

Using social network analysis metrics of virtual forums to predict performance in e-learning courses

Henrique Lemos dos Santos
Instituto de Informática

Universidade Federal do Rio Grande do Sul
Porto Alegre, RS, Brasil.
Email:hlsantos@inf.ufrgs.br

Cristian Cechinel
Campus Araranguá

Universidade Federal de Santa Catarina
Araranguá, SC, Brasil.
Email:cristian.cechinel@ufsc.br

Ricardo Matsumura Araujo
CDTec

Universidade Federal de Pelotas
Pelotas, RS, Brasil.
Email: ricardo@inf.ufpel.edu.br

Emanuel Marques Queiroga
CaVG

Instituto Federal Sul-rio-grandense - IFSul
Pelotas, RS, Brasil.
Email: emanuelqueiroga@cavg.ifsul.edu.br

Abstract—The present article proposes the use of social network metrics extracted from forums interactions in distance education courses in order to predict students failing. Eight centrality metrics from forums were used as input information for training and testing five different classifiers able to early predict at-risk students. The initial findings indicate these attributes are informative and useful for prediction, however predictive models performance vary considerably across courses and depending on the amount of data collected.

Index Terms—social networks, EDM, sistemas híbridos.

I. INTRODUÇÃO

Os cursos superiores no formato de Educação a Distância (EAD) têm crescido significativamente no Brasil durante os últimos anos, especialmente porque a EAD é vista por muitos estudantes como uma modalidade de ensino flexível e cuja qualidade é frequentemente semelhante ao método presencial [1]. Outro aspecto importante para o desenvolvimento de cursos EAD no Brasil foi a criação da Universidade Aberta do Brasil (UAB)¹ em 2006 e da rede e-Tec² em 2007, somente em 2015 e 2016 a UAB foi responsável pela entrada de 250.000 estudantes em cursos superiores EAD espalhados por mais de 550 cidades em diversas regiões do Brasil. Somam-se a esses fatos os recentes estudos, nacionais e internacionais, que constataram desempenhos comparáveis e até mesmo superiores de estudantes de ensino EAD em relação a estudantes provenientes do ensino presencial [2] [3].

Apesar dessas notáveis qualidades, a EAD ainda enfrenta problemas significativos no que diz respeito a evasão e reprovação de estudantes, conforme apontado no Censo EAD realizado em 2015 [4]. O censo contou com a participação de 339 instituições formadoras e contabilizou um total de 5.048.912 alunos matriculados em cursos EAD no Brasil, um

aumento de 30,5% em relação ao ano anterior. Entretanto, o Censo EAD também reportou que 40% das instituições que ofereceram cursos totalmente a distância apresentaram taxas de evasão na casa dos 26%-50% e 7% dessas instituições apresentaram índices de evasão acima de 51%, enquanto que apenas 0,73% das instituições de cursos presenciais apresentaram uma taxa de evasão acima de 50%. Esses dados reafirmam a tendência de expansão da EAD no Brasil mas também expõem suas dificuldades.

Para enfrentar esses problemas, pode-se tirar proveito do inerente viés digital do processo de ensino EAD, que permite o fácil armazenamento e posterior análise dos dados produzidos por alunos, professores e tutores durante a execução de aulas e atividades. Em geral, esse processo se dá em torno de um Ambiente Virtual de Aprendizagem (AVA). Alguns dos AVAs mais utilizados no Brasil não apenas na EAD mas também como suporte à educação presencial são: TelEduc, Moodle, AulaNet e SOLAR. Entre outras funcionalidades administrativas e de coordenação, os AVAs disponibilizam uma ou mais formas de comunicação aos seus usuários, tais como: fóruns de discussão, salas de bate-papo, mensagens privadas, entre outras [5].

Dadas as facilidades proporcionadas pelos AVAs na geração e captura de dados educacionais, duas áreas de pesquisa desenvolveram-se substancialmente nos últimos anos a fim de prover soluções e análises para o processo de ensino (seja ele a distância ou presencial): a Mineração de Dados Educacionais (MDE), de acordo com [6], consiste na exploração de técnicas de estatística, mineração de dados e aprendizagem de máquina em dados educacionais; já a Analítica da Aprendizagem (AL) é definida pela *Society for Learning Analytics Research - SoLAR*³ como sendo ” (...) a medição, coleta, análise e geração de relatórios sobre dados de estudantes e seus contextos, com objetivo de entender e otimizar o processo de ensino e o

¹<http://www.capes.gov.br/uab>

²<http://portal.mec.gov.br/rede-e-tec-brasil>

³<https://solaresearch.org/>

ambiente no qual ele ocorre.”

