Mineração de Dados Educacionais: Uso de Redes Neurais Artificiais na Predição do Perfil Acadêmico do Aluno IFAL Campus Maragogi

Ewerton H. L. Silva[1], Jailton Cardoso da Cruz[1]

[1]Coordenadoria de Tecnologia da Informação – Instituto Federal de Alagoas (IFAL)
Rua Barão de Atalaia, Centro – Maceió – AL – Brasil

ewertonhallan@gmail.com, jailton@ifal.edu.br

Abstract. *This paper presents a case study on data mining use, through artificial neural networks application in the prediction of students profile. The study was carried out at the Maragogi campus of Alagoas Federal Institute, using academic data for deep learning in tool R application. The results indicated each student profile identification, thus reducing the institution spent time to detect possible teaching deficiencies, allowing the adoption of stimulus proactive actions to the students, aiming to surpass the evasion and disapproval high rates.*

Resumo. *Este artigo apresenta um estudo de caso sobre o uso de mineração de dados, mediante a aplicação de redes neurais artificiais na predição de perfis dos discentes. O estudo foi realizado no campus Maragogi do Instituto Federal de Alagoas, utilizando dados acadêmicos para a aplicação de deep learning na ferramenta R. Os resultados sinalizaram para a identificação do perfil de cada discente, reduzindo assim o tempo gasto pela instituição para detectar possíveis deficiências de ensino, permitindo a adoção de ações proativas de estímulo aos discentes, visando superar os altos índices de evasão e reprovação.*

1. Introdução

O Instituto Federal de Alagoas, Campus Maragogi, possui um papel importantíssimo na profissionalização e capacitação técnica dos jovens da região norte do estado. Compromete-se para a redução da desigualdade na educação e proporciona um desenvolvimento humano, econômico e social para as famílias dos jovens, formando profissionais qualificados para o mercado turístico da região.

O Campus atende aos jovens dos municípios de São Luiz do Quitunde, Passo do Camaragibe, Matriz do Camaragibe, São Miguel dos Milagres, Porto Calvo, Porto de Pedras, Japaratinga, Maragogi, dentre outros do estado de Alagoas, além de municípios circunvizinhos, pertencentes ao estado de Pernambuco.

Como abrange vários municípios e cada um deles possui suas dificuldades na educação básica (ensino fundamental), o campus fica na obrigação de tentar melhorar o nível de aprendizado destes jovens, reduzindo o alto nível de evasão e reprovação na

região, que atualmente gira em torno de 30%, conforme fonte do censo escolar 2014/2015[1].

As dificuldades que cada município possui na execução do projeto educacional, principalmente no ensino fundamental, gera uma deficiência no aprendizado destes alunos e é um dos fatores que proporciona um aumento da evasão escolar, visto que o estudante possui grandes dificuldades de acompanhar o conteúdo programático estabelecido pelo MEC.

Atualmente, o levantamento e identificação das dificuldades de cada um dos alunos que ingressam no IFAL – Maragogi, efetuado pelo setor pedagógico, demanda mais de seis meses, o que atrasa a adoção de ações de estímulo a tais discentes. Não bastasse isso, a promoção de qualquer medida demora cerca de dois meses após sua definição para surtir qualquer efeito, não tendo eficácia no ano letivo em curso, levando a altos índices de reprovação.

Diante deste quadro, tem-se que o emprego de Mineração de Dados Educacionais (MDE) tem sido efetivado em diversas áreas da educação, no sentido de dar suporte ao setor pedagógico [Baker e Yacef 2009]. Através de sua aplicação, se espera identificar possíveis fragilidades no campus/curso/disciplina do IFAL - Maragogi, mediante a análise de um vasto banco de dados na busca por padrões ou tendências que permitam analisar o rendimento do estudante no decorrer de toda a sua vida acadêmica.

Vale salientar que tal ferramenta visa avaliar o passado, utilizando análises multivariadas para, com o aprendizado, auxiliar os gestores na tomada de decisão.

