DESCOBERTA DE CONHECIMENTO ATRAVÉS DE MÉTODOS DE APRENDIZAGEM DE MÁQUINA SUPERVISIONADOS APLICADOS AO SIGAA/UFPI

KNOWLEDGE DISCOVERY THROUGH SUPERVISED MACHINE LEARNING METHODS APPLIED TO SIGAA/UFPI

Aline Montenegro Leal Silva, Miguel Lino Ferreira Neto, Francisco das Chagas Imperes Filho, Raimundo Santos Moura, Vinícius Ponte Machado
Universidade Federal do Piauí (UFPI)
alineleal5@yahoo.com.br, miguelneto0019@gmail.com, fcoimperes@hotmail.com, rsm@ufpi.edu.br, vinicius@ufpi.edu.br

Resumo: O presente trabalho exibe um processo de descoberta de conhecimento através de métodos de Aprendizagem de Máquina (AM) supervisionados aplicados ao ensino superior a distância, mais especificamente à base de dados do SIGAA/UFPI, cujos registros foram coletados a partir do segundo semestre de 2014. Neste processo de descoberta de conhecimento, procurou-se a identificação de perfis dos alunos do curso de Licenciatura em Computação na modalidade a distância, a partir de uma correlação entre o IRA (Índice de Rendimento Acadêmico) e os aspectos sociais de tais alunos. Foram utilizados três algoritmos de AM supervisionados com diferentes paradigmas: *J48* (simbólico), *Naive Bayes* (estatístico) e *IBK* (baseado em exemplos). Os perfis descobertos podem auxiliar os gestores do sistema de educação a distância na tomada de decisões em relação a melhorias no processo de ensino-aprendizagem, já que através da mineração de dados tem-se uma ideia do desempenho do aluno, ao mostrar que a deficiência acadêmica possui correlações com aspectos sociais. A partir dessas informações é possível definir-se estratégias diferenciadas em relação a esses alunos, como um acompanhamento presencial por parte dos tutores nos polos de apoio do sistema de educação a distância.

Palavras-chave : Descoberta de Conhecimento; Aprendizagem de Máquina; Educação a Distância.

Abstract: The present work shows a process of knowledge discovery through supervised Machine Learning (ML) methods applied to distance higher education, more specifically to the SIGAA/UFPI database, whose records were collected from the second semester of 2014. In this process of Knowledge discovery, we sought to identify students' profiles of the Degree in Computing in the distance modality, to discover the correlation between the AII (Academic Income Index) and the social aspects of such students. Three supervised ML algorithms with different paradigms were used: J48 (symbolic), Naive Bayes (statistical) and IBK (based on examples). The discovered profiles can assist managers of the distance education system in decisions making regarding improvements in the teaching learning process, since through the data mining it has the sample of the student's development, by showing that the deficiency has correlations with social aspects. Based on this information, it is possible to define differentiated strategies in relation to these students, as a face-to-face monitoring by the tutors in the poles of support of the distance education system.

Keywords: Knowledge Discovery, Machine Learning, Distance Learning.

I. INTRODUÇÃO

As transformações mais marcantes ocorridas na Educação a Distância (EaD) refletem a mudança de um ambiente informacional para um ambiente de conhecimento, onde a figura do professor tutor é considerada como mediador da aprendizagem e o facilitador do acesso ao conhecimento com base no diálogo e na interação. Dessa forma, essa modalidade de educação propicia a criação de um ambiente educacional em que o aluno precisa ser um indivíduo capaz de mostrar autonomia e comprometimento com a aquisição de conhecimento, estimulando assim, o processo ensino-aprendizagem, já que ele ocorre em lugares e/ou tempos diferentes, o que pode dificultar o acompanhamento do aluno (FARIAS, 2013).

Na EaD, os ambientes de gestão da aprendizagem (LMS, do inglês *Learning Management System*) ou ambientes virtuais de aprendizagem (AVAs), a exemplo do SIGAA (Sistema Integrado de Gestão de Atividades Acadêmicas), abrangem funcionalidades para armazenar, distribuir e gerenciar conteúdos de aprendizado, de forma interativa e gradativa. Os LMS são desenvolvidos para permitir abordagens didáticas que auxiliem a promoção do ensino e da aprendizagem em situações de mediação virtual ou semi presencial e acumulam muitos dados também, já que todas as atividades do aluno encontram-se armazenadas em um banco de dados (BD), sendo geridas pelos LMS (CARVALHO et al., 2012)

Atualmente, é alarmante a distância crescente entre a geração de dados e a capacidade de analisá-los e compreendê-los. A medida que o volume de dados aumenta, a proporção dos dados que é analisada e entendida pelas pessoas diminui e escondido entre todo este volume de dados está a informação potencialmente útil (BATISTA, 2003). Existe, portanto, a necessidade de uma nova geração de técnicas e ferramentas que possibilite os analistas humanos compreenderam estas grandes bases de dados, as quais são objetos de estudo de uma área de pesquisa chamada Descoberta de Conhecimento em Base de Dados (DCBD). Para isso, usam-se ferramentas que deveriam ser capazes de criar por si próprias, a partir da experiência passada, uma hipótese ou função, capaz de resolver o problema que se deseja tartar. A esse processo dá-se o nome de Aprendizado de Máquina (AM) (FACELLI et al., 2011). Já a mineração de dados constitui uma das principais fases do processo de DCBD, em que algoritmos de AM são utilizados para a descoberta de conhecimento.

O presente trabalho exibe um processo de descoberta de conhecimento através de métodos de Aprendizagem de Máquina (AM) supervisionados aplicados ao ensino a distância, mais especificamente à base de dados do SIGAA/UFPI, cujos registros foram coletados a partir do segundo semestre de 2014. Neste processo de descoberta de conhecimento, procurou-se a identificação de perfis dos alunos do curso de Licenciatura em Computação na modalidade a distância, a partir de uma correlação entre o IRA (Índice de Rendimento Acadêmico) e os aspectos sociais de tais alunos. Foram utilizados três algoritmos de AM supervisionados com diferentes paradigmas: *J48* (simbólico), *Naive Bayes* (estatístico) e *IBK* (baseado em exemplos). Os perfis descobertos podem auxiliar os gestores do sistema de educação a distância na tomada de decisões em relação a melhorias no processo de ensino-aprendizagem, já que através da mineração de dados tem-se uma ideia do desempenho do aluno, ao mostrar que a deficiência acadêmica possui correlações com aspectos sociais. A partir dessas informações, é possível definir-se estratégias diferenciadas em relação a esses alunos, como um acompanhamento presencial por parte dos tutores nos polos de apoio do sistema de educação a distância.

II. FUNDAMENTAÇÃO TEÓRICA

A. *Descoberta de Conhecimento em Base de Dados*

Existem diversas definições para DCBD, mas uma das mais utilizadas é a proposta por FAYYAD et al. (1996), que define Descoberta de Conhecimento em Base de Dados como um processo não trivial para identificar padrões válidos, novos, potencialmente úteis e compreensíveis em dados existentes. Por não trivial é entendido que alguma busca ou inferência é utilizada. Os padrões descobertos pelo processo de DCBD devem ser validados com novos dados. Esses padrões também devem ser novos e potencialmente úteis, isto é, devem levar algum benefício ao usuário ou à aplicação. Por fim, os padrões devem ser compreensíveis aos analistas humanos, se não imediatamente, ao menos após algum pós-processamento (ARAUJO, 2014).