Unindo o aspecto de adaptação automatizada característico da MDE e o foco na provisão de informação valiosa a professores e tutores característico da AL [7], o presente trabalho tem como objetivo principal verificar a viabilidade do uso de interações entre alunos, tutores e professores em fóruns de disciplinas de cursos superiores EAD para construção de redes sociais, que por sua vez terão métricas extraídas e utilizadas para treinar algoritmos de classificação, com o objetivo de identificar alunos em risco de evasão/reprovação⁴ durante o andamento da disciplina. Essa identificação, quando feita com antecedência, pode ser útil a professores e tutores que desejam oferecer intervenções a estudantes em risco. Além desse propósito inicial, deseja-se também comparar o desempenho de diferentes algoritmos classificadores na tarefa de prever a situação final de estudantes a partir de métricas sociais.

Esse trabalho está organizado da seguinte forma. A Seção II trará um breve compêndio de trabalhos correlatos nacionais e internacionais, que focaram tanto na questão de predição de desempenho de alunos de EAD de maneira geral como também no uso específico de métricas sociais para possibilitar a identificação de alunos em risco. Na Seção III os dados utilizados serão descritos assim como o passo-a-passo desde a construção das redes sociais até a execução e validação dos classificadores. Os resultados obtidos pelos algoritmos selecionados serão apresentados e discutidos na Seção IV e, por fim, as conclusões sobre a presente pesquisa e uma listagem de possibilidades de trabalhos futuros serão desenvolvidas na Seção V.

II. TRABALHOS RELACIONADOS

A tentativa de modelagem comportamental de estudantes para determinar seu desempenho futuro tem sido bastante abordada nos últimos anos. O amplo espectro de possibilidades de dados preditores (variáveis de entrada para os modelos de classificação) também ajuda a impulsionar a diversidade de trabalhos desenvolvidos nessa área. Além disso, o uso de análise de redes sociais (ARS) para otimização do processo de *e-learning* aumentou significativamente nos últimos anos, como pode ser visto na revisão elaborada por [8]. No restante dessa seção citaremos alguns exemplos de trabalhos voltados a esse objetivo, que utilizaram diversos tipos de dados, incluindo dados sociais, tal qual o presente trabalho.

Dentre esses trabalhos relacionados, um dos pioneiros foi desenvolvido por [9]. Nele os autores utilizaram dados estudantis (presença, avaliações, etc.) do curso EAD de Informática da *Hellenic Open University* para alimentar algoritmos de aprendizagem de máquina com objetivo de prever o desempenho final dos alunos. A acurácia alcançada com o classificador Naive Bayes foi a melhor, atingindo 83% antes da metade do período acadêmico. Mais recentemente, [10] coletaram dados navegacionais (cliques), demográficos e avaliativos do AVA de 3 cursos EAD da *Open University*⁵. Com-

binando esses dados, os autores conseguiram prever com até 75% de precisão a desistência de alunos dos seus cursos. Além disso, os autores identificaram que cada um dos 3 tipos de dados tem maior impacto na predição de acordo com o período do curso que está sendo analisado.

No cenário nacional, destaca-se o trabalho de [11], que propõe o uso de interações de estudantes dentro do AVA e de seus atributos derivados como variáveis preditoras de desempenho de um curso EAD. Os autores identificaram que o uso da contagem de interações (por exemplo: visualizações de postagens, mensagens postadas, etc.) em conjunto com atributos derivados (média, mediana, etc.) fez com que algoritmos de classificação, em especial redes bayesianas, obtivessem uma precisão de 0.71 na predição de reprovação dos alunos já na terceira semana de execução do curso. Com foco específico na predição de evasão de alunos EAD ainda pode-se destacar o trabalho de [12] no qual os autores verificaram que a simples contagem de interações dos alunos ao longo do total de semanas de dois cursos EAD foi suficiente para algoritmos de classificação atingirem até 98.67% de acurácia na identificação de alunos evadidos (taxa de verdadeiros negativos), na metade de um dos cursos analisados.

No que diz respeito ao uso de dados sociais como variáveis preditoras de desempenho estudantil, destaca-se o recente trabalho de [13] no qual os autores extraíram 3 centralidades (Grau, Intermediação e Proximidade) da rede social gerada pelas interações entre alunos em um fórum do AVA Moodle e verificaram a existência de correlações lineares (fortes e médias) entre as centralidades e os desempenhos finais dos alunos de um curso técnico de uma instituição de ensino federal. Por fim, ressalta-se o trabalho de [14], onde os autores utilizaram conexões sociais entre estudantes em um fórum do AVA bem como informações qualitativas geradas a partir da análise sobre o conteúdo das mensagens trocadas entre os alunos. Essa análise qualitativa permitiu uma seleção prévia de dados que compuseram a entrada de algoritmos classificadores e de agrupamento, resultando em uma acurácia de cerca de 82% para predição da performance final dos estudantes, utilizando dados coletados até a metade do curso.

