UNISINOS

Programa Interdisciplinar de Pós-Graduação em

Computação Aplicada

Mestrado Acadêmico

João Luiz Cavalcante Ferreira

MD-PREAD: UM MODELO PARA PREDIÇÃO DE REPROVAÇÃO DE APRENDIZES NA EDUCAÇÃO A DISTÂNCIA USANDO ÁRVORE DE DECISÃO

São Leopoldo

2016

UNIVERSIDADE DO VALE DO RIO DOS SINOS – UNISINOS
UNIDADE ACADÊMICA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA
NÍVEL MESTRADO

JOÃO LUIZ CAVALCANTE FERREIRA

MD-PREAD: UM MODELO PARA PREDIÇÃO DE REPROVAÇÃO DE APRENDIZES NA EDUCAÇÃO A DISTÂNCIA USANDO ÁRVORE DE DECISÃO

São Leopoldo
2016

João Luiz Cavalcante Ferreira

MD-PREAD: UM MODELO PARA PREDIÇÃO DE REPROVAÇÃO DE APRENDIZES NA EDUCAÇÃO A DISTÂNCIA USANDO ÁRVORE DE DECISÃO

Dissertação de Mestrado apresentada como requisito parcial para a obtenção do título de Mestre, pelo Programa Interdisciplinar de Pós-Graduação em Computação Aplicada da Universidade do Vale do Rio dos Sinos – UNISINOS

Orientador: Prof. Dr. Jorge Luis Victoria Barbosa

Coorientador: Prof. Dr. Sandro José Rigo

São Leopoldo

2016

Ficha Catalográfica

F368m

Ferreira, João Luiz Cavalcante
MD-PREAD: um modelo para predição de reprovação de aprendizes na educação a distância usando árvore de decisão / João Luiz Cavalcante Ferreira. – São Leopoldo, 2016.
74 f.: il. color.

Dissertação (Mestrado em Computação Aplicada) – Universidade do Vale do Rio dos Sinos, 2016.
Orientador: Prof. Dr. Jorge Luis Victoria Barbosa.
Coorientador: Prof. Dr. Sandro José Rigo.

1. Sistema de Recomendação Educacional 2. Ferramenta computacional 3. Educação a Distância 4. Predição I. Barbosa, Jorge Luis Victoria II. Rigo, Sandro José III. Universidade do Vale do Rio dos Sinos IV. Título.

CDD 005.12

(Elaborada pelo Bibliotecário/Documentalista Odimar José Ferreira Porto – CRB-AM 496/11)

João Luiz Cavalcante Ferreira

"MD-PREAD: UM MODELO PARA PREDIÇÃO DE REPROVAÇÃO DE APRENDIZES NA EDUCAÇÃO A DISTÂNCIA USANDO ÁRVORE DE DECISÃO"

Dissertação apresentada à Universidade do Vale do Rio dos Sinos – Unisinos, como requisito parcial para obtenção do título de Mestre em Computação Aplicada.

Aprovado em 25 de fevereiro de 2016

BANCA EXAMINADORA

Prof. Dr. Jorge Luis Victória Barbosa - UNISINOS

Prof. Dr. Sandro José Rigo – UNISINOS

Prof. Dr. João Carlos Gluz - UNISINOS

Prof.ª Dr.ª Patrícia Brandalise Scherer Bassani - FEEVALE

Prof. Dr. Jorge Luis Victória Barbosa (Orientador)

Prof. Dr. Sandro José Rigo(Cóorientador)

Visto e permitida a impressão

São Leopoldo,

Prof. Dr. Sandro José Rigo

Coordenador PPG em Computação Aplicada

AGRADECIMENTOS

Agradeço primeiramente aos meus pais, Francisco André (falecido) e Euclair, que estiveram sempre ao meu lado torcendo para que tudo corresse bem.

Agradeço especialmente a minha esposa e filho, Rosineide e Jonathas por me apoiarem durante todo o tempo não me deixando fraquejar nos momentos mais difíceis.

Ao meu orientador e Coorientador Prof. Dr. Jorge Luis Victória Barbosa e Prof. Dr. Sandro José Rigo, por terem nos motivado e prestado todas as orientações necessárias à conclusão deste trabalho. Avante!

Ao meu colega de caminhada André Aloise que não mediu esforços para me motivar e caminhar ao meu lado ao longo desses 24 meses.

A professora Msc. Ana Maria Alves Pereira que igualmente me motivou incessantemente a permanecer firme olhando sempre para o horizonte.

Ao me colega de trabalho Carlos Tiago Garantizado, que com suas sugestões ajudou no processo de construção do Modelo proposto.

Aos meus amigos, familiares, colegas do IFRR e Campus Coari que contribuíram de alguma forma para que fosse possível lograr êxito neste mestrado.

A Deus, que me deu saúde, perseverança, sabedoria e consolo nos momentos mais difíceis.

RESUMO

A Educação a Distância (EaD) no Brasil tem se consolidado com diversos estudantes optando por essa modalidade de ensino para ampliar suas formações e realização profissional, no entanto ela enfrenta alguns obstáculos, como a resistência de educandos e educadores, desafios organizacionais, custos de produção e a questão da reprovação ou retenção de alunos. Um dos principais diferenciais dos cursos EaD é a grande quantidade de dados gerados pelas interações no ambiente educacional, o que abre novas possibilidades para estudar e compreender estas interações. A Mineração de Dados educacionais (MDE) é uma área de pesquisa interdisciplinar que lida com o desenvolvimento de métodos para explorar dados originados no contexto educacional. A *Learning Analytics* (LA) é outra área de pesquisa emergente. Ela busca medir, coletar, analisar e relatar dados sobre estudantes. O desafio dos pesquisadores é desenvolver métodos capazes de prever o desempenho dos estudantes de modo a possibilitar a intervenção de professores e tutores visando resgatar o estudante antes que reprove. Esta dissertação propõe o MD-PREAD, um modelo para predição de grupos de risco de reprovação em um ambiente de Educação a Distância. A técnica de árvore de decisão foi utilizada para possibilitar um diferencial quanto à possibilidade de interpretação dos dados gerados pelo uso dos métodos de predição, pois outros métodos, tais como Redes Neurais Artificiais possuem como deficiência justamente a dificuldade de identificar as causas que levam aos resultados das predições. O modelo foi prototipado na ferramenta de mineração *RapidMiner*. Um experimento foi realizado no Instituto Federal de Educação, Ciência e Tecnologia do Amazonas, no programa Universidade Aberta do Brasil, no Curso de Filosofia da educação. Foram feitas coletas de dados históricos de 10 disciplinas de um grupo de 30 aprendizes em dois semestres consecutivos, 2014/2 e 2015/1, o total de alunos matriculados foi de 125, o total de interações levantadas foi de 41070, o cálculo de predição considerou as médias das avaliações de 30 aprendizes, os desvios padrões das interações e suas respectivas situações. Estes dados serviram para compor o conjunto de treinamento necessário para a definição da regra de classificação que teve como predominante a acurácia de 55% e a confiabilidade *Kappa* de 0,22. Foi realizado um segundo processo de validação, após o experimento, considerou-se os 125 alunos e o melhor classificador encontrado foi o J48 com a acurácia de 84,05%, precisão de 77,08% e *recall* de 50,23%. Concluiu-se que o MD-PREAD é uma ferramenta de auxílio no prognóstico de grupos de risco de reprovação, uma vez que possibilitou a geração e disponibilização semanal destes grupos a um sistema de recomendação educacional externo.

Palavras-Chave: EaD; predição; árvore de decisão; *learning analytics.*

ABSTRACT

E-learning in Brazil has been established with many students opting for this type of education to expand their training and professional achievement, however it faces some obstacles, such as resistance from students and educators, organizational challenges, production costs and the question of failure or retention of students. One of the main advantages of e-learning courses is the large amount of data generated by the interactions in the educational environment, which opens up new possibilities to study and understand these interactions. Educational Data Mining (EDM) is an area of interdisciplinary research that deals with the development of methods to explore data that originates in the educational context. Learning Analytics (LA) is another area of emerging research. It seeks to measure, collect, analyze and report data on students. The challenge for researchers is to develop methods to predict the performance of students in order to allow the intervention of teachers and tutors aiming to retrieve the student before failing. This thesis proposes the MD-PREAD, a model for predicting failure of risk groups in a e-learning environment. The decision tree technique was used to enable a difference as to whether the interpretation of the data generated by the use of prediction methods, since other methods such as Artificial Neural Networks that has as disability difficulty in identifying precisely the causes that lead to predictions results. The model was prototyped in RapidMiner mining tool. An experiment was conducted at the Federal Institute of Education, Science and Technology of Amazonas, the Open University of Brazil program in course Philosophy of education. Historical data collection of 10 disciplines from a group of 30 apprentices were made in two consecutive semesters, 2014/2 and 2015/1, the total number of enrolled students was 125, the total raised interactions were 41070, the prediction calculation considered average of 30 apprentices ratings, the standard deviations of the interactions and their situations. These data served to compose the training set required for classification rule defining which had as predominant accuracy of 55% and Kappa reliability 0.22. A second validation process was carried out after the experiment. It was considered the total amount of 125 apprentices and the best classifier found was the J48 with the accuracy of 84.05%, 77.08% of classification precision and recall of 50.23%. It was concluded that the MD-PREAD is a support tool in the prognosis of failure risk groups, since it enabled the generation and weekly availability of these groups to a recommendation system.

Keywords: *E-learnig; prediction; decision tree; learning analytics.*

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
PIPCA	Programa Interdisciplinar de Pós-Graduação em Computação Aplicada
EaD	Educação a Distância
EDM	*Educational Data Mining*
LA	*Learning Analytics*
LAK	*Learning Analytics and Knowledge*
AVA	Ambientes Virtuais de Aprendizagem
TIC	Tecnologias da Informação e da Comunicação
IFAM	Instituto Federal de Educação, Ciência e Tecnologia do Amazonas.
MOODLE	*Modular Object-Oriented Dynamic Learning Environment*
IDH	Índice de Desenvolvimento Humano
IDEB	Índice de Desenvolvimento da Educação Básica
UAB	Universidade Aberta do Brasil
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
e-Tec	Escola Técnica Aberta do Brasil
Secadi	Secretaria de Educação Continuada, Alfabetização, Diversidade e Inclusão.
STIs	Sistemas Tutores Inteligentes
JEDM	*Journal of Educational Data Mining.*
AIEd	*Artificial Intelligence in Education*
KDD	*Knowledge-Discovery in Databases*
SGA	Sistema de Gestão Acadêmica

SUMÁRIO

1 INTRODUÇÃO

A Educação a Distância (EaD) baseia-se nos princípios da igualdade e do ensino permanente, acessível a qualquer pessoa, independentemente do seu perfil, a qualquer hora e em qualquer lugar. Segundo (DETONI; ARAUJO; CECHINEL, 2014) a EaD no Brasil tem se consolidado, com diversos estudantes optando por essa modalidade de ensino para ampliar suas formações e realização profissional.

A demanda por essa modalidade de ensino pode ser acompanhada por pesquisas realizadas periodicamente pela Associação Brasileira de Educação a Distância (ABED). Segundo dados do Censo EaD ("CensoEaD", 2014), o número amostral de matrículas foi de 290.497, esta forma de apresentação dificultou a comparação com anos anteriores, pois era representada com o número total da população. Não observou-se a preocupação com alunos reprovados ou retidos, apenas os evadidos. A evasão é apontada por grande parte das instituições como um dos maiores obstáculos enfrentados nos diferentes tipos de cursos EAD. Em todos os tipos de curso, nenhuma instituição apontou taxas de evasão superiores a 75% e, na maioria dos casos, a evasão identificada se concentra na faixa de até 25%. A falta de tempo para estudar ou participar do curso é apontada pela maioria das instituições como principal motivo para evasão nas diferentes modalidades de EAD pesquisadas.

Um importante desafio de pesquisa é desenvolver métodos capazes de prever o comportamento dos estudantes, de modo a possibilitar a intervenção de professores/tutores, ou demais envolvidos, visando resgatar o estudante antes que ele reprove (MACFADYEN, L.P., DAWSON, 2010).

.

1.1 Motivação

Segundo (FARIA; SALVADORI, 2010) desde 1840 a educação a distância, vem sendo utilizada, passando por várias gerações de avanços tecnológicos. Ela iniciou com a educação por correspondência, e em seguida programas de televisão e rádio passaram a dar suporte a esta modalidade. Na etapa seguinte foi possível combinar as funções de texto, áudio e televisão. Com a chegada das redes de computadores foi possível estabelecer uma melhor comunicação com os usuários através de e-mails e ambientes virtuais de aprendizagem. Nos dias atuais o avanço continua com a agregação do uso de dispositivos móveis.

O ambiente virtual de aprendizagem é rico em dados que não analisados em tempo hábil para que intervenções sejam feitas durante a realização do curso. Para (RIGO et al., 2014) a utilização dos Ambientes Virtuais de Aprendizagem na prática de Educação a Distância proporciona uma grande quantidade de dados que são armazenadas para fins de controle, no entanto a análise deles é feita com atrasos substanciais retardando ações e perdendo oportunidades de intervenções.

No Ambiente Virtual de Aprendizagem (AVA) o aprendiz tem sua curva de aprendizado personalizada, pois poderá avançar ou retroceder nas atividades didáticas, de modo a assimilar o conteúdo no seu ritmo. No entanto há a necessidade de se aferir este aprendizado e precisa-se pensar em modelos de predições que auxiliem neste processo informando aos gestores quais os alunos pertencem ao grupo de risco de reprovação bem como suas lacunas de aprendizado a fim de que sejam sanadas.

A demanda por esta modalidade cresce continuamente, por se tratar de um ambiente onde a relação professor/aluno se estabelece através de recursos disponíveis, a permanência

do aprendiz neste ambiente dependerá da criatividade dos professores e tutores em tornar o ambiente o mais atrativo possível, do contrário o aprendiz tenderá a desistir ou reprovar nas disciplinas.

Para melhorar a gestão do ensino a distância, é necessária à existência de mecanismos que monitorem as interações dos alunos com o ambiente e que alertem quanto a possíveis desvios durante o processo, a fim de serem corrigidos tempestivamente e não apenas ao final do curso quando os benefícios somente serão levados a turmas futuras. A grande quantidade de dados gerada pelas interações no ambiente educacional abre novas possibilidades para estudar e compreender estas interações (HÄMÄLÄINEN, W., VINNI, 2010)(T.; D.; GUPTA, 2014).

Algumas áreas de pesquisas surgiram nos últimos anos com intuito de auxiliar o ensino a distância. A *Educational Data Mining (EDM)* é uma área de pesquisa interdisciplinar que lida com o desenvolvimento de métodos para explorar dados originados no contexto educacional (ROMERO; VENTURA, 2010). Juntamente com a EDM existe a *Learning Analytics* (LA) (DRACHSLER *et al.*, 2012), outra área de pesquisa emergente.

Neste âmbito o presente trabalho propõe um modelo de predição de reprovação de aprendizes em uma disciplina que funcione em um Ambiente Virtual de Aprendizagem, considerando as interações dos alunos e suas médias de avaliações parciais. A ferramenta de mineração utilizada foi o *RapidMiner* (COSTA, JFA, 2011). Trata-se de uma descoberta de conhecimento que se utiliza da taxonomia de mineração de dados (MAIMON, O. & LIOR, 2010), bem como de recursos de previsão, mais especificamente os classificadores e dentre eles a árvore de decisão.

Após a fase de preparação dos dados, seguindo o ciclo de vida do CRISP-DM(MAIMON, O. & LIOR, 2010), a ferramenta *RapidMiner* foi configurada através de seus operadores, para fazer a leitura do conjunto de dados de treino e do conjunto de dados de teste. O operador *Decision Tree* fez o processamento de treino do algoritmo com o critério *Information_Gain* e encontrou a regra a ser aplicada, o operador *Apply Model* aplicou a regra no conjunto de teste e classificou os grupos de predição de aprovação e reprovação, o operador *Filter Examples* selecionou o grupo de predição de reprovação e o operador *Write CSV* gerou o arquivo resultante do processo em formato csv. Os dados obtidos foram disponibilizados para os gestores educacionais e sistemas de recomendação, a fim de servirem de apoio à decisão para os encaminhamentos que precisassem ser feitos. O processo foi repetido semanalmente para fins de acompanhamento dos resultados.

A *Learning Analitycs* concentra várias áreas de pesquisa. No desenvolvimento deste trabalho o foco foi a utilização Mineração de Dados Educacionais, para auxiliar na construção de um modelo de predição de reprovação em ambiente virtual de aprendizagem utilizando classificadores de árvore de decisão.

1.2 Problema e Questão de Pesquisa

Segundo (DETONI; ARAUJO; CECHINEL, 2014) a EaD enfrenta obstáculos a serem ultrapassados como, a resistência de educandos e educadores, desafios organizacionais e custos de produção. Outros empecilhos são a dificuldade do aluno em absorver os conteúdos, a falta de estímulo e motivação para o estudo, de acordo com (MARI et al., 2011).

Diante deste problema torna-se importante prever o grupo de risco de reprovação. A vantagem em utilizar este método é que de posse dessa informação, há a possibilidade de se

fazer recomendações no intuito de erradicar ou minimizar o índice de reprovação em disciplinas regulares cursadas em ambientes virtuais de aprendizagem.

Este trabalho propõe o MD-PREAD, um modelo de predição de reprovação de aprendizes em um ambiente de Educação a Distância utilizando árvore de decisão. Esta técnica foi escolhida por possibilitar um diferencial quanto à possibilidade de interpretação dos dados gerados pelo uso dos métodos de predição, pois outros métodos usados, tais como Redes Neurais Artificiais possuem como deficiência justamente a dificuldade de identificar as causas que levam aos resultados das predições.

A questão de pesquisa deste trabalho é: Quais os benefícios de um modelo de predição de reprovação de aprendizes em EaD baseado em um classificador de árvore de decisão?

1.3 Objetivos

O principal objetivo deste trabalho foi especificar, implementar e avaliar o modelo de predição de reprovação de aprendiz denominado MD-PREAD. As principais características do MD-PREAD são: (i) obter registros históricos de alunos, relacionados às suas interações e avaliações, para servir de base de treinamento, (ii) obter registros correntes de alunos relacionados às suas interações e avaliações para servir de base de teste, (iii) utilizar árvore de decisão para geração de regra de classificação, (iv) aplicar a regra na base de teste para extração dos grupos de predição de aprovação e reprovação, (v) selecionar o grupo de risco de reprovação para serem disponibilizados aos gestores educacionais e ou sistemas de recomendação.

Os objetivos específicos são os que seguem:

- definir os critérios para a classificação;
- estabelecer regras para a classificação;
- modelar o MD-PREAD;
- delinear um cenário de aplicação do MD_PREAD;
- desenvolver um protótipo;
- avaliar o MD-PREAD.

1.4 Metodologia

A viabilidade dos objetivos propostos por este trabalho se deram por meio de uma metodologia que teve como primeira etapa um estudo geral sobre a Educação a distância, *Learning Analytics* e Árvore de decisão.

Na segunda etapa foram realizados estudos de modelos já existentes para predição de situações de reprovação utilizando os mesmos conceitos, com a finalidade de fazer uma comparação e identificar a contribuição do MD-PREAD.

Na terceira etapa foi modelado o MD-PREAD (Sistema de Predição de Reprovação usando Árvore de Decisão) indicando sua arquitetura e discutindo seus principais artefatos.

Na quarta etapa foi definido um cenário da aplicação utilizando os recursos do MD-PREAD. O objetivo desta etapa foi avaliar o modelo e averiguar sua usabilidade.

Na quinta etapa constou a aplicação do protótipo com todos os recursos necessários para a execução do cenário.

Na sexta etapa foram avaliados os resultados do protótipo que proporcionaram os resultados deste trabalho.

1.5 Organização do Texto

Este trabalho está estruturado da seguinte forma.

O capítulo 2 descreve conceitos básicos sobre a Educação a Distância, *Learning Analytics* e Árvore de decisão, no contexto de Mineração de Dados Educacionais.

O capítulo 3 apresenta os trabalhos relacionados ao escopo deste trabalho e uma comparação entre eles.

No capítulo 4 é apresentado o MD-PREAD o modelo para predição que utiliza árvore de decisão, uma visão geral e sua arquitetura.

No capítulo 5 são explanados aspectos de implementação do protótipo e sua aplicação no experimento.

No capítulo 6 são abordados os aspectos de avaliação do MD-PREAD.

No capítulo 7 são feitas as considerações finais, tratando de conclusões, contribuições e trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentados os temas considerados relevantes para a pesquisa, os quais são os seguintes:

- Educação a Distância
- *Learning Analytics*
- Mineração de dados Educacionais

Entender o contexto histórico da Educação a Distância é importante, já que sem a valorização do passado não é possível conseguir obter parâmetros para mensurar o quanto se melhorou. A análise do aprendizado e a mineração de dados educacionais também são fundamentais, uma vez que o perfil procurado será a busca pela informação escondida no banco de dados, que poderá mudar a situação do aprendiz para melhor.

2.1 Educação a Distância

Conforme (“Ministério da Educação”, 2015) educação a distância é a modalidade educacional na qual alunos e professores estão separados, física ou temporalmente e, por isso, faz-se necessária a utilização de meios e tecnologias de informação e comunicação. Essa modalidade é regulada por uma legislação específica e pode ser implantada na educação básica (educação de jovens e adultos, educação profissional técnica de nível médio) e na educação superior.

Segundo (ARAGÃO, 2004) a EaD não é uma modalidade recente de educação, ela é utilizada há muitos anos no Brasil com o apoio de tecnologias como o rádio e o material impresso. Ela passa por reformulações devido à emergência e utilização cada vez mais constante das Tecnologias da Informação e da Comunicação (TIC). De acordo com (RUMBLE, 2000), o desenvolvimento tecnológico da educação a distância passou por gerações. Os sistemas de primeira geração eram baseados em textos impressos ou escritos à mão. O termo educação por correspondência descreve este sistema adequadamente. Eram típicas desse período as escolas e faculdades por correspondência comercial, desenvolvidas a partir de 1840.

