A Model for Predicting Disapproval of Apprentices in Distance Education Using Decision Tree

Um Modelo para Predição de Reprovação de Aprendizes na Educação a Distância usando Árvore de Decisão

João Luiz Cavalcante Ferreira
Instituto Federal de Educação, Ciência e Tecnologia do Amazonas (IFAM)
Manaus, Amazonas, Brasil
lcavalcantef@gmail.com

André Filipe Aloise
Instituto Federal de Educação, Ciência e Tecnologia do Amazonas (IFAM)
Manaus, Amazonas, Brasil
aaloise@gmail.com

Vítor Kehl Matter
Universidade do Vale do Rio dos Sinos (UNISINOS)
São Leopoldo, Rio Grande do Sul
Brasil
vitorkmatter@gmail.com

Jorge Luis Victória Barbosa
Universidade do Vale do Rio dos Sinos (UNISINOS)
São Leopoldo, Rio Grande do Sul
Brasil
jbarbosa@unisinos.br

Sandro José Rigo
Universidade do Vale do Rio dos Sinos (UNISINOS)
São Leopoldo, Rio Grande do Sul
Brasil
rigo@unisinos.br

Kleinner S. Farias de Oliveira
Universidade do Vale do Rio dos Sinos (UNISINOS)
São Leopoldo, Rio Grande do Sul
Brasil
kleinnerfarias@unisinos.br

ABSTRACT

This paper proposes the MD-PREAD, a model that uses the decision tree technique for predicting apprentices with risk of failure. The capability of choosing the decision tree as a way to generate a greater set for educators is the highlight of this project. After the data was collected and processed, it was possible to generate a list of students that had the greatest chance to fail, this data would give the opportunity to help the students to recover their grades before the end of the course. Finally, to evaluate the model, the indexes of the classifiers were compared and the J48 algorithm stood out with an accuracy predominance of 84.5%, precision of 85.52%. It was concluded that the MD-PREAD model can assist in the prognosis of groups at risk of failure.

CCS CONCEPTS

• **Information systems** → **Computing platforms**.

KEYWORDS

Prediction, Decision Tree, Educational Data Mining

ACM Reference Format:
João Luiz Cavalcante Ferreira, André Filipe Aloise, Vítor Kehl Matter, Jorge Luis Victória Barbosa, Sandro José Rigo, and Kleinner S. Farias de Oliveira. 2019. A Model for Predicting Disapproval of Apprentices in Distance Education Using Decision Tree. In *XV Brazilian Symposium on Information Systems (SBSI'19), May 20–24, 2019, Aracaju, Brazil.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3330204.3330223

SBSI'19, May 20–24, 2019, Aracaju, Brazil

ACM ISBN 978-1-4503-7237-4/19/05…$15.00
https://doi.org/10.1145/3330204.3330223

1 INTRODUÇÃO

A Educação a Distância (EaD) baseia-se nos princípios da igualdade e do ensino permanente, acessível a qualquer pessoa. A EaD no Brasil tem se consolidado com diversos estudantes optando por utilizar essa modalidade de ensino [4]. O Censo EAD.BR 2015 registrou uma evasão entre 26% e 50%, com 40% das ocorrências nas instituições que oferecem cursos regulamentados totalmente a distância [17].

Um dos diferenciais de cursos de EAD é a grande quantidade de dados gerada pelas interações no ambiente educacional, o que abre novas possibilidades para estudar e compreender estas interações. 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 afirma Romero et al. [16]. De acordo com Rigo et al. [13] a Learning Analytics 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.

Segundo Evandro et al. [2] a área emergente de Mineração de Dados Educacionais (EDM) 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).

Um dos desafios dos pesquisadores nesta área 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 [4, 9] ou desista [6]. Segundo Cristobal e Sebastian [15], existe a necessidade de se desenvolver ferramentas específicas e fáceis

de usar, para que professores/tutores não familiarizados com as técnicas EDM também possam se valer das descobertas das áreas.

Uma das dificuldades relatadas na área está associada com a dificuldade, com uso dos métodos atuais de predição, de gerar subsídios para os professores de forma a apoiar a sua ação no atendimento dos alunos. Outra dificuldade citada se refere ao tratamento de grande volume de dados gerado na mediação digital.

