Programa Interdisciplinar de Pós-Graduação em

Computação Aplicada

Mestrado Acadêmico

Paulo Fernando Benetti Marcon

Modelagem Generalista ou Individualizada na Construção de Modelos Preditivos para a Identificação de Insucesso Acadêmico

São Leopoldo, 2017

Paulo Fernando Benetti Marcon

MODELAGEM GENERALISTA OU INDIVIDUALIZADA NA CONSTRUÇÃO DE MODELOS PREDITIVOS PARA A IDENTIFICAÇÃO DE INSUCESSO ACADÊMICO

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

Orientador: Dr. João Francisco Valiati

São Leopoldo

2017

M321m Marcon, Paulo Fernando Benetti
Modelagem generalista ou individualizada na construção de modelos preditivos para a identificação de insucesso acadêmico / por Paulo Fernando Benetti Marcon. – 2017.
66 f. : il. ; 30 cm.

Dissertação (mestrado) — Universidade do Vale do Rio dos Sinos, Programa de Pós-graduação em Computação Aplicada, São Leopoldo, RS, 2017.

"Orientador: Dr. João Francisco Valiati."

1. Mineração de dados (computação). 2. Sistemas de alerta antecipado. 3. Predição de insucesso acadêmico. 4. Construção de modelos computacionais. I. Título.

CDU: 004.421

Catalogação na Publicação:
Bibliotecário Alessandro Dietrich - CRB 10/2338

Paulo Fernando Benetti Marcon

Modelagem Generalista ou Individualizada na
Construção de Modelos Preditivos para a
Identificação de Insucesso Acadêmico

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

Aprovado em 31 de março de 2017.

BANCA EXAMINADORA

Prof. Dr. João Francisco Valiati – UNISINOS

Prof. Dr. Sandro José Rigo – UNISINOS

Prof. Dr. Sandro da Silva Camargo – UNIPAMPA

Prof. Dr. João Francisco Valiati

Visto e permitida a impressão
São Leopoldo,

Prof. Dr. Sandro José Rigo
Coordenador PPG em Computação Aplicada

Aos meus pais, Neiva e Fernando Marcon, que sempre me apoiaram; a minha noiva, Juliana Castilhos, que esteve ao meu lado em toda a jornada; e ao meu irmão e amigo, Francisco Marcon.

AGRADECIMENTOS

À DEUS por ter me dado força.

Ao meu orientador, prof. Dr. João Francisco Valiati, por todo o apoio dado e experiência transmitida.

Ao meu amigo Rodrigo de Moraes, pelo direcionamento.

Aos meus amigos Wagner Cambruzzi e Gilmar Piaia, por terem acreditado no meu trabalho.

Ao professor Dr. Wilson Gavião, pelas inspirações.

A todos os meus colegas de mestrado da turma de 2015/1 que me ajudaram.

RESUMO

O uso de recursos tecnológicos para auxiliar nas tarefas de ensino e aprendizagem é uma realidade. A disseminação de ambientes virtuais de aprendizado, como meio de promover a realização de cursos *on-line*, demonstra franca expansão. Além de tarefas que propiciam a ampliação dos meios de ensino, tais sistemas permitem o registro completo de todas as interações dos alunos no decorrer da realização de disciplinas. Essa gama de informação produzida pode ser utilizada para predição de estudantes em situação de risco enquanto a disciplina ocorre, o que para instituições de ensino pode representar redução nos índices de reprovação e evasão. Entretanto o número elevado de variáveis envolvidas, ainda mais quando várias disciplinas são consideradas, dificulta a construção de modelos computacionais eficientes. Desta forma, este trabalho visa investigar a construção de modelos generalistas – treinados com dados de diversas disciplinas disponíveis – contrapondo a construção de modelos individualizados – treinados individualmente com dados de cada disciplina. Para isto um amplo conjunto de dados educacionais foi extraído, obtido de uma instituição de ensino superior, composto de diferentes cursos, disciplinas e períodos letivos, não sendo utilizadas variáveis que invadissem a privacidade dos estudantes. Uma vez definidas as características e transformações dos dados que contribuíam à identificação de insucesso acadêmico no decorrer da disciplina então foram aplicados algoritmos clássicos de Mineração de Dados seguindo ambas as abordagens, generalista e individualizada, e a cada unidade de conteúdo das disciplinas. Os resultados obtidos demonstram vantagens e desvantagens de ambas as abordagens e que dadas as circunstâncias os modelos individualizados podem ser melhores, obtendo taxas de acerto maiores, e que em outras circunstâncias modelos generalistas apresentam um custo menor para a obtenção e manutenção dos modelos preditivos, mesmo com uma queda nos índices de acerto.

Palavras-Chave: Mineração de Dados Educacionais; Sistemas de Alerta Antecipado; Predição de Insucesso Acadêmico; Construção de Modelos Computacionais.

ABSTRACT

The use of technological resources to assist teaching and learning tasks is a reality. The dissemination of virtual learning environments, as a mean of promoting online courses, shows a clear expansion. In addition to tasks that allow the expansion of teaching resources, such systems allow the complete recording of all the interactions of the students inside the courses. This range of information produced can be used to predict at-risk students while the course is taking place, which for educational institutions may represent a reduction in failure and dropout rates. However, the high number of variables involved, especially when several courses are considered, makes it difficult to construct efficient computational models. In this way, this work aims to investigate the construction of generalist models – trained with data from several available courses – counterposing the construction of individualized models – individually trained with data from each course. In this way, a broad set of educational data was extracted, obtained from a higher education institution, composed of different undergraduate programs, courses and academic periods, not using variables that invaded students' privacy. Once the characteristics and transformations of the data that contributed to the identification of academic insuccess during the course were defined, then classical data mining algorithms were applied following both generalist and individualized approaches and to each content unit of the course. The results obtained demonstrate the advantages and disadvantages of both approaches and that given the circumstances the individualized models may be better, obtaining higher hit rates, and that in other circumstances generalist models present a lower cost for the obtaining and maintenance of the predictive models, even with a drop in hit rates.

Keywords: Educational Data Mining; Early Warning Systems; Academic Insuccess Prediction; Computational Models Building.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

p.p.	Pontos percentuais
Neg	Classe Negativa
Pos	Classe Positiva
FN	Falso Negativo
FP	Falso Positivo
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

LISTA DE SIGLAS

AVA	Ambientes Virtuais de Aprendizado
CART	*Classification And Regression Tree*
DM	Mineração de Dados (*Data Mining*)
DT	Árvores de Decisão (*Decision Trees*)
EDM	Mineração de Dados Educacionais *(Educational Data Mining*)
EUA	Estados Unidos da América
EWS	Sistemas de Alerta Antecipado (*Early Warning Systems*)
GD	Gradiente Descendente
GR	Razão de Ganho (*Gain Ratio*)
IG	Ganho de Informação (*Information Gain)*
KDD	Descoberta de Conhecimento em Bases de Dados *(Knowledge Discovery in Databases*)
MOOC	*Massive Open Online Course*
NB	*Naïve Bayes*
RNAs	Redes Neurais Artificiais
SCG	*Scaled Conjugate Gradient*
SMOTE	*Synthetic Minority Over-Sampling Technique*
SVM	Máquina de Vetores de Suporte (*Support Vector Machine*)
UNISINOS	Universidade do Vale do Rio dos Sinos

SUMÁRIO

1 INTRODUÇÃO

Atualmente é comum que instituições de ensino superior tenham as interações entre os professores e alunos dirigidas por sistemas eletrônicos de informações, os quais possibilitam acompanhar o progresso do estudante. Ferramentas como Ambientes Virtuais de Aprendizado (AVA) são utilizadas pelos professores como suporte no processo de aprendizado e incluem recursos de entrega de conteúdo, módulos de pergunta-resposta, submissão de tarefas, informação de notas, entre outros. Assim, principalmente em cursos *on-line* todas as interações feitas pelos estudantes, professores, e possivelmente até outros integrantes da instituição, são registrados no software (AGUDO-PEREGRINA et al., 2014; ANTUNES, 2011; ZAFRA; VENTURA, 2012).

Conforme ocorrem as interações dos alunos dentro do AVA, os professores e demais interessados podem observar o desempenho dos estudantes nas tarefas, descobrir aqueles que têm mais dificuldades na matéria e identificar aqueles que estão em risco de insucesso acadêmico, ou seja, de evadirem ou reprovarem. O acompanhamento constante dos estudantes conduz à possibilidade de gerar retornos às áreas responsáveis, em um tempo hábil, para ajuda-los a ter um bom desempenho escolar e atender os patamares de ensino e aprendizado estipulados pelas instituições de ensino.

Sabe-se que as interações dos estudantes com o professor, com seus colegas e material de aula, levam à troca de informações e à estimulação intelectual, entretanto é difícil de mensurar o peso que cada tipo de atividade e comportamento oferece para o processo de aprendizado daquele individuo, tornando-se difícil a análise de quais métricas e indicativos observar e acompanhar dentro destas ferramentas (AGUDO-PEREGRINA et al., 2014). Como forte indicativo do seu progresso, tem-se comumente as notas obtidas em avaliações e trabalhos, entretanto este critério muitas vezes só é conhecido num momento mais tardio da disciplina e nem sempre em tempo satisfatório para desencadear ações preventivas.

As instituições, adicionalmente, também estão preocupadas com o progresso dos estudantes, pois, de um ponto de vista econômico, enfrentam com a sua evasão a possível perda de verbas provenientes do pagamento das mensalidades. Além disto, muitos programas, fundos e organizações de fomento à educação trabalham atreladas a índices que medem a taxa de evasão e/ou reprovação. Desta forma, é crucial identificar estudantes que estão em risco e agir sobre eles melhorando os índices de retenção e aprovação (JAYAPRAKASH et al., 2014; THAMMASIRI et al., 2014).

A tarefa de acompanhamento e de identificação dos inúmeros fatores que levam ao insucesso acadêmico, como dito, é uma tarefa dispendiosa, difícil e que consome tempo. Tecnologias que possibilitem apontar os estudantes que têm mais probabilidades de evadir, permitindo os professores e demais responsáveis focarem seus esforços nestes alunos, estão tornando-se ferramentas indispensáveis (DANGI; SRIVASTAVA, 2014; MÁRQUEZ-VERA; MORALES; SOTO, 2013).

Os registros de uso gerados pelo AVA nestes casos podem ser utilizados para extrair informações úteis para construir sistemas que possibilitem identificar estudantes em risco de insucesso acadêmico. A Mineração de Dados Educacionais (*Educational Data Mining* - EDM) aborda isso, trazendo o desenvolvimento de métodos para explorar os tipos únicos de informações que descrevem os estudantes, permitindo entender, otimizar o problema e descobrir padrões em grandes volumes de dados (PEÑA-AYALA, 2014b).

Adicionalmente, Sistemas de Alerta Antecipado (*Early Warning Systems* - EWS) podem ser construídos para identificar estudantes que estão em risco de insucesso acadêmico tão cedo quanto for possível (MÁRQUEZ-VERA et al., 2016). Juntamente com EDM, os EWS possibilitam que se construam modelos preditivos que permitam acompanhar o progresso dos estudantes enquanto a disciplina ocorre ou o curso é realizado, gerando alertas previamente e possibilitando que ações que combatam o mau desempenho sejam tomadas em tempo hábil para reversão do quadro (MACFADYEN; DAWSON, 2010; XING et al., 2016).

1.1 Motivações

Visto que os dados são inerentemente relacionados ao progresso dos estudantes dentro de uma disciplina em especifico e dada a complexidade dos dados em função do problema – ao passo que são observadas as diferentes disciplinas e cursos, estilos de aprendizado e ensino, progresso dos estudantes e assim por diante – considera-se a construção de modelos individualizados.

A abordagem de construir modelos individualizados, para cada disciplina, com um conjunto de diferentes parâmetros para cada um, implica num elevado custo para obter e manter modelos eficientes capazes de predizer alunos em risco. Desta forma, desenhar modelos genéricos pode ajudar a reduzir a necessidade intensa de modelagem de dados, tempo de implantação dos modelos e recursos de máquina para treinar os algoritmos, o que pode ser facilmente traduzido em economia, principalmente em sistemas implantados num modelo de computação em nuvem. Entretanto, tratar tudo como um conjunto único de dados, construindo um modelo genérico pode diminuir o seu poder preditivo (GAŠEVIĆ et al., 2016), eventualmente, obtendo-se sistemas não tão confiáveis que não alcançam seus objetivos centrais de identificar com confiança alunos em risco.

Adicionalmente, não é amplamente difundido na literatura a confluência dos temas EDM e EWS considerando a predição enquanto a disciplina ocorre, além da discussão sobre o modelo preditivo empregado.

Trabalhos como o de Gašević et al. (2016), que consideram aspectos de generalização de modelos, possuem uma abordagem estatística e pouco direcionado para a Mineração de Dados (*Data Mining* - DM). Outros trabalhos, como o de Hu, Lo e Shih (2014), apesar de discutirem os pontos acima em conjunto, trazem um estudo sem denotar a relevância da variabilidade e dos diferentes e possíveis agrupamentos de dados (disciplinas, cursos, etc.), além do impacto para o resultado final do classificador.

Além disso, grande parte dos trabalhos de EDM costuma gerar modelos de classificação ou regressão que predizem o desempenho do estudante no curso (PEÑA-AYALA, 2014b), em muitos casos desprezando o aspecto temporal do problema.

A privacidade também tem sido outro tópico de debate, principalmente no uso de conjuntos de dados para outros fins que não para o qual ele foi criado e liberado para uso (HORVITZ; MULLIGAN, 2015). Assim, levantam-se dúvidas da permissão de uso de dados invasivos, ou seja, informações exclusivamente pessoais, não advindas da interação do aluno com a instituição de ensino. Por exemplo, identificadores do perfil demográfico e vida pregressa do estudante, como encontrado em Márquez-Vera et al. (2016).

1.2 Objetivos

O objetivo geral deste trabalho é confrontar a modelagem individualizada, ou seja, utilizando dados de cada disciplina individualmente, e a modelagem generalista, que usa os dados de todas as disciplinas disponíveis em conjunto, para a construção de modelos preditivos à identificação de alunos em risco de insucesso acadêmico enquanto a disciplina ocorre, avaliando vantagens e desvantagens de cada modelagem adotada. Para tanto utilizando-se de variáveis com características temporais e não invasivas de dados educacionais extraídos de um sistema de gerenciamento de aprendizagem.

Para isso definem-se como objetivos específicos os seguintes:

- Consolidar um amplo conjunto de dados educacionais, não invasivo, abrangendo disciplinas de diferentes áreas de conhecimento e vários períodos letivos, separado ainda por unidades de conteúdo;
- Definir as agregações e transformações aplicadas aos atributos do AVA, bem como técnicas de classificação de dados que contribuem à identificação de insucesso acadêmico no decorrer da disciplina;
- Indicar a modelagem de dados generalista ou específica por disciplina, que aparenta, mediante os critérios de avaliação empregados, ser mais promissora à criação de modelos preditivos para EWS.

O presente estudo é caracterizado por ser realizado numa parceria empresa-universidade, sendo que parte dos modelos aqui descritos já estão sendo empregados num projeto que, através de um sistema computadorizado, trabalha a gestão da retenção em instituições de ensino.

1.3 Organização do Trabalho

Este trabalho aborda no capítulo 2 a contextualização dos temas EDM e EWS, e o processo de Descoberta de Conhecimento em Bases de Dados. Já no capítulo 3 são apresentados trabalhos relacionados seletos, no sentido de posicionar o estudo desenvolvido. No capítulo 4 é apresentada a abordagem experimental empregada, bem como a descrição dos dados e suas modelagens, parametrizações e técnicas utilizadas para a obtenção dos resultados, além da apresentação dos resultados obtidos, bem como sua discussão. Por fim o capítulo 5 traz as conclusões do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo traz uma contextualização dos termos e técnicas empregadas no decorrer deste trabalho. Inicialmente, na subseção 2.1 é abordada a área de EDM, seguida do tema EWS, com o propósito de situar este estudo na sua área de conhecimento.

Na subseção 2.2 é detalhado o processo de Descoberta de Conhecimento em Bases de Dados, que guia a metodologia proposta. A subseção 2.3 apresenta técnicas de pré-processamento e transformação de dados, e a subseção 2.4 a etapa de Mineração de Dados em função de algumas técnicas de classificação, com a subseção 2.5 atendo-se em apresentar as métricas de mensuração dos resultados obtidos por classificadores.

Ao final, na subseção 2.6, são feitas considerações frente ao apresentado, traçando os relacionamentos entre cada um dos itens do capítulo, com o todo.

