UM MODELO PREDITIVO PARA ESTUDO DA EVASÃO NA GRADUAÇÃO UTILIZANDO MINERAÇÃO DE DADOS

João Paulo Funchal[1]; Alex Sandro de Paula Rodrigues[2]; Eduardo Nunes Borges[3].

Resumo
A evasão estudantil é um problema recorrente nos mais diversos níveis educacionais do Brasil, em especial, no ensino superior. Além do abandono por parte dos alunos, os investimentos realizados por órgãos federais são prejudicados, visto que, o retorno deste investimento não é retornado. Buscando entender este cenário o presente trabalho, o presente trabalho analisou os dados de alunos do curso d Engenharia de Computação da Universidade Federal do Rio Grande (FURG), buscando compreender as prováveis motivações para a ocorrência de evasão dos discentes no curso. Para efetivar tal tarefa foi empregada a descoberta de conhecimento em bases de dados com objetivo de descobrir padrões implícitos no conjunto de dados, transformando os dados sem significado em informação e conhecimento útil. Por meio desta análise, pode-se constatar que alunos que possuem bom desempenho a disciplinas ligadas a ciências humanas apresentam uma tendência maior a deixar o curso.
Palavras-chave: Mineração de dados, Evasão escolar.

Abstract
Student evasion is a recurrent problem in the most diverse educational levels in Brazil, especially in higher education. In addition to student abandonment, investments made by federal agencies are impaired, since the return of this investment is not returned. The present work analyzed the data of students of the Computer Engineering course of the Federal University of Rio Grande (FURG), seeking to understand the probable motivations for the occurrence of students' avoidance in the course. To accomplish this task, the discovery of knowledge in databases was used to discover patterns implicit in the data set, transforming the data without meaning into information and useful knowledge. Through this analysis, it can be seen that students who perform well in subjects related to the humanities have a greater tendency to leave the course**.**
Keywords: Data Mining, School Evasion.

Introdução

De acordo com Silva Filho et al., 2007, a evasão estudantil é um dos principais problemas que atingem as instituições de ensino em geral. Este problema resulta em perdas sociais, acadêmicas e econômicas, visto que, recursos são desperdiçados e retornos não são obtidos. Ainda conforme o autor, tanto instituições privadas, como publicas apontam como principal motivo para evasão a falta de auxílios financeiros para o estudante manter-se em atividade.

[1] Mestrando em Engenharia de Computação pela Universidade Federal do Rio Grande - FURG; e-mail: joaofunchal@furg.br.
[2] Mestrando em Engenharia de Computação pela Universidade Federal do Rio Grande - FURG; e-mail: alexrodrigues@furg.br.
[3] Prof. Dr. do Centro de Ciências Computacionais - C3 pela Universidade Federal do Rio Grande - FURG; e-mail: eduardoborges@furg.br.

No entanto, outros fatores devem ser considerados, tais como a expectativa do aluno com relação a sua formação, a integração do estudante em relação a instituição, bem como, as diferentes características sociais, econômicas e culturais que apresentam os discentes.

Buscando compreender este contexto, o presente artigo tem por intuito constatar as prováveis causas para a ocorrência de evasão de discentes no curso de Engenharia de Computação, por meio, do histórico de desempenho acadêmico, bem como, por intermédio das notas obtidas no processo seletivo para o ingresso na graduação. Por fim espera-se entender se ações, como bolsas de pesquisa, ensino e extensão possuem relação com a permanência do estudante no curso escolhido.

1 Materiais e Métodos

Esta seção apresenta os conceitos necessários para entendimento da pesquisa realizada, a base de dados foco do estudo realizado e a metodologia empregada.

