Minerando Dados de Ambientes Virtuais de Aprendizagem para Predição de Desempenho de Estudantes

Henrique Lemos dos Santos*, Fabiane Nunes Prates Camargo† e Sandro da Silva Camargo*

*Campus Bagé, Universidade Federal do Pampa, CEP 96413-170, Bagé, Brasil
†Campus Avançado Santana do Livramento, Instituto Federal Sul Rio-Grandense, CEP 97573-010, Santana do Livramento, Brasil

Abstract. Computer Engineering undergraduate students of Universidade Federal do Pampa face severe dificulties in introductory courses. In order to contributing on decreasing this problem, it started to be used a Course Management System to provide Learning Objects and formative evaluation resources. This system started to collect a huge amount of data about students' behaviour and performance. In this scope, this paper describes the use of data mining algorithms in order to discover knowledge about data obtained in the Moodle Course Management System. Based on these discoveries, it will be possible providing useful information to professors and managers as a mean to creating preventive actions to aim increase students' performance in summative evaluations.

Keywords: Data Mining, Course Management Systems, Performance Prediction, Computer Engineering

INTRODUÇÃO

Em âmbito mundial, a Educação a Distância vem apresentado um crescimento constante, tendo ocorrido um incremento de 9% no número de matrículas no ano de 2011 [1]. No Brasil, a Educação a Distância já responde por 14,6% das matrículas a nível de graduação, conforme dados do censo da Educação Superior de 2010 [2]. Aderindo a esta tendência, a partir de 2004, o Ministério da Educação também regulamentou a possibilidade de oferta de até 20% da carga horária dos cursos superiores na modalidade semipresencial. Dentro deste escopo, o Curso de Engenharia de Computação da UNIPAMPA prevê em seu Projeto Pedagógico de Curso a existência de atividades semipresenciais em várias disciplinas, dentre elas está a disciplina de Introdução a Arquitetura de Computadores, oferecida na primeira fase do curso [3]. Esta escolha foi feita em virtude da constatação que os estudantes deste curso apresentam severas dificuldades em disciplinas introdutórias. Assim, esta disciplina é ministrada em formato híbrido, com aulas presenciais e a semipresenciais. A carga horária dedicada as atividades semipresenciais tem sido explorada principalmente com enfoque na realização de avaliações formativas, a fim de permitir o acompanhamento mais preciso da evolução do conhecimento dos estudantes dentro do primeiro módulo da disciplina. Estatísticas globais mostram que, dentre os cursos mediados por tecnologia, 27% se enquadram na categoria híbrida [1].

Como consequência do uso desta mediação tecnológica, são coletadas enormes quantidades de dados a respeito dos padrões de comportamento e do desempenho dos estudantes [4]. Assim, a utilização de técnicas de mineração de dados pode ser utilizada visando, dentre outras coisas, predizer o desempenho dos estudantes. Dado que as universidades visam melhorar a qualidade do ensino, o uso de técnicas de mineração de dados com foco na educação superior pode auxiliar Universidades, gestores e professores a contribuir com ações focadas nos estudantes com maior dificuldade, de forma a contribuir com uma melhora de seu desempenho [5]. Dentro deste escopo, este trabalho relata um estudo de caso sobre a aplicação de técnicas de mineração de dados que permitem, em estágios anteriores às avaliações somativas, identificar alunos que têm maior risco de reprovação. Os dados que sustentam a abordagem proposta são oriundos de avaliações formativas aplicadas no decorrer da disciplina através do Ambiente Virtual de Aprendizagem *Moodle*. Resultados preliminares mostram que os modelos criados permitem a identificação da propensão à reprovação com taxa de acerto em torno de 69%.

O restante deste trabalho está organizado da seguinte forma: a seção **Material e Métodos** descreve o problema tratado e os dados utilizados nos experimentos. A seção seguinte, chamada **Experimentos e Resultados** é dedicada à descrição dos experimentos realizados e discussão a respeito dos resultados obtidos. Já a seção **Conclusão e Trabalhos Futuros** descreve as conclusões e aponta os trabalhos futuros.

MATERIAL E MÉTODOS

Nesta seção são descritos o problema que está sendo tratado e os dados que foram utilizados nos experimentos.