O processo de DCBD considerado neste trabalho encontra-se estruturado em 5 (cinco) fases: a) coleta: obtenção do conjunto de dados, b) pré-processamento: realiza a remoção de ruídos (dados inconsistentes) e balanceamento de classes, c) transformação: formatação dos dados para a aplicação dos algoritmos de Aprendizagem de Máquina, d) mineração de dados: aplicação dos algoritmos de AM para a obtenção de padrões, e) avaliação e interpretação dos resultados: corresponde à descoberta do conhecimento adquirido, como mostra a Figura 1.

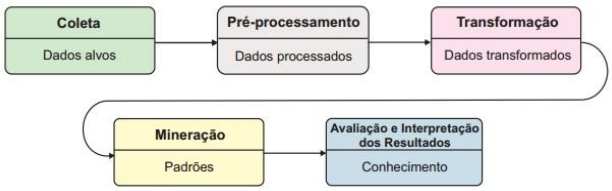

Figura 1. Fases do processo de descoberta do conhecimento (Adaptado de FAYYAD et al., 1996).

B. *Aprendizagem de Máquina*

A Aprendizagem de Máquina, do inglês *machine learning*, é utilizada na fase de mineração de dados do processo de descoberta do conhecimento em base de dados e surgiu da percepção de criar programas computacionais que aprendam um determinado comportamento ou padrão automaticamente, a partir de exemplos ou observações. A ideia por trás da aprendizagem é que percepções devem ser usadas não apenas para agir, mas também para melhorar a habilidade do agente para agir no futuro (RUSSEL & NORVIG, 2009).

Existem várias estratégias de aprendizagem que podem ser utilizadas por um sistema computacional como, por exemplo, o aprendizado por hábito, por instrução, por dedução, por analogia e por indução (COPPIN, 2010), como apresenta a Figura 2.

Figura 2. Estratégias do Aprendizado de Máquina (ARAUJO, 2014).

C. *Hierarquia do Aprendizado*

O aprendizado por indução é caracterizado pelo raciocínio que parte do específico para o geral. É um modo de inferência lógica que permite obter generalizações a partir de exemplos para induzir um conceito que pode ou não preservar a verdade. Assim, mesmo que as premissas sejam verdadeiras, pode-se chegar a conclusões falsas. Justamente por ser essa uma estratégia de aprendizado complexa, uma vez que o aprendizado desempenha a maior parte do esforço para a aquisição do conhecimento, esta permite que conceitos muitos mais amplos possam ser aprendidos. Portanto, constitui uma das estratégias de aprendizado de interesse para as pesquisas relacionadas ao aprendizado de máquina (ARAUJO, 2014).

O Aprendizado Indutivo pode ser dividido basicamente em Aprendizado Supervisionado e Não Supervisionado (MONARD AND BARANAUSKAS, 2003). No primeiro, busca-se a criação de um modelo preciso em relação à predição de valores para novos dados enquanto que no segundo o objetivo é encontrar características que podem resumir os dados. A diferença básica entre esses dois modos

de aprendizagem, é a presença ou não do atributo que rotula os exemplos do conjunto de dados. No Aprendizado Supervisionado esse rótulo é conhecido, ao passo que no aprendizado não supervisionado os exemplos não estão previamente rotulados. Adicionalmente, existe um terceiro modo de aprendizagem, conhecido como aprendizado semi-supervisionado, no qual somente poucos exemplos encontram-se rotulados. Esse fato impossibilita o uso direto de algoritmos de Aprendizado Supervisionado, pois esse modo de aprendizagem requer um número razoável de exemplos rotulados (BLUM & MITCHELL, 1998).

Portanto, no aprendizado de máquina supervisionado, o objetivo é induzir um classificador (ou hipótese), por meio de um conjunto expressivo de dados previamente rotulados, para classificar novos exemplos ainda não rotulados. Se os rótulos das classes possuem valores discretos, o problema é conhecido como classificação. Caso as classes possuam valores contínuos, o problema é conhecido como regressão. A Figura 3 mostra a hierarquia do aprendizado descrita.

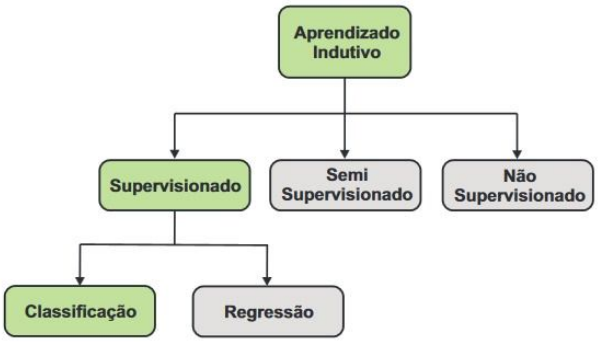

Figura 3. Hierarquia do Aprendizado (ARAUJO, 2014)

Como o atributo classe considerado nesse trabalho (Faixa_IRA) possui valores discretos, optou-se por fazer uma descoberta de conhecimento a partir de uma classificação previamente conhecida (IRA), portanto optou-se pelos algoritmos supervisionados.

D. Algoritmos de Aprendizado Supervisionado

Algoritmos de AM podem ser vistos como mecanismos que extraem um padrão de comportamento a partir de experimentações (MACHADO, 2011). Na classificação, que é utilizada nesse trabalho, um conjunto de treinamento com dados pré-classificados é usado para a criação de regras de classificação que permitem atribuir uma classe aos novos registros de dados.

A seguir são descritos alguns dos algoritmos de aprendizado supervisionados propostos na literatura, de acordo com os principais paradigmas desse aprendizado, já que esses algoritmos são indutivos e permitem obter-se generalizações, além da facilidade de interpretação dos resultados, principalmente o J48 através de suas regras de produção bastante intuitivas.

- *J48:* algoritmo de paradigma simbólico, surgiu da necessidade de recodificar o algoritmo C4.5 para a linguagem Java. Ele tem a finalidade de gerar uma árvore de decisão baseada em um conjunto de dados de treinamento, sendo este modelo usado para classificar as instâncias no conjunto de teste (WITTEN & FRANK, 2005).

- *Naive Bayes:* algoritmo de paradigma estatístico, baseia-se na aplicação da teoria Bayesiana para o cálculo das probabilidades necessárias para a classificação (MITCHELL, 1997; HAN & KAMBER, 2006). Tem como função predizer a probabilidade de um exemplo pertencer a uma determinada classe. Esse classificador considera a hipótese de que todos os atributos são independentes, dado a variável classe.

- *IBK:* algoritmo de paradigma baseado em exemplos, fundamenta-se no algoritmo *K-Nearest Neighbors (K-NN)*, o qual parte da ideia de que em um espaço n-dimensional, um ponto P e os seus K vizinhos mais próximos pertencem à mesma classe. A proximidade entre dois pontos nesse espaço é calculada por meio da distância euclidiana entre eles (SILVA, 2007).