2.1 Mineração de Dados Educacionais e Sistemas de Alerta Antecipado

O crescimento da oferta de cursos, principalmente *on-line*, e o aumento no uso de diferentes sistemas educacionais, que capturam todos os aspectos das interações dos estudantes, fizeram com que se criassem grandes e ricos repositórios de dados educacionais dentro das instituições de ensino. A EDM usa estes conjuntos de dados para entender os aprendizes e o processo de aprendizado, visando desenvolver abordagens computacionais combinando dados e teoria, para revertê-los em benefícios.

Desta forma, EDM pode ser dita como a área que explora os diferentes tipos de dados educacionais para resolver os problemas de pesquisa educacional (ROMERO et al., 2011). Sendo uma nova aplicação de DM[1], focada na descoberta de conhecimento, tomada de decisões e recomendação, orientada para desenhar modelos, tarefas, métodos e algoritmos para explorar informações de conjuntos de dados educacionais (PEÑA-AYALA, 2014b).

Os tópicos que a EDM abrange, segundo Peña-Ayala (2014b), podem ser observados na tabela 1, juntamente com uma breve descrição do objetivo de cada um.

Tabela 1 - Tópicos de pesquisa em EDM segundo Peña-Ayala (2014b).

Tópico	Objetivo
Modelagem do Estudante	Modelar diferentes domínios que caracterizam o estudante
Modelagem do comportamento do Estudante	Caracterizar os padrões de comportamento para adaptar o sistema às tendências dos estudantes
Modelagem do desempenho do Estudante	Predizer o quão bem irá o estudante na disciplina, curso ou determinada situação de aprendizado
Avaliação	Supervisão e avaliação da aquisição de conhecimento pelo estudante
Suporte ao estudante e *feedback*	Rastreamento dos *feedbacks* dos usuários para disparar ações
Curriculum	Montagem de currículos de aulas em função dos alunos
Ferramentas	Desenho, desenvolvimento e teste de ferramentas específicas

Fonte: Adaptado de Peña-Ayala (2014b).

A Modelagem do Desempenho do Estudante, em especial, é um tópico de pesquisa de grande importância, pois, segundo Heppen e Therriault (2008), o problema do baixo desempenho escolar e evasão já é diagnosticado como uma crise nacional nos EUA, ademais eles elencam que EWS são uma alternativa importante para sanar este problema.

[1] DM - Mineração de Dados *(Data Mining*): também conhecido por Descoberta de Conhecimento em Bases de Dados, para maiores detalhes vide subseção 2.2.

A adoção de EWS por governos de vários estados dos EUA, que inclusive demonstra a sua crescente popularidade, destaca a capacidade deste tipo de sistema em permitir focar os indivíduos ou grupos (isto é, disciplinas, cursos, instituições de ensino, etc.) mais necessitados de auxílio para reverter um quadro de baixo desempenho escolar (UEKAWA et al., 2010).

Os EWS definem-se por serem desenhados para alertar responsáveis por decisões de um perigo em potencial, antes que se torne um perigo real. No contexto educacional, ele contempla um conjunto de procedimentos para identificação precoce de estudantes que estão em risco de baixo desempenho escolar e de medidas interventivas para evitar tais problemas (HU; LO; SHIH, 2014). Este tipo de sistema já foi usado inclusive em áreas como ataques militares, prevenção de conflitos, crises econômicas, desastres ambientais, epidemias humanas, entre outros (MÁRQUEZ-VERA et al., 2016).

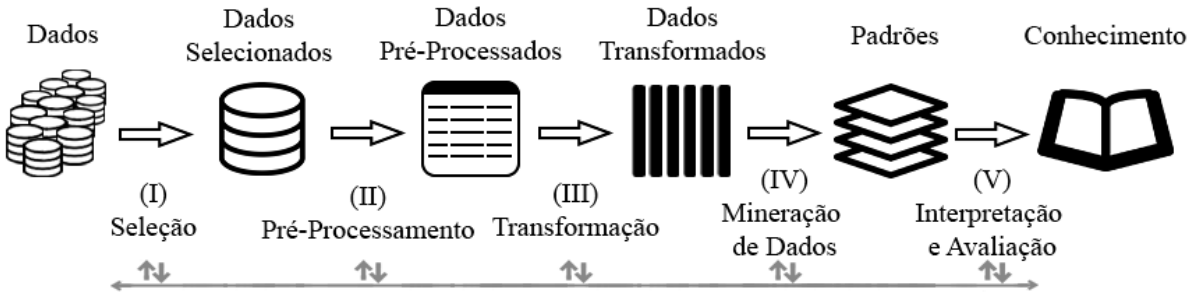
Os dados de entrada ou indicadores de observação de EWS são os aspectos relacionados ao desempenho acadêmico dos estudantes que refletem o seu risco, ou seja, dados rotineiros, disponíveis nas instituições de ensino. Preferencialmente informações atuais relacionadas às disciplinas correntes, pois, predizem melhor o desempenho acadêmico do que características passadas (HEPPEN; THERRIAULT, 2008). Entretanto detectar as situações de risco é um problema complexo, pois não existe somente uma única razão do porquê os estudantes não logram sucesso nas atividades acadêmicas, nem limites fixos nos indicadores observáveis (MÁRQUEZ-VERA et al., 2016). Abordagens estatísticas como a de Uekawa et al.(2010) e Heppen e Therriault (2008) podem não capturar toda a complexidade, pois focam-se em um conjunto de variáveis pequeno para compor um indicador estático. O emprego de técnicas de EDM tornam-se justamente de grande valia nestes sistemas, pois permitem realizar uma análise multivariada abrangendo diversas características do estudante e criam um indicador dinamicamente atualizado e moldado pelos dados.

2.2 Descoberta de Conhecimento em Bases de Dados

A inferência de novos conhecimentos a partir de conjuntos de dados é uma tarefa extensa e complexa e, para possibilitar que a partir de conjuntos de dados chegue-se em novos conhecimentos passíveis de serem compreendidos por humanos, adota-se um processo como guia à exploração dos dados até a obtenção de resultados desejados. Assim, toma-se como base o processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* - KDD), que preocupa-se em dar sentido as informações contidas nestes conjuntos de dados. Em outras palavras, ele tem como objetivo compactar grandes volumes de dados para formas mais abstratas, resumidas e genéricas que sejam possíveis de serem absorvidas por seres humanos (HAN; KAMBER; PEI, 2012).

Segundo Fayyad et al. (1996, p. 40), a Descoberta de Conhecimento é o "Processo não trivial de identificar padrões de informação, válidos, novos, potencialmente úteis e por fim compreensíveis". Esse processo é interativo e iterativo, pois envolve decisões feitas pelo usuário e ciclos de repetição de etapas do processo.

Figura 1 - Processo do KDD.



Fonte: Adaptado de Fayyad, Piatetsky-Shapiro, Smyth (1996).

A figura 1 ilustra o processo do KDD, com cada uma de suas etapas. Seguindo a figura, no início existe a **(I) Seleção de Dados**, que busca ter um entendimento inicial do domínio de aplicação e dos objetivos desejados, passando então para selecionar um conjunto de dados, seja através da seleção em um banco de dados ou focando-se em um subconjunto de um conjunto existente, deste conjunto que irá se extrair o conhecimento (FAYYAD et al., 1996).

Na etapa seguinte, em virtude das bases de dados reais serem altamente suscetíveis a ruídos, dados faltantes e inconsistentes, muitas vezes em virtude de seu tamanho ou das múltiplas fontes de dados, é realizado o **(II)Pré-Processamento** para melhorar a qualidade dos dados. Esta etapa apesar de inicial, pode consumir a maior parte do tempo do processo (WITTEN; FRANK; HALL, 2011).

A **(III)Transformação de Dados** aparece como uma etapa intermediária entre o Pré-Processamento e a Mineração de Dados, justamente por utilizar técnicas que melhor preparem os dados, em função das características do conjunto de dados e também das peculiaridades dos algoritmos que serão utilizados na etapa de DM (HAN; KAMBER; PEI, 2012). Pode-se ainda aplicar estas técnicas em diferentes porções dos dados, levando a se ter diferentes modelagens.

Na quarta etapa ocorre a **(IV)Mineração de Dados** que envolve ajustar algoritmos através da aplicação de técnicas de aprendizado de máquina, para determinar padrões dos dados observados (WITTEN; FRANK; HALL, 2011), ainda segundo Fayyad et al. (1996, p. 41), Mineração de Dados "consiste na aplicação de algoritmos específicos, sob alguma limitação aceitável de eficiência computacional, para produzir uma lista particular de padrões sobre os dados". Estes algoritmos compõem técnicas que são escolhidas e aplicadas conforme o objetivo e a tarefa que se deseja. São exemplos de tarefas: associar, agrupar e classificar, entre outros. Em especial para o contexto deste estudo trabalha-se com a tarefa de classificação, que rotula um item de acordo com rótulos pré-definidos, denominado classes.

Ao final, é difícil saber quando o modelo está suficientemente ajustado e, por conseguinte a sua corretude, bem como a qualidade com que se executou todo o processo. Entretanto é necessário que se mensure os resultados gerados pela aplicação do KDD, para poder avaliá-lo, justamente a etapa de **(V)Interpretação e Avaliação** de resultados possibilita realizar a análise da qualidade da técnica de DM, bem como todo o processo adotado anterior a ela.

2.3 Técnicas de Pré-Processamento e Transformação de Dados

O pré-processamento é uma das etapas mais importantes e dispendiosas do KDD, pois visa tratar os problemas que podem existir nos dados, eventualmente até desconhecidos, e que podem afetar a qualidade futura da análise.

A limpeza de dados é um componente fundamental no tratamento dos dados e, em especial, dois pontos devem ser ressaltados: (a) valores faltantes, que devem ser tratados removendo as amostras, preenchendo os valores manualmente, usando uma constante global, entre outras opções, e (b) dados ruidosos, que são erros aleatórios ou variância anormal dos dados, devendo ser removidos ou tratados, empregando técnicas como regressão, *clusterização*, etc. (HAN; KAMBER; PEI, 2012).

As técnicas de transformação de dados objetivam transformar ou consolidar os dados para formas mais apropriadas para a etapa de DM (HAN; KAMBER; PEI, 2012; WITTEN; FRANK; HALL, 2011), usando-se de alisamento, agregação, construção de atributos, entre outros.

Além dessas etapas básicas, é importante destacar outras técnicas específicas essenciais à realização desse trabalho, e que são descritas nas próximas subseções.

2.3.1 Ganho de Informação

Os dados podem conter inúmeros atributos, muitos dos quais irrelevantes para o objetivo em questão. Manter este tipo de atributo pode inclusive atrapalhar o algoritmo de DM que será usado, confundindo-o, assim a redução de dimensionalidade torna-se um passo importante. (PEÑA-AYALA, 2014a). Neste caso, técnicas de Seleção de Atributos podem ser aplicadas objetivando a seleção e remoção dos atributos irrelevantes para o problema (HAN; KAMBER; PEI, 2012). A técnica Ganho de Informação (*Information Gain* - IG) justamente objetiva a seleção e remoção de atributos irrelevantes a fim de obter uma representação dos dados reduzida, com o mínimo de atributos, mas que a distribuição de probabilidade dos dados seja igual ou próxima a do conjunto original com todos os atributos.

O funcionamento da técnica ocorre por meio da atribuição de um peso para cada atributo da base de dados. Esse peso é obtido através do cálculo de Entropia (SHANNON, 1948). Dessa forma, cada atributo da base de dados é avaliado e os que diminuem a impureza do conjunto, apresentando por exemplo seus valores para somente uma classe, recebem um valor maior. Objetiva-se, com isto, selecionar os atributos mais representativos da base de dados a fim de remover os demais (FELDMAN; SANGER, 2006; HAN; KAMBER; PEI, 2012).

Tecnicamente o IG é definido como a diferença entre a entropia do conjunto original e o atributo que se avalia. Através da seguinte função obtém-se o valor de IG para um atributo:

$$IG(A) = Info(D) - Info_A(D), \tag{2.1}$$

onde *A* refere-se a um atributo e *D* ao conjunto de dados. *Info(D)* é definido da seguinte forma:

$$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i), \tag{2.2}$$

sendo que p_i é a probabilidade que uma certa amostra em *D* pertença a classe C_i, estimado por $\frac{|C_{i,D}|}{|D|}$, e *m* é o número de distintos valores da classe. $Info_A(D)$ é definido por:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), \tag{2.3}$$

onde v é o número de valores distintos de *A* e D_j o conjunto de dados particionado por *J*.

De posse do valor de cada atributo pode-se ordena-los de forma decrescente e escolher um número determinado de atributos do topo da lista, descartando os demais.

2.3.2 SMOTE

Para contornar o problema da desbalanceamento entre as classes, fazendo inclusive com que certos classificadores possam tender em favor da classe majoritária (CHAWLA et al., 2002), pode-se reamostrar o conjunto de dados, realizando *undersampling* que remove amostras da classe majoritária, ou *oversampling* que adiciona amostras para a classe minoritária, ou ainda uma mistura de ambas abordagens.

A técnica de *oversampling* SMOTE (*Synthetic Minority Over-sampling Technique*) proposto por Chawla et al. (2002) busca criar amostras sintéticas da classe minoritária fazendo com que os classificadores criem regiões de decisões maiores e menos específicas.

O seu funcionamento inicia-se buscando um número determinado de vizinhos mais próximos de cada amostra, posteriormente adiciona-se uma nova amostra sintética para cada vizinho da amostra. Esta amostra sintética será criada pela diferença entre os dois vetores (amostras), multiplicada por um número randômico entre 0 e 1, fazendo com que assim ela fique em algum lugar no caminho que vai de uma amostra até um de seus vizinhos mais próximos. A quantidade de vizinhos e número de amostras a se criar são parâmetros passados para o algoritmo.

2.3.3 Normalização de Dados

As unidades de medidas dos atributos tendem a influenciar os algoritmos de mineração de dados, assim atributos que têm métricas que apresentam valores maiores em relação a outros podem vir a ter um maior efeito ou peso. Visando-se a independência de unidades de medidas ou atributos com faixas de valores diferentes, pode-se transformar os dados, para que todos os atributos possam estar dentro de um mesmo espaço de valores.

As técnicas de normalização de dados atuam independentemente em cada atributo fazendo com que o intervalo de dados seja igual para todos. Desta forma, os atributos daquele conjunto de dados passam a ter um mesmo peso. Uma técnica de normalização muito comum de ser empregada é a que traduz os dados para o intervalo [0,1] ou [-1,1] através dos mínimos e máximos do atributo, à qual é atingida realizando a diferença da amostra em relação ao valor mínimo daquele atributo, dividido pela diferença entre o maior e menor valor do atributo (WITTEN; FRANK; HALL, 2011).

2.4 Técnicas de Classificação de Dados

A classificação é uma importante tarefa, utilizada para predizer rótulos categóricos dos dados (HAN; KAMBER; PEI, 2012). Nesta tarefa, o funcionamento das técnicas de DM é definido por dois passos:

a) Primeiro, o algoritmo é alimentado com dados de classes conhecidas e então, ele descobre e aprende os padrões. Esta etapa é chamada de aprendizagem ou treinamento;

b) Num segundo momento, novas amostras de dados para teste são apresentadas a este algoritmo, dito treinado, assume-se temporariamente que não se sabe a classe das amostras, e uma saída é então obtida do algoritmo, representando as classes escolhidas por ele para aquelas amostras. Nesta etapa os dados utilizados devem ser independentes

de todo o processo do KDD, exceto quando for necessário adequá-los aos dados de treinamento, devido ao algoritmo.

De acordo com as características de cada classificador pode-se escolher as melhores alternativas para tratar o problema de classificação dos dados, assim as três subseções seguintes trazem algoritmos de DM que possuem características peculiares a cada um.

2.4.1 Naïve Bayes

O algoritmo *Naïve Bayes* (NB) é uma técnica dita ingênua por considerar que os atributos representativos são condicionalmente independentes uns dos outros, mesmo assim é amplamente utilizada em função de sua simplicidade e velocidade nas etapas de treinamento e teste (CHAKRABARTI, 2003), além disso já foi utilizado em diversos trabalhos da área de EDM (BARBER; SHARKEY, 2012; ER, 2012; MÁRQUEZ-VERA et al., 2016).

Na etapa de treinamento ele funciona através da atribuição de probabilidades aos atributos em função das classes. Já na etapa de testes é feita a probabilidade *a posteriori*, definindo a probabilidade de uma certa amostra pertencer a uma dada classe (WITTEN; FRANK; HALL, 2011), este cálculo é realizado através da função abaixo:

$$Pr[H|E] = \frac{Pr[E|H]\,Pr[H]}{Pr\,[E]}, \tag{2.4}$$

onde $Pr[E]$ ou $Pr[H]$é a probabilidade *a priori* de acontecer *E* ou *H*, $Pr[E|H]$ a probabilidade de *E* ocorrer condicional a *H* e $Pr[H|E]$ a probabilidade *a posteriori* de *H* ocorrer condicional a *E*. *H* é uma hipótese a ser testada, diga-se a classe que se deseja verificar, e *E* é a combinação de atributos da amostra a se verificar, desta forma se estendendo para E_1, E_2 ... E_n, com *n* sendo a quantidade de atributos existentes na amostra, menos a classe. Deve-se a aplicar a função para cada valor distinto da classe e verificar aquele que apresenta maior *likelihood* para uma dada amostra para decidir a classe a que pertence.