Este trabalho explora a técnica de Árvore de Decisão [10] para apoiar na geração de dados sobre as predições, de modo que os professores possam guiar de forma mais adequada a sua metodologia de atuação. Além disso, o trabalho avalia as possibilidades da análise de componentes principais, que segundo Johnson et al. [8] permite reduzir a dimensionalidade de dados sem perder as informações relevantes, em uma etapa intermediária de cálculo para uma análise posterior com diversas tarefas, tais como agrupamentos, classificação, discriminação, redes neurais, entre outras. Assim, a aplicação de componentes principais sob um conjunto de dados, pode 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.

O principal objetivo deste artigo é propor o MD-PREAD, um modelo de predição de um grupo de risco de reprovação, levando em conta a necessidade de geração de subsídios para os educadores e tratando a dimensionalidade dos dados. O modelo foi avaliado com dados reais de uma disciplina cursada em ambiente virtual de aprendizagem e sua avaliação através de prototipagem no RapidMiner[1] uma ferramenta de mineração de dados.

[1] https://rapidminer.com

Tabela 1: Comparação de trabalhos relacionados

Trabalhos	Estratégias para a predição					Foco		Serviços	
	Abordagem com interações	Acoplamento de Classificado	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 et al. 2014]	Sim	Não	Não	Não	Não	Sim	Não	Não	Não
Predição do Desempenho do Aluno usando Sistemas de Recomendação e Acoplamento de Classificadores[Gotardo et al. 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
Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina [De Brito et al. 2014]	Sim	Não	Sim	Não	Não	Não	Não	Não	Não
Um modelo preditivo para diagnóstico de evasão baseado nas interações de alunos em fóruns de discussão [Silva et al. 2015]	Sim	Não	Sim	Não	Sim	Sim	Não	Não	Não
Utilização de técnicas de Mineração de Dados Educacionais para a predição de desempenho de alunos de EaD em Ambientes Virtuais de Aprendizagem [Rabelo et al. 2017]	Sim	Não	Sim	Não	Sim	Sim	Não	Não	Não

O ambiente da pesquisa é o Instituto Federal de Educação, Ciência e Tecnologia do Amazonas (IFAM), com aprendizes do programa Universidade Aberta do Brasil (UAB). Foram realizadas coletas de dados históricos de 10 disciplinas de um grupo de 30 aprendizes em dois semestres consecutivos, sendo que o total de alunos matriculados foi de 125 e o total de interações obtidas foi de 41070. Os dados foram submetidos ao algoritmo classificador de árvore de decisão que aplica as regras encontradas em um conjunto de testes, a metodologia utilizada foi a CRISP-DM[1].

O artigo está organizado em 5 seções. A seção 1 apresenta a motivação do trabalho e uma revisão da literatura. A seção 2 aborda os trabalhos relacionados. A seção 3 apresenta o modelo proposto. Por sua vez a seção 4 trata da implementação. Finalmente na seção 5 são feitas as considerações finais e sugeridos trabalhos futuros.

2 TRABALHOS RELACIONADOS

Nesta seção são apresentados os trabalhos relacionados estudados como parte da contextualização e justificativa do modelo.

Douglas et al. [4] 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 então ajustar seus instrumentos pedagógicos para evitar que estes estudantes reprovem ou evadam.

No trabalho de Reginaldo et al. [5] verifica-se que o processo de aprendizado de máquina apresenta algumas abordagens a respeito de como decidir sobre novos dados adquiridos e seus relacionamentos. 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.

Rodrigo et al. [14] chegaram à conclusão de que o 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. Daniel et al. [3] compararam alguns algoritmos, entre eles: Naive Bayes, IBk, SMO, RandomForest e Multipayer Perceptron, dos quais destacou-se o primeiro com 75% de precisão.

Francisco et al. [17] apresentaram uma proposta de modelo focada nas interações com os fóruns de discussão e suas respectivas notas, submetendo o conjunto de treinamento a cinco classificadores dos quais o J48 alcançou em alguns casos o melhor índice com 73% de precisão reforçando que técnicas baseadas em árvore de decisão são adequadas para a predição em ambientes virtuais de aprendizagem no entanto seu objetivo era buscar um diagnóstico acerca das causas de evasão.