O Problema

Estudantes de graduação em Engenharia de Computação da Universidade Federal do Pampa (UNIPAMPA) apresentam severas dificuldades em disciplinas introdutórias, onde a taxa de aprovação, historicamente, tem sido em torno de 40%. Essa situação é típica de programas de engenharia ao redor do mundo [6, 7, 8]. Visando a obtenção de um nível mais profundo de compreensão do problema a fim de serem obtidos subsídios para superar esta realidade, foram monitorados 59 estudantes matriculados na disciplina de Introdução à Arquitetura de Computadores, que está sendo ministrada no primeiro semestre de 2012. A abordagem adotada dentro desta disciplina vai além da avaliação convencional, baseada no rendimento alcançado pelo aluno em poucas provas, para assim imputar uma nota ao seu desempenho. Estas avaliações são denominadas de avaliações somativas. Além desta abordagem, também foi realizado um conjunto de avaliações formativas, que tem uma função controladora, e são realizadas durante todo o transcorrer do período letivo. Dentro deste escopo, este trabalho considerou apenas o primeiro módulo da disciplina, que aborda o tema *sistemas de numeração*, que é um dos assuntos onde os estudantes encontram um alto nível de dificuldade. Desta forma, utilizando recursos do Ambiente Virtual de Aprendizagem *Moodle*, são realizadas avaliações formativas relativas ao conteúdo de cada aula, visando um acompanhamento mais preciso do aprendizado do estudante. O tempo estimado para realização de cada uma das avaliações formativas é de 3 horas/aula. Nas 5 primeiras semanas da disciplina, que antecedem a primeira avaliação, são realizadas 8 avaliações formativas sobre os 8 diferentes tópicos abordados dentro do tema selecionado, totalizando 24 horas/aula. Tais avaliações contém entre 10 e 25 questões. Dados os 59 estudantes matriculados na disciplina e as 8 avaliações formativas, tem-se 472 avaliações a serem corrigidas em 5 semanas, totalizando mais de 5500 exercícios a serem avaliados.

É relevante salientar que a realização de avaliações formativas aula a aula, utilizando a carga horária presencial, além de impraticável, geraria uma considerável carga de trabalho adicional ao docente. Já a utilização do ambiente *Moodle* permite que o estudante, ao concluir a avaliação formativa, não só seja informado de sua nota final mas também receba o resultado individual de cada uma das questões que compõe o teste. Por outro lado, tal quantidade de avaliações formativas gera uma enorme quantidade de dados. A análise e interpretação destes dados pode ser explorada a fim de fornecer ao docente um melhor discernimento sobre os perfis de desempenho dos estudantes.

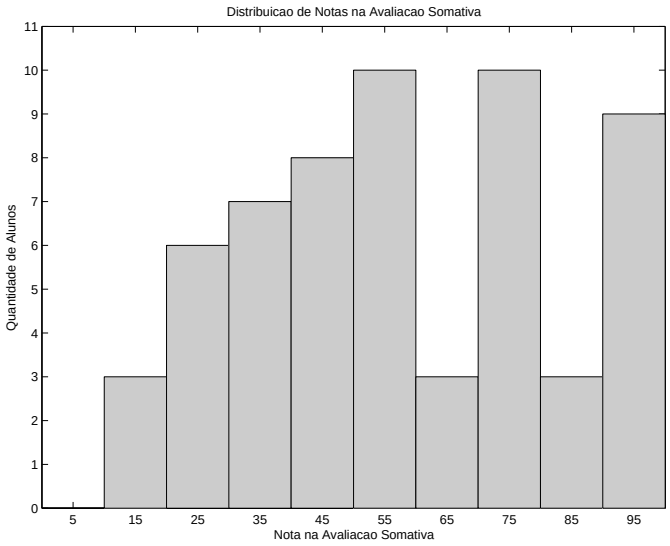

FIGURA 1. Histograma com a Distribuição de Notas na Avaliação Somativa

Após a realização das 8 avaliações formativas, é realizada uma avaliação somativa, cuja distribuição das notas é apresentada na Figura 1. Os 59 estudantes matriculados na disciplinas realizaram tal avaliação somativa, onde a média

das notas foi de 58, sendo que os valores possíveis da avaliação estavam na faixa entre 0 e 100. Para efeitos de análise, os estudantes foram agrupados em dois conjuntos distintos. Os 25 estudantes que alcançaram uma nota igual ou superior a 60, que é a nota requerida para aprovação, foram então rotulados como **aprovados**. Já os 34 estudantes restantes, que obtiveram uma nota inferior a 60, foram rotulados como **reprovados**. A análise das notas da avaliação somativa revelou uma taxa de sucesso de apenas 42% dentre os estudantes da disciplina. Tal taxa de sucesso demonstra claramente a dificuldade encontrada por estes estudantes, e condiz com as taxas de sucesso dos estudantes de semestres anteriores desta disciplina. Também é importante salientar que uma realidade similar de dificuldade é encontrada em outras disciplinas introdutórias deste Curso.

