USANDO A MINERAÇÃO DE DADOS PARA PREDIÇÃO DE DESEMPENHO DE ALUNOS NAS DISCIPLINAS DE PORTUGUÊS E MATEMÁTICA

OLIVEIRA, Manuel Joaquim Silva[1]
CAETANO, Guedes[2]
DANIEL, Eulanda Maria Pedro[3]

Resumo - Este trabalho relata a aplicação de técnicas de mineração de dados para a predição do desempenho dos alunos nas disciplinas de Português e Matemática. A mineração de dados educacionais produz métodos e técnicas que objetivam a descoberta de padrões que fornecem conhecimentos utilizáveis na predição dos processos de ensino e aprendizagem. O experimento utiliza dados reais de duas escolas portuguesas do ensino médio relativos às variáveis relacionadas com o sucesso escolar que depende, em larga escala, de diferentes fatores associados às caraterísticas demográficas, sociais e relacionadas à escola, as quais foram tomadas como a base do presente estudo.

Palavras chave: Mineração de dados. Desempenho de alunos.

Introdução

O objetivo deste artigo é analisar o desempenho dos alunos do ensino médio nas disciplinas de Português e Matemática em duas escolas portuguesas que são nomeadas: Escola Secundária Gabriel Pereira e Escola Básica e Secundária Mouzinho da Silveira.

Foram coletados dados a respeito dos padrões de comportamento e do desempenho dos alunos, que incluem notas, caraterísticas demográficas, sociais e relacionadas à escola. Dois conjuntos de dados são fornecidos sobre o desempenho em dois assuntos distintos: Matemática e Língua Portuguesa. Assim, a utilização de técnicas de mineração de dados pode ser utilizada visando, dentre outras coisas, predizer o desempenho dos alunos. Sendo que as escolas têm como tarefa principal de transmitir o conhecimento e educação, mas também devem garantir uma boa qualidade de ensino, assim como bons resultados no final de todo um processo.

O uso de técnicas de mineração de dados com foco na educação pode auxiliar as escolas, gestores e professores a contribuir com ações focadas nos alunos com maior dificuldade, de forma

[1] Doutorando em Informática na Educação na Universidade Federal do Rio Grande do Sul, Mestrado em Informática na Educação pela Universidade Pedagógica de Moçambique e Licenciado em Ensino de Matemática e Física. Email: jocasiloliveira79@gmail.com.
[2] Doutor em Informática na Educação pela Universidade Federal do Rio Grande do Sul, Professor Auxiliar na Universidade Pedagógica. Email: guedyscaetano@gmail.com.
[3] Mestranda em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande Sul e Licenciada em Engenharia Informática pela Universidade Pedagógica de Moçambique. Email: eulandaniel@gmail.com.

a criar condições de melhoria de seu desempenho. Neste sentido, este trabalho relata um estudo de caso sobre a aplicação de técnicas de mineração de dados que permitem, em estágios anteriores às avaliações das disciplinas em causa, identificar alunos que têm maior risco de reprovação.

Trabalhos relacionados

A previsão do empenho acadêmico dos alunos é um assunto bastante estudado entre pesquisadores das várias áreas de educação. Detoni, Araujo, & Cechinel (2014) estudaram a possibilidade de prever com antecedência o risco de reprovação de um estudante em um curso a distância utilizando contagens de interações. Eles mostram que redes bayesianas são adequadas ao problema e que as introduções de atributos derivados das contagens (exemplo: médias) são úteis para previsões mais precisas, quando a quantidade de dados é esparsa.

De Brito et al. (2014) propuseram a utilização de técnicas de Mineração de Dados para tentar predizer o desempenho dos alunos no primeiro período do curso de Ciência da Computação da Universidade Federal da Paraíba, através das suas notas de ingresso no vestibular. Os resultados mostraram que é possível inferir o desempenho dos estudantes com uma acurácia superior a 70%, sendo esta informação útil para a realização de ações para evitar a evasão, aprimorando o sistema de ensino.

Laci Mary Barbosa Manhães (2015) apresenta uma proposta de arquitetura baseada em Mineração de Dados Educacionais (EDM) para predição do desempenho acadêmico de graduandos com objetivo de fornecer aos gestores educacionais das universidades públicas brasileiras, não especialista em EDM, uma abordagem que oferece informações úteis sobre o desempenho acadêmico dos graduandos e predizer os que estão em risco de abandonar o sistema de ensino. Os resultados experimentais mostram que a arquitetura proposta é capaz de predizer o desempenho acadêmico dos graduandos a cada semestre letivo com precisão em torno de 80%. Além da predição, também foi possível identificar as principais variáveis que distinguem os estudantes que obtém sucesso ou não na conclusão do curso de graduação.