III. SIGAA

Desde 2003, a Universidade Federal do Rio Grande do Norte iniciou um projeto denominado Bases de Dados Integradas que tinha como propósito construir um único banco de dados que integrasse a área acadêmica, administrativa e de recursos humanos e que tal base de dados servisse como repositório de informações para qualquer atividade da sua comunidade. Mais tarde, esse sistema passou a ser chamado de SIGAA.

Esse sistema informatiza os procedimentos da área acadêmica através dos módulos de: graduação, pós-graduação (stricto e lato sensu), ensino técnico, ensinos médio e infantil, submissão e controle dos projetos de ensino (monitoria e inovações), registro e relatórios da produção acadêmica dos docentes, atividades de ensino a distância e um ambiente virtual de aprendizagem denominado Turma Virtual. Atualmente o SIGAA está presente em mais de 29 universidades públicas do Brasil, incluindo a Universidade Federal do Piauí.

IV. FERRAMENTA WEKA

Para apoiar o desenvolvimento deste trabalho, fez-se uso da ferramenta *Weka*, que contém uma coleção de algoritmos de Aprendizagem de Máquina e ferramentas de pré-processamento de dados projetadas para que possam experimentar rapidamente os métodos existentes em novos conjuntos de dados. Ainda fornece suporte para todo o processo de mineração de dados experimental, incluindo a preparação dos dados de entrada, avaliação de sistemas de aprendizagem estatística, visualização dos dados de entrada e o resultado da aprendizagem (BOUCKAERT et al., 2015).

V. TRABALHOS RELACIONADOS

Para auxiliar o processo de descoberta de conhecimento proposto, foi realizado um mapeamento sistemático com o objetivo de identificar e sumarizar os algoritmos de aprendizagem de máquina existentes que oferecem suporte à

descoberta de padrões na área de sistema de gerenciamento de aprendizagem ou ambientes virtuais de aprendizagem.

A. Mapeamento Sistemático (MSE)

Um MSE é um método de pesquisa que permite identificar de forma sistemática a literatura em uma área de conhecimento, apresentando na maioria das vezes um resumo visual (mapa) dos resultados encontrados (PETERSEN et al. 2008).

Para apoiar o desenvolvimento do trabalho em questão, fez-se uso da ferramenta The End, desenvolvida pelo Laboratório de Engenharia de Software e Informática Industrial (Ea-SII), vinculado a Universidade Federal do Piauí. Ela oferece um serviço Web capaz de auxiliar o planejamento, execução e sumarização dos resultados de um MSE, tornando a pesquisa mais ágil e replicável (BRAGA et al. 2015). A Figura 4 exibe as etapas que compõem o processo de Mapeamento Sistemático implementado pela ferramenta The End.

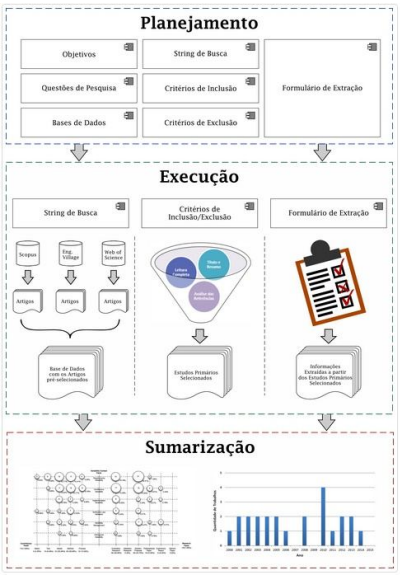

Figura 4. Etapas do processo de MSE - Fonte: Ferramenta The End.

Os resultados do MSE apresentados nesse trabalho seguem um conjunto de diretrizes propostas por (KITCHENHAM et al. 1995), com o objetivo de identificar e sumarizar os algoritmos de aprendizagem de máquina existentes que oferecem suporte à descoberta de padrões na área de sistema de gerenciamento de aprendizagem. A seguir apresenta-se o protocolo com as diretrizes que guiaram o estudo em questão.

B. Planejamento

Nessa etapa, os pesquisadores devem planejar a execução do estudo, definindo os objetivos, as questões de pesquisa, as bases de dados que serão utilizadas, a *string* de busca a ser usada nas bibliotecas digitais selecionadas, os critérios para inclusão e exclusão dos trabalhos obtidos, formulários para extração de informações relevantes, dentre outros aspectos.

B.1. Objetivos

O objetivo do MSE é identificar as principais técnicas de Inteligência Computacional (IC) e os algoritmos de Aprendizagem de Máquina para a determinação de padrões em base de dados relacionada a sistemas de gerenciamento de aprendizagem ou ambientes virtuais de aprendizagem.

B.2. Questões de Pesquisa

As questões de pesquisa foram derivadas a partir dos objetivos do mapeamento.

Estudos primários correspondem a investigações originais, que constituem as publicações encontradas nas bibliotecas digitais.

Visando estimular o levantamento e compreensão dos Estudos Primários (EP) sobre o assunto foco desse estudo, a seguinte questão de pesquisa (QP) direcionou esse trabalho.

QP1: Quais técnicas de Inteligência Computacional e algoritmos de Aprendizagem de Máquina apoiam pesquisadores e gestores na descoberta de padrões em relação aos sistemas de gerenciamento de aprendizagem? Essa questão orientou a elaboração das seguintes sub-questões de pesquisa:

- Quais técnicas de Inteligência Computacional foram utilizadas?
- Quais algoritmos de aprendizagem de máquina foram usados?

B.3. Base de Dados

No trabalho em questão, os artigos científicos foram pesquisados tomando como referência as bases de dados digitais: *Engineering Village*, *Scopus e Web of Science*, pois essas bases publicam artigos das principais conferências e autores da área de Inteligência Artificial.

B.4. Strings de Busca

A pesquisa inicia-se pela formação das *strings* de busca, que são os termos utilizados nas bibliotecas digitais para a busca de estudos primários. Visando tornar a busca mais refinada, convém utilizar os mecanismos de pesquisa avançada disponibilizados por cada base de dados. A tabela 1 mostra as *strings* de busca utilizadas na pesquisa em questão.

Tabela 1. *Strings* de busca aplicadas às bibliotecas digitais.

Base de Dados	***Strings* de Busca**
Engineering Village	“machine learning” and “learning management system”
Scopus	“machine learning” and “learning management system”
Web of Science	“machine learning” and “learning management system”

A Tabela 2 apresenta a quantidade de resultados obtidos ao aplicar as strings de busca nas bases de dados digitais.

Tabela 2. Resultado das pesquisas nas bibliotecas digitais.

Base de Dados	Resultados
Engineering Village	19
Scopus	29
Web of Science	7
Total	55

Dos 55 artigos candidatos, 19 estavam em duplicidade, restando, portanto, 36 artigos para serem analisados de acordo com os critérios de inclusão e exclusão.

B.5. Critérios de Inclusão

A inclusão de um trabalho é determinada pela relevância em relação às questões levantadas. Baseado nisso, decidiu-se que seriam incluídos na pesquisa os estudos que:

- Devem ser escritos em Inglês AND;
- Devem estar publicados em workshop, conferência, revista ou jornal entre os anos de 2000 e 2016 AND;
- Devem apresentar técnicas de Inteligência Computacional aplicadas à área de Educação a Distância (EaD) AND;
- Devem utilizar algoritmos de aprendizagem de máquina supervisionados ou não supervisionados aplicados ao Sistema de Gerenciamento de Aprendizagem.

