Predição de reprovação na educação a distância: um estudo comparativo

Patrícia Takaki Neves[1], Seiji Isotani [2], Bruno Elias Penteado[3]

Resumo

Este trabalho teve como problema de pesquisa predizer a reprovação de alunos de um curso superior a distância de modo a viabilizar intervenções pedagógicas com antecedência. O objetivo foi investigar empiricamente 6 algoritmos de classificação, com 3 opções de balanceamento de dados utilizando 2 conjuntos normalizados de dados (com notas de 3 e 5 semanas de aula, de um total de 6 semanas) e, por fim, comparar 5 métricas selecionadas. Com o software WEKA foram realizados 36 testes. Os resultados mostraram que o SVM (Máquina de Vetor de Suporte) obteve os melhores valores de acurácia (94,1%) e índice kappa *(0,868)(ambos com o algoritmo SMOTE de balanceamento) e de especificidade (99%)(sem balanceamento). As melhores sensibilidade (70,6%) e* g-means *(81,3%) ficaram com o DT (*Decision Table*) (com* Class Balancer*). Dados de 3 semanas (50% da disciplina) apresentaram resultados próximos daqueles obtidos com 5 semanas. Análises e possíveis aplicações são apresentadas visando minimizar as reprovações dos alunos.*

Abstract

The present study had as a research problem to predict the failure of students of a higher education course in the distance in order to allow pedagogical interventions previously. The objective was to empirically investigate 6 classification algorithms, with 3 data balancing options and using 2 normalized datasets *(with grades of 3 and 5 weeks of class, out of a total of 6 weeks) and, finally, compare 5 metrics selected. With the WEKA software, 36 tests were performed. The results showed that the SVM (Support Vector Machine) obtained best values for accuracy (94.1%) and* kappa *index (0.868) (both with the SMOTE balancing algorithm) and specificity (99%) (without balance); getting the best sensitivity (70.6%) and* g-means *(81.3%) with the DT (*Decision Table*) model (with* Class Balancer*). The 3-week (50% of the discipline) presented results close to those obtained at 5 weeks. Analyses and possible applications are presented in order to minimize failure of students.*

[1]Docente do DCC/CCET/UNIMONTES, Doutoranda no PGCIn/UFSC, Mestre em Ciência da Com-putação pelo IC/UNICAMP, Especialista em Computação Aplicada à Educação pelo ICMC/USP, patri-cia.takaki@usp.br

[2]Orientador1, Professor Titular do ICMC/USP, sisotani@icmc.usp.br

[3]Orientador2, Tutor do ICMC/USP, brunopenteado@usp.br

Cite as: Neves, P. & Isotani, S. & Penteado, B. (2020). Predição de reprovação na educação a distância: um estudo comparativo. Anais dos Trabalhos de Conclusão de Curso. Pós-Graduação em Computação Aplicada à Educação Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo.

1. Introdução

A computação aplicada à educação tem o potencial de contribuir para a integração de diferentes áreas do conhecimento na construção de soluções computacionais para as demandas educacionais.

Para fazer frente ao desafio constante de ensinar e aprender em um mundo cada vez mais dinâmico e moderno, a comunidade científica têm se apropriado de um novo paradigma tecnológico e desenvolvido pesquisas teóricas e empíricas nesta interface interdisciplinar que se forma entre a computação e a educação.

É preciso agilidade e eficácia na extração de informações e a produção de conhecimentos para lidar com questões sociais e econômicas do século 21 (ISOTANI; BITTENCOURT, 2015). A Mineração de Dados Educacionais (MDE) constitui uma das principais abordagens das pesquisas em Mineração de Dados (MD) para lidar com os contextos e práticas informacionais da educação na escala desejada.

O conceito de MDE, do inglês *Educational Data Mining*, representa o campo de pesquisa multidisciplinar dedicado ao desenvolvimento de métodos para explorar os tipos de dados presentes nos ambientes educacionais (ROMERO; VENTURA, 2013). Coma MDE é possível "compreender de forma mais eficaz e adequada os alunos, como eles aprendem, o papel do contexto na qual a aprendizagem ocorre, além de outros fatores que influenciam a aprendizagem" (BAKER; ISOTANI; CARVALHO, 2011, p.4).

Para Luckin *et al.* (2016, p.17) a MDE é o "Desenvolvimento e uso de métodos para analisar e interpretar grandes quantidades de dados (big data) coletados de, por exemplo, ambientes virtuais de aprendizagem ou de sistemas de gerenciamento de escolas, faculdades e universidades".

Inserido na temática da MDE, o presente trabalho visa aplicar a tarefa da predição (classificação) para identificar, com antecedência, os alunos em risco de reprovação. Este problema de pesquisa está definido para os alunos de um curso superior ofertado na modalidade de educação a distância (EaD) por uma universidade pública via sistemática Universidade Aberta do Brasil (UAB).

A reprovação de alunos em cursos superiores se apresenta de diferentes formas e depende de vários fatores, sendo uma questão educacional que envolve diferentes atores (AGUIAR *et al.*, 2014). Uma reprovação pode causar desmotivação no aluno, perda do fluxo de aprendizagem esperado, desperdício de recursos financeiros, impactos negativos aos indicadores da instituição, diminuição do fomento disponibilizado e ainda pode colaborar fortemente com a elevação dos índices de evasão do curso.

Segundo Silva Filho *et al.* (2007) a evasão no ensino superior representa um problema para todas as nações, causando prejuízos sociais, científicos e econômicos. Ora são recursos públicos que não apresentam um retorno efetivo do investimento, ora são importantes perdas de receitas no setor privado. Quando um aluno abandona sua oportunidade de formação e de crescimento intelectual ele também é excluído de todo um processo educacional que o prepara não só para o mercado de trabalho mas também para o exercício da cidadania.

Tendo em vista que os cursos ofertados na sistemática UAB têm ofertas únicas,

ou seja, não têm ingressos anuais ou semestrais, caso um aluno seja reprovado em uma disciplina, ele não terá a oportunidade de se matricular em uma próxima turma dado que ela dificilmente será novamente ofertada até a integralização do seu curso. Este aluno, ainda que existam estratégias pedagógicas para recuperar o seu aprendizado e sua aprovação, terá maiores chances de evadir-se do curso, impactando negativamente a sua formação, seu futuro profissional, o fomento da IES e até mesmo o desenvolvimento social e econômico da localidade onde o aluno está inserido. Na melhor das hipóteses ele irá optar por dar continuidade à sua graduação em uma outra IES.

É neste cenário que se insere a motivação deste trabalho, que busca realizar uma pesquisa inédita na IES em questão, na tentativa de disponibilizar aos envolvidos no processo de ensino-aprendizagem (alunos, professores e gestores) um modelo preditivo com informações adicionais sobre o risco de cada acadêmico(a) vir a reprovar-se em determinada disciplina. Esta iniciativa tem o potencial de minimizar tais reprovações.

Apoiados em alertas elaborados com base em análises preditivas de reprovações em disciplinas, tanto gestores, educadores (coordenadores de curso, de polo, professores, tutores presenciais e a distância) quanto educandos, poderiam ser favorecidos ao se viabilizar um processo de tomada de decisão tempestivo que possa reverter tal risco.

O presente trabalho tem como objetivo identificar empiricamente o(s) modelo(s) de classificação com os melhores resultados para a predição de reprovação de alunos de um curso na modalidade de educação a distância, tendo em vista um estudo comparativo de cinco métricas selecionadas. Como objetivos específicos, tem-se: coletar e consolidar os relatórios de notas dos alunos disponíveis no AVA (Ambiente Virtual de Aprendizagem) Moodle utilizado; realizar o pré-processamento dos dados, incluindo anonimizações, inserções e exclusões de dados, normalização e separação entre dados de treinamento e de testes; realizar testes de predição de reprovação utilizando seis diferentes algoritmos, três opções de balanceamento de dados (sem filtro, com SMOTE e com *Class Balancer*) edois conjuntos de dados (de 3 e 5 semanas de aula, de um total de 6); reunir e comparar as métricas de acurácia, sensibilidade, especificidade, índice *kappa* e média geométrica (*g-means*) dos modelos preditivo, analisar as predições do(s) melhor(es) algoritmos com dados semanais dos alunos; e, por fim, analisar os resultados obtidos com vistas à sua utilização.

Este artigo está organizado como segue. Nas seções 2 e 3 são apresentados os referenciais teóricos sobre MDE, predição de reprovação na EaD e trabalhos correlatos. Na seção 4 é descrita a metodologia desenvolvida para a descoberta dos modelos preditivos com melhores resultados. Na seção 5 são apresentados e discutidos os resultados obtidos e suas limitações e, por fim, na seção 6 estão as conclusões e os trabalhos futuros.

