UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ – UTFPR
DEPARTAMENTO ACADÊMICO DE COMPUTAÇÃO
CURSO SUPERIOR DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

VALDEIR LUIZ DA SILVA

ANÁLISE E MINERAÇÃO DE DADOS DOS CURSOS DE PÓS-GRADUAÇÃO DO ENSINO À DISTÂNCIA DA UTFPR - CÂMPUS MEDIANEIRA

TRABALHO DE DIPLOMAÇÃO

MEDIANEIRA
2015

VALDEIR LUIZ DA SILVA

ANÁLISE E MINERAÇÃO DE DADOS DOS CURSOS DE PÓS-GRADUAÇÃO DO ENSINO À DISTÂNCIA DA UTFPR – CÂMPUS MEDIANEIRA

Trabalho de Diplomação apresentado à disciplina de Trabalho de Diplomação, do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas – COADS – da Universidade Tecnológica Federal do Paraná – UTFPR, como requisito parcial para obtenção do título de Tecnólogo.

Orientador: Prof. M. Sc. Cesar Alfredo Cardoso
Coorientador: Prof. Dr. Arnaldo Candido Júnior

MEDIANEIRA
2015

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Diretoria de Graduação e Educação Profissional
Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas

TERMO DE APROVAÇÃO

Análise e Mineração de Dados dos Cursos de Pós-Graduação do Ensino à Distância da UTFPR - Câmpus Medianeira

Por

Valdeir Luiz da Silva

Este Trabalho de Diplomação (TD) foi apresentado às 13:00h do dia 16 de novembro de 2015 como requisito parcial para a obtenção do título de Tecnólogo no Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas, da Universidade Tecnológica Federal do Paraná, Câmpus Medianeira. O acadêmico foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

Prof. M. Sc. Cesar Alfredo Cardoso
UTFPR – Câmpus Medianeira
(Orientador)

Prof. M. Sc. Glória Patricia Lopez Sepulveda
UTFPR – Câmpus Medianeira
(Convidado)

Prof. M. Sc. Jorge Aikes Junior
UTFPR – Câmpus Medianeira
(Convidado)

Prof. M. Sc. Juliano Rodrigo Lamb
UTFPR – Câmpus Medianeira
(Responsável pelas atividades de TCC)

A folha de aprovação assinada encontra-se na Coordenação do Curso.

RESUMO

DA SILVA, Valdeir Luiz. ANÁLISE E MINERAÇÃO DE DADOS DOS CURSOS DE PÓS-GRADUAÇÃO DO ENSINO À DISTÂNCIA DA UTFPR - CAMPUS MEDIANEIRA. Trabalho de Diplomação (Tecnologia em Análise e Desenvolvimento de Sistemas). Universidade Tecnológica Federal do Paraná. Medianeira 2015.

A grande quantidade de dados produzido pelo uso dos sistemas de informação demanda sistemas adequados para exploração destes dados. A área de mineração de dados tem este propósito, porém é necessário seguir uma metodologia completa e sistemática para não pôr em risco a credibilidade do conhecimento obtido. Algumas técnicas e métodos disponíveis para mineração de dados foram experimentadas e aplicadas ao imenso volume de dados produzidos pelos acadêmicos de pós-graduação na modalidade de ensino à distância da UTFPR do período de agosto de 2014 a agosto de 2015, com objetivo de prever uma possível evasão de um aluno, que utiliza o sistema Moodle em um curso de Ensino à Distância. Constatou-se para este caso singular o melhor classificador como sendo o algoritmo J48 pela razão de obter uma acurácia de 81,29% nas 1048 instâncias analisadas.

Palavras-Chave: Mineração de Dados. Sistema de Descoberta de Conhecimento. *Business intelligence*.

ABSTRACT

DA SILVA, Valdeir Luiz. ANÁLISE E MINERAÇÃO DE DADOS DOS CURSOS DE PÓS-GRADUAÇÃO DO ENSINO À DISTÂNCIA DA UTFPR - CAMPUS MEDIANEIRA. Trabalho de Diplomação (Tecnologia em Análise e Desenvolvimento de Sistemas). Universidade Tecnológica Federal do Paraná. Medianeira 2015.

The large amount of data produced by the use of information systems demand adequate systems for exploiting these data. The data mining area has this purpose, but is must follow a complete and systematic methodology to not jeopardize the credibility of the obtained knowledge. Some methods and techniques available for data mining have been experienced and applied to the immense volume of data produced by postgraduate students in teaching distance mode of UTFPR the period August 2014 to August 2015. The objective to provide for a possible evasion of a student, who uses Moodle system at a Distance Learning course. Was found in this singular case the best classifier as the J48 algorithm considering of it obtained an accuracy of 81.29% in 1048 instances analyzed.

Keywords: Data Mining. Knowledge Discovery System. Business intelligence.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE SIGLAS

ABED	Associação Brasileira de Educação
AMB	Especialização em Gestão Ambiental em Municípios
ARFF	Attribute Relation File Format
AVA	Ambiente Virtual de Aprendizagem
CIE	Especialização em Ensino de Ciências
CRISP-DM	Cross Industry Standard Process for Data Mining
DBM	Data Base Manager
EAD	Ensino à Distância
ELPL	Especialização em Ensino de Língua Portuguesa e Literatura
GPL	General Public License
GAM	Gestão Ambiental em Municípios
GPM	Especialização em Gestão Pública Municipal
IIAE	Especialização em Informática Instrumental Aplicada a Educação
IBM	International Bussines Machine
LMS	Learning Management System
KDD	Knowledge Discovery in Databases
MLP	Mult-layer Perceptron
mSQL	Mini Structured Query Language
PHP	Hypertext Preprocessor
PHP/FI	Hypertext Preprocessor/ Forms Interpreter
SMO	Sequencial Minimal Optimization
UAB	Universidade Aberta do Brasil
UTFPR	Universidade Tecnológica Federal do Paraná
WEKA	Waikato Environment Knowledge Analysis
ROC	Recieiver Operating Characteristic

SUMÁRIO

1 INTRODUÇÃO

O ensino à distância (EAD) tem se tornado uma opção importante para graduações e pós-graduações no Brasil como alternativa de complementar o ensino presencial, em função da evolução das tecnologias de informação e comunicação. A linha que distinguia as duas modalidades está desaparecendo, porém, alguns desafios surgem e são necessárias estratégias e planejamento para vencê-los.

Desta forma como os sistemas de gestão, os meios utilizados para prover o ensino na modalidade a distância geram enorme quantidade de dados aos quais aplicam-se análises e técnicas adequadas possibilitando antecipar tendências de evasão ou de baixo desempenho, ou também identificar o perfil destes estudantes.

São inúmeros os casos de desistência no ensino a distância, essa é uma informação com base no censo da Associação Brasileira de Educação (ABED) o qual estima que número de desistentes dos cursos de EAD no ano de 2013 acrescidos de mais de 5% em relação ao período anterior. Para tentar resolver este problema pode-se utilizar o processo de mineração de dados, definido por Fayyad et al. (1996) como um processo incomum, que possibilite identificar e descobrir nos dados analisados, padrões valiosos, que sejam novos e fáceis de compreender.

A mineração de dados tem sido utilizada em diversas áreas de negócio e tem revelado resultados satisfatórios. De acordo com Olson; Delen (2008) essas são algumas das áreas: Bancária, Medicina, Eleitoral, Segurança, Tomada de decisão, Supermercados, Marketing, agências de viagens.

1.1 OBJETIVO GERAL

Desenvolver um estudo sobre os dados históricos produzidos pelos acadêmicos do EAD da UTFPR-MD durante um período de tempo, neste caso do mês de agosto de 2014 a agosto de 2015 e através deste conhecimento descobrir algumas regras e os padrões, que servirão de suporte aos os gestores no planejamento do ensino a distância na tomada de decisão.

1.2 OBJETIVOS ESPECÍFICOS

Este Trabalho será desenvolvido por meio dos objetivos específicos abaixo:

- Seleção dos dados observando sua natureza;
- Seleção da informação necessária para o conhecimento que se deseja extrair;
- Transformação e preparação dos dados para atender aos métodos de mineração de dados e das ferramentas escolhidas para análise dos dados;
- Mineração dos dados utilizando os algoritmos disponíveis nas ferramentas de mineração de dados.
- Relatório e apresentação dos padrões, indicadores de probabilidade e perfis relevantes.
- Apontar caminhos para novas pesquisas nesta área que poderá servir de apoio a trabalhos futuros.

1.3 JUSTIFICATIVA

Dentre os maiores problemas que as universidades que utilizam a modalidade a distância enfrentam, encontram-se a evasão e a retenção de alunos, problema este de complexa solução. Segundo o censo da Associação Brasileira de Educação ABED (2013), a evasão do ano 2012 foi de 11,74% em relação ao inscritos nos cursos autorizados pelo Ministério da Educação (MEC) e em 2013 foi de 16,94%.

Alguns autores como Amidaci et al. (2004) definem está evasão como sendo a desistência definitiva do estudante em qualquer fase do curso, sem o ter concluído com sucesso.

Segundo Coelho (2010) algumas das causas que sustentam a evasão nos cursos de ensino à distância são: insuficiente domínio técnico do uso do computador especialmente da Internet, falta do contato pessoal entre professor e estudantes, dificuldade de expor ideias numa comunicação escrita e a ausência de um

agrupamento de pessoas em uma instituição física. A evasão é considerada por Abbdad et al. (2010); frequente e crescente nos cursos de ensino à distância sendo ela atrelada a varios fatores como: falta de tempo, dificuldades financeiras, falta de condições de estudo no local de trabalho ou em casa, dificuldade de redigir textos, falta de habilidade para adimistrar o tempo de estudo, falta de habilidade para utilizar os recursos da internet.

Sendo o EAD provido por ambientes virtuais de aprendizagem (AVA), são armazenados registros dos dados gerados pelos estudantes, neste contexto manifesta-se a oportunidade de utilizar estes dados para a descoberta de conhecimento.

Em muitas situações que envolvem a mineração de dados pode ser necessária a construção de um Armazém de Dados. A ideia inicial foi desenvolvida pela IBM em 1960 a qual trazia o nome de *Information Warehouse*, porém o termo Armazém de Dados apenas foi utilizado em 1992 por Willian Harvey Inmon, em seu livro "*Developing the Data Warehouse"*. Kimball (2002) um precursor deste conceito definiu-o como um banco de dados projetado de forma consolidada para consulta e análise, com objetivo de obtenção de informação que possam facilitar na tomada de decisão. Kimball defende o princípio que um Armazém de Dados deve ser desenvolvido para ser compreensível e rápido. Uma definição de armazém de dados foi dita por Inmon (2002), "Coleção de dados orientada a assuntos, integrada, variável no tempo e não volátil, que é usada para o apoio à tomada de decisão".