No trabalho de Humberto et al. [11] foi apresentado uma demonstração sobre a eficácia das técnicas de mineração de dados, sendo utilizado um ambiente Virtual de Aprendizagem, aplicadas técnicas de refinamento e selecionados seis indicadores de desempenho com base nas interações. Foram criados elementos de uma tabela de classificação com três valores: regular, bom e ótimo. Para avaliar comparou-se os resultados de dois classificadores, o ID3 e J48, concluindo que o segundo obteve resultados melhores com acurácia de até 96,5%.

Ao analisar a comparação dos trabalhos relacionados na Tabela 1, pode-se perceber que 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. Ressalta-se que o Modelo pode ser configurado para operar com qualquer classificador, preferencialmente aquele de melhores índices.

3 MODELO PROPOSTO

Nesta seção é apresentado o modelo de predição de reprovação. Esse 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.

O MD-PREAD é um modelo de sistema de predição educacional para gestores educacionais e sistemas de recomendação, com foco na predição da 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.

3.1 Arquitetura

A Figura 1 apresenta 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.

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, o que possibilita a geração e a disponibilização de uma lista de possíveis reprovados em um relatório para exportação e utilização em outros sistemas.

3.2 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

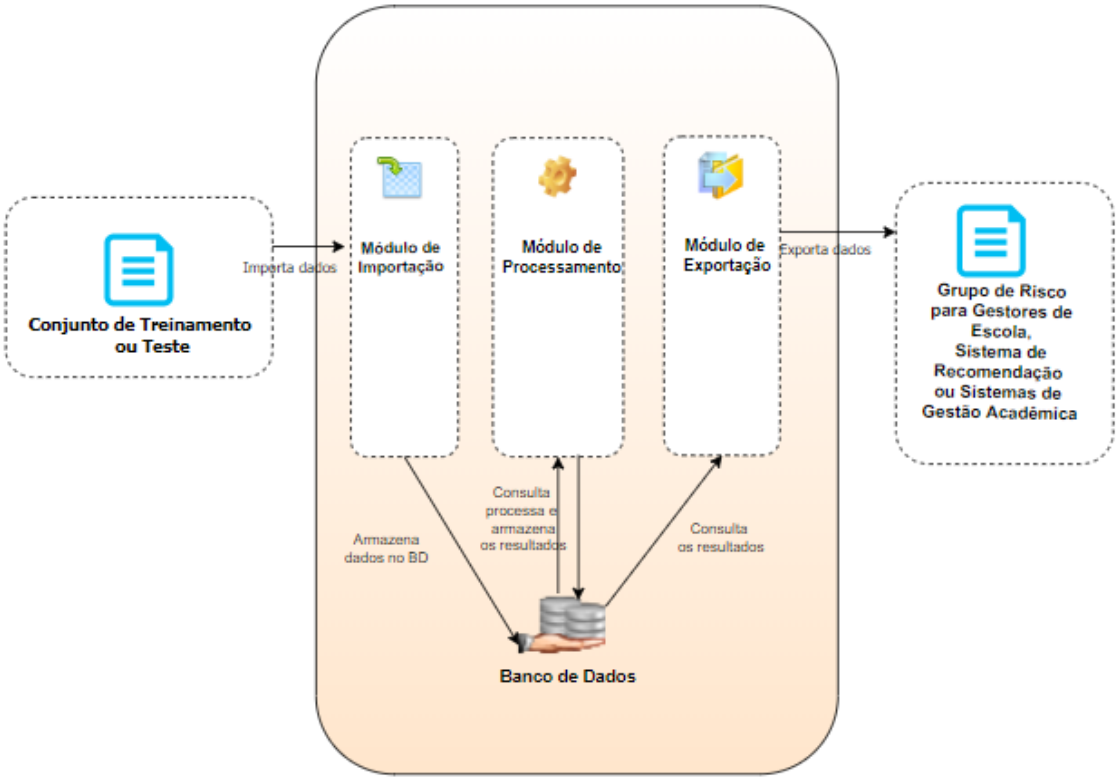

Figura 1: Arquitetura do MD-PREAD

geração de regras e 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, uma segunda listagem contendo as interações dos aprendizes no ambiente virtual de aprendizagem e uma terceira listagem contendo as médias parciais das avaliações na disciplina cursada pelos aprendizes no ambiente virtual de aprendizagem.