B.6. Critérios de Exclusão

Trabalhos sem relevância em relação ao assunto de pesquisa devem ser excluídos. Portanto, decidiu-se que seriam excluídos da pesquisa os EP que:

- Não fossem escritos em Inglês OR;
- Não estivessem publicados em workshop, conferência, revista ou jornal entre os anos de 2000 e 2016 OR;
- Não apresentassem técnicas de Inteligência Computacional aplicadas à área de Educação a Distância (EaD) OR;
- Não utilizassem algoritmos de aprendizagem de máquina supervisionados ou não supervisionados aplicados ao Sistema de Gerenciamento de Aprendizagem.

B.7. Processo de Seleção de Estudos

O processo de seleção de estudos foi realizado pelas fases a seguir:

1. Leitura de títulos e resumo: nessa etapa, dois pesquisadores (primeiro e segundo autores) aplicaram os critérios de inclusão e exclusão no título e resumo de todos os 55 trabalhos candidatos à EP identificados durante a pesquisa. Desses 55 trabalhos, 25 foram aceitos pelos dois pesquisadores, 11 foram rejeitados e 19 trabalhos encontravam-se duplicados, resultando na eliminação dos trabalhos rejeitados e duplicados.

2. Leitura completa dos trabalhos: nessa etapa, dos 25 trabalhos aceitos, 7 foram removidos por estarem disponíveis somente em forma de abstracts ou resumos expandidos e 8 foram removidos após a leitura das seções de introdução e conclusão, resultando em 10 trabalhos aceitos.

3. Análise de Referências: Realizou-se uma análise das referências dos trabalhos aceitos e decidiu-se pela inclusão de 1 trabalho, totalizando 11 trabalhos aceitos.

B.8. Processo de Extração de Dados

A extração das informações foi realizada por dois pesquisadores (autores respectivamente) com o auxílio da ferramenta The End e de planilhas eletrônicas contendo campos gerais (título, autores, ano etc.) e campos específicos para cada questão de pesquisa, tais como técnicas de IC e algoritmos de AM utilizados.

C. Resultados

Esta seção apresenta a sumarização do MSE baseada nos 11 trabalhos incluídos como EP da pesquisa (Tabela 3), fornecendo uma identificação das técnicas de Inteligência Computacional e algoritmos de Aprendizagem de Máquina para a determinação de padrões em base de dados relacionada a sistemas de gerenciamento de aprendizagem ou ambientes virtuais de aprendizagem (Figura 5), que é o objetivo deste trabalho.

Tabela 3. Estudos primários selecionados.

Título	Autor	Ano
Analyzing learning concepts in intelligent tutoring systems	Gunel, K. and Polat, R. and Kurt, M.	2016
Educational Data Mining on Learning Management Systems using SCORM	Chandra, Deka Ganesh and Raman, Anupama C.	2014
Using Feature Selection and Association Rule Mining to Evaluate Digital Courseware	Shaveen Singh,Sunil Pranit Lal	2013
Towards improvements on domain-independent measurements for collaborative assessment	Anaya, A.R. and Boticario, J.	2013
Educational Courseware Evaluation Using Machine Learning Techniques	Singh, S. and Lal, S.P.	2013
Application of knowledge based decision technique to predict student enrollment decision	Malaya Dutta Borah, Rajni Jindal, Daya Gupta, Ganesh Chandra Deka	2011
Multidimensional adaptations for open learning management systems	Baldiris, S. and Santos, O.C. and Huerva, D. and Fabregat, R. and Boticario, J.G.	2008
Modelling collaborative competence level using machine learning techniques	Valetts, L.M. and Navarro, S.B. and Gesa, R.F.	2008
Question recommender with ML business logic	Burdescu, D.D. and Mihescu, M.C. and Logoftu, B.	2008
Creating glossaries using pattern-based and machine learning techniques	Westerhout, Eline and Monachesi, Paola	2008
Enhancing the assessment environment within a learning management systems	Burdescu, D.D. and Mihaescu, M.C.	2007

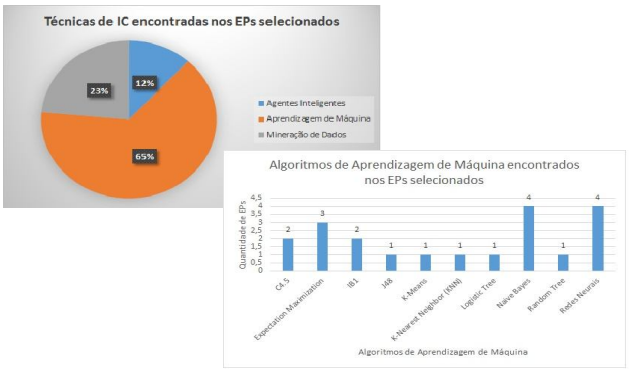

Figura 5. Técnicas de IC e Algoritmos de AM encontrados nos EPs.

Apesar da busca ter sido limitada a um período de tempo (2000 a 2016), observa-se que a maioria dos trabalhos concentraram-se nos últimos nove anos. Isto evidencia que pesquisas visando a identificação das técnicas de IC e dos algoritmos de Aprendizagem de Máquina utilizados para a determinação de padrões em base de dados relacionada a sistemas de gerenciamento de aprendizagem ou ambientes virtuais de aprendizagem têm aumentado na última década. O gráfico da Figura 6 ilustra a concentração dos estudos por ano.

Figura 6. Distribuição dos estudos primários ao longo dos anos.

Trabalhos propostos na literatura que utilizam técnicas de classificação de dados no processo de DCBD serviram como influência para a realização deste, alguns inclusive utilizaram a educação a distância como suporte.

No trabalho de VALETTS et al. (2008), foi apresentado um modelo de usuário para definir níveis de competência colaborativa dos alunos da modalidade a distância através do comportamento de colaboração. O modelo utilizou técnicas de Aprendizado de Máquina e gerou ações diferentes em uma plataforma de aprendizagem, tais como realizar recomendações para os professores proporem atividades para reforçar o nível do estudante ou para apoiar a tarefa de criação de grupos.

Em BORAH et al. (2011), foi realizado uma descoberta de conhecimento através de técnicas de seleção de atributos aplicadas ao algoritmo C.4.5 DT, na tentativa de otimizá-lo para uma tomada de decisão em relação a escolha do melhor ramo de engenharia a ser escolhido por um estudante de nível superior.