A fim de evitar que atributos em que a probabilidade condicional seja 0, levando a uma multiplicação por 0, e causando problemas na fórmula, em geral utiliza-se a contagem de ocorrências de valores iniciando em 1 para todos os atributos (WITTEN; FRANK; HALL, 2011). Para atributos faltantes no teste, basta ignorá-lo na aplicação da função e para faltantes no treino ignora-se ele na contagem.

No caso de haver atributos numéricos pode-se considerar que eles possuem uma distribuição normal e calcular a função densidade de probabilidade, conforme abaixo:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{2.5}$$

onde x é o atributo da amostra a ser testada, μ e σ, respectivamente a média e o desvio padrão amostral do atributo calculado sobre os dados de treino.

2.4.2 C4.5

As Árvores de Decisão (*Decision Trees* - DT) são populares pelo fato de o conhecimento adquirido ser entendível por humanos, por conseguir manusear dados de altas dimensões e pelas etapas de treino e teste serem fáceis e rápidas (HAN; KAMBER; PEI, 2012). Uma DT é como uma estrutura em árvore de um fluxograma, em que cada nodo é um teste em um atributo; cada ramo, uma saída do teste e cada folha uma classe. A classificação de uma amostra se dá por

percorrer esta árvore testando os atributos da amostra contra as condições dos nodos e prosseguindo pelo ramo correto até chegar na folha que indica a classe.

A forma de treinamento de uma DT é feita por recursivas divisões do conjunto de treinamento, conforme a árvore é construída. Um algoritmo de indução de uma DT é dado no quadro 1.

Quadro 1 - Algoritmo básico para indução de DT.

```
Algoritmo: Gera_árvore_de_decisão
Entrada: D – conjunto de treinamento com as classes associadas
         E – Lista de atributos
    M – critério de divisão que melhor particiona as amostras dentro de classes individuais.
Saída: Árvore de Decisão
Método:
    1. crie um nodo N;
    2. se amostras em D são todas da mesma classe H, então
    3.       retorne N como um nodo folha rotulado com a classe H;
    4. se E é vazio então
    5.       retorne N como nodo folha rotulado com a classe majoritária de D;
    6. aplique M(D,E) para encontrar o melhor critério_de_divisão
    7. rotule nodo N com o critério_de_divisão;
    8. se atributo de divisão é de valor discreto e árvore não restrita à binária então
    9.       E = E - atributo_de_divisão
    10. para cada resultado j do critério_de_divisão faça
    11.      deixe Dj ser o conjunto de amostras em D satisfazendo o resultado j
    12.      se Dj é vazio então
    13.            anexe a folha rotulada com a classe majoritária em D para o nodo N
    14.      caso contrário
    15.            anexe o nodo retornado por Gera_árvore_de_decisão(Dj,E) para o nodo N;
    16. retorne N
```

Fonte: Adaptado de Han, Kamber e Pei (2012).

Para o treinamento da DT pode-se utilizar a técnica C4.5 (QUINLAN, 1993), como visto em trabalhos da área (ER, 2012; HU; LO; SHIH, 2014). Ela é sucessora da técnica ID3 (QUINLAN, 1986), adotando uma estratégia *top-down*, com busca gulosa através do espaço de possíveis DT. Para o uso do C4.5, deve-se utilizar a medida Razão de Ganho (*Gain Ratio* - GR) para a escolha do critério de divisão no algoritmo do quadro 1.

A medida GR é uma extensão da medida IG[2] (QUINLAN, 1986) utilizada no ID3, que tenta sobrepujar o problema do IG de selecionar para divisão, atributos que têm um grande número de valores distintos, levando a separar em pequenos conjuntos, que não agregam à tarefa de classificação. Para tanto, ela incorpora uma função que é sensível ao quanto a informação é ampla e uniforme (MITCHELL, 1997), definida pela seguinte fórmula:

$$SplitInfo_E(D) = -\sum_{j=1}^{v} \frac{|D_j|}{D} \times log_2\left(\frac{|D_j|}{D}\right). \quad (2.6)$$

Este valor representa o potencial de informação gerado por dividir o conjunto de dados D em v partições, correspondendo a v resultados do teste no atributo E. Nota-se que esta medida é a entropia de D em relação a E, diferente do IG que é em relação a classe.

[2] Ganho de Informação (*Information Gain* – IG): Neste casso refere-se somente a medida apresentada na equação 2.1, e não a toda técnica de Seleção de Atributos apresentada na subseção 2.3.1.

Desta forma pode-se obter o GR através da seguinte função:

$$GR(E) = \frac{IG(E)}{SplitInfo(E)}. \quad (2.7)$$

Atenta-se que o uso do GR leva a um problema quando o denominador tende a 0, levando a um comportamento indefinido. Desta forma, pode-se adotar uma limitação que dita que o valor de IG deve ser maior ou igual a média de IG de todos os atributos examinados (WITTEN; FRANK; HALL, 2011).

Para contornar problemas de *overfitting*[3] e a construção de ramos modelando possíveis anomalias e *outliers*[4], a técnica C4.5 adota uma estratégia de poda chamada pessimista que usa a estimativa da taxa de erro em cima do conjunto de treinamento para decidir os cortes (HAN; KAMBER; PEI, 2012). O processo funciona por considerar o conjunto de dados que chega a cada nodo e então imagina-se que a classe majoritária é escolhida para representar aquele nodo, este procedimento dá um certo número de erros do total do número de instâncias, assim se a taxa de erro diminuir mantém-se a poda.

Quando os atributos são numéricos, deve-se primeiro transformá-los em nominais para serem aplicados no classificador C4.5. A abordagem tomada pela técnica funciona ordenando as amostras em ordem crescente e rotulando com um identificador único as amostras que repetem o seu valor de classe, sem alterná-la por uma determinada contagem; caso a classe mude depois de uma determinada contagem, associa-se um novo identificador para aquelas amostras, e assim por diante, até que todos os dados tenham sido transformados (WITTEN; FRANK; HALL, 2011).

2.4.3 Redes Neurais Artificiais de Múltiplas Camadas *Feedforward* com *Backpropagation*

As Redes Neurais Artificiais (RNAs) são algoritmos inspirados nos neurônios biológicos, constituídos por unidades interconectadas que possuem um peso associado. Tais unidades fundamentais, analogamente a sua inspiração, são chamadas de neurônios e as suas conexões de sinapses (HAYKIN, 2001).

Adicionalmente, os neurônios possuem um *bias*, com o efeito de aumentar ou diminuir a entrada líquida da função de ativação, e uma função de ativação que, por sua vez, é utilizada para restringir a amplitude da saída do neurônio, usualmente definida por uma função logística sigmóide. A função de ativação, sendo não linear e diferenciável, permite que as RNAs modelem problemas de classificação que são linearmente inseparáveis (MITCHELL, 1997).

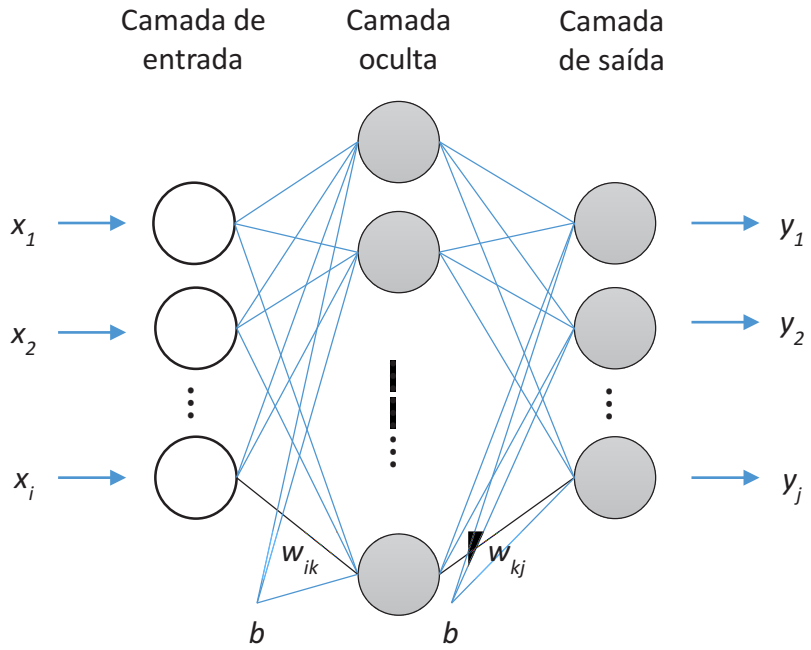
Diversos neurônios que estão conectados entre si compõem as RNAs. A característica *feedforward* se refere à direção com que os dados fluem através desses neurônios, sendo da esquerda para a direita e sem ocorrer ciclos. Já a característica de múltiplas camadas refere-se a como estão dispostos os neurônios na RNA, sendo por camadas e existindo então uma ou mais camadas entre a de entrada e a de saída (HAYKIN, 2001).

---

[3] *Overfitting*: ocorre quando o modelo, na fase de treinamento, incorpora fortemente os detalhes dos dados, passando a não conseguir generalizar e obter bons resultados quando aplicado a outros conjuntos de dados (WITTEN; FRANK; HALL, 2011).

[4] *Outliers:* amostras que não estão em conformidade com o comportamento geral dos dados (HAN; KAMBER; PEI, 2012)

Figura 2 - Exemplo de uma RNA de múltiplas camadas *feedforward*.



Fonte: Elaborado pelo autor.

Na figura 2 apresenta-se um modelo de RNA com duas camadas computacionais, sendo que a primeira camada é dita de entrada, a do meio camada oculta, e a última, camada de saída, *x* e *y* referem-se, respectivamente, as entradas e saídas, *b* à entrada do *bias*, W_{ik} o peso associado à sinapse que liga o neurônio *i*, da camada de entrada, ao neurônio k, da camada oculta, e W_{kj} o peso da sinapse do neurônio *k* em relação ao neurônio *j*, da camada de saída.

A aprendizagem da RNA ocorre pelos ajustes dos pesos, com um sinal de erro obtido pela comparação entre o valor predito pela RNA e o valor correto da classe para as amostras, com isso pode-se corrigir os pesos da RNA até que ela chegue num estado de aprendizagem.

Um algoritmo para realizar os ajustes dos pesos da RNA de múltiplas camadas *feedforward* é o *Backpropagation*. Esta configuração de RNA já mostrou bons resultados no trabalho de Lykourentzou et al. (2009).

O processo de treinamento da RNA com este algoritmo inicia-se com a decisão do número de camadas ocultas e a quantidade de neurônios que as compõem, além de outros parâmetros abordados no decorrer do texto. Cabe denotar que não há um consenso das melhores topologias e parametrizações, sendo empregado em geral metodologias por tentativa e erro (HAN; KAMBER; PEI, 2012). Após estas escolhas, os pesos iniciais dos neurônios são aleatoriamente gerados, comumente num intervalo de -1 a 1 ou -0,5 a 0,5. O processo continua com a alimentação das amostras de treinamento, para cada amostra todos os seus atributos são apresentados para a camada de entrada e calculados em cada neurônio das camadas subsequentes. Cada passagem de todo o conjunto de treinamento pela RNA é chamada de época.

O objetivo do treinamento é diminuir o erro da RNA ajustando os pesos. Assim, após a passagem de cada amostra, que irá gerar uma saída do classificador, ocorre o processo de *Backpropagation*, buscando encontrar o erro mínimo global através do método de Gradiente Descendente (GD). Para tanto, o primeiro passo é obter o erro, confrontando o resultado

esperado com o obtido na saída da RNA, conforme a fórmula abaixo para cada neurônio da saída.

$$Err_j = y_j(1 - y_j)(T_j - y_j), \tag{2.8}$$

onde, *y* é a saída e *T* é o valor conhecido da classe. Denota-se que $y_j(1 - y_j)$ é a derivativa de uma função logística.

O erro da camada anterior também é calculado, portanto para o cálculo de erro de um neurônio *k* da camada oculta utiliza-se a função:

$$Err_k = y_k(1 - y_k) \sum_j Err_j w_{kj}, \tag{2.9}$$

onde w_{kj} é o peso da conexão da unidade *k* para a unidade *j* na camada posterior e Err_j é o erro da unidade *j*.

Os pesos de cada sinapse, incluindo o *bias,* são então atualizados seguindo as funções abaixo, que obtêm a variação e depois atualiza-os:

$$\Delta w_{ik}(t) = (l) Err_k y_i + (m) \Delta w_{ik}(t - 1), \tag{2.10}$$

$$w_{ik} = w_{ik} + \Delta w_{ik}, \tag{2.11}$$

onde w_{ik} é o peso da unidade *i,* na camada anterior, até *k*. Nota-se nesta função a inclusão dos parâmetros *l* e *m*, respectivamente taxa de aprendizado e *momentum*, que são duas constantes que vão de 0 a 1, bem como *t* e *t-1* para representar, respectivamente, o gradiente atual e passado.

A finalização do processo de treinamento ocorre quando a variação dos pesos nas épocas anteriores alcança um patamar, a saída da função de erro diminui até um valor desejado ou uma quantidade pré-especificada de épocas ocorre.

Cabe observar que a inclusão da taxa de aprendizado tem por objetivo ajudar a evitar o algoritmo ficar preso a mínimos locais, assim se a taxa de aprendizagem for muito baixa o aprendizado irá ocorrer lentamente e pode ficar preso a um mínimo local, do contrário oscilações entre soluções inadequadas podem ocorrer, bem como nunca chegar numa solução ótima (HAYKIN, 2001). Adicionalmente o *momentum* fará com que o processo de busca tenha uma inércia, evitando excessivas oscilações (ROJAS, 1996), com isto o gradiente anterior (*t-1*) é somado com o peso atual, ponderando-se então com a sua constante.

Estas duas parametrizações apesar de serem benéficas para o algoritmo, incorrem em um problema que é a escolha dos seus valores. Como elas são altamente dependentes da tarefa de aprendizagem, não desenvolveram-se abordagens generalistas (ROJAS, 1996).

Por fim, o funcionamento da RNA na classificação de novas amostras se dá pela passagem dos atributos destas através dos neurônios da camada de entrada, que serão então calculados nos neurônios das camadas ocultas, até chegar nos neurónios da camada de saída, os quais irão determinar a classe que pertence a amostra. Cabe denotar que, para atributos com valores discretos, pode-se separar com seus valores únicos, múltiplos neurônios na camada de entrada e tornar a entrada destes binária.

O cálculo no neurônio é feito pela multiplicação do valor que está entrando por cada sinapse pelos seus respectivos pesos, incluindo o *bias* que em geral tem um valor fixo de entrada. Após soma-se todos estes termos e aplica-se a função de ativação escolhida.

2.5 Interpretação e Avaliação de Modelos de Classificadores

O desempenho dos classificadores, bem como todo o processo, deve ser avaliado segundo alguma métrica, visando determinar qual o melhor modelo de classificador e se ele está obtendo êxito em classificar novas amostras depois de treinado.

Para tanto, separa-se uma quantidade de amostras do conjunto de dados para apresentar para o classificador como teste, esperando assim poder confrontar a saída obtida do classificador, leia-se a classe à qual ele tomou para associar à amostra, com a classe que realmente a amostra pertence, obtida do conjunto de dados original. Existem diversas técnicas para fazer a separação entre as bases de treino e teste, como *cross-validation*, *leave-one-out*, *bootstrap*, etc., quando há um volume significativo de dados para o domínio do problema é comum separar com base num percentual do total, em geral a base de treinamento sendo maior que a de teste (WITTEN; FRANK; HALL, 2011). Por fim, o cruzamento da saída do classificador com o valor da classe real obtida do conjunto de dados, permite gerar a Matriz de Confusão, apresentada na tabela 2, que contabiliza as amostras classificadas correta e incorretamente pelo modelo, em função da classe esperada.

Tabela 2 - Matriz de Confusão.

		Classe Predita	
		Positiva	Negativa
Classe Correta	Positiva	Verdadeiros Positivos (VP)	Falsos Negativos (FN)
	Negativa	Falsos Positivos (FP)	Verdadeiros Negativos (VN)

Fonte: Adaptado de Witten, Frank e Hall (2011).

