplinas e efetivamente explorar os recursos nelas disponíveis.
- Baixo incentivo e necessidade, por parte dos alunos, de participar, visualizar ou interagir em discussões nos fóruns.

Estas constatações são potencialmente úteis no auxílio à gestores no processo de tomada de decisão, no que diz respeito a eventuais intervenções em cursos, disciplinas ou até mesmo diretamente com professores e alunos, afim de verificar e sugerir adequações no AVA, que visem garantir a dinâmica de ensino-aprendizagem e o aperfeiçoamento e reforço das potencialidades do ensino a distância.

## 5 CONCLUSÃO

A quantidade de pessoas matriculadas em cursos de nível superior na modalidade à distância, tem aumentado significativamente nos últimos anos. Um dos grandes desafios é de como realizar a avaliação e o acompanhamento do aprendizado desses alunos. Uma saída é a captura e análise das informações armazenadas em AVA.

Neste trabalho, foi apresentada uma abordagem de classificação dos alunos de Educação a Distância com base nas interações dos mesmos com os recursos disponíveis no AVA. Para tanto, foi utilizada a base de um AVA real, e nela foi aplicada o processo da DCBD e os conceitos da Mineração de Dados. Observa-se através dos resultados obtidos, que as interações do usuário e as ações que ele pratica no AVA tem correlação com o seu desempenho.

Um dado que pode ser considerado alarmante, é que a interação do aluno com as ferramentas de avaliação disponíveis na disciplina, está diretamente relacionada com a possibilidade de aprovação ou reprovação, ou até mesmo na previsão de sua nota, não sendo necessário levar em conta quantos acessos este aluno teve na disciplina e outras interações como visualização do material disponibilizado, interações no fórum com colegas e tutores, visita a unidades e conteúdos da disciplina por exemplo.

Dessa maneira, os resultados apresentados neste trabalho, podem sugerir a realização de uma leitura dos métodos de avaliação, utilizados pelos professores dentro do AVA, fornecendo parâmetros para subsidiar a utilização de outras ferramentas disponíveis dentro da ferramenta Moodle, não tão somente o questionário e envios de arquivos, para diversificar e/ou aprimorar a avaliação de um aluno.

Uma consideração positiva, é que a Mineração de Dados pode ser utilizada para a previsão do desempenho de um acadêmico, aprovação, reprovação e nota, tendo como base a interação produzida dentro do AVA. Pode se dizer também que a metodologia adotada para o estudo, o processo DCBD proposta por Fayyad et al.(1996), se mostra eficiente quando aplicada em base de dados educacionais, mais especificamente, em bases de dados de AVA.

Este trabalho pode contribuir para monitorar o desempenho dos alunos, como por exemplo, a criação de um Sistema de Apoio a Decisão, utilizando técnicas e conceitos da MD, para construir relatórios personalizados sobre o desempenho do aluno com base em sua interação com o AVA.

Um contribuição deste trabalho, foi a realização do estudo a partir de uma base de dados real e atual, onde descreve-se todas as etapas do processo de DCDB, que tiveram o intuito de identificar resultados relevantes para a tarefa de classificação. Portanto, pode ser utilizado como um referencial para a aplicação do processo em outra base de dados proveniente de AVA e para criação de estratégias para monitorar o comportamento do alunos.

Outra consideração importante, é sobre o poder da ferramenta Weka, que é uma ferramenta gratuita com uma ampla gama de recursos e algoritmos para a execução e aplicação de tarefas e técnicas de Mineração de Dados, portanto este software pode ser de grande valia para a aplicação do processo de DCBD.

Como trabalhos futuros, apresentam-se alguns desafios:

- Aplicar a DCBD em um novo conjunto alvo utilizando ainda a base de dados AVA Moodle, em busca de confirmação do conhecimento descoberto,.
- Utilizar o mesmo conjunto alvo, mas agora mudando os atributos e a abordagem para a extração dos mesmos.
- Aplicar outras tarefas e técnicas de mineração de dados e assim complementar os resultados, através das tarefas de agrupamento ou predição de comportamento destes alunos por exemplo.
- Investigar junto aos professores, qual é a perspectiva de avaliação mais utilizadas nos demais cursos do EAD da instituição, afim de validar o resultado encontrado.
- Planejar um Sistema de Apoio a decisão, voltada a EAD, com base nesse estudo.

As áreas de Mineração de Dados e Sistemas de Apoio a Decisão podem contribuir para a modalidade de Ensino a Distância. Os Desafios para a ampliação e consolidação desta modalidade, são muitos, principalmente no que se diz quanto a avaliação do aprendizado dos alunos. Estudos que integram essas áreas e utilizam a grande disponibilidade de dados disponíveis em AVA, são uma grande vantagem nessa corrida e podem contribuir potencialmente para vencer esse desafio, garantindo assim essa e outras atividades inerentes do sistema educacional, se valendo da ótica computacional para auxiliar o processo de tomada de decisão.

## REFERÊNCIAS

- ALMEIDA, M. E. B. **Educação a distância na internet: abordagens e contribuições dos ambientes digitais de aprendizagem**. São Paulo: [s.n.], 2003. 327-340 p.
- BERRY, M. J. A.; LINOFF, G. **Data Mining Techniques: For Marketing, Sales, and Customer Support**. New York: Wiley Computer Publishing, 1997.
- BLACKBOARD. **Blackboard - Ensino Superior**. 2015. Disponível em: <http://blackboard.grupoa.com.br/mercados/ensino-superior/>. Acesso em: 03 mar. 2015.
- CONTI, F. de. **Mineração de dados no Moodle: Análise de prazos de entrega de atividade**. 68 p. Dissertação de Mestrado, 2011.
- DAMASCENO, M. **Introdução a Mineração de Dados utilizando o Weka**. [S.l.], 2010. Disponível em: <http://connepi.ifal.edu.br/ocs/index.php/connepi/CONNepi2010/paper/viewFile/258/207>. Acesso em: 02 abr. 2015.
- DEMARCO, D. J. Um balanço do programa nacional de formação em administração pública (pnap) como estratégia de fortalecimento da gestão pública : o caso da escola de administração da ufrgs. in: Congresso consad de gestão pública. **Congresso Consad de Gestão Pública**, p. 1 – 27, 2013.
- DEMARCO, D. J.; VIEIRA, A. Programa nacional de formação em administração pública (pnap): Um balanço da implementação pela escola de administração da ufrgs. in: Esud 2014. **Xi Congresso Brasileiro De Ensino Superior A Distância**, Florianópolis, 2014.
- DETONI, D.; ARAUJO, R. M. de; CECHINEL, C. Predição de reprovação de alunos de educação a distância utilizando contagem de interações. In: **Simpósio Brasileiro de Informática na Educação, 2014**. Dourados: Anais do 3º Congresso Brasileiro de Informática na Educação, 2014. v. 1, p. 1–8.
- DUNKEL, B. et al. Systems for kdd: From concepts to practice. **Future Generation Computer Systems**, v. 13, p. 231–242, 1997.
- FAYYAD, U. et al. **Advances in Knowledge discovery and data mining**. Menlo Park: Mit Press, 1996.
- FMUP. **Métodos de Estimação de Reprodutividade de Medidas**. 2015. Disponível em: <http://users.med.up.pt/joakim/intromed/estatisticakappa.htm>. Acesso em: 06 mai. 2015.
- FRAWLEY, W. J.; SHAPIRO, G. P.; MATHEUS, C. J. **Knowledge Discovery in Databases: an Overview**. 1992. 57-70 p.
- GOEBEL, M.; GRUENWALD, L. A survey of data mining and knowledge discovery software tools. **SIGKDD Explorations**, v. 1, p. 20–33, 1999.
- GOLDSCHMIDT, R.; PASSOS, E. L. **Data Mining: um guia prático: conceitos, técnicas, ferramentas e aplicações**. [S.l.]: Elsevier, 2005.
- GOMES, A. S. et al. Amadeus: Novo modelo de sistema de gestão de aprendizagem. v. 8, p. 1–15, 2009.
- HARRISON, T. **Intranet data warehouse: ferramentas e tecnicas para a utilizacao do data warehouse na intranet**. [S.l.]: Berkerley/ABDR, 1998. ISBN 9788572514606.

IARALHAM, L. C. **Contribuição da tecnologia da informação na educação a distância no instituto universal brasileiro: um estudo de caso**. 2009. Disponível em: <http://www.fam2011.com.br/site/revista/pdf/ed4/art3.pdf>.