Os sistemas de segunda geração eram baseados em televisão e áudio. Eles contavam com a televisão e o rádio para captar leituras ao vivo na sala de aula e as transmitir a outros grupos de alunos, que poderiam seguir a lição de uma sala de aula distante por meio da televisão ou do rádio.

Os sistemas de terceira geração trouxeram os sistemas de primeira e segunda fase juntos, em uma abordagem multimídia, com base em textos, áudio e televisão. Entretanto, havia diferenças: a transmissão tendia a ser usada como um meio suplementar de apoio ao material impresso, não sendo meio predominante como no caso dos sistemas de segunda geração.

Os sistemas de quarta geração foram desenvolvidos em torno de comunicações mediadas por computador, como conferência por computador e correio eletrônico, associados ao acesso a bancos de dados, bancos de informações e bibliotecas eletrônicas, com a utilização de instruções orientadas por computador. O avanço da informática e da tecnologia de redes de computadores, principalmente da Internet, deu nova dimensão à EaD, tendo em

vista tornar possível formar mais pessoas, independentemente da localização geográfica ou temporal dos sujeitos potencialmente partícipes dos processos de ensino e aprendizagem.

No Brasil estão sendo realizadas pesquisas no sentido de viabilizar os processos de aprendizagem à distância através de dispositivos *wireless*. Pode-se denominar esse momento como o início da quinta geração da EaD, viabilizada por meio de estratégias de comunicação portáteis e sem fio, os chamados *Mobile Learning* ou *mlearning* (SACCOL, AMAROLINDA I. C. Z. ; BARBOSA, JORGE L. V. ; SCHLEMMER, ELIANE ; REINHARD, 2011). A EaD vem sendo desenvolvida para atender a demandas diversificas da agenda no novo cotidiano, a possibilidade de se ter acesso a informações de cunho educacional.

A queda das barreiras de espaço e tempo são o principal desafio e trunfo para a expansão da EaD, entendida como um processo educativo que envolve diferentes meios de comunicação capazes de ultrapassar esses limites e permitir a interação dos sujeitos com as diversas fontes de informação.

Assim, alteram-se papéis tradicionalmente cristalizados: o aluno deixa de ser um receptor passivo e se torna responsável por sua aprendizagem, com direito a trabalhar num ritmo individualizado, sem perder, no entanto, a possibilidade de interagir com seus pares e com o professor, esse deixa de ser o dono do saber e o controlador da aprendizagem, para ser um orientador que estimula a curiosidade, o debate e a interação com os participantes do processo.

2.1.1 Universidade Aberta do Brasil

A Universidade Aberta do Brasil (UAB) é um sistema integrado por universidades públicas que oferece cursos de nível superior para camadas da população que têm dificuldade de acesso à formação universitária, por meio do uso da metodologia da educação a distância. O público em geral é atendido, mas os professores que atuam na educação básica têm prioridade de formação, seguidos dos dirigentes, gestores e trabalhadores em educação básica dos Estados, Municípios e do Distrito Federal.

O Sistema UAB foi instituído pelo Decreto nº 5.800, de 8 de junho de 2006, para o desenvolvimento da modalidade de educação a distância, com a finalidade de expandir e interiorizar a oferta de cursos e programas de educação superior no País. Ele fomenta a modalidade de educação a distância nas instituições públicas de ensino superior, bem como apoia pesquisas em metodologias inovadoras de ensino superior respaldadas em tecnologias de informação e comunicação. Além disso, incentiva a colaboração entre a União e os entes federativos e estimula a criação de centros de formação permanentes por meio dos polos de apoio presencial em localidades estratégicas.

Deste modo, o Sistema UAB propicia a articulação, a interação e a efetivação de iniciativas que estimulam a parceria dos três níveis governamentais (Federal, Estadual e Municipal) com as universidades públicas e demais organizações interessadas, enquanto viabiliza mecanismos alternativos para o fomento, a implantação e a execução de cursos de graduação e pós-graduação de forma consorciada. Ao dar início a esta modalidade de ensino com a universidade pública de qualidade em locais distantes e isolados, incentiva o desenvolvimento de municípios com baixos IDH e IDEB.

Neste contexto, funciona como um eficaz instrumento para a universalização do acesso ao ensino superior e para a requalificação do professor em outras disciplinas,

fortalecendo a escola no interior do Brasil, minimizando a concentração de oferta de cursos de graduação nos grandes centros urbanos e evitando o fluxo migratório para as grandes cidades.

2.2 Learning Analytics

Segundo (RIGO et al., 2014) mesmo possuindo uma grande quantidade de dados sobre os estudantes, as instituições de ensino superior têm sido tradicionalmente ineficientes no uso desses dados, muitas vezes realizando análises com atrasos substanciais, retardando ações e perdendo oportunidades de intervenções.

Este contexto torna-se mais evidente ao ser constatado que a utilização de recursos de mediação digital e a melhoria dos sistemas de suporte, tais como os sistemas acadêmicos, repositórios digitais e ambientes virtuais de aprendizagem de fato fazem parte do cotidiano dos cursos de graduação em nível universitário, além de ser observado em menor escala em outros níveis de ensino.

De acordo com (RIGO et al., 2014) LA não é uma nova área de pesquisa, mas ela pode ser considerada uma síntese de técnicas existentes em diversas áreas de pesquisa convergentes com o uso da tecnologia para melhoria do processo de ensino e aprendizagem. A relação entre LA e as áreas de pesquisa correlatas foram referenciadas no trabalho de Chatti (M. A. CHATTI., AL. L. DYCKOFF, U.SCHROEDER; THÜS., 2012) e incluem *Learning Analytics*, *action research*, Mineração de Dados Educacionais, sistemas de recomendação e aprendizagem personalizada ou adaptativa.

Desta forma, as iniciativas de LA podem envolver combinações de recursos disponibilizados a partir de diversas outras áreas, tais como o Aprendizado de Máquina, a Inteligência Artificial, o resgate de informações, recursos de Estatística ou de Visualização de Dados, entre outros. O seu crescente interesse pode ser observado em diversos projetos e eventos destacados a partir da criação da *Society for Learning Analytics Research*[1].

Os autores (M. A. CHATTI., AL. L. DYCKOFF, U.SCHROEDER; THÜS., 2012) propuseram um modelo de referência para LA baseado em quatro dimensões, com o objetivo de identificar o conjunto completo de elementos necessários para construir sistemas a partir da abordagem de análise em LA. Estas quatro dimensões seriam o tipo de dados coletados, o público-alvo da análise, o objetivo da análise dos dados e, por fim, a técnica utilizada. Já no trabalho de Greller e Drachsler (W. GRELLER, 2012) são analisadas questões que devem ser consideradas para minimizar desafios em potencial e permitir uma exploração benéfica de dados educacionais.

Segundo (R. FERGUSON., 2012), o desenvolvimento e adoção dos recursos de LA, e também dos recursos de MDE, passam por alguns desafios importantes descritos em quatro grandes grupos, que seriam: a ampliação da conexão com ciências do aprendizado, tais como a cognição, a meta-cognição e a pedagogia, a necessidade de desenvolver métodos para tratamento de maiores conjuntos de dados e dados mais diversos, tais como dados de computação móvel, dados biométricos e dados descrevendo emoção ou humor.

[1] http://solaresearch.org/

2.3 Mineração de dados Educacionais

Segundo (COSTA, EVANDRO *et al.*, 2012) a área emergente de Mineração de Dados Educacionais procura desenvolver ou adaptar métodos e algoritmos de mineração existentes, de tal modo que se preste a compreender melhor os dados em contextos educacionais. Estes dados são produzidos principalmente por estudantes e professores, extraídos dos ambientes em que interagem, tais como os ambientes virtuais de aprendizagem AVAs.

Assim, há a necessidade, por exemplo, de adequação dos algoritmos de mineração de dados existentes para lidar com especificidades inerentes aos dados educacionais, tais como a não independência estatística e a hierarquia dos dados.

Por outro lado, há uma necessidade significativa e urgente no provimento de ambientes computacionais apropriados para mineração de dados educacionais, oferecendo facilidades de uso para cada um dos atores envolvidos, notadamente ao professor.

2.3.1 Origens da Mineração de Dados Educacionais

Segundo (COSTA, EVANDRO *et al.*, 2012) apesar de existirem algumas iniciativas com *workshops* específicos dentro das conferências sobre *Artificial Intelligence in Education (AIEd),* foi somente em 2005, em Pittsburgh, EUA, que foi organizado o primeiro *Workshop on Educational Data Mining*, como parte do *20th National Conference on Artificial Intelligence (AAAI 2005).* Daí em diante, houve mais algumas realizações deste workshop entre 2006 e 2007. Seguindo-se, em 2008 lança-se, em Montreal, Canadá, a primeira conferência em *EDM: First International Conference on Educational Data Mining*, evento este que se estabeleceu e ganhou regularidade de realização anual, esteve em 2012 na sua quinta edição.

Em 2009, esta sociedade investiu na criação de um periódico e publicou o seu primeiro volume do *JEDM - Journal of Educational Data Mining*. Em 2011 constituiu-se a sociedade científica para *EDM (International Educational Data Mining Society*[2]*)*. Enfim, a área de EDM está bem consolidada internacionalmente, mas, ainda dando os seus primeiros passos no Brasil, ficando a produção de trabalhos por conta de algumas poucas iniciativas de pesquisas isoladas.

2.3.2 Tarefas para Mineração de Dados Educacionais

Há diversas tarefas envolvidas em EDM, notadamente as que decorrem diretamente da análise de dados gerados nas interações dos estudantes com os ambientes de aprendizagem. Dessa análise surgem demandas para responder questões relacionadas a como melhorar a aprendizagem do estudante, como desenvolver ambientes educacionais mais eficazes que contribuam efetivamente para os estudantes aprenderem mais e em menos tempo.

Em outra perspectiva, pretende-se saber qual o perfil dos cursistas que vem reprovando nos últimos períodos, quais os caminhos que percorreram, as atividades que realizaram ou não, a fim de determinar este perfil, através de um sistema de predição. Este perfil pode então ser comparado a comportamentos análogos e nos proporcionar um grupo de risco, e esta base de conhecimento propiciar insumos para um sistema de recomendação a fim de mitigar o índice de reprovação.

[2] Ver detalhes em http://www.educationaldatamining.org/

Do ponto de vista computacional, alguns desafios práticos que se apresentam em vários contextos educacionais estão relacionados, por exemplo, a falta de padronização dos dados, o que acaba exigindo grande esforço de pré-processamento (BAKER, R.S.J.D., 2011). Além disso, há a necessidade de adequação dos algoritmos clássicos de mineração de dados para lidar com especificidades inerentes aos dados educacionais, tais como a não independência estatística e a hierarquia dos dados (MENZIES, 2003).

Segundo (COSTA, EVANDRO *et al.*, 2012) a tarefa de classificação diz respeito ao processo de encontrar um modelo que descreve e distingue classes de dados ou conceitos. Os modelos são derivados com base nas análises de coleções de dados, denominadas conjuntos de treinamentos, os quais correspondem a objetos de dados para os quais os rótulos de classes são conhecidos.

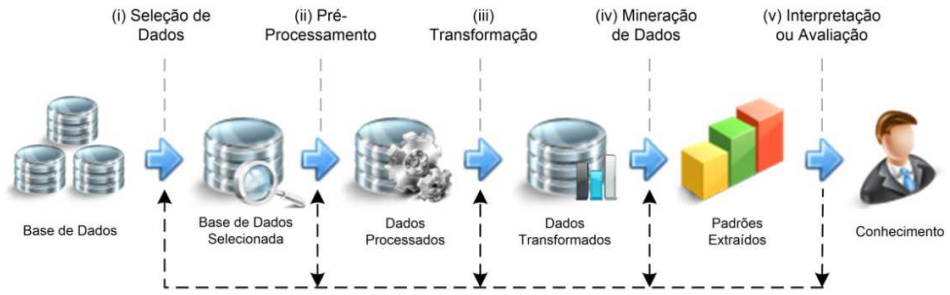
O modelo é usado para predizer o rótulo de classe de objetos para os quais o rótulo de classe é desconhecido. Ele associa um item de dado a uma ou várias classes predefinidas. Para (HAN, J. AND KAMBER, 1992) os modelos derivados podem ser representados em várias formas, tais como: árvore de decisão, regras de classificação, funções matemáticas e redes neurais. Enquanto na classificação a predição é feita para um atributo classificador que assume valores discretos, em modelos de regressão a variável alvo é continua, ou seja, associa um item de dado a uma ou mais variáveis de predição de valores reais.

Segundo (COSTA, EVANDRO *et al.*, 2012) a análise de agrupamento de dados procura associar um item de dado com um ou vários agrupamentos determinados pelos dados, valendo-se principalmente de medidas de similaridades. Já a abordagem de mineração de regras de associação busca encontrar possíveis relações interessantes entre atributos de uma base de dados.

2.3.3 Tarefas e Algoritmos de Mineração de Dados Educacionais

Nesta seção serão apresentadas as etapas a serem cumpridas a fim de se obter as informações e as regras necessárias, a serem aplicadas para a predição. O processo pode ser definido como sendo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis conforme ilustrado na Figura 1.

Figura 1 Etapas do processo de DCBD



Fonte: adaptado por (RIGO et al., 2014) de (FAYYAD, 1996)

Os itens a seguir apresentam as etapas do processo de descoberta de conhecimento em banco de dados - DCBD:

- Seleção de dados: a partir do entendimento do domínio da aplicação, é realizada a seleção de um conjunto de dados de acordo com os objetivos do processo;
- Pré-processamento: os dados disponíveis normalmente não estão em um formato preparado para a análise final, podendo apresentar inconformidades tais como: duplicidade, falta de consistência, ruídos entre outros. Dessa forma torna-se necessária a aplicação de métodos para tratamento, limpeza e redução de volume de dados para que os dados estejam em perfeitas condições para as próximas etapas de análise;
- Transformação: nesta etapa os dados já pré-processados, serão transformados para que possam ser utilizados nos algoritmos de extração de padrões, conforme a tarefa e objetivo da mineração de dados;
- Mineração de dados: nesta etapa consiste na aplicação do algoritmo de mineração de dados selecionado, que é o responsável pela extração do conhecimento implícito no conjunto de dados selecionados. Caracteriza-se aqui a descoberta do conhecimento e padrões;
- Interpretação ou avaliação: nesta etapa o conhecimento extraído é analisado quanto a conhecimento útil e novo, e caso caracterize-se como tal, poderá ser utilizado no suporte ao processo de tomada de decisão na área de domínio da aplicação.

Segundo (FAYYAD, 1996) as tarefas de mineração são de associação, classificação, agrupamento e a sumarização. Quanto ao objetivo podem ser:

- Descritiva: objetiva gerar padrões descritivos através da avaliação do comportamento dos dados;
- Preditiva: utiliza algumas variáveis da base de dados como atributos para prever valores desconhecidos ou futuros de outras variáveis de interesse.

Conforme (COSTA, EVANDRO *et al.*, 2012) na tarefa de predição, a meta é desenvolver modelos que façam inferência sobre aspectos específicos dos dados por meio da análise e associação dos diversos aspectos encontrados nos dados. Um modelo preditivo pode ser entendido como uma função $f(X, \beta) \approx Y$, onde X é um conjunto de variáveis preditoras, β são parâmetros desconhecidos e Y é a variável preditiva. Em outras palavras, deseja-se estimar o valor de Y por meio da descoberta de β utilizando-se X. No processo de predição, é fundamental que boa parte dos dados sejam rotulados manualmente, ou seja, a aprendizagem do modelo ocorrerá de forma supervisionada e dar-se-á utilizando um conjunto de treinamento com valores previamente conhecidos de Y.

2.3.4 Árvore de decisão

De acordo com (HAN, J. AND KAMBER, 2000) uma árvore de decisão possui uma estrutura de árvore, onde cada nó interno pode ser entendido como um atributo de teste, e cada nó-folha possui um rótulo de classe. Árvores de decisão são modelos estatísticos que

utilizam treinamento supervisionado para classificação e predição dos dados, ou seja, no conjunto de treinamento as variáveis preditivas Y são conhecidas.

Basicamente uma árvore de decisão AD permite dividir recursivamente um conjunto de dados de treino até que cada divisão forneça uma classificação para a instância. As AD conforme apresentado na Figura 2 consistem em nós que formam uma árvore, o que significa que, existe um nó-raiz que não tem ramos de entrada, ao contrário dos restantes nós. Cada nó intermediário específico é um teste para o atributo, e cada ramo descendente desse nó corresponde ao valor possível do atributo. Esse conjunto de regras é seguido até ser atingido o nó-terminal ou folha (MAIMON, O. & LIOR, 2010).

Segundo (RYSZARD S MICHALSKI, IVAN BRATKO, 1998) as árvores geradas seguem a seguinte estrutura:

- Folhas: correspondem às classes;
- Nós: são os atributos nos quais estão ligadas subárvores;
- Ramos: são os valores dos atributos.

Figura 2 Exemplo de uma árvore de decisão

Fonte: (RYSZARD S MICHALSKI, IVAN BRATKO, 1998)

A árvore de decisão é, portanto um dos métodos utilizados para se alcançar o objetivo de classificação na predição, a árvore pode implementar vários algoritmos, a seleção do algoritmo mais adequado pode ser obtida comparando a acurácia e confiabilidade do índice *Kappa* dos resultados obtidos na fase de treinamento e assim aplicar a regra gerada no conjunto de teste.

3 TRABALHOS RELACIONADOS

Este capítulo tem como objetivo apresentar trabalhos relacionados à área de Mineração de Dados Educacionais. A seleção se deu através de pesquisa em bases científicas conhecidas, como a Revista Brasileira de Informática na Educação (RBIE), Simpósio Brasileiro de Informática na Educação (SBIE), Springer, ACM, IEEE, entre outras.

Foram utilizadas, para fins de filtro as seguintes palavras-chave "Predição", "Educação a Distância", "Árvore de Decisão" e "Reprovação" para as bases brasileiras; e "*Prediction*", "*e-learning*", "*Decision Tree*" e "*Disapproval*" para as bases internacionais.

3.1 Predição de Reprovação de Alunos de Educação a Distância Utilizando Contagem de Interações.

Este trabalho de autoria de Douglas Detoni e Ricardo Matsumura Araújo (DETONI; ARAUJO; CECHINEL, 2014) teve como contexto a Universidade Federal de Pelotas (UFPEL) que oferece cursos de graduação, pós-graduação e extensão na modalidade a distância.

Observa-se que houve uma análise comparativa de modelos existentes e a verificação de seus melhores índices ao longo das sete semanas de observação como demonstrado na Tabela 1.

De forma geral, as Redes Bayesianas obtiveram os melhores resultados, com exceção da primeira e segunda semana, onde as Florestas Aleatórias obtiveram melhores resultados.

Tabela 1 Precisões para classificação de reprovados – experimento sem utilização de atributos

Caso	Modelo	Semanas						
		S1	S2	S3	S4	S5	S6	S7
Entre Semestres	Rede Bayesiana	0,0	0,22	0,44	0,51	0,60	0,63	**0,64**
	Rede Neural	0,0	0,16	0,33	0,45	0,49	0,51	0,51
	J48	0,0	0,21	0,37	0,40	0,52	0,56	0,56
	Floresta Aleatória	0,1	0,33	0,36	0,40	0,48	0,51	0,54
Entre Turmas	Rede Bayesiana	0,0	0,24	0,45	0,51	0,58	0,62	**0,66**
	Rede Neural	0,0	0,14	0,33	0,46	0,49	0,53	0,56
	J48	0,0	0,20	0,39	0,42	0,51	0,59	0,60
	Floresta Aleatória	0,1	0,28	0,33	0,41	0,50	0,56	0,61

Fonte: (DETONI; ARAUJO; CECHINEL, 2014)

Os autores concluíram que a possibilidade de prever com antecedência se um estudante de educação a distância corre o risco de não concluir uma disciplina ou curso é de grande valia para professores e tutores, que podem ajustar seus instrumentos pedagógicos para evitar que estes estudantes reprovem ou evadam.

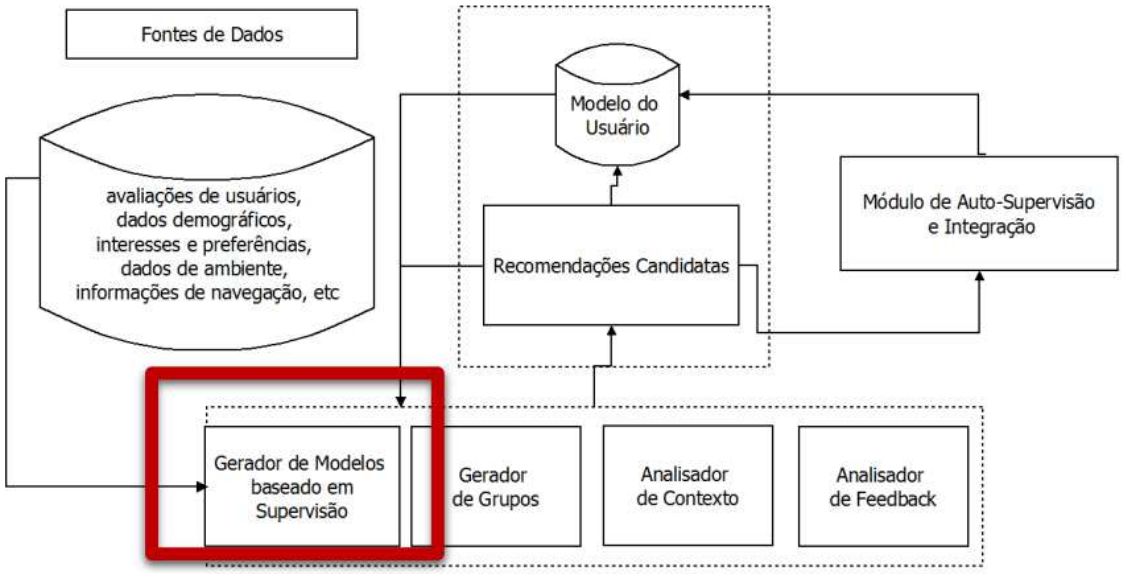
3.2 Predição do Desempenho do Aluno usando Sistemas de Recomendação e Acoplamento de Classificadores

Segundo os autores Reginaldo Gotardo, Paulo Roberto Massa Cereda e Estevam Rafael Hruschka Junior (GOTARDO; CEREDA; JUNIOR, 2013) o processo de aprendizado de máquina apresenta algumas abordagens a respeito de como decidir sobre novos dados adquiridos e seus relacionamento. O aprendizado supervisionado usa dados já rotulados previamente para predizer os rótulos de novos dados. A maior vantagem do aprendizado supervisionado é a consistência estabelecida pelas relações descobertas.