Com os valores da Matriz de Confusão pode-se calcular métricas mais descritivas que permitem avaliar e comparar o desempenho do classificador, como: (a) acurácia, que permite a avaliação geral de acertos, considerando tanto amostras negativas quanto positivas, (b) *precision*, que indica quanto das instâncias classificadas como uma classe são realmente daquela classe, sem considerar quantas amostras deixaram de ser classificadas com aquela classe, (c) *recall*, que mensura a quantidade de amostras que foram rotuladas como uma classe, mas sem se importar com a quantidade de amostras que foi rotulada incorretamente para aquela classe, (d) e a medida F_1 que traz uma média harmônica das medidas *precision* e *recall,* mensurando assim precisão e completude do classificador (HAN; KAMBER; PEI, 2012). Cabe lembrar que certas medidas podem ser calculadas sob a perspectiva de uma única classe, possibilitando analisar a tendência do modelo naquela métrica para cada classe. As suas formulações são trazidas abaixo, juntamente com o cálculo da medida F_1 com a média ponderada sobre as duas classes.

$$Acurácia = \frac{VP+VN}{VP+VN+FP+FN}, \tag{2.12}$$

$$Recall_{Pos} = \frac{VP}{VP+FN}, \quad Recall_{Neg} = \frac{VN}{VN+FP}, \tag{2.13}$$

$$Precision_{Pos} = \frac{VP}{VP+FP}, \quad Precision_{Neg} = \frac{VN}{VN+FN}, \tag{2.14}$$

$$F_{1-Pos} = 2\frac{Precision_{Pos}*Recall_{Pos}}{Precision_{Pos}+Recall_{Pos}}, \quad F_{1-Neg} = 2\frac{Precision_{Neg}*Recall_{Neg}}{Precision_{Neg}+Recall_{Neg}}, \tag{2.15}$$

$$F_1 = \frac{F_{1-Pos} * (VP+FN) + F_{1-Neg} * (VN+FP)}{(VP+VN+FP+FN)}, \quad (2.16)$$

onde *Pos* denota a classe positiva e *Neg* a classe negativa. Cabe observar que a medida $F_{1,}$ pode receber outros valores para a constante multiplicativa tanto no denominador como no numerador, caso se deseje dar mais ênfase para uma das métricas, usualmente é utilizado com os valores apresentados fazendo com que o *recall* e a *precision* tenham o mesmo peso (HAN; KAMBER; PEI, 2012).

2.6 Considerações

Os EWS são sistemas genéricos e que podem ser aplicados à área educacional, sua concepção básica é através do simples monitoramento de variáveis, assim o campo de DM, através de sua aplicação em dados educacionais pelo nome de EDM, colabora oferecendo análises preditivas multivariadas. Em especial, o tópico de pesquisa dentro de EDM que este trabalho se atem é a Modelagem do Desempenho do Estudante, conforme foi apresentado na tabela 1.

A geração do indicador para alimentar o EWS, através da aplicação de DM envolve a aplicação de diversas técnicas, e por este motivo deve ser estruturada em torno de um processo, tomando-se assim o KDD como linha guia para gerar o conhecimento necessário. O pré-processamento é visto dentro deste processo como etapa fundamental na obtenção de bons resultados, razão pela qual foi dado ênfase a ele destacando-o, juntamente com a etapa de transformação de dados, numa subseção separada.

A subseção de técnicas de classificadores trouxe o cerne da relação entre EDM e EWS, pois permite criar modelos para determinar, preditivamente, se um certo aluno está em risco ou não. Para tanto, três técnicas já estudadas nas áreas de EDM e EWS foram abordadas, cada qual com uma vantagem sobre os outras. Inicialmente foi apresentado o classificador NB que é reconhecido por sua simplicidade, velocidade e por basear-se nas probabilidades (CHAKRABARTI, 2003), após as DT com o algoritmo C4.5 e a característica de trazer regras possíveis de serem compreendidas por seres humanos com fundamentação na teoria da entropia (QUINLAN, 1993) e por fim as RNAs *feedforward* com o algoritmo *Backpropagation,* apresentando-se como uma técnica tolerante a ruído, passível de solução de problemas não lineares e que consegue aproximar qualquer função matemática (HAYKIN, 2001).

Por fim, a avaliação de resultados objetivou demonstrar métricas para medir o quão assertivo são os modelos de classificadores gerados com estas técnicas, possibilitando mensurar os erros e acertos nas escolhas de suas parametrizações e modelagens.

3 TRABALHOS RELACIONADOS

Este capítulo objetiva situar este trabalho frente à literatura. Na subseção 3.1 apresenta-se dois trabalhos que buscam através de métodos estatísticos definir a correlação entre as variáveis de entrada e a predição do desempenho dos estudantes, um trazendo o aspecto temporal e outro o impacto de modelos generalistas contrapondo com modelos específicos por disciplina. Já na subseção 3.2 são abordados dois trabalhos que realizam análises preditivas com DM, durante o período das aulas.

Por fim, a subseção 3.3 aborda brevemente outros estudos de menor importância para o contexto deste trabalho, pontuando críticas a respeito de cada um. E na subseção 3.4 são trazidas considerações e comentários frente aos principais estudos, considerando EDM empregado em EWS.

3.1 Abordagens Estatísticas à Análise do Desempenho dos Estudantes

O trabalho de Gašević et al. (2016) apesar de não focar nos aspectos temporais do problema, assim não fazendo predições enquanto ocorrem as aulas, traz um estudo singular verificando o impacto de modelos generalistas em contraponto com os separados por disciplina.

Os dados são provenientes de uma universidade, sendo que as turmas das 9 disciplinas selecionadas possuíam um perfil de alto insucesso acadêmico, são de primeiro ano, cada uma com mais de 150 alunos e de diferentes áreas de conhecimento, reunindo assim 4134 estudantes extraídos de um histórico de cinco anos de dados.

As variáveis vieram em sua maioria do AVA, que foi utilizado como apoio às aulas presenciais, representando assim o perfil do estudante, através do uso de fóruns, *logins,* submissões de trabalho, *quizzes*, etc. Adicionalmente usou-se de informações pessoais como idade, casa em área rural, gênero, entre outras. O objetivo de definir o sucesso acadêmico dos estudantes foi estudado através de duas modelagens em função da classe: (a) uma contendo uma variável de saída contínua indo de 0% a 100% e (b) outra modelagem com uma variável categórica, com dois valores possíveis: passou (atingiu 50% ou mais) ou falhou (atingiu menos de 50% ou desistiu).

A análise prosseguiu através do uso de ferramentas estatísticas, visando verificar a relação das variáveis de entrada com a de saída. Para as variáveis contínuas de saída utilizou-se regressão linear múltipla, criando dois novos modelos, um somente com variáveis de características pessoais dos estudantes e outro adicionando dados do AVA, para cada um destes modelos, utilizou-se os dados de todas as 9 disciplinas em conjunto e também separados por disciplina. Para as variáveis de saída categórica os mesmos modelos foram gerados, mas utilizando regressão logística.

Os resultados apontaram que existe uma diferença muito grande da intensidade e do uso dos recursos do AVA entre as diferentes disciplinas, mesmo naquelas da mesma área de conhecimento, bem como a adição dos dados do AVA, quase em todos os cenários melhorou a análise. Interessante perceber que as variáveis mais importantes para o modelo geral não foram as mesmas para os modelos por disciplina, segundo os autores, isto se deve ao fato que algumas variáveis são gerais para todos as disciplinas, e que por outro lado outras que são mais específicas de determinadas disciplinas acabam tendo maior peso. Ao fim, é pontuado que os modelos gerais obtiveram resultados piores do que quando utilizado modelos separados por disciplina.

O estudo desenvolvido em You (2016) visa encontrar uma correlação entre os dados extraídos do AVA e o possível sucesso acadêmico de 530 estudantes que realizaram uma disciplina *on-line*.

A disciplina em questão foi disponibilizada por uma universidade e era composta por 13 unidades de conteúdo, cada qual com conteúdo específico sendo liberado ao longo do tempo. Durante o decorrer da disciplina duas provas foram realizadas de forma presencial no campus e quatro trabalhos a serem entregues *on-line* foram solicitados. Alunos que desistiram ou não terminaram a disciplina foram removidos do conjunto de dados.

Duas modelagens de dados foram adotadas, uma que considerou os dados das semanas 1 a 8, e outra que inclui os dados das semanas 1 a 15, ambas testadas contra o resultado final da nota do estudante. A investigação foi conduzida através de análise de correlação com regressão hierárquica.

As conclusões do autor são que as variáveis escolhidas foram o suficiente para correlacionar com a nota final, bem como as predições antecipadas também foram possíveis de serem feitas utilizando somente as seguintes: assiduidade, número de *logins*, submissões atrasadas, leitura de avisos e nota da prova intermediária. Ele ainda explicitou como vantagem que o estudo dele não se utiliza de variáveis invasivas ou limitadas em adquirir o comportamento atual dos estudantes.

3.2 EWS Utilizando-se de Técnicas de DM

O trabalho de Hu, Lo e Shih (2014) traz o desenvolvimento de um EWS focando-se no aspecto temporal das variáveis, necessário para se trabalhar com predição enquanto ocorre a disciplina.

Para tanto, foram utilizados os dados de alunos que cursaram uma disciplina ofertada em dois semestres consecutivos pelo mesmo professor. Os dados foram extraídos dos registros de uso do AVA.

As variáveis extraídas, ao total 14, estavam relacionadas a: número de acessos ao AVA; tempo *on-line*; média de tempo por sessão; tempo de leitura e uso dos materiais da disciplina; atraso e não entrega de tarefas; participação e réplicas de fórum. O atributo usado como classe descrevia se o aluno passou ou falhou na disciplina, sendo que a base toda continha 300 estudantes dos quais somente 16 falharam.

Os autores enfatizam que é difícil construir um sistema que preveja o comportamento do estudante durante o semestre com base no seu comportamento passado, assim eles criaram três conjuntos de dados extraindo-os do início da disciplina até a quarta, oitava e décima terceira semana de aula, respectivamente.

O estudo incluiu a reamostragem dos registros utilizando *undersampling* e *oversampling* randômico a fim de igualar ambas as classes. Para cada conjunto de dados, 30 ciclos de reamostragem foram executados, variando a semente dos números aleatórios para escolher diferentes amostras a cada ciclo. Além disto, na execução dos experimentos aplicou-se *10-fold cross-validation* para cada base, sempre agrupando os resultados pela média.

Os classificadores utilizados foram: C4.5, CART (BREIMAN et al., 1984), Regressão Logística (SUMNER; FRANK; HALL, 2005), adicionalmente em conjunto com estes classificadores utilizou-se a técnica *AdaBoost* (FREUND; SCHAPIRE, 1996). A melhor técnica, segundo os autores, foi o CART com o *AdaBoost* que alcançou na primeira predição,

na 4ª semana, uma acurácia de 0,972 (0,009 erro tipo I[5] e 0,048 erro tipo II[6]), e na segunda predição, na 13ª semana, uma acurácia de 0,980 (0,000 erro tipo I[5] e 0,041 erro tipo II[6]).

Outro trabalho relevante é o de Lykourentzou et al. (2009), que traz uma predição focada em disciplinas *on-line* ofertados por uma universidade da Grécia. Os dados são referentes a duas disciplinas da área de computação, que consistem em 8 unidades cada, sendo ofertados duas vezes por ano. Ao longo das 7 unidades, diversos testes são aplicados e na unidade 8 um exame final é exigido para, em conjunto com as outras notas, determinar o sucesso acadêmico ou não, do estudante.

O número total de estudantes, oriundos das 3 edições de cada disciplina, é de 193, sendo que 109 terminaram alguma das disciplinas. As variáveis utilizadas são relacionadas a aspectos pessoais como gênero, experiência de emprego, nível educacional, entre outras; e a dados extraídos do AVA, como atividades da sessão e notas dos testes. A variável de saída prediz se o estudante irá evadir ou não.

As técnicas empregadas foram RNA com *Backpropagation,* Máquina de Vetores de Suporte (*Support Vector Machine* - SVM) (BURGES, 1998), e *probalistic ensemble simplified fuzzy* ARTMAP (LOO; RAO, 2005), além de uma técnica desenvolvida pelos autores que emprega a saída combinada destes três classificadores para definir a nova saída. Os dados da última edição das disciplinas foram selecionados para teste, resultando em 39 alunos para uma disciplina e 91 para a outra, utilizando o restante para treino. Para a primeira predição, os algoritmos foram alimentados somente com os dados pessoais dos estudantes, após foi adicionado ao modelo os dados extraídos do AVA da unidade 1, em seguida os dados que descreviam as atividades da unidade 2, e assim por diante até se obter a predição da unidade 7, de cada disciplina.

As conclusões dos autores são que o uso dos dados extraídos do AVA melhoraram a predição e o melhor classificador foi aquele construído por eles, o qual considerava o aluno como evadido se ao menos uma das três técnicas de base apontava ele como evadido. A abordagem tomada neste algoritmo resultou numa acurácia de 85% (74% *recall* e 93% *precision* para a classe evadido) na primeira unidade de uma das disciplinas, chegando a ~97% de acurácia (~95% *recall* e 100% *precision* para a classe evadido) na quarta unidade da mesma disciplina. Para a outra disciplina o mesmo algoritmo alcançou na primeira unidade 75% de acurácia (~90% *recall* e ~65% *precision* para a classe evadido) e chegou a ~96% de acurácia (~100% *recall* e ~90% *precision* para a classe evadido) na quarta unidade.

3.3 Outros Trabalhos com EWS

Estudos que possuem uma ligação mais fraca com o objetivo deste estudo, em função do modo que a predição é gerada ou o contexto dos dados empregados, são apresentados na tabela 3 e ao longo desta subseção. Na tabela 3 também foi incluído os quatro trabalhos citados nas duas seções anteriores, além disso, destaca-se a coluna "Estuda a generalização dos modelos" para apontar os trabalhos que estudam a abordagem generalista entre disciplinas, ou aqueles que o façam, mas não dentro da disciplina, sinalizados com o rotulo "Parcialmente".

---

[5] Erro tipo I: FP/(FP+VN) (HU; LO; SHIH, 2014), para maiores informações das abreviaturas vide subseção 2.5.

[6] Erro tipo II: FN/(VP+FN) (HU; LO; SHIH, 2014), para maiores informações das abreviaturas vide subseção 2.5.

Tabela 3 - Trabalhos relacionados.

Trabalho	Predição enquanto as aulas estão em andamento	Foco da predição: Aluno na disciplina	Usa dados com riscos de privacidade	Trata desbalanceamento	Usa dados de AVAs	Estuda a generalização dos modelos	Aborda dados de estudantes de nível superior	Técnicas			
								RNA	NB	DT	Outros
Kampff (2008)	X	X	X	-	X	Parcialmente	X	-	-	X	Regras de Associação
Lykourentzou et al. (2009)	X	X	X	-	X	-	X	X	-	-	*Fuzzy ARTMAP*, SVM, *Ensemble*
Barber e Sharkey (2012)	X	-	X	-	X	-	X	-	X	-	Regressão Logística
Er (2012)	X	X	-	-	X	-	X	-	X	X	*K-Star*, *Ensemble*
Halawa, Greene e Mitchell (2014)	X	X	-	-	-	-	-	-	-	-	Manual
Hu, Lo, e Shih (2014)	X	X	-	X	X	-	X	-	-	X	Regressão Logística, *AdaBoost*
Jayaprakash et al. (2014)	X	-	X	X	X	Parcialmente	X	-	X	X	Regressão Logística, SVM/SMO
Lara et al. (2014)	X	X	-	-	X	Parcialmente	X	-	-	-	Distância Euclidiana
Fonti (2015)	X	X	X	-	-	-	-	-	-	-	*Inductive Logic Programming*
Villagrá-Arnedo et al. (2015)	X	X	-	-	-	-	X	-	-	-	SVM
Gašević et al. (2016)	-	X	X	-	X	X	X	-	-	-	Regressão Linear Múltipla e Regressão Logística
Márquez-Vera et al. (2016)	X	-	X	X	-	-	-	-	X	X	*Classification Rules*, *Instance-based Learning*, Programação Genética, *Interpretable Classification Rule Mining*, SVM
Sanchez-Santillan et al. (2016)	X	X	-	-	X	Parcialmente	-	-	-	X	JRIP, *BayesNet*
Xing et al. (2016)	X	X	-	-	-	-	-	-	-	X	*Stacking Ensemble*, *Generalized Bayesian Network*
You (2016)	X	X	-	-	X	-	X	-	-	-	Regressão Hierárquica

Fonte: Elaborado pelo autor.

O primeiro trabalho apresentado na tabela 3 é o de Barber e Sharkey (2012) que usa-se de dados do ensino superior, com variáveis extraídas do AVA. Pouco detalhamento é dado sobre a definição do conjunto de dados, além disto os modelos priorizaram as variáveis não dependentes do tempo.