Desta forma, os dados tratados são os seguintes: a) dados de identificação dos aprendizes, tais como Matrícula, Nome, E-mail, Curso, Matriz curricular, Nascimento, Período letivo inicial, Renda familiar, Sexo, Tipo de escola de origem, Turma; b) interações dos aprendizes no ambiente virtual de aprendizagem, tais como Curso, Hora, Endereço IP, Nome completo, Ação, Informação; c) médias parciais das avaliações na disciplina cursada pelos aprendizes no ambiente virtual de aprendizagem, como os dados de Disciplina, Nome, Sobrenome, Endereço de e-mail, Média das atividades.

Antes de iniciar o processo de importação que disponibiliza o conjunto de treinamento para identificar a regra de predição, é necessário preparar o conjunto de dados de teste, o que é feito selecionando os atributos que são usados para a predição conforme ilustrado na Figura 2.

Este mesmo processo de Extração, Tratamento e Limpeza (ETL) é análogo para os dados de treinamento e teste, sendo 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 final 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 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 demonstrado na Figura 3.

3.3 Módulo de Processamento

O módulo de processamento apresentado na Figura 4 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.

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, precisão e recall, os dados disponíveis para este

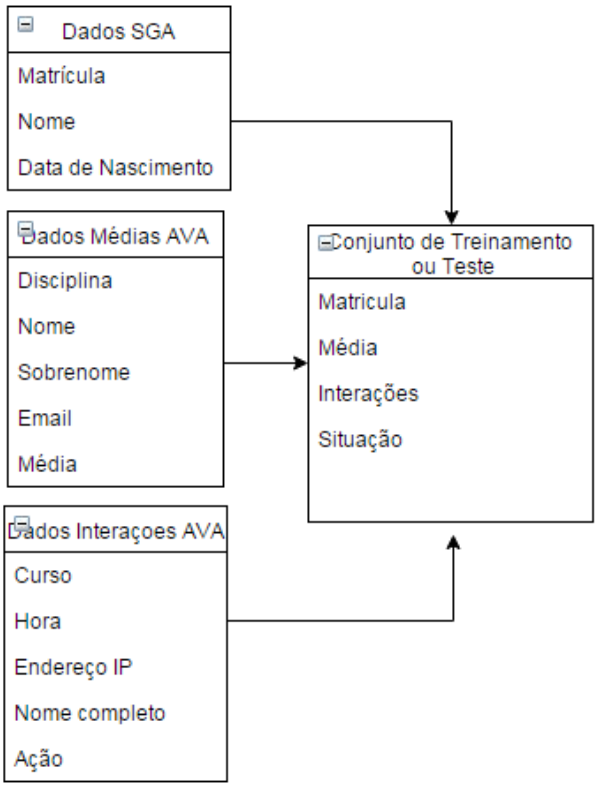

Figura 2: Modelagem dos dados

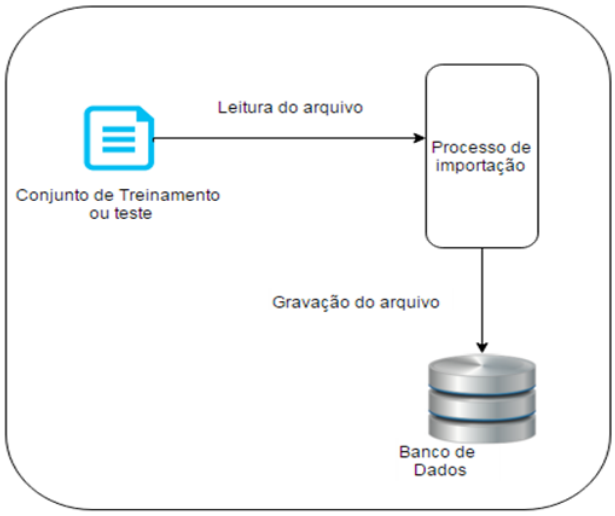

Figura 3: Módulo de Importação

modelo são as médias dos alunos e o desvio padrão das interações, são estes que ao final do processo resultarão em um padrão de comportamento.