Os autores apresentaram uma abordagem de aprendizado contínuo usando múltiplos algoritmos e visões para garantir que os dados obtidos se aproximem ao máximo do conjunto real e suas relações existentes.

A abordagem geral do trabalho é apresentada na Figura 3. No trabalho foi apresentado o funcionamento do componente *Supervision-based Model Generator*, que está em destaque por ser o modelo de supervisão proposto, e resultados obtidos com o teste do mesmo.

Figura 3 Abordagem Geral



Fonte: (GOTARDO; CEREDA; JUNIOR, 2013)

A auto supervisão é feita internamente através da própria integração de algoritmos diferentes, colaborando entre si, e externamente ao componente através de regras que associam os resultados dos algoritmos.

Para aumentar o tamanho da base de dados disponível foi utilizada a abordagem de *Bootstrapping* na qual o conjunto de treinamento utilizado vai sendo incrementado escolhendo-se as melhores instâncias classificadas do conjunto de teste. Nos testes foram usadas probabilidades de acerto do classificador com o *NaiveBayes*.

A partir de um determinado tamanho de conjunto de treinamento o classificador passava a diminuir continuamente sua taxa de acerto.

A solução usa a abordagem de aumentar o tamanho do conjunto de treinamento a cada execução do algoritmo. Para diminuir este efeito foi usado mais de um algoritmo trabalhando em conjunto e rotulando os dados de outro algoritmo, de forma cruzada. Segundo (BLUM, A, 1988) este acoplamento permite um modelo com melhor desempenho.

Os resultados obtidos mostram que o comportamento do algoritmo na abordagem proposta é similar ao uso dos dados originais da amostra. Através destes resultados criaram-se recomendações que pudessem indicar preferências do aluno que obtém melhores resultados.

3.3 Modelo de Regressão Linear aplicado à previsão de desempenho de estudantes em ambiente de aprendizagem

Segundo os autores Rodrigo Lins Rodrigues, Francisco P. A. de Medeiros e Alex Sandro Gomes(RODRIGUES, R. L., DE MEDEIROS, F. P. A., & GOMES, 2013) extrair informações relevantes que auxiliem a gestão da aprendizagem e viabilizem o acompanhamento efetivo de estudantes em cursos mediados por tecnologia é um desafio.

O objetivo principal deste trabalho foi investigar a viabilidade da utilização do modelo de regressão linear para a obtenção de inferências em etapas iniciais da realização de cursos online, como forma de apoiar a tomada de decisão por parte de professores e gestores.

Nesse sentido, foi proposta a utilização da técnica de regressão linear para estimar o desempenho de alunos baseados em suas interações dentro da plataforma virtual de aprendizagem. Mais uma vez verifica-se que não há um modelo novo, mas a utilização de um modelo existente para aferir se é adequado na estimativa do desempenho dos alunos.

Os autores chegaram à conclusão de que modelo linear foi considerado um bom modelo para explicar que existe uma relação entre a quantidade de interações via fórum de discussão e o desempenho dos alunos.

3.4 Previsão de Desempenho de Estudantes em Cursos EaD Utilizando Mineração de Dados: uma Estratégia Baseada em Séries Temporais.

Segundo os autores Ernani Gottardo, Celso Kaestner, Robinson Vida Noronha (GOTTARDO; KAESTNER; NORONHA, 2012) prover informações relevantes que auxiliem o desenvolvimento de processos de acompanhamento efetivo de estudantes em cursos a distância é um dos objetivos da comunidade científica de Informática na Educação. Neste trabalho, técnicas de mineração de dados educacionais foram utilizadas para geração de inferências sobre o desempenho de estudantes a partir de dados coletados em séries temporais. O objetivo principal foi investigar a viabilidade de obtenção destas informações em etapas iniciais de realização do curso, deforma a apoiar a tomada de ações proativas. Nos experimentos realizados foram utilizados os algoritmos de classificação "*Random Forest*" e "*Multilayer Perceptron*" (WITTEN, I.H., FFRANK E., HALLL, 2011).

Nos dois experimentos o maior percentual de acurácia foi obtido na parte intermediária das divisões realizadas. Entretanto, apenas no experimento 2 a taxa obtida neste período foi estatisticamente superior ao período inicial. A explicação deste fato poderia ser investigada através de uma análise mais detalhada da organização, estrutura e atividades desenvolvidas, preferencialmente envolvendo pessoas ligadas ao curso, como coordenadores, professores, tutores, entre outros.

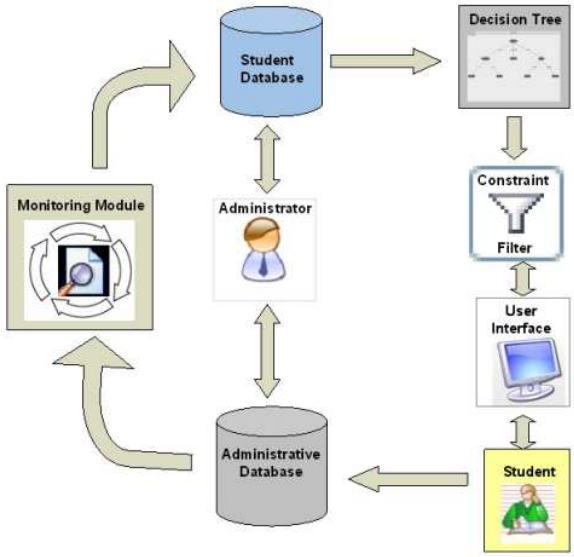
3.5 A decision tree based system for student academic advising and planning in information systems programmers.

Este trabalho de autoria de Naoufel Werghi and Faouzi Kamoun (WERGHI; KAMOUN, 2010) apresenta um Sistema de Apoio à Decisão (SAD) para aconselhamento do estudante. Os autores descrevem o novo paradigma que modela o aconselhamento do estudante como um problema de pesquisa, em que o espaço de busca é representado por uma árvore de decisão que incorpora praticamente todas as instâncias de um plano acadêmico do aluno.

O sistema implementa implicitamente, através da árvore de decisão, muitas regras acadêmicas. Ele permite uma busca sistemática e exaustiva das diferentes instâncias do plano do estudante, uma avaliação metodológica e medição da adequação de um plano acadêmico a um determinado estudante.

O SAD é composto por três componentes padrões: o banco de dados, o modelo e interface do usuário, como apresentado na figura 4. O banco de dados é um repositório de dados do aluno. A partir deste banco de dados pode-se extrair um perfil do estudante, que engloba todas as informações necessárias para os assessores.

Figura 4 Arquitetura DSS e do fluxo de informações



Fonte: (WERGHI; KAMOUN, 2010)

O banco de dados administrativo é um banco de dados relacional que armazena as informações estatísticas sobre os alunos, informações sobre o curso, horários e professores. A configuração do perfil pode ser feita por um administrador ou pode ser delegada a um módulo de monitoramento, atuando como mediador entre o banco de dados administrativo e o banco de dados do aluno.

A unidade de modelagem composta da árvore de decisão e do filtro de restrição é o núcleo do sistema. Funciona como um consultor virtual e incorpora os processos associados com as tarefas do conselheiro, como busca por diferentes alternativas e opções para o planejamento de curso, e selecionando os cursos que são compatíveis com as normas acadêmicas, proporcionando a melhor correspondência com o perfil do estudante.

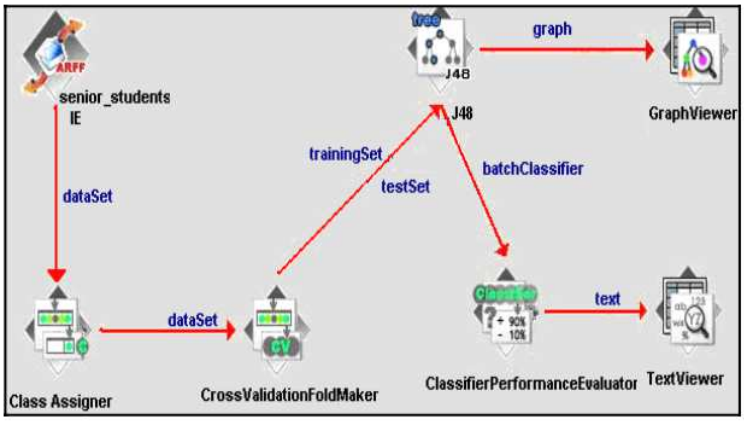
3.6 Analysis and predictions on students behavior using decision trees in weka environment

Este trabalho de autoria de Bresfelean, Vasile Paul (BRESFELEAN, 2007) representa uma implementação de uma ferramenta de análise de algoritmo J48 em dados coletados a partir de pesquisas em diferentes alunos de especialização de uma faculdade, com o objetivo de prever suas escolhas em continuar sua educação com estudos de pós-graduação, através de árvores de decisão.

Para o presente trabalho, optou-se por usar o J48 porque tem desempenho melhor do que ID3 em quase todas as circunstâncias.

Optou-se por componentes do WEKA e estes foram relacionados em um grafo gerado do processo de investigação que é iniciado com a carga do arquivo dos alunos do curso de Informática Econômica representado pelo arquivo IE.arff na ferramenta , que utiliza a classe *Assigner* para classificar os atributos. Seguindo o fluxo valida os atributos com a classe *CrossValidationFoldMaker*, encaminha o conjunto de treinamento e posteriormente o conjunto de teste para o classificador J48 que gera o lote classificado como é apresentado na Figura 5.

Figura 5 O modelo gráfico construído.



Fonte: (WITTEN, I.H., FFRANK E., HALLL, 2011)

3.7 Comparações

Para comparar os trabalhos pesquisados, foram consideradas as estratégias utilizadas para fazer a predição, o foco de aplicação e os serviços entregues. A Tabela 2 apresenta um comparativo entre os trabalhos e enfoca de que forma cada uma das estratégias foram apresentadas. A seguir é feita a descrição de cada critério da comparação:

- Abordagem com interações: As interações com as atividades realizadas no ambiente virtual de aprendizagem são relevantes por quantificarem a incidência de procura pela atividade pedagógica dessa forma todos os modelos nesse ambiente se utilizam desta abordagem;
- Acoplamento de Classificadores: identifica se foram utilizados vários classificadores no processo de predição;

- Séries Temporais: identifica se foram utilizadas séries temporais no processo de predição;
- Regressão Linear: identifica se foi utilizada regressão linear no processo de predição;
- Árvore de Decisão: identifica se foi utilizada a técnica de mineração de dados árvore de decisão no processo de predição;
- Educação a Distância: identifica se foi utilizado em Educação a distância;
- Foco em Reprovação: identifica se todo tem foco em reprovação;
- Lista com as notas das atividades do grupo de risco**:** identifica se existe a preocupação em apresentar um arquivo de saída contendo o grupo de risco gerado na predição para fins de disponibilização a um sistema de recomendação;
- Percentual individual de indicador de reprovação**:** identifica se existe algum indicador de taxa de predição de reprovação individualizado, indicando o quanto o aprendiz tende a reprovar ou aprovar.

Tabela 2 Comparação dos trabalhos

Trabalhos	**Estratégias para a predição**					**Foco**		**Serviços**	
	Abordagem com interações	Acoplamento de Classificadores	Séries Temporais	Regressão Linear	Árvore de Decisão	EaD	Reprovação	Fornece lista com as médias das atividades do grupo de risco	Fornece percentual indicador de reprovação
Predição de Reprovação de Alunos de Educação a Distância Utilizando Contagem de Interações (DETONI; ARAUJO; CECHINEL, 2014)	Sim	Não	Não	Não	Não	Sim	Não	Não	Não
Predição do Desempenho do Aluno usando Sistemas de Recomendação e Acoplamento de Classificadores(GOTARDO; CEREDA; JUNIOR, 2013)	Sim	Sim	Sim	Não	Não	Sim	Não	Não	Não
Modelo de Regressão Linear aplicado à previsão de desempenho de estudantes em ambiente de aprendizagem(RODRIGUES, R. L., DE MEDEIROS, F. P. A., & GOMES, 2013)	Sim	Não-	Não	Sim	Não	Sim	Não	Não	Não
Previsão de Desempenho de Estudantes em Cursos EaD Utilizando Mineração de Dados: uma Estratégia Baseada em Séries Temporais(GOTTARDO; KAESTNER; NORONHA, 2012)	Sim	Não	Sim	Não	Não	Sim	Não	Não	Não

Trabalhos	Estratégias para a predição					Foco		Serviços	
	Abordagem com interações	Acoplamento de Classificadores	Séries Temporais	Regressão Linear	Árvore de Decisão	EaD	Reprovação	Fornece lista com as médias das atividades do grupo de risco	Fornece percentual indicador de reprovação
A decision-tree-based system for student academic advising and planning in information systems programmes. (WERGHI; KAMOUN, 2010)	Não	Não	Sim	Não	Sim	Não	Não	Não	Não
Analysis and predictions on students' behavior using decision trees in weka environment(BRESFELEAN, 2007)	Sim	Não	Sim	Não	Sim	Sim	Não	Não	Não

Fonte: Próprio autor

No primeiro trabalho, observou-se a comparação de modelos tradicionais de aprendizado de máquina como Rede Bayesiana, Rede Neural, J48 e Floresta Aleatória concluindo por fim que Rede Bayesiana em se tratando de predição tendo como atributo exclusivamente a contagem de interações apresentou melhor precisão segundo (DETONI; ARAUJO; CECHINEL, 2014).

O segundo trabalho (GOTTARDO; KAESTNER; NORONHA, 2012) tem como foco a previsão não de reprovação, mas de desempenho do estudante, como este trabalho busca coletar informações que sirvam de insumos para os sistemas de apoio a decisão julgou-se relevante sua seleção.

O terceiro trabalho (RODRIGUES, R. L., DE MEDEIROS, F. P. A., & GOMES, 2013) igualmente não tem foco em reprovação, mas de forma análoga ao segundo trabalho se preocupa com o desempenho do aluno e melhorar a atuação do aluno está diretamente relacionado ao seu estímulo e melhoria dos seus resultados que por fim corroboram com a redução da reprovação. Ele usa a regressão linear para estimar o desempenho de alunos baseado em interações dentro da plataforma.

O quarto trabalho (GOTARDO; CEREDA; JUNIOR, 2013) mostra outra iniciativa em predição não de reprovação, mas de desempenho, no entanto seus resultados também são análogos ao que este trabalho se propõe. Utilizou a técnica de acoplamento que partindo de um conjunto mínimo consegue simular e prever uma situação real com boas taxas de precisão.

O quinto trabalho (WERGHI; KAMOUN, 2010) não tem relação com ambientes virtuais de aprendizagem, nem se preocupa com predição de desempenho ou de reprovação de alunos, mas se utiliza do classificador de árvore de decisão que é o foco deste trabalho, para auxiliar alunos nas suas escolhas de disciplinas contribuindo como facilitador da montagem do plano de curso de cada aluno.

O sexto trabalho (BRESFELEAN, 2007) é o que mais se aproxima desta proposta, pois utiliza a ferramenta de mineração WEKA e utiliza o classificador J48 para fazer a predição de comportamento de estudantes.

O MD-PREAD traz um diferencial quanto à possibilidade de interpretação dos dados gerados pelo uso dos métodos de predição, pois outros métodos usados, tais como Redes Neurais Artificiais possuem como deficiência justamente a dificuldade de identificar as causas que levam aos resultados das predições.

A coleta de registros históricos das avaliações e interações dos aprendizes com o ambiente permitiu identificar a regra de classificação pelo uso do algoritmo de árvore de decisão *Information_Gain*, apresentando dados que logo após as primeiras atividades avaliadas pelo professor, puderam auxiliar na tomada de decisão quanto à melhoria do estímulo e motivação do aluno e melhoraria da sua absorção de conteúdo.

Isso foi possível em função de que o modelo gerou um arquivo com as informações do grupo de risco de reprovação que foi utilizado em um sistema de recomendação desenvolvido por um mestrando da UNISINOS, auxiliando assim o gestor educacional na tomada de decisão.

Assim, foi possível identificar a possibilidade de reprovação na disciplina e fazer o encaminhamento do grupo de risco de reprovação ao Sistema de Recomendação de forma antecipada. Este trabalho contribuiu para que houvesse um acompanhamento da reprovação dos alunos no decorrer de disciplinas.

4 MODELO MD-PREAD

Neste capítulo será apresentado o modelo de predição de reprovação. Este modelo explora a mineração de dados, tendo como público-alvo gestores educacionais e sistemas de recomendação. O modelo tem o intuito de fornecer informações que contribuam na tarefa de minimizar a reprovação de aprendizes em disciplinas, e recebe o nome de MD-PREAD.

A primeira seção apresenta uma visão geral sobre o modelo e suas principais características. Na segunda seção é abordada a arquitetura do modelo. A terceira seção trata do módulo de importação dos dados, que realiza buscas de elementos em um banco de dados. Na quarta seção é descrito o módulo de processamento dos dados, o qual trata as informações obtidas a fim de identificar as regras de classificação responsáveis pela predição. Na quinta seção é descrito o funcionamento do módulo de exportação que gera os arquivos de acordo com os perfis de reprovação encontrados a ser disponibilizados para sistemas de recomendações educacionais externos.

4.1 Visão Geral

O MD-PREAD é um modelo de sistema de predição educacional para gestores educacionais e sistemas de recomendação, com foco na reprovação em disciplinas. Suas principais características são:

- Suporte a vários níveis de EaD: apesar da abordagem a cursos de graduação e pós-graduação neste trabalho, o modelo não restringe o nível de ensino a ser analisado;
- Gerenciamento de predições: permite que durante o andamento de uma disciplina, várias predições possam ocorrer;
- Gerenciamento do perfil do aluno: permite a utilização de perfis de aprendizes para o gerenciamento das predições;
- Suporte a gerenciamento de geração de arquivos de lote: permite gerar arquivos contendo o grupo de risco de reprovação.

Com base nessas características, através dos seus módulos internos, o modelo permite que seja possível usar as informações de dados do Sistema de Gestão Acadêmica e do Ambiente Virtual de Aprendizagem para fazer predição de reprovação de aprendizes em disciplinas. As informações são relevantes para gestores educacionais e sistemas de recomendação.

4.2 Arquitetura

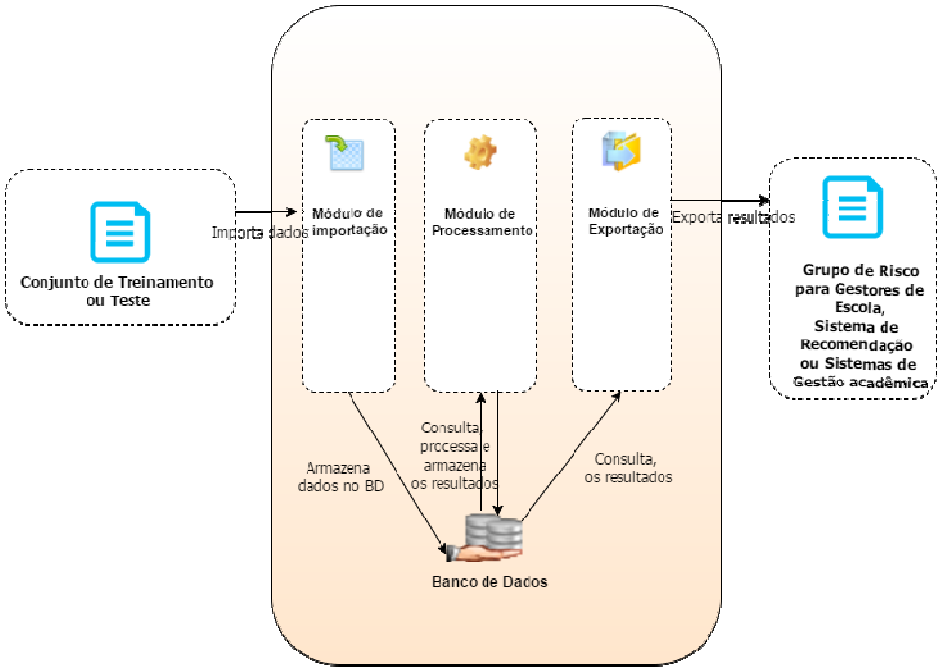
A Figura 6 apresenta o MD-PREAD, em que se pode observar um ambiente externo de entrada, o modelo e um ambiente externo de saída. O primeiro representado pelo conjunto de treinamento ou conjunto de testes, simbolizado pela figura de um arquivo de dados de entrada pré-selecionados do Ambiente Virtual de Aprendizagem e do Sistema de Gestão Acadêmica da Instituição.

O segundo é o MD_PREAD organizado em três componentes, chamados de módulos: Módulo de importação, Módulo de processamento e Módulo de exportação. O modelo propõe um sistema de predição centrado na disciplina. Antes do processo preditivo são extraídos dados históricos do ambiente virtual para fins de treinamento de um algoritmo classificador de árvore de decisão que busca uma regra para classificar os possíveis reprovados com base nas interações e médias das avaliações parciais. Neste trabalho foram consideradas interações os acessos dos aprendizes aos recursos e avaliações do ambiente virtual.