A partir desta realidade, passou-se a tentar identificar se, dentre os dados disponíveis no ambiente *Moodle* com o histórico das avaliações formativas, haveriam evidências que pudessem predizer o sucesso ou não de um estudante na avaliação somativa. A subseção seguinte descreve os dados utilizados nos experimentos.

Dados Utilizados

Um passo essencial para o processo de mineração de dados é a seleção dos dados. Na Tabela 1 são apresentados todos os dados utilizados nos experimentos obtidos, incluindo tanto aqueles obtidos a partir da ferramenta *Moodle* quanto os dados externos a ferramenta, que são a quantidade de presenças e as notas na avaliação somativa. São apresentados o número de ordem do dado, uma descrição curta de seu conteúdo, o seu tipo de dado e o domínio de valores que o dado pode assumir. Faz-se importante salientar o ajuste necessário ao valor da variável 11. Quando o *Moodle* gera a nota média das avaliações, considera somente as notas das avaliações respondidas. Supondo-se que o aluno tenha respondido somente uma das avaliações e tenha obtido nota 100, seria atribuída a ele a média final 100, apesar de outras sete avaliações não terem sido respondidas. O ajuste executado, considera a nota 0 para avaliações não respondidas, atribuindo média 12.5 para o aluno citado anteriormente.

TABELA 1. Descrição dos dados disponíveis

Ordem	Descrição	Tipo de Dado	Domínio
1	Nome do estudante	String	
2	Sobrenome do estudante	String	
3	Endereço de e-mail do estudante	String	
4	Indicador de repetência	Booleano	0 - Calouro, 1 - Repetente
5	Quantidade de presenças nas aulas presenciais	Inteiro	[0,20]
6	Percentual de presenças nas aulas presenciais	Inteiro	[0,100]
7	Quantidade de avaliações formativas respondidas	Inteiro	[0,8]
8	Percentual de avaliações formativas respondidas	Inteiro	[0,100]
9	Quantidade de total de presenças	Inteiro	[0,44]
10	Percentual de presenças	Inteiro	[0,100]
11	Nota média das avaliações formativas	Inteiro	[0,100]
12	Nota média ajustada das avaliações formativas	Inteiro	[0,100]
13*	Nota da avaliação somativa	Inteiro	[0,100]
14†	Rótulo da nota da avaliação somativa	Booleano	0 - Reprovado, 1 - Aprovado
15	Nota na avaliação formativa 1	Inteiro	[0,100]
16	Nota na avaliação formativa 2	Inteiro	[0,100]
17	Nota na avaliação formativa 3	Inteiro	[0,100]
18	Nota na avaliação formativa 4	Inteiro	[0,100]
19	Nota na avaliação formativa 5	Inteiro	[0,100]
20	Nota na avaliação formativa 6	Inteiro	[0,100]
21	Nota na avaliação formativa 7	Inteiro	[0,100]
22	Nota na avaliação formativa 8	Inteiro	[0,100]

* Dado a ser previsto
† Dado a ser previsto

A Figura 2 mostra a quantidade de avaliações formativas respondidas pelos alunos. 24 estudantes, representando aproximadamente 41% do total, responderam todas as avaliações formativas propostas. Este número é muito próximo à proporção de alunos que obtiveram sucesso na avaliação somativa. Por outro lado, aproximadamente 59% deixaram de responder uma ou mais avaliações formativas.

A distribuição de notas nas avaliações formativas é mostrada na Figura 3. A figura mostra dois extremos dentre as 472 avaliações formativas passíveis de serem respondidas, 210 obtiveram nota igual ou superior a 90 representando aproximadamente 44% e, no sentido oposto, em torno de 30% ou não foram respondidas ou atingiram nota inferior a 10.