INEP. **Censo da Educação Superior - CENSUP 2013**. 2014. Disponível em: [http://download.inep.gov.br/educacao\\_superior/censo\\_superior/apresentacao/2014/coletiva\\_censo\\_superior\\_2013.pdf](http://download.inep.gov.br/educacao_superior/censo_superior/apresentacao/2014/coletiva_censo_superior_2013.pdf). Acesso em: 03 mar. 2015.

JUNIOR, J. de O.; NORONHA, R. V.; KAESTENER, C. A. A. Análise da correlação da evasão de cursos de graduação com o empréstimo de livros em biblioteca. In: **Simpósio Brasileiro de Informática na Educação, 2014**. Dourados: Anais do 3º Congresso Brasileiro de Informática na Educação, 2014. v. 1, p. 1–10.

LE MOS, E. P. **Análise de crédito bancário com o uso de data mining: redes neurais e árvores de decisão**. Tese (Doutorado) — Universidade Federal do Paraná, 2003.

LOPES, C. C. **Um sistema de apoio à tomada de decisão no acompanhamento do aprendizado em Educação a Distância**. 2003.

MARQUES, F. C. S.; CARVALHO, J. O. F. de. Prospecção dos tipos de sistemas de apoio à tomada de decisão utilizados nas organizações e identificação das formas de apresentação de suas informações. In: **XVIII Encontro de Iniciação Científica da PUC-Campinas**. São Paulo: Anais do XVIII Encontro de Iniciação Científica da PUC-Campinas, 2013. v. 1, p. 1–5.

MEC. **Secretaria de Educação a Distância : Referenciais de qualidade para a educação superior a distância**. 2007. Disponível em: <http://portal.mec.gov.br/seed/arquivos/pdf/legislacao/refead1.pdf>. Acesso em: 06 fev. 2015.

MIRANDA, G. Q. Mec/uab – programa universidade aberta do brasil. **IV Simpósio Internacional - O Estado e as Políticas Educacionais no tempo presente**, Uberlândia, 2008. Disponível em: <http://www.simposioestadopoliticas.ufu.br/imagens/anais/pdf/DC28.pdf>. Acesso em: 06 fev. 2015.

MOODLE. **Site Oficial do Moodle - Documentação**. 2015. Disponível em: [https://docs.moodle.org/28/en/Main\\_page](https://docs.moodle.org/28/en/Main_page). Acesso em: 21 jan. 2015.

MORAN, J. M. **Modelos e avaliação do ensino superior a distância no Brasil**. 2009. Disponível em: <http://www.eca.usp.br/prof/moran/site/textos/educacaoonline/modelos1.pdf>. Acesso em: 07 fev. 2015.

NUNES, A. C. de O. Fada –ferramenta de apoio à decisão acadêmica. In: **13º Congresso Internacional de Educação a Distância**, Curitiba, Anais do Evento, p. 1–11, 2007.

OLIVEIRA, D. de Pinho Rebouças de. **Sistemas de informações gerenciais: estratégias, táticas, operacionais**. Atlas, 2005. 299 p. ISBN 9788522446131. Disponível em: <https://books.google.com.br/books?id=OsWHNAAACAAJ>. Acesso em: 11 fev. 2015.

PEREIRA, T. R.; CHAVES, D. A. Moodle: Um experimento on-line para potencializar um ambiente de apoio à aprendizagem. **XVIII SIMPÓSIO NACIONAL DE GEOMETRIA DESCRITIVA E DESENHO TÉCNICO**, 2007.

PETERS, O. **A educação a distância em transição: tendências e desafios**. São Leopoldo: Editora Unisinos, 2003. 400 p.

- PNAP. **Programa Nacional de Formação em Administração Pública**. 2014. Disponível em: <http://www.pnap.ufsc.br>. Acesso em: 14 jan. 2015.
- POLLONI, E. G. F. **Administrando sistemas de informação: estudo de viabilidade**. 2. ed. São Paulo: Futura, 2001.
- POSTGRESQL. **Sobre o PostgreSQL**. 2015. Disponível em: <http://www.postgresql.org.br/old/sobre>. Acesso em: 26 mar. 2015.
- RAMOS, J. L. C. **Requisitos para ferramentas de avaliação em ambientes virtuais de ensino**. Dissertação de Mestrado, 2006.
- RIGO, S. J. et al. Minerando dados educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. **Revista Brasileira de Informática na Educação**, v. 22, p. 132–145, 2014.
- RODRIGUES, C. A. F.; SCHIDMIT, L. M. **Introdução à Educação à Distância**. Ponta Grossa: NUTEAD/UEPG, 2010.
- ROMERO, C.; VENTURA, S. Data mining in education. wiley interdisciplinary reviews: Data mining and knowledge discovery. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, John Wiley & Sons, Inc., v. 3, 2013.
- SANTOS, H. L. dos; CAMARGO, F. N. P.; CAMARGO, S. S. Predizendo o sucesso de estudantes através do uso avaliações formativas em avas. In: **Simpósio Brasileiro de Informática na Educação, 2014**. Rio de Janeiro: Anais dos Workshops do Congresso Brasileiro de Informática na Educação, 2012. v. 1, p. 1–10.
- SILVA, I. A. F. **Descoberta de conhecimento em base de dados de monitoramento ambiental para a avaliação da qualidade da água**. Dissertação de Mestrado, 2010. Disponível em: [http://pgfa.ufmt.br/index.php?option=com\\_docman&task=doc\\_download&gid=92&Itemid=37](http://pgfa.ufmt.br/index.php?option=com_docman&task=doc_download&gid=92&Itemid=37). Acesso em: 08 abr. 2015.
- TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao data mining: mineração de dados**. Rio de Janeiro: Editora Ciência Moderna, 2009. 900 p.
- TELEDUC. **TelEduc: Quem Somos**. 2015. Disponível em: <http://www.teleduc.org.br/?q=historico>. Acesso em: 03 mar. 2015.
- WEKA. **Weka 3: Data Mining Software in Java**. 2015. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 15 fev. 2015.
- WITTEN, I. H.; FRANK, E. **Data Mining: Practical machine learning tools and techniques**. [S.l.]: Morgan Kaufmann, 2005.

## **APÊNDICE A – SQL UTILIZADAS PARA EXTRAÇÃO DE DADOS**

## Consultas SQL para extração de dados do AVA

Neste apêndice apresentamos algumas consultas utilizadas para extração dos dados do banco do AVA, o prefixo das tabelas é "ava", podendo ser configurado na instalação do sistema Moodle, sendo o prefixo padrão "mdl".

Consulta para listar todos os *logs* do AVA:

```
select * ava_log
```

Consulta Sql para listar todas as ações e o número de vezes que foram realizadas dentro do AVA:

```
select
concat( module , ' ' , action) as acao , count(*) as total
from ava_log
group by module, action
order by module ,action;
```

Consulta para listar todas as ações agrupadas por usuário:

```
select
userid,
concat( module , ' ' , action) as acao , count(*) as total
from ava_log where course = 2451 and userid = 4108
group by module, action , userid
order by module ,action;
```