1.1 Descoberta de Conhecimento em Base de Dados

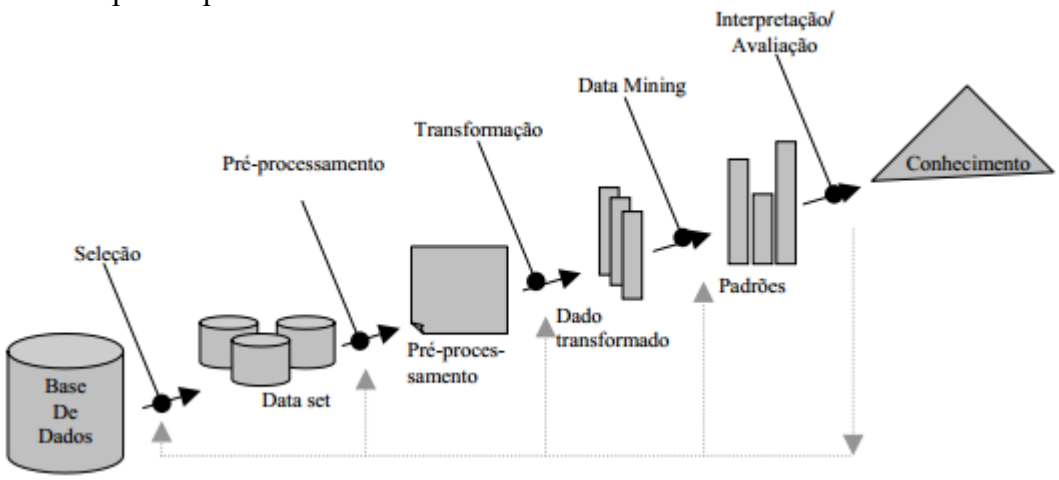
Segundo Tan et al., 2009, a Descoberta de Conhecimento em Bases de Dados ou Knowledge Discovery in Databases (KDD) é uma metodologia que combina métodos tradicionais de análise de dados com algoritmos sofisticados para processar grandes volumes de dados. Esta metodologia consiste em uma série de etapas de transformação, pré-processamento dos dados, até chegar à fase de pós-processamento e interpretação dos resultados. Os dados usados neste processo podem estar armazenados em diversos formatos e serem originários de uma fonte de dados central ou até mesmo distribuída (Tan et al., 2009). Na fig.1 estão representadas todas as fases do processo de descoberta de conhecimento. As etapas do KDD serão resumidas a seguir (TAN et al., 2009; WITTEN; FRANK, 2005; FAYYAD et al., 1996):

- **Pré-processamento** é a transformação dos dados de entrada em um formato apropriado para a mineração. Resumidamente, o pré-processamento abrange as etapas de integração dos dados de múltiplas fontes, a limpeza dos dados para remoção de ruídos e dados duplicados e a seleção de registros e características que serão relevantes nas etapas seguintes.
- **Mineração de dados** consiste em investigar grandes quantidades de dados com o objetivo de encontrar padrões previamente desconhecidos, úteis e relacionados entre os dados. Um conjunto de fatores motivou o desenvolvimento de técnicas para a mineração de dados,

principalmente a necessidade diária de lidar com bancos de dados, com centenas ou milhares de atributos, sendo necessárias técnicas para o tratamento dessa alta dimensionalidade. Atualmente, há diferentes tarefas de mineração de dados, tais como classificação, regressão, agrupamento e associação.

• **Pós-processamento** é a etapa que avalia os resultados da mineração de dados. Podem ser utilizados métodos estatísticos para desconsiderar resultados não legítimos. Geralmente, essa etapa é realizada com o apoio de especialistas que avaliam a qualidade e conhecimento obtido a partir dos modelos de mineração de dados.

Figura 1 - Etapas do processo de descoberta de conhecimento em base de dados



Fonte: adaptado de Fayyad et al. (1996).

A mineração de dados é a principal etapa do processo de KDD, sendo separada em dois tipos de tarefas de aprendizado, as preditivas e descritivas. Nas tarefas preditivas, o objetivo é encontrar funções que a partir de conjuntos de dados possam ser utilizadas para prever o valor ou classe de novos entrantes, com base nos valores de seus atributos. Para isto, os objetos do grupo de treinamento devem ter os mesmos atributos de entrada e saída. Neste tipo de tarefa, é utilizado o aprendizado supervisionado, no qual, se tem um "supervisor externo", que sabe a saída apropriada para cada objeto (FACELI, 2011).

Enquanto nas tarefas descritivas, o intuito é investigar ou detalhar um conjunto de dados. Nestes algoritmos a classe de cada objeto de treinamento é desconhecida, assim como, o número ou conjunto de classes que compõem a base de dados. Por causa disso, tem-se uma aprendizagem não supervisionada (FACELI, 2011; WITTEN; FRANK, 2005).

Na seção subsequente do presente trabalho será abordado o *dataset* usado para obtenção dos resultados.

1.2 Conjuntos de dados analisado

A base de dados analisada foi extraída do sistema acadêmico da Universidade Federal do Rio Grande e contém dados a respeito de todos os estudantes do curso Engenharia de Computação, período compreendido entre de 1994 e 2015.

Com relação ao esquema relacional fornecido pelo Núcleo de Tecnologia da Informação (NTI) da universidade onde foram realizadas uma série de consultas, por meio, da linguagem SQL (*Structured Query Language*), com intuito, de realizar operações de pré-processamento a fim de preparar os dados para a mineração. Destacam-se as operações para preenchimento de dados faltantes, remoção de instâncias inadequadas ou dados mal cadastrados (ruído), integração de atributos em uma única tabela, alteração de tipos de dados, entre outras.

1.3 Ferramentas e tecnologias

Foram utilizados para armazenamento e pré-processamento dos dados o Sistema Gerenciador de Banco de Dados (SGBD) PostgreSQL[4], bem como, o software Weka[5] na etapa de mineração de dados. Por fim, alguns *scripts* foram codificados por intermédio da linguagem PHP (*Hypertext Preprocessor*).