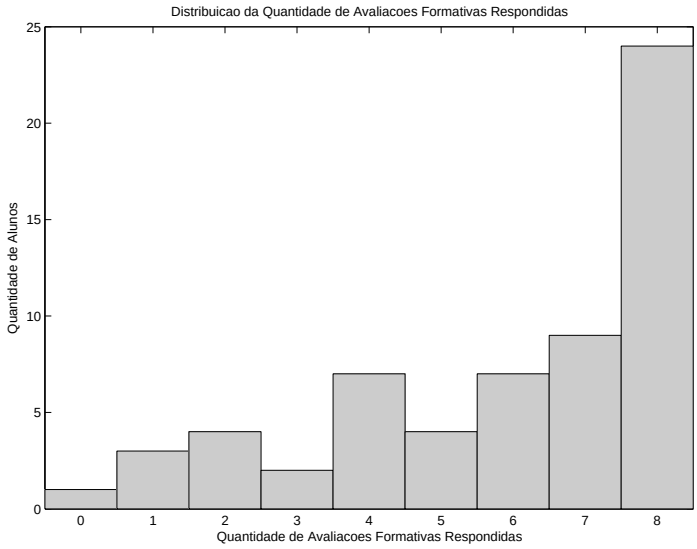

FIGURA 2. Histograma da Quantidade de Avaliações Formativas Respondidas

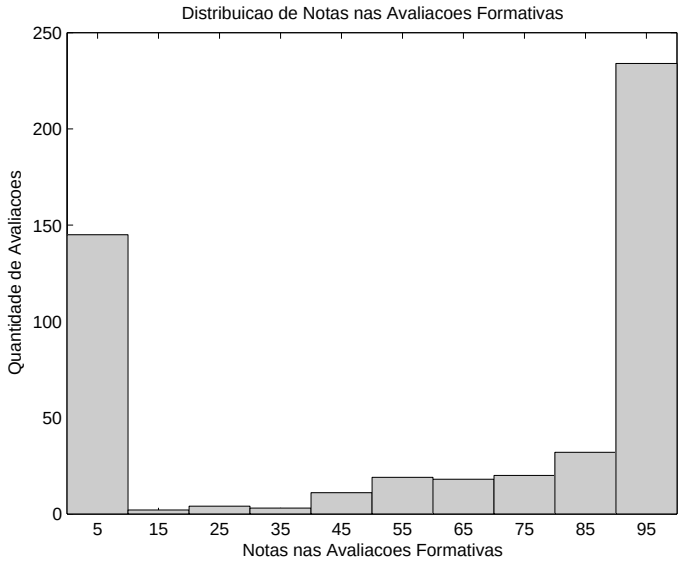

FIGURA 3. Histograma das Notas nas Avaliações Formativas

EXPERIMENTOS E RESULTADOS

A fim de buscar a identificação dos fatores que contribuem para o baixo desempenho dos estudantes na avaliação somativa, foram realizados experimentos de análise de correlação, clusterização e classificação. Os experimentos e resultados são descritos a seguir.

Análise de Correlação

A primeira análise realizada foi a respeito do coeficiente de correlação entre a nota na avaliação somativa e as demais variáveis existentes. O coeficiente de correlação visa medir a dependência linear entre duas variáveis. Este experimento visava identificar quais variáveis, dentre as coletadas, tem uma maior influência na nota da avaliação somativa. Como resultado, foram obtidos os coeficientes de correlação apresentados na tabela 2. É possível verificar que todas as variáveis tem uma correlação pequena ou média com a variável alvo, não existindo nenhuma variável com correlação que possa ser considerada forte [9]. Por outro lado, todas as variáveis apresentam uma correlação positiva, indicando que o crescimento de seus contribui para o crescimento da nota na avaliação somativa. Os resultados também mostram que a nota média ajustada das avaliações formativas é a variável com maior correlação com o resultado da avaliação somativa. Isto mostra que quanto maior é o grau de comprometimento do estudante com a realização das avaliações formativas, melhor será seu desempenho na avaliação formal. Esta conclusão também é fortalecida pela variável que representa o percentual de avaliações formativas respondidas, que contém o terceiro maior valor de correlação com as variáveis analisadas. A figura 4 mostra o modelo linear probabilístico generalizado [9] a partir da variável alvo e da variável com maior coeficiente de correlação, 0.46. Dado que a correlação é apenas média, o uso desta métrica estatística é pouco conclusivo em relação ao problema em questão.

