

Predição de Problemas no Desempenho de Estudantes de um Curso Técnico de Nível Médio: Uma Análise Comparativa de Algoritmos de Mineração de Dados

Renan Aleixo Paganatto¹, Ana Maria Monteiro¹

¹Centro Universitário Campo Limpo Paulista (UNIFACCAMP) – Campo Limpo Paulista – SP – Brasil

{renan.ccn@gmail.com, ana.monteiro@faccamp.br}

Abstract. *Education plays a fundamental role in the development of a society, especially in a country like Brazil. For those in charge of managing educational activities and for teachers, it is very important to be able to predict possible problems in the students' performance, which eventually may cause some of them to leave the educational environment. In general, experienced teachers are able to predict the future performance of their students. But in addition to the contribution of these teachers, to better understand how to predict student performance, the growing amount of electronic data available in educational institutions can be used. From the application of mining techniques on the available data, meaningful information can be extracted to be used by educational managers and teachers, to make decisions and apply preventive measures when necessary to improve student performance and the quality of the educational process. In this research, data from an Integrated High School Internet Computing course at a State Technical School (ETEC) were used. The data originates from the first years of the course with 477 students, containing 28 attributes, second years of the course with 396 students, containing 27 attributes and third years of the course with 385 students, containing 27 attributes, the data was collected between 2014 and 2020. To predict student performance, experiments were conducted using the Naive Bayes, J48 (decision tree) and IBk (Knn) algorithms, all available in the WEKA environment. The metrics used to evaluate the algorithms are accuracy, sensitivity, F-measure and the area under the ROC curve. The objective of the experiments was to determine which of the algorithms used can best classify the data of the students who are attending the institution. Students are classified according to whether they are approved, approved with pending for the next year, failed, or dropped out. With the algorithms used in the experiments, the best classification results were obtained with the application of the SMOTE technique and Costs, with respect to the accuracy of the results.*

Keywords: *Student performance prediction, educational data mining, decision trees, Naive Bayes algorithm, Knn.*

Resumo. *A educação tem um papel fundamental no desenvolvimento de uma sociedade, em especial em um país como o Brasil. Para os encarregados de gerenciar atividades educacionais e para os professores é muito importante poder prever possíveis problemas no desempenho dos estudantes o que, eventualmente, pode fazer com que alguns deles abandonem o ambiente*

educacional. Em geral, professores experientes conseguem prever o desempenho futuro de seus estudantes. Mas além da contribuição desses professores, para entender melhor como antever o desempenho dos alunos, pode ser utilizada a crescente quantidade de dados eletrônicos disponíveis nas instituições de ensino. A partir da aplicação de técnicas de mineração nos dados disponíveis, pode ser extraída informação significativa a ser utilizada por gestores educacionais e professores, para tomar decisões e aplicar medidas preventivas quando necessário para melhorar o desempenho dos alunos e a qualidade do processo educacional. Nesta pesquisa foram utilizados dados provenientes de um curso de Informática para Internet Integrado ao Ensino Médio de uma Escola Técnica Estadual (ETEC). Os dados são originários dos primeiros anos do curso com 477 alunos, contendo 28 atributos, segundos anos do curso com 396 alunos, contendo 27 atributos e terceiros anos do curso com 385 alunos, contendo 27 atributos, os dados foram coletados entre 2014 e 2020. Para a previsão de desempenho dos alunos foram realizados experimentos utilizando os algoritmos Naive Bayes, J48 (árvore de decisão) e IBk (Knn), todos disponíveis no ambiente WEKA. As métricas utilizadas na avaliação dos algoritmos são a precisão, sensibilidade, medida-F e a área embaixo da curva ROC. O objetivo dos experimentos foi determinar qual dos algoritmos utilizados consegue classificar melhor os dados dos alunos que estão cursando na instituição. Os alunos são classificados segundo eles venham a ser aprovados, aprovados com pendências para o próximo ano, reprovados ou evadidos. Com os algoritmos utilizados nos experimentos, os melhores resultados na classificação foram obtidos com a aplicação da técnica SMOTE e Custos, com relação à precisão dos resultados.

Palavras-chave: *Previsão do desempenho de estudantes, mineração de dados educacionais, árvores de decisão, algoritmo Naive Bayes, Knn.*