No transcorrer da disciplina, considerando as interações dos aprendizes nas diversas atividades do ambiente virtual e suas avaliações semanais, ocorre um processo de extração dessas informações que são alimentadas no sistema de predição, este por sua vez aplica a regra encontrada na fase de treinamento.

Após essa etapa, já é possível realizar a exportação do grupo de risco que é o conteúdo do arquivo de saída que será destinado a um sistema de recomendação educacional externo.

Figura 6 Arquitetura do MD-PREAD



Fonte: Próprio Autor

4.3 Módulo de Importação

O módulo de importação é responsável por importar os arquivos de treinamento ou teste para que seja dado início ao processo de geração de regra ou predição. O arquivo precisa ser preparado e são necessários arquivos de saída extraídos a partir de dois sistemas, o de Gestão Acadêmica e os do Ambiente virtual de aprendizagem, através de três listagens, a primeira contendo dados de identificação dos aprendizes conforme descrito a seguir:

- Matrícula: identificação da matrícula do aluno no registro escolar indicando que se encontra regularmente matriculado na instituição de ensino;
- Nome: descrição do nome do aluno permite que a identificação não seja apenas de um código, mas de forma personalizada podemos saber quem é o aprendiz em situação de possível reprovação;
- E-mail: descrição do e-mail é relevante para possibilitar o envio de mensagens caso seja necessário pelos gestores acadêmicos ou sistemas de recomendação;
- Curso: descrição do curso importante porque além de identificar o curso a que pertence o aprendiz permite agrupar os alunos na elaboração relatórios;
- Matriz curricular: descrição da matriz curricular em que o aprendiz tem vincula, considerando que no ambiente virtual as disciplinas não estão atreladas a cursos, essa informação é relevante para fins de organização dos dados;
- Nascimento: descrição da data de nascimento que nos possibilita calcular a idade do aprendiz e usar essa informação para agregar na geração do perfil;
- Período letivo inicial: descrição do período letivo inicial essa informação é importante porque permite situar o aprendiz ao longo do cumprimento das disciplinas do curso;
- Renda familiar: informação da renda familiar, essa informação é importante porque pode auxiliar na definição do perfil do aprendiz;
- Sexo: descrição do sexo é importante porque pode contribuir para contribuir com a construção do perfil do aprendiz;
- Tipo de escola de origem: descrição da Escola de origem, se pública, particular ou mista, igualmente importante pois pode contribuir na definição do perfil do aprendiz;
- Turma: descrição da turma, este é mais um atributo que possibilitar agrupamentos em elaboração de relatórios.

Uma segunda listagem contendo as interações dos aprendizes no ambiente virtual de aprendizagem conforme descritas seguir.

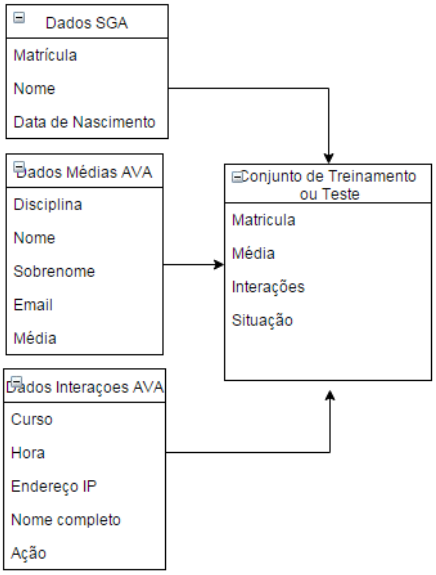
- Curso: descreve a disciplina no AVA, pois neste ambiente o que ele chama de curso na verdade são disciplinas que cada professor administra;
- Hora: a hora da Interação, registro que nos permite saber o momento em que a interação ocorreu;
- Endereço IP: identificação do Protocolo de Identificação do Hardware que originou a interação do aprendiz;
- Nome completo: identifica o aluno;
- Ação: descrição das atividades e ou recursos utilizados pelo aprendiz no ambiente;
- Informação: Descreve várias informações que podem servir de filtros em consultas como o nome do professor, o nome da disciplina o nome da atividade ou nome do recurso.

E uma terceira listagem contendo as médias parciais das avaliações na disciplina cursada pelos aprendizes no ambiente virtual de aprendizagem conforme a seguir.

- Disciplina: descrição da Disciplina, nome da disciplina na qual o aprendiz está fazendo parte;
- Nome: descrição do primeiro nome, o AVA na exportação da tabela de notas separa o nome do sobrenome o que requer um tratamento de concatenação para que se possa relacionar as tabelas;
- Sobrenome: descrição do sobrenome essa informação é importante porque precisará ser concatenada para que o nome seja composto e possa ser relacionado;
- Endereço de e-mail: descrição do Endereço de e-mail que é importante pois auxilia na correlação das tabelas;
- Média das atividades: descrição da média das atividades, como os alunos fazem várias avaliações, são consideradas as médias para fins de cálculo das regras de perfil de reprovação.

Antes de iniciar o processo de importação que disponibiliza o conjunto de treinamento para identificar a regra, é necessário preparar o conjunto de teste, o que é feito selecionando os atributos que são usados para a predição conforme ilustrado na Figura 7. É importante ressaltar que os dados utilizados são exportados do AVA e pode-se observar que na exportação de médias não exporta a matrícula como identificador, apenas nome, sobrenome e e-mail. Já na tabela de interações o identificador é o nome completo, havendo a necessidade, portanto, de se realizar um preparo dos dados a fim de se garantir a consistência e confiabilidade dos dados utilizados.

Figura 7 Preparação dos dados

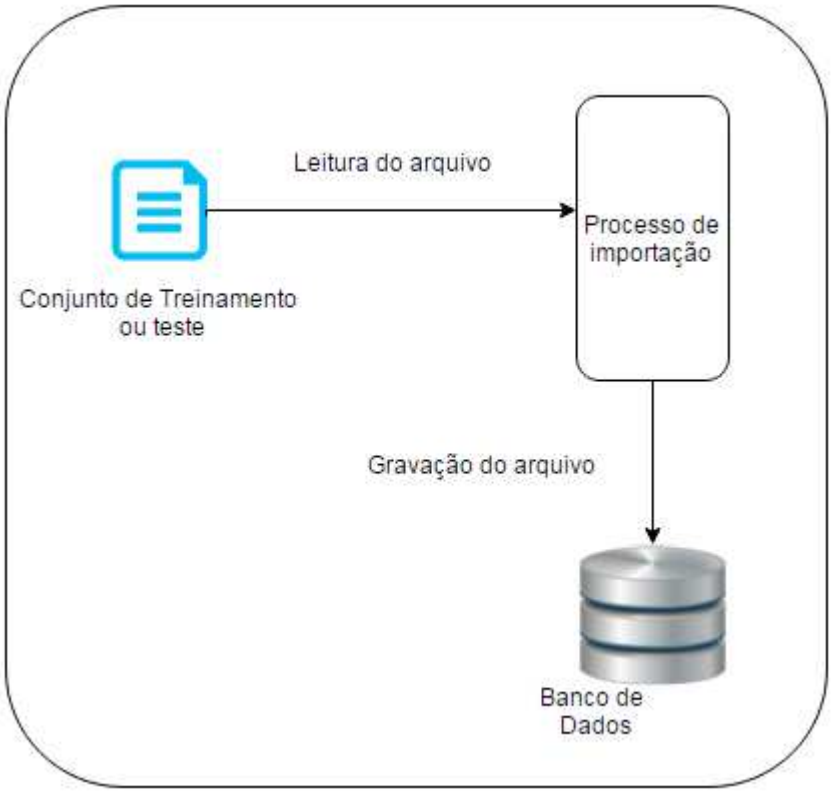


Fonte: Próprio autor

Este mesmo processo de preparação é análogo para os dados de treinamento e teste, a diferença é que para fins de treinamento os dados são completos e já se sabe a situação final do aprendiz enquanto que no conjunto de teste a informação da situação não é conhecida.

Após a etapa de preparação dos dados, estes são importados para o modelo e torna-se possível o prosseguimento do processo de treinamento ou predição pelo módulo de processamento. O Módulo de importação faz a leitura dos arquivos de treinamento para a definição da regra e posteriormente dos arquivos do conjunto de testes para a predição. A Figura 8 apresenta o Módulo de importação.

Figura 8 Módulo de Importação



Fonte: Próprio autor

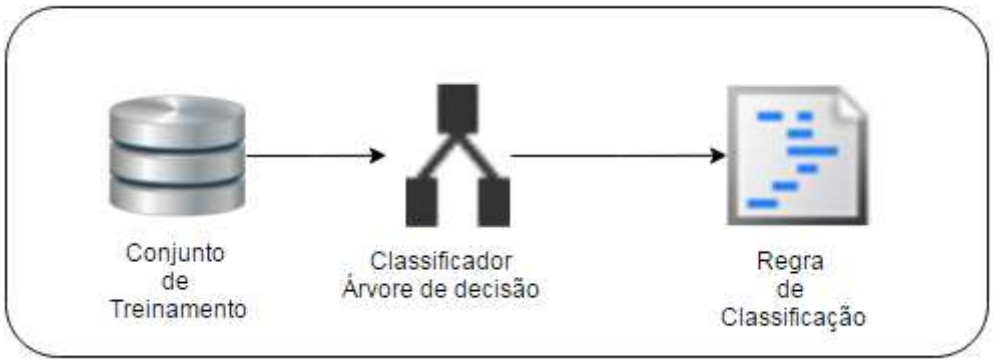
Para realizar este trabalho, não são considerados para fins de treinamento, estudantes desistentes, tendo em vista a indisponibilidade do resultado da situação de aprovação ou reprovação bem como a nota final nestes casos. A escolha dos atributos para compor o conjunto de dados experimentais sobre os estudantes selecionados é baseada na tabela de Notas e interações exportada do Ambiente Virtual de Aprendizagem bem como de dados do Sistema de Gestão Acadêmica.

4.4 Módulo de Processamento

A Figura 9 apresenta o processo de geração da Regra de Decisão através do treinamento. O módulo de processamento é o mais relevante pois é ele que define com base em um conjunto de treinamento a regra de classificação. Os dados do conjunto de treinamento

são entregues ao algoritmo de árvore de decisão que de forma dinâmica encontra a regra a ser aplicada no conjunto de teste.

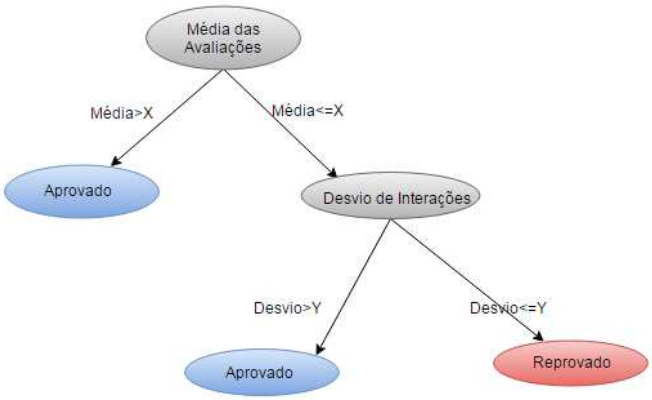
Figura 9 Processo de Treinamento



Fonte: Próprio autor

Para a classificação, é usada uma ferramenta de mineração de dados que tenha componentes que possibilitem a leitura do arquivo de treinamento e o processo de geração da regra de classificação. A escolha do classificador deve levar em consideração os melhores índices de acurácia e confiabilidade *Kappa* (LANDIS; KOCH, 1977). Como os dados disponíveis para este modelo são as médias dos alunos e o desvio padrão das interações (número de acessos), são estes que ao final do processo resultarão em um padrão de comportamento. A árvore de decisão apresenta os nós e as condições a serem testadas para a predição, onde X é a média das avaliações e Y o desvio padrão das interações. A Figura 10 apresenta a árvore de decisão gerada para a predição.

Figura 10 Árvore de Decisão do MD-PREAD



Fonte: Próprio autor

A regra pode ser expressa por:

Se Média das Avaliações for maior que um valor de média encontrado X;

A situação da predição é Aprovado;

Se não se o desvio padrão for maior que um valor encontrado Y;

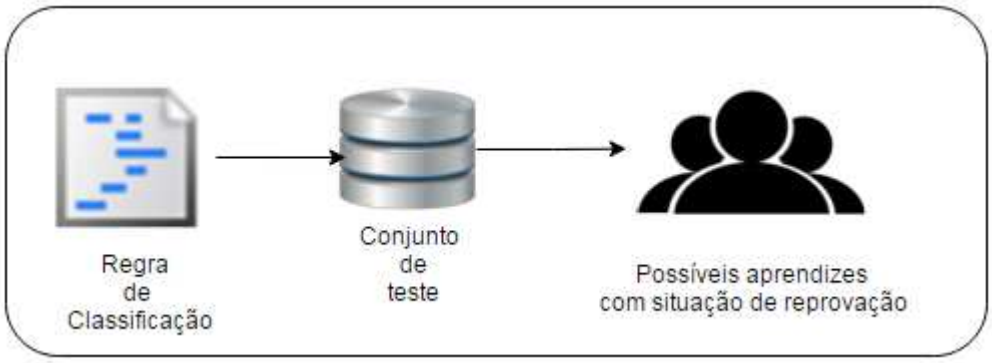
A situação da predição é Aprovado;

Se não se o desvio padrão for menor ou igual Y;

A situação da predição é Reprovado.

O processo de classificação ocorre da seguinte forma, após o processo de treinamento e armazenamento da regra de classificação, o processo de importação faz a leitura do arquivo de teste e este passa pela regra de classificação selecionada gerando como saída a lista do grupo de risco contendo os aprendizes com probabilidade de reprovação. Este processo é apresentado na Figura 11 e deve ser executado a cada semana.

Figura 11 Processo de Predição

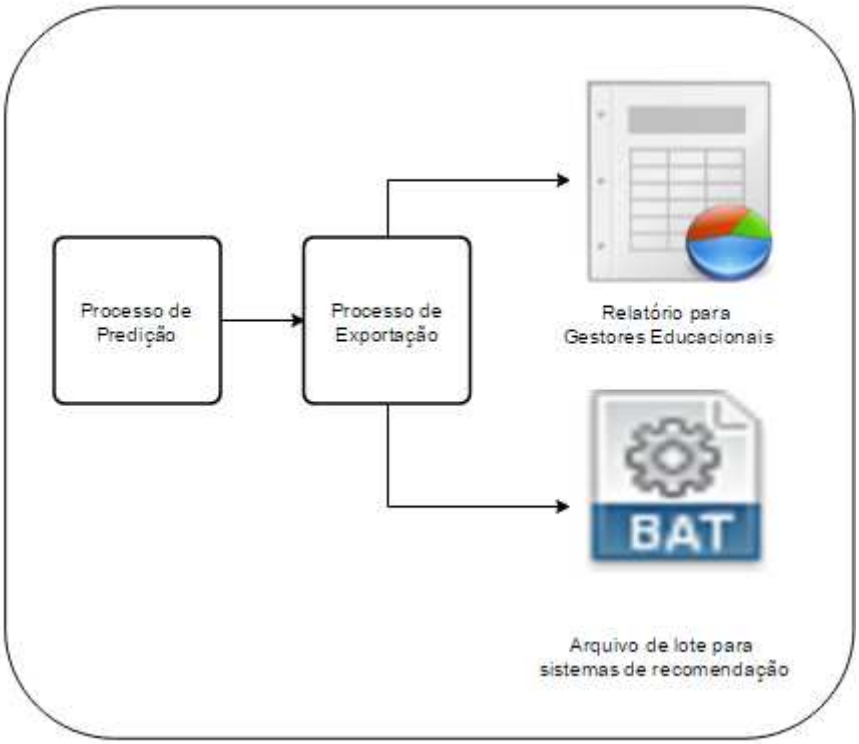


Fonte: Próprio autor

4.5 Módulo de Exportação

O módulo de exportação é o responsável em receber o resultado da predição e gerar os arquivos de saída, no caso de entrega para sistemas de recomendação e gestores educacionais. É composto basicamente por dois processos, o processo de predição que aplica as regras na base de teste e entrega o resultado para o processo de exportação que por sua vez gera os arquivos de saída conforme ilustrado na Figura 12.

Figura 12 Módulo de exportação



Fonte: Próprio autor

5 ASPECTOS DE IMPLEMENTAÇÃO

Neste capítulo são apresentados os aspectos de implementação do modelo. Para permitir a avaliação do MD-PREAD, foi necessário configurar o protótipo em uma ferramenta de mineração de dados capaz de auxiliar nos procedimentos e métodos necessários para cumprir as etapas na busca do perfil do aprendiz com tendência de reprovar em uma disciplina.

Foram feitas coletas de dados históricos de 10 disciplinas de um grupo de 30 aprendizes em 02 semestres consecutivos, 2014/2 e 2015/1, o total de alunos matriculados foi de 125, o total de interações levantadas foi de 41070, foi considerado no cálculo da predição as médias das avaliações de de uma amostra de 30 aprendizes com uma margem de erro de 13% para mais ou para menos, os desvios padrões das interações e suas respectivas situações. Estes dados serviram para compor o conjunto de treinamento necessário para a definição da regra de classificação que teve como predominante a acurácia de 55% e a confiabilidade *Kappa* de 0,22.

5.1 Implementação do MD-PREAD

Nesta seção são apresentadas as etapas de prototipagem do MD-PREAD. Para testar o modelo foi escolhida a ferramenta de mineração de dados *RapidMiner* que possibilitou configurar o modelo e executar a aplicação até a etapa de exportação do arquivo.

Segundo Costa (COSTA, JFA, 2011) o *RapidMiner* é uma aplicação líder mundial dos sistemas *open-source* para DM. Esta ferramenta está disponível como uma aplicação *standalone* para análises de dados, e como um motor de Mineração de Dados para a integração dos seus próprios produtos.

O *RapidMiner* apresenta um conjunto de características, tais como:

- integração de dados, ETL (*Extract*, *Transform*, *Load*), análise de informação e produção de relatórios, tudo numa única suíte;
- tem uma parte gráfica poderosa e intuitiva para análises de processos;
- reconhecimento de erros *on-the-fly* e correções Rápidas.

5.1.1 Configuração do Modelo de Predição

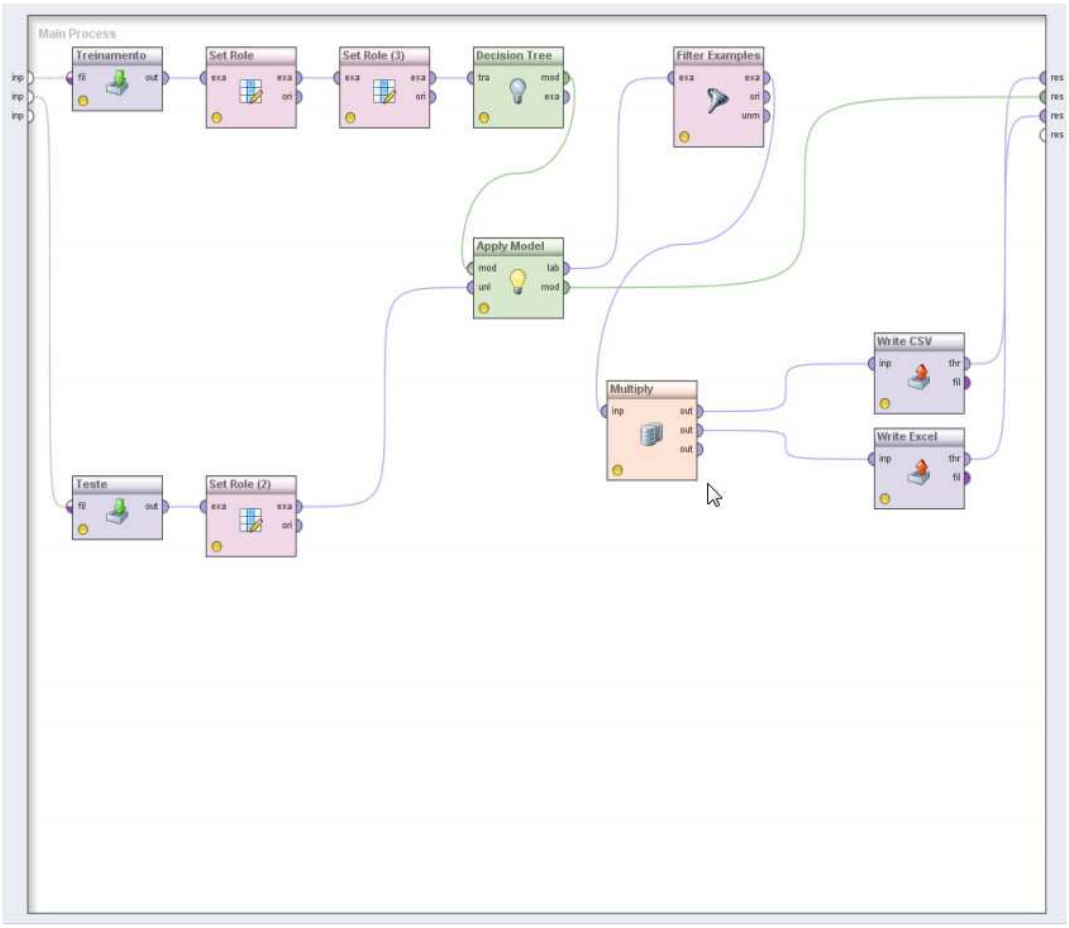
O *RapidMiner* disponibiliza para o usuário processos e operadores, assim criou-se o processo de predição que é composto de vários operadores como descritos a seguir:

- Operador *Read*: possibilita vincular ao processo um arquivo. Neste caso foram usados dois operadores deste tipo vinculados aos conjuntos de treinamento e de testes;
- Operador *Set Role*: configura tanto os atributos que não devam ser considerados, como o atributo de predição. Utilizaram-se três operadores deste tipo, dois para desconsiderar o atributo matrícula e um para indicar o atributo situação como o *Label* (Atributo de predição);

- *Decision Tree*: configura o algoritmo classificador usado na predição;
- *Appy Model*: possibilita aplicar as regras de um classificador a um conjunto de testes;
- *Filter Examples*: possibilita configurar um filtro no resultado da predição;
- *Multiply*: possibilita ramificar as saídas do processo para outros operadores;
- *Write*: recebe a saída do processo e gera um arquivo em um diretório previamente especificado.