III. DESCRIÇÃO DOS DADOS E METODOLOGIA

A. Descrição dos dados

Para a realização desse trabalho foram obtidos e anonimizados dados provenientes dos AVAs (Moodle) do curso de Licenciatura em Educação do Campo (CLEC), da Universidade Federal de Pelotas (UFPel), e dos cursos Técnico em Administração (ADM), Técnico em Agroindústria (AGRO), Técnico em Biocombustíveis (BIO) e Técnico em Contabilidade (CONT), do IFSul - Campus Visconde da Graça. Uma importante diferença em relação a disciplinas de cursos presenciais é que as disciplinas desses cursos EAD ocorrem em um período mais compactado, que normalmente compreende de 5 a 7 semanas. Esses dados consistiram nas mensagens trocadas nos fóruns dos AVAs e a situação final de cada aluno na disciplina (variável binária - conclusão com sucesso ou não). Essa situação final foi extraída de registros acadêmicos

⁴No presente trabalho ambos desfechos são tratados de maneira conjunta.

⁵<http://www.open.ac.uk/>

e da presença do aluno em disciplinas no semestre seguinte⁶ no caso do curso CLEC da UFPel, e do próprio Moodle (nota das atividades) no caso dos cursos do IFSul. As disciplinas do curso CLEC utilizadas no presente trabalho foram executadas no 1º e 2º semestre de 2013, enquanto que as disciplinas dos 4 cursos do IFSul ocorreram entre o 2º semestre de 2013 e o 2º semestre de 2014.

A Tabela I apresenta uma descrição quantitativa dos cursos. Nessa tabela ficam visíveis algumas diferenças entre as duas instituições no que diz respeito ao gerenciamento dos cursos EAD. Enquanto na UFPel há um alto número de professores cadastrados nos AVAs das disciplinas, as disciplinas do IFSul contém quase sempre apenas um professor cadastrado e uma série de tutores. Outra diferença importante ocorre nos percentuais médios de reprovação (soma dos percentuais em cada disciplina do curso dividida pelo número de disciplinas do curso): enquanto na UFPel apenas cerca de 14.5% dos alunos reprovaram ou evadiram, no IFSul esse número gira em torno de 47% a 59%, dependendo do curso. Além disso, destaca-se o maior uso dos fóruns como ferramenta de comunicação pelo curso CLEC que, embora possua o menor número de alunos matriculados, possui também o maior número de interações (ou conexões) únicas coletadas durante 7 semanas.

O padrão decrescente, ao longo das semanas, de interações nos fóruns das disciplinas pode ser observado na Figura 1. Na figura cada linha representa a média das N disciplinas de cada curso.

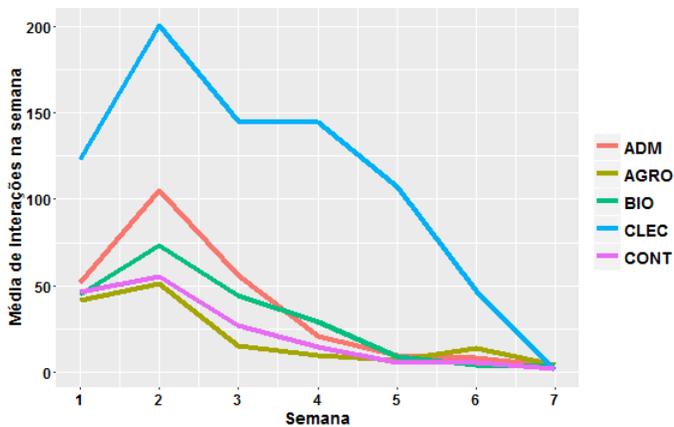


Fig. 1. Distribuição média das interações coletadas ao longo de 7 semanas para disciplinas dos 5 cursos

Uma interação única pode ser definida como uma conexão direcional estabelecida por meio de uma mensagem no fórum (ou resposta, por exemplo, do aluno x para o professor y). Por ser única, mesmo que um determinado aluno tenha enviado diversas mensagens, em tempos diferentes, a um mesmo professor (ou outro aluno) nos fóruns de uma mesma disciplina, a interação foi contada apenas uma vez, para fins de descrição quantitativa.

⁶No formato EAD adotado pela UFPel, se um aluno reprova em uma disciplina, ele será desligado do curso no semestre seguinte.