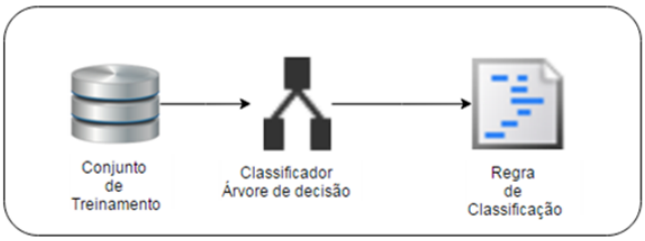

Figura 4: Módulo de Processamento

3.4 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 5.

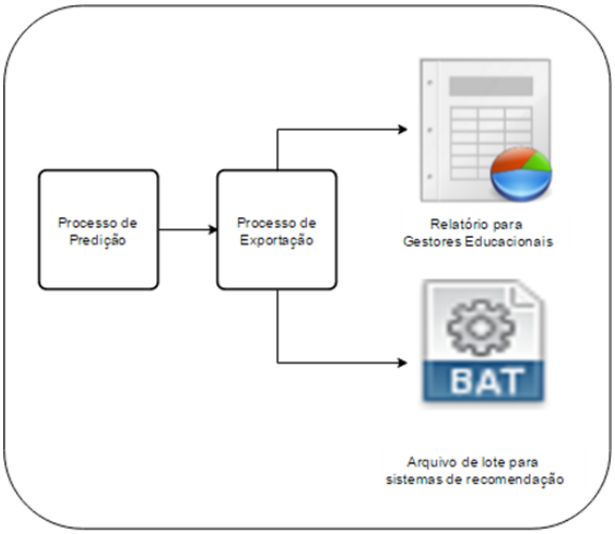

Figura 5: Módulo de Exportação

A Figura 6 ilustra o conjunto de regras obtidas com o uso dos dados e o algoritmo de árvore de decisão utilizado. Pode-se observar que com este recurso as diversas variáveis tratadas ficam evidenciadas com relação ao seu impacto e podem ser usadas pelos professores como subsídio para guiar a sua ação com os alunos.

4 IMPLEMENTAÇÃO E DISCUSSÃO DOS RESULTADOS

Nesta seção são apresentados os aspectos de implementação do modelo e discutidos os resultados.

4.1 Aspectos de implementação

Para permitir a avaliação do MD-PREAD, foi necessário configurar o protótipo em uma ferramenta de mineração de dados capaz de

```
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
```

Figura 6: Árvore de decisão gerada

auxiliar nos procedimentos e métodos necessários para cumprir as etapas na busca do perfil do aprendiz com tendência de reprovar em uma disciplina.

A seguir 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.

O RapidMiner é uma aplicação open-source para DM [12]. 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.

A configuração do processo de predição no RapidMiner utilizou os operadores tais como o "Read" para a leitura dos conjuntos de teste e treinamento, os operadores "Set" para a configuração dos atributos, o "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 arquivos de saída.

Verificou-se a necessidade de reduzir o número de recursos avaliativos, para tanto utilizou-se a técnica de análise dos principais componentes (PCA). Segundo Richard e Dean [7] 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. Com o auxílio do PCA, método utilizado para encontrar os principais componentes, foram selecionados 3 atributos: o "fórum view discussion", "resource view" e "forum view forum". Estes atributos foram utilizados como filtro para o cálculo do desvio padrão no conjunto de treinamento.

4.2 Processo de avaliação

O processo de avaliação 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.

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.

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. Desta vez utilizando os algoritmos descritos na tabela acima. Para este volume de dados destacou-se o J48 com acurácia de 84,5%, precisão na classificação de reprovados de 85,52% conforme apresentado na Tabela 3.

A regra utilizada no MD-PREAD foi gerada em função do algorítmo de árvore de decisão J48 que foi aquele avaliado como o melhor dentre os selecionados para a predição do grupo de risco de reprovar destacando-se os índices de acurácia de 84,5% e de predição para reprovar de 85,52% comparado com os demais algoritmos utilizados para árvore de decisão. A classificação foi baseada no conjunto de treinamento e levou em consideração os polos, as disciplinas, as médias e o desvio padrão das interações nos componentes selecionados pelo PCA. Verificou-se que nem sempre o desvio padrão maior determina a predição de aprovação, mas na maioria dos polos a afirmativa é verdadeira.