2. Fundamentação Teórica

Mineração de Dados (MD) consiste em um esforço conjunto entre homem e máquina, através da exploração de dados armazenados, na busca de informações valiosas, com o objetivo de descobrir padrões e correlações. Tem como entrada os dados e como saída o conhecimento [Côrtes, Porcaro e Lifschitz 2002].

A MD faz parte de um processo chamado de Descoberta de Conhecimento em Banco de Dados *(Knowledge Discovery in Database -- KDD)*, no qual se utilizam conceitos de banco de dados, métodos estatísticos, ferramentas de visualização e técnicas de inteligência artificial. Pode-se dizer que a MD é vista como a etapa principal do KDD, onde ocorre a transformação dos dados e a geração do conhecimento, sendo imprescindível para o processo de tomada de decisão [Galvão e Marin 2009].

---

[1] https://g1.globo.com/educacao/noticia/abandono-no-ensino-medio-alcanca-11-do-total-de-alunos-apontam-dados-do-censo-escolar.ghtml

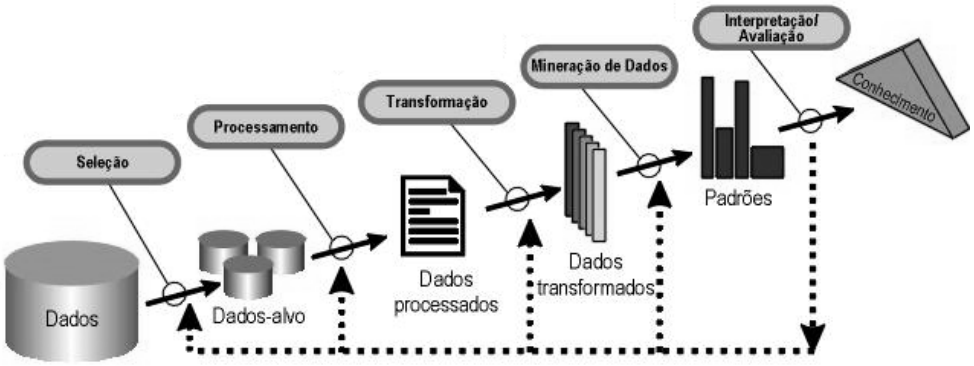

Figura 1. Etapas do KDD

Entretanto, para ser viabilizada, faz-se necessária uma preparação prévia dos dados. Etapas como seleção, pré-processamento e transformação têm um grande peso no resultado final da aplicação de uma mineração de dados, na medida em que é imprescindível identificar quais dados (atributos) serão selecionados e sua localização.

Dessa forma, tem-se que a preparação dos dados se inicia com a definição dos atributos a serem analisados, fase seguida do pré-processamento, no qual a preocupação se direciona ao conteúdo dos dados, com o tratamento dos nulos, duplicados, fora de contexto, valores incompletos, etc.

A transformação é a etapa que antecede a aplicação do algoritmo e é responsável por preparar os atributos de forma que possam atender aos requisitos do algoritmo de mineração. Há várias técnicas de transformação, dentre as quais se destacam: sumarização, agrupamentos, eliminação de valores fora do padrão, redução na quantidade de atributos. As etapas que antecedem a mineração visam melhorar a qualidade dos dados para aumentar a acurácia e eficiência do algoritmo [Côrtes, Porcaro e Lifschitz 2002].

A fase seguinte consiste na mineração de dados propriamente dita. Para sua aplicação, é importante definir inicialmente qual o objetivo na utilização do sistema de mineração de dados. Segundo [Costa, et al. 2012], há dois tipos de meta: verificação ou descoberta. Na primeira, o sistema se limita a verificar a ocorrência das hipóteses estabelecidas pelo usuário. Já a segunda busca encontrar novos padrões não previstos anteriormente, relacionando-se, geralmente, com os métodos de predição e descrição [Baker, Isotani e Carvalho 2006].