Consulta para listar o total de vezes de cada ação realizada por cada aluno, essa consulta cria uma coluna para o identificador do usuário e uma coluna para cada ação realizada:

```
select usuario ,
sum(survey_view_all) as survey_view_all,
sum(survey_view_graph) as survey_view_graph,
sum(user_change_password) as user_change_password,
sum(forum_view_forum) as forum_view_forum,
sum(assignment_view_all) as assignment_view_all,
sum(quiz_attempt) as quiz_attempt,
sum(forum_delete_post) as forum_delete_post,
sum(upload_upload) as upload_upload,
sum(forum_add_discussion) as forum_add_discussion,
```



```
sum(user_view) as user_view,
sum(course_user_report) as course_user_report,
sum(forum_add_post) as forum_add_post,
sum(forum_delete_discussion) as forum_delete_discussion,
sum(quiz_close_attempt) as quiz_close_attempt,
sum(forum_view_forums) as forum_view_forums,
sum(quiz_continue_attemp) as quiz_continue_attemp,
sum(assignment_view) as assignment_view,
sum(quiz_view) as quiz_view,
sum(course_view) as course_view,
sum(forum_view_discussion) as forum_view_discussion,
sum(quiz_review) as quiz_review,
sum(resource_view_all) as resource_view_all,
sum(resource_view) as resource_view,
sum(forum_update_post) as forum_update_post,
sum(quiz_view_all) as quiz_view_all,
sum(survey_view_form) as survey_view_form,
sum(forum_user_report) as forum_user_report,
sum(discussion_mark_read) as discussion_mark_read,
sum(survey_submit) as survey_submit,
sum(forum_subscribe) as forum_subscribe,
sum(forum_search) as forum_search,
sum(user_update) as user_update,
sum(user_view_all) as user_view_all ,
sum(assignment_upload) as assignment_upload
from (
select usuario ,
case acao when 'survey view all' then total else 0 end as survey_view_all,
case acao when 'survey view graph' then total else 0 end as survey_view_graph,
case acao when 'user change password' then total else 0 end as user_change_password,
case acao when 'forum view forum' then total else 0 end as forum_view_forum,
case acao when 'assignment view all' then total else 0 end as assignment_view_all,
case acao when 'quiz attempt' then total else 0 end as quiz_attempt,
case acao when 'forum delete post' then total else 0 end as forum_delete_post,
case acao when 'upload upload' then total else 0 end as upload_upload,
case acao when 'forum add discussion' then total else 0 end as forum_add_discussion,
case acao when 'user view' then total else 0 end as user_view,
case acao when 'course user report' then total else 0 end as course_user_report,
case acao when 'forum add post' then total else 0 end as forum_add_post,
case acao when 'forum delete discussion' then total else 0 end
as forum_delete_discussion,
case acao when 'quiz close attempt' then total else 0 end as quiz_close_attempt,
case acao when 'forum view forums' then total else 0 end as forum_view_forums,
case acao when 'quiz continue attemp' then total else 0 end as quiz_continue_attemp,
case acao when 'assignment view' then total else 0 end as assignment_view,
case acao when 'quiz view' then total else 0 end as quiz_view,
case acao when 'course view' then total else 0 end as course_view,
case acao when 'forum view discussion' then total else 0 end
```

```
as forum_view_discussion,
case acao when 'quiz review' then total else 0 end as quiz_review,
case acao when 'resource view all' then total else 0 end as resource_view_all,
case acao when 'resource view' then total else 0 end as resource_view,
case acao when 'forum update post' then total else 0 end as forum_update_post,
case acao when 'quiz view all' then total else 0 end as quiz_view_all,
case acao when 'survey view form' then total else 0 end as survey_view_form,
case acao when 'forum user report' then total else 0 end as forum_user_report,
case acao when 'discussion mark read' then total else 0 end as discussion_mark_read,
case acao when 'survey submit' then total else 0 end as survey_submit,
case acao when 'forum subscribe' then total else 0 end as forum_subscribe,
case acao when 'forum search' then total else 0 end as forum_search,
case acao when 'user update' then total else 0 end as user_update,
case acao when 'user view all' then total else 0 end as user_view_all,
case acao when 'assignment upload' then total else 0 end as assignment_upload
from
(
select
userid usuario,
concat( module , ' ' , action) acao,
count(*) total
from ava_log
group by module, action ,userid
order by userid, module ,action
) a group by usuario , acao , total
) b
group by usuario
order by usuario;
```

Consulta para listar as categorias e os cursos cadastrados no AVA:

```
select
av4.parent , av4.id , av4.name,
av3.parent , av3.id , av3.name,
av2.parent , av2.id , av2.name,
av1.parent , av1.id , av1.name
from ava_course_categories av1
left join ava_course_categories av2 on av1.parent = av2.id
left join ava_course_categories av3 on av2.parent = av3.id
left join ava_course_categories av4 on av3.parent = av4.id
order by av2.name
```

Consulta para listar as ações dos usuários ordenadas por data de uma determinada disciplina:

```
select
to_timestamp( time ) , userid, module , action
from ava_log where course = ?
order by time asc;
```

## **APÊNDICE B – DESCRIÇÃO DOS DADOS**

## Descrição Dos Dados

Os dados estudados possuem as seguintes características: Nota-se que dos 237 alunos estudados a maioria dos alunos é do sexo feminino.

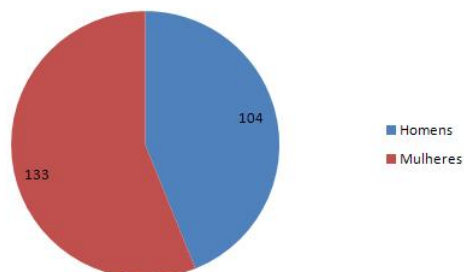


Figura 17 – Distribuição de alunos por sexo

A distribuição dos alunos por sexo entre os polos, observa-se que o polo de Ponta Grossa, é o que mais possui alunos e o de Congonhinhas é o que menos possui alunos matriculados e somente em Palmeira os homens são a maioria.

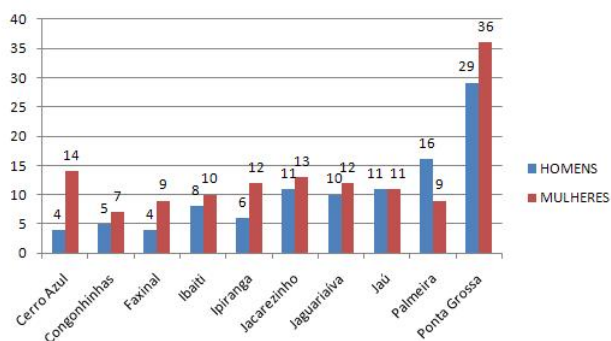


Figura 18 – Distribuição de alunos nos polos por sexo

A escolaridade dos alunos nos mostra que apenas 33 alunos já possuíam ensino superior completo quando ingressaram no curso.

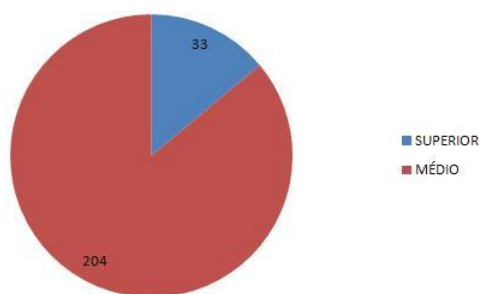
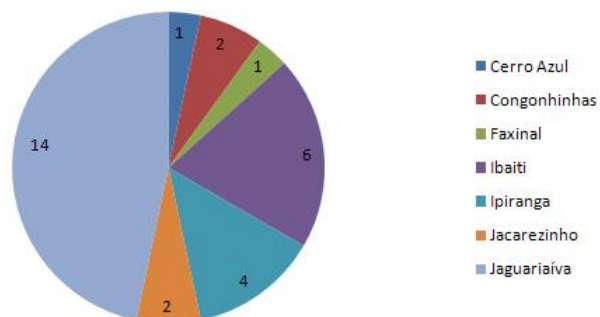


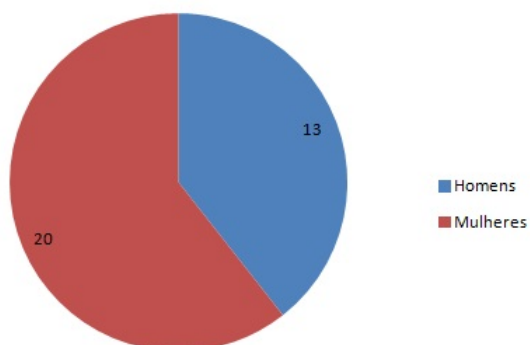
Figura 19 – Escolaridade dos alunos

Dentre os 10 polos, o polo de Ponta Grossa é o que possui o maior numero de alunos que já possuem nível superior completo.



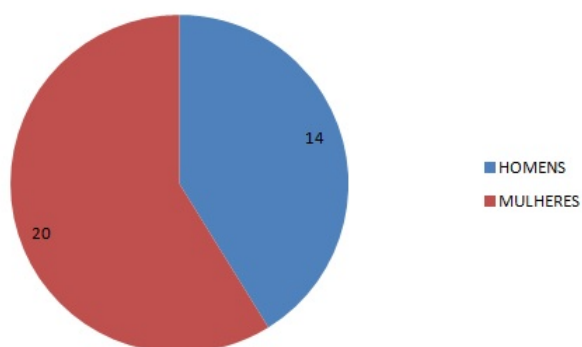
**Figura 20 – Alunos com ensino superior completo**

A distribuição por sexo dos alunos que já possuem o nível superior completo, mostra que a maioria destes é do sexo feminino.



**Figura 21 – Quantidade de Alunos com nível superior por sexo**

Quanto as disciplinas estudadas, a quantidade de reprovações por sexo, nos mostra que as mulheres reprovaram mais que os homens.



**Figura 22 – Quantidade de reprovações por sexo**

A seguir são apresentadas a quantidade de reprovações em cada disciplina que varia no máximo de 2 alunos da maior para a menor quantidade.