A cada semana o grupo de risco com predição de reprovação era selecionado pelo algoritmo de árvore de decisão utilizado no experimento. Observou-se que na primeira predição feita o índice foi de 100% dos aprendizes na lista do grupo de risco de reprovar, o que se justifica por não ter havido nenhuma atividade realizada pelos aprendizes, na segunda predição o percentual foi de 43% do grupo inicial, na terceira predição aumentou para 66% em relação a predição anterior e na quarta predição evidenciou-se uma redução de 22% em relação à predição anterior conforme demonstrado na Figura 7.

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. Ressalta-se que o professor possui acesso às variáveis que foram utilizadas e percebe o interesse dos aprendizes pelos recursos "fórum view discussion", "resource view" e "forum view forum", de modo a ter a oportunidade de explorá-los melhor afim de tornar o ambiente virtual de aprendizagem mais atrativo reformulando assim suas metodologias.

Tabela 2: Análise dos Índices dos algoritmos de classificação

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	**84,5%**	**77,08%**	**85,52%**	**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%

De posse da lista de risco de reprovação o professor pode buscar encontrar as lacunas dos aprendizes e adequar sua metodologia com o objetivo de melhorar o desempenho do aluno bem como seu interesse pela disciplina, contribuindo assim para a redução do índice de reprovação da turma.

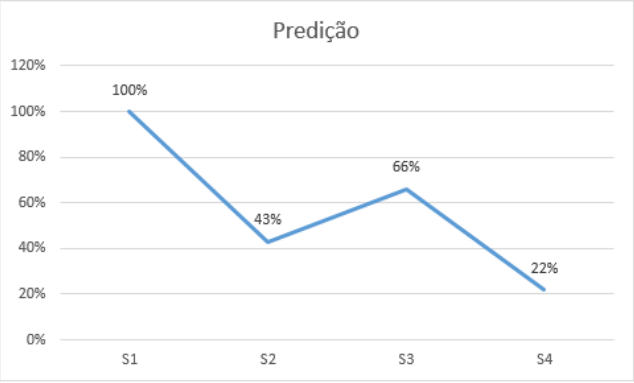

Figura 7: Evolução da Predição

5 CONSIDERAÇÕES FINAIS

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

Após o processo foi possível gerar uma lista de possíveis reprovados que teriam a oportunidade de serem recuperados antes do final da disciplina. Finalmente para avaliação do modelo foram comparados os índices dos classificadores e destacou-se o J48 com uma predominância de acurácia de 84,5%, precisão de 85,52%.

Concluiu-se que o MD-PREAD pode auxiliar no prognóstico de grupos de risco de reprovação, uma vez que possibilitou a geração e a disponibilização de uma lista de possíveis reprovados em um relatório para exportação e utilização em outros sistemas. A percepção de utilidade da árvore de decisão também foi considerada positiva, mesmo tendo sido feita uma avaliação limitada ao grupo técnico de desenvolvimento, sendo um ponto de contribuição deste trabalho.

Como trabalho futuro sugere-se o desenvolvimento de um plugin no Moodle que possa absorver ou encapsular a ideia do modelo e disponibilizar essa funcionalidade de modo amplo. Também sugere-se a realização de avaliações mais longas e com grandes grupos de professores, para consolidar a percepção de utilidade dos resultados da árvore de decisão como apoio adicional à ação do educador na tentativa de reversão das tendências apontadas pelos procedimentos de mineração e predição.

AGRADECIMENTOS

Os autores agradecem à Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Universidade do Vale do Rio dos Sinos (Unisinos) e ao Instituto Federal de Educação, Ciência e Tecnologia do Amazonas (IFAM) pelo apoio ao desenvolvimento desse trabalho. Os autores reconhecem especialmente o apoio do Programa de Pós-Graduação em Computação Aplicada (PPGCA) e do Laboratório de Computação Móvel (Mobilab) da Unisinos.