TABELA 2. Análise de Correlação entre a Nota da Avaliação Somativa e as demais variáveis

Ordem[*]	**Variável**	**Coeficiente de Correlação**
4	Indicador de repetência	0.06
6	Percentual de presenças nas aulas presenciais	0.27
8	Percentual de avaliações formativas respondidas	0.44
10	Percentual de presenças	0.42
11	Nota média das avaliações formativas	0.23
12	Nota média ajustada das avaliações formativas	0.46[†]
15	Nota na avaliação formativa 1	0.27
16	Nota na avaliação formativa 2	0.36
17	Nota na avaliação formativa 3	0.35
18	Nota na avaliação formativa 4	0.42
19	Nota na avaliação formativa 5	0.45
20	Nota na avaliação formativa 6	0.18
21	Nota na avaliação formativa 7	0.18
22	Nota na avaliação formativa 8	0.34

[*] De acordo com a ordem apresentada na Tabela 1
[†] Mais alto Coeficiente de Correlação encontrado

É importante ressaltar que algumas variáveis foram eliminadas neste experimento, seja por conterem valores do tipo *string*, caso das variáveis de 1 a 3, seja por conterem informação redundante com alguma outra variável já analisada, caso dos seguintes pares de variáveis: 5 e 6; 7 e 8; 9 e 10.

Clusterização

Algoritmos de clustering são largamente utilizados em mineração de dados exploratória, além de ser uma tarefa comum em análise estatística. Na prática, a clusterização é o processo de agrupar objetos em subconjuntos de acordo com seu nível de similaridade, sendo uma forma eficiente de representação das características gerais da população sendo amostrada [10]. No contexto deste trabalho, foi aplicado o algoritmo de clusterização *k-means*, através da utilização da ferramenta *Weka* [11]. Este algoritmo foi aplicado com a finalidade de permitirem a identificação dos diferentes perfis de padrão de desempenho de alunos, com base nos dados coletados sobre seu desempenho nas diversas avaliações formativas.

Os resultados gerados com base na análise no percentual de frequência do aluno nas atividades presenciais (P), percentual de avaliações formativas respondidas (F), nota média ajustada nas atividades formativas (M) e nota da avaliação somativa (N), gerou dois *clusters* com as seguintes características médias para seus centróides:

1. P = 93%, F = 86%, M = 74% e N $\geq$ 6 (29 alunos).
2. P = 81%, F = 63%, M = 52% e N $<$ 6 (32 alunos).

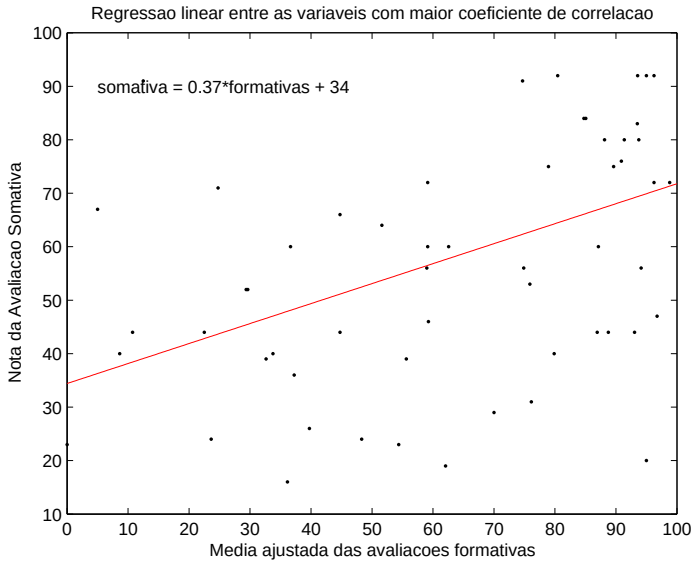

FIGURA 4. Média ajustada nas avaliações formativas x Nota na avaliação somativa

As figuras 5 e 6 mostram a distribuição gráfica das amostras dentro dos *clusters* sob duas perspectivas diferentes. Na figura 5 é possível verificar que há uma maior concentração de estudantes com rótulo aprovado na parte superior do eixo y, que representa a média das notas das avaliações formativas e, eventualmente, são encontrados alunos aprovados que tiveram um baixo desempenho neste quesito. Esta figura também permite identificar uma concentração de pontos próximos a uma diagonal imaginária, de forma que há uma tendência que alunos com maior assiduidade presencial tendem a ter um maior percentual de avaliações formativas respondidas, assim como uma maior média nas notas das avaliações formativas. Já a figura 6 permite evidenciar que também há uma maior concentração de alunos com rótulo aprovado com alto percentual de participação em atividades presenciais. Também é possível evidenciar que muitos estudantes reprovados tem elevada assiduidade e comprometimento com as avaliações formativas. Por outro lado, são mais raros os casos de estudantes aprovados com baixa assiduidade e que não respondam as avaliações formativas. Esta figura também demonstra a ausência de alunos que tenham um alto percentual de avaliações formativas respondidas aliada a um baixo percentual de assiduidade presencial. Adicionalmente pode ser percebida uma alta concentração de alunos com alto percentual de avaliações formativas respondidas e alta assiduidade presencial, o que pode ser explicado pelo fato da assiduidade mínima requerida para aprovação ser de 75% para ambas variáveis.