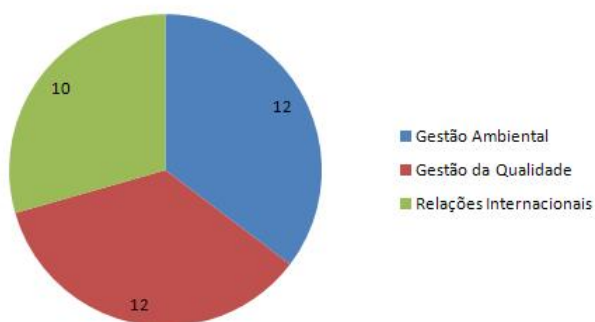


Figura 23 – Quantidade de reprovações por disciplina

A quantidade de reprovações por sexo e disciplina, nos revela que as mulheres reprovaram em igual ou maior quantidade que os homens .

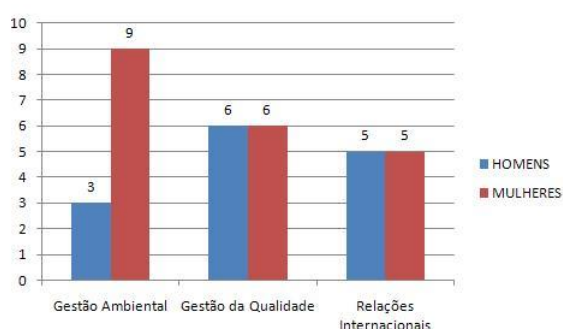


Figura 24 – Reprovações por polo e sexo na disciplina de Gestão Ambiental

Na disciplina Gestão Ambiental, as reprovações por polo e sexo, nos mostram que no polo de Palmeira e Jaguariaíva só tivemos 1 homem reprovado, os outros 2 homens reprovados aparecem no polo de Ponta Grossa, enquanto nos outros 3 polos, encontram-se o maior número de mulheres reprovadas.

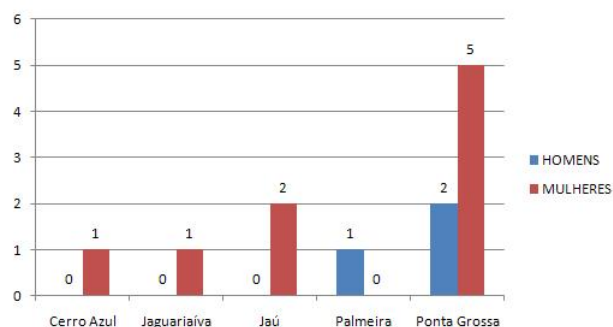
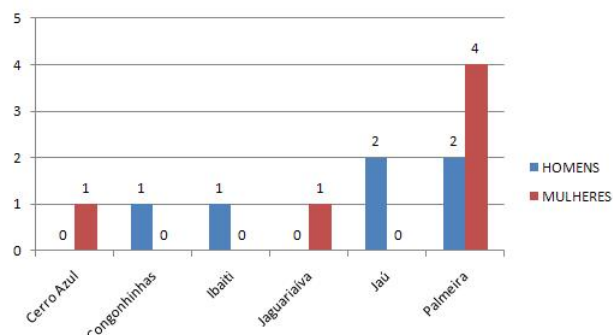


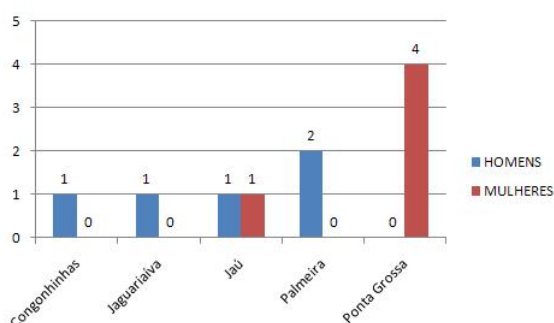
Figura 25 – Reprovações por polo e sexo na disciplina de Gestão Ambiental

Na disciplina Gestão da Qualidade, as reprovações por polo e sexo, nos mostram que nos polos de Congonhinhas, Ibaiti e Palmeira somente homens foram reprovados, em Cerro Azul e Jaú, uma mulher reprovado cada, e em Ponta Grossa 4 mulheres e 2 homens reprovados.



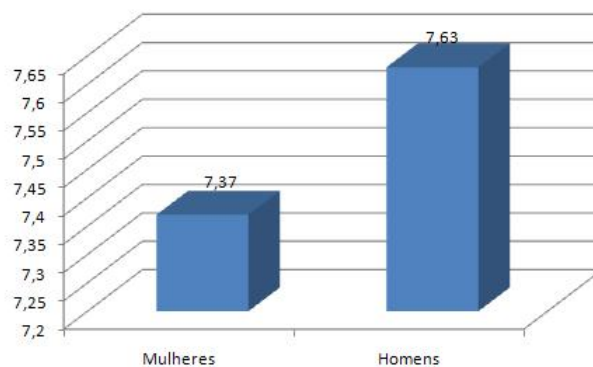
**Figura 26 – Reprovações por polo e sexo na disciplina de Políticas Públicas**

Na disciplina Relações Internacionais, as reprovações por polo e sexo, nos mostram que nos polos de Congonhinhas, Jaguariáiva e Palmeira tiveram somente homens reprovados, em Jaú teve 1 homem e uma mulher reprovados, e em Ponta Grossa apenas 4 mulheres reprovadas.



**Figura 27 – Reprovações por polo e sexo na disciplina de Relações Internacionais**

Nota-se que tanto os homens quanto as mulheres ficaram com a média de notas acima de 7,0, porém os homens superaram as notas das mulheres em 0,26 décimos.



**Figura 28 – Média de Nota por Sexo**



Quanto as médias dos polos por sexo, nota-se que dois polos se destacaram com médias maiores que 8.0, o polo de Jacarezinho com média de 8.4 seguido do polo de Faxinal com média de 8.22, observa-se também que em 7 dos 10 polos a média masculina é superior à feminina, porém a maior diferença é no polo de Congonhinhas, onde as mulheres tem 1.10 pontos a mais que os homens.

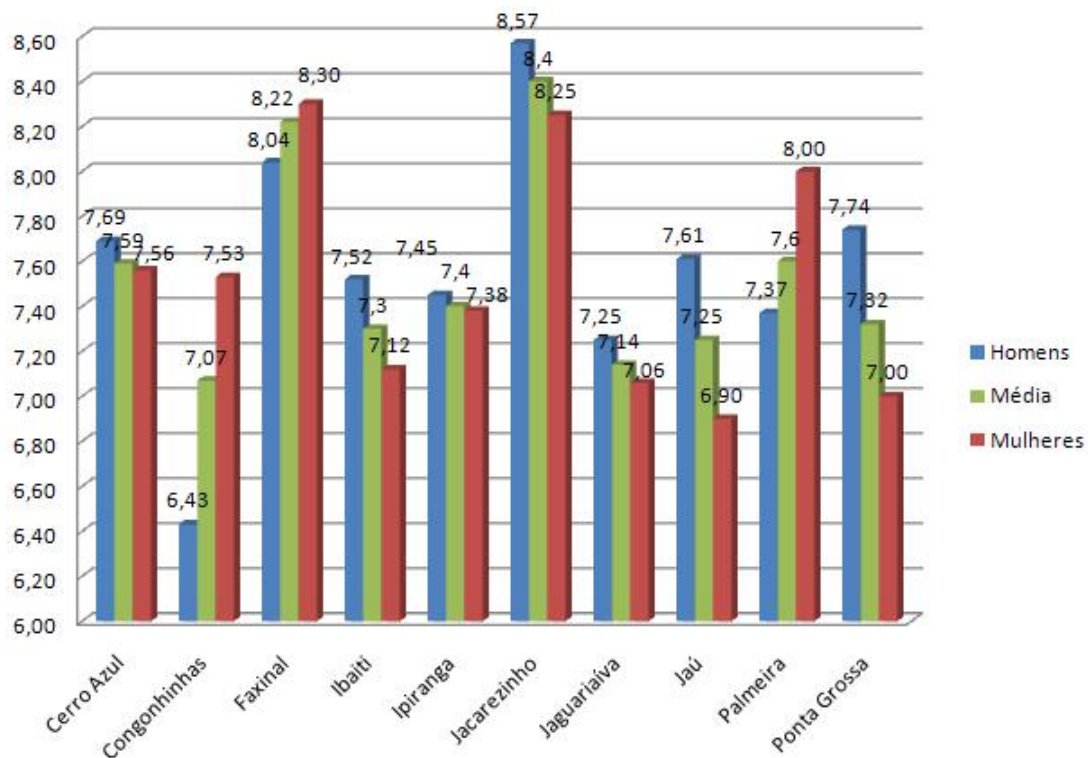


Figura 29 – Médias das notas dos polos por sexo

A média de idades por sexo, mostra que em geral, nos polos a média de idade das mulheres é mais alta do que a dos homens. .