Er (2012) apresenta um trabalho utilizando NB, C4.5 e K-Star, com predições feitas em três períodos específicos, sem utilizar variáveis não dependentes do tempo. Pouco cuidado é dado no critério de seleção dos alunos, tomando-se uma abordagem randômica, além disto os períodos de predição sempre contemplaram ao menos uma das notas das avaliações, desta forma somente quando a disciplina completou a sua quarta semana de aula foi realizada a primeira predição.

Fonti (2015) apresenta uma tese de doutorado focada no desenvolvimento de um EWS utilizando *Inductive Logic Programming*, com pouca atenção dada na modelagem de dados e criação dos modelos preditivos. Halawa, Greene e Mitchell (2014) e Xing et al. (2016) também desenvolveram EWS, entretanto com dados oriundos de MOOCs. Um sistema *gamificado* de ensino com EWS é apresentado em Villagrá-Arnedo et al. (2015), utilizando técnicas de DM para detectar alunos com problemas, de forma temporal.

O trabalho de Kampff (2008) busca descobrir regras de associação para implementar alertas visando verificar o sucesso ou insucesso acadêmico dos estudantes, tendo como base a sua nota do meio da disciplina.

O termo EWS é amplamente abordado em Márquez-Vera et al. (2016), inclusive com exemplos de sistemas em produção. Entretanto apesar deste foco, o trabalho apresenta dados na sua maioria não dependentes do tempo, afetando a capacidade de predições temporais, ademais as variáveis utilizadas são invasivas e talvez não prontamente avaliáveis.

Sanchez-Santillan et al. (2016) apresenta uma breve experimentação, ainda em andamento, que mostra que nem sempre juntar dados de diferentes anos traz melhores resultados.

Alternativas mais bem elaboradas de EWS com EDM são apresentadas em Lara et al. (2014). Neste trabalho os alunos são classificados semanalmente com base na Distância Euclidiana à dois conjuntos de referências, um para cada classe, criados através dos dados históricos da disciplina. Pontos críticos são a falta de estudos para se criar modelos generalistas resultando em quatro modelos, um por disciplina, além disto os resultados obtidos com o classificador são distantes de outros algoritmos clássicos.

Outro extenso trabalho é apresentado por Jayaprakash et al. (2014), onde são abordadas diferentes técnicas de pré-processamento, desbalanceamentos das bases e técnicas de classificação, inclusive estudando a generalização de modelos entre instituições de ensino. Entretanto apesar dos esforços, pouquíssimas variáveis dependentes do tempo foram utilizadas, o que acaba penalizando a dinamicidade do sistema em ser capaz de predizer alunos em risco durante a execução da disciplina.

3.4 Considerações

Quatro trabalhos foram detalhados por tratarem de dados do ensino superior, além de outros de importância menor que foram discutidos na subseção anterior. O primeiro trabalho foi selecionado, pois, especialmente levanta a questão da variabilidade dos dados, e os três seguintes em função de trazerem o desenvolvimento de predições que possibilitem averiguar o desempenho dos estudantes enquanto as aulas ocorrem.

Os aspectos abordados nos trabalhos apresentados e as suas respectivas discussões, ignoraram os resultados numéricos obtidos por eles, pois como todos utilizaram-se de base de dados diferentes, seria insustentável fazer uma comparação neste nível.

Iniciando com uma linha estatística, o trabalho de Gašević et al. (2016) traz importantes considerações acerca da variabilidade dos dados. Os resultados alcançados são muito interessantes, pois pontuam que, a princípio, as técnicas não conseguem sobrepor os diferentes domínios de dados e que assim é melhor modelar os dados a fim de obter conjuntos de dados por disciplina. Outro ponto interessante, é que mesmo que o AVA tenha sido utilizado como sistema de apoio, conseguiu-se viabilizar a análise, demonstrando ser melhor utilizar estes dados, do que não os utilizar.

Um ponto que causa estranheza na análise é o uso de dados de outros anos, dos quais não se realizou análise para ver a relação com os atuais, questionando se talvez somente com os dados mais atuais não se alcançariam outras conclusões.

O trabalho de You (2016) também não faz o emprego de técnicas de DM, assim procura-se a correlação dos indicadores de comportamento com a classe. Por não separar uma parte da base de dados para treino e teste, torna-se difícil a comparação com os outros trabalhos e da validação do modelo em novos dados.

Mesmo assim, ele traz um estudo suportado unicamente por dados advindos do AVA, trazendo uma modelagem interessante de atributos, principalmente para medir a assiduidade do estudante, em que pontos são dados conforme critérios definidos pelos autores. Este tipo de abordagem é interessante, pois permite introduzir conceitos subjetivos advindos do analista, como entrada dos algoritmos. Outro ponto de destaque foi o uso da análise de regressão hierárquica que permitiu fazer a análise dos dados ano a ano.

(HU; LO; SHIH, 2014) é de certa forma único para o contexto deste trabalho, pois traz uma análise criteriosa dos requisitos para se desenvolver EWS, inclusive com o desenvolvimento do sistema. Entretanto, ele faz uso de poucos dados e de baixa qualidade para a extração de resultados confiáveis, pois possui somente uma disciplina e esta ainda possui um fator de desbalanceamento muito alto, tão pouco ele deixa de explorar variações que os dados podem apresentar e passa a agrupar todas as edições dela fazendo *cross-validation* para os testes, sem se questionar se as últimas edições apresentam as mesmas distribuições de dados que as primeiras e se o modelo iria classificar corretamente quando o fossem apresentados dados das últimas edições.

Com isto a análise de resultados também ficou comprometida, visto que somente uma ínfima parte da base de dados possuía alunos apresentando insucesso acadêmico. Considerando que o *cross-validation* mantivesse a proporção de desbalanceamento em cada *fold,* e lembrando que o processo de reamostragem só pode ser aplicado no conjunto de treino, haveria menos de duas amostras da classe insucesso acadêmico para testar, a cada aplicação do classificador. Adicionalmente levantam-se também dúvidas se o processo de reamostragem foi realmente só aplicado ao treino.

A modelagem de sumarizar todos os dados, até aquela semana pode fazer com que nas semanas finais o sistema demore mais para predizer comportamentos de alunos que tinham uma boa média de desempenho e que subitamente passam a ter problemas no aprendizado, seja decorrente de uma nova matéria ou problemas particulares, pois não são apresentados isoladamente os dados da unidade atual às técnicas e assim pode demorar até que os números somados caiam num padrão de mal desempenho.

O estudo de Lykourentzou et al. (2009) trouxe uma gama de classificadores bem interessantes, entretanto a partir da análise de resultados é difícil assumir qual o melhor classificador, sendo que teria sido mais propício uma análise por classe. O melhor classificador segundo os autores foi aquele projetado por eles, pondera-se que como esta técnica classifica qualquer estudante como evadido quando qualquer dos três, tidos como base, tenha classificado assim, ele está beneficiando a classe evadido em total detrimento da outra, questionando-se assim a robustez do modelo em acertar as duas classes.

Atenta-se que a classe de predição só representa aqueles estudantes que irão evadir, sem se preocupar, se eles terão sucesso acadêmico, chegando num mínimo para serem aprovados. Outro ponto de atenção é o uso de dados invasivos para realizar as predições.

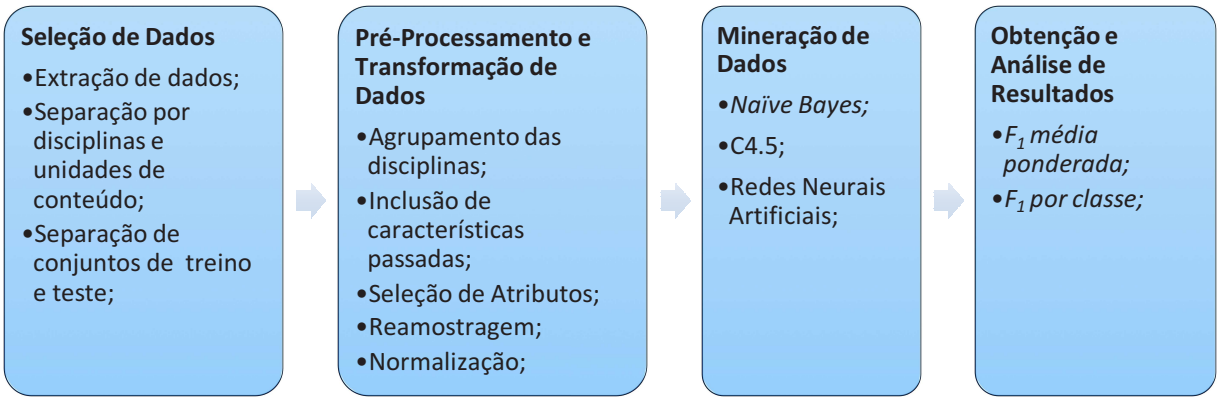
Lykourentzou et al. (2009) ao contrário do trabalho de Hu, Lo e Shih (2014), tomaram o cuidado de separar como teste a última edição. A modelagem dos dados para alimentar os classificadores incluiu a apresentação da unidade de aprendizado corrente, junto com todas as unidades que já foram apresentadas, isto pode novamente atrapalhar a identificação de mudanças de comportamentos bruscas dos estudantes, pois todas as unidades terão o mesmo peso.

Por fim, frente a todos os trabalhos que foram expostos, percebe-se que a literatura deixa uma lacuna para um trabalho que, utilizando-se de dados do ensino superior, não invasivos, advindos do AVA, traga um estudo discutindo o impacto dos classificadores para EWS, principalmente considerando a generalização dos modelos, para tanto com a aplicação de técnicas de pré-processamento e validação com as devidas métricas. Assim, o presente trabalho se posiciona frente à literatura tentando cobrir essa lacuna.

4 ABORDAGEM EXPERIMENTAL

Amparada pelo KDD, a presente abordagem experimental está baseada em 4 etapas, cada uma com procedimentos ou técnicas selecionadas fundamentando-se na literatura e numa análise preliminar dos dados de entrada. A figura 3 apresenta esta abordagem, pontuando os procedimentos ou técnicas.

Figura 3 – Abordagem experimental utilizada.



Fonte: Elaborado pelo Autor.

Inicialmente, a etapa de Seleção de Dados visa definir o conjunto de dados de trabalho, para tanto, extraindo-os através de uma seleção do conjunto completo. A separação deste conjunto de dados por disciplinas individuais, bem como por unidades de conteúdo das disciplinas, define a granularidade da análise, possibilitando também a análise do progresso da predição ao longo do tempo. Outra separação é a que divide as partes destinadas para o treino e para o teste.

Na etapa de Pré-Processamento e Transformação de Dados os procedimentos voltam-se para ajustar o conjunto de dados. Inicialmente é feita a criação do agrupamento dos dados de todas as disciplinas, possibilitando a comparação com uma abordagem generalista. Prosseguindo, há uma preocupação com a inclusão das características das unidades passadas, permitindo assim que os algoritmos de Mineração de Dados considerem também o conhecimento do progresso do estudante. Igualmente, preocupa a dinamicidade do uso dos recursos do AVA dentro das disciplinas, as quais possuem comportamentos diferentes entre si e ao longo do tempo, trazendo a importância da Seleção de Atributos em apontar as variáveis mais relevantes nos diferentes momentos. Adicionalmente, o procedimento de reamostragem entra para corrigir alguma tendência que o classificador possa ter em função de uma classe majoritária e a normalização de dados para a correção de disparidades entre os atributos, bem como readequação do intervalo numérico, principalmente em função das RNAs.

Prosseguindo, após a seleção e ajuste dos dados, a etapa de Mineração de Dados atua na classificação das amostras de teste, depois de ser realizado o processo de treino. Utiliza-se para tanto, três diferentes classificadores.

Ao final, cabe à etapa de Obtenção e Análise de Resultados comparar os resultados obtidos da saída das técnicas com as classes corretas a fim de obter o percentual de acertos e erros de ambas as classes calculando as devidas métricas, possibilitando assim, analisar e comparar os modelos sob a ótica dos objetivos propostos.

As próximas subseções trazem a aplicação desta abordagem experimental, com o detalhamento e discussão dos passos adotados e os resultados obtidos. Adicionalmente, a primeira subseção traz as ferramentas empregadas para o desenvolvimento dos experimentos.

4.1 Ferramentas e Bibliotecas Utilizadas

O desenvolvimento do programa para realizar as experimentações foi feito inteiramente na linguagem de programação R (R CORE TEAM, 2014), com o auxílio da biblioteca *plyr* (WICKHAM, 2011). As técnicas de pré-processamento IG e SMOTE, e de classificação NB e C4.5, foram utilizadas do pacote RWeka (HORNIK; BUCHTA; ZEILEIS, 2009), que é uma interface em R para o programa Weka (WITTEN; FRANK; HALL, 2011).

As RNAs foram executadas através da biblioteca RSNNS (BERGMEIR; BENÍTEZ, 2012), utilizando as bibliotecas *foreach* (REVOLUTION ANALYTICS; WESTON, 2015a), *doParallel* (REVOLUTION ANALYTICS; WESTON, 2015b) e *doMC* (REVOLUTION ANALYTICS; WESTON, 2015c) para a sua paralelização.

A análise e comparação dos resultados foram efetuadas através do software Excel e da biblioteca de gráficos *ggplot2* (WICKHAM, 2009).

Os experimentos foram executados numa máquina que conta com um processador Intel Core i7 de 8 núcleos, com *clock* de 3,4 GHz, e 32 GB de RAM trabalhando a 1600 MHz.

4.2 Descrição e Seleção do Conjunto de Dados

Os dados foram extraídos de uma universidade situada no sul do Brasil. Esta instituição de ensino possui oferta de cursos de nível superior, tanto "a distância", como presenciais. Na modalidade "a distância", as aulas são totalmente *on-line* e as disciplinas ocorrem principalmente de forma bimestral. Durante esse período uma avaliação presencial é exigida, a qual compõe a nota final, juntamente com as notas das tarefas. Assim, a cada unidade de conteúdo da disciplina, que normalmente corresponde ao período de uma semana, novos materiais são disponibilizados e tarefas são solicitadas pelos professores, visando também incentivar o engajamento e interação dos estudantes na comunidade da disciplina.

As variáveis extraídas vieram principalmente do AVA, composto pelo software Moodle[7], assim, cabe destacar que nenhuma informação intrusiva a vida particular do estudante foi utilizada, nem dados que não pudessem estar prontamente disponíveis. As únicas variáveis que não foram extraídas do AVA são aquelas obtidas do histórico de aproveitamento de disciplinas cursadas pelo estudante, a citar "QuantidadeMatriculasValidas" e "MatriculasPercentualAproveitamento".

O atributo classe e as duas variáveis de aproveitamento de disciplinas não se alteram de uma unidade de conteúdo para outra, assim, excluindo estas três variáveis, os atributos restantes são definidos como variáveis dependentes do tempo (HU; LO; SHIH, 2014).

Os tipos de dados das variáveis são todos numéricos. Recursos que não foram utilizados eventualmente, foram representados com o valor 0. A tabela 4 apresenta todas as variáveis com uma breve descrição de cada uma.

[7] Moodle: Disponivel em https://www.moodle.org/

Tabela 4 - Descrição das variáveis utilizadas.

	Atributo	Descrição
1	Forum_Quantidade_Post	Quantidade de postagens em fóruns
2	Forum_Quantidade_Visualizacoes	Quantidade de visualizações de fóruns
3	Forum_TempoUso*	Tempo de visualização ou uso dos fóruns
4	Turno_TempoUsoMadrugada*	Tempo de uso, turno da madrugada
5	Turno_TempoUsoManha*	Tempo de uso, turno da manhã
6	Turno_TempoUsoTarde*	Tempo de uso, turno da tarde
7	Turno_TempoUsoNoite*	Tempo de uso, turno da noite
8	TempoUsoTotal*	Tempo de uso, todos os turnos
9	Numero_Dias_Acessados_Modulo	Quantidade de dias distintos que houve acesso
10	Questionario_Quantidade	Quantidade de questionários respondidos
11	Questionario_Tentativas	Quantidade de tentativas de respostas aos questionários
12	Questionario_TempoUso*	Tempo de visualização ou uso dos questionários
13	Assignment_Post_Quantidade	Quantidade de submissões as tarefas
14	Assignment_View_Quantidade	Quantidade de visualizações das tarefas
15	Assignment_View_TempoUso*	Tempo de visualização das tarefas
16	Resource_View_Quantidade	Quantidade de visualizações a objetos de aprendizagem, como arquivos, *links*, etc.
17	Resource_View_Tempo*^	Tempo de visualização dos objetos de aprendizagem
18	Post_Quantidade	Quantidade total de postagens e submissões
19	View_Quantidade	Quantidade total de visualizações
20	Click_Quantidade	Quantidade total de registros de cliques
21	QuantidadeMatriculasValidas+	Quantidade de matrículas em disciplinas
22	MatriculasPercentualAproveitamento+	Percentual de aprovação das matriculas válidas
23	Classe+	Sucesso ou insucesso acadêmico

* Tempos de uso são calculados com base no tempo entre os cliques.
^ Alguns objetos de aprendizagem não possibilitam extrair o tempo de uso do usuário, por serem recursos separados do Moodle.
\+ Variáveis não dependentes do tempo e que não tiveram os atributos somados.