A Figura 13 apresenta a configuração do processo de predição no *RapidMiner*. A figura mostra os operadores já citados anteriormente, ou seja, o *read* para a leitura dos conjuntos de teste e treinamento, os operadores *set* para a configuração dos atributos, o operador *Decision Tree* para configurar o classificador da árvore, o operador *Apply Model* para aplicar a regra ao conjunto de treinamento, o operador *filter* para filtrar apenas os aprendizes com predição para reprovar, o operador *Mutiply* para possibilitar a exportação no operador de exportação *write* responsável pela geração dos arquivosde saída.

Figura 13 Processo de Predição no RapidMiner



Fonte: Próprio autor

5.1.2 Processo de Validação

O processo de validação utilizado foi o de Validação Cruzada que é um processo de aprendizagem supervisionada em mineração de dados. Após o pré-processamento e a formatação, os dados são fragmentados em dois subconjuntos, denominados base de treinamento e base de testes. Numa primeira etapa um algoritmo de indução de conhecimento é aplicado à base de treinamento. Com isso se obtém um modelo "treinado", que representa o conhecimento extraído. Numa segunda etapa o modelo obtido é aplicado ao fragmento da base de dados denominado base de testes. Como a base de testes também é previamente rotulada, se pode medir a taxa de acerto do modelo, comparando-se o resultado obtido com a rotulação disponível na base de testes de acordo com (CABENA, P., HADJINIAN, P., STADLER, R., VERHEES, J E ZANASI, 1997). A técnica de Validação Cruzada consiste em dividir a base de dados em x partes, destas, x-1 partes são utilizadas para o treinamento e uma serve como base de testes. O processo é repetido x vezes, de forma que cada parte seja usada uma vez com o conjunto de testes. Ao final, a correção total é calculada pela média dos resultados obtidos em cada etapa, obtendo-se assim uma estimativa da qualidade do modelo de conhecimento gerado e permitindo análises estatísticas. Neste experimento o processo foi repetido 10 vezes.

5.1.3 Escolha do algoritmo classificador de árvore de decisão

A escolha do algoritmo foi feita com base em comparações apresentadas na Tabela 3, realizadas utilizando a base de teste e a verificação dos índices de acurácia, percentual de reprovação e índice Kappa (LANDIS; KOCH, 1977) onde o *Information_Gain* foi o que se destacou.

Tabela 3 Análise dos índices dos algoritmos de classificação

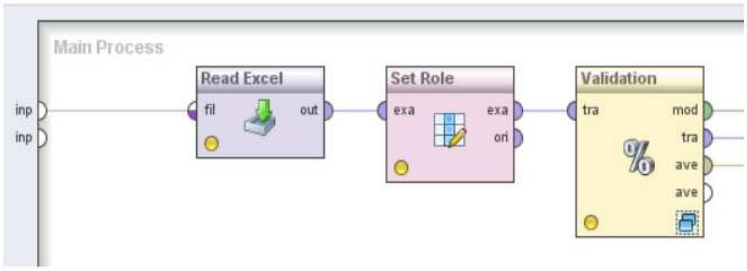
Algoritmos	Critério	Acurácia	Precisão de classificação de reprovação	Kappa
Decision tree	*Gain_ratio*	55,00%	33,33%	0,187
	Information_Gain	**55,00%**	**35,29%**	**0,22**
	Gini_index	49,17%	0,00%	-0,217
	Accuracy	58,33%	0,00%	-0,117
ADTree	-	51,67%	16,67%	0,067
W-J48	-	57,50%	35,71%	0,15

Fonte: Próprio autor

Os números apresentados na Tabela 3 foram extraídos do processo Validação do Processo de Predição através da configuração do operador *Validation* na ferramenta *RapidMiner* apresentado na Figura 14, que quando executado gera o resultado apresentado na Figura 15 para acuraria e Figura 16 para índice *Kappa*. O processo de validação foi executado para todos os parâmetros listados na tabela 3 predominando o *Information Gain* como aquele de melhor confiabilidade.

A Figura 14 apresenta a configuração do Processo de validação, que utiliza o operador *Read* para fazer a leitura do conjunto de treinamento, o *Set Role* para setar o atributo *label* e o operador *Validation* para gerar as estatísticas de acurácia, precisão e confiabilidade.

Figura 14 Configuração do Processo de Validação no RapidMiner

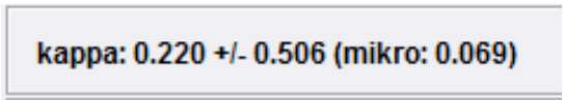


Fonte: Próprio autor

Figura 15 Resultado da Acurácia após execução do processo de validação do Algoritmo

Fonte: Próprio autor

Figura 16 Resultado do Índice Kappa após execução do Processo de Validação do Algoritmo



Fonte: Próprio autor

A seguir são listados alguns dos parâmetros do operador de árvore de decisão mencionados na Tabela 3 que demonstra a seleção do parâmetro com os melhores índices de acurácia:

- *Gain_Ratio*: é uma variante do ganho de informação. Ele ajusta o ganho da informação para cada atributo a fim de permitir que haja amplitude e uniformidade nos valores de atributo;
- *Information_Gain*: com este critério o algoritmo calcula a entropia de todos os atributos. O atributo com a entropia mínima é selecionado para dividir. Este método tem uma tendência para selecionar atributos com um grande número de valores;
- *Gini_Index*: trata a medida de impureza de um *ExampleSet*, divide os atributos escolhidos e faz uma redução no índice gini média dos subconjuntos resultantes;
- *Accuracy*: este algoritmo busca maximizar a precisão da árvore;
- *Kappa*: segundo Landis (LANDIS; KOCH, 1977) é um índice que indica o grau de concordância conforme apresentado na Tabela 4 (LANDIS; KOCH, 1977).

Tabela 4 Índices de Confiabilidade

Values of Kappa	Interpretação
<0	*No agreement*
0-0,19	*Poor agreement*
0,2 – 0,39	*Fair agreement*
0,4 – 0,59	*Moderate agreement*
0,6 – 0,79	*Substantial agreement*
0,8 – 1,00	*Almost perfect agreement*

Fonte:(LANDIS; KOCH, 1977)

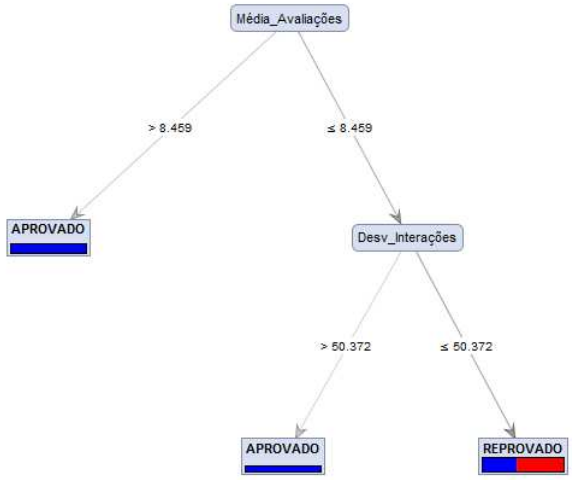
A Figura 17 apresenta a configuração do parâmetro *information_gation* que foi o que apresentou os melhores índices para a predição

Figura 17 Configuração de parâmetro da Árvore de Decisão

Fonte: Próprio autor

A Figura 18 apresenta a árvore de decisão gerada após a execução do processo de validação, mostrando que os alunos com média de avaliações maiores que 8,459 são possíveis aprovados e que as notas inferiores a este valor caso o desvio padrão de todas as interações no sistema seja menor que 50 são possíveis reprovados.

Figura 18 Árvore de decisão encontrada no treinamento



Fonte: Próprio autor

Após a execução do processo de validação também se pode conferir a Regra de classificação gerada que é apresentada na Figura 19.

Figura 19 Regra de classificação encontrada no treinamento

```
Média_Avaliações > 8.459: APROVADO
{APROVADO=13, REPROVADO=0}
Média_Avaliações ≤ 8.459
|   Desv_Interações > 50.372: APRO-
VADO {APROVADO=5, REPROVADO=0}
|   Desv_Interações ≤ 50.372: RE-
PROVADO {APROVADO=8, REPROVADO=12}
```

Fonte: Próprio autor

Os atributos escolhidos para o cálculo da regra foram as médias das avaliações e o desvio padrão das interações no ambiente virtual. A Figura 20 mostra a lista dos possíveis alunos com situação de “reprovar”, já considerando a regra da predição, estes resultados foram obtidos com dados simulados de um conjunto de teste, considerando que nesta fase foi avaliada a viabilidade do modelo.

Figura 20 Resultado do 1º Grupo de Risco - Simulação

Row No.	Matricula	Situação	prediction(Situação)	confidence(APROVADO)	confidence(REPROVADO)	Média_Avaliações	Desv_Interações
1	[redacted]	?	REPROVADO	0.400	0.600	0	23.551
2		?	REPROVADO	0.400	0.600	7.670	43.301
3		?	REPROVADO	0.400	0.600	4.330	13.856
4		?	REPROVADO	0.400	0.600	8	0
5		?	REPROVADO	0.400	0.600	0	0
6		?	REPROVADO	0.400	0.600	8.330	3.536
7		?	REPROVADO	0.400	0.600	8	21.213
8		?	REPROVADO	0.400	0.600	1	42.032
9		?	REPROVADO	0.400	0.600	8.330	14.142
10		?	REPROVADO	0.400	0.600	8.330	0
11		?	REPROVADO	0.400	0.600	3	28.284
12		?	REPROVADO	0.400	0.600	7.330	0
13		?	REPROVADO	0.400	0.600	7.330	0

Fonte: Próprio autor

5.2 Implementação do MD-PREAD pós-experimento

Considerando a baixa acurácia no primeiro treinamento que foi feito com uma amostra aleatória simples, resolveu-se após o experimento realizar um segundo treinamento coletando os dados de toda a população dos 125 alunos que realizaram 115.776 acessos nos períodos de 2014/2 e 2015/1 em 10 disciplinas ofertadas e suas respectivas médias das atividades. Para tanto foram extraídos os dados de forma análoga a seção 5.1 e foram identificados nos acessos os seguintes recursos e atividades:

- *assignment upload*
- *assignment view*
- *blog view*
- *course recent*
- *folder view*
- *forum add discussion*
- *forum add post*
- *forum search*
- *forum update post*
- *forum view discussion*
- *forum view forum*
- *glossary add entry*
- *glossary update entry*

- *glossary view*
- *journal add entry*
- *journal update entry*
- *journal view*
- *quiz attempt*
- *quiz review*
- *resource view*
- *wiki comments*
- *wiki edit*
- *wiki history*
- *wiki view*

Verificou-se a necessidade de reduzir a dimensão do problema em componentes. Utilizou-se a técnica de análise dos principais componentes. Segundo (JOHNSON, RICHARD ARNOLD, 1992), a análise de componentes principais pode ser uma etapa intermediária de cálculo para uma análise posterior, como: análise de agrupamentos, classificação, discriminação, redes neurais, entre outras. Assim, a aplicação de componentes principais sob um conjunto de dados, poderá ser extremamente útil, a fim de gerar soluções para uma classe de problemas em mineração de dados que exige a redução de dimensionalidade, como uma etapa de pré-processamento.

A Figura 21 apresenta o número de interações agrupadas por polo após a coleta das informações. São exibidos 24 atributos e seus respectivos quantitativos.

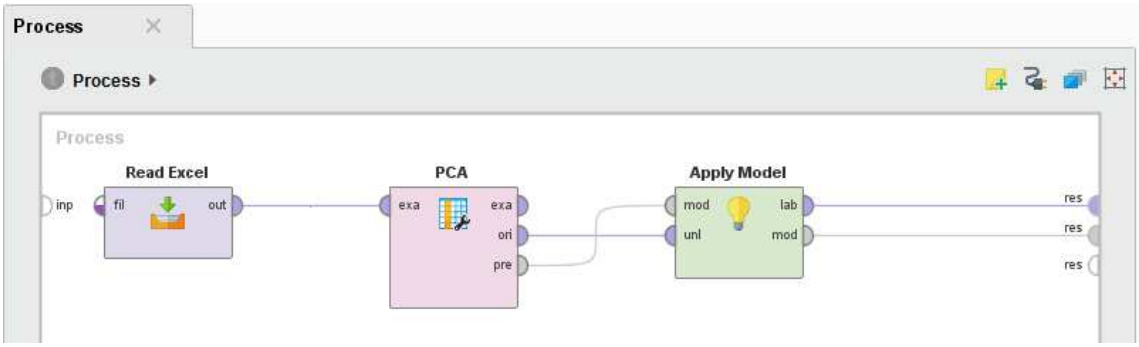
Figura 21 Número de acessos por polos

Polo	assignment upload	assignment view	blog view	course recent	folder view	forum add discussion
Boa Vista	107	1047	1	368	748	164
Caracaraí	88	880	3	207	661	152
Manaus	133	1494	34	286	1186	155
Tefé	108	909	2	205	603	143

Fonte: Próprio autor

Havia a necessidade de reduzir essa dimensão, assim estes dados foram submetidos ao processo de análise dos principais componentes conforme apresentado na Figura 22.

Figura 22 Processo PCA



Fonte: Próprio autor

Após o processo, dois componentes foram apontados conforme apresentado na Figura 23.

Figura 23 N° de componentes apontados pelo PCA

ExampleSet (4 examples, 0 special attributes, 2 regular attributes)

Row No.	pc_1	pc_2
1	142.356	-770.859
2	-1041.777	-0.228
3	3185.279	342.176
4	-2285.858	428.910

Fonte: Próprio autor

O processo considera o desvio padrão, proporções de variância e a variância acumulada para a indicação dos principais componentes, conforme apresentado na Figura 24.

Figura 24 Lista dos Componentes

Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC 1	2343.552	0.942	0.942
PC 2	546.282	0.051	0.994
PC 3	192.188	0.006	1.000
PC 4	0.000	0.000	1.000
PC 5	0.000	0.000	1.000
PC 6	0.000	0.000	1.000
PC 7	0.000	0.000	1.000
PC 8	0.000	0.000	1.000
PC 9	0.000	0.000	1.000
PC 10	0.000	0.000	1.000
PC 11	0.000	0.000	1.000
PC 12	0.000	0.000	1.000
PC 13	0.000	0.000	1.000
PC 14	0.000	0.000	1.000
PC 15	?	-0.000	1.000

Fonte: Próprio autor

Com o auxílio do PCA foram selecionados portanto 3 atributos: o "*forum view discussion*", *resource view*" e "*forum view forum*". Estes atributos foram utilizados como filtro para o cálculo do desvio padrão no conjunto de treinamento.

Figura 25 Componentes indicados pelo PCA

Attribute	PC 1 ↓	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10
forum view discussion	0.601	0.405	0.608	-0.229	0.158	0.001	0.073	-0.013	0.015	0.011
resource view	0.597	-0.781	-0.009	0.012	-0.131	0.023	-0.097	0.030	-0.008	-0.001
forum view forum	0.458	0.296	-0.559	0.470	0.258	0.101	0.146	-0.094	-0.101	0.010
glossary view	0.131	0.234	0.013	0.358	-0.608	-0.196	-0.001	0.204	0.177	-0.095
assignment view	0.116	0.105	-0.296	-0.183	0.187	0.234	-0.206	0.339	0.336	0.025
journal view	0.114	0.180	-0.278	-0.517	-0.348	0.439	-0.243	-0.327	-0.015	-0.114
folder view	0.110	0.098	-0.086	-0.016	-0.210	-0.539	-0.213	-0.489	0.051	0.038
quiz review	0.109	0.097	-0.173	-0.342	-0.131	-0.265	0.294	0.441	-0.290	-0.017
journal update entry	0.041	0.061	-0.073	-0.048	-0.176	-0.196	-0.261	0.248	0.227	-0.015
forum add post	0.032	-0.026	-0.186	-0.206	0.161	-0.307	0.150	-0.261	0.244	-0.156

Fonte: Próprio autor

Foram considerados para o treinamento 979 registros de médias e desvios padrões de interações (número de acessos). Esse conjunto foi submetido ao processo de validação supramencionado na seção 5.1 deste trabalho. Desta vez utilizando os algoritmos descritos na tabela abaixo. Para este volume de dado destacou-se o J48 com acurácia de 84.05%, precisão na classificação de reprovados de 85,52% e o *class Recall* de 50,23%, conforme apresentado na Tabela 5.

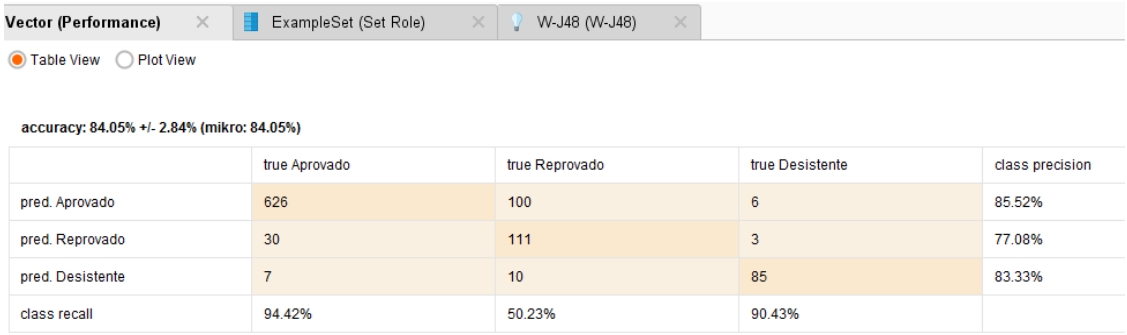
Tabela 5 Tabela de Classificadores

Algoritmo	Acurácia	Class Precision			Class Recall		
		Pred Aprovado	Pred Reprovado	Pred Desistente	True Aprovado	True Reprovado	True Desistente
Decicion Tree – Gain Ratio	84,47%	84,43%	81,10%	89,36%	96,39%	46,61%	89,36%
Decicion Tree – Information Gain	82,11%	84,55%	70,78%	81,82%	92,46%	49,32%	86,17%
BFTree	83,23%	84,57%	74,13%	86,46%	94,27%	47,96%	88,30%
J48	**84,05%**	**85,52%**	**77,08%**	**83,33%**	**94,42%**	**50,23%**	**90,43%**
Random Forest	80,07%	80,42%	80%	76,32%	97,89%	34,39%	61,7%
Random Tree	74,85%	83,38%	48,86%	75,26%	83,26%	48,42%	77,66%

Fonte: Próprio autor

Na Figura 26 observa-se a matriz de confusão que representa nas linhas as predições de Aprovação, Reprovação e Desistentes. Nas colunas a indicação dos valores Verdadeiros Aprovados, Verdadeiros Reprovados e Verdadeiros Desistentes. Na diagonal é possível observar um destaque nos números apresentados que fornecem o índice *recall* na parte inferior da figura. À direita se vê a classificação de precisão. Na parte superior observa-se a acurácia do classificador.

Figura 26 Matriz de Confusão



	true Aprovado	true Reprovado	true Desistente	class precision
pred. Aprovado	626	100	6	85.52%
pred. Reprovado	30	111	3	77.08%
pred. Desistente	7	10	85	83.33%
class recall	94.42%	50.23%	90.43%	

Fonte: Próprio autor

A regra encontrada para uso no MD-PREAD está descrita abaixo:

W-J48

J48 pruned tree

Média <= 0

| Periodo = 2014/2: Reprovado (83.11/12.0)

| Periodo = 2015/1

| | Polo = Boa Vista

| | | DesvioInt <= 13: Desistente (27.16)

| | | DesvioInt > 13

| | | | DesvioInt <= 56.824291: Aprovado (6.79/3.79)

| | | | DesvioInt > 56.824291: Reprovado (9.05/1.05)

| | Polo = Caracaraí

| | | DesvioInt <= 20.008332

| | | | DesvioInt <= 15.011107: Desistente (14.97)

| | | | DesvioInt > 15.011107: Reprovado (5.76/0.76)

| | | DesvioInt > 20.008332: Desistente (17.27/2.0)

| | Polo = Manaus: Desistente (35.45/8.45)

| | Polo = Tefé

| | | DesvioInt <= 44.105933: Reprovado (15.22/1.0)

| | | DesvioInt > 44.105933: Aprovado (3.0)

Média > 0
| Média <= 6.67
| | Disciplina = Filosofia da Educação Brasileira
| | | DesvioInt <= 16.041613: Reprovado (7.13/1.0)
| | | DesvioInt > 16.041613: Aprovado (24.0/7.0)
| | Disciplina = Filosofia da Educação Medieval: Aprovado (1.0)
| | Disciplina = Filosofia na Antiguidade
| | | DesvioInt <= 18.248288
| | | | Média <= 5.333333: Reprovado (7.0/1.0)
| | | | Média > 5.333333: Aprovado (4.0/1.0)
| | | DesvioInt > 18.248288: Aprovado (18.0/4.0)
| | Disciplina = INTRODUÇÃO À MODALIDADE DE EDUCAÇÃO A DISTÂNCIA
| | | Polo = Boa Vista: Reprovado (4.0/1.0)
| | | Polo = Caracaraí
| | | | Média <= 4: Aprovado (4.0/1.0)
| | | | Média > 4: Reprovado (2.0)
| | | Polo = Manaus: Aprovado (0.0)
| | | Polo = Tefé: Aprovado (6.0/1.0)
| | Disciplina = Metodologia Científica: Aprovado (23.0/8.0)
| | Disciplina = Epistemologia do Ensino de Filosofia da Educação: Aprovado (29.0/2.0)
| | Disciplina = Filosofia da Educação Contemporânea: Aprovado (14.0/1.0)
| | Disciplina = Métodos e Técnicas de Pesquisa em Educação
| | | Média <= 4.74: Reprovado (8.18)
| | | Média > 4.74
| | | | DesvioInt <= 28.676355: Reprovado (10.5/2.0)
| | | | DesvioInt > 28.676355: Aprovado (12.0/1.0)
| | Disciplina = Tópicos Especiais em Antropologia da Educação: Aprovado (22.0)
| | Disciplina = Tópicos Especiais em Ética: Aprovado (26.0/6.0)
| Média > 6.67: Aprovado (538.4/52.4)

Number of Leaves : 30

6 ASPECTOS DE AVALIAÇÃO

Nesta seção será apresentada a instituição que serviu de cenário para a realização do experimento, a abrangência dos níveis de ensino alcançados pela modalidade de Educação a Distância, a disciplina escolhida para a avaliação do modelo e o acompanhamento dos resultados alcançados a cada semana.