2. Mineração de Dados Educacionais

A Mineração de Dados (MD), do inglês *Data Mining* (DM), pode ser compreendida como uma abordagem exploratória, analítica e indutiva (ANGELI *et al.*, 2017) capaz de revelar informações relevantes e significativas para apoiar a tomada de decisão por meioda identificação de padrões em dados (WITTEN; FRANK; HALLE, 2011).

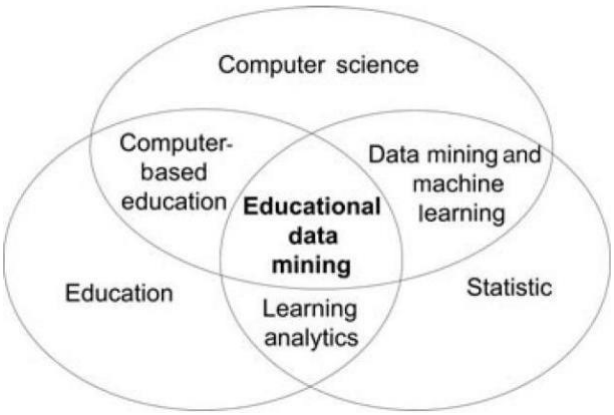
A MD pode ser compreendida como uma das fases de um processo mais abrangente,

o *Knowledge Discovery in Databases* (KDD), juntamente com as etapas de pré e pós-processamento. O KDD é um processo não convencional de descoberta de conhecimento em bases de dados que identifica novos padrões compreensíveis, válidos e potencialmente úteis (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

No cenário educacional, a Mineração de Dados Educacionais é um campo de pesquisa e desenvolvimento cujo objetivo está focado na expansão e no avanço dos horizontes educacionais como forma de contribuir com a evolução do processo educacional (RODRIGUES; ISOTANI; ZáRATE, 2018).

Segundo Romero e Ventura (2013), a MDE pode ser definida como uma combinação de três principais áreas: ciência da computação, educação e estatística (Figura 1).

Figura 1. Áreas envolvidas na Mineração de Dados Educacionais



Fonte: (ROMERO; VENTURA, 2013, p.13))

Existem diferentes taxonomias para classificar as tarefas da MDE. Mohamed e Tasir (2013) utilizam cinco tipos de tarefas básicas: clusterização, classificação, padrões sequenciais, predição e análises de regra de associação. Romero e Ventura (2013) identificam as técnicas de classificação, clusterização e análises de associação como sendo as principais. De forma análoga, Silva e Silva (2015) classificam as tarefas em modelo preditivo, análise de agrupamento e regras de associação.

Analisar dados educacionais por meio de tarefas clássicas da MDE como classificação, agrupamento e regras de associação pode ajudar professores, gestores e alunos a melhor compreenderem os cenários informacionais complexos da educação e, assim, tomarem melhores decisões durante todo o processo de ensino-aprendizagem em que estão inseridos (SILVA; SILVA, 2015).

Diversas métricas educacionais podem ser obtidas por meio do rápido processamento de grande quantidade de dados educacionais, em especial dos Ambientes Virtuais de Aprendizagem (AVA), onde o armazenamento dos dados é naturalmente facilitado.

Técnicas de MDE podem ajudar educadores e gestores a estabelecer uma base pedagógica para as decisões tomadas no planejamento ou na modificação de um ambiente educacional ou abordagem de ensino, aproximando-se da abordagem intitulada *Learning Analytics (LA)* (ROMERO; VENTURA, 2007, 2020).

3. Predição de reprovação na EaD

Segundo Baker, Isotani e Carvalho (2011) a meta da predição é criar modelos que inferem aspectos específicos dos dados por meio de variáveis *preditivas* ao analisar e agrupar diversos aspectos encontrados nos dados, por meio de variáveis *preditoras* .

A predição de reprovação de alunos em disciplinas é uma das possíveis aplicações deste campo de pesquisa que tem recebido especial atenção em diversos níveis e contextos educacionais, dada a relevância do tema (ARAQUE; ROLDÁN; SALGUERO, 2009; MARQUEZ; ROMERO; VENTURA, 2011). Revisões sistemáticas e publicações de estado da arte na área de MDE registram pesquisas diversas envolvendo aspectos relacionados à reprovação de alunos (ROMERO; VENTURA, 2007; MOHAMED; TASIR, 2013; RODRIGUES; ISOTANI; ZáRATE, 2018; ROMERO; VENTURA, 2020).

As possibilidades de melhorias educacionais que se abrem ao se investigar o fenômeno da reprovação de alunos, em seus específicos contextos de aprendizagem, vêm de encontro com as recomendações uníssonas que as pesquisas nesta área fornecem. Embora a reprovação ainda seja um fenômeno pouco explorado pela literatura acadêmica, ela aparece como variável observável que influencia a evasão do aluno, dentre outras (SILVA, 2013).

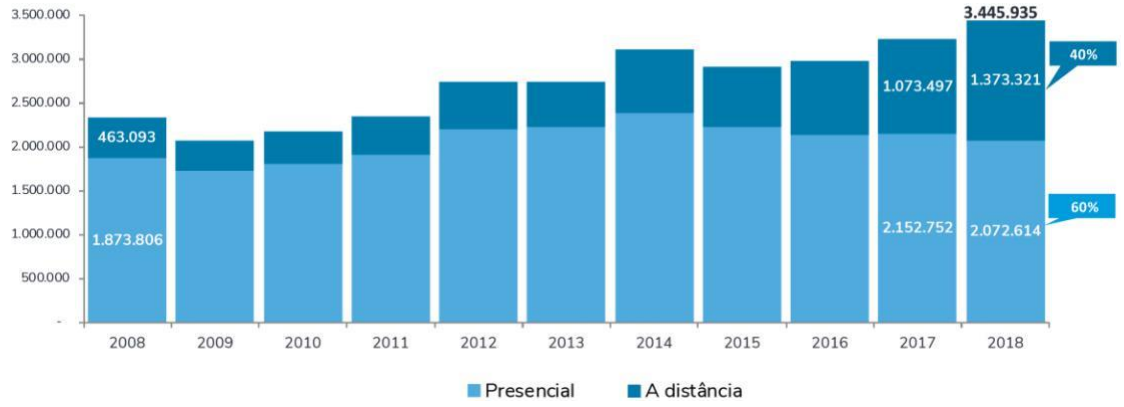
Os índices de reprovação e evasão na Educação a Distância (EaD) são ainda maiores que a educação presencial (KOTSIANTIS, 2009). Dentre as diversas definições de EaD tem-se que o Art. 1o do Decreto no. 9.057, de 25 de maio de 2017, define EaD como uma "modalidade educacional na qual a mediação didático-educacional nos processosde ensino e aprendizagem ocorra com a utilização de meios e tecnologias de informação e comunicação [...] por estudantes e profissionais da educação que estejam em lugarese tempos diversos". O dispositivo destaca ainda a necessidade de pessoal qualificado, políticas de acesso, acompanhamento e avaliação compatíveis, dentre outros requisitos.

A EaD tem se desenvolvido como uma modalidade educacional democrática e dinâmica, fortemente baseada nas tecnologias de informação e comunicação (TIC) e outros recursos educacionais próprios (BELLONI, 2015). Ela ocupa uma posição de destaque na educação superior brasileira. A ascensão desta modalidade é registrada anualmente no Censo da Educação Superior 2018, conforme ilustra a Figura 2.

É possível perceber que entre 2017 e 2018 o número de ingressantes no ensino su-perior aumentou graças ao ingressos da modalidade a distância, cuja variação foi de 27,9% em relação ao ano anterior, ao passo que nos cursos presenciais houve uma diminuição de - 3,7%. O número de ingressos na EaD triplicou (196,6%) em 10 anos (2008 a 2018), representando 40% dos ingressantes em cursos de graduação de todo o Brasil em 2018 (INEP, 2019).

A UAB é uma política pública de educação que desempenha um importante papel social de acesso à educação superior no Brasil (CLíMACO, 2011; ARRUDA; ARRUDA, 2015; MENDONÇA *et al.*, 2020). Embora a EaD esteja colaborando com a popularização, o acesso, a democratização e a inovação da educação superior por meio de programas de formação gratuita a distância, o problema da evasão foi e ainda é sentido nas diversas iniciativas no país e no mundo (ALEJANDRA; BEHAR, 2009).