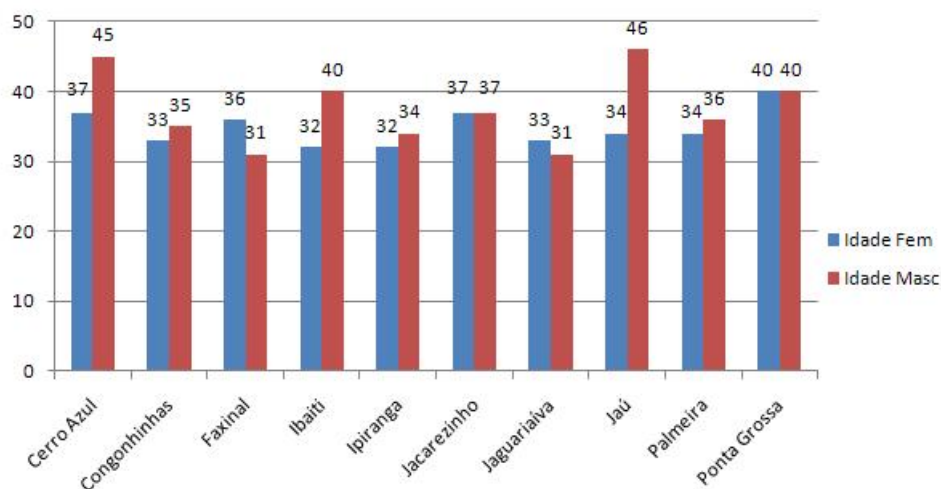


Figura 30 – Média de Idade por polo e sexo

Abaixo podemos verificar que a idade média dos alunos passa dos 30 anos de idade em todos os polos, sendo que nos polos de Ponta Grossa e de Jaú a média de idade dos alunos é de 40 anos.

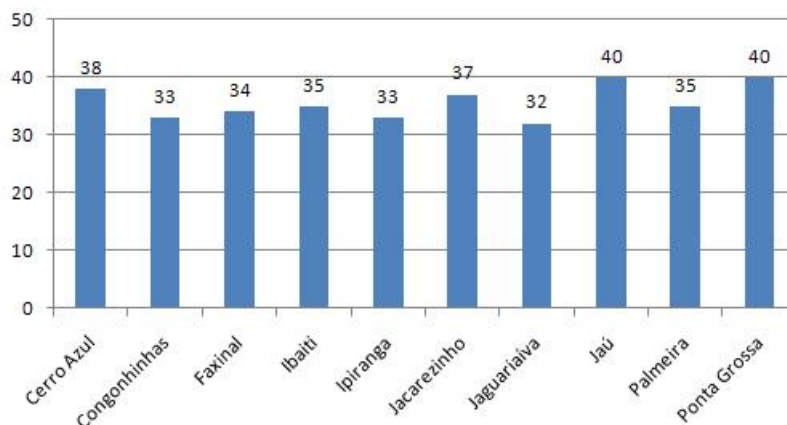


Figura 31 – Média de idade por polo

A maioria dos alunos tiveram médias acima de 7, sendo a maior média nessa matéria 9,5, a quantidade de acessos variou entre 4 e 105 acessos, porém, somente 10 alunos tiveram uma quantidade de acessos maior que 60. A maior nota foi obtida por alunos que tiveram a quantidade de acessos entre 9 e 32, entre os alunos que ficaram com notas entre 5 e menos que 7 as quantidades de acessos variaram entre 10 e 64.

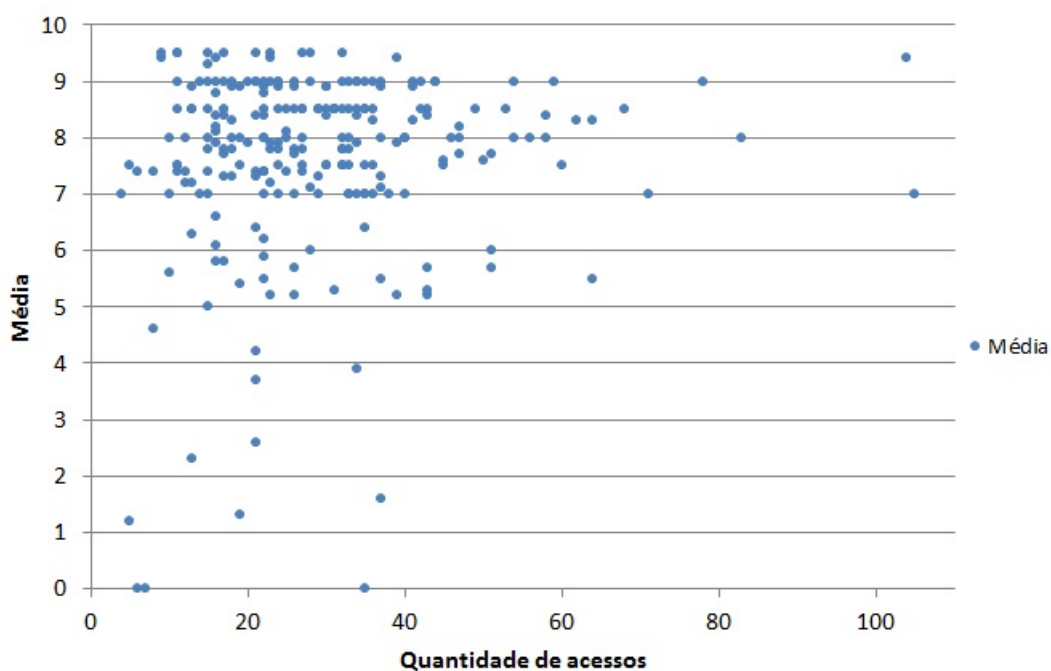


Figura 32 – Gestão Ambiental - Médias por quantidade de acessos

No gráfico de médias por polo vemos que o polo que mais teve notas altas foi o de Jacarezinho, com 8 alunos com médias de 9,5, e apenas um aluno com média abaixo de 7, com a quantidade de acessos variando entre 4 e 104 acessos.

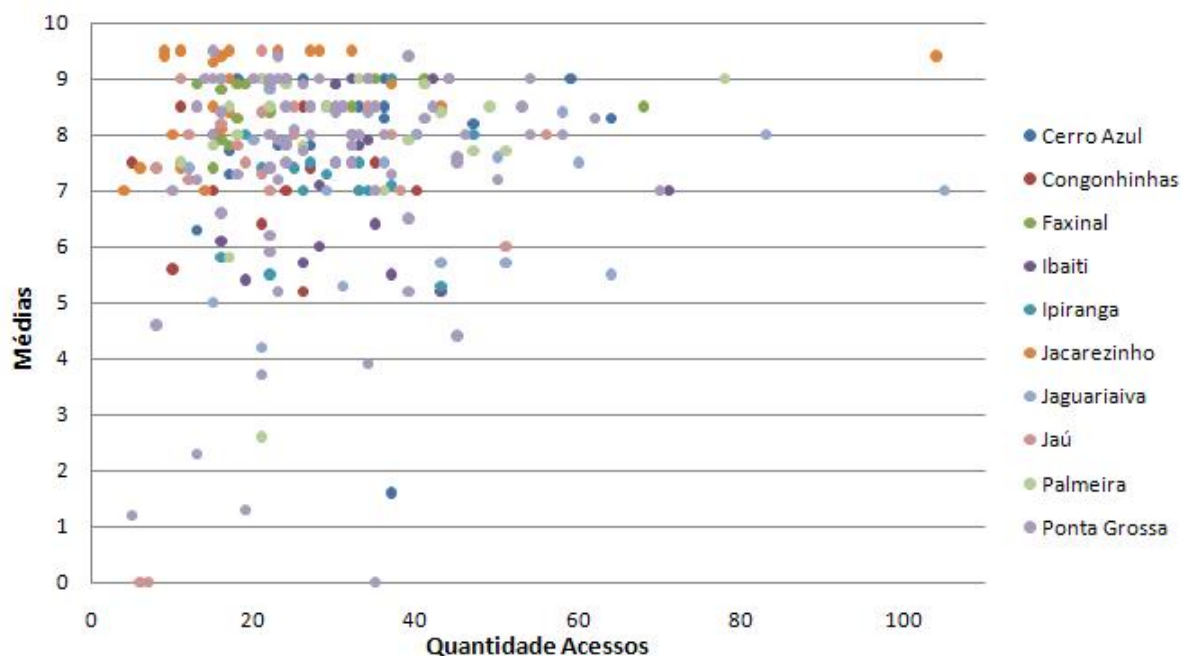


Figura 33 – Gestão Ambiental - Médias por polo pela quantidade de acessos

No gráfico de médias por sexo, tem-se que a maioria dos alunos com médias abaixo de 7 são do sexo feminino. Nota-se que a maioria dos alunos tiveram uma quantidade de acessos entre 10 e 39.