Mineração de dados para Fayyad et al. (1996) trata-se de como um processo incomum, que possibilite identificar e descobrir nos dados analisados, padrões valiosos, que sejam novos e fáceis de compreender, desta forma é possível compreender que o armazém de dados como o próprio nome já diz é onde armazenam-se os dados e mineração dos dados é o processo para identificar os padrões contidos nese armazém.

O objetivo da mineração de dados é a descoberta de conhecimento escondido atrás de grandes volumes de dados. De acordo com Braga (2005), mineração de dados está inserida em um processo maior chamado de descoberta de conhecimento em banco de dados o (*KDD - Knowledge Discovery in Databases*). Sendo assim definiu-a como um processo para descoberta de conhecimento de grandes bancos de

dados e fez menção à um modelo genérico de etapas para um projeto de “Mineração de dados” como sendo as seguintes:

- Definição do problema;
- Aquisição e avaliação dos dados;
- Extração de características e realce;
- Plano de prototipagem, Prototipagem e desenvolvimento do modelo;
- Avaliação do modelo;
- Implementação;
- Avaliação do retorno do investimento (Pós-projeto).

De acordo com Coordenador do Programa Universidade Aberta do Brasil (UAB) na UTFPR, Prof. M. Sc. Cesar Alfredo Cardoso, esta modalidade de ensino sofre em várias turmas com a evasão. O Diretor elaborou e forneceu um relatório das turmas de 2013, nele foi constatado um alto índice de evasão que pode ser visualizado na Tabela 1. Nela estão separados: por curso, a quantidade de alunos que concluíram as disciplinas, incluídos os que reprovaram, o número de desistentes e o total de pessoas que realizaram a matricula em cada curso.

Tabela 1 – Relatório Alunos EAD UTFPR Turmas 2013.

Curso	*Concluídos*	*Desistentes*[1]	*Matriculados*
Ciências (CIE)	90	90	180
Educação (EDU)	128	112	240
Gestão Ambiental em Municípios (GAM)	85	144	229
Gestão Pública (GP)	93	107	200
Gestão Pública Muncipal (GPM)	44	106	150
Total Geral	440	559	999

Na Figura 1 é possível observar um gráfico onde identifica-se a grande diferença entre os que realizaram a matrícula e os que concluíram o curso.

[1] Inclui reprovados.

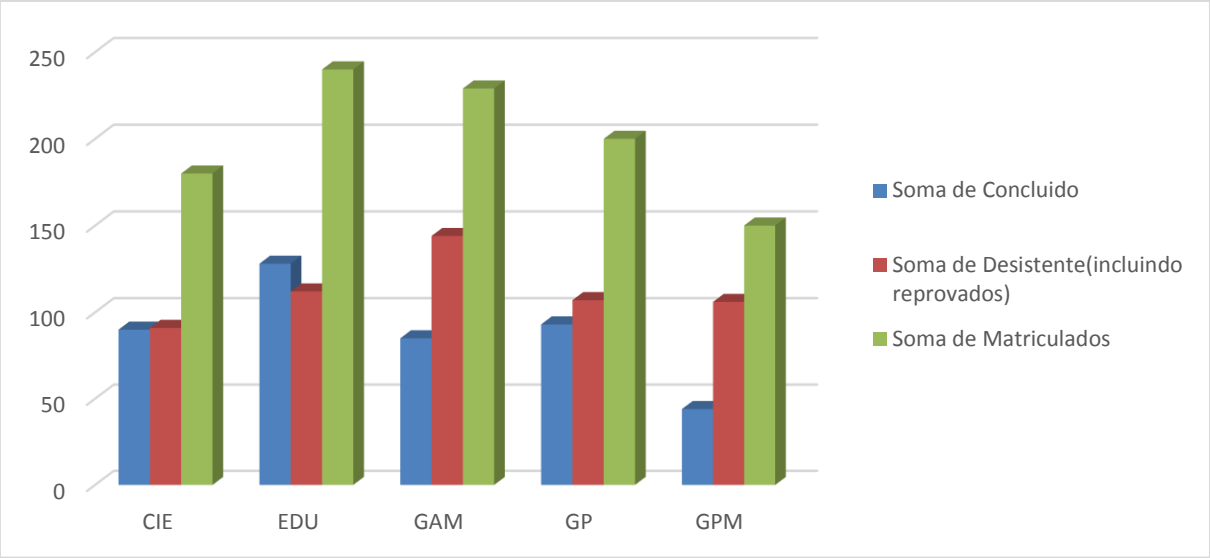


Figura 1 - Gráfico de alunos EAD UTFPR 2013.

1.4 CONTEXTUALIZAÇÃO

Este trabalho englobou a leitura de livros, artigos científicos, buscas na rede mundial de computadores, pesquisa sobre ferramentas disponíveis e tecnologias utilizadas para mineração de dados.

Com o acesso aos dados do Ensino à Distância dos cursos de pós-graduação da UTFPR tornou possível na primeira fase definir a estratégia para extração dos dados do AVA (Ambiente Virtual de Aprendizagem) Moodle. Fez parte da estratégia aproveitar os módulos disponíveis no sistema Moodle, assim facilitaria uma futura agregação ao próprio sistema AVA.

A segunda fase compreendeu a transformação e a organização dos dados obtidos. O formato de arquivo escolhido foi ARFF (*Attribute-Relation File Format*), pois era o formato necessário para interpretação da ferramenta de mineração Weka descrita no capítulo seguinte em fundamentação teórica. Os dados foram extraídos da base de dados através do próprio sistema Moodle, apenas foi realizado uma modificação no módulo de exportação de notas, adequando este módulo para geração dos arquivos que posteriormente seriam utilizados na ferramenta de mineração de

dados. Neste ponto do trabalho foram definidos os atributos e os classificadores os quais foram utilizados na ferramenta Weka.

A terceira fase correspondeu a fase de mineração dos dados, foi optado pela utilização da ferramenta Weka por estar disponível e ser distribuída livremente sobre a licença *General Public License* (GPL). Nesta fase foram utilizados os algoritmos disponíveis na ferramenta Weka para realizar análise sobre os dados o que possibilitou a descoberta de conhecimento.

1.5 ESTRUTURA DO TRABALHO

O Capítulo 2 está contido a fundamentação teórica sobre o trabalho desenvolvido, no Capítulo 3 está contida os materiais e métodos que foram utilizados para no decorrer do trabalho, o Capítulo 4 compreende os resultados obtidos e discussões sobre estes, o Capítulo 5 contém as considerações finais deste trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão relatados os conceitos e as tecnologias utilizadas que tornaram possível a realização deste trabalho.

2.1 MOODLE

Segundo o site oficial[2] Moodle (2015), “Moodle é um sistema on-line gratuito de gestão de aprendizagem, ou LMS”, ele é uma plataforma de aprendizagem projetado para fornecer educadores, administradores e alunos com um único sistema podendo ser integrado ao sistema acadêmico já existente na instituição, com ele é possível criar ambientes de aprendizagem personalizados. O sistema Moodle é disponibilizado sobre a licença de software livre, deste modo esta continuamente em desenvolvimento pela comunidade de inúmeros programadores em todo o mundo, de acordo com o site oficial ele está presente em 155 países e algumas universidades tem ele como base em toda sua estratégia de ensino à distância.

O Moodle é um dos sistemas de gestão de ensino a distância. Ele é desenvolvido na linguagem PHP (*Hypertext Preprocessor*), fornece suporte a vários tipos de banco de dados e sua instalação é possível em qualquer sistema operacional Moodle (2015).

A versão utilizada do sistema Moodle pelo EAD é a 1.9.3, nesta versão existe um módulo de exportação de notas, o qual os professores utilizam para importar arquivos nos formatos: CSV (*Comma-Separated Values*), TXT (*TeXT file*), XML (*Extensible Markup Language*), XLS (*eXceL Spreadsheet*) e ODS (*Open Document Spreadsheet*), deste modo foi realizado uma modificação do *script* PHP da exportação de notas no formato TXT.

---

[2] https://docs.moodle.org/29/en/About_Moodle_FAQ#What_is_Moodle.3F

2.2 PHP

De acordo com a documentação disponível no site oficial[3] The PHP Group (2015), o Hypertext Preprocessor, é uma linguagem de programação que segundo The Php Group (2015), é o sucessor do PHP/FI (*Forms Interpreter*) o qual foi desenvolvido por Ramus Lerdof, sua primeira definição era: simples conjunto de binários escrito em linguagem de programação C, originalmente usado para acompanhamento de visitas para seu próprio currículo online, ele chamou este conjunto de scripts por "*Personal Home Page Tools*", frequentemente referenciado por "PHP Tools". Com o passar do tempo, segundo o mesmo autor, Ramus reescreveu o PHP Tools, incorporando interações com Banco de Dados, fornecendo uma estrutura na qual os usuários poderiam desenvolver aplicações web simples e dinâmicas, como um livro e páginas pessoais. Em 08 de junho de 1995, Ramus abriu o código fonte do PHP Tools para a comunidade, isso permitiu e incentivou os usuários a oferecerem correções e de forma geral aperfeiçoá-lo.

O código teve uma completa reforma, em abril de 1996, Ramus introduziu o PHP/FI. Esta mudança realmente evoluiu o PHP de um conjunto de ferramentas para sua própria linguagem de programação, recebeu status de PHP 2.0. A nova geração incluía suporte aos bancos de dados: DBM (*Data Base Manager*), Msql (*Mini Structured Query Language*), e Postgres95, aceitava cookies, funções de apoio definidas pelo usuário, dentre outras.

Assim de acordo com Morin; Brown (1999) PHP é uma linguagem interpretada por *software* e é desenvolvida na forma de *scripts*.

2.3 DESCOBERTA DE CONHECIMENTO

[3] http://php.net/manual/pt_BR/history.php.php

Para Fayyad (1996), o modelo clássico desenvolvido no propósito de transformação de dados em conhecimento fundamenta-se em um processo manual no qual os especialistas organizam todas as informações e, posteriormente, são analisadas e produzidos relatórios, porém este processo acaba sendo evitado, pelo motivo na maioria das vezes, ser realizado sobre um enorme volume de dados, o que torna este processo inviável. Porém Fayyad relata que o KDD (*Knowledge Discovery in Databases*) surge com o propósito de tentar solucionar o problema da sobrecarga de dados. Para alguns autores como Wang (2005) e Han; Kamber (2006) KDD e mineração de dados significam a mesma coisa, porém para Fayyad (1996) trata o KDD como todo o processo de Descoberta de conhecimento e a mineração de dados como uma tarefa deste processo. A representação do processo KDD feita por Fayyad pode ser visualizado na figura 2.