Figura 2. Ingresso em cursos de graduação, por modalidade de ensino



Fonte: Censo da Educação Superior 2018 (INEP, 2019, p.15)

Diversas causas levam à evasão na EaD. Coelho (2004) cita a falta de fluência nos recursos tecnológicos, a dificuldade de lidar com espaços-tempos diferentes, a ausência física do professor, os obstáculos para adequar o tempo a ser dedicado aos estudos, as participações sem regularidade nos ambientes virtuais de aprendizagem, dentre outros. Este problema é complexo e é conhecido como *"the one hundred factors problem"* (MARQUEZ; ROMERO; VENTURA, 2011)

Os trabalhos de Sousa e Maciel (2016), Santos (2013) e Pacheco, Nakayama e Rissi (2015) apresentam análises da evasão de cursos superiores no Programa UAB. Estes números têm sido uma preocupação de governos e instituições públicas em geral. Para que sejam concebidas estratégias para enfrentar este problema, as causas começam a ser investigadas (SANTOS, 2013; BITTENCOURT; MERCADO, 2014; PACHECO; NAKAYAMA; RISSI, 2015; SOUSA; MACIEL, 2016).

O Centro de Educação a Distância da Universidade Estadual de Montes Claros (CEAD/Unimontes) integra a Sistemática UAB desde as primeiras ofertas ainda em 2008. Atualmente (outubro de 2020) a UAB/Unimontes conta com 8 cursos de graduação sendo ofertados a distância em 27 cidades de Minas Gerais. Nos dois últimos processos seletivos foram ofertadas 1850 vagas: 900 vagas em 7 graduações iniciadas em abr/2018 e 950 vagas em 5 graduações iniciadas em ago/2020. O Curso de Pedagogia detém a maior quantidade de vagas em ambas as ofertas (250 e 230, respectivamente) e por isso foi escolhido como objeto de análise para a realização desta pesquisa. Considerou-se os dados dos alunos ingressantes em 2018, dada a disponibilidade dos dados necessários. Estes alunos estão distribuídos em 5 polos de apoio presencial (Buritis, Monte Azul, Nova Serrana, Urucuia e Várzea da Palma).

Para além dos desafios naturais da educação, a EaD acumula outros relacionados aos elementos próprios da modalidade, como a questão do acesso às tecnologias de informação e comunicação (TIC) e o desenvolvimento de competência informacional (*Information Literacy (IL)*)(VITORINO; PIANTOLA, 2009, 2011; VITORINO, 2016).

3.1. Trabalhos relacionados

Diversos trabalhos têm experimentado a aplicação das técnicas de Mineração de Dados Educacionais na predição de reprovação e/ou evasão de alunos. O estudo de Kotsiantis (2009) dedicou-se em utilizar e adaptar as técnicas de *Machine Learning* para predizer a desistência de alunos em cursos superiores abertos e ofertados na educação a distância propondo uma abordagem específica para lidar com dados desbalanceados. Foram compa-radas as médias geométricas (g-means) das acurácias das classes minoritária e majoritária de alunos.

Marquez, Romero e Ventura (2011) analisaram dados de estudantes mexicanos do ensino médio para predizer suas reprovações utilizando 10 algoritmos de classificação com o WEKA. Foram comparadas as métricas de acurácia geral, verdadeiro positivo (aprovado), verdadeiro negativo (reprovado) e g-means. Eles testaram um conjunto de 77 atributos, provenientes de três fontes diferentes de dados e depois testaram com um conjunto reduzido de 15 atributos, selecionados por meio de técnicas de redução da dimensionalidade dos dados.

Aguiar *et al.* (2014) propuseram medidas de engajamento baseadas em medidas de acessos aos portifólios eletrônicos dos estudantes do primeiro ano de um curso de engenharia (num estudo de coorte) para predizer o desempenho deles no curso. Foram utilizados cinco diferentes métodos de classificação, comparando não só a acurácia geral do modelo mas também as acurácias de cada classe, bem como a curva ROC destes resultados. Adicionalmente, foi comparado o uso do algoritmo de balanceamento SMOTE (CHAWLA*et al.*, 2002), que não apresentou resultados significativos.

Detoni, Cechinel e ARAÚJO (2015) desenvolveram um modelo de predição de reprovação baseado na contagem de interações no Ambiente Virtual de Aprendizagem utilizado, e seus atributos derivados, durante as sete semanas de oferta das disciplinas. Eles analisaram as taxas de verdadeiros-positivos da classe minoritária (reprovados), também chamada de sensibilidade, uma vez que os dados são desbalanceados e, por isso, prever a reprovação é naturalmente mais difícil. Os atributos derivados melhoraram as predições nas primeiras semanas.

O trabalho de Costa *et al.* (2017) comparou a efetividade de diferentes técnicas da MDE na predição de reprovação na disciplina de introdução à programação de alunos de cursos presenciais e a distância de uma universidade pública brasileira. Eles utilizaram as ferramentas Pentaho para o pré-processamento e WEKA para os testes dos algoritmos, com e sem ajustes nos parâmetros. Foram comparadas as médias harmônicas entre precisão e revocação (*F-measure*) dos modelos preditivos e comparados seus resultados.

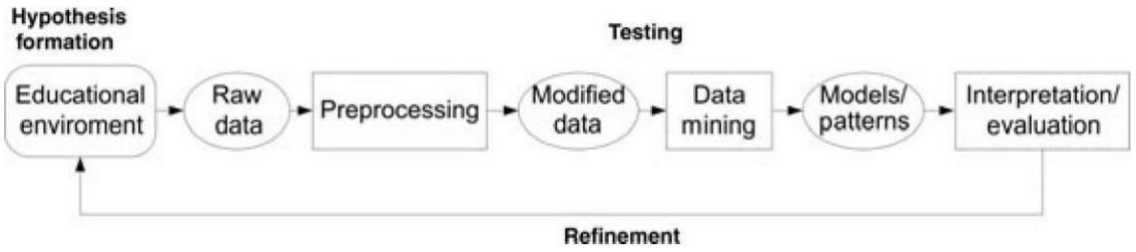
Diversos outros trabalhos relacionados a este demonstram abordagens e resultados interessantes e relevantes para o escopo desta pesquisa, mas não serão detalhados por restrições de espaço (DEKKER; PECHENIZKIY; VLEESHOUWERS, 2009; ARAQUE; ROLDÁN; SALGUERO, 2009; SANTOS, 2013; SILVA, 2013; BITTENCOURT; MER-CADO, 2014; PACHECO; NAKAYAMA; RISSI, 2015; ADEJO; CONNOLLY, 2018; BARROS *et al.*, 2019).

4. Metodologia

A metodologia exploratória desta pesquisa longitudinal utilizou 4396 conjuntos de dados de desempenho de 254 alunos em 21 disciplinas do Curso de Pedagogia ofertado na modalidade a distância pelo Centro de Educação a Distância da Universidade Estadual de Montes Claros (CEAD/Unimontes), via Sistemática UAB da CAPES/MEC. Estes dados correspondem às disciplinas que encontravam-se concluídas em março de 2020, início do então 4o período do curso.

O percurso metodológico desenvolvido neste trabalho é uma combinação do framework metodológico de KDD proposto por Fayyad, Piatetsky-Shapiro e Smyth (1996) e o processo de mineração de dados para descoberta de conhecimentos educacionais de Romero e Ventura (2013), que é apresentado na Figura 3:

Figura 3. Processo de mineração de dados educacionais



Fonte: (ROMERO; VENTURA, 2013, p.19)

Foram reunidos 105 arquivos de relatórios de notas das salas virtuais do Moodle (ead.unimontes.br) referentes a 21 disciplinas do curso. Foram escolhidas as disciplinas de 60h pois representam 84% (21 de um total de 25) das disciplinas concluídas até o momento da coleta de dados. Elas possuem a mesma quantidade de atividades avaliativas e mesma distribuição de notas. Disciplinas de 60h têm duração de 6 semanas.

Diversas transformações iniciais de pré-processamento foram realizadas sobre as planilhas provenientes do Relatório de Notas do Moodle. Tendo reunido todas as planilhas, foram selecionados os dados referentes às notas obtidas nas Atividades Colaborativas (AC) 1 e 2 divididas entre apresentação e entrega, Fóruns de Discussão (FD) de 1 a 4 e Atividade Individual (AI) 1. Foram inseridas as colunas de STATUS e POLO. As notas da "Atividade Individual" 2 (9pts) e "Avaliação Presencial" (40pts), que completam a distribuição das notas das disciplinas, foram removidas pois estas ocorrem no penúltimo e último dias de aula, respectivamente, e suas notas só são registradas após o término da disciplina. Os dados foram anonimizados, os valores numéricos transformados para padrão americano (decimas com ".") e conversão em arquivo .csv. Na aba "Preprocess" do WEKA foi realizada a normalização dos dados numéricos das notas, que variavam de 3pts (FD) a 9pts (AI). Isso foi feito com o uso do filtro do tipo *"unsupervised"* subtipo *"attribute"* denominado *"Normalize"* (WITTEN; FRANK; HALLE, 2011, p.434).