VIER et al. (2015) discutiu questões relevantes do uso de redes Bayesianas, integrantes do paradigma estatístico e que utiliza métodos de Aprendizagem de Máquina supervisionados para o ensino através de ambientes virtuais de aprendizagem. Propôs-se um Modelo de Aluno Probabilístico, que fornece informações importantes sobre o domínio do aluno para um Sistema Tutor Inteligente, voltado ao ensino de programação. As Redes Bayesianas (RB), também utilizadas nesse trabalho, são modelos probabilísticos que permitem lidar de forma rigorosa com a representação de conhecimentos em domínios onde existe incerteza. A estrutura da rede bem como suas probabilidades foram montadas a partir de entrevistas com especialistas, professores com longa experiência no ensino de programação. A rede proposta foi avaliada com a técnica de alunos virtuais. Através de simulações de 30 alunos virtuais, foi possível verificar que os resultados da rede ficaram muito próximos dos resultados das avaliações, em geral, apresentando uma diferença menor que 5% ou no máximo de até 35%. Esse trabalho mostrou que o uso de modelos probabilísticos é uma forma eficiente de se inferir o conhecimento do aluno.

Em ARAUJO (2014), foi realizado o aprendizado automático do processo de regulação médica/odontológica para reduzir os custos das operadoras de plano de saúde e aumentar sua eficiência na tomada de decisões, favorecendo assim a redução de custos também para os beneficiários, além de gerar mais agilidade durante a realização de exames/procedimentos/tratamentos. Para isso, retratou métodos de AM supervisionados, incluindo os paradigmas simbólico, estatístico, baseado em exemplos e conexionista. Para a realização do aprendizado a que se propõe, foi utilizado um *ensemble* para combinar o resultado dos três melhores classificadores (*RandomTree, RandomForest e KNN*). Vale ressaltar que os resultados obtidos pelo ensemble foram superiores aos obtidos por todos os outros classificadores individualmente.

FILHO et al. (2016) propôs um novo método para classificação no reconhecimento facial: o Classificador de Novidades. Para testar e avaliar o desempenho do classificador proposto, o mesmo foi comparado ao Vizinho mais Próximo (*KNN*), que corresponde ao *IBK* do *Weka* e também utilizado nesse trabalho, o que se obteve um resultado muito eficaz.

VI. MÉTODO

Neste trabalho, a metodologia empregada considerou o esquema de Descoberta do Conhecimento em Base de Dados proposto por HOLZINGER (2014) apud FAYYAD et al. (1996). Para a obtenção dos perfis dos alunos do curso de Licenciatura em Computação na modalidade a distância, a partir de uma correlação entre o IRA (Índice de Rendimento Acadêmico) e os aspectos sociais de tais alunos, executou-se sequencialmente todas as fases do esquema de DCBD considerado (coleta, pré-processamento, transformação, mineração de dados, avaliação e interpretação dos resultados).

Inicialmente, durante a fase de coleta, constatou-se que os dados encontravam-se disponibilizados em uma tabela com 19 atributos e 322 instâncias, cujos registros foram coletados a partir do segundo semestre de 2014, provenientes da base real do SIGAA fornecida pela Universidade Federal do Piauí.

Ao longo da fase de pré-processamento, observou-se que alguns atributos não tinham relevância para a pesquisa em questão, pois o objetivo era obter uma correlação entre o IRA dos alunos do curso de Licenciatura em Computação na modalidade a distância e os aspectos sociais de tais alunos. Portanto, utilizou-se o ganho de informação, que é uma técnica de seleção de atributos relevantes do *Weka*, que calculou a razão de ganho para cada um dos atributos e aqueles com razão de ganho nulo foram eliminados. Com isso, 13 atributos foram desconsiderados (Matrícula,

Código_situação, Situação, Curso, Código_estado_civil, Código_raça, Tipo_necessidade_especial, Tipo_sanguíneo, Código_cidade_residência, Cidade_residência, Bairro, País_residência, Nacionalidade) de um total de 19, restando apenas 6 (Polo, Sexo, Raça, Estado_civil, Ano_conclusao e IRA).

O atributo Faixa_IRA, resultante de uma discretização do atributo IRA, foi escolhido para ser o atributo classe pelo fato desse trabalho visar uma correlação entre o IRA e os aspectos sociais dos alunos. A princípio, o atributo classe recebeu as faixas de valores baixo (correspondente ao IRA de 0 a 3,9), regular (IRA de 4,0 a 6,9), bom (IRA de 7,0 a 8,9) e ótimo (IRA de 9,0 a 10,0), conforme será vista na seção Avaliação dos Resultados.

Notou-se também, ainda na fase de pré-processamento, que um dos atributos (Ano_conclusão), no caso se refere à conclusão do Ensino Médio, possuía 19 instâncias preenchidas de forma default, que após o tratamento de valores desconhecidos, foram ignoradas. A base resultante passou a ter 303 instâncias. Considerando ainda essa fase, percebeu-se que algumas classes encontravam-se desbalanceadas, sendo assim utilizado o filtro *Resample* da ferramenta *Weka*, que balanceia um conjunto de dados por meio de uma amostragem com reposição, mantendo o número de exemplos no conjunto de treinamento constante.

Já na fase de transformação, houve a normalização do atributo Estado_civil em Solteiro(a) e Não_Solteiro(a) visando realizar um equilíbrio dos dados, pois inicialmente tínhamos Solteiro(a), Casado(a), Divorciado(a), Separado Consensualmente, Separado Judicialmente e Outro, sendo que os quatro últimos campos citados foram encontrados em poucas instâncias. Ainda nessa fase, foi feita a discretização do atributo IRA em Faixa_IRA, tornando-o atributo classe, segundo já comentado anteriormente nessa seção.

A etapa seguinte consistiu na execução da fase de mineração de dados, onde algoritmos supervisionados foram utilizados para a classificação de padrões. Visando a obtenção da extração do conhecimento para a identificação dos perfis, foram escolhidos três algoritmos de paradigmas diferentes: *J48*, *Naive Bayes e IBK*, já que esses algoritmos são indutivos e permitem obter-se generalizações, além da facilidade de interpretação dos resultados, principalmente o *J48* através de suas regras de produção bastante intuitivas. Eles foram submetidos ao método de teste *Cross Validation* (Validação Cruzada), que divide o conjunto de dados em K (por padrão K=10) partições (*Folds*) e depois separa uma parte para teste e realiza o treinamento com as demais; este procedimento é repetido para todas as partes. Para avaliar a qualidade da classificação foram utilizadas duas métricas: a acurária, para medir a quantidade de instâncias corretamente classificadas (taxa de acerto) e com isenção de erros e a estatística *Kappa* (LANDIS & KOCH, 1977), que expressa a medida da diferença entre a concordância dos dados de referência e a classificação automática e a probabilidade de concordância entre os dados de referência e a classificação aleatória. Conceitualmente, o índice *Kappa* pode ser definido de acordo com a equação 1.

$$\kappa = \frac{N\sum X_{ii} - \sum X_{i+}X_{+i}}{N^2 - \sum X_{i+}X_{+i}} \quad (1)$$

X_{ii} = concordância observada

X_{+i} e X_{i+} (produto das marginais), sendo a concordância esperada

N = total de elementos observados

A Tabela 4 mostra a interpretação dos valores da estatística *Kappa*.

Tabela 4. Interpretação dos valores da estatística *Kappa*.

Valores de Kappa	Interpretação
<0	Sem concordância
0 – 0,20	Pobre
0,20 – 0,39	Ligeira
0,40 – 0,59	Moderada
0,60 – 0,79	Substancial
0,80 – 1,0	Excelente

Por último, na fase de avaliação dos resultados, foi apresentado o resultado obtido pelos algoritmos de classificação *J48*, *Naive Bayes e IBK* utilizados na fase de Mineração de Dados.