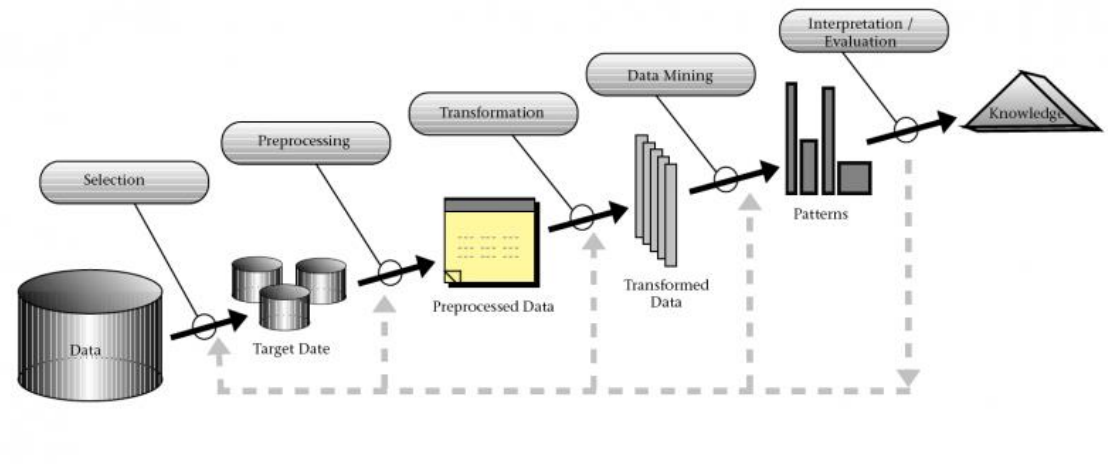


Figura 2 - Processo KDD
Fonte: Fayyad(1996).

Na fase de seleção, definida por Fayyad (1996) é onde ocorre a escolha do conjunto de dados contendo todas as características, atributos, casos e observações que serão utilizados no processo. Levando em consideração a diversidade dos tipos de dados das fontes, isto poderá tornar esta fase bastante complexa, necessitando desenvolver um programa que lide com as diferentes aplicações e os diferentes tipos de dados (Fayyad 1996).

A fase de pré-processamento caracteriza-se pela limpeza dos dados, nesta parte do processo são reparados dados incompletos, eliminados os redundantes e inconsistentes, pois deve-se considerar a qualidade dos dados pois são indispensáveis para determinar um resultado eficiente. Dunkel et al. (1997) destaca a problemática de definir os dados que são inapropriados e quais retirar do conjunto selecionado, pois para isto depende da estrutura e aplicação deste dado.

A transformação dos dados corresponde a fase que os dados são pré-processados, os dados adquiridos nas anteriores necessitam estar em um formato adequados para a próxima fase para que os algoritmos de mineração de dados possam ser aplicados (Fayyad 1996).

A fase de Mineração de dados é onde são aplicadas as técnicas de mineração de dados, aqui utiliza-se os algoritmos de mineração com proposito da descoberta de conhecimento, também são escolhidos quais algoritmos utilizar de acordo com o objetivo da descoberta, a definição de mineração de dados é destacada em 1997 por Berry, M. J. A. ; Linoff, G.:

> Data Mining 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 que as possibilitem montar melhores estratégias de marketing, vendas, suporte, melhorando assim os seus negócios (Berry, M. J. A. ; Linoff, G. 1997).

Na última fase a de interpretação são escolhidas a técnicas para avaliar os resultados obtidos, podendo-se utilizar técnicas e ideias diferentes para diversas áreas por exemplo: administradores de banco de dados, profissionais da área de estatísticas e pesquisadores de inteligência artificial utilizam técnicas diferentes para avaliar os resultados, porém o objetivo é o mesmo: obter informação (Fayyad 1996).

Outro processo de descoberta de conhecimento foi desenvolvido em 1996 pelas empresas que proviam auxilio à clientes que necessitavam ferramentas de mineração de dados. A preocupação destas empresas era desenvolver processos ou metodologias que servissem de suporte e orientação para as aplicações de mineração de dados, sendo assim através de um consórcio firmado entre a empresas e os potenciais compradores das ferramentas de mineração de dados, foi desenvolvido um modelo de processo de mineração de dados chamado de *Cross Industry Standard*

Process for Data Mining, ou simplesmente CRISP-DM, este processo descreve as abordagens que são geralmente adotadas por especialistas em mineração de dados para descoberta de conhecimento, além do mais sua base é o processo KDD de Fayyad (CHAPMAN et al, 1999).

De acordo com Olson (2008) o processo CRISP-DM agrega seis fases de maneira cíclica, e multidirecional podendo-se avançar e retroceder entre elas, podendo ser visualizado na Figura 3.

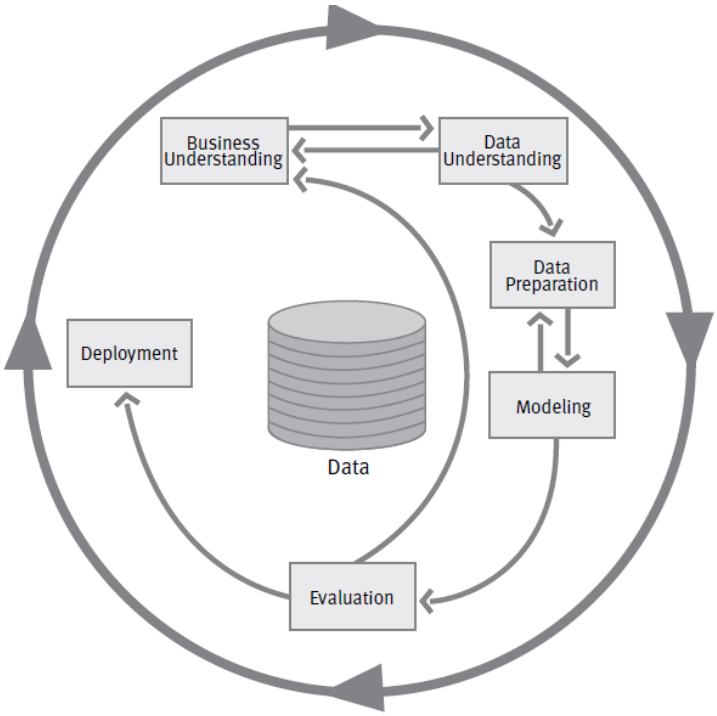


Figura 3 - Processo CRISP-DM
Fonte: Chapman (1999).

Entendimento do negócio trata-se de conhecer sobre aquilo que a organização ou cliente espera da mineração de dados, de acordo com Turban et al. (2009) uma análise de mineração de dados utilizando o processo CRISP-DM pode tomar até 60% do tempo nas fases que compreendem o Entendimento do Negócio e o Entendimento dos Dados.

Levando em consideração que a origem dos dados poderá ser de inúmeras fontes é necessário definir objetivos para conhecimento destes dados. Olson; Delen (2008), relataram em sua obra que para obter o conhecimento dos Dados deve se

proceder objetivando descrever de forma sucinta o problema bem como apontar os dados mais significativos que resultaram na solução do problema e ainda comprovar que as mutações relevantes do problema não sejam independentes.

Pelo fato que os Dados podem ser oriundos de diversas fontes, é normal que estes não estejam prontos para aplicação dos métodos e técnicas da mineração de dados, assim eles poderão sofrer ajustes como de limpeza, combinação de dados ou o preenchimento de valores vazios.

A Modelagem é onde são aplicadas as técnicas de mineração de dados, são escolhidas dentre as disponíveis na ferramenta de mineração, a mais adequada de acordo com o resultado esperado.

Na Avaliação são necessárias: a participação dos integrantes da tomada de decisão, os que tem conhecimento do negócio e especialistas nos dados envolvidos. Através do uso de ferramentas gráficas de visualização e análise dos resultados serão realizados testes e validações para demonstração da confiabilidade dos dados bem como validações conhecidas na área de estatística: validação cruzada, teste de suporte, teste de treinamento, divisão de porcentagem e para auxiliar a análise dos resultados e necessário recorrer taxa de acerto e taxa de erro de instâncias mineradas, a matriz de confusão, erro relativo médio, precisão dentre outros (Han; Kamber, 2005).

Após a execução do modelo estabelecido é necessário que os resultados sejam distribuídos ao envolvidos para assim tomar conhecimento dos resultados.

2.4 WEKA

Weka (*Waikato Environment for Knowledge Analysis*) é um pacote de software desenvolvido usando linguagem de programação Java, pela Universidade de Waikato da Nova Zelândia em 1993 e adquirida por uma empresa chamada *Pentaho* em 2006, porém ele possui versão disponível sobre a licença GPL, permitindo fazer alterações em seu próprio código fonte. Nas palavras de Witten e Frank:

> Este pacote Weka possui uma interface gráfica de usuário que leva você através de tarefas de mineração de dados e possui excelentes ferramentas de visualização de dados que ajudam a compreender os modelos. (WITTEN; FRANK, Prefácio V 2005).

De acordo com os mesmos autores, a ferramenta Weka fornece implementações de algoritmos de aprendizado que são facilmente aplicáveis a um conjunto de dados; o pacote inclui métodos para os problemas considerados padrões em mineração de dados como: regressão, classificação, agrupamento, regras de associação e seleção de atributos, além de facilitar a visualização e fornecer ferramentas para o pré-processamento de dados.

2.4.1 Algoritmo J48

O algoritmo J48 é uma implementação desenvolvida utilizando a linguagem de programação Java do algoritmo C4.5, este por sua vez é definido por Wu et al (2007), como sendo um algoritmo que produz classificadores expressos por árvores de decisão, em cada nó da árvore é avaliado individualmente o significado e a existência de cada atributo, ou seja o algoritmo escolhe um atributo que melhor subdivide o conjunto das amostras em subconjuntos, para ressaltar a informação contida nesta escolha e subdivisão (Quilan, 1993; HALL et al., 2009).

A construção da árvore do classificador J48 é feita do topo para a base, por outro lado este algoritmo também pode construir classificadores na forma de regras compreensíveis. Ele é considerado o melhor dispõe o resultado na quanto a montagem de árvores de decisão, com base em um conjunto de dados de treinamento (Witten; Frank, 2005).