6.1 Sobre a Instituição

O Instituto Federal de Educação Ciência e Tecnologia do Amazonas (IFAM) oferece cursos de graduação, pós-graduação e especiais modalidade de Educação a distância. Na Universidade Aberta do Brasil, são mais de 920 alunos espalhados por diversos polos.

A seguir apresentasse a lista dos cursos gerenciados pelo IFAM relacionados ao programa UAB:

CURSOS ESPECIAIS – A DISTÂNCIA:

- Formação Pedagógica de Docentes para Educação Básica.

GRADUAÇÃO – A DISTÂNCIA (1ª LICENCIATURA):

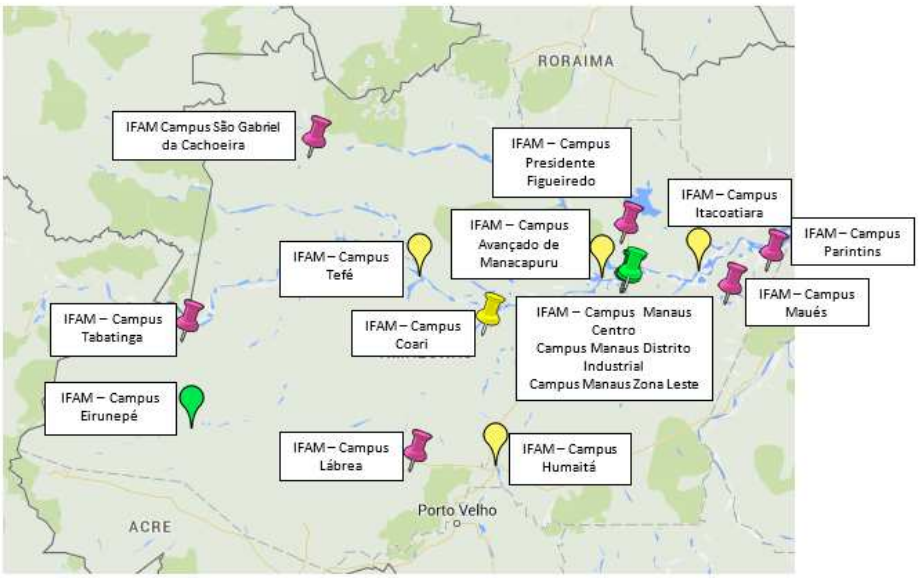
Licenciatura em Física.

PÓS-GRADUAÇÃO LATO SENSU:

- Especialização em Educação do Campo;
- Especialização em Educação Musical;
- **Especialização em Filosofia da Educação;**
- Especialização em Gestão Pública;
- Especialização em História e Cultura Africana e Afro-brasileira;
- Especialização em Informática na Educação.

A Figura 27 apresenta a abrangência territorial alcançado pelo Instituto Federal de Educação, Ciência e Tecnologia do Amazonas.

Figura 27 Campus do IFAM



Fonte: Próprio autor

O curso escolhido foi o de Filosofia da Educação EAD/UAB – Integral que teve seu início no segundo semestre de 2014, este curso é composto por 14 disciplinas conforme apresentado na Figura 28. Os números de matrículas iniciais foram de 30 alunos no polo de Manaus, 35 alunos no polo de Boa vista, 30 alunos no polo de Caracaraí e 30 alunos no polo de Tefé, totalizando 125 alunos conforme apresentado nas Figuras 29 e 30.

O motivo de sua escolha foi em razão de que houve uma oferta da disciplina **Língua Brasileira de Sinais** no período de 05 a 30 de outubro, momento oportuno paras se observar a turma de alunos conforme apresentado na Figura 31.

Figura 28 Matriz Curricular do Curso de Filosofia da Educação

CAMPUS MANAUS-CENTRO
COORDENAÇÃO DE CONTROLE ACADÊMICO
Matrizes Curriculares

Matriz: 8635 - CMC - FILOSOFIA DA EDUCAÇÃO EAD/UAB - INTEGRAL (2014/2)

Curso: 3133 - CMC - FILOSOFIA DA EDUCAÇÃO EAD/UAB - INTEGRAL

Nível	Periodicidade	Regime	Situação	Per. Letivo Inicial	C.H. Disciplinas
Pós-Graduação	Semestre	Seriado	Matriz em Vigor	2014/2	450

Per.	Componentes Curriculares								Carga Horária	Co-Requisitos	Pré-requisitos
	Código	Descrição	Núcleo	OPT	Hab.	Cred.	Cred. Nesc.				
1	CEAD.192	INTRODUÇÃO À MODALIDADE EAD	COM	N	2688	30	0		30		
1	CEAD.193	METODOLOGIA CIENTÍFICA	COM	N	2688	30	0		30		
1	CEAD.313	FILOSOFIA DA EDUCAÇÃO NA ANTIGUIDADE	COM	N	2688	2	0		30		
1	CEAD.314	FILOSOFIA DA EDUCAÇÃO MEDIEVAL	COM	N	2688	2	0		30		
1	CEAD.323	FILOSOFIA DA EDUCAÇÃO BRASILEIRA	COM	N	2688	3	0		30		
2	CEAD.301	TÓPICOS DE ANTROPOLGIA DA EDUCAÇÃO	COM	N	2688	3	0		30		
2	CEAD.316	MÉTODOS E TÉCNICAS DE PESQUISA NA EDUCAÇÃO BÁSICA E NO	COM	N	2688	3	0		30		
2	CEAD.317	TÓPICOS ESPECIAIS EM ÉTICA	COM	N	2688	2	0		30		
2	CEAD.324	FILOSOFIA DA EDUCAÇÃO CONTEMPORÂNEA	COM	N	2688	3	0		30		
2	CEAD.325	EPISTEMOLOGIA DO ENSINO DE FILOSOFIA DA EDUCAÇÃO	COM	N	2688	7	0		30		
	CEAD.306	ELABORAÇÃO DE TCC/MONOGRAFIA, PROJETO E ARTIGO CIENTÍFICO	COM	N	2688	30	0		30		
3	CEAD.307	TRABALHO DE CONCLUSÃO DE CURSO	COM	N	2688	60	0		60		
3	CEAD.308	LÍNGUA BRASILEIRA DE SINAIS	COM	N	2688	2	0		30		
3	CEAD.312	TÓPICOS ESPECIAIS EM FILOSOFIA POLÍTICA	COM	N	2688	2	0		30		

Código	Sigla	Habilitação	Básica	Créd. Obrig.	Carga Horária				
					Estágio	Optativa	Compl.	Proj. Fin.	Min. Créd.
2688		Disciplinas Básicas	Sim						

Fonte: Sistema de Gestão Acadêmica do IFAM 2015

Figura 29 Turmas de Manaus, Boa Vista e Caracarai

CAMPUS MANAUS-CENTRO
COORDENAÇÃO DE CONTROLE ACADÊMICO
Estatística de Alunos por Situação

Filtros Utilizados para Gerar este Relatório:

Estrutura de Curso: 2013 - Educação a Distância
Curso: CMC - FILOSOFIA DA EDUCAÇÃO EAD/ UAB - INTEGRAL
Ano Letivo: 2014
Per. Letivo: 2

Turma	Aprov.	Aprovado	Rep. Falta	Reprovado	Total
20142.3133. MAO1	7	18	5	0	30
20142.3133.BV1	6	12	15	2	35
20142.3133.CAR1	10	6	13	1	30
Totais:	23	36	33	3	95

Fonte: Sistema de Gestão Acadêmica do IFAM 2015

Figura 30 Turma de Tefé

IFAM CAMPUS TEFE
COORDENAÇÃO DE REGISTRO ACADÊMICO
Estatística de Alunos por Situação

Filtros Utilizados para Gerar este Relatório:

Estrutura de Curso: 2013 - Educação a Distância
Curso: CTEFE - FILOSOFIA DA EDUCAÇÃO EAD/UAB - INTEGRAL
Ano Letivo: 2014
Per. Letivo: 2

Turma	Aprov.	Aprovado	Rep. Falta	Reprovado	Total
20142.3257.1	7	16	6	1	30
Totais:	7	16	6	1	30

Fonte: Sistema de Gestão Acadêmica do IFAM 2015

Figura 31 Apresentação da Disciplina Libras no AVA

Fonte: Portal EaD do IFAM 2015

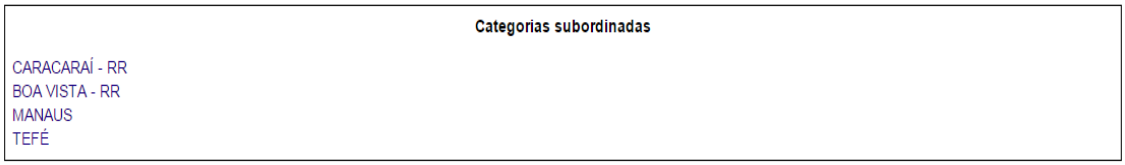
O Ambiente Virtual de Aprendizagem do IFAM tem como plataforma o Moodle, nele os Cursos são Categorias como se pode observar na Figura 32.

Figura 32 Lista de Categorias

Fonte: Portal EaD do IFAM 2015

Os polos também são identificados como categorias e subordinados as categorias dos cursos, conforme apresentado na Figura 33.

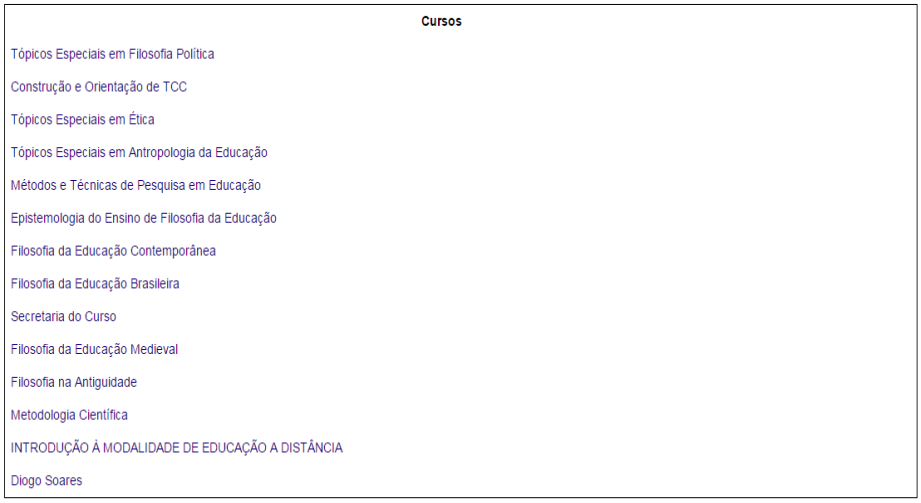
Figura 33 Categorias subordinadas



Fonte: Portal EaD do IFAM 2015

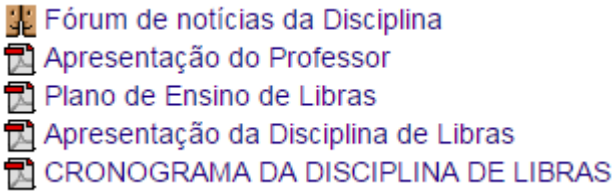
As disciplinas são tratadas no ambiente como cursos conforme a Figura 34, o professor pode disponibilizar o material de orientação para o início das atividades conforme apresentado na Figura 35, também pode apresentar a metodologia conforme apresentado na Figura 36 e as avaliações que ocorrerão na disciplina conforme apresentado na Figura 37.

Figura 34 Cursos



Fonte: Portal EaD do IFAM 2015

Figura 35 Apresentação do Curso/Disciplina



Fonte: Portal EaD do IFAM 2015

Figura 36 Metodologias e Recursos de Ensino

Metodologias, Técnicas e Recursos de Ensino.
O programa será desenvolvido com base em:
- Leitura Analítica e Trabalhos de Pesquisa;
- Bibliotecas Digitais Online;
- Hiperlinks com vídeos complementares;
- Salas de Aulas Virtuais;
- Fóruns e *Chats*;
- Vídeo Conferência;
- Seminário;
- Atividade Prática.

Fonte: Plano de Ensino do Curso de Pós-Graduação em Filosofia da Educação IFAM 2015

Figura 37 Avaliação da aprendizagem

Avaliação da Aprendizagem
As atividades Avaliativas serão:

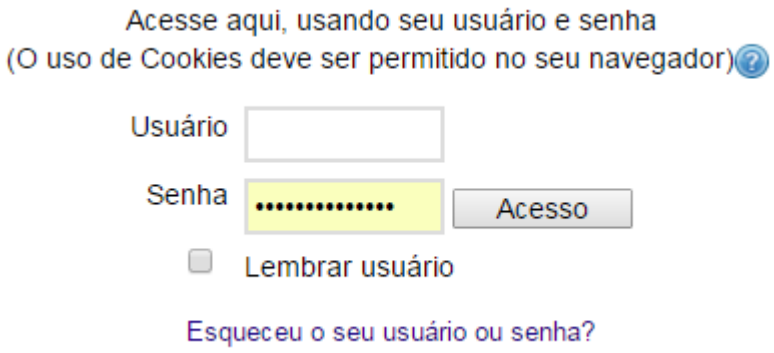
- Fórum
- Questionário
- Envio de Arquivo único Individual
- Envio de Arquivo único em Grupo
- Apresentação de seminário

Fonte: Plano de Ensino do Curso de Pós-Graduação em Filosofia da Educação IFAM 2015

6.2 Extração dos dados

Os dados foram extraídos diretamente do ambiente de produção através do link http://ead.ifam.edu.br/uab/login/index.php conforme apresentado na Figura 38.

Figura 38 Tela de acesso ao ambiente



Fonte: Portal EaD do IFAM 2015

Foi acessada a disciplina que no ambiente é tratada como curso e descrita como LIBRAS, cada polo está assim identificado no ambiente e possui uma pasta de relatórios que contém os links que leva aos logs dos usuários conforme apresentado na Figura 39.

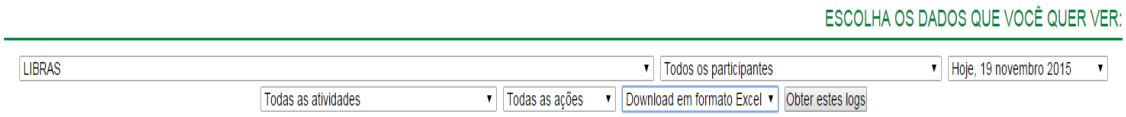
Figura 39 localização da pasta de Logs no ambiente

Fonte: Portal EaD do IFAM 2015

Ao clicar no link dos Logs o sistema apresenta a possibilidade de filtros para a exportação, os filtros possibilitam ver os logs de um aprendiz ou todos, pode-se optar por ver os logs de um dia específico ou todos os dias, de forma análoga pode-se filtrar logs de uma

atividade ou todas, as opções de saída são: mostrar na página, download em formato Excel, download em formato ODS e download em formato *Text*, conforme apresentado na Figura 40. Para este experimento optou-se pela exportação em planilha do Excel.

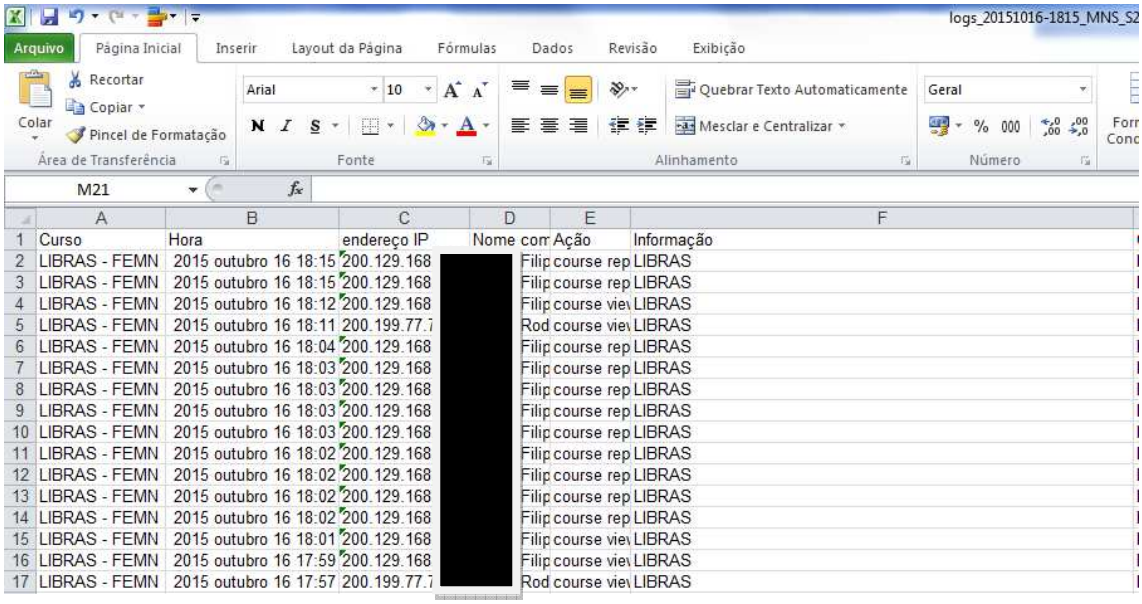
Figura 40 Tela de exportação do ambiente virtual de aprendizagem



Fonte: Portal EaD do IFAM 2015

Ao se clicar em obter *logs* o arquivo é gerado contendo o nome da disciplina/curso, a data e hora do acesso, o endereço IP do *hardware*, o nome do usuário e uma informação. A extração foi feita a cada semana para todas as 04 turmas envolvidas no experimento conforme apresentado na Figura 41.

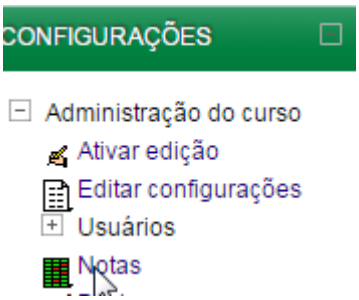
Figura 41 Arquivo de logs de atividades



Fonte: Portal EaD do IFAM 2015

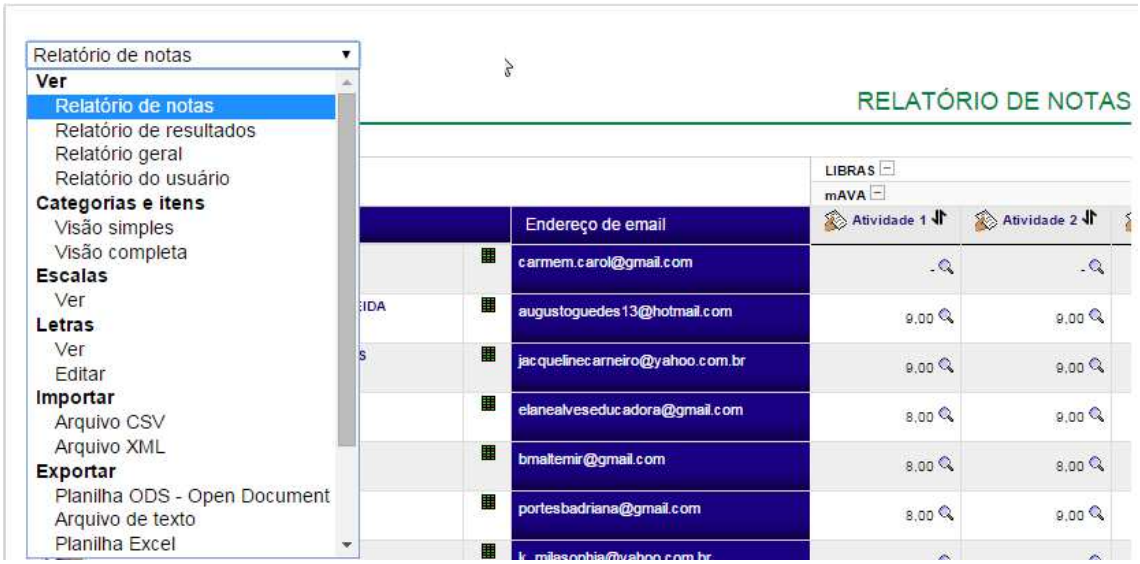
As notas ficam na guia configurações do curso selecionado e o link tem a descrição “Notas” conforme apresentado nas Figura 42 e 43. Ao se clicar em notas o sistema apresenta as opções: Ver, Categorias e itens, Escalas, Letras, Importar e Exportar. Neste experimento optou-se por utilizar a exportação planilha do Excel.