VII. AVALIAÇÃO DOS RESULTADOS

Nesta seção, são apresentados os resultados de três classificadores com diferentes paradigmas (*J48, Naive Bayes e IBK*), aplicados à base de dados do SIGAA/UFPI. Para chegarmos ao resultado obtido, foi feita uma sequência de cinco testes (T1, T2, T3, T4 e T5) na tentativa de encontrar os perfis dos alunos do curso de Licenciatura na modalidade a distância, a partir de uma correlação entre o IRA (Índice de Rendimento Acadêmico) e os aspectos sociais de tais alunos. As diferenças entre os testes realizados se deu por causa da discretização da Faixa_IRA. Em todos os classificadores considerou-se os mesmos atributos (Polo, Sexo, Raca, Estado_civil, Ano_conclusao, Faixa_IRA), quantidade de instâncias (303) e o atributo classe Faixa_IRA.

A. Classificador de Paradigma Simbólico

Inicialmente, os testes foram realizados com o algoritmo J48. No primeiro momento a discretização do atributo classe foi realizada com uma casa decimal, como mostra a Tabela 5 (Teste T1).

Tabela 5. Discretização inicial do atributo Faixa_IRA (Teste T1).

IRA	Faixa_IRA
[0,0;3,9]	Baixo
[4,0;6,9]	Regular
[7,0;8,9]	Bom
[9,0;10,0]	Ótimo

Em T1, a acurácia obtida foi de **53.4653%** e a estatística *Kappa* de **0.2157**, o que gera classificação não muito confiável. Tentando reduzir a perda de informação gerada pela discretização, já que cada faixa de valores em T1 estava com tamanho considerável e isso dificultava o conhecimento sobre a exata nota do aluno e pretendendo-se obter uma classificação mais confiável, realizou-se um novo teste de acordo com a tabela 6 (Teste T2).

Tabela 6. Nova Discretização do atributo Faixa_IRA (Teste T2).

IRA	Faixa_IRA
[0,0;1,1]	Insuficiente
[1,2;3,4]	Baixo
[3,5;6,2]	Regular
[6,3;7,8]	Bom
[7,9;10,0]	Ótimo

Nesse segundo teste, as 303 instâncias foram divididas em 5 grupos e a acurácia obtida foi de **32.0132%** e a estatística *Kappa* de **0.1485**. Uma nova sequência de testes foi realizada (Teste T3), dessa vez dividindo o total de instâncias em 10 grupos, na tentativa de reduzir ainda mais a perda de informação e gerar uma melhor classificação. A tabela 7 mostra como foi feita a discretização do atributo Faixa_IRA.

Tabela 7. Discretização posterior do atributo Faixa_IRA (Teste T3)

IRA	Faixa_IRA
[0,0;0,9]	Um
[1,0;1,9]	Dois
[2,0;2,9]	Três
[3,0;3,9]	Quatro
[4,0;4,9]	Cinco
[5,0;5,9]	Seis
[6,0;6,9]	Sete
[7,0;7,9]	Oito
[8,0;8,9]	Nove
[9,0;10,0]	Dez

A acurácia alcançada através do Teste T3 foi de **21.4521%** e a estatística *Kappa* de **0.0877**. Como pode-se observar, a medida em que aumentamos o conjunto de dados na discretização do atributo classe, a acurácia e a estatística *Kappa* tendem a diminuir. Então como a classificação foi piorando, decidiu-se voltar a faixas de valores mais restritas.

Novos testes foram feitos (Teste T4), dessa vez considerando menos conjuntos de dados para o Faixa_IRA, de acordo com a Tabela 8.

Tabela 8. Discretização do atributo Faixa_IRA com um conjunto de dados reduzidos (Teste T4)

IRA	Faixa_IRA
[0,0;3,9]	Baixo
[4,0;6,9]	Regular
[7,0;10,0]	Bom

Dessa vez, a acurácia obtida por meio de T4 foi de **54.4554%** e a estatística *Kappa* de **0.2316**. Então, finalmente, resolveu-se realizar a discretização com um conjunto menor de dados, dentre todos já testados. A tabela 9 mostra a discretização final do atributo Faixa_IRA (Teste T5), já que o perfil Reprovativo ou ainda Exame Final ou Aprovativo é suficiente para nosso processo.

Tabela 9. Discretização final do atributo Faixa_IRA (Teste T5).

IRA	Faixa_IRA
[0,0;3,9]	Reprovativo
[4,0;10,0]	Exame Final ou Aprovativo

Com esse novo teste (T5), conseguimos a maior acurácia alcançada até o momento para o atributo classe Faixa_IRA. A acurácia foi de **67.6568%** e a estatística *Kappa* **0.3377**. A partir desse resultado, resolveu-se considerar diversos subconjuntos aleatórios para teste, dessa vez com 200 instâncias, na tentativa de encontrar um conjunto representativo. O resultado disso ultrapassou a acurácia do teste anterior, obtendo os valores de **73%** para acurácia e **0.4325** para a estatística *Kappa* no melhor caso. Portanto, optou-se por não realizar mais nenhuma discretização em nenhum atributo para que o resultado pudesse refletir exatamente a realidade do curso de Licenciatura em Computação da Universidade Federal do Piauí na modalidade a distância, que possui níveis de aprovação elevados em determinados polos de apoio presencial do sistema de Educação a Distância, em contraposição a níveis baixos em outros polos.

Permitiu-se apenas aplicar o filtro *Resample* sobre o atributo classe, que faz uma subamostragem estratificada do conjunto de dados, o que gerou uma acurácia de 94,5% e índice *Kappa* de 0.8836. Verificou-se que os resultados após o balanceamento tiveram uma melhora significativa. Os demais algoritmos também farão uso do teste T5 com um conjunto de 200 instâncias.

B. Classificador de Paradigma Estatístico

O *Naive Bayes* foi o algoritmo de paradigma estatístico utilizado para classificar o conjunto de dados. A acurácia obtida por esse algoritmo através da discretização do atributo Faixa_IRA (teste T5) e considerando o mesmo subconjunto aleatório obtido pelo J48, que também indicou ser o melhor dentre todos os testes realizados pelo Naive Bayes, foi de **70,5%** e o índice *Kappa* de **0.3786**. Posteriormente, observou-se que a aplicação do filtro *Resample* à base de dados resultou

em uma acurácia de **79%** e estatística *Kappa* de **0.5566**. Constatou-se que os resultados alcançados pelo *Naive Bayes* foram um pouco inferiores ao do melhor teste do *J48*, tanto antes como após o balanceamento e os resultados apresentados por esse algoritmo não foram suficientemente claros para o processo.

C. Classificador de Paradigma Baseado em Exemplos

Utilizou-se o *IBK* como algoritmo de paradigma baseado em exemplos para a classificação de padrões. A acurácia atingida por esse classificador através do teste T5 e em seguida considerando o mesmo subconjunto aleatório obtido pelo *J48*, que também foi o melhor dentre todos os testes realizados pelo *IBK*, foi de **68%** e o índice *Kappa* de **0.2948**, para K=5. Em um segundo momento, notou-se que o filtro *Resample* quando aplicado à base de dados, gerou uma acurácia de **83,5%** e estatística *Kappa* de **0.6648**.