2.4.2 Algoritmo Part

De acordo com Frank; Witten (1993), é um algoritmo que utiliza o método separar para conquistar, ele produz um conjunto ordenado de regras chamadas de

“listas de decisão”, cada vez um novo dado é comparado com cada regra da lista e o item é atribuído a categoria da primeira regra de correspondência.

Em outras palavras o algoritmo J48.Part constrói em cada interação uma árvore de decisão parcial do algoritmo J48 e deixa a melhor regra em uma folha desta árvore, as regras com um índice de cobertura mais alto são apresentadas para o usuário e as restantes são descartadas, assim ele produz regras a partir de uma árvore de decisão como o J48 (Frank; Witten, 1993).

Este algoritmo fornece um resultado claro e simples porque são apresentadas em forma de condições para o usuário (Rezende, 2003).

2.4.3 Algoritmo SMO

Para explicar o algoritmo SMO (*Sequencial Minimal Optimization*) é necessário falar do SVM (*Support Vector Machine*) conforme Cortes, Vapnik (2005) relatam, o SVM como sendo um conjunto de métodos de aprendizado supervisionado que através de uma análise sobre estes dados reconhece tendências, podendo ser usado em técnicas de classificação e para análise de regressão.

De acordo com Wintten; Frank (2005), o algoritmo SVM pertence aos classificadores lineares desenvolvidos por Vapnik para solução de problemas de classificação reconhecimento de padrões, seu objetivo é reduzir riscos estruturais, com o mínimo de erros na classificação empírica e o máximo de margem geométrica entre as instâncias. Ainda estes mesmos autores definem o conceito do algoritmo SMO como: desenvolvido por John Platt, implementado utilizando a lógica do algoritmo SVM no software Weka e utiliza memória linear para treinamento, permitindo gerenciar um grande número de arquivos de treinamento.

2.4.4 Algoritmo MLP

O algoritmo MLP (*Multi-Layer Perceptron*) algoritmo utiliza um modelo que mapeia conjuntos de dados de entrada em relação a um conjunto de saídas

apropriadas. Em outras palavras ele é composto por várias camadas de nós em um grafo direcionado e uma destas camadas totalmente conectado ao próximo elemento do grafo, com exceção aos nós de entrada, cada nó é um neurônio artificial, que processa de forma não linear seus valores de entrada, resultando em uma saída (Vellasco, 2007).

O MLP utiliza uma aprendizagem supervisionada chamada de retro propagação para treinar a rede, é uma modificação do padrão linear *perceptron* e pode distinguir os dados que são linearmente inseparáveis.

2.4.5 Medidas de Avaliação

As medidas de avaliação podem ser obtidas da Matriz de Confusão, nela estão incluídos dados sobre as classificações reais dos dados analisados e as previstas pelo algoritmo de classificação (VISA et al., 2011). Um exemplo desta matriz é demonstrado no Tabela 2.

Tabela 2 – Exemplo de Matriz de Confusão.

		Previsão	
		Negativo	Positivo
Atual	Negativo	VP	FP
	Positivo	FN	VN

Ainda de acordo com Visa, S. et al., 2011, a definição para cada uma das medidas de avaliação é:

Taxa de verdadeiro positivo (TP Rate): taxa de verdadeiros positivos, ou seja, casos classificados corretamente.

Taxa de falso positivo (*FP Rate*): taxa de falsos positivos, casos falsamente classificados como uma determinada classe.

Precisão (*Precision*): a precisão corresponde a proporção de casos que são de uma classe dividido pelo total de casos classificados como aquela classe.

Medida F1 (*F-Measure*): refere-se a medida combinada entre a precisão e a cobertura.

Cobertura (*Recall*): proporção de casos classificados como uma determinada classe dividido pelo total real nessa classe (equivalente a taxa VP).

Estas medidas são definidas pelas seguintes fórmulas:

- Taxa de verdadeiro positivo: VN / (FN + VN).
- Taxa de falso positivo: FP / (FN + VN).
- Precisão: VN / (FP + VN).
- Medida F1: 2 * precisão * cobertura / (precisão + cobertura).
- Cobertura: VN / (FN + VN).

Para Margotto, P. R. (2010), a medida: Curva ROC é utilizada na área de comunicações como uma demonstração entre o sinal e o ruído, interpretando-se o sinal como os verdadeiros positivos e o ruído como os falsos positivos.

3 MATERIAL E MÉTODOS

Este capítulo apresenta os materiais e métodos utilizados no desenvolvimento deste trabalho. A utilização desses no desenvolvimento da mineração de dados resultou na descoberta de algumas regras de classificação, dentre outras informações que poderão ser utilizadas em uma aplicação para auxílio na tomada de decisão.

3.1 DESENVOLVIMENTO DOS SCRIPTS PHP

Foram desenvolvidos scripts PHP para a aquisição dos dados através do sistema Moodle, assim fazendo uso do modulo de exportação de notas era executado o script em cada curso, sendo eles: Especialização em Gestão Ambiental em Municípios (AMB), Especialização em Ensino de Ciências (CIE), Especialização em Ensino de Língua Portuguesa e Literatura (ELPL), Especialização em Informática Instrumental Aplicada a Educação (IIAE), Especialização em Gestão Pública Municipal (GPM). O script realizou consultas no banco de dados do Moodle trazendo informações dos alunos de todos os cursos disponíveis, e adaptou e organizou os dados na forma dos atributos que ficaram definidos como:

- “media_notas”;
- “media_semanal_cliques”;
- “porcentagem_tarefas”;
- “dia_mais_acessado”;
- “dia_mais_tarefas_entregues”;
- “curso”;
- “mes”.
- “desistente”

Estes atributos são os atributos encontrados no arquivo ARFF, parte deste arquivo pode ser visualizado no Apêndice C. Os atributos foram definidos conforme debates feitos entre: Prof. M. Sc. Cesar Alfredo Cardoso, Prof. Dr. Arnaldo Candido

Júnior e algumas pessoas envolvidas com o suporte técnico dos cursos do EAD. O atributo “media_notas” traz a média do aluno desde seu ingresso no curso até o mês atual, o atributo “media_semanal_cliques” refere-se a média semanal do registro de log de cada aluno, o atributo “dia_mais_acessado” tem a informação de qual dia é mais acessado pelo aluno, o atributo “dia_mais_tarefas_entregues” registra o dia em que o estudante mais entrega as tarefas, o atributo “curso” define de forma abreviada qual curso pertence aquele aluno, o atributo “mes” informada de 1 a 13 qual o mês que está atrelado aquela linha do arquivo, e o atributo “desistente” registar com o caracter “y” se aquele aluno é ou com “n” caso o aluno não seja desistente.

Os atributos têm relevada importância na predição de um desistente como pode ser observado a seguir: quando o atributo “media_das_notas” está com um índice baixo é um indicativo que o aluno está desmotivado e prestes a desistir. Para o atributo “media_semanal_cliques”: representando uma contabilidade baixa sugere que o aluno está pouco ativo no contexto do curso. Em relação ao atributo “mês”: alguns meses podem ser considerados mais difíceis para o aluno, a exemplo dos de prova de prova caracterizando talvez um fator desmotivador. Quanto ao atributo “dia_mais_acessado”: eventualmente o aluno que deixa para acessar a plataforma no último dia de entrega das tarefas tem a predisposição de realiza-la na pressa em consequência disso ele recebe notas menores puxando sua média para baixo.

De maneira geral o processo de implementação ficou definido conforme a figura 4.

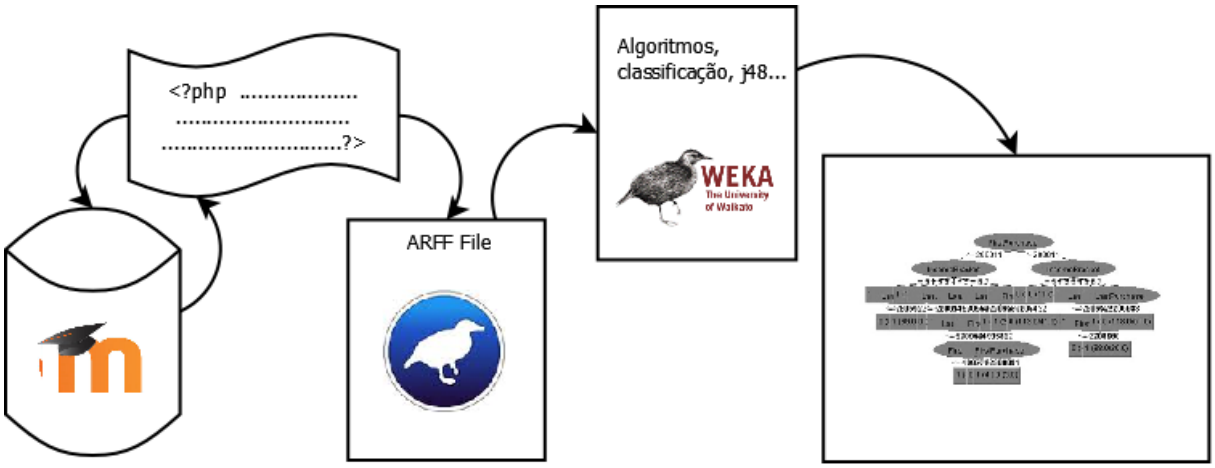


Figura 4 - Processo de implementação.

O processo de implementação compreendeu a importação dos dados por meio dos scripts PHP, através destes scripts foi produzido também o arquivo ARFF, na etapa seguinte foi utilizado a ferramenta Weka para o uso dos algoritmos de descoberta de conhecimento, para na seguinte realizar a análise sobre os resultados obtidos.

4 RESULTADOS E DISCUSSÕES

Neste capítulo será apresentada os resultados obtidos com a etapa de mineração de dados e o que foi necessário para que a resultado deste processo fosse satisfatório.

4.1 MINERAÇÃO

Após a utilização dos scripts para importação e a transformação dos dados foi iniciada a fase mineração de dados. Em parte do processo de Mineração foi realizado o pré-processamento dos dados.

A respeito do arquivo ARFF produzido ressalta-se o seguinte: os alunos desistentes estavam sendo registrados junto com os reprovados nos dados encontrados no sistema Moodle, assim como objetivo era tentar prever uma possível desistência era necessário a separação. Desta forma os alunos desistentes estão separados dos reprovados e os reprovados estão junto com os aprovados no arquivo ARFF.

O período analisado foi 1 ano, que compreendeu de agosto de 2014 a agosto de 2015. No arquivo está contido um grupo de cada mês um registro de cada aluno. Já os alunos que concluíram o curso aparecem 12 vezes no arquivo como não desistente. Outra situação é com o aluno que desiste no sexto mês do curso, seus dados aparecem no arquivo ARFF uma vez como desistente e os 5 meses anteriores como não desistente. Quanto a outros períodos eles são análogos ao caso de 6 meses.