Fonte: Elaborado pelo autor.

As disciplinas selecionadas para formar o conjunto de trabalho totalizaram 10 diferentes disciplinas bimestrais, advindas da grade curricular de diferentes cursos, sendo que 7 disciplinas são da área de conhecimento das Ciências Sociais Aplicadas e 3 das Exatas. Cabe denotar que cada disciplina pode ocorrer separada por duas ou mais turmas, para um mesmo período.

As unidades de conteúdo que dividem estas disciplinas somam um total de 9, todas divididas por intervalos de uma semana. Assim, a unidade 1 dos dados, por exemplo, representa todas as interações efetuadas do início da disciplina até o final da 1ª semana de aula, da mesma forma que a unidade 2 representa somente os dados da 2ª semana de aula, e assim por diante.

A classe dos dados representa se o estudante teve sucesso acadêmico, atingindo a nota mínima para ser aprovado na disciplina, ou insucesso acadêmico, quando ele cancelou a disciplina ou não atingiu a nota mínima.

A tabela 5 apresenta a quantidade de amostras por períodos letivos e disciplinas. Cada amostra ou registro corresponde a um estudante que cursou aquela disciplina naquele período letivo, alerta-se que o número de registros não corresponde ao número de alunos, e sim, ao

número de matrículas efetuadas nas disciplinas. Adicionalmente, também é possível observar a distribuição das amostras por classe.

Tabela 5 - Distribuição das amostras por período letivo, disciplina e classe, incluindo o conjunto com a abordagem generalista.

Área	Disciplina	Classe	Período Letivo										Total
			2013/1	2013/2	2013/3	2013/4	2014/1	2014/2	2014/3	2014/4	2015/1	2015/2	2015/3	2015/4	
Ciências Sociais Aplicadas	1	Sucesso	122	-	77	-	66	-	73	-	71	-	69	-	**478**
		Insucesso	40	-	28	-	26	-	23	-	24	-	26	-	**167**
		Taxa Sucesso	**75,31%**	**-**	**73,33%**	**-**	**71,74%**	**-**	**76,04%**	**-**	**74,74%**	**-**	**72,63%**	**-**	**74,11%**
Ciências Sociais Aplicadas	2	Sucesso	53	93	85	91	69	111	82	66	79	75	80	76	**960**
		Insucesso	21	45	14	28	29	50	17	29	20	19	17	21	**310**
		Taxa Sucesso	**71,62%**	**67,39%**	**85,86%**	**76,47%**	**70,41%**	**68,94%**	**82,83%**	**69,47%**	**79,80%**	**79,79%**	**82,47%**	**78,35%**	**75,59%**
Exatas	3	Sucesso	-	37	4	30	36	53	35	50	37	60	33	45	**420**
		Insucesso	-	18	10	17	14	24	16	21	11	25	16	9	**181**
		Taxa Sucesso	**-**	**67,27%**	**28,57%**	**63,83%**	**72,00%**	**68,83%**	**68,63%**	**70,42%**	**77,08%**	**70,59%**	**67,35%**	**83,33%**	**69,88%**
Ciências Sociais Aplicadas	4	Sucesso	106	-	77	-	66	38	65	26	74	35	51	27	**565**
		Insucesso	29	-	26	-	26	11	29	11	23	6	17	16	**194**
		Taxa Sucesso	**78,52%**	**-**	**74,76%**	**-**	**71,74%**	**77,55%**	**69,15%**	**70,27%**	**76,29%**	**85,37%**	**75,00%**	**62,79%**	**74,44%**
Ciências Sociais Aplicadas	5	Sucesso	--	51	-	42	-	23	-	39	-	41	-	39	**235**
		Insucesso	-	68	-	67	-	74	-	74	-	52	-	56	**391**
		Taxa Sucesso	**-**	**42,86%**	**-**	**38,53%**	**-**	**23,71%**	**-**	**34,51%**	**-**	**44,09%**	**-**	**41,05%**	**37,54%**
Ciências Sociais Aplicadas	6	Sucesso	91	-	70	-	76	46	43	65	69	54	30	44	**588**
		Insucesso	21	-	16	-	19	22	13	23	28	36	12	29	**219**
		Taxa Sucesso	**81,25%**	**-**	**81,40%**	**-**	**80,00%**	**67,65%**	**76,79%**	**73,86%**	**71,13%**	**60,00%**	**71,43%**	**60,27%**	**72,86%**
Ciências Sociais Aplicadas	7	Sucesso	122	22	57	49	58	49	63	36	61	75	48	38	**678**
		Insucesso	37	14	21	25	33	29	26	27	24	19	24	19	**298**
		Taxa Sucesso	**76,73%**	**61,11%**	**73,08%**	**66,22%**	**63,74%**	**62,82%**	**70,79%**	**57,14%**	**71,76%**	**79,79%**	**66,67%**	**66,67%**	**69,47%**
Exatas	8	Sucesso	16	40	26	24	20	39	33	19	24	35	21	20	**317**
		Insucesso	42	89	46	51	28	98	60	38	46	62	32	55	**647**
		Taxa Sucesso	**27,59%**	**31,01%**	**36,11%**	**32,00%**	**41,67%**	**28,47%**	**35,48%**	**33,33%**	**34,29%**	**36,08%**	**39,62%**	**26,67%**	**32,88%**
Exatas	9	Sucesso	33	-	30	-	41	-	44	-	38	-	35	-	**221**
		Insucesso	37	-	53	-	74	-	53	-	65	-	64	-	**346**
		Taxa Sucesso	**47,14%**	**-**	**36,14%**	**-**	**35,65%**	**-**	**45,36%**	**-**	**36,89%**	**-**	**35,35%**	**-**	**38,98%**
Ciências Sociais Aplicadas	10	Sucesso	355	119	215	58	233	179	157	152	237	165	166	117	**2153**
		Insucesso	67	20	56	15	56	57	39	38	55	41	30	35	**509**
		Taxa Sucesso	**84,12%**	**85,61%**	**79,34%**	**79,45%**	**80,62%**	**75,85%**	**80,10%**	**80,00%**	**81,16%**	**80,10%**	**84,69%**	**76,97%**	**80,88%**
Abordagem Generalista	--	Sucesso	898	362	641	294	665	538	595	453	690	540	533	406	**6615**
		Insucesso	294	254	270	203	305	365	276	261	296	260	238	240	**3262**
		Taxa Sucesso	**75,34%**	**58,77%**	**70,36%**	**59,15%**	**68,56%**	**59,58%**	**68,31%**	**63,45%**	**69,98%**	**67,50%**	**69,13%**	**62,85%**	**66,97%**
Total por Período Letivo (desconsiderando Abordagem Generalista)			**1192**	**616**	**911**	**497**	**970**	**903**	**871**	**714**	**986**	**800**	**771**	**646**	**9877**

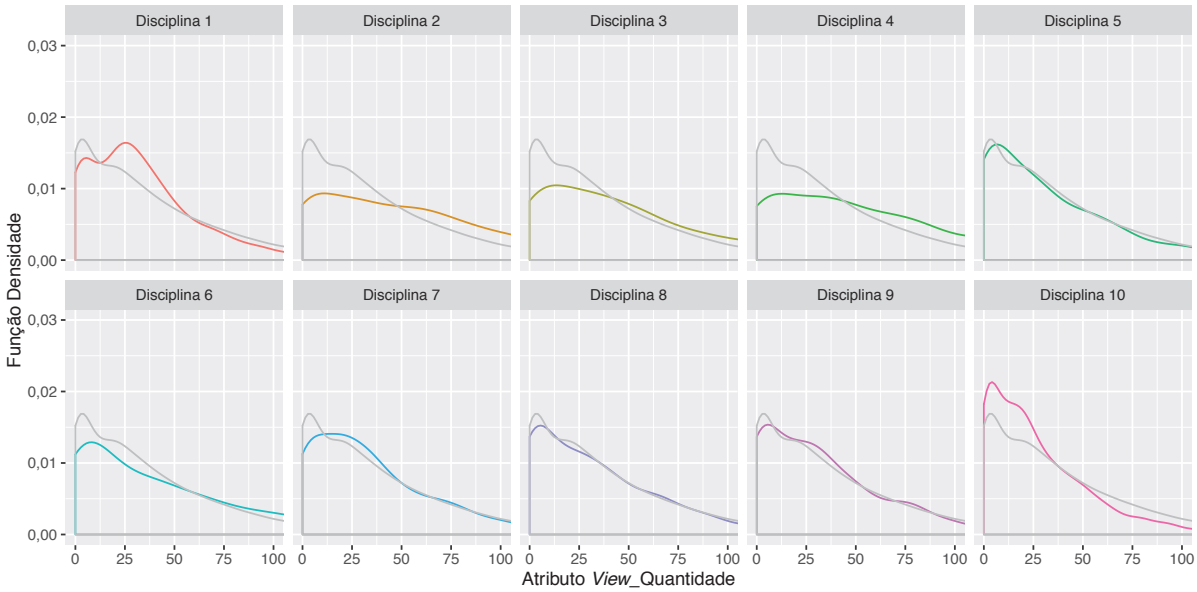
* O período letivo é representado pelo ano e um número sequencial, e o campo "Taxa Sucesso" indica a quantidade de amostras da classe sucesso acadêmico dividido pelo total de ambas as classes.

Fonte: Elaborado pelo autor.

As diferentes unidades de conteúdo formaram novos conjuntos de dados separados, selecionando-se para trabalho os conjuntos das unidades 1, 2, 3, 4, 5 e 6, por ainda oportunizarem a predição antecipada. Esta seleção se deve em função da unidade de conteúdo 6 servir para desencadear ações contra o insucesso acadêmico na semana 7, a última antes do teste final.

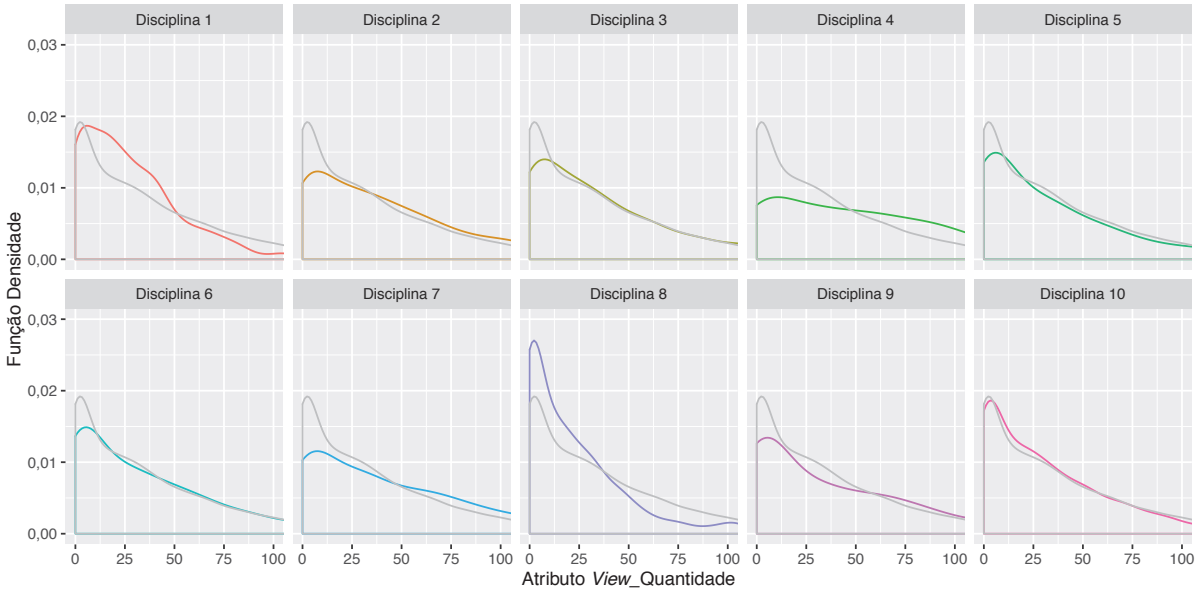
As figuras 4 e 5 apresentam a distribuição da variável "View_Quantidade", respectivamente, para as unidades de conteúdo 3 e 5, este atributo apresentou-se como um dos mais relevantes apontado pela técnica Ganho de Informação (*Information Gain* – IG, vide subseção 2.3.1). Estes gráficos exibem a diferença entre os dados agrupados e não agrupados, principalmente visto no intervalo inicial, bem como a diferença do uso do AVA entre as disciplinas e unidades de conteúdo. Atenta-se que para facilitar a visualização foi feito um corte no ponto 100 do eixo *x* dos gráficos. É possível observar que a maioria dos estudantes tende a ver o AVA poucas vezes, com o número ficando ao redor do zero, exceto nas disciplinas 2, 3 e 4 da unidade de conteúdo 3 e também da disciplina 4 na unidade de conteúdo 5, locais onde este padrão foi mais fraco. Esta tendência que foi fortemente seguida pelos dados agrupados, e que conforme as disciplinas avançaram para a unidade de conteúdo 5 cresceu, mostrando que os estudantes estão aparentemente mais engajados nas unidades de conteúdo iniciais do que as do meio. Interessante também observar a disciplina 4, que apresentou o comportamento mais balanceado considerando ambos gráficos.

Figura 4 - Função de densidade para o atributo "*View*_Quantidade" da unidade de conteúdo 3, com a linha cinza representando o conjunto generalista e as linhas coloridas os conjuntos individualizados.



Fonte: Elaborado pelo autor.

Figura 5 - Função de densidade para o atributo "*View*_Quantidade" da unidade de conteúdo 5, com a linha cinza representando o conjunto generalista e as linhas coloridas os conjuntos individualizados.



Fonte: Elaborado pelo autor.

Os valores do eixo *y* nos gráficos das figuras 4 e 5 são a probabilidade relativa da variável em questão ("*View*_Quantidade") tomar um dado valor de uma faixa.

O critério de seleção dos dados de treino e teste adotado foi a reserva dos dados referentes à 2013, 2014 e os dois primeiros períodos letivos de 2015 para treinamento, e somente os últimos dois períodos letivos de 2015 reservados para teste. Eventualmente, alguns grupos de teste tiveram apenas um período selecionado, ao invés de dois, pois, nem todas as disciplinas ocorrem em todos os períodos letivos, como pode se observar na tabela 5. Esta separação dos dados de treino e teste com base num corte temporal, baseou-se na suposição que, ao selecionar as edições mais recentes das disciplinas para teste, estas representem melhor o dado atual e consequentemente o modelo de classificação esteja mais bem ajustado a novos dados. Este corte, apesar de agrupar as edições das disciplinas ao longo dos períodos letivos, não descarta a separação dos conjuntos de dados por disciplinas ou grupo, e por unidades de conteúdo.

Nos conjuntos que tiveram disciplinas agrupadas foi realizado o treino com todos os dados de treino do conjunto, entretanto as métricas de desempenho sobre os dados de teste foram calculadas separadamente por disciplina, possibilitando averiguar o desempenho do modelo generalista com os dados de teste por disciplina. O desempenho dos inúmeros modelos foi mensurado através da média ponderada e por classe da métrica F_1.

Adicionalmente, é importante reforçar que a aplicação dos algoritmos de pré-processamento não considerou o conjunto de teste, sendo somente feito adaptações a esse em relação ao conjunto de treino, onde necessário.

4.3 Pré-processamento e Transformação de Dados

Inicialmente o conjunto de dados foi expandido criando-se conjuntos separados por disciplina e grupo das disciplinas. Estas divisões ou agrupamento visam verificar o impacto de abordagens generalistas que se usam do grupo de disciplinas para o treinamento, contrapondo com alternativas que geram um modelo para cada disciplina, conforme visto no trabalho de Gašević et al. (2016). Desta forma, além dos 10 conjuntos por disciplina, um novo conjunto foi criado, com todas as disciplinas.

A introdução de características de unidades passadas à unidade atual, conforme foi discutido no capítulo de trabalhos relacionados, foi feita criando-se um novo atributo para cada uma das variáveis dependentes do tempo. As variáveis criadas representam o somatório das unidades passadas com a unidade corrente, de cada uma das variáveis originais. Cabe destacar que as variáveis originais permaneceram, juntamente com as novas variáveis somadas, nos seus respectivos conjuntos de dados. As variáveis que não eram dependentes do tempo e, portanto, não foram utilizadas para a criação de novas variáveis foram sinalizadas na tabela 4. Desta forma, cabe destacar que o número final de variáveis disponíveis, que era de 23, ficou em 43.