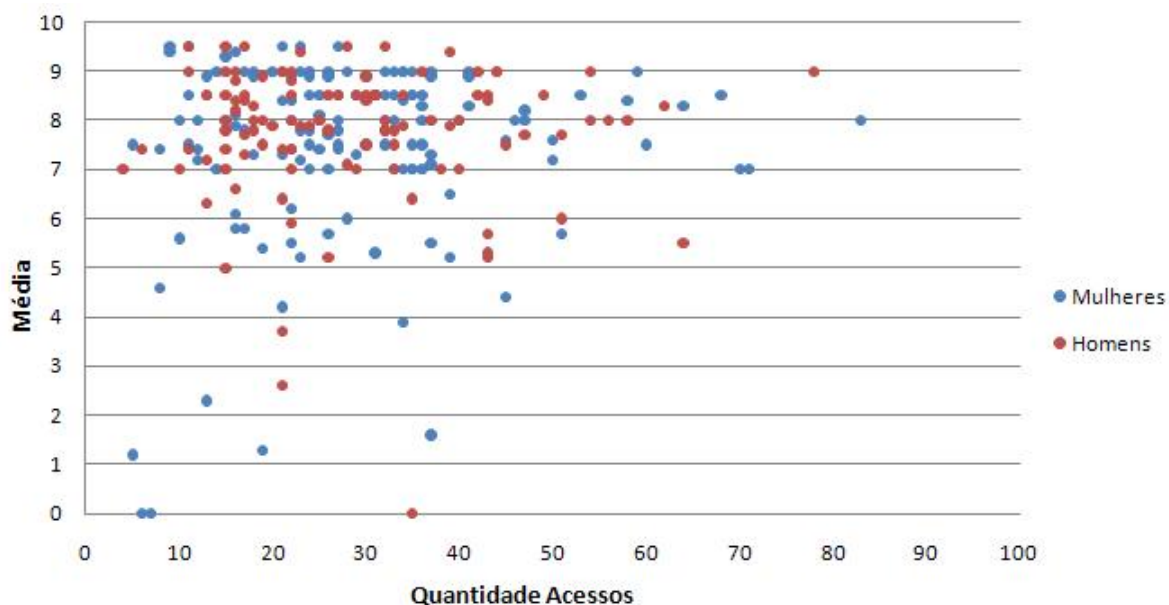


Figura 34 – Gestão Ambiental - Médias por sexo

Dos alunos que tiveram notas abaixo de 5,5, todos têm menos de 46 anos, os que alcançaram a nota máxima da matéria têm entre 24 e 50 anos.

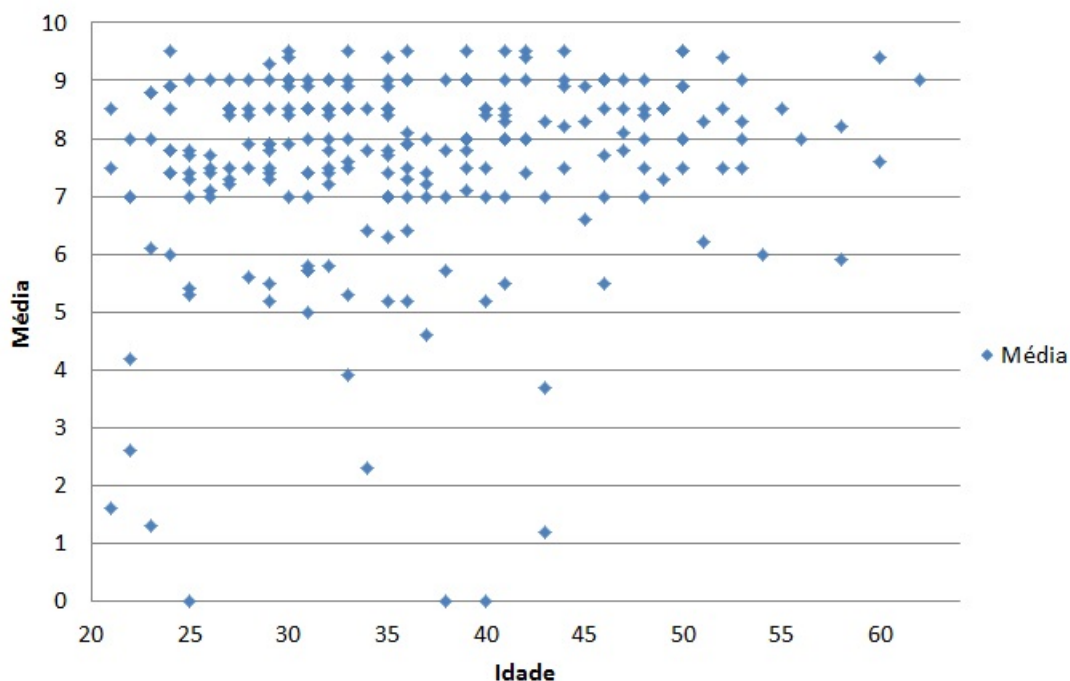


Figura 35 – Gestão Ambiental - Médias por idade

Na matéria de Gestão da Qualidade observa-se que a maioria dos alunos tem médias entre 7 e 8 e que vários alunos com quantidade de acessos acima de 80.

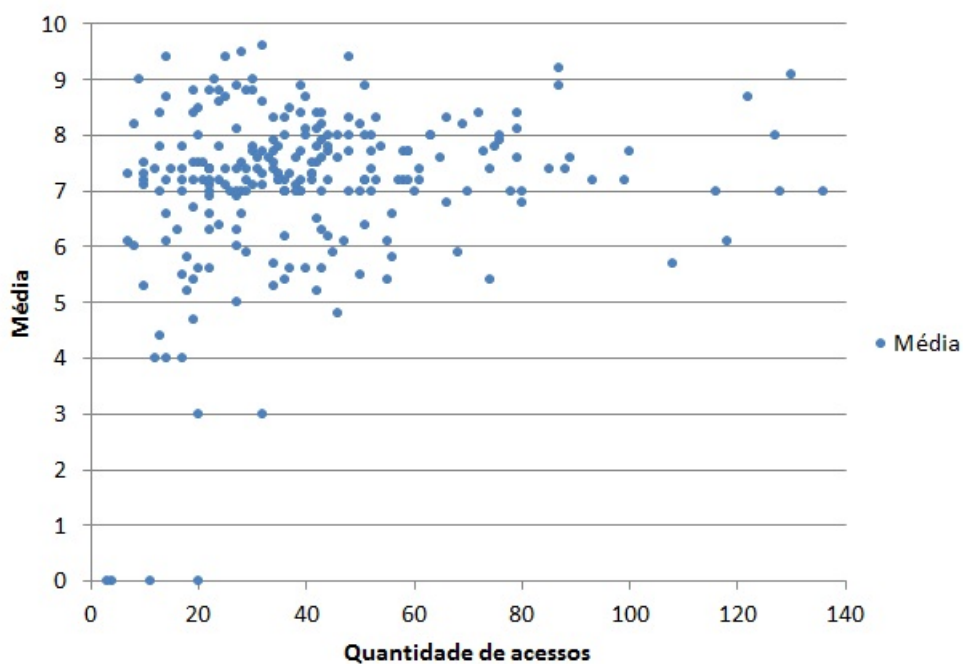


Figura 36 – Gestão da Qualidade - Médias por quantidade de acessos

Nota-se um predomínio das notas do polo de Jacarezinho, sendo que apenas 4 alunos desse polo tiveram média abaixo de 8, desses somente um aluno teve média abaixo de 7.