B. Geração das redes sociais e extração de métricas

O experimento realizado focou na criação de uma rede social (RS) a cada semana para cada disciplina, durante as 7 semanas nas quais a disciplina foi executada. Dessa forma, ao final foi possível verificar com quanta antecedência, em termos de semanas, os algoritmos classificadores conseguiram atingir um resultado satisfatório. Além disso, essas RS foram geradas de maneira cumulativa, isto é, por exemplo, a RS da semana 2 de uma disciplina contém todas as interações que ocorreram até o final da semana 2 nos fóruns dessa disciplina.

Para o presente trabalho, todos os usuários (alunos, professores e tutores) foram tratados de maneira equivalente, sem atributos extras devido a sua classe; esses usuários consistiram nos **vértices** das RS geradas. As **arestas** (ou conexões) por sua vez foram geradas a partir das respostas, nos fóruns, dadas pelos usuários a outros usuários; todas as arestas tiveram como atributos a direção e o peso. A direção diferencia uma resposta dada pelo usuário x ao usuário y da resposta do usuário y ao usuário x . Para calcular o peso de cada conexão optou-se por utilizar a simples contagem de palavras contidas nas respostas coletadas a fim de diminuir a importância de mensagens diminutas (por exemplo, saudações ou confirmações). Ainda, a fim de valorizar conexões que comprovam maior vínculo entre os envolvidos, a soma de conexões que ocorreram em duas semanas consecutivas foi feita com um bônus de 0.25 para o peso da última conexão, para exemplificação: se houve uma conexão $x \rightarrow y$ com peso 50 na semana 1 e na semana 2 a conexão se repetiu com um peso 70, então a RS da semana 2 conterá $x \rightarrow y$ com peso $50 + 70 * 1.25 = 137.50$, onde 0.25 acrescido ao 1.0 é o bônus de peso por interação consecutiva em 2 semanas. Esse bônus se mantém o mesmo a cada sequência de 2 semanas consecutivas e baseia-se parcialmente no trabalho de [15] onde os autores ressaltam o desenvolvimento de "cadeias de interações" (seqüências) como elemento particularmente importante para se determinar a efetividade de interações em grupos *online*.

A partir das redes sociais semanais de cada disciplina foram então extraídas 8 diferentes métricas de centralidade. Em geral, a centralidade visa identificar nodos mais ativos na rede [13] quantificando sua importância em função de algum cálculo sobre suas conexões e os atributos destas. As centralidades utilizadas nesse trabalho foram (centralidades marcadas por um * consideraram o peso das arestas no seu cálculo): C1 - Centralidade de grau clássica*, C2 - Centralidade de grau dos k -caminhos mínimos ($k=2$), C3 - Centralidade de grau de entrada*, C4 - Centralidade de grau de saída*, C5 - Centralidade de intermediação, C6 - Centralidade de proximidade*, C7 - Centralidade de autovetor e C8 - PageRank*.

C. Classificadores

Os 7 arquivos contendo as 8 métricas sociais para cada aluno de uma determinada disciplina todavia precisaram passar por um processo de reamostragem uma vez que, em geral, a quantidade de alunos aprovados era significativamente maior (em algumas disciplinas essa diferença ultrapassou 100%) do que a quantidade de alunos de reprovados. Como exposto

TABELA I
DESCRIZAÇÃO QUANTITATIVA GERAL DOS DADOS

Instituição	Curso	Disciplinas (Polos)	Professores	Tutores	Alunos	Percentual médio de reprovação ou evasão (%)	Interações únicas coletadas (em 7 semanas)
[?]	CLEC	10 (4)	33	19	141	14.5	5482
	ADM	16 (18)	15	122	900	47.1	3500
[?]	AGRO	15 (12)	14	92	1221	51.3	1900
	BIO	12 (15)	12	121	1357	59.8	2226
	CONT	15 (12)	14	122	1187	53.7	1910

por [16], não é apropriado avaliar o desempenho de algoritmos classificadores em *datasets* desbalanceados, nesse artigo os autores também propuseram o algoritmo de reamostragem SMOTE, utilizado no presente trabalho por meio da API Weka⁷. Todos os *datasets* foram reamostrados de maneira que ao final houvesse um balanceamento perfeito (50% de cada classe - aprovados e reprovados). Para a etapa de classificação, também realizada com auxílio da API Weka, foram escolhidos algoritmos utilizados por [11] e por [12], sendo eles: *Bayes Net*, *J48*, *REPTree*, *Random Forest*, e *Multilayer Perceptron* (MLP). Os modelos gerados para os 5 algoritmos foram testados apenas dentro da disciplina onde foram treinados em uma determinada semana. O método escolhido foi a validação cruzada, com número de *folds* igual a 10.