O método de descrição consiste em descrever padrões e tendências reveladas pelos dados e geralmente oferece uma possível interpretação para os resultados obtidos. É muito utilizada para exploração de dados, com a finalidade de comprovar a influência de certas variáveis no resultado revelado [Camilo e Silva 2009].

Já o método de predição tem como objetivo escolher o melhor resultado possível baseado na análise de dados históricos. A previsão deste resultado depende da relação existente entre os dados analisados [Camilo e Silva 2009].

Partindo deste contexto, tem-se que a Mineração de Dados Educacionais (MDE) pode ser definida como um ramo emergente da MD e se preocupa com o desenvolvimento de métodos para exploração dos tipos de dados únicos provenientes dos ambientes educacionais, bem como utiliza esses métodos para entender melhor os alunos e as características de como eles aprendem [Baker e Yacef 2009].

Na MDE, o método preditivo é amplamente recomendado com o uso da regressão, que tem como uma de suas técnicas mais populares as redes neurais artificiais (RNA), por ser uma forte opção para a busca de conhecimento baseada em experiências anteriores [Dias 2014].

As RNA compreendem procedimentos computacionais que envolvem o desenvolvimento de estruturas matemáticas com habilidade de aprendizado. São programas que implementam uma detecção sofisticada de padrões, através de algoritmos de aprendizado de máquina. Este modelo foi inspirado na estrutura neural humana, que adquire conhecimento através da experiência [Côrtes, Porcaro e Lifschitz 2002].

Uma das propriedades mais importantes de uma rede neural artificial é a capacidade de aprender por intermédio de exemplos e fazer inferências sobre o que aprendeu, melhorando gradativamente o seu desempenho [Ferneda 2006].

Tal aprendizado é feito através do treinamento da RNA, o que pode ser promovido mediante o algoritmo de retropropagação ou *backpropagation*, o mais comumente utilizado, tanto em redes multicamadas como simples [Santos, et al. 2005]. A retropropagação consiste em fazer ajustes nos pesos dos neurônios, no sentido inverso (da camada de saída para a camada entrada) [Iyoda 2000].

3. Metodologia

A presente pesquisa detém uma abordagem quantitativa, na medida em que busca padrões nos dados acadêmicos coletados e estudados, criando um perfil do aluno através de análise estatística, permitindo a identificação dos discentes com base no rendimento escolar.

Quanto ao objetivo, a pesquisa se classifica como descritiva, pois visa identificar as possíveis fragilidades dos alunos com base na análise histórica das características acadêmicas dos estudantes que adentraram na instituição entre os anos de 2015 e 2017.

Os dados utilizados neste estudo são provenientes de relatórios arquivados e material eletrônico disponível no sistema acadêmico do IFAL (SIGAA[2]), referentes aos anos de 2015 a 2017, dos cursos de hospedagem e agroecologia do campus Maragogi.

Após a coleta dos dados percebeu-se que era necessário concentrar toda a informação em um único meio físico, providência que fora adotada através da convergência de todos os dados para uma mesma estrutura, neste caso banco de dados relacional (MySQL[3]).

---

[2] Sistema Integrado de Gestão de Atividades Acadêmicas - IFAL

[3] Sistema de Gerenciamento de Banco de Dados - MySQL

Depois de convergir todos os dados, iniciou-se o tratamento, transformação e preparação dos mesmos para trabalhar com a RNA. Neste momento, foi incluída a etapa da verificação da relevância existente entre as variáveis coletadas.

A aplicação da RNA multicamada se deu através da utilização do algoritmo de reconhecimento de padrões *h20 deep learning* com retropropagação, onde foi adotado o uso de 1 (uma) a 3 (três) camadas internas, com até 8 (oito) neurônios em cada uma delas.

Adotou-se a ferramenta estatística R para trabalhar com RNA multicamadas, por possuir diversos recursos que agilizam o trabalho da aplicação da RNA em diversas configurações de camadas internas e quantidade de neurônios.