Foi necessário balancear este conjunto de dados do ARFF, uma classe como a de não desistentes sendo predominante em relação a outra, pode gerar um vício em alguns algoritmos, como a que foi explicado no parágrafo anterior. Assim de acordo com pesquisas realizadas o filtro “*resample”* poderia contornar este problema.

Para o balanceamento dos dados foi utilizando o filtro de instância supervisionado “*resample*” disponível na ferramenta Weka. Nas propriedades do filtro

foi definido o fator *Bias* para distribuição uniforme de classe (*biasToUniformClass*) em: 100, *noReplacement* escolheu-se a opção: “*True*” para não substituir os dados, e o “*sampleSizePercent*” em 7%.

Em outras palavras este balanceamento faz com que a lógica dos algoritmos seja realizada apenas em parte das instâncias, o que equilibrou a comparação entre 496 desistentes e 552 não desistentes.

Considerando que a ferramenta Weka traz valores padrões razoáveis para a maioria dos algoritmos testados, optou-se pelo uso dos valores fornecidos.

Para este conjunto de dados o algoritmo de classificação de árvores J48 obteve uma acurácia de 81,29% correspondendo a 852 instâncias, de certa forma um desempenho razoavelmente bom para estes dados específicos, já para outros dados este algoritmo poderá apresentar uma performance contraria. A árvore de decisão deste algoritmo está disponível no Apêndice A. O Quadro 1 representa a matriz de confusão do J48. A matriz de confusão do J48 demonstra que o algoritmo classificou 83 das 413 instâncias de desistentes incorretamente e 113 das 439 instâncias de não desistentes.

Ao selecionar apenas alguns atributos: “media_notas”, “media_semanal_cliques”, “porcentagem_tarefas”, “dia_mais_acessado” e “desistente” identificou-se uma piora no rendimento de acurácia do algoritmo para 72,51% representando 760 instâncias classificadas corretamente e 27,48% equivalendo 288 instâncias incorretas, este resultado pode-se visualizar no Apêndice A.

```
=== Matriz de Confusão ===

   a   b   <-- classified as
 413  83 |   a = y
 113 439 |   b = n
```

Quadro 1 - Matriz de Confusão J48.

Um detalhamento da acurácia do algoritmo J48 incluindo: A taxa de verdadeiro positivo, o falso positivo, a precisão, a cobertura, a medida-f e a curva ROC (*Recieiver Operating Characteristic*), elas estão definidas por classe para este algoritmo visualização está disponível na Tabela 3. Pode-se observar que este algoritmo privilegiou levemente a cobertura.

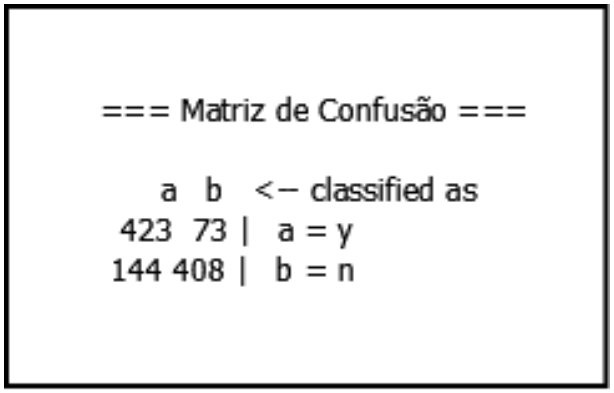
Tabela 3 - Acurácia detalhada por classe

Taxa VP	Taxa FP	Precisão	Cobertura	Medida F1	Curva ROC	Classe
0.833	0.205	0.785	0.833	0.808	0.843	desistente
0.795	0.167	0.841	0.795	0.818	0.843	não-desistente
0.813	0.185	0.815	0.813	0.813	0.843	

Já o algoritmo de classificação de regras Part atingiu aproximadamente 79,29% de acurácia e classificou 20,70% das instâncias incorretamente, desempenho um pouco inferior se comparado com o J48. Parte do relatório gerado pela ferramenta Weka para este algoritmo encontra-se no Apêndice B onde define a regra deste algoritmo. O Quadro 2 representa a matriz de confusão do algoritmo Part.

Na matriz de confusão do algoritmo Part é possível visualizar a incorreta classificação de 73 das 423 instâncias dos desistentes e 144 das 408 dos não desistente.

Ao retirar os atributos: “dia_mais_tarefas_entregues”, “curso” e o atributo “mes” e realizar a execução do algoritmo Part percebeu-se a redução para 69,65% de acurácia representando 730 das instâncias e 318 instâncias equivalente a 30,34% das instâncias classificadas incorretamente, sua lista de regras pode-se visualizar no Apêndice B.



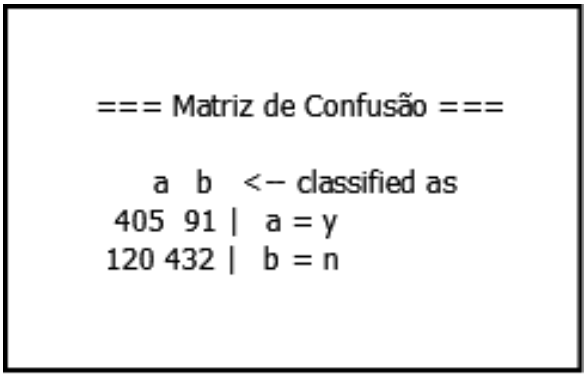
Quadro 2 - Matriz de confusão algoritmo Part.

O detalhe da acurácia do algoritmo de regras Part está detalhado na Tabela 4, onde visualiza-se: A taxa de verdadeiro positivo, o falso positivo, a precisão, a cobertura, a medida F1 e a curva ROC. Na Tabela 4 também pode se observar que o algoritmo Part privilegiou levemente a cobertura.

Tabela 4 - Detalhamento da acurácia do algoritmo Part

Taxa VP	Taxa FP	Precisão	Cobertura	Medida F1	Curva ROC	Classe
0.853	0.261	0.746	0.853	0.796	0.831	desistente
0.739	0.147	0.848	0.739	0.79	0.831	não-desistente
0.793	0.201	0.8	0.793	0.793	0.831	

O algoritmo de classificação de funções *Multi-layer Perceptron*, classificou com a melhor calibragem para este caso corretamente 837 instâncias o que corresponde à 79,86% do total, veja na Tabela 5, e classificou incorretamente 211 sendo 20,13% do total, pode-se reparar que também atingiu menos acurácia que o J48. Pode-se visualizar a matriz de confusão deste algoritmo no Quadro 3, nela destaca-se a que o algoritmo classificou 91 das 405 dos desistentes incorretamente, e 120 dos 432 não desistentes incorretamente.



Quadro 3 - Matriz de confusão *MultilayerPerceptron.*

Detalhamento da acurácia por classe na Tabela 6 para o algoritmo *MultilayerPerceptron*.

No caso particular do algoritmo de redes neurais *Multi-layer Perceptron*, em uma abordagem exploratória realizados testes e definidos os valores disponíveis na Tabela 5.

Tabela 5 - Testes algoritmo MLP

Taxa de aprendizagem	***Momentum***	**Tamanho do conjunto de validação**	**Número de neurônios escondidos**	**Acurácia**
0.1	0.0	0(1)	3	78,53%
0.3(1)	0.2(1)	10	6	78,24%
0.6	0.4	20	10	78,96%
0.3(1)	0.2(1)	0(1)	Automático(1)	79,86%
0.1	0.2(1)	0(1)	Automático(1)	77,95%
0.6	0.2(1)	0(1)	Automático(1)	79,00%
0.3(1)	0.0	0(1)	Automático(1)	79,48%
0.3(1)	0.4	0(1)	Automático(1)	79,85%
0.3(1)	0.2(1)	10	Automático(1)	79,10%
0.3(1)	0.2(1)	20	Automático(1)	79,58%
0.3(1)	0.2(1)	0(1)	3	76,52%
0.3(1)	0.2(1)	0(1)	6	78,43%
0.3(1)	0.2(1)	0(1)	10	79,19%

Notas:
(1) valores padrões do calibrador.

Estas mudanças na calibração do algoritmo resultaram em uma piora no índice de acerto ficando sempre entre 76,5% e 79,8%, deste modo optou-se pelo uso dos valores padrões de calibragem que produziram 79,86% de acerto e os resultados da Tabela 6.

Tabela 6 - Detalhamento da acurácia por classe MLP

Taxa VP	Taxa FP	Precisão	Cobertura	Medida F1	Curva ROC	Classe
0.817	0.217	0.771	0.817	0.793	0.86	desistente
0.783	0.183	0.826	0.783	0.804	0.86	não-desitente
0.799	0.2	0.8	0.799	0.799	0.86	

Utilizando o *Kernel PolyKernel o* algoritmo de função SMO atingiu uma acurácia de 76,43% classificando 801 instâncias corretamente e 23,56% o que representou 247 instâncias incorretamente, o que ficou abaixo da eficiência do algoritmo J48. Pode-se visualizar a matriz de confusão para este algoritmo no quadro 4, foram classificadas 117 instâncias de alunos desistentes incorretamente e 130 dos não desistentes incorretamente.

```
=== Matriz de Confusão ===

   a   b   <-- classified as
 379 117 |   a = y
 130 422 |   b = n
```

Quadro 4 - Matriz de confusão SMO.

Pode-se verificar a acurácia detalhada para o algoritmo SMO na Tabela 7.

Tabela 7 - Acurácia detalhada algoritmo SMO

Taxa VP	Taxa FP	Precisão	Cobertura	Medida F1	Curva ROC	Classe
0.764	0.236	0.745	0.764	0.754	0.764	desistente
0.764	0.236	0.783	0.764	0.774	0.764	não-desistente
0.764	0.236	0.765	0.764	0.764	0.764	

4.2 DISCUSSÃO DE RESULTADOS

Vale ressaltar que foi necessário o balanceamento dos dados pela forma com que estavam os dados no arquivo ARFF: cada aluno concluinte é visualizado 12 vezes no arquivo (inclusive reprovados), cada aluno desistente no primeiro mês conta uma vez no arquivo como desiste, cada aluno desistente no segundo mês conta duas vezes: uma vez como desistente e uma como não desistente, o aluno que desiste no sexto mês conta 5 vezes como não desistente e uma vez como desistente. Tudo isso gerou forte desbalanceamento para a classe não desistente.