A planilha resultante totalizou 4396 conjuntos de dados, que foram divididos de modo que as notas das 16 primeiras disciplinas concluídas foram utilizadas para treinamento (3348 dados - 76%) dos modelos e as 5 últimas disciplinas para os testes (1048 dados - 24%).

A ferramenta utilizada foi o WEKA *(Waikato Environment for Knowledge Analysis)* (WITTEN; FRANK; HALLE, 2011). Na aba *"Preprocess"* foram aplicados os filtros de balanceamento *Class Balancer* e SMOTE, ambos do tipo *"supervised"* subtipo *"instance"* (WITTEN; FRANK; HALLE, 2011, p.443). No filtro SMOTE foram utilizados os parâmetros *"classvalue"* e *"percentage"* "2" (representando o índice da classe minoritária para a qual se deseja gerar os dados sintéticos) e "450" (de modo que as quantidades de dados de ambas as classes ficassem equivalentes) (CHAWLA *et al.*, 2002).

Os testes com cada modelo de predição foram realizados na aba *"Classify"*, onde foi escolhida a opção *"Supplied test set"* e fornecido o caminho para o arquivo do *dataset* de testes, igualmente normalizado.

Os algoritmos classificadores utilizados foram *Naive Bayes*, *Logistic Regression*, *Suport Vector Machine*, *Decision Table*, *J48* (Decision Tree) e *OneR*, como *baseline*. Eles representam diferentes classes de algoritmos e permitem analisar os padrões por eles revelados (*"transparent box"* (WITTEN; FRANK; HALLE, 2011, p.5) ou *"white box"* (MARQUEZ; ROMERO; VENTURA, 2011, p.2)). Suas saídas podem revelar alguma estrutura dos dados e fornecer informações compreensíveis sobre os dados para a tomada de decisões por seus usuários. O *OneR* foi utilizado como *baseline* por ser um classificador simples baseado em apenas uma regra (*One Rule*).

Estes algoritmos são frequentemente utilizados nas pesquisas citadas na seção 3.1 e em ampla literatura especializada da área (KOTSIANTIS, 2009; MARQUEZ; ROMERO; VENTURA, 2011; AGUIAR *et al.*, 2014; DETONI; CECHINEL; ARAÚJO, 2015; COSTA *et al.*, 2017; MELLO; PONTI, 2018). Todos os algoritmos foram treinados e testados utilizando dois *datasets*: um com dados de 3 semanas de aula e outro com dados de 5 semanas de aula (de um total de 6 semanas).

Cada algoritmo foi treinado com 3 diferentes configurações de filtros de balancea-mento (sem filtro, com SMOTE e com *Class Balancer*). Os 36 resultados obtidos foram registrados em uma planilha eletrônica, onde foi gerada a métrica de média geométrica (*g-means*) e confeccionados os gráficos.

Por fim, os resultados foram comparados sob diferentes métricas:

1. **acurácia geral**: quantidade total de acertos sobre a quantidade total de instâncias;
2. **sensibilidade**: acurácia da classe minoritária (reprovados) ou verdadeiro-negativo (TN) ou ainda *"recall"*;
3. **especificidade**: acurácia da classe majoritária (aprovados) ou verdadeiro-positivo (TP) ou ainda *"precision"*;
4. ***g-means***: média geométrica entre TN e TP, ou seja $\sqrt{TN * TP}$ e
5. **índice *kappa***: coeficiente estatístico que compara a acurácia esperada por uma classificação ao acaso com a acurácia geral do modelo avaliado.

Ao final, os melhores algoritmos foram testados utilizando-se as notas semanais dos alunos, obtendo-se o desempenho semanal deles numa sequência cronológica de dados. Algumas análises foram elaboradas quanto à interpretabilidade destes resultados sob diferentes perspectivas com relação à questão da reprovação dos alunos.

5. Resultados e Discussão

A exploração inicial dos dados coletados indicou que os 254 alunos ingressantes obtiveram, ao longo de 3 semestres letivos, distribuições desbalanceadas entre aprovações e reprovações nas 21 disciplinas cursadas, como se esperava. Estes dados corroboram com as estatísticas apresentadas em toda a literatura pesquisada, demonstrando uma clara concentração de registros de aprovação dos alunos nas disciplinas cursadas, como pode ser percebido na Tabela 1.

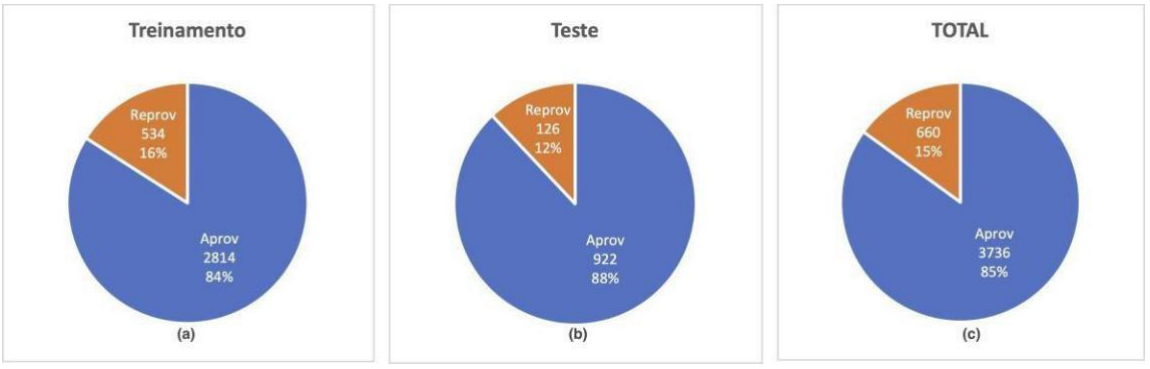
Tabela 1. Aprovações e reprovações

STATUS	Treinamento	Teste	TOTAL
Aprov	2814	922	**3736**
Reprov	534	126	**660**
TOTAL	**3348**	**1048**	**4396**

Fonte: dados da pesquisa.

Os dados de treinamento consistiram das 16 primeiras disciplinas de 60h ofertadas, enquanto os dados de testes corresponderam aos das 5 últimas disciplinas concluídas. Os gráficos que integram a Figura 4 representam, visualmente, a dimensão do desbalancea-mento dos dados utilizados na mineração.

Figura 4. Distribuições entre aprovados e reprovados nos conjuntos de dados (a) Treinamento, (b) Testes e (c) TOTAL



Fonte: dados da pesquisa.

Após executar os algoritmos classificadores selecionados, os resultados de acurácia geral, sensibilidade, especificidade e coeficiente *kappa* foram coletados das saídas dos modelos e registradas numa planilha eletrônica, onde a métrica *g-means* foi calculada.

As tabelas 2 e 3 a seguir apresentam as métricas para ambos os *datasets* utilizados (3 e 5 semanas, respectivamente). Em negrito estão destacados os maiores valores alcançados em cada métrica por cada algoritmo. Em vermelho estão os maiores valores alcançados em cada métrica considerando todos os algoritmos.

Os resultados obtidos com os dados disponíveis na 3a semana de aula são apresentados na Tabela 2 (o algoritmo *OneR* não é habilitado quando se utiliza o *Class Balancer*).

Tabela 2. Resultados com dados de 3 semanas

Algoritmo	Filtro para balanceamento	dataset 3 semanas				
		Acurácia	Sensibilidade	Especificidade	G-means	Kappa
Naive Bayes	sem	**88,8%**	60,3%	**92,7%**	74,8	0,501
	com SMOTE	88,7%	**61,9%**	92,4%	**75,6**	**0,505**
	com ClassBalancer	88,5%	**61,9%**	92,2%	75,5	0,500
Logistic	sem	**92,4%**	55,6%	**97,4%**	73,6	**0,595**
	com SMOTE	88,7%	61,9%	92,4%	**75,6**	0,505
	com ClassBalancer	82,7%	**66,7%**	84,9%	75,3	0,387
SMO (SVM)	sem	92,6%	45,2%	**99,0%**	66,9	0,557
	com SMOTE	**93,1%**	65,1%	97,0%	**79,5**	**0,656**
	com ClassBalancer	81,9%	**66,7%**	83,9%	74,8	0,371
Decision Table	sem	90,5%	52,4%	**95,7%**	70,8	0,516
	com SMOTE	83,4%	61,9%	86,3%	73,1	0,381
	com ClassBalancer	**90,9%**	**70,6%**	93,7%	**81,3**	**0,600**
OneR	sem	**92,8%**	**59,5%**	**97,4%**	**76,1**	**0,627**
	com SMOTE	90,2%	48,4%	95,9%	68,1	0,488
	com ClassBalancer					
J48	sem	**92,8%**	52,4%	**97,7%**	71,6	**0,578**
	com SMOTE	77,1%	65,9%	78,6%	72,0	0,291
	com ClassBalancer	76,3%	80,2%	75,8%	**78,0**	0,334

Fonte: dados da pesquisa.