IV. RESULTADOS E DISCUSSÕES

Apesar do presente trabalho ter interesse na predição de insucesso de alunos EAD, é importante que os classificadores tenham não apenas altas taxas de verdadeiro-positivo como também baixas taxas de falso-positivo. Nesse sentido utilizou-se como métrica de avaliação a *área sob a curva ROC* (*Receiver Operating Characteristic*). A curva ROC é um gráfico de sensibilidade (taxa de verdadeiros positivos - TVP) versus 1-especificidade (taxa de falsos positivos - TFP) para os diferentes valores de corte de um teste. Considerando a classe de alunos que evadiram ou reprovaram, a TVP é dada pela divisão do número de alunos com estes desfechos (reprovaram ou evadiram) e foram classificados corretamente pelo número total de alunos com o desfecho de fato, já a TFP corresponde a divisão do número de alunos que não apresentaram o desfecho mas foram classificados como tal pelo número total de alunos que concluíram com sucesso. A API do Weka varia os diferentes pontos de corte de uma curva ROC a partir das estimativas que as instâncias têm de pertencer a uma dada classe, nesse caso, classe de alunos que reprovaram ou evadiram⁸. Já a *área sob a curva ROC* dimensiona de maneira numérica, a partir de uma curva ROC, a capacidade discriminativa do classificador, isto é, a classificação correta dos alunos que apresentaram o desfecho negativo e também dos que finalizaram a disciplina com sucesso. Como referência, pode-se admitir que valores para

a *área sob a curva ROC* são excelentes entre 0.9 e 1, bons entre 0.8 e 0.9, razoáveis entre 0.7 e 0.8, ruins entre 0.6 e 0.7, e fracassados entre 0.5 e 0.6 [17].

Na Figura 2 são apresentados os resultados médios para cada curso. É importante ressaltar que os resultados mostrados estão agrupados, em forma de médias, para os 5 cursos dos quais se obteve os dados. Dessa forma, algumas disciplinas apresentaram áreas significativamente diferentes das que estão expostas nessa figura inicial. Assim, é possível perceber que o curso CLEC foi o curso cujas disciplinas proporcionaram o melhor desempenho médio para os classificadores, em especial o algoritmo *Random Forest*, que para esse curso, na semana 4 atingiu uma média na área sob a curva ROC de 0.9375. Para esse mesmo curso, outros algoritmos (*REPTree* e *Bayes Net* também destacaram-se, com áreas consideradas boas (entre 0.85 e 0.9) também já na semana 4, logo após a disciplina ter atingido 50% dos seus dias letivos.

Os 4 cursos do IFSul (ADM, AGRO, BIO e CONT), todavia, apresentaram um desempenho médio inferior. Isso se deu muito em função da existência de disciplinas que possuíam pouquíssimas interações nos seus fóruns, fazendo com que as RS geradas fossem extremamente esparsas e, por consequência, suas métricas de centralidade inócuas de informação relevante para os classificadores. Apesar disso, destaca-se que mesmo nesses cursos o algoritmo *Random Forest* obteve os melhores resultados. Por exemplo, no curso BIO esse algoritmo já a partir da semana 3 atingiu um desempenho médio de 0.80 (bom). Já no curso CONT, na semana 4 o classificador *Random Forest* obteve uma área média de 0.76, considerada razoável de acordo com a escala de referência.

Como dito anteriormente, os algoritmos apresentam resultados bastante discrepantes dentro de cada curso, devido a diferença no uso dos fóruns em cada disciplina. Para exemplificar esse fato, focamos nos dois algoritmos que obtiveram melhor desempenho médio (Fig. 2) e em dois cursos (ADM e CLEC). Na Figura 3 pode-se observar o diferente comportamento (em forma de desvio padrão) dos algoritmos para o curso onde as disciplinas tiveram muitas diferenças entre si (ADM) e para o curso onde as disciplinas foram mais homogêneas entre si (CLEC). Essas diferenças se deram tanto na questão de uso dos fóruns quanto no resultado dos classificadores, uma vez que este é causa direta de métricas