Observou-se que o *IBK* obteve resultados inferiores ao do melhor teste do *J48*, tanto antes quanto depois do balanceamento, mas superiores ao resultado do *Naive Bayes* após o balanceamento e os resultados apresentados por esse algoritmo também não foram suficientemente transparentes para o processo.

A tabela 10 apresenta as acurácias e índices *Kappa* obtidos através dos três classificadores aplicados à base de dados do SIGAA/UFPI.

Tabela 10. Resultado da avaliação dos classificadores *J48, Naive Bayes e IBK.*

Classificador	***J48***	***Naive Bayes***	***IBK***
Acurácia sem filtro	73%	70,5%	68%
Estatística Kappa sem filtro	0,4325	0,3786	0,2948
Acurácia com filtro	94,5%	79%	83,5%
Estatística Kappa com filtro	0,8836	0,5566	0,6648

VIII. INTERPRETAÇÃO DOS RESULTADOS

Considerando que o *J48* apresentou os melhores resultados dentre os classificadores considerados, a figura 7 mostra o resultado das regras de produção geradas por ele.

Podemos observar a identificação dos perfis dos alunos do curso de Licenciatura em Computação na modalidade a distância, a partir de uma correlação entre o IRA (Índice de Rendimento Acadêmico) e os aspectos sociais de tais alunos, conforme proposto.

Pôde-se observar que no polo da cidade de Bom Jesus, os alunos que obtiveram um melhor desempenho acadêmico foram aqueles que concluíram o ensino médio até 2009.

Já no polo da cidade de Inhuma, o pior desempenho foi dos alunos cujo ano de conclusão do ensino médio foi a partir de 2013.

No polo da cidade de Marcos Parente, constatou-se que as mulheres com ano de conclusão do ensino médio até 2001 tiveram um desempenho abaixo dos homens, considerando essa mesma época.

No polo da cidade de Pio IX, os alunos com ano de conclusão do ensino médio até 2011 e raça branca, se sobressaíram nos estudos em relação aos negros e pardos nesse mesmo período.

Já no polo da cidade de Piripiri, por exemplo, observou-se que as pessoas não-solteiras possuem maiores dificuldades para estudar.

No polo da cidade de São João do Piauí, os alunos com pior desempenho foram aqueles que concluíram o ensino médio até 2004.

Finalmente, no polo da cidade de Teresina, os alunos que obtiveram um melhor desempenho foram aqueles do sexo feminino que concluíram o ensino médio até 1997.

```
Polo = Bom Jesus
|   Ano_conclusao <= 2009: EXAME_FINAL_OU APROVATIVO
|   Ano_conclusao > 2009: REPROVATIVO
Polo = Inhuma
|   Ano_conclusao <= 2013: EXAME_FINAL_OU_APROVATIVO
|   Ano_conclusao > 2013: REPROVATIVO
Polo = Marcos Parente
|   Ano_conclusao <= 2001
|   |   Sexo = F: REPROVATIVO
|   |   Sexo = M: EXAME_FINAL_OU_APROVATIVO
|   Ano_conclusao > 2001: EXAME_FINAL_OU_APROVATIVO
Polo = Pio IX
|   Ano_conclusao <= 2011
|   |   Raca = Branco: EXAME_FINAL_OU_APROVATIVO
|   |   Raca = Negro: REPROVATIVO
|   |   Raca = Pardo: REPROVATIVO
|   Ano_conclusao > 2011: REPROVATIVO
Polo = Piripiri
|   Estado_civil = Nao_Solteiro(a): REPROVATIVO
|   Estado_civil = Solteiro(a)
|   |   Sexo = F: EXAME_FINAL_OU_APROVATIVO
|   |   Sexo = M: REPROVATIVO
Polo = Sao Joao do Piaui
|   Ano_conclusao <= 2004: REPROVATIVO
|   Ano_conclusao > 2004: EXAME_FINAL_OU_APROVATIVO
Polo = Teresina
|   Ano_conclusao <= 1997
|   |   Sexo = F: EXAME_FINAL_OU_APROVATIVO
|   |   Sexo = M: REPROVATIVO
|   Ano_conclusao > 1997: REPROVATIVO
```

Figura 7. Regras de produção geradas pelo classificador *J48*.

Os perfis descobertos podem auxiliar os gestores do sistema de educação a distância na tomada de decisões em relação a melhorias no processo de ensino-aprendizagem, já que através da mineração de dados tem-se uma ideia do desempenho do aluno, ao mostrar que a deficiência acadêmica possui correlações com aspectos sociais. No polo da cidade de Inhuma, por exemplo, os alunos com pior desempenho foram aqueles cujo ano de conclusão do ensino médio foi a partir de 2013. Nesse caso, isso nos faz refletir sobre a política educacional a partir desse período. Com base nessas informações, é possível definir-se estratégias diferenciadas em relação a esses alunos, como um acompanhamento presencial

por parte dos tutores nos polos de apoio do sistema de educação a distância.

IX. CONCLUSÕES

Este trabalho apresentou um processo de descoberta de conhecimento através de métodos de Aprendizagem de Máquina (AM) supervisionados aplicados à base de dados do SIGAA/UFPI, cujos registros foram coletados a partir do segundo semestre de 2014. Neste processo de descoberta de conhecimento, identificou perfis dos alunos do curso de Licenciatura em Computação na modalidade a distância, a partir de uma correlação entre o IRA (Índice de Rendimento Acadêmico) e os aspectos sociais de tais alunos. Foram utilizados três algoritmos de AM supervisionados com diferentes paradigmas: *J48* (simbólico), *Naive Bayes* (estatístico) e *IBK* (baseado em exemplos).

Constatou-se que a acurácia e o índice *Kappa* obtidos por meio do *J48* antes do balanceamento não foram tão satisfatórios, no caso foi de 73% e 0.4325 respectivamente, mas optou-se por gerar um resultado que pudesse refletir exatamente a realidade do curso em questão, que possui níveis de aprovação elevados em determinados polos de apoio presencial do sistema de Educação a Distância, em contraposição a níveis baixos em outros polos. Após o balanceamento, a acurácia passou a ser de 94,5% e o índice *Kappa* de 0.8836.

Observou-se que o algoritmo *J48* apresentou uma performance melhor em relação aos outros classificadores, tanto antes quanto depois do balanceamento, mostrando regras de produção bastante concisas que melhor representam a correlação do IRA com os demais atributos. Portanto, optou-se por utilizar as regras obtidas através do *J48* para o reconhecimento de padrões sobre os perfis, por serem mais confiáveis.