De acordo com a descoberta de conhecimento obtida neste estudo, o algoritmo J48 teve melhor eficiência, foi possível observar que se o período de tempo analisado for menor ou igual aos 4 primeiros meses de curso, o classificador J48 não obteve bons resultados em colocar 123 dos registros como desistentes onde apenas 17 realmente foram. Este problema pode ser atribuído ao volume de dados que foi analisado, porém o classificador encontrou uma regra com bastante acerto para alguns cursos, por exemplo para o curso AMB: considerando os 4 primeiros meses do curso, media de notas sendo maior que 7.46, a média semanal de cliques ser menor que 68,1 o aluno é desistente. O algoritmo acertou todas as instâncias com essa situação. Em outras palavras significa que mesmo o aluno obtendo uma média razoável, porém com pouca atividade na plataforma de aprendizagem ele pode se tornar um possível desistente. Em outra situação: caso o mês seja maior ou igual ao nono mês do curso, o curso seja CIE, a média das notas maior menor ou igual a 7,54, o dia com mais frequência de entrega de tarefas seja menor ou igual a quarta-feira, e a média das notas menor ou igual a 1,43 o aluno é considerado desistente, repara-se

algo já esperado que se a média do aluno for muito baixa ele tem grande chance de evadir do curso, o que pode ser feito neste caso é gerar notificações através do sistema para o aluno que obteve nota muito baixa em algumas atividades, ou até mesmo para os tutores para contatar sobre as dificuldades encontradas pelo aluno em tal tarefa, propondo assim alguns estudos para compensar em futuras avalições.

```
mes <= 4
|   curso = ELPL: y (123.0/17.0)
|   curso = IIAE
|   |   dia_mais_tarefas_entregues <= 3
|   |   |   dia_mais_tarefas_entregues <= 1
|   |   |   |   dia_mais_tarefas_entregues <= -1
|   |   |   |   |   media_notas <= 0.41: n (3.0)
|   |   |   |   |   media_notas > 0.41: y (4.0/1.0)
|   |   |   |   dia_mais_tarefas_entregues > -1: n (13.0/1.0)
|   |   |   dia_mais_tarefas_entregues > 1
|   |   |   |   mes <= 2: y (2.0)
|   |   |   |   mes > 2
|   |   |   |   |   media_notas <= 0.88: y (2.0)
|   |   |   |   |   media_notas > 0.88: n (5.0/1.0)
|   |   dia_mais_tarefas_entregues > 3: n (8.0)
|   curso = CIE
|   |   dia_mais_tarefas_entregues <= 3: y (81.0/10.0)
|   |   dia_mais_tarefas_entregues > 3: n (7.0/1.0)
|   curso = AMB
|   |   media_notas <= 7.46: y (78.0/15.0)
|   |   media_notas > 7.46
|   |   |   mes <= 2
|   |   |   |   media_semanal_cliques <= 68.17: y (3.0)
|   |   |   |   media_semanal_cliques > 68.17: n (2.0)
|   |   |   mes > 2: n (6.0)
```

Quadro 5 - Amostra árvore J48

O Quadro 5 representa parcialmente a árvore de decisão do J48, a árvore completa está disponível no Apêndice A.

Por outro lado, observou-se que os alunos utilizaram a terça-feira em todos os cursos como o dia que mais se entregou a atividade tarefa conforme a figura 5, considerando que o prazo para entrega das tarefas é semanal: ele inicia segunda-feira as 8h00m e finalizam domingo as 23h00m, contrariando o esperado que fosse

domingo o dia que teria maior fluxo de entrega de tarefas. Isso pode auxiliar os responsáveis pela estrutura do servidor em fazer algumas modificações para este dia da semana ou até mesmo a equipe de suporte ficar de prontidão para uma demanda maior de situações adversas. Na figura 5 o número: -1 significa nenhum, 0 domingo, 1 segunda-feira, 2 terça-feira, 3 quarta-feira, 4 quinta-feira, 5 sexta-feira e 6 para sábado.

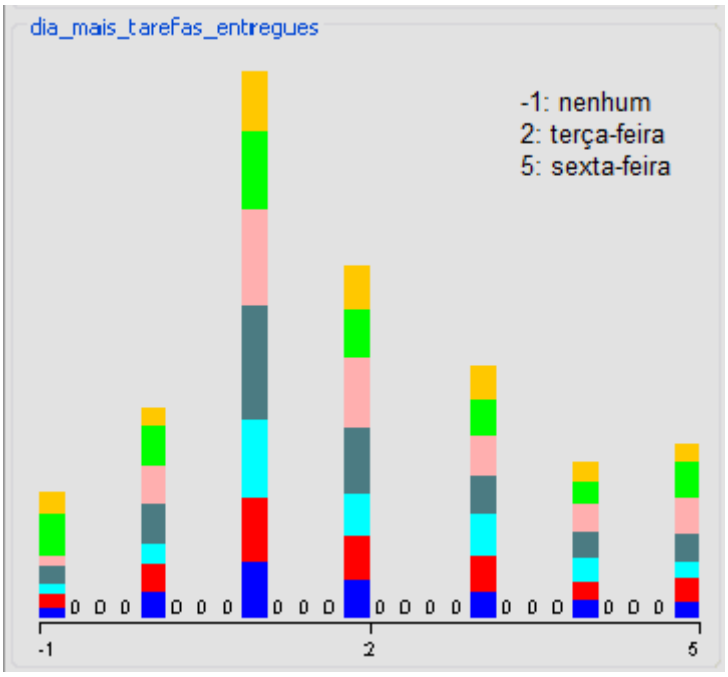


Figura 5 - Gráfico dia com mais tarefas entregues.

Considerando a dificuldade de obter dados de mais períodos, os resultados foram considerados promissores pelo coordenador do programa UAB (Universidade Aberta do Brasil) na UTFPR.

5 CONSIDERAÇÕES FINAIS

Neste capítulo serão apresentadas as conclusões do trabalho realizado, bem como as melhorias que poderão ser realizadas no objetivo de obter melhores resultados na descoberta de conhecimento.

5.1 CONCLUSÃO

Neste trabalho, dados da plataforma Moodle foram extraídos e resultados para criar 4 algoritmos classificadores. O classificador J48 foi o melhor no conjunto de dados analisados, o algoritmo SMO foi considerado o pior de acordo com a performance obtida. Os classificadores podem ser aplicados por docentes da instituição para descobrir regras ou ajudar na descoberta da causa de evasão.

Os pontos fortes do trabalho são: a identificação de possíveis tendências na área de ensino a distância o que permite o tomador de decisão em antecipar movimentos que trarão benefício à instituição, a identificação do dia da semana com maior fluxo de envio de atividades e a referência como um modelo de estratégia na descoberta de conhecimento sobre este tipo de dado. Os pontos fracos foram: a dificuldade de obtenção dos dados o que não possibilita novos testes para validações do conhecimento descoberto.

O objetivo deste trabalho foi alcançado, pois as técnicas empregadas resultaram em descoberta de algumas regras e padrões. Pelo número restrito de dados que tínhamos disponíveis pode ser que estas regras necessitem uma nova abordagem para aumentar sua confiabilidade

5.2 TRABALHOS FUTUROS/CONTINUAÇÃO DO TRABALHO

É proposto para a continuação deste trabalho a obtenção de um volume maior de dados para poder realizar teste de treinamento de dados e também inclusão de outros atributos que possam melhorar a descoberta de conhecimento, isso resultara em uma melhora e aumento na confiabilidade dos resultados. Também pode ser interessante desenvolver umas algumas adaptações para atender outras versões do sistema Moodle ou para outro sistema AVA que se tenha necessidade, pois pode ser que outras instituições utilizem um sistema distinto do Moodle Como os cursos do EAD da UTFPR são anuais, pode-se construir um *Data Warehouse* com os dados de vários anos de graduações fortalecendo as regras que poderiam ser descobertas.

REFERÊNCIAS BIBLIOGRÁFICAS

BERRY, M. J. A.; LINOFF, G. **Data Mining Techniques**: For Marketing, Sales, and Customer Support, Wiley Computer Publishing, New York , 1997.

CHAPMAN, P. et al, **CRISP-DM 1.0** Step-by-step data mining guide, 1999. Disponível em <http://www.statoo.com/CRISP-DM.pdf>. Acesso em:29 de outubro de 2015.

CORTES, C; VAPNIK, V. **Support vector networks**, Machine Learning, n.20, p. 273–296, Kluwer Academic Publishers, 1995. Disponível em <http://csee.wvu.edu/~xinl/library/papers/comp/ML/svm.pdf>. Acesso em: 25 de outubro de 2015.

DAVID, M. W, **Evaluation: From Precision, Recall And F-Measure To Roc, Informedness, Markedness & Correlation**, Australia, 2011. Disponível em < http://www.bioinfopublication.org/files/articles/2_1_1_JMLT.pdf> Acesso em 31 de outubro de 2015.

DUNKEL, B. et al, **Systems for KDD**: From concepts to practice. Future Generation Computer Systems, n. 13, p. 231-242, 1997.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. From Data Mining to Knowledge Discovery in Databases**. AI Magazine**, v. 17, p. 37-54, 1996. Disponível em:<http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230/1131> Acessado em: 16 de junho de 2015.

FRANK, E.; WITTEN, I. H., **Generating Accurate Rule Sets Without Global Optimization**, p.144--151, Morgan Kaufmann, 1998. Disponível em <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.47.6014&rep=rep1&type=pdf>. Acesso em: 25 de outubro de 2015.

HALL, M.et al. **The Weka Data Mining Software: an update**, SIGKDD Explorations Newsletter, v.11, n.1, p.10-18, 2009.

HAN, J; KAMBER, M., **Data Mining:** Concepts and Techniques. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005.

INMON, W. H., **Building the Data Warehouse**, 3ª edição, New York: John Wiley & Sons, 2002

KIMBALL, R.; ROSS, M.. **The Data Warehouse Toolkit:** The Complete Guide to Dimensional Modeling. Segunda edição, John Wiley & Sons, Inc., New York, NY, USA, 2002.

MARGOTTO, P. R., **Curva ROC**: Como fazer e interpretar no SPSS, 2010. Disponível em:<http://www.paulomargotto.com.br/documentos/Curva_ROC_SPSS.pdf>. Acesso em 26 de outubro de 2015.

MOODLE. **About Moodle FAQ**. Disponível em: <https://docs.moodle.org/29/en/About_Moodle_FAQ#What_is_Moodle.3F> Acesso em 25 Agosto de 2015.

MORIN, R.; BROWN, V., **Scripting Languages**, 1999. Disponível em: <http://www.mactech.com/articles/mactech/Vol.15/15.09/ScriptingLanguages/index.html>. Acesso em: Agosto de 2015.