⁷<http://www.cs.waikato.ac.nz/ml/weka/>

⁸<https://weka.wikispaces.com/ROC+curves>

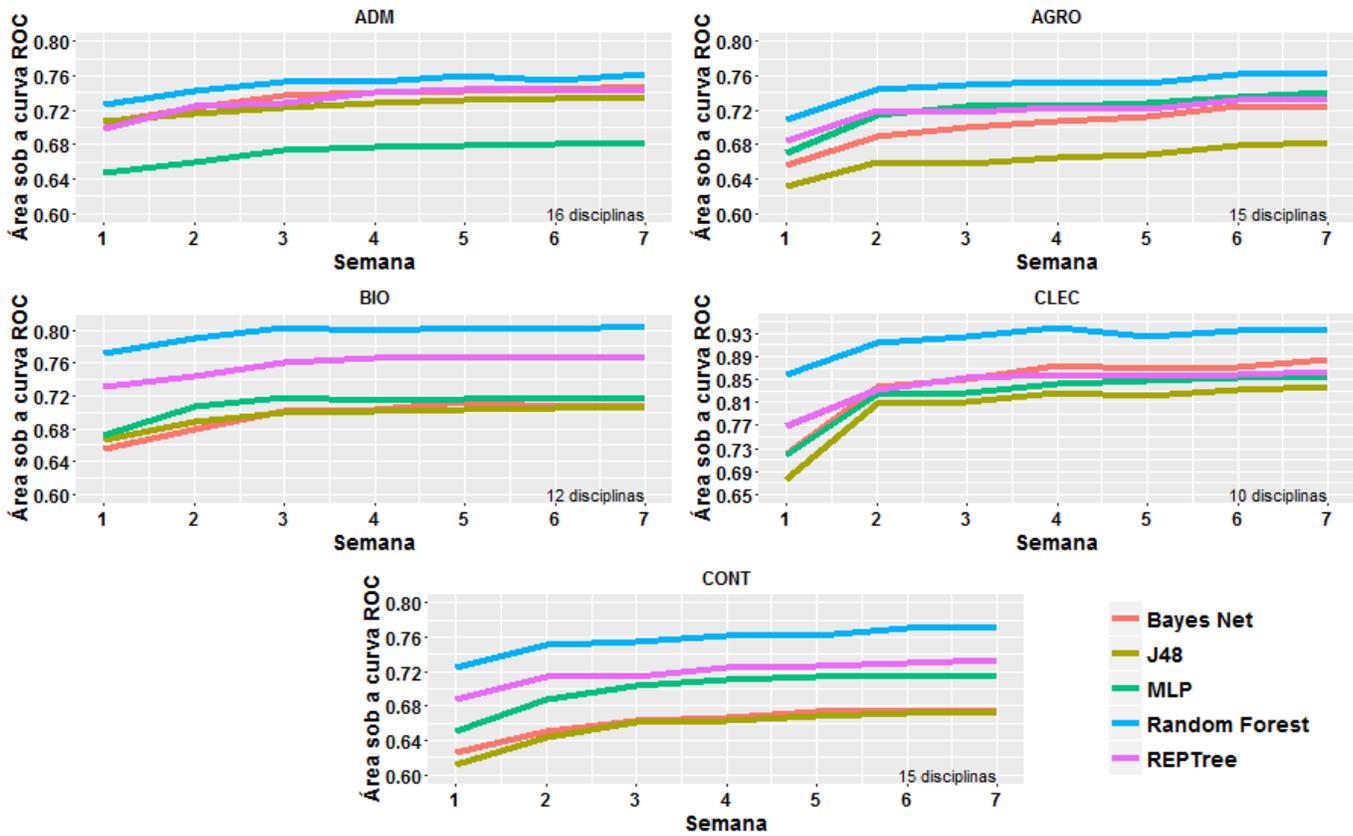


Fig. 2. Resultados médios gerais obtidos por curso

que dependem do uso dos fóruns.

Analisando o melhor desempenho de cada algoritmo, a cada semana, é possível perceber, a partir da Tabela II, que mesmo os algoritmos que tiveram desempenhos médios inferiores, J48 e MLP, conseguiram obter bons resultados isolados, na semana 1 e 3, respectivamente. O algoritmo J48 na primeira semana de uma disciplina do curso BIO alcançou uma TVP de 0.950 para alunos aprovados e 0.875 para alunos reprovados. Já na terceira semana, as redes neurais geradas pelo MLP em uma disciplina do curso CONT atingiram uma TVP de 0.994 para alunos aprovados e 0.962 para alunos reprovados. Nas semanas seguintes analisadas, 5 e 7, o algoritmo *Random Forest* superou os demais no que diz respeito ao melhor desempenho isolado.

V. CONSIDERAÇÕES FINAIS

Neste trabalho foi apresentada uma metodologia para predição de evasão ou reprovação em disciplinas de cursos de educação a distância a partir de métricas de redes extraídas de interações registradas entre alunos, professores e tutores. Estas interações foram agrupadas em semanas e modeladas como grafos direcionados e valorados representando a rede social subjacente. Oito métricas de centralidade foram então extraídas para cada aluno de cada um dos cursos, sendo utilizadas para treinar classificadores com o objetivo de prever evasão ou reprovação.