Percebeu-se, por exemplo, que no polo da cidade de Bom Jesus os alunos que obtiveram um melhor desempenho acadêmico foram aqueles que concluíram o ensino médio até 2009. Já no polo da cidade de Inhuma, o pior desempenho foi dos alunos cujo ano de conclusão do ensino médio foi a partir de 2013. No polo da cidade de Marcos Parente, notou-se que as mulheres com ano de conclusão do ensino médio até 2001 tiveram um desempenho abaixo dos homens, considerando essa mesma época. No polo da cidade de Pio IX, por exemplo, os alunos que concluíram o ensino médio até 2011 e que possuem a raça negra ou parda, têm um desempenho acadêmico muito baixo. No polo da cidade de Piripiri, observou-se que as pessoas não-solteiras possuem maiores dificuldades para estudar. No polo da cidade de São João do Piauí, os alunos com pior desempenho foram aqueles que concluíram o ensino médio até 2004. Constatou-se também que no polo da cidade de Teresina, os alunos que obtiveram um melhor desempenho foram aqueles do sexo feminino que concluíram o ensino médio até 1997.

Conforme visto, os perfis descobertos podem auxiliar os gestores do sistema de educação a distância na tomada de decisões em relação a melhorias no processo de ensino-aprendizagem, já que através da mineração de dados tem-se uma ideia do desempenho do aluno, ao mostrar que a deficiência acadêmica possui correlações com aspectos sociais. No polo da cidade de Inhuma, por exemplo, os alunos com pior desempenho foram aqueles cujo ano de conclusão do ensino médio foi a partir de 2013. Nesse caso, isso nos faz refletir sobre a política educacional a partir desse período. Com base nessas informações, é possível definir-se estratégias diferenciadas em relação a esses alunos, como um acompanhamento presencial por parte dos tutores nos polos de apoio do sistema de educação a distância.

Por fim, acredita-se que uma correlação entre o desempenho do trabalho do professor tutor em cada polo de apoio presencial do sistema de Educação a Distância e a performance dos alunos daquele polo constitui potenciais pontos para uma avaliação futura, bem como o uso de outros algoritmos de classificação para comparar com os resultados alcançados.

REFERÊNCIAS

[1] Farias, S. Os benefícios das tecnologias de informação e comunicação (TIC) no processo de educação a distância (EaD). Rev. digit. bibliotecon. cienc. inf., Campinas, SP, v.11, n.3, páginas 15-29, set/dez, 2013. ISSN 1678-765X.

[2] Carvalho, R., Filho, I., Vidal, T., Melo, R. & Gomes, A. Integração entre o sistema de gestão acadêmica e o sistema de gestão da aprendizagem: identificando necessidades e prototipando requisitos favoráveis a pratica docente. Revista Brasileira de Computação Aplicada (ISSN 2176-6649), Passo Fundo, v. 4, n. 1, páginas 81-91, Março, 2012.

[3] Batista, G. Pré-processamento de dados em aprendizado de máquina supervisionado. Simpósio Brasileiro de Automação Inteligente, Natal, Brasil, 2003.

[4] Facelli, K., Lorena, A. C., Gama, J., & Carvalho, A. Inteligência artificial: Uma abordagem de aprendizado de máquina. LTC, 2011.

[5] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. Knowledge discovery and data mining: Towards a unifying framework. Second International Conference on Knowledge Discovery and Data Mining (KDD-96), pages 82-88, 1996.

[6] Araujo, F. Descoberta de conhecimento em base de dados para o aprendizado da regulação médica/odontológica em operadora de plano de saúde. Dissertação de Mestrado. Maio, 2014.

[7] Russel, S., & Norvig, P. *Artificial intelligence: A modern approach*. 3rd. Prentice Hall, 2009.

[8] Coppin, B. *Inteligência artificial*. LTC, Rio de Janeiro - Brasil., 2010.

[9] Monard, M., & Baranauskas, J. *Conceitos sobre aprendizado de máquina*. Volume 1, Rezende, 1a edition, 2003.

[10] Blum, A. & Mitchell, T. Combining labeled and unlabeled data with co-training. Proceedings of the Eleventh Annual Conference on Computational Learning Theory, pages 92-100, New York, NY, USA. ACM, 1998.

[11] Machado, V. P. *Inteligência Artificial*. Editora EDUFPI, 2011.

[12] Witten, I. & Frank, E. Data mining: practical machine learning tools and techniques. Elsevier, 2th edition, 2005.

[13] Mitchell, T. *Machine learning*. McGraw-Hill, 1997.

[14] Han, J. & Kamber, M. *Data mining: concepts and techniques*. 2nd ed. San Francisco, CA: Morgan Kaufmann, 2006.

[15] Silva, E. R. J. Investigação de técnicas de extração e seleção de características e classificadores aplicados ao problema de classificação de dígitos manuscritos de imagens de documentos históricos. Dissertação de mestrado, Universidade Federal de Pernambuco, 2007.

[16] WEKA website. Disponível em: http://www.cs.waikato.ac.nz/ml/weka/ Acesso em: 02/03/2015.

[17] Bouckaert, R., Frank, E., Hall, M., Kirkby, R., Reutemanm, P., Seewald, A. & Scuse, D. *Weka manual for version 3-6-13,* 2015.

[18] Petersen, K., Feldt, R.M.S., Mattsson, M. *Systematic mapping studies in software engineering*. 12th International Conference on Evaluation and Assessment in Software Engineering, 2008.

[19] THE END website. Disponível em: http://easii.ufpi.br/theend/home Acesso em 10/02/2016.

[20] Braga, R., Oliveira, P., Souza, M., Santos, P., Rabelo, R. & Britto, R. *Ferramentas para o desenvolvimento de Sistemas Baseados em Inteligência Computacional: Um Mapeamento Sistemático*. Conferência: Simpósio Brasileiro de Automação Inteligente, Natal, Brasil, 2015.

[21] Kitchenham, B., Pickard, L., Pfleeger, S.L. *Case studies for method and tool evaluation*. IEEE software, 1995.

[22] ENGINEERING VILLAGE website. Disponível em: https://www.engineeringvillage.com/search/quick.url. Acesso em: 20/03/2016.

[23] SCOPUS website. Disponível em: http://www.scopus.com. Acesso em: 14/04/2016.

[24] WEB OF SCIENCE website. Disponível em: https://www.webofknowledge.com/. Acesso em: 18/05/2016.

[25] Valetts, L., Navarro, S. & Gesa, R. Modelling collaborative competence level using Machine Learning Techniques. IADIS International Conference e-Learning, 2008.

[26] Borah, M., Jindal, R., Gupta, D. & Deka, G. Application of knowledge based decision technique to Predict student enrollment decision. International Conference on Recent Trends in Information Systems, 2011.

[27] Vier, J., Gluz, J. & Jaques, P. Empregando redes bayesianas para modelar automaticamente o conhecimento dos aprendizes em lógica de programação. Revista Brasileira de Informática na Educação, Volume 23, Número 2, 2015.

[28] Filho, C., Falcão, T. & Costa, M. Comparação do desempenho do classificador de novidades com o classificador do vizinho mais próximo no reconhecimento facial. Simpósio Brasileiro de Automação Inteligente, Fortaleza, Brasil, Outubro, 2013.

[29] Holzinger, A., Dehmer, M. & Jurisica, I. Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directionsk. BMC Bioinformatics,15 Suppl 6:I1. doi: 10.1186/1471-2105-15-S6-I1, 2014.

[30] Landis, J., & Koch, G. The measurement of observer agreement for categorical data. Biometrics, Vol. 33, No. 1, pages 159-174, Março, 1977.