Classificação

A fim de identificar a tendência dos estudantes a terem sucesso na avaliação somativa, foram aplicados algoritmos de classificação sobre os dados disponíveis. Estes experimentos foram realizados na tentativa de predizer o conteúdo da variável 14, que contém o *Rótulo da nota da avaliação somativa*, que pode assumir os valores *aprovado* ou *reprovado*. Dentre os algoritmos de classificação disponíveis foram selecionados os que geram seus resultados na forma de árvores, tendo em vista a fácil interpretação de seus resultados. Dentre os algoritmos utilizados, são apresentados aqui apenas os resultados gerados pelo *REPTree* e pelo *J48* que geraram árvores com poucos nodos e maior capacidade preditiva. Todos os experimentos de classificação realizados utilizaram a técnica de validação cruzada *Leave-One-Out* para medir a capacidade de generalização dos modelos [11].

O algoritmo *REPTree* constrói árvores de decisão para classificação ou regressão com base no ganho de informação/variância e poda esta árvore usando uma poda guiada por erro [11]. A aplicação do algoritmo *REPTree* gerou a árvore mostrada na figura 7, que indica que se o Percentual de presenças totais for menor que 84,09%, o aluno deverá ser reprovado (Nota Somativa<6), caso contrário, ele deverá ser aprovado (Nota Somativa$\geq$6). Esta árvore permite a classificação correta de 60,5% dos estudantes, porém a estatística *Kappa* de 0,2137, que considera a qualidade da ár-