4. Resultados

O presente estudo tem como meta identificar o perfil de cada aluno do IFAL - Maragogi, de acordo com critérios a serem definidos pelo setor de ensino da instituição, apontando suas fragilidades no aprendizado, de modo a servir de subsídio na busca de soluções que minimizem os problemas de desempenho.

Os dados acadêmicos para estudo foram obtidos de uma amostra dos registros acadêmicos do campus Maragogi, compreendendo os anos de 2015 a 2017. De um quantitativo de 9.431 registros, foram utilizados 4.560, alusivos aos 2 cursos existentes. Após a etapa de extração e transformação, foram eliminados os registros inconsistentes, restando cerca de 48% dos dados originais para estudo.

As variáveis utilizadas na pesquisa foram: curso, ano letivo, disciplinas, notas, classificação no exame de seleção, tipo de concorrência, idade, município de origem. Tais dados foram trabalhados para garantir sua consistência, integridade e limpeza através de programa desenvolvido pelo autor para esta finalidade, não sendo adotado nenhum software de extração de dados (ETL) já existente.

Na preparação dos dados para aplicação da RNA, fez-se necessário efetuar um tratamento com a finalidade de detectar as relevâncias de cada variável com a resposta de retorno da RNA. Para tanto, foi utilizada a Análise de Agrupamento para identificar as relevâncias através do método euclidiano, usando a soma quadrada do erro.

Na figura 2 é mostrada a aplicação do *cluster analisys* antes do tratamento para melhorar a relevância das variáveis estudadas e o resultado obtido após este tratamento.

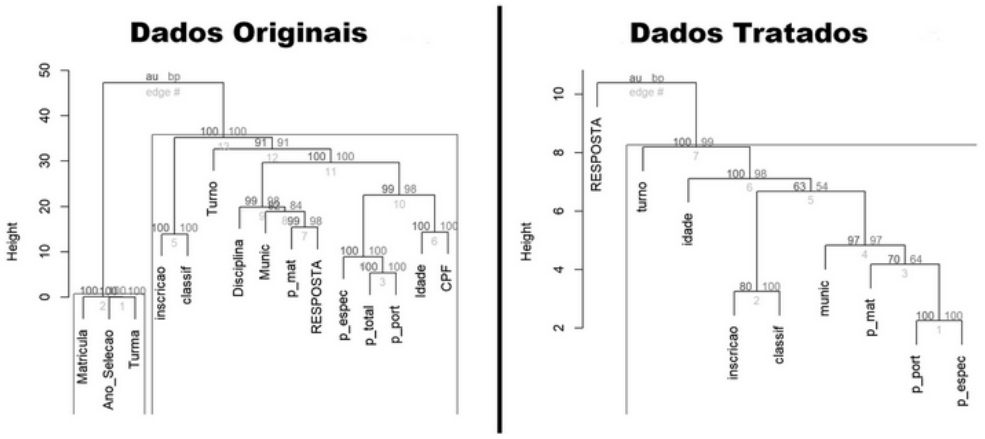

Figura 2. Cluster Analisys

Verificou-se, portanto, que, após o tratamento, as variáveis utilizadas no grupo de **dados tratados** possuíam um alto grau de relevância, proporcionando uma maior qualidade no aprendizado da RNA.

Ao implementarmos uma RNA, faz-se necessário separar um conjunto de dados para treinamento e outro para teste [Santos, et al. 2005]. No presente estudo, o conjunto de teste foi definido de forma aleatória no quantitativo de 20% da amostra total. O restante da amostra foi destinado ao treinamento da RNA.

Quanto à definição da quantidade de neurônios usados na camada oculta, esta deve ser feita de forma empírica, de acordo com a experiência do pesquisador e conhecimento teórico. Deve-se considerar que um número excessivo de neurônios e/ou camadas pode acarretar em problemas de generalização. Adotamos, no presente caso, a aplicação de um intervalo entre 1 a 7 neurônios por camada, com no máximo 3 camadas internas [Binoti, et al. 2014].