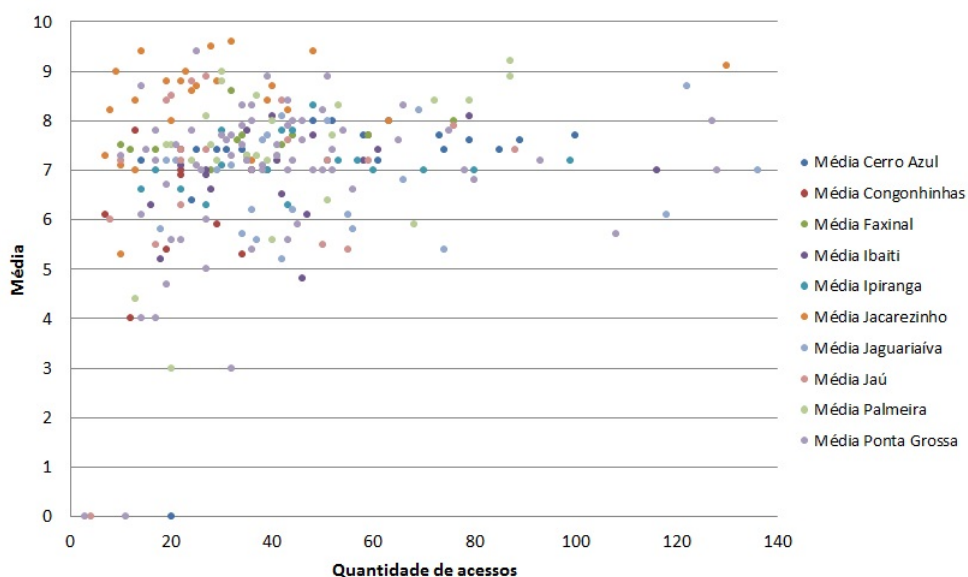


Figura 37 – Gestão da Qualidade - Médias por polo pela quantidade de acessos.

Observa-se no gráfico de acessos por sexo, que os alunos que tiraram notas acima de 8 a maioria é do sexo masculino e dos alunos que tiveram médias entre 5 e 7 a maioria é do sexo feminino.

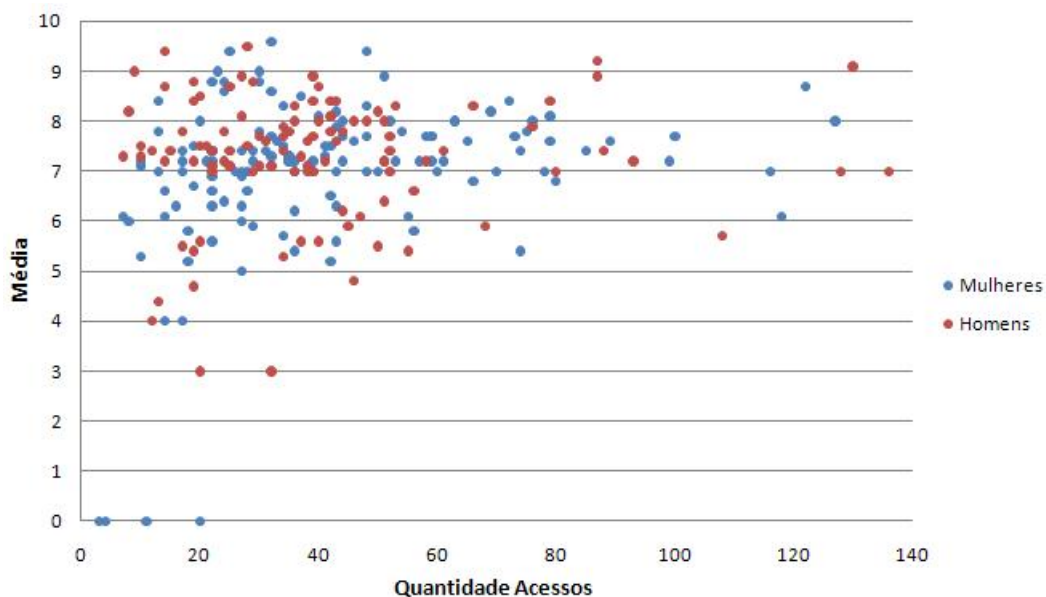


Figura 38 – Gestão da Qualidade - Médias por sexo

Entre os alunos que obtiveram notas acima de 9,0, a idade mínima é 28 anos, e a máxima 60. Os alunos com idades acima de 43 nenhum teve notas abaixo de 5,0.

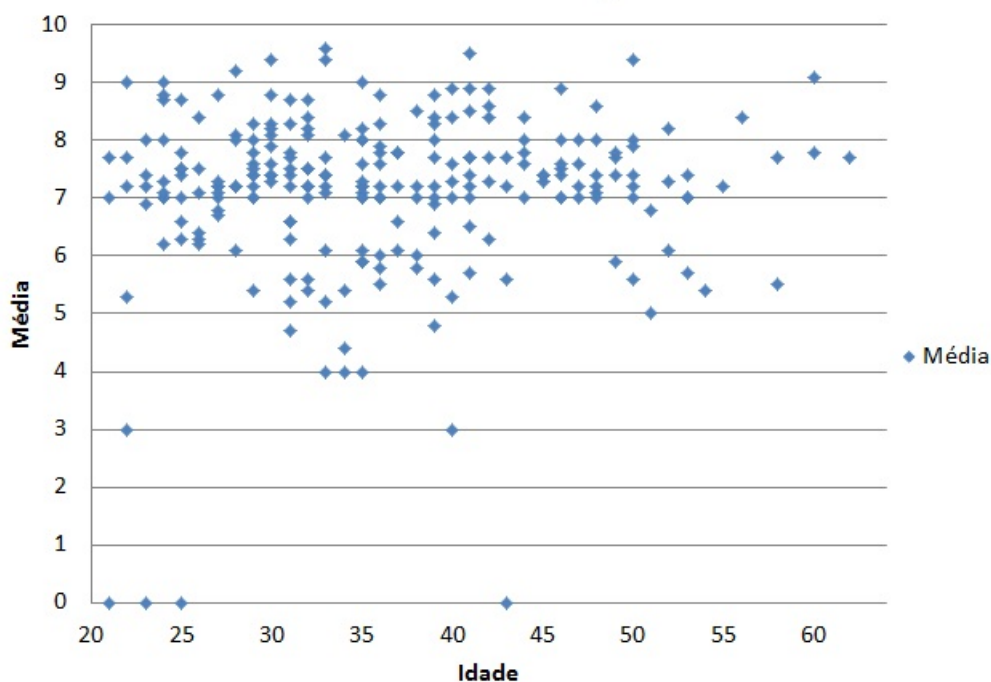


Figura 39 – Gestão da Qualidade - Médias por idade

Na matéria de Relações Internacionais, a maioria das médias está entre 8 e 10, sendo que poucas as medias foram abaixo de 7. A maioria predominante de acessos fica entre 7 e 40, com poucos alunos com quantidade de acessos maior que 60.

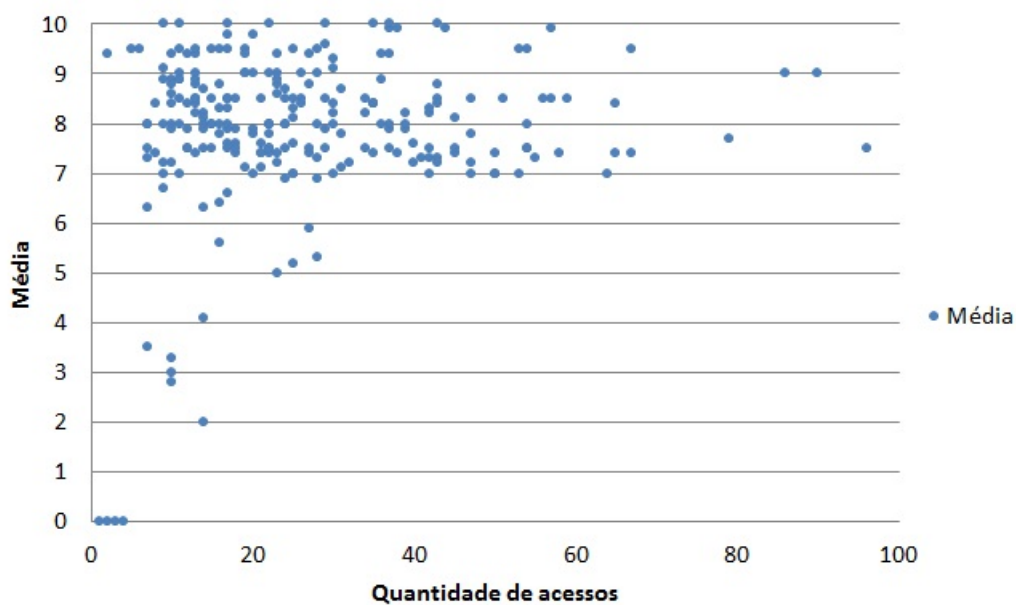


Figura 40 – Relações Internacionais - Médias por quantidade de acessos

Nota-se que a maioria dos alunos do polo de Jacarezinho tem médias acima de 8 e a maioria dos alunos do polo de Ponta Grossa tiveram médias entre 7 e 8.