REFERÊNCIAS

[1] Pete Chapman. 2000. *CRISP-DM 1.0: Step-by-step Data Mining Guide.* SPSS.
[2] Evandro Costa, Ryan S J Baker, Lucas Amorim, and Jonathas Magalhães. 2012. Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. *Jornada de Atualização em Informática na Educação - JAIE 2012* (2012), 1–29.
[3] Daniel Miranda De Brito, Iron Araújo de Almeida Júnior, Eduardo Vieira Queiroga, and Thaís Gaudencio Do Rêgo. 2014. Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina. *Anais do XXV SBIE* (2014), 882. https://doi.org/10.5753/cbie.sbie.2014.882
[4] Douglas Detoni, Ricardo Matsumura Araujo, and Cristian Cechinel. 2014. Predição de Reprovação de Alunos de Educação a Distância Utilizando Contagem de Interações. *Anais do XXV SBIE* 896–905 (2014), 896. https://doi.org/10.5753/cbie.sbie.2014.896
[5] Reginaldo Gotardo, Paulo Roberto Massa Cereda, and Estevam Rafael Hruschka Junior. 2013. Predição do Desempenho do Aluno usando Sistemas de Recomendação e Acoplamento de Classificadores. *Anais do XXIV SBIE* 24, 1 (2013), 657–666. https://doi.org/10.5753/CBIE.SBIE.2013.657
[6] L. Heidrich, J. L. V. Barbosa, W. Cambruzzi, Rigo S. J., M. G. Martins, and R. B. S. dos Santos. 2018. Diagnosis of learner dropout based on learning styles for online distance learning. *Telematics and Informatics* 35, 6 (2018), 1593–1606. https://doi.org/10.1016/j.tele.2018.04.007

[7] Richard Arnold Johnson and Dean W Wichern. 1998. *Applied multivariate statistical analysis (4th. ed.)* (4th ed.). Upper Saddle River, N.J Prentice Hall.
[8] H. Johnson, L., Adams Becker, S., Cummins, M., Estrada, V., Freeman, A., and Ludgate. 2013. *Horizon Report. 2013 Higher Education Edition.* Technical Report. Austin, Texas, USA.
[9] L. P. Macfadyen and S. Dawson. 2010. Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education* 54, 2 (2010), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008
[10] Oded Maimon and Lior Rokach. 2010. *Data Mining and Knowledge Discovery Handbook (2nd. ed.)* (2nd ed.). Springer US, New York, USA.
[11] Humberto Rabelo, Aquiles Burlamaqui, Ricardo Valentim, Danieli Silva de Souza Rabelo, and Soraya Medeiros. 2017. Utilização de técnicas de mineração de dados educacionais para predição de desempenho de alunos de EaD em ambientes virtuais de aprendizagem. *Anais do XXVIII SBIE* (2017), 1527. https://doi.org/10.5753/cbie.sbie.2017.1527
[12] RapidMiner 2018. RapidMiner. Retrieved November 5, 2018 from https://rapidminer.com
[13] S. J. Rigo, W. Cambruzzi, J. L. V. Barbosa, and S. C. Cazella. 2014. Minerando Dados Educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. *Revista Brasileira de Informática na Educação* 22 (2014), 132–146.
[14] Rodrigo Lins Rodrigues, Francisco P. A. De Medeiros, and Alex Sandro Gomes. 2013. Modelo de Regressão Linear aplicado à previsão de desempenho de estudantes em ambiente de aprendizagem. *Anais do XXIV SBIE* (2013), 607–616. https://doi.org/10.5753/CBIE.SBIE.2013.607
[15] Cristóbal Romero and Sebastián Ventura. 2010. Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews* 40, 6 (Nov. 2010), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532
[16] Cristobal Romero and Sebastian Ventura. 2013. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 3 (2013), 12–27. https://doi.org/10.1002/widm.1075
[17] Francisco Silva, Josenildo Da Silva, Reinaldo Silva, and Luís Carlos Fonseca. 2015. Um modelo preditivo para diagnóstico de evasão baseado nas interações de alunos em fóruns de discussão. *Anais do XXVI SBIE* (2015), 1187. https://doi.org/10.5753/cbie.sbie.2015.1187