O uso de um número pequeno de neurônios na camada oculta favorece um melhor poder de generalização, reduzindo o risco de *Overfitting e Overtrainnig* (problemas de generalização) [Santos, et al. 2005].

Os resultados obtidos após a aplicação da RNA identificaram os padrões e tendências que deram origem a estudantes bem sucedidos, permitindo a repetição de casos de sucesso, bem como as dificuldades dos demais alunos, para que as mesmas sejam sanadas.

Após a transformação dos dados e aplicação das Redes Neurais na ferramenta R utilizando o algoritmo *h20 deep learning*, foram avaliadas diversas configurações de números de camada oculta e quantidade de neurônios em cada camada oculta, chegando-se aos seguintes resultados:

Tabela 1. Distribuição de neurônios por camadas ocultas, com percentual de acertos

Id	Conf. Camada Oculta	Percentual de Acertos
01	2 Camadas [2, 2]	83%
02	**3 Camadas [2, 2, 2]**	**93%**
03	3 Camadas [3, 3, 3]	90%
04	2 Camadas [4, 4]	80%
05	2 Camadas [5, 5]	80%
06	3 Camadas [5, 5, 5]	80%
07	3 Camadas [6, 6, 6]	87%
08	3 Camadas [7, 7, 7]	87%
09	2 Camadas [2, 3]	83%
10	3 Camadas [2, 3, 4]	90%
11	3 Camadas [3, 4, 5]	80%
12	**2 Camadas [4, 5]**	**93%**
13	**3 Camadas [4, 5, 6]**	**93%**

Destaque-se que a tabela 1 só contempla os resultados que obtiveram percentual de acurácia maior e igual a 80% de significância.

Dentre as diversas configurações aplicadas, as que melhor representaram o perfil do aluno com o menor erro foram as RNA de id's número 2, 12 e 13, com acerto de aproximadamente 93% do resultado previsto.

Percebe-se, ainda, que a RNA com 3 camadas ocultas produziu mais resultados com nível de acurácia melhor do que quando utilizadas apenas 1 ou 2 camadas. Destaca-se aqui a RNA com 2 camadas de configuração de id número 12 (camadas [4,5]), caso único dentre as configurações estudadas, onde o ganho na velocidade de processamento não foi fator relevante para se determinar o uso eficiente de 2 camadas ocultas como solução.

Desta forma, pode-se afirmar que a RNA proporcionou o resultado esperado para identificar e apontar as fragilidades de ensino a serem sanadas no Campus IFAL - Maragogi.

Através da RNA implementada com o presente estudo, é possível determinar, de forma rápida e ágil, o perfil de cada discente que entra na instituição, de acordo com critérios preestabelecidos pelo setor pedagógico, possibilitando definir quais ações de estímulo serão adotadas, tornando-as mais eficientes, na medida em que, desde o primeiro dia de aula, já se terá em mãos tais definições, ao contrário do que ocorria normalmente quando já se tinha passado quase 6 meses para poder identificar os problemas e dificuldades, além da enorme quantidade de trabalho para efetuar este tipo de levantamento.

5. Conclusão

A Mineração de Dados Educacionais, por intermédio das Redes Neurais Artificiais, é um mecanismo que merece ser reconhecido por sua habilidade de promover a predição de fatos escolares através da análise de dados históricos, com capacidade de aprendizado.

Com o presente trabalho, considerando o treinamento e teste da rede neural projetada, será possível, quando alimentada com dados de um novo indivíduo proveniente da região atendida pelo IFAL - Maragogi, predizer o futuro desempenho do novo aluno, classificando-o em perfis predispostos pela administração do campus, como forma de permitir melhorar a atuação dos profissionais responsáveis por sanar os problemas educacionais e acompanhar o coeficiente acadêmico.

Fazendo o uso de RNA na ferramenta R, utilizando o algoritmo *deep learning* com retropropagação e 03 (três) camadas ocultas internas, conseguiu-se uma rede neural artificial capaz de predizer, com precisão de até 93% (noventa e três por cento), o provável coeficiente acadêmico de cada aluno admitido na instituição, apenas com base nas notas obtidas no exame de seleção, idade, curso escolhido e município de origem.