Ainda que com dados de apenas 3 semanas de aula, os algoritmos foram capazes de fornecer importantes resultados, demonstrando um desempenho elevado com considerável antecedência. O algoritmo baseado em Máquina de Vetor de Suporte (SVM) obteve as melhores métricas para acurácia (93,1%) com uso do filtro SMOTE de balanceamento, especificidade (99%) e índice *kappa* (0,656), com destaque para a elevada especificidade (sem uso de filtro de balanceamento). O SVM oferece maior garantia teórica de gene-ralização e é um indicativo de separabilidade linear entre as classes (MELLO; PONTI, 2018).

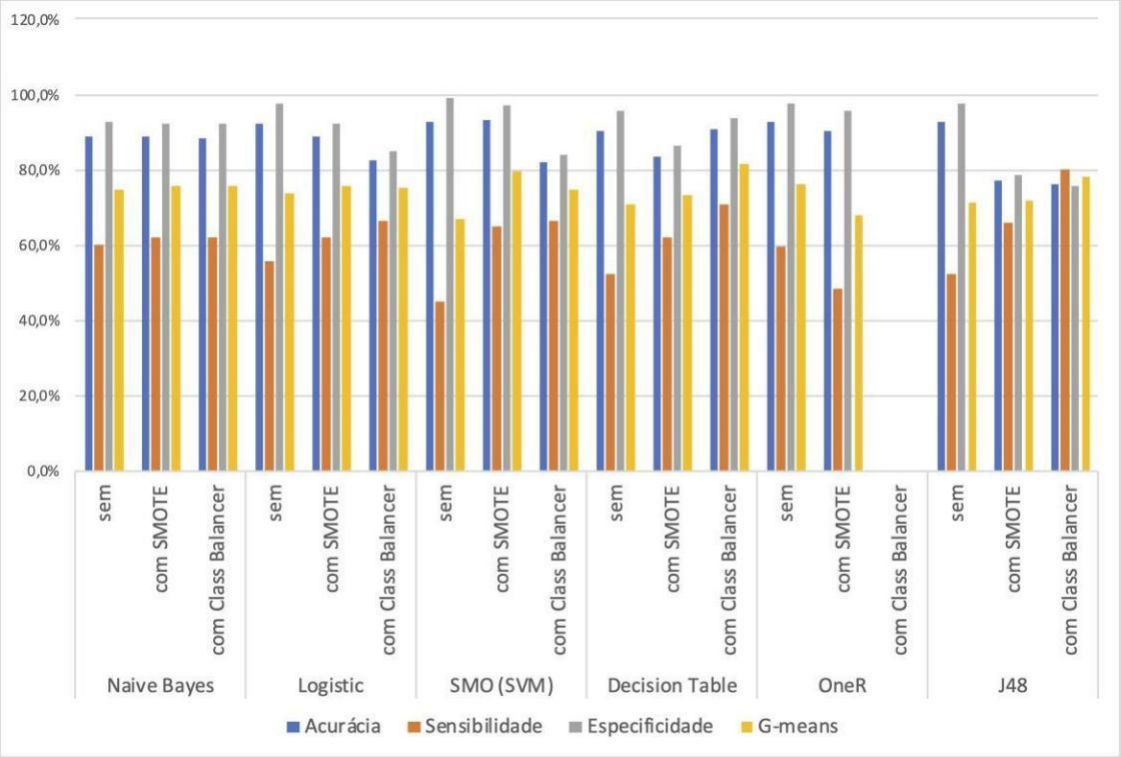
Embora o algoritmo baseado em Árvore de Decisão (*J48*) tenha alcançado 80,2% de sensibilidade, o maior valor dentre todos os algoritmos, devido ao seu baixo índice *kappa*, de apenas 0,334, o destaque de melhor sensibilidade foi atribuído ao algoritmo *Decision Table*, com 70,6% utilizando o filtro *Class Balancer*, situação em que também obteve a maior g-means, de 81,3.

Considerando que a predição da reprovação está implicitamente mais preocupada com alunos com risco de reprovação, ou seja, está focada na classe minoritária, é natural que as métricas de sensibilidade e g-means sejam analisadas com especial atenção.

Neste aspecto, tem-se que a utilização dos filtros de balanceamento SMOTE e *Class Balancer* melhoraram as métricas de sensibilidade e g-means em todos os algoritmos, com exceção do *OneR*, *baseline*. Este é um dado importante uma vez que demonstra a efetividade do filtro SMOTE para este contexto de dados, onde se pretende predizer a ocorrência de uma nova instância na classe minoritária.

A Figura 5 representa graficamente os dados apresentados da tabela 2, com exceção do índice *kappa*, devido à diferença de escala.

Figura 5. Comparativo das métricas dos algoritmos com dados de 3 semanas



Fonte: dados da pesquisa.

Dando prosseguimento aos testes, foram realizados os testes dos algoritmos com os dados de 5 semanas de aula. Os resultados constam na Tabela 3.

Assim como nos dados de 3 semanas, o algoritmo baseado em Máquina de Vetor de Suporte (SVM) obteve as melhores métricas para acurácia e índice *kappa* utilizando o filtro SMOTE e para especificidade sem utilizar filtros.

De forma também análoga aos resultados com os dados de 3 semanas, embora o algoritmo baseado em Árvore de Decisão (*J48*) tenha alcançado 80,2% de sensibilidade, o maior valor dentre todos os algoritmos, o destaque de melhor sensibilidade foi atribuído ao algoritmo *Decision Table* com *Class Balancer*, com 70,6%, devido ao seu índice *kappa* do *J48* ter sido baixo, com apenas 0,324. O *Decision Table* com *Class Balancer* também obteve a maior g-means, de 81,3%, mesmo valor de 3 semanas.

Seguindo com uma análise mais detalhada sobre as métricas mais sensíveis à classe minoritária, tem-se que os filtros de balanceamento melhoraram os resultados de sensibilidade e *g-means* em todos os algoritmos, exceto o *OneR*. Ademais, o *Class Balancer* melhorou a sensibilidade destes algoritmos superando todos resultados obtidos com o SMOTE. Já na métrica *g-means*, o SMOTE foi melhor que o *Class Balancer* em 3 algoritmo, perdendo para o *Class Balancer* em um (*J48*). A Figura 5 representa graficamente os dados apresentados da Tabela 3, com exceção do índice *kappa* devido à

Tabela 3. Resultados com dados de 5 semanas

Algoritmo	Filtro para balanceamento	dataset 5 semanas				
		Acurácia	Sensibilidade	Especificidade	G-means	Kappa
Naive Bayes	sem	88,9%	60,3%	92,8%	74,8	0,504
	com SMOTE	**90,0%**	61,1%	**93,9%**	**75,7**	**0,538**
	com Class Balancer	88,6%	**61,9%**	92,3%	75,6	0,502
Logistic	sem	92,2%	57,1%	97,0%	74,4	0,594
	com SMOTE	**93,5%**	66,7%	**97,2%**	**80,5**	**0,676**
	com Class Balancer	83,9%	**67,5%**	86,1%	76,2	0,413
SMO (SVM)	sem	92,3%	45,2%	**98,7%**	66,8	0,546
	com SMOTE	**94,1%**	62,7%	98,4%	**78,5**	**0,686**
	com Class Balancer	82,6%	**66,7%**	84,8%	75,2	0,385
Decision Table	sem	**93,5%**	62,7%	**97,7%**	78,3	**0,663**
	com SMOTE	85,5%	61,9%	88,7%	74,1	0,425
	com Class Balancer	90,9%	**70,6%**	93,7%	**81,3**	0,600
OneR	sem	**92,8%**	**59,5%**	**97,4%**	**76,1**	**0,627**
	com SMOTE	90,2%	48,4%	95,9%	68,1	0,488
	com Class Balancer					
J48	sem	**92,3%**	52,4%	**97,7%**	71,6	**0,578**
	com SMOTE	76,8%	69,0%	77,9%	73,3	0,300
	com Class Balancer	75,7%	80,2%	75,1%	**77,6**	0,324

Fonte: dados da pesquisa.

diferença de escala.

Para obter ainda mais informações sobre o comportamento destes dados e a capacidade de predição destes algoritmos, escolheu-se os dois algoritmos que obtiveram os melhores resultados nas métricas selecionadas, com seus respectivos filtros, e utilizou-se os dados semanais das disciplinas. Com isso foi possível perceber, em uma série temporal de dados, como o SVM e o *Decision Table* predizem as reprovações ao longo de cada semana de aula. A Tabela 4 exibe os resultados do algoritmo SVM, com o SMOTE, sobre os dados semanais dos alunos. Estes resultados estão representados na Figura 7.