Da Costa et al. (2014) apresentam uma pesquisa onde se buscou, através da aplicação do processo de descoberta de conhecimento em bases de dados, explicitar padrões de evasão nos cursos de educação permanente em modalidade EAD para profissionais da saúde, promovidos pela Universidade Aberta do SUS (UNA-SUS). Os dados foram analisados aplicando a tarefa de regras de classificação utilizando a técnica de árvores de decisão e como resultado obtiveram, um modelo

preditivo de fácil entendimento com 97,6% de acertos na classificação do conjunto de treinamento.

Rabelo et al. (2017) neste trabalho relatam a aplicação de técnicas de mineração de dados educacionais para a predição do desempenho de alunos de EaD em ambiente virtual de aprendizagem (AVA) utilizando o Moodle como plataforma para realização de cursos de graduação à distância. O experimento utilizou dados reais de uma base histórica contendo treze turmas de cursos de graduação, sendo parte de um estudo que visa melhorar o processo de ensino à distância da Universidade Federal do Rio Grande do Norte (UFRN).

Base de dados

Um passo essencial para o processo de mineração de dados é a seleção dos dados. Informações sobre alunos das escolas são armazenadas em dois tipos de registros: os registros dos dados sobre caraterísticas demográficas, sociais e o histórico escolar dos alunos. As informações contidas em cada um deles estão na Tabela 1. Nos experimentos do registro do aluno foram utilizadas inicialmente as informações dos atributos.

Na tabela estão apresentados o número de ordem do dado, uma descrição curta de seu conteúdo, o seu tipo de dado e o domínio de valores que o dado pode assumir.

TABELA 1 - Descrição dos dados disponíveis.

Ordem	Descrição	Tipo de Dado	Domínio
1	Escola do aluno	Binário	'GP' - Gabriel Pereira ou 'MS' - Mousinho da Silveira
2	Sexo do aluno	Binário	'F' - feminino ou 'M' – masculino
3	Idade do aluno	Numérico	[15, 22]
4	Tipo de endereço residencial	Binário	'U' - urbano ou 'R' – rural
5	Tamanho da família	Binário	'LE3' - menor ou igual a 3 ou 'GT3' - maior que 3
6	Status de coabitação dos pais	Binário	'T' - morando juntos ou 'A' – separados
7	Educação materna	Numérico	0 - nenhuma, 1 - educação primária (4ª série), 2 - 5ª a 9ª série, 3 - educação secundária ou 4 - educação superior
8	Educação do pai	Numérico	0 - nenhum, 1 - ensino primário (4 ª série), 2 - 5 ª a 9 ª série, 3 - ensino secundário ou 4 - ensino superior
9	Trabalho da mãe	Nominal	
10	Trabalho do pai	Nominal	
11	Razão para escolher esta escola	Nominal	próximo de 'casa', escola 'reputação', 'curso' preferência
12	Tutor do aluno	Nominal	'mãe', 'pai' ou 'outro')
13	Tempo de deslocamento da	Numérico	1 - <15 min., 2 - 15 a 30 min., 3 - 30 min. A 1 hora