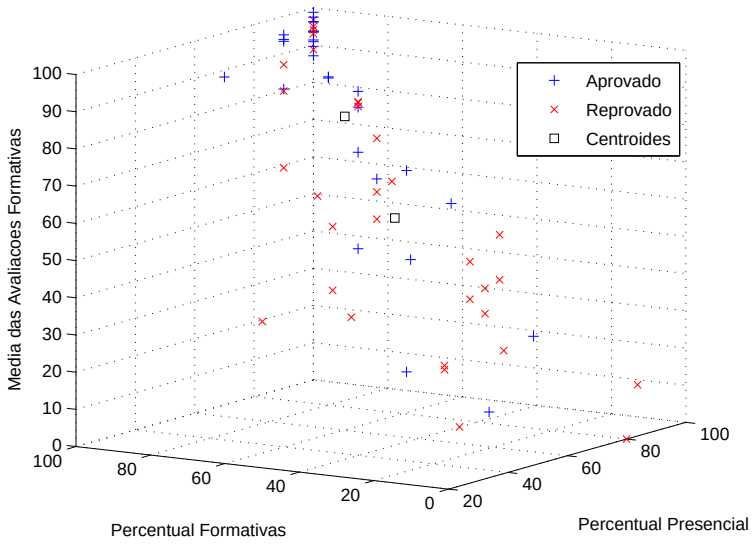

FIGURA 5. Distribuição gráfica das amostras

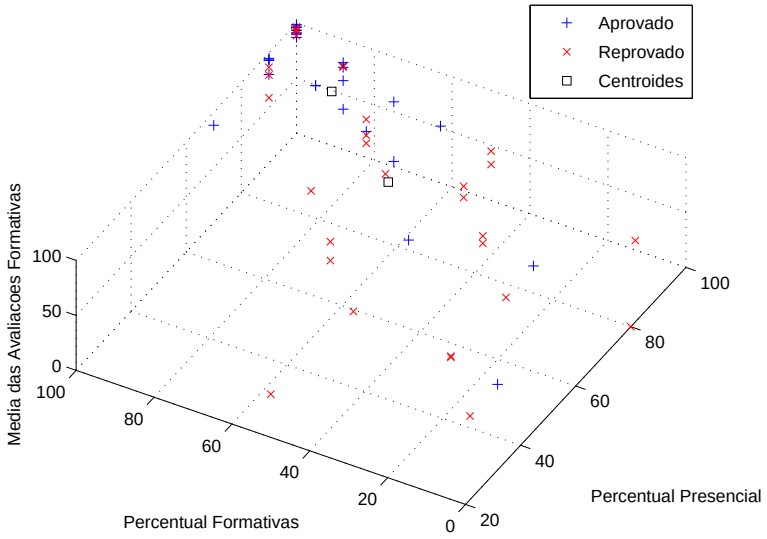

FIGURA 6. Distribuição gráfica das amostras

vore, é considerada apenas regular [12]. A estatística *Kappa* é uma medida de concordância usada em escalas nominais que fornece uma medida do quanto o resultado gerado pelo classificador se afasta de uma classificação meramente aleatória. De qualquer forma, fica muito claro que a assiduidade do estudante, tanto nos compromissos presenciais quanto não presenciais, é um dos aspectos relevantes para o seu sucesso na avaliação somativa. Adicionalmente é relevante mencionar que as árvores geradas atingiram em torno de 62% de sensibilidade e 60% de especificidade.

O outro algoritmo utilizado foi o *J48*, que é uma implementação do algoritmo C4.5, proposto por Quinlan [13], para geração top-down de árvores de decisão. Este algoritmo usa uma técnica de busca gulosa, determinando a cada passo qual o atributo, dentre os disponíveis, que é mais preditivo, e dividindo um nodo da árvore com base neste atributo [11]. Assim, a aplicação do algoritmo *J48* gerou a árvore apresentada na figura 8, que contém as seguintes

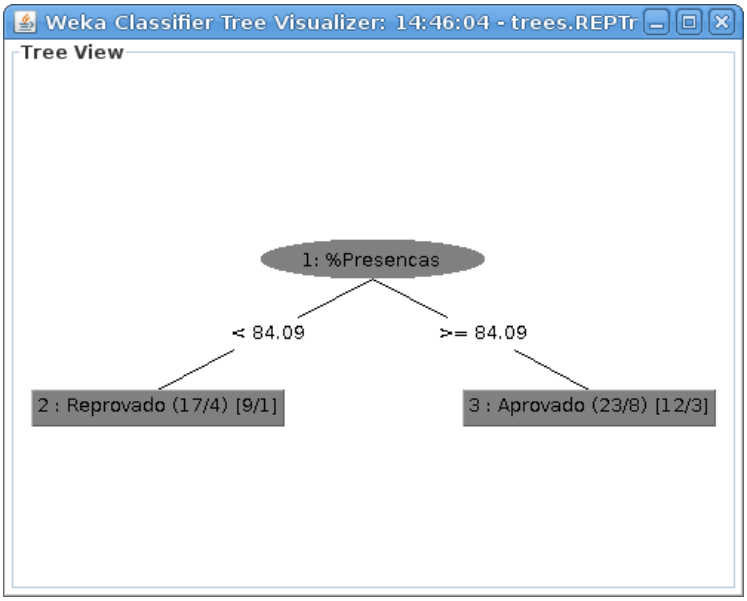

FIGURA 7. Árvore gerada pelo algoritmo *REPTree*

variáveis: Percentual total de presenças, Percentual de assiduidade presencial e Média ajustada das notas nas atividades formativas. Esta árvore permite a classificação correta de aproximadamente 69% dos estudantes. De acordo com a árvore gerada, estudantes que cumprem 72,73% ou menos dos seus compromissos presenciais e semipresenciais tendem a ser reprovados. Caso o valor tenha sido superior a este limiar, o próximo atributo a ser considerado seria o percentual de compromissos em atividades presenciais que, caso for maior que 90%, sugeriria a aprovação do aluno. Caso contrário, passa então a ser considerado limiar em torno de 80% de aproveitamento nas atividades formativas para indicar a aprovação ou reprovação do estudante. Adicionalmente é relevante mencionar que as árvores geradas atingiram em torno de 76% de sensibilidade, 62% de especificidade e a estatística de *Kappa* de 0,38. Desta forma, a árvore gerada pelo algoritmo *J48* apresenta resultados melhores que o *REPTree*.