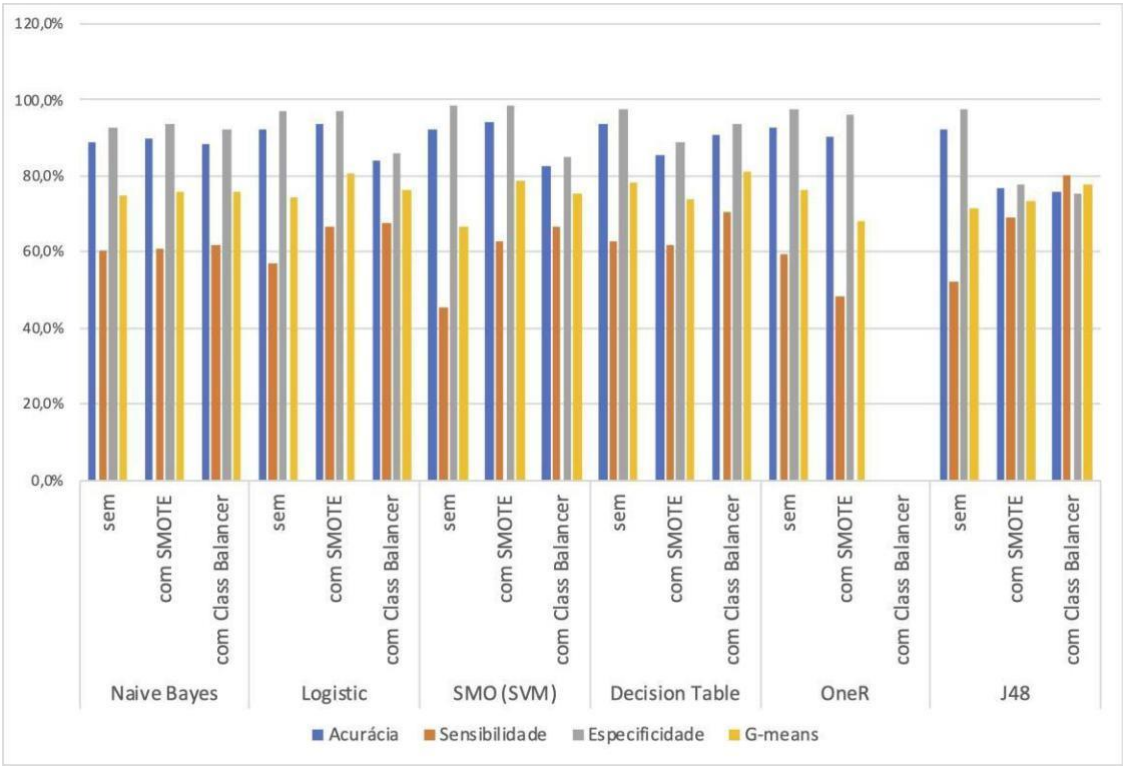
Tabela 4. Execuções do SVM com SMOTE com dados de 1 a 5 semanas

SMO (SVM) com SMOTE				
	Acurácia	Sensibilidade	Especificidade	G-means
1sem	77,1%	55,6%	80,0%	66,7
2sem	78,7%	55,6%	81,9%	67,5
3sem	93,1%	**65,1%**	97,0%	**79,5**
4sem	93,9%	61,9%	98,3%	78,0
5sem	**94,1%**	62,7%	**98,4%**	78,5

Fonte: dados da pesquisa.

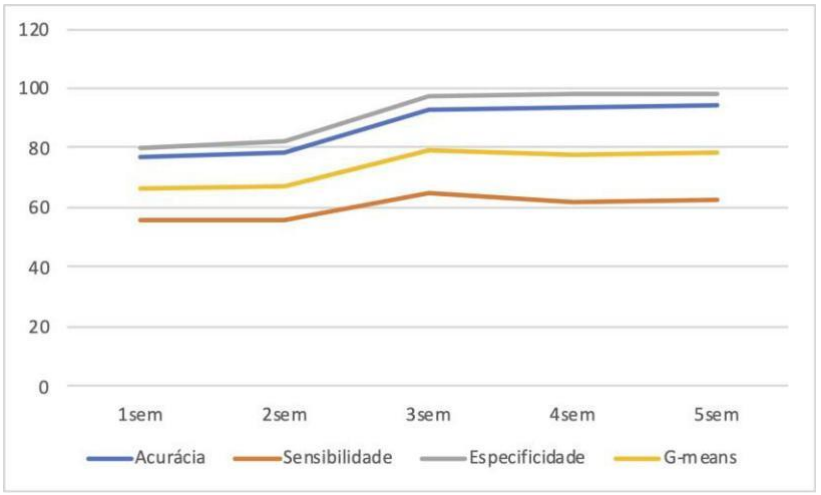
A Tabela 5 exibe os resultados do algoritmo *Decision Table*, com o *Class Balancer*,

Figura 6. Comparativo das métricas dos algoritmos com dados de 5 semanas



Fonte: dados da pesquisa.

Figura 7. Análise temporal do SVM com SMOTE com dados de 1 a 5 semanas



Fonte: dados da pesquisa.

sobre os dados das 5 semanas. Os resultados estão representados na Figura 8.

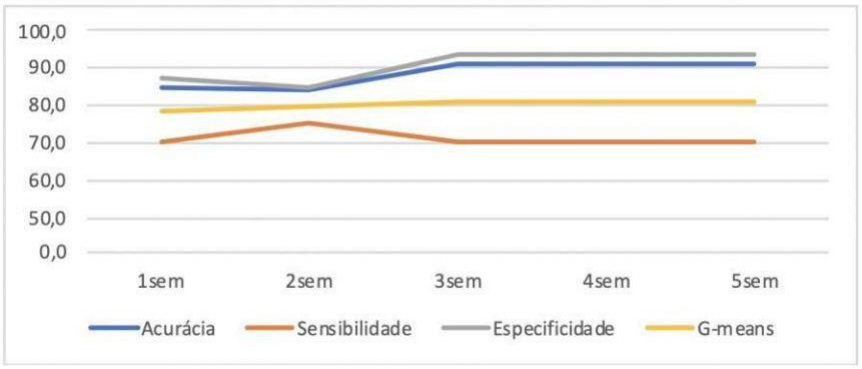
Analisando estas informações adicionais, referentes às sequência cronológicas de dados disponibilizados semanalmente pelo modelo de oferta do curso, é possível perceber que na 3a semana de aula, exatamente na metade das disciplinas, tem-se modelos eficientes de predição da reprovação dos alunos.

Tabela 5. Execuções do *Decision Table* com *Class Balancer* com dados de 1 a 5 semanas

Decision Table com Class Balancer				
	Acurácia	Sensibilidade	Especificidade	G-means
1sem	85,1%	70,6%	87,1%	78,4
2sem	84,0%	75,4%	85,1%	80,1
3sem	90,9%	70,6%	93,7%	81,3
4sem	90,9%	70,6%	93,7%	81,3
5sem	90,9%	70,6%	93,7%	81,3

Fonte: dados da pesquisa.

Figura 8. Análise temporal do *Decision Table* com *Class Balancer* com dados de 1 a 5 semanas



Fonte: dados da pesquisa.

Em suma, os resultados obtidos corroboram com a literatura especializada e mostram que as técnicas de mineração de dados utilizadas são capazes identificar, com antecedência e acurácias satisfatória, os alunos com chances de reprovação nas disciplinas. O algoritmo SMO que utiliza SVM (máquina de vetor de suporte) obteve os melhores valores para acurácia e índice *kappa*, ambos utilizando o SMOTE, e especificidade, sem balanceamento. O DT (*Decision Table*) obteve o melhor valor de sensibilidade, utilizando o *Class Balancer*.

Os filtros de balanceamento melhoraram quase todas as métricas de sensibilidade e *g-means* mas não tiveram este mesmo ganho nas métricas de acurácia, especificidade e índice *kappa*. Dados de 3 e 5 semanas obtiveram os desempenhos semelhantes, demonstrando que os modelos foram capazes de aprender a predizer a reprovação com antecedência de 50% da disciplina.

Para interpretar as métricas no contexto em questão, tem-se que a sensibilidade indica o quanto o algoritmo classificou corretamente os alunos que foram reprovados (classe minoritária). Por outro lado a especificidade indica a acurácia do algoritmo em classificar os aprovados. A acurácia engloba as ambas as classes, assim como a métrica *g-means*, que combina sensibilidade e especificidade em uma média geométrica. O índice *kappa* foi incluído para indicar o quanto cada algoritmo foi melhor que um classificador

aleatório.

5.1. Limitações da pesquisa

Embora este experimento forneça evidências interessantes sobre a efetividade dos algoritmos de mineração de dados educacionais selecionados em predizer, antecipadamente, os alunos com risco de reprovação, é importante registrar suas limitações.