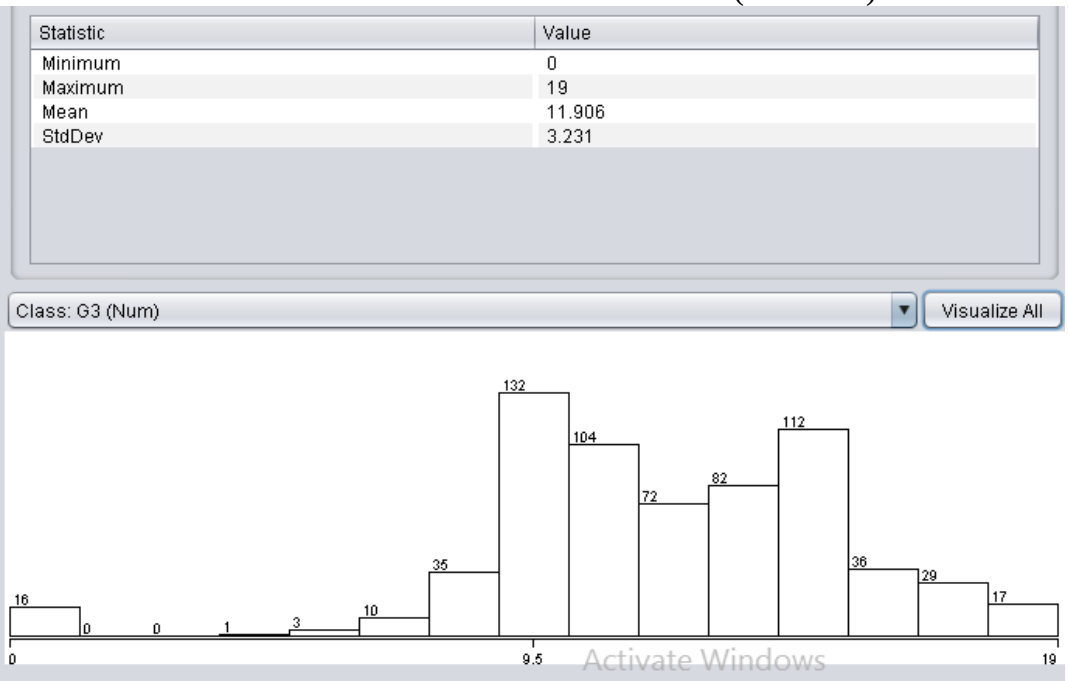
	escola para casa		ou 4 -> 1 hora)
14	Tempo de estudo - tempo de estudo semanal	Numérico	1 - <2 horas, 2 - 2 a 5 horas, 3 - 5 a 10 horas, ou 4 -> 10 horas)
15	Número de falhas de classe anteriores	Numérico	n se 1 <= n <3, se não 4
16	Suporte educacional extra	Binário	sim ou não
17	Suporte educacional familiar	Binário	sim ou não
18	aulas extra pagas dentro do curso (Matemática ou Português)	Binário	sim ou não
19	Atividades extra-curriculares	Binário	sim ou não
20	Creche assistida	Binário	sim ou não
21	Quer ter ensino superior	Binário	sim ou não
22	Acesso à Internet em casa	Binário	sim ou não
23	Com um relacionamento amoroso	Binário	sim ou não
24	Qualidade das relações familiares	Numérico	de 1 - muito ruim a 5 – excelente
25	Tempo livre depois da escola	Numérico	de 1 - muito baixo a 5 - muito alto
26	Sair com os amigos	Numérico	de 1 - muito baixo a 5 - muito alto
27	Consumo de álcool durante a jornada de trabalho	Numérico	de 1 - muito baixo a 5 - muito alto
28	Consumo de álcool no final de semana	Numérico	de 1 - muito baixo a 5 - muito alto
29	Estado de saúde atual	Numérico	de 1 - muito ruim a 5 - muito bom
30	Número de faltas escolares	Numérico	de 0 a 93
31	Nota do primeiro período	Numérico	de 0 a 20
32	Nota do segundo período	Numérico	de 0 a 20
33	Nota final	Numérico	de 0 a 20

Fonte: Autores.

Experimentos e análise dos resultados

O treinamento do modelo foi realizado com um conjunto de 649 instâncias a fim de buscar a identificação dos fatores que contribuem para o desempenho dos estudantes na nota final. Vale salientar que neste estudo os alunos aprovados são todos os que têm nota média, igual ou superior a ($9{,}5 \approx 10$) numa escala de (0 a 20) valores, sendo um total de 584 aprovados correspondente a 89,98 % num universo de 649 alunos, conforme ilustra o gráfico 1, onde a nota mínima foi de 0 valor e máxima de 19 valores, com uma média de 11,906 e um desvio padrão de 3,231.

GRÁFICO 1 – Resultados da variável alvo (nota final)



Fonte: Autores.

Neste estudo foram realizados experimentos de análise de correlação e classificação. Os experimentos e resultados são descritos a seguir:

Análise de correlação

A primeira análise realizada foi a respeito do coeficiente de correlação entre a nota na avaliação final (G3) e as demais variáveis existentes. A medida que mostra o grau de relacionamento entre duas variáveis, é chamada de coeficiente de correlação. É também conhecida como medida de associação, de interdependência, de intercorrelação ou de relação entre as variáveis (LIRA, 2004).

Este experimento visa identificar quais variáveis, dentre as coletadas, tem uma maior influência na nota final (G3). Como resultado, foram obtidos os coeficientes de correlação apresentados na tabela 2. É possível verificar que quase todas as variáveis têm uma correlação pequena ou média com o variável alvo, sendo as únicas variáveis com correlação forte as notas do primeiro período (G1) e do segundo período (G2). Por outro lado, todas as variáveis apresentam uma correlação positiva, com exceção do número de faltas, indicando que o seu crescimento contribui para o crescimento da nota final.

Neste experimento as variáveis do tipo *string* foram eliminadas sendo as numéricas as únicas utilizadas.