Os 66 conjuntos de dados gerados até este ponto, ainda foram tratados por técnicas de Seleção de Atributos para a remoção de atributos irrelevantes, por exemplo, em casos onde não houve o uso de um determinado recurso. Para tanto optou-se pelo algoritmo IG (QUINLAN, 1986), utilizado também no trabalho de Romero et al. (2013), parametrizado para a geração de 4 novos conjuntos de dados, com 10, 15, 20 e 25 atributos melhores selecionados, para cada conjunto já existente.

Verificando as abordagens vistas em Hu, Lo e Shih (2014) e Jayaprakash et al. (2014) que utilizaram-se de reamostragem randômica dos dados, e especialmente seguindo a abordagem de Márquez-Vera et al. (2013), que utilizou-se do algoritmo SMOTE (CHAWLA et al., 2002), o qual segundo Chawla et al. (2002) apresenta melhores resultados que abordagens puramente randômicas, equiparou-se o percentual de amostras com sucesso e insucesso acadêmico, criando novas amostra sintéticas da classe minoritária.

Adicionalmente à criação de novas variáveis, as técnicas de transformação de dados também foram utilizadas para a normalização dos atributos através dos máximos e mínimos de cada conjunto.

4.4 Mineração de Dados

Os diferentes conjuntos de dados que foram gerados na etapa anterior serviram de entrada para três técnicas de classificação de dados. A citar NB, C4.5 e RNA de múltiplas camadas *feedforward* com *Backpropagation*.

As técnicas NB e C4.5 foram utilizadas com as parametrizações padrão da biblioteca, replicando as abordagens vistas em Barber e Sharkey (2012), Er (2012), Hu, Lo e Shih (2014), Márquez-Vera et al. (2016) e Xing et al. (2016).

As RNAs foram parametrizadas utilizando uma camada oculta com 3, 9, 17, 25 e 31 neurônios, com o critério de parada definido para 100, 300, 400, 500 e 700 épocas, taxa de aprendizado com 0,001, 0,02 e 0,05 e valor de *momentum* com 0, 0,5 e 1, sendo que esta definição de valores para os parâmetros foi baseada parcialmente em testes preliminares. Adicionalmente o algoritmo *Scaled Conjugate Gradient* (SCG) (MØLLER, 1993) também foi utilizado para treino, como alternativa para o GD, pois o SCG não exige a parametrização nem da taxa de aprendizagem e nem do valor de *momentum*. Cada valor de parametrização foi

cruzado com todas as outras possibilidades de parametrização dos outros campos e cada configuração foi repetida 5 vezes.

4.5 Apresentação e Análise de Resultados

A combinação de todas as parametrizações descritas dos algoritmos de Mineração de Dados, bem como as quantidades de atributos selecionados pelo IG, gerou 1.008 possíveis modelos de classificadores. Multiplicando pelas 66 modelagens dos dados, por disciplina e em grupo, além da separação em módulos, construiu-se 66.528 modelos de classificadores. Realizando a separação por disciplina dos grupos na aplicação dos testes, conforme exposto anteriormente, obteve-se 120.960 resultados, cada um mensurado através das 3 métricas elencadas no capítulo anterior. Assim, cada disciplina, para cada unidade de conteúdo, possui 6.048 resultados. Adicionalmente, para cada configuração das RNAs, foram executadas 5 repetições.

Os gráficos da figura 6 mostram os melhores resultados ao longo das unidades de conteúdo por disciplina, treinados individualmente (individualizada) e com os dados de todas as disciplinas (generalista). Percebe-se primeiramente que não há um claro indicativo que agregar mais dados, de mais disciplinas, conduz a melhores resultados, de fato na disciplina 9 a abordagem generalista obteve um resultado com uma diferença de 10 p.p. para menos quando observadas as unidades de conteúdo iniciais e a classe de insucesso acadêmico, adicionalmente o oposto também não se confirma totalmente, já que em muitas situações a abordagem de considerar os dados de todas as disciplinas no treino levou a melhores resultados.

As diferentes densidades, vistas nos gráficos das figuras 4 e 5, também poderiam explicar os diferentes resultados alcançados, pois percebe-se através do comportamento da variável em foco que as disciplinas têm distribuições diferentes, inclusive entre as unidades de conteúdo e conjuntos, entretanto desta forma era esperado, por exemplo, que as disciplinas 2 e 4, por apresentarem densidades distantes quando visualizadas separadas por disciplina e em conjunto, obtivessem resultados piores com a abordagem generalista, o que não ocorreu, assim percebe-se que mais testes são necessários neste sentido, visto que explicar os resultados sob a perspectiva de uma única variável parece demasiadamente simplista.

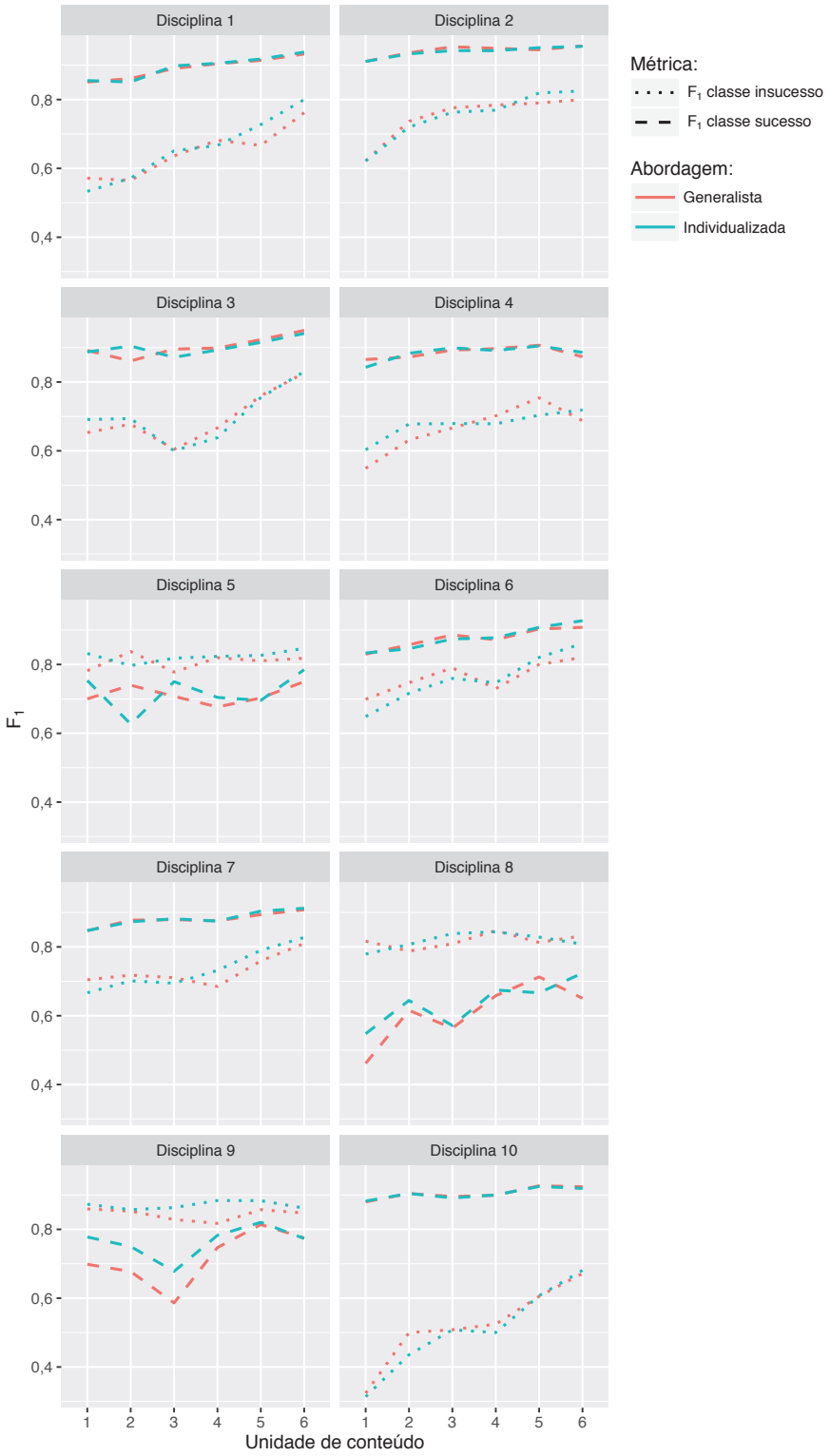
Outra percepção, que já era esperada ser vista em todas as disciplinas, foi o aumento na taxa de acerto das predições ao longo das unidades de conteúdo, por de fato a maioria das disciplinas apresentou este comportamento, mas é interessante notar que a subida não é linear, ao contrário, uma pequena descida ou estagnação é observada próxima as unidades de conteúdo centrais (3 e 4), isso foi atribuído possivelmente como uma mudança no comportamento do estudante no uso do Moodle que não foi capturado pelas técnicas, ou possivelmente mascarado pelos atributos que representam os somatórios das unidades passadas.

No geral a classe de estudantes de insucesso acadêmico foi mais difícil de ser predita, alcançando na maior parte dos casos piores resultados que a outra classe. Os casos onde a classe de insucesso acadêmico alcançou melhores resultados foi onde a taxa de sucesso (como pode ser visto na tabela 5) caiu abaixo de 50%, mostrando o impacto do desbalanceamento de classes.

Os dois piores resultados foram obtidos com a disciplina 8 na classe de alunos com sucesso acadêmico e na disciplina 10 na classe insucesso, cabe salientar que estas disciplinas possuem, respectivamente, a menor e maior taxa de sucesso de estudantes. Do lado oposto, considerando que o principal objetivo é identificar alunos com risco de insucesso acadêmico, pode-se observar que a disciplina 9 alcançou os melhores resultados, e desconsiderando-se os casos onde a taxa de sucesso caiu abaixo de 50%, as disciplinas 6 e 7 alcançaram os melhores

resultados nas unidades de conteúdo iniciais e a disciplina 6 foi a melhor na unidade de conteúdo final.

Figura 6 - Melhores resultados por disciplina ao longo das unidades de conteúdo medido pela F_1.



Fonte: Elaborado pelo autor.

Os gráficos da figura 7 trazem a média ponderada da F_1 dos mesmos resultados da figura 6, entretanto sob a perspectiva das técnicas de classificação utilizadas. Inicialmente percebe-se a predominância das RNAs, especialmente aquelas que usam o GD, comportamento que também pode ter sido influenciado pelo grande número de parametrizações utilizadas. Continuando, as RNAs com SCG também tiveram bom desempenho, e mesmo sendo ligeiramente menor que as RNAs com GD, a escolha por esta traz a vantagem de necessitar menos parâmetros.

Figura 7 - Melhores resultados por técnica separadas por disciplina e abordagem, ao longo das unidades de conteúdo, medido pela média ponderada da *F_1*.

Fonte: Elaborado pelo autor.

O algoritmo NB em comparação com o C4.5 alcançou resultados melhores, mas quando o NB teve resultados melhores foi por uma pequena diferença. O oposto mostrou que quando o NB foi pior, ele teve resultados mais imprevisíveis com uma diferença maior em relação ao C4.5, principalmente como visto na disciplina 9.

Ao visualizar somente os resultados da abordagem generalista, a fim de observar as técnicas com maior poder de generalização, em geral as RNAs com GD apresentam-se em primeiro lugar. As RNAs com SCG tem resultados aproximados as das RNAs com GD, entretanto poucas vezes chega a obter resultados melhores que essa. O C4.5 e NB alternam-se ao longo dos resultados, com somente alguns momentos onde o C4.5 obtém resultados melhores que o das RNAs com SCG.

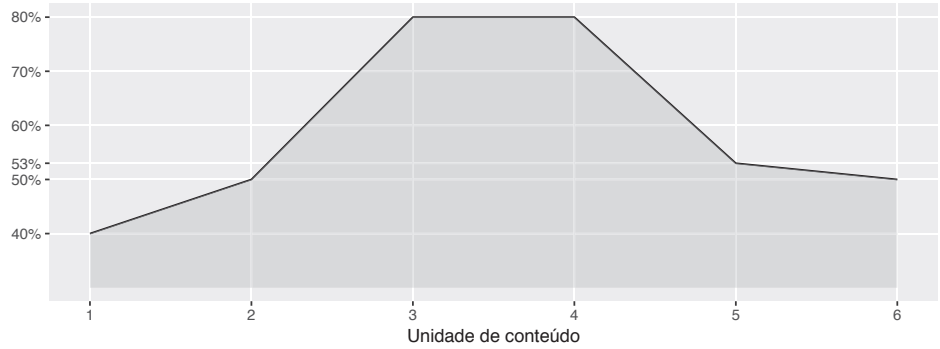
O gráfico da figura 8 e a tabela 6 apresentam a quantidade de atributos em comum dos melhores resultados escolhidos com base na média ponderada da F_1, agrupados, respectivamente, por unidades de conteúdo e disciplinas. Cabe lembrar que os atributos, os quais compõem os melhores modelos, foram escolhidos pela técnica de Seleção de Atributos IG.

Desta forma, o gráfico da figura 8 apresenta as 6 unidades de conteúdo, com cada ponto representando a porcentagem de atributos selecionadas pelo IG que todas as 10 disciplinas apresentaram em comum. Para a obtenção destes valores, por exemplo, considere que uma disciplina A teve selecionado para a unidade de conteúdo 1 os atributos *a*, *b*, *c*; considere que uma outra disciplina B, para a mesma unidade de conteúdo, teve selecionado os atributos *c*, *d*, *f*, *g*; considere ainda que outra disciplina C, também para a mesma unidade de conteúdo, teve selecionado os atributos *a*, *c*, *d*; portanto o único atributo que se tem em comum a todas as 3 disciplinas do exemplo é o "c", chegando-se assim a 33% de atributos em comum. Caso os conjuntos de dados de cada disciplina apresentem quantidades diferentes de atributos selecionados (como é o caso da disciplina B do exemplo), considera-se como denominador da razão, para a porcentagem, a quantidade de atributos do conjunto que apresentou a menor quantidade de atributos (neste exemplo 3 atributos), assim espera-se trabalhar com o pior caso, pois para apresentar alguma variação necessariamente todas as disciplinas terão que apresentar alguma diferença num provável conjunto reduzido de atributos, em relação a quantidade máxima de atributos de cada disciplina. Na abordagem generalista, neste casso, se alcança 100% para todas as unidades de conteúdo, visto que o conjunto de dados para cada unidade de conteúdo será o mesmo, sem ter o que variar exceto a quantidade de atributos.

A tabela 6 apresenta a mesma abordagem do gráfico da figura 8, entretanto representa em cada linha somente uma disciplina, com a porcentagem de atributos em comum para todas as unidades de conteúdo daquela disciplina.

As duas análises contribuem no sentido de explicitar que, quando vistos de diferentes ângulos e considerando os melhores resultados que se obteve, existe uma variabilidade nos dados entre as disciplinas e unidades de conteúdo, e principalmente entre as diferentes abordagens.

Figura 8 - Atributos em comum dos melhores resultados, através das disciplinas, agrupado por unidades de conteúdo.



Fonte: Elaborado pelo autor.

O gráfico da figura 8 representa possíveis diferenciações do uso dos recursos do AVA ou padrões de acesso em cada disciplina que impactam para o sucesso ou insucesso do estudante, assim pode-se ter uma disciplina ministrada por um professor que deu mais ênfase a utilização dos fóruns ou outro professor, de outra disciplina, que deu mais ênfase a submissão de tarefas, até mesmo pode-se considerar que em função do dia da semana que a disciplina ocorre, por exemplo na segunda-feira, possibilita que os alunos, com o intuito de entregar as tarefas, tenham mais acessos no dia anterior, domingo, e assim em outros turnos que não o noturno, período comumente associado aos alunos que trabalham durante o dia.

Nesta figura também se percebe que, inicialmente a primeira e segunda unidades de conteúdo apresentaram valores baixos com uma diferenciação maior que as unidades de conteúdo 3 e 4, provavelmente em função de ainda não haver um padrão claro nos dados por ser o início da disciplina. Assim como as duas últimas unidades, comportamento que se relacionou a uma maior diferenciação nos recursos utilizados entre os professores das diferentes disciplinas, conforme elas avançam para o fim da disciplina.

Tabela 6 - Comparação dos atributos em comum da abordagem individualizada e generalista por unidades de conteúdo, agrupado por disciplinas, considerando os melhores resultados.

Disciplina	Abordagem individualizada	Abordagem generalista
1	73%	100%
2	80%	93%
3	80%	93%
4	80%	93%
5	90%	100%
6	80%	100%
7	90%	100%
8	100%	80%
9	40%	100%
10	70%	100%

Fonte: Elaborado pelo autor.