Os dados utilizados representam os alunos que ingressaram em um curso na modalidade de educação a distância de uma universidade pública estadual que integra a Sistemática UAB, seguindo seu modelo de fomento. Assim, os resultados não podem ser generalizados. Embora este trabalho tenha comparado 5 diferentes métricas (acurácia, sensibilidade, especificidade, média geométrica e índice *kappa*) outras métricas conhecidas como *f-measure* e curva *ROC* ou mesmo ponderações combinadas destas métricas, como apontado em Dekker, Pechenizkiy e Vleeshouwers (2009) e MÁRQUEZ-VERA *et al.* (2016) podem ser utilizadas.

Todos os algoritmos foram utilizados com a parametrização padrão da ferramenta WEKA. Ajustes podem impactar significativamente nos resultados.

6. Conclusões e trabalhos futuros

Predizer a reprovação de alunos com antecedência e acurácia é uma tarefa complexa para a jovem e dinâmica área de pesquisa em mineração de dados educacionais (MDE). Tendo em vista que os contextos educacionais são muito diversos, especialmente na EaD, as pesquisas em MDE estarão sempre buscando extrair informações úteis e tempestivas para melhor apoiar o processo de tomada de decisão educacional, favorecendo a aprendizagem de diversas formas.

Este trabalho realizou um estudo comparativo de 6 algoritmos de classificação para predizer a reprovação de alunos de um curso na EaD. Os dados são prevenientes de 105 relatórios de notas do Moodle obtidos nas salas virtuais de 21 disciplinas do Curso de Pedagogia, que iniciou com 254 alunos em abril de 2018. O curso é ofertado na modalidade a distância pela UAB/Unimontes em 5 polos de apoio presencial no estado de MG.

Os dados foram pré-processados, normalizados e separados em *datasets* de treinamento e de teste. Foram utilizados, na ferramenta WEKA, os algoritmos *Naive Bayes*, *Logistic*, *SMO(SVM)*, *Decision Table*, *OneR* e *J48*. Os testes foram executados sem filtros de balanceamento, com o filtro SMOTE e com o *Class Balancer*.

Considerando, nesta ordem, os dados de 3 e de 5 semanas, tem-se que o *SMO(SVM)* sem filtro de balanceamento alcançou os maiores valores de especificidade com 99% e 98,7%, respectivamente, o que significa que, dentre os aprovados, o algoritmo foi capaz de classificar corretamente em 99% e 98,7% dos casos; e com o filtro SMOTE obteve os maiores valores para acurácia geral, com 93,1% e 94,1%, e índice *kappa* de 0,656 e 0,686. *Decision Table* com *Class Balancer* foi o melhor algoritmo em termos de sensibilidade, com 70,6% para 3 e 5 semanas, o que significa que, dentre os reprovados, o algoritmo foi capaz de classificar corretamente em 70,6% dos casos; e média geométrica *g-means* de 81,3% para 3 e 5 semanas. Embora o *J48* com o *Class Balancer* tenha obtido o maior valor de sensibilidade (80,2%), os valores do índice *kappa* nestes casos foram os menores

(0,334 e 0,324).

Analisando a evolução semanal destes 2 melhores algoritmos foi possível perceber que a 3a semana reúne as melhores condições que favorecem as predições mais assertivas, sugerindo que os dados que seguem na 4a e 5a semanas ou não ajudam ou mesmo pioram as predições de reprovações.

Como esperado, os filtros de balanceamento melhoraram a maioria dos resultados obtidos para as métricas de sensibilidade e *g-means*, diminuindo, em geral, os valores de acurácia geral e especificidade. Dados de 3 e 5 semanas não tiveram diferenças significativas.

A interpretação das métricas obtidas com vistas à mitigação da reprovação dos alunos deve levar em consideração os objetivos e contextos educacionais envolvidos. As-sim, caso o objetivo seja optar por um modelo de predição que minimize classificações incorretas de alunos com grandes chances de se reprovarem como se estes fossem se aprovarem, ou seja, caso se deseje promover ao máximo as intervenções junto aos alunos que irão se reprovar visando reverter esta tendência, deve-se priorizar os modelos que alcançam as melhores métricas de sensibilidade (que registra a acurácia na classe mino-ritária, ou seja, quantos alunos reprovados foram classificados como reprovados). Caso o objetivo seja optar por um modelo que não sobrecarregue professores e tutores com alertas desnecessários de reprovações quando os alunos na realidade têm altas chances de se aprovarem, deve-se priorizar os modelos que alcançam as melhores métricas de especificidade (que registra a acurácia na classe majoritária, ou sejam a porcentagem dos alunos aprovados que foram classificados como aprovados). Obviamente, a utilização com-binada das métricas permitirá priorizar outros objetivos educacionais, como balancear as acurácias dos verdadeiros positivos e negativos conjuntamente, ou explorar as informações adquiridas pelos modelos, como os valores associados aos atributos que mais interferem na reprovação do aluno, ou ainda a identificar o atributo que mais contribui com a reprovação do aluno, dentre outros.

Como trabalhos futuros tem-se a análise dos desempenhos alcançados por estes algoritmos por meio de métricas que combinem diferentes pesos aos falsos positivos e falsos negativos a partir de consultas às opiniões de especialistas. Outro possível trabalho futuro inclui a avaliação das saídas dos classificadores por especialistas em EaD, especialmente professores e tutores envolvidos, de modo que suas expertises em identificar alunos com risco de reprovação possam ser complementadas com informações obtidas pelos algoritmos. É possível ainda utilizar dados dos outros 6 cursos também iniciados em abril de 2018 e dos 5 cursos iniciados em agosto de 2020, que totalizam 1850 vagas no ensino superior público ofertadas e em curso atualmente. Pretende-se também avaliar os resultados preditivos acrescentando dados demográficos e outros dados coletados por meio de questionários específicos e, por fim, implementar estratégias para incorporar alertas de predição de reprovação às salas virtuais utilizadas na IES.

Referências

ADEJO, O. W.; CONNOLLY, T. Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher*

Education, Emerald Publishing Limited, v. 10, n. 1, p. 61–75, 2018. ISSN 2050-7003.

AGUIAR, E. *et al.* Engagement vs performance: Using electronic portfolios to predict first semester engineering student persistence. *Journal of Learning Analytics*, v. 1, n. 3, p. 7–33, Nov. 2014. Disponível em: https://learning-analytics.info/index.php/JLA/article/view/4076. Acesso em: 17 mai. 2020.

ALEJANDRA, P.; BEHAR, C. *Modelos pedagógicos em educação a distâncias*. São Paulo: ARTMED, 2009.

ANGELI, C. *et al.* Data mining in educational technology classroom research: Can it make a contribution? *Computers & Education*, v. 113, p. 226–242, 2017. Disponível em: https://ro.uow.edu.au/cgi/viewcontent.cgi?article=1222&context=smartpapers. Acesso em: 20 nov. 2019.

ARAQUE, F.; ROLDÁN, C.; SALGUERO, A. Factors influencing university drop out rates. *Computers & Education*, v. 53, n. 3, p. 563–574, 2009. ISSN 0360-1315. Disponível em: http://www.sciencedirect.com/science/article/pii/S0360131509000815. Acesso em: 10 out. 2020.

ARRUDA, E. P.; ARRUDA, D. E. P. Educação à (SIC) distância no Brasil: políticas públicas e democratização do acesso ao ensino superior. *Educação em Revista*, v. 31, p. 321–338, set. 2015. ISSN 0102-4698. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-46982015000300321&nrm=iso. Acesso em: 11 out. 2020.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de dados educacionais: Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação*, v. 19, n. 02, p. 03–13, 2011. ISSN 2317-6121. Disponível em: https://br-ie.org/pub/index.php/rbie/ article/view/1301. Acesso em: 07 nov. 2019.

BARROS, R. *et al.* Predição do rendimento dos alunos em lógica de programação com base no desempenho das disciplinas do primeiro período do curso de ciências e tecnologia utilizando métodos de aprendizagem de máquina. *Simpósio Brasileiro de Informática na Educação - SBIE*, v. 30, n. 1, p. 1491–1500, 2019. ISSN 2316-6533. Disponível em: https://br-ie.org/pub/index.php/sbie/article/view/8882. Acesso em: 12 out. 2020.

BELLONI, M. L. *Educação a DIstância*. 7. ed. [S.l.]: Editora Autores Associados, 2015. 144 p.

BITTENCOURT, I. M.; MERCADO, L. P. L. Evasão nos cursos na modalidade de educação a distância: estudo de caso do curso piloto de administração da UFAL/UAB. *Ensaio: Avaliação e Políticas Públicas em Educação*, Sscielo, v. 22, p. 465–504, 06 2014. ISSN 0104-4036. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-40362014000200009&nrm=iso. Acesso em: 05 out. 2020.

