

Análise preditiva do desempenho dos alunos do curso de pedagogia no âmbito da educação à distância no ENADE

Palabras clave: Mineração de dados - Floresta Aleatória - Aprendizado de Máquina - software R

Uma das formas de analisar as instituições de ensino superior e o desempenho dos estudantes no Brasil é através do Exame Nacional de Desempenho dos Estudantes (ENADE). O ENADE busca avaliar o grau de conhecimento dos estudantes em relação aos conteúdos programáticos previstos nas diretrizes curriculares de seus respectivos cursos de graduação. A partir dos seus resultados é possível tomar decisões inteligentes para a melhoria do processo de ensino-aprendizagem.

O Exame é composto por uma prova, bem como um questionário aplicado aos alunos com o intuito de coletar informações quanto ao seu perfil socioeconômico e aspectos relacionados à sua formação. Portanto, o Enade gera um grande volume de dados que permite estudos visando à melhoria da qualidade do ensino.

A ideia deste trabalho foi identificar fatores que possam influenciar no processo de aprendizagem dos discentes. Mais especificamente, buscam-se relações entre as respostas dadas ao questionário e o resultado obtido na prova. O enfoque é dado aos alunos do curso de Licenciatura em Pedagogia na modalidade de Educação a distância (EaD), oferecido através do Consórcio Centro de Educação Superior à Distância do Estado do Rio de Janeiro (CEDERJ). O Cederj é um consórcio formado por sete universidades públicas do Estado do Rio de Janeiro (UERJ; UENF; UNIRIO; UFRJ; UFF; UFRRJ; IFF) e um centro universitário (CEFET-RJ) em parceria com a Secretaria de Estado de Ciência, Tecnologia e Inovação do Rio de Janeiro, por intermédio da Fundação Centro de Ciências e Educação Superior à Distância do Estado do Rio de Janeiro (CECIERJ), com o objetivo de oferecer cursos de graduação a distância, na modalidade semipresencial para todo o Estado do Rio de Janeiro.

A EaD traz ao ensino superior, alguns desafios e potencialidades a serem explorados, como uma nova forma de se obter conhecimento e qualificação profissional. Sendo assim, como a educação superior tem um papel estratégico na formação, na produção do conhecimento e na informação, a constituição de um sistema de educação superior com modalidade EaD pode contribuir para que os países se adaptem as mudanças de larga escala em curso no mundo todo.

Este trabalho teve como objetivo utilizar a técnica de mineração de dados (TAN et al., 2009) com a utilização de métodos de aprendizado de máquina (MONARD e BARANAUSKAS, 2003). A mineração de dados têm como principal objetivo extrair o máximo de informação de base de dados extensas e, para isso, utiliza-se de técnicas como análise exploratória básica (DA CUNHA e CARVAJAL, 2009), particionamento recursivo (STEINER et al., 2004), análise de agrupamentos (KAUFMAN e ROUSSEEUW, 2009), modelagem de regressão (BISHOP, 2006; FOX, 1997), entre outros. Com os avanços tecnológicos de armazenamento de dados, essa área têm ganhado mais visibilidade (HAND, 2006).

O aprendizado de máquina é um ramo de inteligência artificial cujos estudos são feitos para que, com o auxílio de algoritmos, o computador tome decisão com base em informações inseridas neles. Para esse trabalho, o algoritmo utilizado é a regressão de floresta aleatória ou em inglês Radom Forest (BREIMAN, 2001), como um modelo de regressão capaz de prever o desempenho do estudante no ENADE através das variáveis socioeconômicas, buscando identificar e analisar o perfil dos estudantes que prestaram a prova.

Dessa maneira, o aprendizado de máquina fornece a base técnica para a mineração de dados que transforma dados brutos em informações de mais fácil compreensão, como previsões, correlações e relações de causalidade, o que, no processo de análise, auxilia na compreensão e explicação de fenômenos. No que tange os pacotes utilizados no *software* R, destaca-se o *tidyverse*, na parte de manipulação de dados, *randomForest* com o objetivo de obter a regressão de florestas aleatórias, *caret* e *Metrics* para obter métricas de avaliação do modelo e, por fim, *ggplot2* na visualização gráfica.

A base de dados utilizada para a realização deste estudo foi adquirida no portal do INEP, onde os dados estão disponíveis para o público através de download (INEP, 2017). Os dados escolhidos são oriundos da base de dados do ENADE 2017, que contém dados relativos aos estudantes que realizaram o exame e responderam

ao questionário.

Referências

- BISHOP, C. M.. Pattern recognition and machine learning. Springer, 2006.
- BREIMAN, L.. Random forests. Machine learning, Springer, v. 45, n. 1, p. 5–32, 2001
- DA CUNHA, S. B.; CARVAJAL, S.. Estatística Basica - a Arte de Trabalhar com Dados. Elsevier Brasil, 2009.
- FOX, J.. Applied regression analysis, linear models, and related methods. Sage Publications, Inc, 1997.
- HAND, D. J. Data Mining. Encyclopedia of Environmetrics , v. 2, John Wiley & Sons, Ltd.,2006.
- KAUFMAN, L.; ROUSSEEUW, P. J.. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, 2009.
- MONARD, M. C.; BARANAUSKAS, J. A.. Conceitos de aprendizado de máquina. In S. O. Rezende, editor, Sistemas Inteligentes - Fundamentos e Aplicações, pages 89–114. Editora Manole, 2003.
- RAMESH, V. PARKAVI, P. e RAMAR, K.(2013) Predicting Student. Performance: A Statistical and Data Mining Approach. International Journal of Computer Applications. [S.l: s.n.].
- STEINER, M. T. A. et al.. Data Mining como Suporte à Tomada de Decisões-uma Aplicação no Diagnóstico Médico. XXXVI Simpósio Brasileiro de Pesquisa Operacional, “O impacto da pesquisa operacional nas novas tendências multidisciplinares”, v. 23, p. 96-107, 2004.
- TAN, PANG-NING et al.. Introduction to Data Mining, Pearson, 2009.