Em testes em cinco cursos reais, estas métricas se mostraram bastante informativas e, mesmo sozinhas, foram capazes de prever com boa sensibilidade e especificidade a ocorrência do desfecho (reprovação ou evasão). Observou-se que o modelo baseado em *Random Forest* foi em média o que apresentou melhor desempenho e que os resultados melhoram substancialmente quando mais informações são utilizadas para calcular as métricas. No entanto, também observou-se que o desempenho varia consideravelmente com a disciplina, refletindo as diferentes formas com que as interações ocorrem em cada uma. Em particular, disciplinas onde há pouca atividade nos fóruns resultam em redes esparsas pouco informativas. Mesmo nestes casos, a predição mostrou-se melhor que escolha aleatória.

A principal contribuição deste trabalho encontra-se na comprovação experimental da utilidade de métricas de redes sociais para predição de evasão ou reprovação. Estas métricas tem como vantagem adicionar informação estrutural a modelos mais simples, e.g. contagem de interações [?], sem exigir grande especificidade sobre o ambiente onde está sendo aplicado (i.e. não há distinção sobre o *tipo* ou o *conteúdo* da interação).

Trabalhos futuros incluem analisar com detalhes cada uma das métricas de centralidade para verificar quais são mais informativas e se há variabilidade entre cursos; comparar diretamente a metodologia proposta com outras técnicas aplicadas

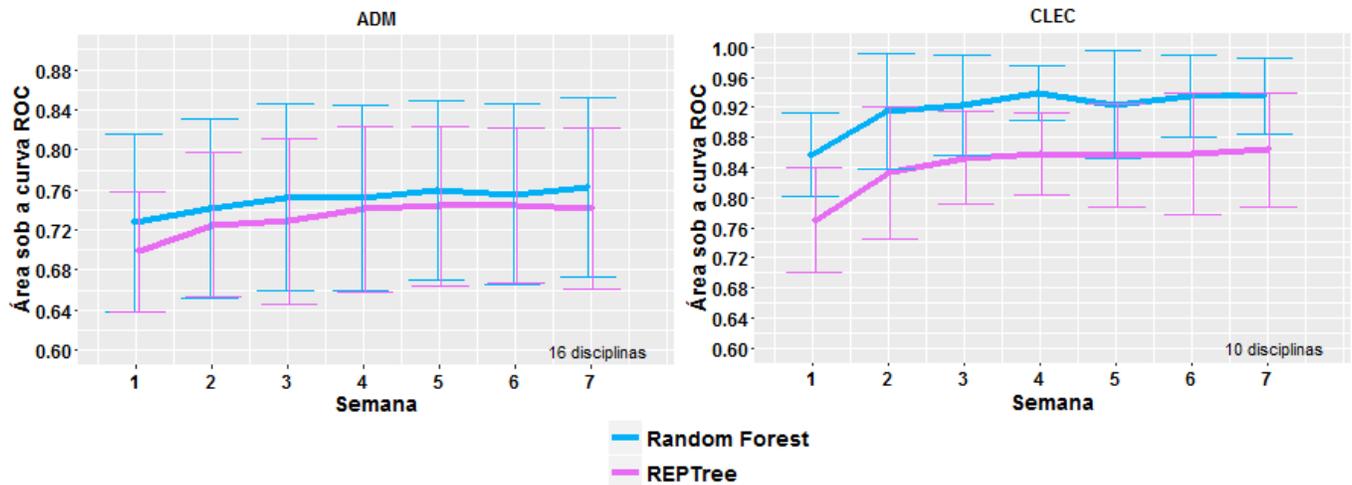


Fig. 3. Média e desvio padrão para os resultados dos 2 melhores algoritmos em 2 cursos

TABELA II
MELHORES RESULTADOS DE CADA ALGORITMO PARA SEMANAS 1, 3, 5 E 7

Algoritmo	Semana 1		Semana 3		Semana 5		Semana 7	
	TVP-APR	TVP-REP	TVP-APR	TVP-REP	TVP-APR	TVP-REP	TVP-APR	TVP-REP
<i>Bayes Net</i>	0.533	0.963	0.983	0.965	0.980	0.850	0.990	0.992
J48	0.950	0.875	0.977	0.965	0.986	0.954	0.970	0.908
MLP	0.830	0.978	0.994	0.962	0.994	0.962	0.994	0.962
<i>Random Forest</i>	0.879	0.904	0.961	0.947	0.990	0.980	1.0	0.990
<i>REPTree</i>	0.850	0.930	0.942	0.959	1.0	0.930	1.0	0.950

na mesma base de dados; e desenvolver meta-modelos que permitam decidir quando a metodologia pode ser aplicada, já que foi demonstrada uma grande variabilidade entre cursos e disciplinas. Adicionalmente, é interessante integrar as métricas aqui utilizadas em outros modelos que utilizam outros atributos (e.g. contagem de interações).