Já a tabela 6 representa a porcentagem de diferentes recursos utilizados ou padrões de acessos dentro da mesma disciplina ao longo das 6 unidades de conteúdo, ou seja, a possível diferenciação em função do uso dos recursos do AVA que possivelmente professores empregaram ao longo da mesma disciplina, ou eventuais padrões de acesso dos alunos ao AVA que alteraram-se em função da aproximação do período de provas e conclusão da disciplina, indicativos que eventualmente tornam-se mais ou menos influentes para o IG, a fim de indicar características de alunos que estão em risco.

Em especial, a tabela 6, indica que a abordagem generalista é menos suscetível as nuances dos dados, quando ao longo das unidades de conteúdo, visto que as variações dos atributos escolhidos foram em geral menores, provavelmente em função do agrupamento das disciplinas estarem ofuscando pequenos comportamentos particulares de um dado momento da disciplina, como por exemplo o uso de somente um determinado recurso de aprendizagem. Certas disciplinas, em função de especificidades dos dados, ainda apresentaram comportamentos atípicos em relação as demais, como é o caso da disciplina 9 que alcançou o valor de 40%, provavelmente em função de uma maior dinamicidade com que os recursos do AVA foram empregados, e a disciplina 8 que os atributos variaram mais com o abordagem generalista, demonstrando que talvez na abordagem generalista o IG tenha considerado características particulares que não eram desta disciplina, e sim das outras que estavam no mesmo conjunto.

Por fim, percebe-se que os resultados de ambas as abordagens, generalista e individualizada, mostraram-se promissoras. Neste estudo, também tentou-se demonstrar o elevado número de experimentos para a obtenção dos resultados, o que para a abordagem individualizada necessitou 60.480 modelos, enquanto que para a generalista foram necessários 10 vezes menos ou 6.048 modelos. Expandindo para o cenário de uma instituição de ensino, a qual possui muito mais do que 10 disciplinas, vislumbra-se um grande ganho com a abordagem generalista com a diminuição considerável da quantidade de modelos para se gerar e manter, sem uma perda significativa na assertividade dos resultados. Adicionalmente, o emprego de técnicas como as RNAs com SCG permitem diminuir ainda mais a quantidade de modelos, passando de 5.400 com as RNAs com GD para 600, quando considerada a abordagem generalista.

5 CONCLUSÃO

A evasão e o baixo desempenho acadêmico de estudantes são de longo tempo um problema (HEPPEN; THERRIAULT, 2008). Modelos preditivos que usam dados provenientes de AVA permitem identificar antecipadamente esses alunos, quando empregados com EWS esses se destacam pela relevância como sistemas de suporte à decisão aos professores e setores estratégicos e de planejamento dentro das instituições de ensino. A maioria dos estudos encontrados na literatura que trabalham nisso não são focados nos modelos preditivos gerados pela Mineração de Dados e poucos deles estão preocupados com a sua generalização, mesmo sendo um assunto relevante, pois impacta diretamente a confiabilidade do sistema, além de gastos financeiros. A confiabilidade do EWS está ligada intimamente a capacidade do modelo preditivo, gerado pela EDM, apontar o mais corretamente possível os estudantes em risco. Ao passo que, caso o sistema aponte diversos estudantes erroneamente, os envolvidos acabam dispendendo tempo sobre alunos que não precisavam de ajuda em detrimento dos que realmente precisavam de auxílio.

O presente trabalho baseado no processo do KDD, aplicando técnicas clássicas de aprendizado de máquina, bem como, distintas abordagens de modelagem de dados, permitiu extrair resultados que demonstram o comportamento dos modelos preditivos a medida que adiciona-se mais dados na fase de treinamento, dessa forma, criando-se abordagens generalistas. Assim permitindo estender e melhorar o entendimento das implicações da forma de uso dos modelos de Mineração de Dados dentro de EWS, ponto que foi inclusive apontado como uma limitação nas conclusões do trabalho de Hu, Lo e Shih (2014).

Os resultados também permitiram avaliar o balanço desejado entre o grau de erro da predição, e o quão cedo na disciplina as instituições de ensino podem começar a agir preventivamente sobre os estudantes, para reverter uma situação futura não desejada. Adicionalmente, foi demonstrado que atingir tais resultados é possível sem a necessidade de invadir a privacidade dos alunos, para tanto somente usando dados advindos da interação do aluno com o AVA.

Na definição do modelo preditivo de identificação de estudantes em risco de insucesso acadêmico, as abordagens de representar a variável classe como os alunos em risco de insucesso acadêmico e acompanhar o progresso dos estudantes dentro da disciplina foram tidas como alternativas mais promissoras, em detrimento de outras vistas na literatura levantada no capítulo 3 e que apenas objetivavam identificar alunos que iriam evadir ou ignoravam as interações enquanto a disciplina estava em progresso.

Os melhores resultados alcançados, principalmente com as RNAs com GD, mostraram que, no contexto deste trabalho, nem sempre mais dados impactam para o modelo mais acurado e que a melhor opção em muitos casos é manter os modelos com os dados de apenas uma disciplina. Tais conclusões são parcialmente similares as do trabalho de Gašević et al. (2016). Entretanto considerando uma aplicação na indústria, a escolha de modelos generalistas e RNAs com SCG apresentaram-se também como opções promissoras em função de uma gama menor de modelos e parâmetros para se ajustar nas RNAs. Assim, apesar de uma assertividade um pouco menor a escolha por estas opções se traduz em menos manutenções, levando a uma diminuição nos custos. Desta forma, uma das grandes contribuições deste trabalho foi demonstrar que ambas as abordagens, individual e generalista, apresentam vantagens e desvantagens, dependendo das circunstâncias.

Em tempo, ainda é importante destacar que buscou-se um conjunto de dados que fosse diversificado, com diferentes disciplinas e cursos, limitando-se a manter os dados quase que

exclusivamente atrelados ao Moodle, que é uma fonte de dados comum a muitos trabalhos da área, visando assim mitigar problemas de reprodutibilidade dos resultados para outros conjuntos de dados.

Por fim, no decorrer do trabalho ainda percebeu-se extensões que podem ser realizadas a este estudo, vindo a melhorar os resultados e trazendo novas compreensões, assim sugere-se como trabalhos futuros os pontos elencados:

- Analisar o impacto de variáveis selecionadas pelo algoritmo de Seleção de Atributos para um determinado modelo tem para outros modelos que não tiveram aquelas variáveis selecionadas, a fim de verificar a importância e peso que o algoritmo de Seleção de Atributos tem na construção do modelo;
- Uso de outras técnicas comumente vistas na área de Mineração de Dados, como SVM, as quais já se mostram promissoras em outros trabalhos e que possam eventualmente melhorar os resultados;
- Uso de técnicas que são específicas para transferência de conhecimento entre domínios de dados, como as advindas da área de *Transfer Learning* (LU et al., 2015; PAN; YANG, 2010), desta forma provendo mais uma forma de generalizar os modelos, trabalhar também com disciplinas mais novas e com menos dados, e tentar ultrapassar os resultados obtidos no treino de forma individualizada;
- Estudar a generalização entre as unidades de conteúdo e as disciplinas com diferentes cursos, a fim de verificar outra possibilidade de diminuição dos modelos de mineração;
- Testar outros valores de balanceamento na reamostragem, visto que alguns resultados estavam relacionados ao desbalanceamento do conjunto de dados, conforme observado na seção 4.5.

REFERÊNCIAS

AGUDO-PEREGRINA, Á. F. et al. Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. **Computers in Human Behavior**, v. 31, n. 1, p. 542–550, fev. 2014.

ANTUNES, C. Anticipating student's failure as soon as possible. In: ROMERO, C. et al. (Eds.). . **Handbook of Educational Data Mining**. Boca Raton, FL: CRC Press Taylor, 2011. p. 353–363.

BARBER, R.; SHARKEY, M. **Course Correction : Using Analytics to Predict Course Success**. Second International Conference on Learning Analytics and Knowledge. **Anais**...Vancouver, BC: 2012.

BERGMEIR, C.; BENÍTEZ, J. M. Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. **Journal of Statistical Software**, v. 46, n. 7, p. 1–26, 2012.

BREIMAN, L. et al. **Classification and regression trees**. Boca Raton, FL: Chapman & Hall/CRC, 1984.

BURGES, C. J. C. C. A Tutorial on Support Vector Machines for Pattern Recognition. **Data Mining and Knowledge Discovery**, v. 2, n. 2, p. 121–167, 1998.

CHAKRABARTI, S. **Mining The Web: Discovering Knowledge from Hypertext Data**. San Francisco, CA: Morgan Kaufmann Publishers, 2003.

CHAWLA, N. V. et al. SMOTE: synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, v. 16, n. 1, p. 321–357, 2002.

DANGI, A.; SRIVASTAVA, S. **Educational data classification using selective Naive Bayes for quota categorization**. 2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE). **Anais**...Patiala: IEEE, dez. 2014

ER, E. Identifying At-Risk Students Using Machine Learning Techniques. **International Journal of Machine Learning and Computing**, v. 2, n. 4, p. 476–480, 2012.

FAYYAD, U. et al. (EDS.). **Advances in Knowledge Discovery and Data Mining**. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37–54, 1996.

FELDMAN, R.; SANGER, J. **The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data**. Cambridge, UK: Cambridge University Press, 2006.

FONTI, M. **A Predictive Modeling System: Early identification of students at-risk enrolled in online learning programs**. 2015. 109 f. Tese (Doutorado) - College of Engineering and Computing, Nova Southeastern University, Florida, 2015.

FREUND, Y.; SCHAPIRE, R. R. E. **Experiments with a New Boosting Algorithm**. International Conference on Machine Learning. **Anais**...1996.

GAŠEVIĆ, D. et al. Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. **Internet and Higher Education**, v. 28, p. 68–84, 2016.

HALAWA, S.; GREENE, D.; MITCHELL, J. Dropout Prediction in MOOCs using Learner Activity Features. **eLearning Papers**, v. 37, p. 1–10, 2014.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3. ed. Waltham, MA: Morgan Kaufmann, 2012.

HAYKIN, S. **Redes Neurais: Princípios e Prática**. 2. ed. [s.l.] Bookman, 2001.

HEPPEN, J. B.; THERRIAULT, S. B. **Developing Early Warning Systems to Identify Potential High School Dropouts**, 2008.

HORNIK, K.; BUCHTA, C.; ZEILEIS, A. Open-source machine learning: R meets Weka. **Computational Statistics**, v. 24, n. 2, p. 225–232, 2009.

HORVITZ, E.; MULLIGAN, D. Data, privacy, and the greater good. **Science**, v. 349, n. 6245, p. 253–255, 17 jul. 2015.

HU, Y.-H.; LO, C.-L.; SHIH, S.-P. Developing early warning systems to predict students' online learning performance. **Computers in Human Behavior**, v. 36, p. 469–478, 2014.

JAYAPRAKASH, S. M. et al. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. **Journal of Learning Analytics**, v. 1, n. 1, p. 6–47, 2014.

KAMPFF, A. J. C.; REATEGUI, E. B.; LIMA, J. V. DE. Mineração de dados educacionais para a construção de alertas em ambientes virtuais de aprendizagem como apoio à prática docente. **Novas Tecnologias na Educação**, v. 6, n. 2, 2008. Não paginado.

LARA, J. A. et al. A system for knowledge discovery in e-learning environments within the European Higher Education Area - Application to student data from Open University of Madrid, UDIMA. **Computers and Education**, v. 72, p. 23–26, 2014.

LOO, C. K.; RAO, M. V. C. Accurate and reliable diagnosis and classification using probabilistic ensemble simplified fuzzy ARTMAP. **IEEE Transactions on Knowledge and Data Engineering**, v. 17, n. 11, p. 1589–1593, 2005.

LU, J. et al. Transfer learning using computational intelligence: A survey. **Knowledge-Based Systems**, v. 80, p. 14–23, 2015.

LYKOURENTZOU, I. et al. Dropout prediction in e-learning courses through the combination of machine learning techniques. **Computers and Education**, v. 53, p. 950–965, 2009.

MACFADYEN, L. P.; DAWSON, S. Mining LMS data to develop an "early warning system" for educators: A proof of concept. **Computers and Education**, v. 54, p. 588–599, 2010.

MÁRQUEZ-VERA, C. et al. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. **Applied Intelligence**, v. 38, n. 3, p. 315–330, 2013.

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, fev. 2016.

MÁRQUEZ-VERA, C.; MORALES, C. R.; SOTO, S. V. Predicting School Failure and Dropout by Using Data Mining Techniques. **IEEE Journal of Latin-American Learning Technologies**, v. 8, n. 1, p. 7–14, fev. 2013.

MITCHELL, T. M. **Machine Learning**. McGraw-Hil ed. [s.l.] McGraw-Hill, 1997.

MØLLER, M. F. A scaled conjugate gradient algorithm for fast supervised learning. **Neural Networks**, v. 6, p. 525–533, jan. 1993.

PAN, S. J.; YANG, Q. A survey on transfer learning. **IEEE Transactions on Knowledge and Data Engineering**, v. 22, n. 10, p. 1345–1359, 2010.

PEÑA-AYALA, A. (ED.). **Educational Data Mining: Applications and Trends**. Cham: Springer International Publishing, 2014a. v. 524.

PEÑA-AYALA, A. Educational data mining: A survey and a data mining-based analysis of recent works. **Expert Systems with Applications**, v. 41, n. 4, part 1, p. 1432–1462, mar. 2014b.

QUINLAN, J. R. Induction of Decision Trees. **Machine Learning**, v. 1, n. 1, p. 81–106, 1986.

QUINLAN, J. R. **C4.5: Programs for Machine Learning**. San Mateo, CA: Morgan Kaufmann, 1993.

R CORE TEAM. **R: A Language and Environment for Statistical Computing.** Vienna, Austria, 2014. Disponivel em: <http://www.r-project.org/>. Acesso em: 29 abr. 2017.

REVOLUTION ANALYTICS; WESTON, S. **foreach: Provides Foreach Looping Construct for R**, 2015a. Disponivel em: <https://cran.r-project.org/package=foreach>. Acesso em: 29 abr. 2017.

REVOLUTION ANALYTICS; WESTON, S. **doParallel: Foreach Parallel Adaptor for the "parallel" Package**, 2015b. Disponivel em: <https://cran.r-project.org/package=doparallel>. Acesso em: 29 abr. 2017.

REVOLUTION ANALYTICS; WESTON, S. **doMC: Foreach Parallel Adaptor for "parallel"**, 2015c. Disponivel em: <https://cran.r-project.org/package=domc>. Acesso em: 29 abr. 2017.

ROJAS, R. **Neural Networks: A Systematic Introduction**. Berlin: Springer-Verlag, 1996.

ROMERO, C. et al. (EDS.). **Handbook of educational data mining**. Boca Raton, FL: CRC Press Taylor, 2011.

ROMERO, C. et al. Predicting students' final performance from participation in on-line discussion forums. **Computers and Education**, v. 68, p. 458–472, 2013.

SANCHEZ-SANTILLAN, M. et al. **Predicting Students' Performance: Incremental Interaction Classifiers**. Proceedings of the Third (2016) ACM Conference on Learning @ Scale. **Anais**...: L@S '16.New York, NY, USA: ACM, 2016.

SHANNON, C. E. A mathematical theory of communication. **The Bell System Technical Journal**, v. 27, p. 379–423, 1948.

SUMNER, M.; FRANK, E.; HALL, M. Speeding Up Logistic Model Tree Induction. In: **Knowledge Discovery in Databases: PKDD 2005**. Berlin: Springer-Verlag, 2005. v. 3721,p. 675–683.

THAMMASIRI, D. et al. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. **Expert Systems with Applications**, v. 41, p. 321–330, 2014.

UEKAWA, K. et al. **Creating an Early Warning System: Predictors of Dropout in Delaware**, 2010.

VILLAGRÁ-ARNEDO, C. et al. **Detección precoz de dificultades en el aprendizaje. Herramienta para la predicción del rendimiento de los estudiantes**. III Congreso Internacional sobre Aprendizaje, Innovación y Competitividad. **Anais**...2015

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. [s.l.] Springer-Verlag New York, 2009.

WICKHAM, H. The Split-Apply-Combine Strategy for Data Analysis. **Journal of Statistical Software**, v. 40, n. 1, p. 1–29, 2011.

WITTEN, I. H.; FRANK, E.; HALL, M. **Data Mining: Practical Machine Learning Tools and Techniques**. 3. ed. Burlington, MA: Morgan Kaufmann, 2011.

XING, W. et al. Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. **Computers in Human Behavior**, v. 58, p. 119–129, 2016.

YOU, J. W. Identifying significant indicators using LMS data to predict course achievement in online learning. **Internet and Higher Education**, v. 29, p. 23–30, 2016.

ZAFRA, A.; VENTURA, S. Multi-instance genetic programming for predicting student performance in web based educational environments. **Applied Soft Computing**, v. 12, p. 2693–2706, 2012.