TABELA 2 - Análise de Correlação entre a nota final e as demais variáveis

Ordem	Variável	Coeficiente de correlação
1	Idade do aluno	0,1368
2	Educação materna	0,1843
3	Educação do pai	0,1843
4	Tempo de deslocamento da escola para casa	0,0542
5	Tempo de estudo semanal	0,2185
6	Número de falhas de classe anteriores	0,4289
7	Qualidade das relações familiares	0,0409
8	Tempo livre depois da escola	0,0907
9	Sair com os amigos	0,0919
10	Consumo de álcool durante a jornada de trabalho	0,1732
11	Consumo de álcool no final de semana	0,1149
12	Estado de saúde atual	0,0694
13	Número de faltas escolares	- 0,0172
14	Nota do primeiro período	0,8262
15	Nota do segundo período	0,9143

Fonte: Autores.

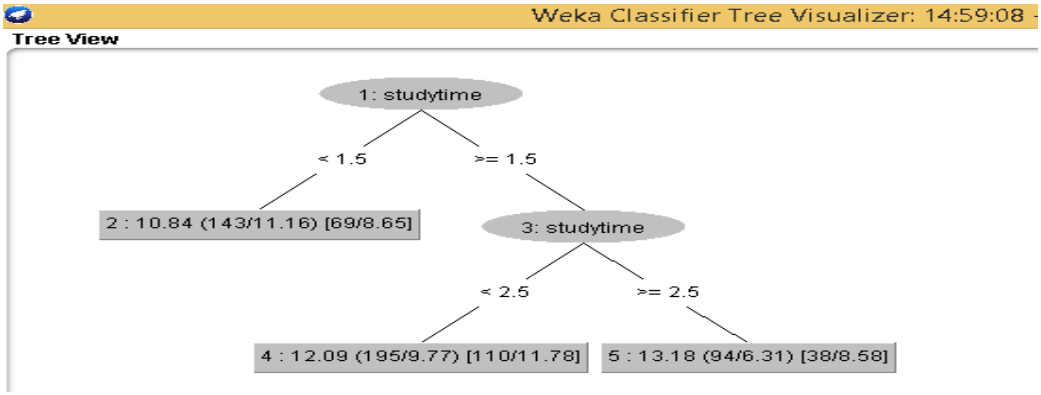
Análise de classificação

A fim de identificar a tendência dos estudantes a terem sucesso na nota final (G3), foram aplicados algoritmos de classificação sobre os dados disponíveis. Estes experimentos foram realizados na tentativa de predizer o resultado final dos alunos aprovados (com a nota superior ou igual a 10 valores e inferior ou igual a 20 valores) ou reprovado (com a nota inferior a 10 valores com aproximação por excesso). Dentre os algoritmos de classificação disponíveis foi selecionado um dos que gera seus resultados na forma de árvores, tendo em vista a fácil interpretação de seus resultados. O algoritmo utilizado foi o REPTree que gerou árvores com poucos nodos e maior capacidade preditiva. Todos os experimentos de classificação realizados utilizaram a técnica de validação cruzada para medir a capacidade de generalização dos modelos. Das variáveis utilizadas foram selecionadas apenas as duas com maior coeficiente de correlação (número de falhas de classe anteriores e tempo de estudo semanal) apresentados na tabela 2, com exceção das notas do primeiro período e segundo período.

Segundo Paulo (*apud* BONINI, 2016), o algoritmo REPtree usa informações de ganho e

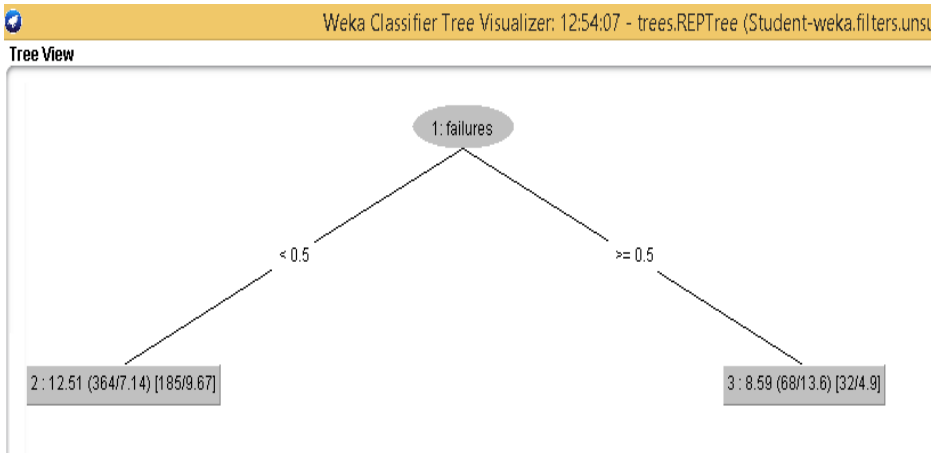
variância para construir a árvore de decisão. Além disso, usa a técnica de poda por redução de erro para podar os ramos da árvore. A poda por redução de erro separa as amostras em dois conjuntos, um de treinamento e outro de validação. Os dados de treinamento são usados para construir a árvore de decisão e os de validação são utilizados para verificar os erros de classificação. Possui também, uma variável que define o número máximo de profundidade da árvore.