Os resultados obtidos demonstraram um grande potencial de aplicação das Redes Neurais Artificiais para a criação do perfil e detecção de possíveis alunos que necessitam de um acompanhamento especial para compensar a defasagem educacional da rede privada, estadual e municipal de ensino.

6. Referências

Baker, R. e Yacef, K. (2009). "The state of educacional data minig: a review and future visions", In: Journal of educational data minig. V. 97, p 320-324. Disponível em: https://doi.org/10.1016/j.sbspro.2013.10.240. Acesso em: 14 set. 2018.

Baker, R. S. J., Isotani, S. e Carvalho, A. M. J. B. (2006). "Mineração de dados educacionais: Oportunidade para o Brasil", In: Revista Brasileira de Informática na Educação - RBIE. V. 19, n. 2. Disponível em: http://br-ie.org/pub /index.php/rbie/article/view/1301/1172. Acesso em: 01 set. 2018.

Binoti, M. L. M. S. et al. (2014). "Redes Neurais Artificiais para Estimação do Volume de Árvore", In: Revista Árvore. V. 38, n. 2, p. 283-288. Disponível em: http://dx.doi.org/10.1590/S0100-67622014000200008. Acesso em: 07 jun. 2018.

Camilo, C. O. e Silva, J. C. (2009). "Mineração de dados: Conceitos, Tarefas, Métodos e Ferramentas". Universidade Federal de Goiás – GO. Disponível em: http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf. Acesso em: 03 ago. 2018.

Côrtes, S. C., Porcaro, R. M. e Lifschitz, S. (2002). "Mineração de dados - funcionalidades, técnicas e abordagens". Pontifícia Universidade Católica do Rio de Janeiro – RJ. Disponível em: https://www.researchgate.net/publication/ 228912907_Mineracao_de_dados-funcionalidades_tecnicas_e_abordagens. Acesso: 3 out. 2018.

Costa, E. et al. (2012). "Mineração de dados educacionais: Conceitos, Técnicas, Ferramentas e Aplicações", In: Jornada de atualização em informática na educação

(JAIE 2012). Disponível em: http://www.br-ie.org/pub/index.php/pie /article/view/2341/2096. Acesso em: 01 jun. 2018.

Dias, M. M. (2014). “Minerando dados educacionais: relato de experiência no ambiente virtual labsql”. Dissertação (Mestrado) - Universidade Federal do Pará. Disponível em:http://repositorio.ufpa.br/jspui/bitstream/2011/9013/1/Dissertacao_MineracaoDadosEducacionais.pdf. Acesso em: 09 jun. 2018.

Ferneda, E. (2006). “Redes neurais e sua aplicação em sistemas de recuperação de informação”. USP - Ribeirão Preto - SP. Revista ibict. V. 35, n. 1, p. 25-30. Disponível em: http://revista.ibict.br/ciinf/article/view/1149/1312. Acesso em: 03 set. 2018.

Galvão, N. D. e Marin, H. F. (2009). “Técnica de mineração de dados: uma revisão da literature”, In: Acta Paulista de Enfermagem. V. 22, n. 5, p. 686-690. Disponível em: http://dx.doi.org/10.1590/S0103-21002009000500014. Acesso em: 9 jun. 2018.

Iyoda, E. M. (2000). “Inteligência computacional no projeto automático de redes neurais híbridas e redes neurofuzzy heterogêneas”. Universidade Campinas de São Paulo. Disponível em: http://www.dca.fee.unicamp.br/~vonzuben/research/emi_mest.html. Acesso em: 13 out. 2018.

Santos, A. M. et al. (2005). “Usando redes neurais artificiais e regressão logística na predição da hepatite a”, In: Revista Brasileira Epidemiológica. V. 8, n. 2, p 117-126. Disponível em: http://dx.doi.org/ 10.1590/S1415-790X2005000200004. Acesso em: 12 set. 2018.