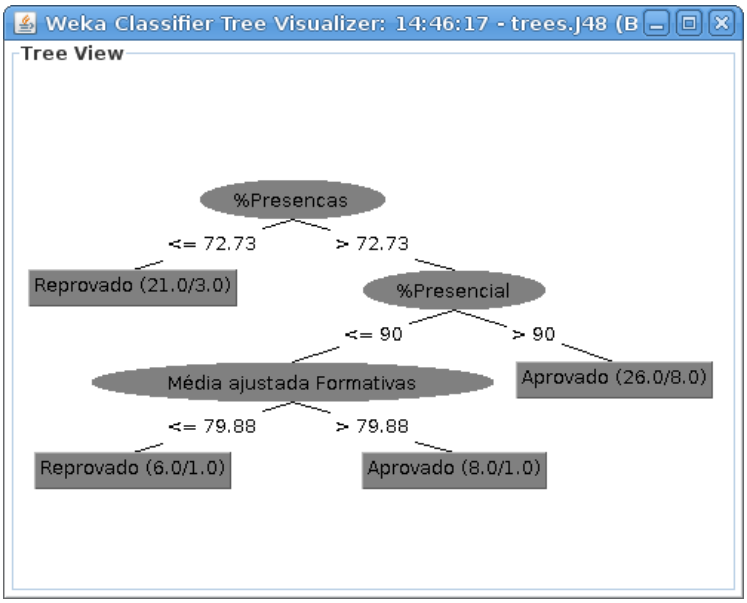

FIGURA 8. Árvore gerada pelo algoritmo *J48*

CONCLUSÕES E TRABALHOS FUTUROS

Os três grupos de experimentos realizados comprovam uma realidade esperada: *Quanto maior a dedicação nas atividades presenciais e semipresenciais, melhor o desempenho do aluno nas avaliações somativas*. Porém, apesar da realidade ser esperada, havia a carência de dados que pudessem comprová-la. Logo, os experimentos aqui relatados mostram que, independentemente da técnica utilizada, os resultados convergem para uma conclusão comum, sendo possível fazer uma predição do desempenho de um aluno, em termos de aprovação, em torno de 69% dos casos.

Desta forma, a fim de aumentar o nível de aprovação na disciplina, o conhecimento prévio sobre os fatores que contribuem para um baixo desempenho poderia ser explorado a fim de serem planejadas, nas semanas que antecedem a avaliação somativa, atividades de monitoria dirigidas ao grupo com tendência a não alcançar o desempenho requerido.

Os próximos trabalhos a serem executados incluem a formulação de estratégias que possam prevenir o baixo desempenho dos alunos na avaliação somativa para serem aplicadas no próximo semestre. O foco principal é, a partir da identificação de que um estudante tem uma alta propensão a ter um baixo desempenho na avaliação, oferecer atividades de adicionais, com supervisão do monitor da disciplina e com aulas adicionais de revisão, nas duas semanas que antecedem a avaliação somativa. Os resultados da avaliação somativa serão analisados a fim de identificar se as estratégias adotadas conduziram à melhora do desempenho dos alunos mais propensos a terem um desempenho não satisfatório.

REFERÊNCIAS

1. 2011 distance education survey results: Trends in e-learning: Tracking the impact of e-learning at community colleges, Tech. rep., ITC - Instructional Technology Council (2012).
2. Censo da educação superior 2010, Tech. rep., INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2011).
3. N. D. E. do Curso de Engenharia de Computação, Ppc - projeto pedagógico de curso, Tech. rep., Universidade Federal do Pampa (2010), URL `http://cursos.unipampa.edu.br/cursos/engenhariadecomputacao/files/2011/01/PPC_Completo_Revisado.pdf`.
4. B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch, "Predicting student performance: an application of data mining methods with the educational web-based system LON-CAPA," in *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference*, 2003.
5. N. Thai-Nghe, L. Drumond, A. Krohn-Grimberghe, and L. Schmidt-Thieme, *Procedia CS* **1**, 2811–2819 (2010).
6. Y. Wang, S. Cui, Y. Yang, and J. ao Lian, *Technology Interface Journal* **10** (2009).
7. D. Budny, and J. Tartt, "Do Engineering Students Fail Because they Don't Know How to Fail?," in *39th ASEE IEEE Frontiers in Education Conference*, 2009.
8. S. Parsons, "Overcoming High Failure Rates in Engineering Mathematics: A Support Perspective," in *Proceedings of the International Conference on Innovation, Good practice and Research in Engineering Education*, 2004, pp. 195–200.
9. D. C. Montgomery, and G. C. Runger, *Applied Statistics and Probability for Engineers*, John Wiley and Sons, New York, NY, 2003, 3 edn.
10. A. K. Jain, and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, New Jersey, 1948.
11. I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Burlington, MA, 2011, 3 edn.
12. J. R. Landis, and G. G. Koch, *Biometrics* **33**, 159–174 (1977).
13. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, 1993.