Para a variável tempo de estudo a aplicação do algoritmo REPTree gerou a árvore mostrada na figura 1, que indica que se o tempo de estudo for inferior a 1,5h, a média dos alunos é de 10.84 valores, se for entre 1,5h e 2,5h, a média é de 12,9 valores e superior a 2,5h, a média é de 13,18 valores, verificando-se aqui o crescimento da nota com o tempo de estudo.



Fonte: Autores.
Figura 1 - Árvore gerada pelo algoritmo *REPtree* relacionando a variável tempo de estudo.

Para a variável número de falhas de classe anteriores a aplicação do algoritmo *REPTree* gerou a árvore mostrada na figura 2, revelou que os alunos que tiveram um número de falhas superior ou igual 0,5% nas classes anteriores tiveram uma média de 8,59 (reprovados) e os que tiveram um número de falhas inferior a 0,5% obtiveram uma média de 12.51 (aprovados).



Fonte: Autores.
Figura 2 - Árvore gerada pelo algoritmo *REPtree* relacionada a variável número de falhas de classe anteriores.

Considerações Finais

Neste trabalho foi possível notar a relevância da relação das variáveis que possam ou não influenciar na nota final dos alunos, permitindo apoiar aos próprios alunos, professores, instituições de ensino, familiares e sociedade no geral, para a tomada de decisão na melhoria do desempenho destes.

Os experimentos realizados comprovam uma realidade esperada: quanto maior a dedicação nas atividades em sala de aula e em horas de estudo extraclasse, melhor o desempenho do aluno na nota final.

Foi possível verificar que o desempenho dos alunos, em termos de aprovação foi em torno de 89,98%, sendo esta uma percentagem aceitável, no entanto não a desejável, que seria a de 100%. Por essa razão, há que tomar em conta as possíveis razões que podem levar ao fracasso ou insucesso dos alunos.

USING THE MINING OF DATA FOR PREDICTION OF PERFORMANCE OF STUDENTS IN PORTUGUESE DISCIPLINES AND MATHEMATICS

Abstract - THIS paper reports the application of data mining techniques to predict students performance in Portuguese and Mathematics subjects. Educational data mining produces methods and techniques that seek the discovery of patterns that provide usable knowledge in predicting the

teaching and learning processes. The experiment uses real data from two Portuguese high school schools on the variables related to school success that depends, in large scale, on different factors associated with the demographic, social and school-related characteristics, which were taken as the basis of the present study.

Keywords: Data mining. Student performance.

Referências

BONINI, J. A. **Aplicação de algoritmos de árvore de decisão sobre uma base de dados de câncer de mama**. Universidade Federal de Santa Maria, p. 57–67, 2016.

DA COSTA SUSANE SANTOS; CAZELLA SILVIO; RIGO SANDRO JOSÉ. Minerando dados sobre o desempenho de alunos de cursos de educação permanente em modalidade EAD: Um estudo de caso sobre evasão escolar na UNA-SUS. **Revista Novas Tecnologias na Educação**, v. 12, n. 1, p. 1–10, 2014.

DE BRITO, D. M. et al. **Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina**. n. Cbie, p. 882, 2014.

DETONI, D.; ARAUJO, R. M.; CECHINEL, C. **Predição de Reprovação de Alunos de Educação a Distância Utilizando Contagem de Interações.** n. Cbie, p. 896, 2014.

LACI MARY BARBOSA MANHÃES. **Predição do desempenho acadêmico de graduandos utilizando mineração de dados educacionais**, 2015.

LIRA, S. A. **Análise de correlação: abordagem teórica e de construção dos coeficientes com aplicações.** Setores de Ciências Exatas e de, p. 209, 2004.

RABELO, H. et al. **Utilização de técnicas de mineração de dados educacionais para predição de desempenho de alunos de EaD em ambientes virtuais de aprendizagem**. n. Cbie, p. 1527, 2017.

Recebido em 29/10/2018.
Aprovado em 21/12/2018.