OLSON, D., L.; DELEN, D. **Advanced Data Mining Techniques**, Primeira edição Springer Publishing Company Incorporated, 2008.

QUINLAN, J. R., C4.5: **Programs for Machine Learning**, Morgan Kaufmann Publishers, Sydney, Australia, 1993

REZENDE, S. O. **Sistemas Inteligentes**: Fundamentos e Aplicações. Barueri, SP: Manole, 2003.

THE PHP GROUP. **PHP:** História do PHP - Manual -PHP Hypertext Preprocessor, 2015. Disponível em: <http://php.net/manual/pt_BR/history.php.php> Acessado em 6 de Setembro de 2015.

TURBAN, E; SHARDA ,R; ARONSON, J., E.; KING, D., **Business Intelligence**: Um Enfoque Gerencial para a Inteligência do Negócio, Bokman, Porto Alegre, RS, 2009.

VELLASCO, M. M. B. R., **Redes Neurais Artificiais**, Laboratório ICA- Inteligéncia Computacional Aplicada PUC, Rio de Janeiro, RJ, 2007. Disponível em < http://www2.ica.ele.puc-rio.br/Downloads/33/ICA-introdu%C3%A7%C3%A3o%20RNs.pdf> Acesso em 29 de outubro de 2015.

VISA, S. et al, **Confusion Matrix-based Feature Selection**, p. 120-127, 2011. Disponível em < http://ceur-ws.org/Vol-710/paper37.pdf> Acesso em 31 de outubro de 2015.

WANG, J. **Encyclopedia of Data Warehousing and Mining**. Segunda edição, Information Science Reference - Imprint of: IGI Publishing, Hershey, PA, 2008.

WITTEN, I. H.; FRANK, E., **Data Mining**: Practical Machine Learning Tools and Technique, Segunda edição, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

WU, X., et al, **Top 10 algorithms in data mining**, Springer-Verlag London Limited, 2007. Disponível em: <http://www.cs.umd.edu/~samir/498/10Algorithms-08.pdf>. Acesso em: 25 de outubro de 2015.

APÊNDICES

APÊNDICE A – Árvore do normal do algoritmo J48 e árvore com seleção de atributos

Árvore normal J48

```
mes <= 4
|   curso = ELPL: y (123.0/17.0)
|   curso = IIAE
|   |   dia_mais_tarefas_entregues <= 3
|   |   |   dia_mais_tarefas_entregues <= 1
|   |   |   |   dia_mais_tarefas_entregues <= -1
|   |   |   |   |   media_notas <= 0.41: n (3.0)
|   |   |   |   |   media_notas > 0.41: y (4.0/1.0)
|   |   |   |   dia_mais_tarefas_entregues > -1: n (13.0/1.0)
|   |   |   dia_mais_tarefas_entregues > 1
|   |   |   |   mes <= 2: y (2.0)
|   |   |   |   mes > 2
|   |   |   |   |   media_notas <= 0.88: y (2.0)
|   |   |   |   |   media_notas > 0.88: n (5.0/1.0)
|   |   dia_mais_tarefas_entregues > 3: n (8.0)
|   curso = CIE
|   |   dia_mais_tarefas_entregues <= 3: y (81.0/10.0)
|   |   dia_mais_tarefas_entregues > 3: n (7.0/1.0)
|   curso = AMB
|   |   media_notas <= 7.46: y (78.0/15.0)
|   |   media_notas > 7.46
|   |   |   mes <= 2
|   |   |   |   media_semanal_cliques <= 68.17: y (3.0)
|   |   |   |   media_semanal_cliques > 68.17: n (2.0)
|   |   |   mes > 2: n (6.0)
|   curso = EDU
|   |   porcentagem_tarefas <= 81.48
|   |   |   dia_mais_acessado <= 0
|   |   |   |   dia_mais_acessado <= -1: y (4.0)
|   |   |   |   dia_mais_acessado > -1: n (5.0)
|   |   |   dia_mais_acessado > 0
|   |   |   |   mes <= 3
|   |   |   |   |   media_semanal_cliques <= 15.35
|   |   |   |   |   |   dia_mais_tarefas_entregues <= 1
|   |   |   |   |   |   |   media_semanal_cliques <= 2.93: y (6.0)
|   |   |   |   |   |   |   media_semanal_cliques > 2.93: n (2.0)
|   |   |   |   |   |   dia_mais_tarefas_entregues > 1: n (3.0)
|   |   |   |   |   media_semanal_cliques > 15.35: y (27.0/2.0)
|   |   |   |   mes > 3: y (8.0)
|   |   porcentagem_tarefas > 81.48: n (16.0/1.0)
|   curso = GPM
|   |   dia_mais_tarefas_entregues <= -1
|   |   |   media_notas <= 0.46: n (5.0)
|   |   |   media_notas > 0.46
|   |   |   |   dia_mais_acessado <= 2: n (4.0/1.0)
|   |   |   |   dia_mais_acessado > 2: y (4.0)
|   |   dia_mais_tarefas_entregues > -1
|   |   |   mes <= 1
|   |   |   |   dia_mais_tarefas_entregues <= 0: y (7.0)
|   |   |   |   dia_mais_tarefas_entregues > 0
|   |   |   |   |   media_notas <= 2.33: y (2.0)
|   |   |   |   |   media_notas > 2.33
|   |   |   |   |   |   media_semanal_cliques <= 87.81: n (8.0/1.0)
|   |   |   |   |   |   media_semanal_cliques > 87.81: y (8.0/1.0)
|   |   |   mes > 1: y (50.0/6.0)
|   curso = TCTE
```

```
|  |  media_notas <= 3.04
|  |  |  dia_mais_acessado <= 3: y (9.0/1.0)
|  |  |  dia_mais_acessado > 3: n (2.0)
|  |  media_notas > 3.04: n (16.0)
mes > 4
|  mes <= 9
|  |  media_notas <= 8.15
|  |  |  curso = ELPL
|  |  |  |  media_notas <= 5.66: n (3.0/1.0)
|  |  |  |  media_notas > 5.66: y (5.0)
|  |  |  curso = IIAE: n (25.0/5.0)
|  |  |  curso = CIE
|  |  |  |  media_notas <= 7.54
|  |  |  |  |  dia_mais_tarefas_entregues <= 3
|  |  |  |  |  |  media_notas <= 1.43: n (2.0)
|  |  |  |  |  |  media_notas > 1.43
|  |  |  |  |  |  |  media_notas <= 6.66: y (17.0)
|  |  |  |  |  |  |  media_notas > 6.66
|  |  |  |  |  |  |  |  media_notas <= 7.02: n (2.0)
|  |  |  |  |  |  |  |  media_notas > 7.02: y (2.0)
|  |  |  |  |  dia_mais_tarefas_entregues > 3: n (2.0)
|  |  |  |  media_notas > 7.54: n (5.0)
|  |  |  curso = AMB
|  |  |  |  dia_mais_tarefas_entregues <= -1: n (2.0)
|  |  |  |  dia_mais_tarefas_entregues > -1: y (38.0/11.0)
|  |  |  curso = EDU
|  |  |  |  mes <= 7
|  |  |  |  |  porcentagem_tarefas <= 78.85: y (14.0/1.0)
|  |  |  |  |  porcentagem_tarefas > 78.85: n (6.0)
|  |  |  |  mes > 7
|  |  |  |  |  media_notas <= 5.3: y (5.0/1.0)
|  |  |  |  |  media_notas > 5.3: n (16.0)
|  |  |  curso = GPM
|  |  |  |  porcentagem_tarefas <= 43.45: n (13.0)
|  |  |  |  porcentagem_tarefas > 43.45: y (10.0/2.0)
|  |  |  curso = TCTE: n (17.0)
|  |  media_notas > 8.15
|  |  |  media_semanal_cliques <= 0.17: n (97.0/9.0)
|  |  |  media_semanal_cliques > 0.17
|  |  |  |  curso = ELPL: y (11.0/2.0)
|  |  |  |  curso = IIAE: y (0.0)
|  |  |  |  curso = CIE: y (0.0)
|  |  |  |  curso = AMB: y (0.0)
|  |  |  |  curso = EDU: n (1.0)
|  |  |  |  curso = GPM: n (2.0)
|  |  |  |  curso = TCTE: n (1.0)
|  mes > 9
|  |  porcentagem_tarefas <= 2.15
|  |  |  mes <= 11: n (2.0)
|  |  |  mes > 11: y (10.0)
|  |  porcentagem_tarefas > 2.15: n (217.0/15.0)
```

Árvore Algoritmo J48 com seleção de atributos: “media_notas”, “media_semanal_cliques”, “porcentagem_tarefas”, “dia_mais_acessado” e “desistente”

```
media_semanal_cliques <= 0.14
|  dia_mais_acessado <= -1: y (47.0/9.0)
|  dia_mais_acessado > -1
|  |  media_notas <= 8.1
|  |  |  porcentagem_tarefas <= 58.23
|  |  |  |  media_semanal_cliques <= 0.08: n (85.0/7.0)
|  |  |  |  media_semanal_cliques > 0.08
|  |  |  |  |  dia_mais_acessado <= 1: n (2.0)
```

```
|  |  |  |  |  dia_mais_acessado > 1: y (4.0/1.0)
|  |  |  porcentagem_tarefas > 58.23
|  |  |  |  media_semanal_cliques <= 0.11
|  |  |  |  |  dia_mais_acessado <= 4
|  |  |  |  |  |  dia_mais_acessado <= 1: n (76.0/27.0)
|  |  |  |  |  |  dia_mais_acessado > 1: y (116.0/55.0)
|  |  |  |  |  dia_mais_acessado > 4
|  |  |  |  |  |  media_notas <= 5.19: y (3.0)
|  |  |  |  |  |  media_notas > 5.19: n (14.0)
|  |  |  |  media_semanal_cliques > 0.11: y (12.0/2.0)
|  |  media_notas > 8.1
|  |  |  dia_mais_acessado <= 1: n (97.0)
|  |  |  dia_mais_acessado > 1
|  |  |  |  media_semanal_cliques <= 0.1
|  |  |  |  |  media_semanal_cliques <= 0.07: n (76.0/9.0)
|  |  |  |  |  media_semanal_cliques > 0.07
|  |  |  |  |  |  media_notas <= 10.27: n (15.0/2.0)
|  |  |  |  |  |  media_notas > 10.27: y (10.0/2.0)
|  |  |  |  media_semanal_cliques > 0.1: n (12.0)
media_semanal_cliques > 0.14
|  media_notas <= 0.01: y (50.0)
|  media_notas > 0.01
|  |  porcentagem_tarefas <= 30
|  |  |  media_semanal_cliques <= 27.69
|  |  |  |  porcentagem_tarefas <= 0.69
|  |  |  |  |  media_notas <= 0.32: n (4.0/1.0)
|  |  |  |  |  media_notas > 0.32: y (5.0)
|  |  |  |  porcentagem_tarefas > 0.69: n (74.0/10.0)
|  |  |  media_semanal_cliques > 27.69: y (14.0/2.0)
|  |  porcentagem_tarefas > 30
|  |  |  dia_mais_acessado <= 1
|  |  |  |  dia_mais_acessado <= 0: y (78.0/33.0)
|  |  |  |  dia_mais_acessado > 0
|  |  |  |  |  media_notas <= 2.78: y (22.0/1.0)
|  |  |  |  |  media_notas > 2.78: n (42.0/18.0)
|  |  |  dia_mais_acessado > 1
|  |  |  |  media_semanal_cliques <= 31.46
|  |  |  |  |  dia_mais_acessado <= 4: y (101.0/40.0)
|  |  |  |  |  dia_mais_acessado > 4
|  |  |  |  |  |  media_semanal_cliques <= 6.69: n (11.0/1.0)
|  |  |  |  |  |  media_semanal_cliques > 6.69
|  |  |  |  |  |  |  media_semanal_cliques <= 23.48: y (3.0)
|  |  |  |  |  |  |  media_semanal_cliques > 23.48: n (3.0/1.0)
|  |  |  |  media_semanal_cliques > 31.46: y (222.0/53.0)
```