VI. AGRADECIMENTOS

O presente trabalho foi realizado com apoio do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) por meio do Edital Universal 01/2016 processo 404369/2016-2 (Projeto: “Comparação entre Diferentes Abordagens na Modelagem e Identificação de Acadêmicos em Risco em Cursos de Educação a Distância”).

REFERENCES

- [1] T. L. Russell, *The no significant difference phenomenon: As reported in 355 research reports, summaries and papers*. North Carolina State University, 1999.
- [2] J. Silva, F. Oliveira, and L. Mourão, “Uma comparação entre cursos a distância e presencial,” in *18º CIAED - Congresso Internacional ABED de EaD - Histórias, Análíticas e Pensamento Aberto – Guias para o Futuro da EAD*, São Luís, Maranhão, 2012.
- [3] K. Swan, “Learning effectiveness online: What the research tells us,” *Elements of quality online education, practice and direction*, vol. 4, pp. 13–47, 2003.
- [4] ABED, “Censo ead.br - relatório analítico da aprendizagem a distância no brasil,” 2015, 2015. Disponível em: <http://www.abed.org.br/arquivos/Censo_EAD_2015_POR.pdf>. Acesso em janeiro de 2017.
- [5] M. Gonzalez, *Fundamentos da Tutoria em Educação a Distância*. São Paulo, SP, Brasil: Avercamp, 2005.
- [6] C. Romero and S. Ventura, “Educational data mining: a review of the state of the art,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 40, no. 6, pp. 601–618, 2010.
- [7] G. Siemens and R. S. J. d. Baker, “Learning analytics and educational data mining: Towards communication and collaboration,” in *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge*, ser. LAK '12. New York, NY, USA: ACM, 2012, pp. 252–254. [Online]. Available: <http://doi.acm.org/10.1145/2330601.2330661>
- [8] K. L. Cela, M. Á. Sicilia, and S. Sánchez, “Social network analysis in e-learning environments: A preliminary systematic review,” *Educational Psychology Review*, vol. 27, no. 1, pp. 219–246, 2015.
- [9] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, “Predicting students’ performance in distance learning using machine learning techniques,” *Applied Artificial Intelligence*, vol. 18, no. 5, pp. 411–426, 2004.
- [10] A. Wolff, Z. Zdrahal, A. Nikolov, and M. Pantucek, “Improving retention: Predicting at-risk students by analysing clicking behaviour in a virtual learning environment,” in *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, ser. LAK '13. New York, NY, USA: ACM, 2013, pp. 145–149. [Online]. Available: <http://doi.acm.org/10.1145/2460296.2460324>
- [11] C. Cechinel, R. Araujo, and D. Detoni, “Modelling and prediction of distance learning students failure by using the count of interactions,” *Brazilian Journal of Computers in Education*, vol. 23, no. 03, p. 1, 2015. [Online]. Available: <http://br-ie.org/pub/index.php/rbie/article/view/3306>
- [12] E. M. Queiroga, C. Cechinel, R. M. Araújo, and G. da Costa Bretanha, “Generating models to predict at-risk students in technical e-learning

courses,” in *2016 XI Latin American Conference on Learning Objects and Technology (LACLO)*, Oct 2016, pp. 1–8.

- [13] J. Cássia, A. Brito, and F. Medeiros, “Aplicação de métricas da análise de redes sociais como apoio a avaliação das interações discentes em fóruns de discussão online de um ambiente virtual de aprendizagem,” in *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, vol. 5, no. 1, 2016, p. 467.
- [14] C. Romero, M.-I. López, J.-M. Luna, and S. Ventura, “Predicting students’ final performance from participation in on-line discussion forums,” *Computers & Education*, vol. 68, pp. 458–472, 2013.
- [15] A. Calvani, A. Fini, M. Molino, and M. Ranieri, “Visualizing and monitoring effective interactions in online collaborative groups,” *British Journal of Educational Technology*, vol. 41, no. 2, pp. 213–226, 2010. [Online]. Available: <http://dx.doi.org/10.1111/j.1467-8535.2008.00911.x>
- [16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [17] R. H. El Khouli, K. J. Macura, P. B. Barker, M. R. Habba, M. A. Jacobs, and D. A. Bluemke, “Relationship of temporal resolution to diagnostic performance for dynamic contrast enhanced mri of the breast,” *Journal of Magnetic Resonance Imaging*, vol. 30, no. 5, pp. 999–1004, 2009.