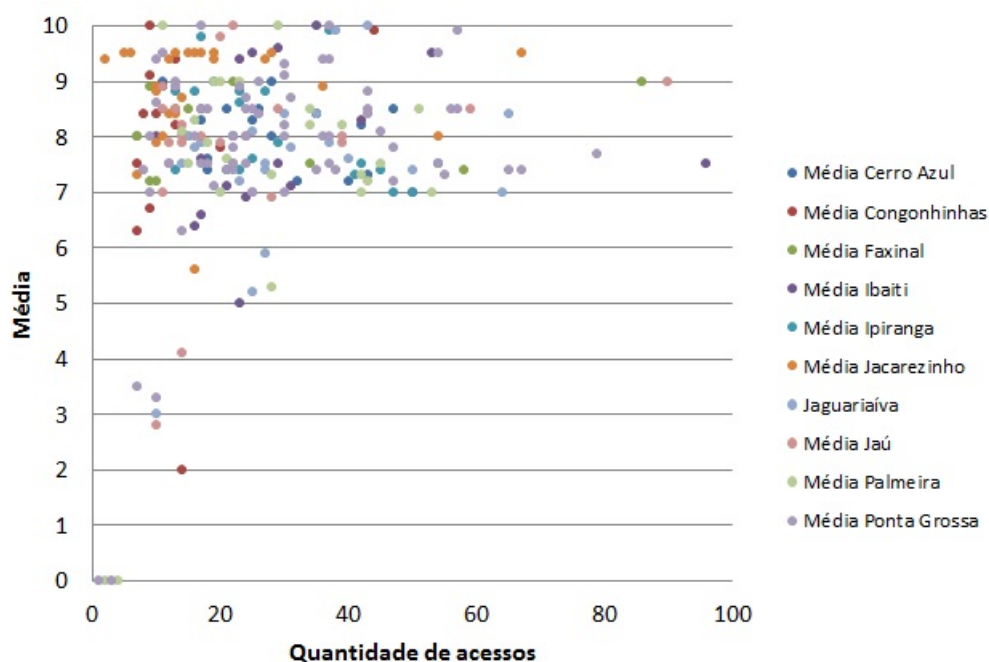


Figura 41 – Relações Internacionais - Médias por polo por quantidade de acessos

Na distribuição das notas por sexo, dos alunos que tiveram médias acima de 9, vemos que a maioria é do sexo masculino, a maior parte desses com médias acima de 9,5.

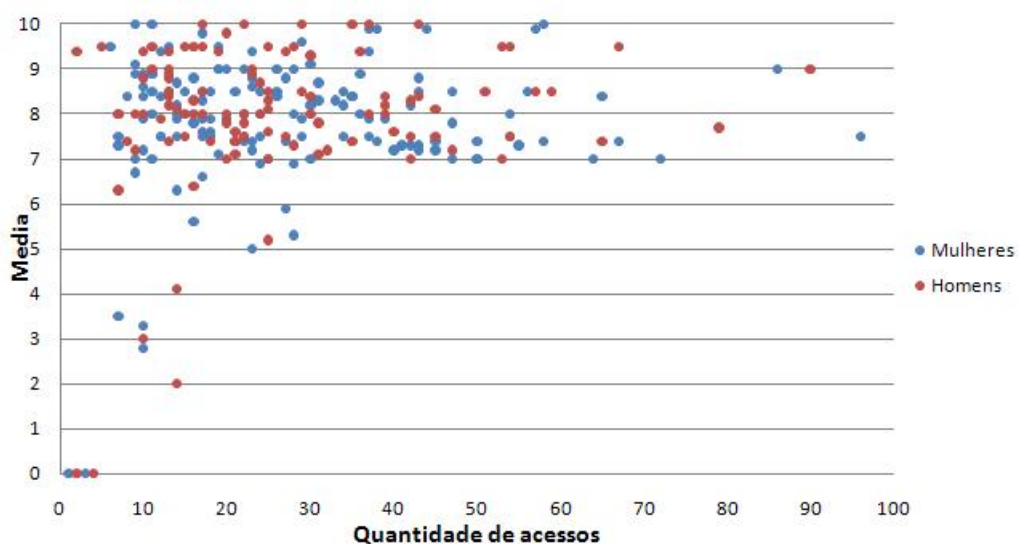


Figura 42 – Relações Internacionais - Médias por sexo

Observa-se que os alunos que obtiveram a nota máxima nessa matéria tem entre 23 e 43 anos, e que nenhum aluno com idade acima de 53 anos teve nota abaixo de 7,0.

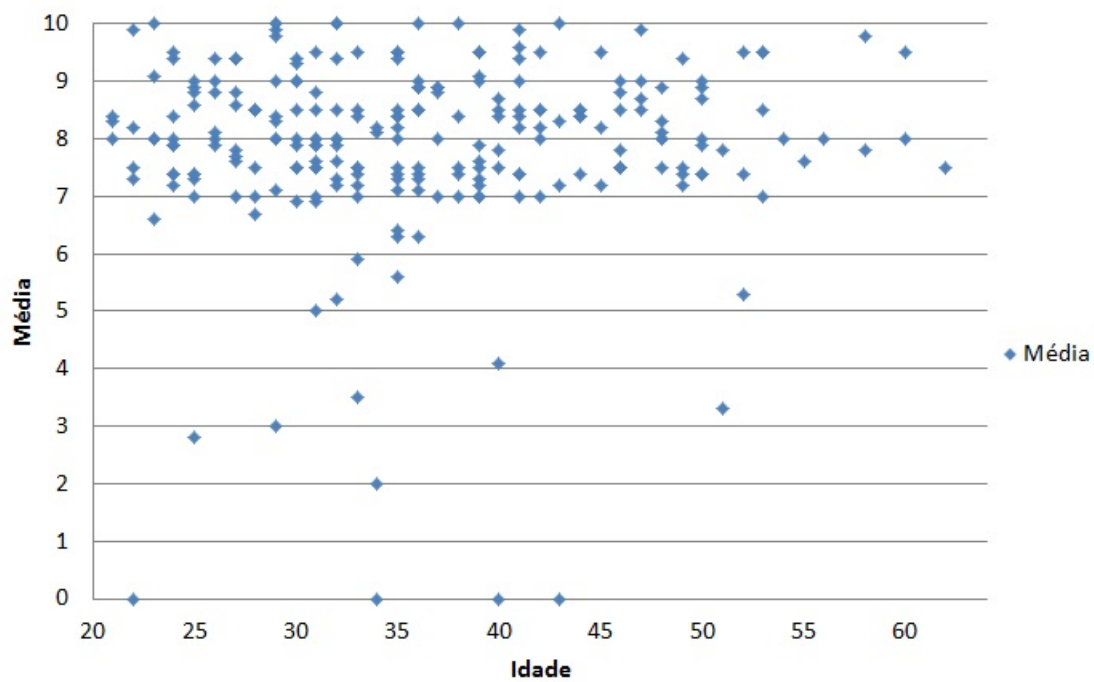


Figura 43 – Relações Internacionais - Médias por idade



**APÊNDICE C – TERMO DE CESSÃO DE DADOS DO AVA**

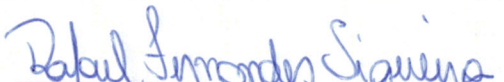
---

Termo de Cessão de Dados NUTEAD

Informamos que estamos cedendo à base de dados do Moodle/NUTEAD-UEPG para RAFAEL FERNANDES SIQUEIRA com finalidade de pesquisa científica.

Informamos que em razão da confidencialidade dos dados cedidos, classificados como informação restrita ou confidencial, estas informações devem ser tratadas com absoluta reserva em qualquer condição e não podem ser divulgadas ou dadas a conhecer a terceiros não autorizados, sem a expressa e escrita autorização da Coordenação do NUTEAD.

Ponta Grossa, 01 de setº 2014 .

  
RAFAEL FERNANDES SIQUEIRA  
RG: 9.427.407-9

  
UNIVERSIDADE ESTADUAL DE PONTA GROSSA  
Núcleo de Tecnologia e Educação Aberta e a Distância  
ELIANE DE FÁTIMA RAUSKI  
Prof. Msc. Eliane de Fátima Rauski  
COORDENADORA GERAL NUTEAD