APÊNDICE B – Lista de decisão normal e com seleção de atributos do algoritmo de regra Part

Lista de decisão normal algoritmo Part

```
mes > 4 AND
mes > 9 AND
porcentagem_tarefas > 2.15 AND
curso = GPM: n (41.0)

mes > 4 AND
mes > 8 AND
porcentagem_tarefas > 0 AND
curso = EDU AND
mes > 9: n (37.0)

mes > 4 AND
media_notas > 8.15 AND
media_semanal_cliques <= 0.15 AND
dia_mais_acessado <= 1: n (74.0)

mes <= 4 AND
media_notas > 0.15 AND
curso = CIE AND
dia_mais_tarefas_entregues <= 3: y (81.0/10.0)

media_notas <= 0.01: y (52.0)

mes > 5 AND
curso = TCTE: n (38.0)

mes > 5 AND
dia_mais_tarefas_entregues > 1 AND
dia_mais_acessado <= 4 AND
curso = ELPL AND
media_semanal_cliques <= 23.48: n (17.0/1.0)

mes <= 3 AND
curso = AMB AND
mes > 1 AND
media_notas <= 6.73: y (30.0/1.0)

mes > 7 AND
media_notas > 0.21 AND
curso = EDU AND
media_notas > 5.3: n (16.0)

mes > 8 AND
media_notas > 0.21 AND
curso = CIE AND
porcentagem_tarefas > 70.73: n (17.0)

curso = ELPL AND
media_semanal_cliques > 0.05: y (98.0/23.0)

media_notas > 8.1 AND
mes > 3 AND
curso = AMB: n (22.0/1.0)

curso = IIAE: n (88.0/19.0)

media_notas > 7.71 AND
mes > 3 AND
curso = EDU AND
dia_mais_tarefas_entregues <= 4: n (23.0/1.0)
```

```
curso = TCTE AND
media_notas > 3.04: n (17.0)

curso = TCTE AND
dia_mais_acessado <= 3: y (9.0/1.0)

curso = EDU AND
porcentagem_tarefas <= 81.48: y (73.0/12.0)

curso = AMB AND
dia_mais_acessado <= 0 AND
mes <= 9: y (23.0/2.0)

curso = GPM AND
mes <= 4 AND
dia_mais_tarefas_entregues > -1 AND
mes > 1: y (50.0/6.0)

curso = AMB: y (91.0/40.0)

curso = EDU: n (14.0/1.0)

curso = GPM AND
mes <= 2 AND
dia_mais_acessado > 2 AND
dia_mais_tarefas_entregues <= 0: y (11.0)

dia_mais_acessado <= 4 AND
curso = GPM AND
dia_mais_acessado > -1 AND
dia_mais_tarefas_entregues > -1 AND
mes <= 8 AND
porcentagem_tarefas > 32.43 AND
media_notas <= 9.55 AND
dia_mais_acessado > 1: y (21.0/4.0)

curso = GPM AND
mes <= 11 AND
media_semanal_cliques <= 87.81: n (42.0/3.0)

curso = CIE AND
dia_mais_acessado <= 4 AND
dia_mais_tarefas_entregues <= 3 AND
media_notas <= 7.54 AND
media_notas > 1.55: y (21.0/1.0)

curso = CIE AND
dia_mais_tarefas_entregues > 0: n (24.0/1.0)

curso = GPM: y (7.0)
: n (11.0/2.0)
```

Lista de regras do algoritmo Part com seleção de atributos: “media_notas”, “media_semanal_cliques”, “porcentagem_tarefas”, “dia_mais_acessado” e “desistente”

```
media_semanal_cliques <= 0.14 AND
dia_mais_acessado <= -1 AND
porcentagem_tarefas <= 7.69 AND
porcentagem_tarefas <= 0.69: y (27.0/7.0)

media_semanal_cliques <= 0.14 AND
media_notas > 8.1 AND
dia_mais_acessado <= 1: n (97.0)
```

media_notas > 0 AND
media_semanal_cliques > 20.7 AND
dia_mais_acessado > 2: y (208.0/50.0)

media_notas <= 0: y (59.0)

media_semanal_cliques <= 0.11 AND
dia_mais_acessado > 4 AND
media_notas > 4.76 AND
porcentagem_tarefas <= 102.86: n (39.0)

porcentagem_tarefas <= 49.66 AND
media_notas > 3.94: n (41.0/1.0)

porcentagem_tarefas <= 30 AND
porcentagem_tarefas > 0.69 AND
dia_mais_acessado > -1 AND
media_notas <= 0.43: n (44.0/2.0)

media_notas > 7.71 AND
media_semanal_cliques <= 0.17 AND
porcentagem_tarefas > 80.56 AND
porcentagem_tarefas <= 94: n (65.0/2.0)

media_semanal_cliques > 0.11 AND
dia_mais_acessado <= 2 AND
media_notas <= 0.5 AND
dia_mais_acessado > 0: y (19.0)

APÊNDICE C – Parte do Arquivo ARFF

```
@Relation EAD_UTFPR
@ATTRIBUTE media_notas REAL
@ATTRIBUTE media_semanal_cliques REAL
@ATTRIBUTE porcentagem_tarefas REAL
@ATTRIBUTE dia_mais_acessado INTEGER
@ATTRIBUTE dia_mais_tarefas_entregues INTEGER
@ATTRIBUTE curso {ELPL, IIAE, CIE, AMB, EDU, GPM, TCTE}
@ATTRIBUTE mes INTEGER
@ATTRIBUTE desistente {y, n}

@Data
3.00,43.34,100.00,2,5,AMB,1,n
3.00,34.76,100.00,2,0,AMB,1,n
3.00,151.24,100.00,3,4,AMB,1,n
3.00,125.28,100.00,1,0,AMB,1,n
3.00,57.79,100.00,1,5,AMB,1,n
3.00,103.39,100.00,3,0,AMB,1,n
3.00,33.18,100.00,3,0,AMB,1,n
3.00,60.72,100.00,5,2,AMB,1,n
3.00,42.21,100.00,2,4,AMB,1,n
1.67,8.58,66.67,2,0,AMB,1,n
1.67,26.86,100.00,0,4,AMB,1,n
3.00,44.47,100.00,0,0,AMB,1,n
1.33,33.41,66.67,1,0,AMB,1,n
3.00,69.75,100.00,1,0,AMB,1,n
3.00,53.50,100.00,0,1,AMB,1,n
3.00,88.26,100.00,4,0,AMB,1,n
2.33,46.50,100.00,0,2,AMB,1,n
3.00,83.75,100.00,5,4,AMB,1,n
3.00,72.69,100.00,0,3,AMB,1,n
0.33,0.00,66.67,-1,-1,AMB,1,y
3.00,32.73,100.00,2,3,AMB,1,n
2.33,2.48,66.67,2,-1,AMB,1,n
2.33,49.89,100.00,0,0,AMB,1,n
3.00,62.53,100.00,0,4,AMB,1,n
2.67,146.28,100.00,1,1,AMB,1,y
1.67,32.51,100.00,5,5,AMB,1,n
3.00,76.52,100.00,0,1,AMB,1,n
3.00,77.20,100.00,0,2,AMB,1,n
3.00,77.88,100.00,2,1,AMB,1,n
1.67,10.84,100.00,5,0,AMB,1,y
0.33,3.84,66.67,3,-1,AMB,1,n
3.00,71.78,100.00,0,4,AMB,1,n
3.00,107.00,100.00,2,1,AMB,1,n
1.67,11.51,100.00,1,0,AMB,1,n
3.00,0.23,100.00,2,5,AMB,1,n
3.00,41.31,100.00,2,3,AMB,1,y
0.33,0.00,66.67,-1,-1,AMB,1,y
0.33,2.93,66.67,2,-1,AMB,1,n
1.67,124.60,100.00,2,3,AMB,1,n
3.00,39.73,66.67,0,4,AMB,1,n
3.00,53.50,66.67,3,4,AMB,1,y
3.00,56.21,100.00,4,3,AMB,1,n
3.00,66.14,100.00,5,0,AMB,1,n
3.00,51.24,100.00,3,0,AMB,1,n
3.00,81.72,100.00,1,4,AMB,1,n
3.00,57.79,100.00,2,3,AMB,1,n
3.00,37.92,100.00,2,5,AMB,1,n
3.00,33.63,100.00,1,3,AMB,1,n
3.00,81.72,100.00,3,1,AMB,1,n
3.00,56.21,100.00,5,0,AMB,1,n
0.33,14.45,66.67,1,-1,AMB,1,y
3.00,85.78,100.00,2,2,AMB,1,n
3.00,62.08,100.00,1,2,AMB,1,n
3.00,34.99,100.00,0,4,AMB,1,y
3.00,52.14,100.00,5,2,AMB,1,n
3.00,32.73,100.00,1,2,AMB,1,n
0.33,6.32,0.00,0,3,AMB,1,n
```

0.33,3.16,66.67,3,-1,AMB,1,n
3.00,56.43,100.00,4,5,AMB,1,n