CHAWLA, N. V. *et al.* SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002. Disponível em: https://www.jair.org/index.php/jair/article/view/10302. Acesso em: 05 ago. 2020.

CLíMACO, J. C. T. de S. Educação a distância: política pública essencial à educação brasileira. *Revista Brasileira de Pós-Graduação*, scielo, v. 8, n. 1, p. 15–28, dez. 2011. ISSN 2358-2332. Disponível em: http://ojs.rbpg.capes.gov.br/index.php/rbpg/article/download/242/231/.

COELHO, M. de L. *A Evasão nos Cursos de Formação Continuada de Professores Universitários na Modalidade de Educação a Distância Via Internet*. ABED - Associação Brasileira de Educação a Distância, 2004. Disponível em: http://www.abed.org.br/site/pt/midiateca/textos_ead/626/a_evasao_nos_cursos_de_formacao_continuada_de_professores_universitarios_na_modalidade_de_educacao_a_distancia_via_internet_. Acesso em: 10 out. 2020.

COSTA, E. B. *et al.* Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, v. 73, p. 247–256, 2017. ISSN 0747-5632. Disponível em: http://www.sciencedirect.com/science/article/pii/S0747563217300596. Acesso em: 09 out. 2020.

DEKKER, G. W.; PECHENIZKIY, M.; VLEESHOUWERS, J. M. Predicting students drop out: A case study. In: *Proceedings of the 2nd International Conference on Educational Data Mining*. Cordoba, Spain: [s.n.], 2009. Disponível em: https://eric.ed.gov/?id=ED539082. Acesso em: 12 out. 2020.

DETONI, D.; CECHINEL, C.; ARAÚJO, R. M. Modelagem e predição de reprovação de acadêmicos de cursos de educação a distância a partir da contagem de interações. *Revista Brasileira de Informática na Educação*, v. 23, n. 3, p. 1–11, 2015.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery em Databases. *AI Magazine*, v. 17, n. 13, p. 37–54, 1996. Disponível em: https://www.aaai.org/ojs/index.php/aimagazine/article/view/1230/1131. Acesso em: 02 fev. 2020.

INEP. *Censo da Educação Superior: Notas Estatísticas*. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais/MEC, 2019. Disponível em: http://download.inep.gov.br/educacao_superior/censo_superior/documentos/2019/censo_da_educacao_superior_2018-notas_estatisticas.pdf. Acesso em: 10 out. 2020.

ISOTANI, S.; BITTENCOURT, I. I. *Dados Abertos Conectados*. Editora Novatec, 2015. 176 p. Disponível em: http://ceweb.br/livros/dadosabertosconectados/. Acesso em: 12 fev. 2020.

KOTSIANTIS, S. Educational Data Mining: a case study for predicting dropout-prone students. *International Journal of Knowledge Engineering and Soft Data Paradigms*, v. 1, n. 2, p. 101–111, 2009. Disponível em: https://www.inderscienceonline.com/doi/abs/10.1504/IJKESDP.2009.022718. Acesso em: 11 out. 2020.

LUCKIN, R. *et al. Intelligence Unleashed. An argument for AI in Education.* London: Pearson, 2016. Disponível em: http://discovery.ucl.ac.uk/1475756/. Acesso em: 10 out. 2020.

MARQUEZ, C.; ROMERO, C.; VENTURA, S. Predicting school failure using data mining. In: *Proceedings of the 4th International Conference on Educational Data Mining*. Eindhoven, The Netherlands: [s.n.], 2011. p. 271–276. Disponível em: https://www. academia.edu/download/30661306/edm2011_paper11_short_Marquez-Vera.pdf. Acesso em: 10 out. 2020.

MELLO, R. F. de; PONTI, M. A. *Machine Learning: A Practical Approach on the Statistical Learmimg Theory*. [S.l.]: Springer, 2018.

MENDONÇA, J. R. C. d. *et al.* Políticas públicas para o ensino superior a distância: um exame do papel da universidade aberta do brasil. *Ensaio: Avaliação e Políticas Públicas em Educação*, scielo, v. 28, p. 156–177, 03 2020. ISSN 0104-4036. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-40362020000100156&nrm=iso.

MOHAMED, S. K.; TASIR, Z. Educational data mining: A review. *Procedia - Social and Behavioral Sciences*, v. 97, p. 320–324, 2013. Disponível em: https://core.ac.uk/download/pdf/82645138.pdf. Acesso em: 10 out. 2020.

MÁRQUEZ-VERA, C. *et al.* Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, v. 33, n. 1, p. 107–124, 2016. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.12135. Acesso em: 08 out. 2020.

PACHECO, A. S. V.; NAKAYAMA, M. K.; RISSI, M. Evasão e permanência dos estudantes de um curso de administração a distância do sistema Universidade Aberta do Brasil: uma teoria multiparadigmática. *Revista de Ciências da Administração*, v. 17, n. 41, p. 65–81, 2015. ISSN 2175-8077. Disponível em: https://periodicos.ufsc.br/index.php/adm/article/view/2175-8077.2015v17n41p65. Acesso em: 01 out. 2020.

RODRIGUES, M. W.; ISOTANI, S.; ZáRATE, L. E. Educational data mining: A review of evaluation process in the e-learning. *Telematics and Informatics*, v. 35, n. 6, p. 1701–1717, 2018. ISSN 0736-5853. Disponível em: http://www.sciencedirect.com/science/article/pii/S0736585317306639. Acesso em: 03 mai. 2020.

ROMERO, C.; VENTURA, S. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, v. 33, n. 1, p. 135–146, 2007. ISSN 0957-4174. Disponível em: http://www.sciencedirect.com/science/article/pii/S0957417406001266. Acesso em: 02 dez. 2019.

ROMERO, C.; VENTURA, S. Data mining in education. *WIREs Data Mining and Knowledge Discovery*, v. 3, n. 1, p. 12–27, 2013. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1075. Acesso em: 02 out. 2020.

ROMERO, C.; VENTURA, S. Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, v. 10, n. 3, p. e1355, 2020. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1355. Acesso em: 05 out. 2020.

SANTOS, A. G. R. A evasão nos cursos de graduação a distância UAB/unimontes no polo de são joão da ponte/mg. *Revista Multitexto)*, v. 2, n. 1, 2013. Disponível em: http://www.ead.unimontes.br/multitexto/index.php/rmcead/article/view/119. Acesso em: 08 out. 2020.

SILVA FILHO, R. L. L. e *et al.* A evasão no ensino superior brasileiro. *Cadernos de Pesquisa*, v. 37, n. 132, p. 641–659, 2007. Disponível em: https://www.scielo.br/pdf/cp/v37n132/a0737132.pdf. Acesso em: 10 out. 2020.

SILVA, G. P. d. Análise de evasão no ensino superior: uma proposta de diagnóstico de seus determinantes. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, scielo, v. 18, n. 2, p. 311–333, jul. 2013. ISSN 1414-4077. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1414-40772013000200005&nrm=iso. Acesso em: 10 out. 2020.

SILVA, L. A.; SILVA, L. Fundamentos de mineração de dados educacionais. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, v. 3, n. 1, p. 568–581, 2015. ISSN 2316-8889. Disponível em: https://br-ie.org/pub/index.php/wcbie/article/view/ 3281. Acesso em: 27 mai. 2020.

SOUSA, A. da S. Q.; MACIEL, C. E. Expansão da educação superior: permanência e evasão em cursos da Universidade Aberta do Brasil. *Educação em Revista*, Scielo, v. 32, p. 175–204, 12 2016. ISSN 0102-4698. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-46982016000400175&nrm=iso. Acesso em: 11 out. 2020.

VITORINO, E.; PIANTOLA, D. Competência informacional – bases históricas e conceituais: construindo significados. *Ciência da Informação*, Brasília, DF, v. 38, n. 3, 2009. ISSN 1518-8353. Disponível em: http://revista.ibict.br/ciinf/article/view/1236.

VITORINO, E. V. Análise dimensional da competência informacional: bases teóricas e conceituais para reflexão. *Revista Ibero-Americana de Ciência da Informação*, Brasília, v. 9, n. 2, p. 421–440, jul./dez. 2016. ISSN 1983-5213. Disponível em: https://periodicos.unb.br/index.php/RICI/article/view/2420.

VITORINO, E. V.; PIANTOLA, D. Dimensões da competência informacional (2). *Ciência da Informação*, v. 40, n. 1, 2011.

WITTEN, I. H.; FRANK, E.; HALLE, M. A. *Data Mining*: Practical machine learning tools and techniques. 3rd. ed. San Francisco, CA: Morgan Kauffman, 2011.