Figura 42 Localização das notas no ambiente



Fonte: Portal EaD do IFAM 2015

Figura 43 Tela de exportação de notas das atividades



Fonte: Portal EaD do IFAM 2015

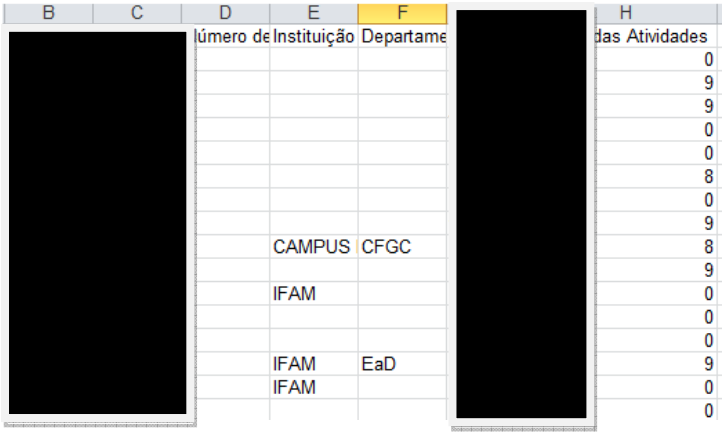
Ao clicar em Planilha do Excel o sistema apresenta as configurações possíveis para a exportação das notas, este curso foi configurado com três atividades, duas chamadas de avaliações presenciais e categorias e total do curso que são campos calculados que guardam as médias parciais e finais respectivamente. Neste experimento foram exportadas as notas das atividades 1, 2 e 3 conforme apresentado na Figura 44, uma a cada semana, considerando a seguinte regra: se houver apenas uma nota ela é a média, duas ou três notas a média é o somatório das notas dividido pelo número de atividades.

Figura 44 Configuração da exportação de notas

Fonte: Portal EaD do IFAM 2015

O arquivo gerado contém o nome e sobrenome do aprendiz, número de identificação, instituição, departamento, endereço de e-mail e média conforme apresentado na Figura 45, o processo de exportação ocorreu semanalmente com todas as turmas.

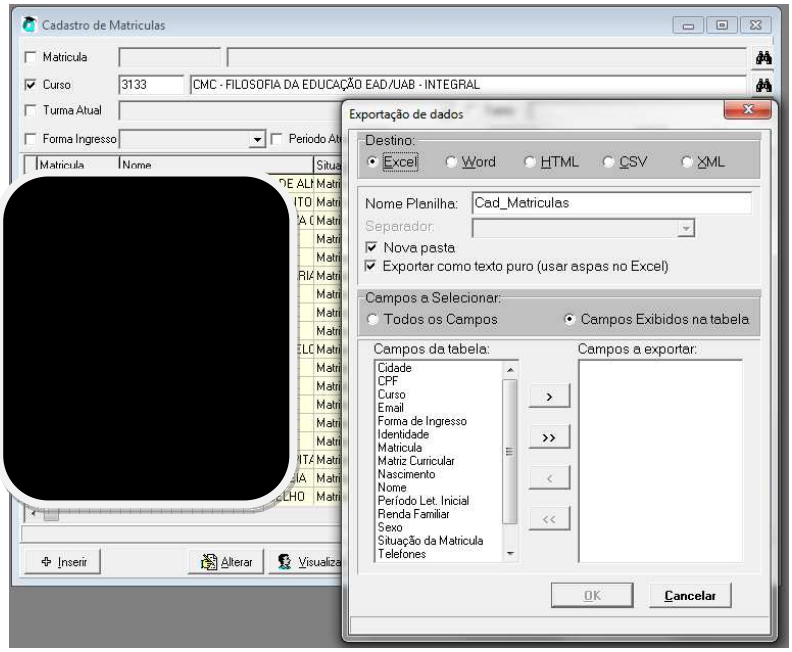
Figura 45 Arquivo de média das avaliações



Fonte: Portal EaD do IFAM 2015

Os dados de matrículas dos alunos foram extraídos do Sistema de Gestão Acadêmica do IFAM que de forma análoga ao ambiente virtual possibilita a exportação de dados em vários formatos dentre eles a planilha do Excel, conforme apresentado na Figura 46.

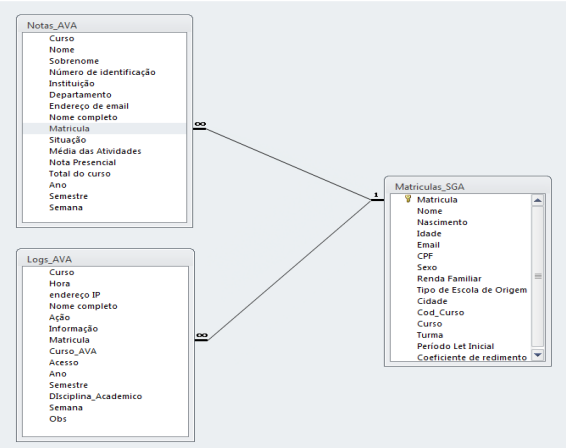
Figura 46 Exportação dos dados de matrículas dos alunos



Fonte: Sistema de Gestão Acadêmica do IFAM 2015

Optou-se por realizar a implementação do banco de dados sobre o programa *ACCES*, da *Microsoft* (CORPORATION, 1994), em função de sua praticidade. A Figura 47 apresenta as tabelas utilizadas no experimento. Os dados foram importados para o Sistema de Gerenciamento de Banco de Dados Relacional e utilizou-se de consultas de atualização para a obtenção dos dados que compõem o arquivo de teste utilizado a cada semana para a predição semanal conforme apresentado na Figura 48.

Figura 47 Banco de dados do MD-PREAD



Fonte: Próprio autor

Figura 48 Arquivo de teste utilizado para a predição

Matricula	Média_Atividades	DesvDeInterações	Situação
[redacted]	9	89	
	8	0	
	9	0	
	8	159	
	9	0	
	9	86	
	9	0	
	0	0	
	0	0	
	8	149	
	9	78	
	8	0	
	8	0	
	0	0	
	0	0	
	8	0	
	0	0	
	9	159	
	9	0	
	9	0	
	8	50	

Fonte: Próprio autor

A Figura 49 apresenta o arquivo de saída contendo as médias, desvio padrão das interações a situação não definida, a identificação do aprendiz os percentuais de predição de aprovação e reprovação e a situação prediction do grupo de risco de reprovar.

Figura 49 Recorte da predição da 4ª Semana

A	B	C	D	E	F	G
Média_At	DesvDeIn	Situação	Matricula	confidence(APROVADO)	confidence(REPROVADO)	prediction(Situação)
,0	,0		[redacted]0031	43%	57%	REPROVADO
,0	,0		0139	43%	57%	REPROVADO
,0	,0		0147	43%	57%	REPROVADO
,0	,0		0155	43%	57%	REPROVADO
,0	,0		0228	43%	57%	REPROVADO
,0	,0		0236	43%	57%	REPROVADO
8,0	32,0		0260	43%	57%	REPROVADO
,0	,0		0279	43%	57%	REPROVADO
,0	,0		0350	43%	57%	REPROVADO
,0	,0		0368	43%	57%	REPROVADO
,0	,0		0376	43%	57%	REPROVADO
,0	23,0		0406	43%	57%	REPROVADO
,0	,0		0414	43%	57%	REPROVADO
,0	,0		0430	43%	57%	REPROVADO
,0	,0		0449	43%	57%	REPROVADO
,0	,0		0503	43%	57%	REPROVADO
,0	,0		0511	43%	57%	REPROVADO
,0	,0		0520	43%	57%	REPROVADO
,0	,0		0546	43%	57%	REPROVADO
,0	,0		0597	43%	57%	REPROVADO
8,0	43,0		0627	43%	57%	REPROVADO
,0	,0		0635	43%	57%	REPROVADO
,0	,0		0643	43%	57%	REPROVADO
,0	19,0		0708	43%	57%	REPROVADO
,0	,0		0716	43%	57%	REPROVADO
0	2,0		0740	43%	57%	REPROVADO

Fonte: Próprio autor

O Regimento do IFAM contém informações acerca de como serão calculadas as avaliações no ambiente virtual e define a fórmula apresentada na Figura 50 orientando a como se chegar a Nota Final do aluno.

No ambiente virtual de aprendizagem o professor tem a liberdade de criar uma ou mais atividades, motivo pelo qual estas informações são variadas no Moodle, para se obterem as notas das avaliações das atividades, serão exportadas somente as notas que compõem a média das atividades.

Figura 50 Fórmula de Cálculo da média Semestral

$$MS = \frac{\frac{\Sigma_{i=1}^{n} AVA_i}{n} + 2.NAP}{3} \geq 6{,}0$$

Onde:
MS = Média Semestral (por disciplina).
AVA = Nota das Atividades do AVA.
NAP = Nota da Avaliação Presencial (Peso 2).

Fonte: Organização didática do IFAM (2012)

6.3 Avaliação do grupo de Teste

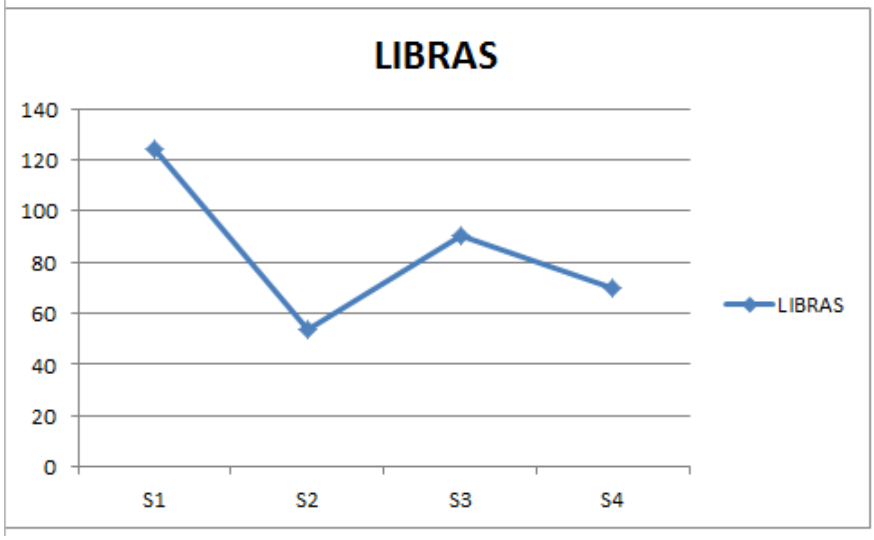
A seguir são apresentados os resultados alcançados a cada semana com o grupo de teste, vale ressaltar que nesta fase não se trata de dados de treinamento os quais se conhece o resultado da situação final, mas de um curso em andamento onde semanalmente foi executado o processo de predição e selecionados os aprendizes com situação de predição Reprovado, a essa lista denominou-se grupo de risco.

A cada semana o grupo de risco com predição de reprovação era selecionado pelo algoritmo de árvore de decisão utilizado no experimento. Observou-se de acordo com a Figura 51 que a primeira predição feita na primeira semana, 100% dos aprendizes apareciam no grupo de risco de reprovar com um número de 124, o que se justifica por não ter havido nenhuma atividade realizada pelos aprendizes, na segunda predição tem-se um grupo menor de 54 aprendizes que representa 43% do grupo inicial, na terceira predição tem-se 90 aprendizes que representa um aumento de 66% em relação a predição anterior e na quarta predição 70 aprendizes, uma redução de 22% em relação à predição anterior. Conforme as atividades são ofertadas para os aprendizes nota-se a variação da curva do comportamento do grupo de risco ao longo das semanas. A predição aponta para um total de 56,45% da turma que é composta por quatro polos, que são municípios onde tutores presenciais podem dar assistência aos aprendizes, que são Caracaraí, Boa Vista, Tefé e Manaus.

Constatou-se um grande número de desistências, com 46 aprendizes constantes da sala que representam 37% do total de alunos e que reflete diretamente na predição, pois estes aprendizes naturalmente aparecem em todas as predições. As desistências não ocorreram em função das atividades no ambiente virtual, o que pode-se concluir é que a secretaria do Curso considerou todos os alunos que já estavam no ambiente para dar celeridade ao processo de matrícula, mas de fato nem todos estavam na sala virtual. A relação com a predição está no fato de que os alunos que não estão participando da sala virtual aparecem em todas as listagens de predição de reprovação, dando um falso resultado de que existem muitos aprendizes no grupo de risco.

Figura 51 Número de Predições a cada semana

	Predição			
Disciplina	S1	S2	S3	S4
LIBRAS	124	54	90	70

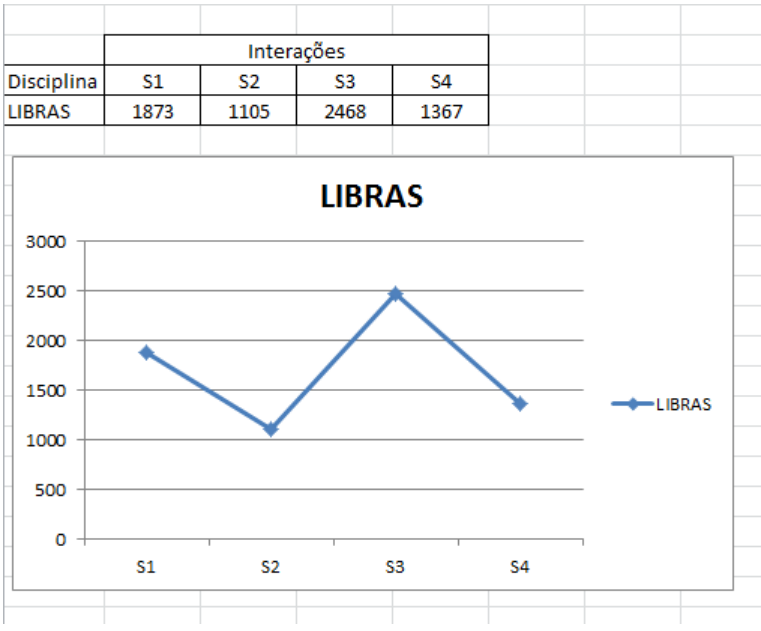


Fonte: Próprio autor

Observando as interações é possível praticar a *Learning Analytic*, a análise dos comportamentos constatados pelas interações. Conforme apresentado na Figura 52, observa-se um acesso inicial de 1873 acessos e tem uma queda de 41% em relação a primeira semana, a terceira semana se destacou com um aumento de 123% e na 4ª semana observou-se uma queda de 44,61%. Aqui é possível verificar o grau de interesse semanal, mais adiante se verá como é esse comportamento em relação a cada atividade.

Figura 52 nº interações dos aprendizes

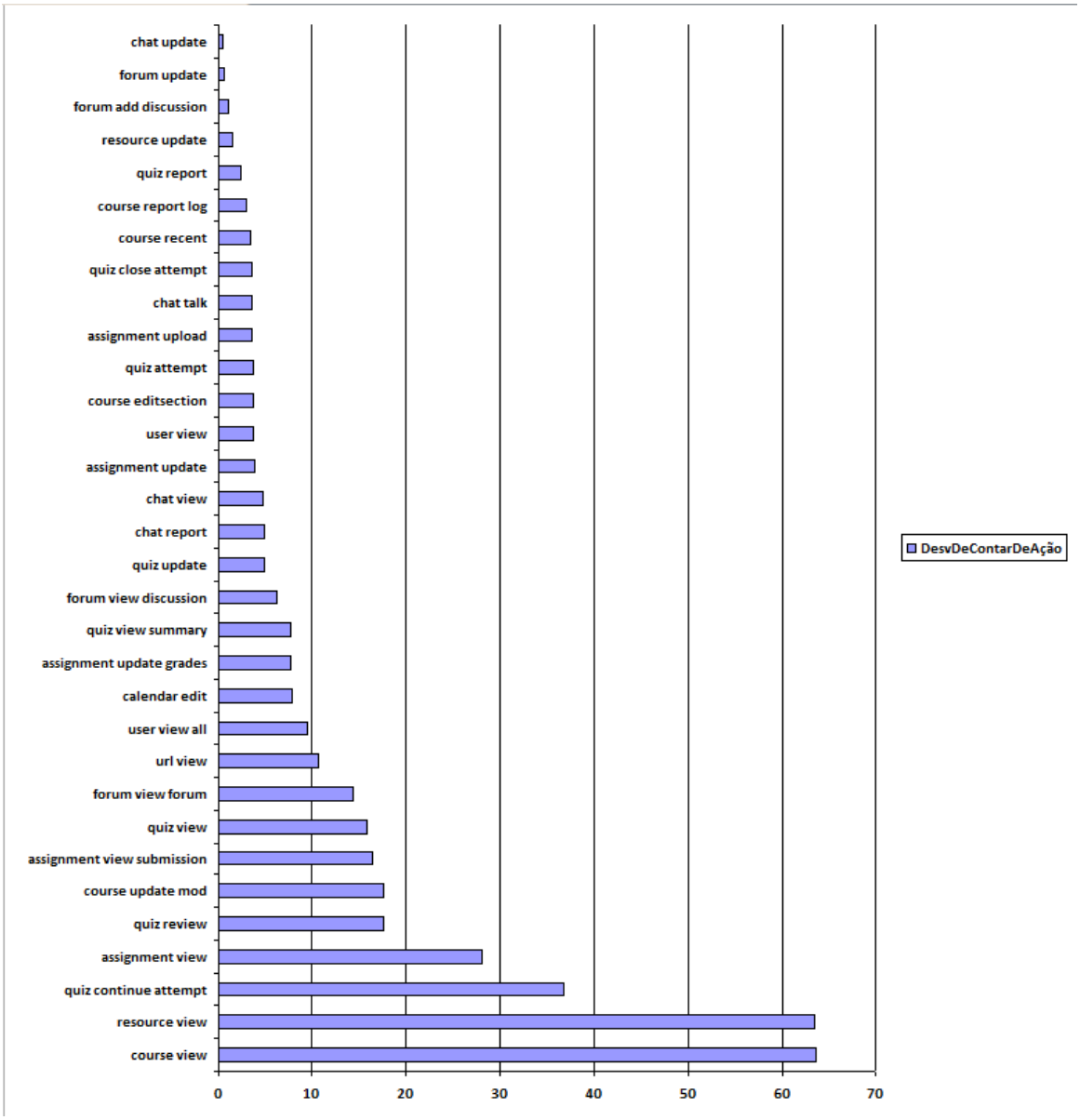
	Interações			
Disciplina	S1	S2	S3	S4
LIBRAS	1873	1105	2468	1367



Fonte: Próprio autor

A Figura 53 apresenta o desvio padrão das interações feitas no ambiente pelos aprendizes e ordenadas pelo grau de interesse em cada atividade ofertada pelo professor, com esta informação é possível dar opções aos gestores educacionais e tutores do ambiente para avaliar o grau de interesse nas atividades ofertadas e pensar em estratégias para valorizar aquelas de maior interesse ou estimular aquelas de menor interesse adequando assim a metodologia à realidade que se apresenta.

Figura 53 Desvio padrão das interações por atividade

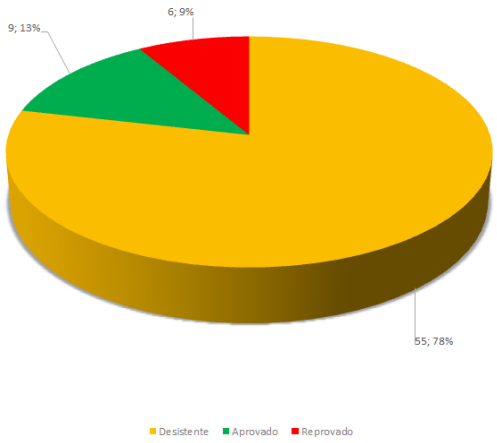


Fonte: Próprio autor

A Figura 54 apresenta os resultados das situações após o final da disciplina, não se trata mais de predição e sim de resultados reais. O foco está no resultado real do grupo de risco da última semana, observou-se um número de desistentes de 55 aprendizes representando 78% do grupo de risco, 09 aprovações representando 13% e 06 reprovações representando 9% do grupo de risco.

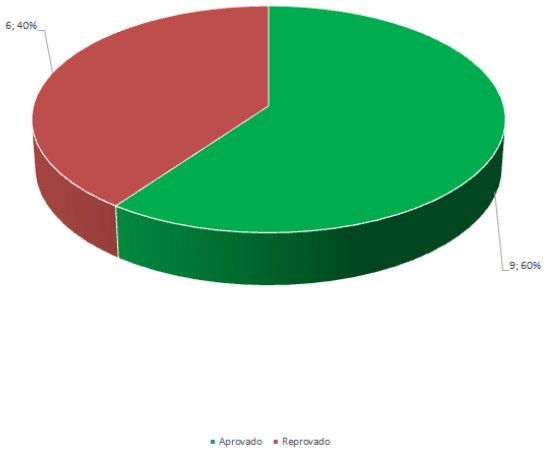
Para que o aprendiz possa ser considerado um reprovado é necessário que o mesmo seja avaliado e participe de alguma forma com o ambiente, o que não se pode configurar com alunos desistentes, assim para termos uma visão mais real dos resultados desta dissertação retirou-se do grupo de risco os desistentes e apresentou-se a Figura 55 que apresenta uma taxa de 40% de aprendizes reprovados e 60% de aprendizes aprovados que estavam contidos no grupo de risco de reprovar, o que poderia levar a deduzir que a taxa de acerto foi de 40%. No entanto não se pode descartar que este processo foi desenvolvido em paralelo com o Protótipo de um sistema de recomendação, e que os arquivos foram enviados para que tratamentos fossem adotados a fim de reduzir o número de reprovados. A acurácia da predição foi de 55% e a confiabilidade 0,2.

Figura 54 Resultado final do grupo de risco



Fonte: Próprio autor

Figura 55 Resultado final do grupo de risco desconsiderando as desistências



Fonte: Próprio autor

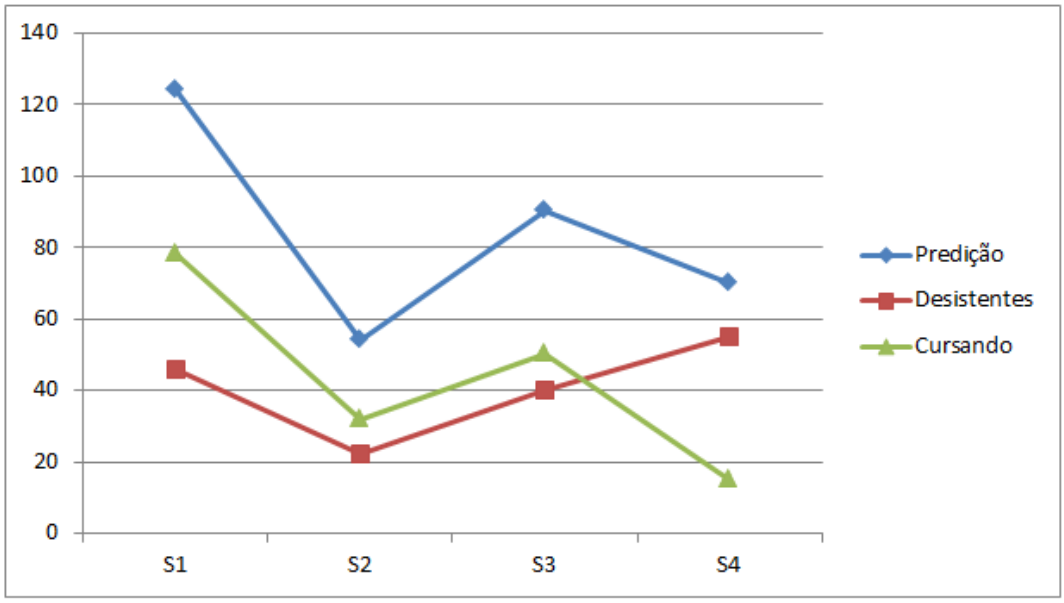
Na Figura 56 observa-se que na primeira semana 100% dos aprendizes aparecem na lista de predição, mas já observa-se um percentual de 37% de aprendizes desistentes, chegou-

se a essa conclusão após a análise realizada ao final da disciplina. Estes aprendizes não interagiram com o ambiente virtual restando apenas 78 alunos frequentando efetivamente. Na segunda semana o grupo de risco já passa a ser menor uma vez que são filtrados apenas aqueles com predição de reprovação constantes do grupo de risco.

O gráfico apresenta o comportamento das linhas de aprendizes com situação de “desistentes” e “cursando” mostrando um comportamento de similaridade até a terceira semana, quando então ocorre um aumento do primeiro e uma diminuição do segundo respectivamente. A inclusão de aprendizes em excesso na sala virtual gerou um resultado falso positivo.

Figura 56 Evolução semanal de comportamento

Semanas	Predição	Desistentes	Cursando
S1	124	46	78
S2	54	22	32
S3	90	40	50
S4	70	55	15



Fonte: Próprio autor

6.4 Avaliação do Classificador J48 no segundo treinamento

Para avaliar os resultados do treinamento do classificador encontrado no segundo treinamento descrito na seção 5.2, foram comparados os índices deste, com os índices dos classificadores dos trabalhos relacionados conforme Tabela 6.

Com relação ao primeiro trabalho o MD-PREAD apresentou um índice de precisão superior, os demais trabalhos não utilizaram este índice para mensurar sua eficiência.

Com relação ao segundo trabalho não foi possível chegar a conclusões em função de não ter sido apresentado o índice de eficiência do classificador.

Com relação ao terceiro trabalho o MD-PREAD teve um índice de normalidade inferior.

Com relação ao quarto trabalho o MD-PREAD teve uma acurácia superior.

Com relação ao quinto trabalho não foi possível chegar a conclusões em função de não ter sido apresentado o índice de eficiência do classificador.

Com relação ao sexto trabalho o MD-PREAD apresentou um índice *Kappa* superior.

Tabela 6 Comparação dos índices com Trabalhos Relacionados

Trabalhos relacionados	Precisão	Acurácia	Recall	Kappa	Shapiro-Wilk Test (Forum)
Predição de Reprovação de Alunos de Educação a Distância Utilizando Contagem de Interações. Redes Bayesianas Precisão(1)	66%	-	-	-	-
Predição do Desempenho do Aluno usando Sistemas de Recomendação e Acoplamento de Classificadores(2)	-	-	-	-	-
Modelo de Regressão Linear aplicado à previsão de desempenho de estudantes em ambiente de aprendizagem(3)	-	-	-	-	**p-value=0,1886**
Previsão de Desempenho de Estudantes em Cursos EaD Utilizando Mineração de Dados: uma Estratégia Baseada em Séries Temporais (4)	-	81,5%	-	-	-
A decision-tree-based system for student academic advising and planning in information systems programmes (5)	-	-	-	-	-
Analysis and predictions on students behavior using decision trees in weka environment (6)	-	-	-	0	-
MD_PREAD	**77,08**	**84,05%**	**50,23**	**0,646**	p-value= 0.000006

Fonte: Próprio autor

7 CONSIDERAÇÕES FINAIS

Neste capítulo serão apresentadas as considerações finais com as conclusões do experimento, as contribuições da pesquisa e os trabalhos futuros.

7.1 Conclusões

Este trabalho propôs um modelo denominado MD_PREAD e para tanto se utilizou da ferramenta de mineração de dados *RapidMiner*, ela permitiu que a configuração de todas as etapas do modelo. O cenário escolhido para a aplicação do modelo foi o Instituto Federal de Educação, Ciência e Tecnologia do Amazonas em seu ambiente virtual de aprendizagem.

Através da técnica de mineração de dados, árvore de decisão, utilizadas em dados históricos foi possível encontrar a regra de classificação mais adequada considerando o conjunto de treinamento utilizado. O classificador de árvore de decisão selecionado foi o *Decision Tree* com o critério *Gain Information* que nos testes de comparação realizados, obteve uma acurácia de 55% e um índice *Kappa* de 0,2 que indica a confiabilidade do classificador em uma escala de 0-1.

Foi feito o experimento com a disciplina de Libras do curso de Filosofia da Educação ofertada no programa Universidade Aberta no Ambiente Virtual de Aprendizagem do Instituto Federal de Educação, Ciência e Tecnologia do Amazonas, com coletas semanais e geração de um arquivo contendo a lista dos aprendizes com o risco de reprovar e enviando este arquivo a um protótipo de Sistema de Recomendação, objeto da dissertação de um mestrando da Universidade do Vale do Rio dos Sinos, que por sua vez, com base nestas listas enviou as recomendações ao Coordenador e tutores da disciplina.

Concluiu-se com os resultados deste trabalho que as predições geradas a cada semana teve resultados dinâmicos, não sendo uma lista estática, variando conforme as interações feitas no ambiente.

Conclui-se ainda que na primeira semana por ser de apresentação da disciplina e disponibilização de materiais, iniciou com cerca de 1800 interações e variou ao longo das semanas. As interações são indicadores de interesse e refletem no resultado das avaliações.

Concluiu-se também que as atividades "*resource view*" e "*forum view discussion*" foram as mais visitadas pelos aprendizes, que com essa informação o coordenador, tutores e professores podem reavaliar sua metodologia para aproveitar o interesse dos aprendizes nas atividades predominantes e estimular a visita a outras atividades pedagogicamente relevantes e que ainda não despertaram o interesse.

Concluiu-se que o modelo é viável de ser aplicado. As taxas de precisão, encontradas no segundo treinamento são boas quanto a acurácia e precisão de classificação e razoáveis em relação ao índice *recall*.

Concluiu-se que a utilização do PCA na etapa de pré-processamento contribui para melhorar os índices de eficiência e que quanto maior o tamanho do conjunto de treinamento melhores serão os resultados.

7.2 Contribuições

Este trabalho contribuiu com a aplicação das técnicas de mineração de dados, mostrando que é possível fazer prognósticos de aprendizes com situação de reprovação utilizando árvore de decisão.

Mostrou ainda, que o modelo, se adotado, pode contribuir para a redução dos índices de reprovação de um grupo de risco de reprovar. Uma contribuição não só científica, mas também social, pois propicia a continuidade do aluno na instituição rumo à conclusão de seus estudos, custeados, no caso das instituições federais de ensino, por recursos públicos.

A Tabela 7 tem o mesmo formato da Tabela 2 e nesta tabela foi incluída uma linha para a comparação dos trabalhos com o modelo proposto. Para comparar os trabalhos, foram consideradas as estratégias utilizadas para fazer a predição, o foco de aplicação e os serviços entregues. A seguir é feita a descrição de cada critério da comparação:

- Abordagem com interações: O MD-PREAD usou a abordagem com interações, para coletar indícios do comportamento dos aprendizes com as atividades ofertadas pelo professor, permitindo aprender com estas interações qual o nível de interesse deles pelas atividades;
- Acoplamento de Classificadores: O MD-PREAD não utilizou o acoplamento de classificadores, foi utilizado apenas um classificador;
- Séries Temporais: O MD-PREAD utilizou as séries temporais para após o pré-processamento, utilizar o conjunto de treino e encontrar o classificador de árvore de decisão mais adequado para a predição;
- Regressão Linear: O MD-PREAD não utilizou regressão linear, apesar de ser este um dos métodos utilizados para predição, optou-se pela abordagem de classificação porque esta apresenta uma série de alternativas para se interpretar dados;
- Árvore de Decisão: O MD-PREAD utilizou a árvore de decisão por esta ser um tipo de classificador que oferece um diferencial quanto à possibilidade de interpretação dos dados gerados pelo uso dos métodos de predição, pois outros, tais como Redes Neurais Artificiais possuem como deficiência justamente a dificuldade de identificar as causas que levam aos resultados das predições;
- Educação a Distância: O MD-PREAD enfoca o contexto de Educação a distância porque este proporciona uma grande quantidade de dados que são armazenadas para fins de controle e que vem sendo utilizado de forma crescente;
- Foco em Reprovação: O MD-PREAD tem o foco em reprovação porque a desistência e a evasão são consequências que se tratadas na fase em que o aprendiz ainda está no ambiente podem ser evitadas;
- Lista com as notas das atividades do grupo de risco**:** O MD-PREAD gera listagem de grupos de risco com predição de reprovar para possibilitar a sistemas de recomendação ou gestores educacionais a intervenção tempestiva;
- Percentual individual de indicador de reprovação**:** O MD-PREAD utiliza o cálculo do percentual de predição para mensurar a tendência do aprendiz a reprovar.

Considerando a Tabela 7, nota-se que o MD-PREAD se diferencia dos demais trabalhos relacionados à predição por usar como conjuntos de teste e treinamento, as médias das notas e o desvio padrão das interações dos aprendizes extraídos do ambiente virtual de aprendizagem, bem como o classificador de árvore de decisão que utiliza como código o algoritmo Decision Tree com o critério *Information Gain*.

Tabela 7 Comparação com os trabalhos relacionados

Trabalhos	Estratégias para a predição					Foco		Serviços	
	Abordagem com interações	Acoplamento de Classificadores	Séries Temporais	Regressão Linear	Árvore de Decisão	EaD	Reprovação	Fornece lista com as médias das atividades do grupo de risco	Fornece percentual indicador de reprovação
Predição de Reprovação de Alunos de Educação a Distância Utilizando Contagem de Interações (DETONI; ARAUJO; CECHINEL, 2014)	Sim	Não	Não	Não	Não	Sim	Não	Não	Não
Predição do Desempenho do Aluno usando Sistemas de Recomendação e Acoplamento de Classificadores(GOTARDO; CEREDA; JUNIOR, 2013)	Sim	Sim	Sim	Não	Não	Sim	Não	Não	Não
Modelo de Regressão Linear aplicado à previsão de desempenho de estudantes em ambiente de aprendizagem(RODRIGUES, R. L., DE MEDEIROS, F. P. A., & GOMES, 2013)	Sim	Não-	Não	Sim	Não	Sim	Não	Não	Não
Previsão de Desempenho de Estudantes em Cursos EaD Utilizando Mineração de Dados: uma Estratégia Baseada em Séries Temporais(GOTTARDO; KAESTNER; NORONHA, 2012)	Sim	Não	Sim	Não	Não	Sim	Não	Não	Não
A decision-tree-based system for student academic advising and planning in information systems programmes. (WERGHI; KAMOUN, 2010)	Não	Não	Sim	Não	Sim	Não	Não	Não	Não
Analysis and predictions on students' behavior using decision trees in weka environment(BRESFELEAN, 2007)	Sim	Não	Sim	Não	Sim	Sim	Não	Não	Não
MD-PREAD: Um modelo para predição de reprovação de aprendizes na educação a distância usando árvore de decisão.	Sim	Não	Sim	Não	Sim	Sim	Sim	Sim	Sim

Fonte: Próprio autor

7.3 Trabalhos Futuros

O MD-PREAD apresenta uma proposta inicial que pode ser melhorada. No decorrer da elaboração deste trabalho, melhorias foram identificadas que apontam para trabalhos futuros. A seguir são listadas algumas sugestões:

- com base neste trabalho é possível vislumbrar outras iniciativas a serem realizadas como, por exemplo, o desenvolvimento de um *plugin* no *Moodle* que possa absorver ou encapsular a ideia do modelo e agregar essa funcionalidade à ferramenta;
- outra possibilidade é o desenvolvimento de um sistema em uma linguagem de programação que agregue APIs da ferramenta de mineração *RapidMinner*;
- realização de testes com mais atributos em conjunto de testes em busca da melhoria dos índices de acurácia e confiabilidade;
- realização de teste com outros classificadores em busca de melhores resultados.
- avaliação do segundo treinamento utilizando o MD-PREAD com o Classificador J48 e verificar os resultados.

REFERÊNCIAS

ARAGÃO, Cláudia Regina Dantas. *EDUCAÇÃO A DISTÂNCIA Contextualização da EaD*. Disponível em: <http://www.ebah.com.br/content/ABAAABE6UAI/educacao-a-distancia-ead>. Acesso em: 15 maio 2014.

BAKER, R.S.J.D., I. S. d. C. A. Mineração de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação.*, 2011. Disponível em: <http://www.br-ie.org/pub/index.php/rbie/article/view/1301>.

BLUM, A, E T Mitchell. Combining labeled and unlabeled data with co -training. 1988, New York, USA: COLT'98 Proceedings of the eleventh annual conference on Computational learning., 1988. Disponível em: <http://dl.acm.org/citation.cfm?doid=279943.279962>.

BRESFELEAN, Vasile Paul. Analysis and predictions on students' behavior using decision trees in weka environment. jun. 2007, [S.l.]: IEEE, jun. 2007. p. 51–56. Disponível em: <http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=4283743>. Acesso em: 6 fev. 2015.

CABENA, P., HADJINIAN, P., STADLER, R., VERHEES, J E ZANASI, A. Discovering Data Mining: From Concept to Implementation,. *Prentice Hall*, 1997. Disponível em: <http://dl.acm.org/citation.cfm?doid=846170.846181>.

CensoEaD.

CORPORATION, MICROSOFT. *Microsoft Access: Sistema de Gerenciamento de Banco de Dados Relacional para Windows, criando aplicativos*. Disponível em: <https://products.office.com/pt-br/access>. Acesso em: 25 nov. 2015.

COSTA, Evandro *et al.* Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. *Jornada de Atualização em Informática na Educação - JAIE*, v. d, p. 1–29, 2012.

COSTA, Jfa. Um ambiente gráfico para facilitar tarefas de data mining via ferramenta R. p. 24–25, 2011. Disponível em: <http://repositorium.sdum.uminho.pt/handle/1822/19829>.

DETONI, Douglas; ARAUJO, Ricardo Matsumura; CECHINEL, Cristian. *Predição de Reprovação de Alunos de Educação a Distância Utilizando Contagem de Interações*. *Anais do Simpósio Brasileiro de Informática na Educação*. [S.l: s.n.]. Disponível em: <http://www.br-ie.org/pub/index.php/sbie/article/view/3026>. Acesso em: 5 fev. 2015. , 2014

DRACHSLER, Hendrik *et al.* The Pulse of Learning Analytics - Understandings and Expectations from the Stakeholders. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, v. 15, n. 3, p. 42–57, 2012. Disponível em: <http://www.ifets.info/journals/15_3/4.pdf\nhttp://lnx-hrl-075v.web.pwo.ou.nl/handle/1820/4506\nhttp://dspace.ou.nl/handle/1820/3850>.

FARIA, Aa; SALVADORI, Angela. A Educação a Distância e seu Movimento Histórico no Brasil. *Revista das Faculdades Santa Cruz*, p. 15–22, 2010. Disponível em: <http://santacruz.br/v4/download/revista-academica/14/08-educacao-a-

distancia-e-seu-movimento-historico-no-brasil.pdf>.

FAYYAD, U. a.-S. *From Data Mining to Knowledge Discovery:An Overview*. Menlo Park: MIT,611 p., 1996.

GOTARDO, Reginaldo; CEREDA, Paulo Roberto Massa; JUNIOR, Estevam Rafael Hruschka. Predição do Desempenho do Aluno usando Sistemas de Recomendação e Acoplamento de Classificadores. *Anais do Simpósio Brasileiro de Informática na Educação*, v. 24, n. 1, 2013. Disponível em: <http://www.br-ie.org/pub/index.php/sbie/article/view/2544>. Acesso em: 5 fev. 2015.

GOTTARDO, Ernani; KAESTNER, Celso; NORONHA, Rv. Previsão de Desempenho de Estudantes em Cursos EAD Utilizando Mineração de Dados: uma Estratégia Baseada em Séries Temporais. *Simpósio Brasileiro de Informática na Educação*, n. SBIE, p. 26–30, 2012. Disponível em: <http://www.br-ie.org/pub/index.php/sbie/article/view/1758>.

HÄMÄLÄINEN, W., VINNI, M. *Classifiers for Educational Data Mining*. Flórida: Chapman & Hall/CRC, Pages: 57-71., 2010.

HAN, J. AND KAMBER, M. Data mining: concepts and techniques. *Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence The MIT Press.*, 1992.

HAN, J. AND KAMBER, M. Data mining: concepts and techniques. *Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.*, 2000.

JOHNSON, RICHARD ARNOLD, And Dean W. Wichern. *Applied multivariate statistical analysis.* 4. ed. Englewood Cliffs, NJ: [s.n.], 1992.

LANDIS, J R; KOCH, G G. The measurement of observer agreement for categorical data. *Biometrics*, v. 33, n. 1, p. 159–174, 1977.

M. A. CHATTI., AL. L. DYCKOFF, U.SCHROEDER, H.; THÜS. A reference model for learning analytics. 2012, [S.l.]: Int. J. Technol. Enhanc. Learn., Inderscience Geneva, Switzerland., 2012. p. 318–331.

MACFADYEN, L.P., DAWSON, S. “Mining LMS Data to Develop an ”Early Warning System“ for Educators: A Proof of Concept”. *Computers & Education, no. 54*, p. 588–599, 2010.

MAIMON, O. & LIOR, R. *Data Mining and Knowledge Discovery Handbook*. New York - USA: Springer, 2010.

MARI, Marcelo M *et al.* ANÁLISE DA EVASÃO E REPROVAÇÃO DE ALUNOS EM CURSOS A DISTÂNCIA : UM ESTUDO EMPÍRICO. 2011, [S.l.]: XXXIX Congresso Brasileiro de Educação em Engenharia, 2011.

MENZIES, Tim. Data Mining for. *In McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of EducationElsevier, Oxford, UK.*, v. 10, 2003.

Ministério da Educação. 2015. Disponível em: <http://portal.mec.gov.br/component/content/article?id=12823:o-que-e->.

R. FERGUSON. Learning analitycs: drivers, developments and challenges. *International Journalof Technology Enhanced Learning (IJTEL)*, v. 4(5/6), p. 304–317, 2012.

RIGO, Sandro José *et al.* Minerando Dados Educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. *Revista Brasileira de Informática na Educação*, v. 22, p. 132–146, 2014. Disponível em: <http://www.br-ie.org/pub/index.php/rbie/article/view/2423>.

RODRIGUES, R. L., DE MEDEIROS, F. P. A., & GOMES, A. S. Modelo de Regressão Linear aplicador à previsão de desempenho de estudantes em ambiente de aprendizagem. *Anais do XXIV SBIE.*, 2013.

ROMERO, Cristobal; VENTURA, Sebastin. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, v. 40, p. 601–618, 2010.

RUMBLE, Greville. Tecnologia da educação a distância em cenários do terceiro mundo. 2000, Cuiabá: In: PRETI, Oreste (Org.). Educação a distância: construindo signifi cados. Cuiabá: NEAD/ IE – UFMT, 2000. p. 43–63.

RYSZARD S MICHALSKI, IVAN BRATKO, Avan Bratko. *Machine Learning and Data Mining; Methods and Applications*. New York, NY, USA: Sons, John Wiley &, 1998.

SACCOL, AMAROLINDA I. C. Z. ; BARBOSA, JORGE L. V. ; SCHLEMMER, ELIANE ; REINHARD, Nicolau . Mobile Learning in Organizations: Lessons Learned from Two Case Studies. *International journal of information and communication technology education, v. 7*, p. 11–24, 2011.

T., S. Mishra; D., Kumar; GUPTA. Mining Students' Data for Prediction Performance. *2014 Fourth International Conference on Advanced Computing & Communication Technologies*, p. 255–262, 2014.

W. GRELLER, H. Drachsler. Translating Learning into Numbers: a generic framework for learning analytics. *Educational Technology & Society*, v. 15 n.3, p. 42–57, 2012.

WERGHI, Naoufel; KAMOUN, Faouzi Kam. A decision-tree-based system for student academic advising and planning in information systems programmes. *International Journal of Business Information Systems*, v. 5, n. 1, p. 1, 2010.

WITTEN, I.H., FFRANK E., HALLL, M.A. *Data Mining: Practical Machine Learning Tools and Techniques.* 3 ed ed. San Francisco: Morgan Kaufmann, 2011.