2 Resultados

Os experimentos realizados no presente artigo tiveram como objetivo identificar possíveis motivações para casos de evasão de estudantes no curso de Engenharia de Computação a partir do histórico do desempenho acadêmico e das notas obtidas no processo seletivo para o ingresso na graduação. Foram selecionados apenas alunos que ingressaram pelo antigo processo seletivo vestibular. Para que o propósito do trabalho fosse alcançado a base de dados teve que passar por algumas modificações, onde se buscou relacionar os alunos com as disciplinas cursadas, afim de se obter o número de disciplinas aprovadas e reprovadas pelo acadêmico. Após essa etapa, a informação obtida foi combinada com as notas das provas do vestibular, totalizando assim, 18.643 registros e gerando uma tabela (tab. 1) com os seguintes atributos.

[4] https://www.postgresql.org/
[5] http://www.cs.waikato.ac.nz/ml/weka/

Tabela 1 - Atributos escolhidos para utilizar na classificação

Atributo	Descrição
reprovacoes	Número de reprovações do aluno
aprovacoes	Número de aprovações do aluno
media	Média para classificação no vestibular
portugues	Nota do vestibular na disciplina Português
literatura	Nota do vestibular na disciplina Literatura
fisica	Nota do vestibular na disciplina Física
biologia	Nota do vestibular na disciplina Biologia
matematica	Nota do vestibular na disciplina Matemática
quimica	Nota do vestibular na disciplina Química
situacao	Situação atual do aluno na universidade

Fonte: Elaborada pelos autores.

O atributo *situacao* pode conter um dos seguintes rótulos:

- desligado por abandono (evadido);
- desligado por transferência (evadido);
- desligado por mudança de curso (evadido);
- desligado a pedido (evadido);
- afastado temporariamente;
- formado;
- em mobilidade acadêmica;
- matriculado;
- em garantia de vaga.

O objetivo específico do trabalho é identificar um destes possíveis valores, como forma de predizer a situação de um estudante a partir das notas do vestibular e número de disciplinas da Engenharia de Computação em que aprovou e reprovou.

O primeiro experimento utilizou a técnica de classificação baseada em árvores de decisão com o algoritmo C4.5 (QUINLAN, 1993). A escolha por esse tipo de algoritmo deve-se ao fato de produzir resultados de fácil compreensão, assim como, o melhor desempenho apresentado quando comparado com outros algoritmos testados, por exemplo: ZeroR e Radom Forest.

O algoritmo J48 foi configurado da seguinte forma: número mínimo de instâncias classificadas por folha igual a 100. O método de teste escolhido foi a validação cruzada com 10 partições. Os resultados podem ser vistos na fig.2.

Figura 2 - Avaliação da classificação utilizando o algoritmo C4.5

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       17513               93.9387 %
Incorrectly Classified Instances      1130                6.0613 %
Kappa statistic                          0.8528
Mean absolute error                      0.0214
Root mean squared error                  0.1032
Relative absolute error                 22.935  %
Root relative squared error             47.8218 %
Total Number of Instances            18643

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.91      0.033      0.83      0.91      0.868      0.977    Desligado por abandono
                 0.998     0.036      0.988     0.998     0.993      0.988    Formado
                 0.397     0.01       0.474     0.397     0.432      0.968    Desligado por Transferência
                 0         0          0         0         0          0.972    Desligado mudança de curso
                 0.808     0.003      0.48      0.808     0.602      0.997    Mobilidade Acadêmica CSF
                 0.361     0.005      0.551     0.361     0.436      0.966    Desligado a Pedido
                 0.8       0.006      0.882     0.8       0.839      0.994    Matriculado
                 0         0          0         0         0          0.619    Garantia de Vaga
                 0         0          0         0         0          0.968    Afastado Temporariamente
Weighted Avg.    0.939     0.032      0.928     0.939     0.933      0.985
```

Fonte: Elaborada pelos autores.

A medida de avaliação geral ou acurácia = 93.9% representa a porcentagem de instâncias corretamente identificadas. Também estão destacadas as medidas para cada classe de dados:

- **Precisão** (*Precision*) que mede a porcentagem de acerto entre as instâncias preditas com um determinado rótulo;
- **Revocação** (*Recall*) que mede a porcentagem de acerto entre as instâncias de um determinado rótulo;
- **Medida F** (*F-measure*) é a média harmônica entre a precisão e a revocação.

Nota-se que para as classes Desligado por abandono, Formado e Matriculado, que juntas são a grande maioria dos alunos, o algoritmo escolhido atingiu medida F igual a 86,8 , 99,3 e 83,9% respectivamente.

Na fig.3 é apresentada a matriz de confusão. Para cada classe (linhas), é apresentado o número de instâncias preditas pelo algoritmo (coluna). A diagonal principal mostra os registros classificados corretamente. Os registros classificados que não se encontram na diagonal principal foram classificados de forma errada. Por exemplo, na coluna *a*, 16 estudantes regularmente matriculados foram confundidos com os desligados por abandono.

Figura 3 - Matriz de confusão da classificação utilizando o algoritmo C4.5

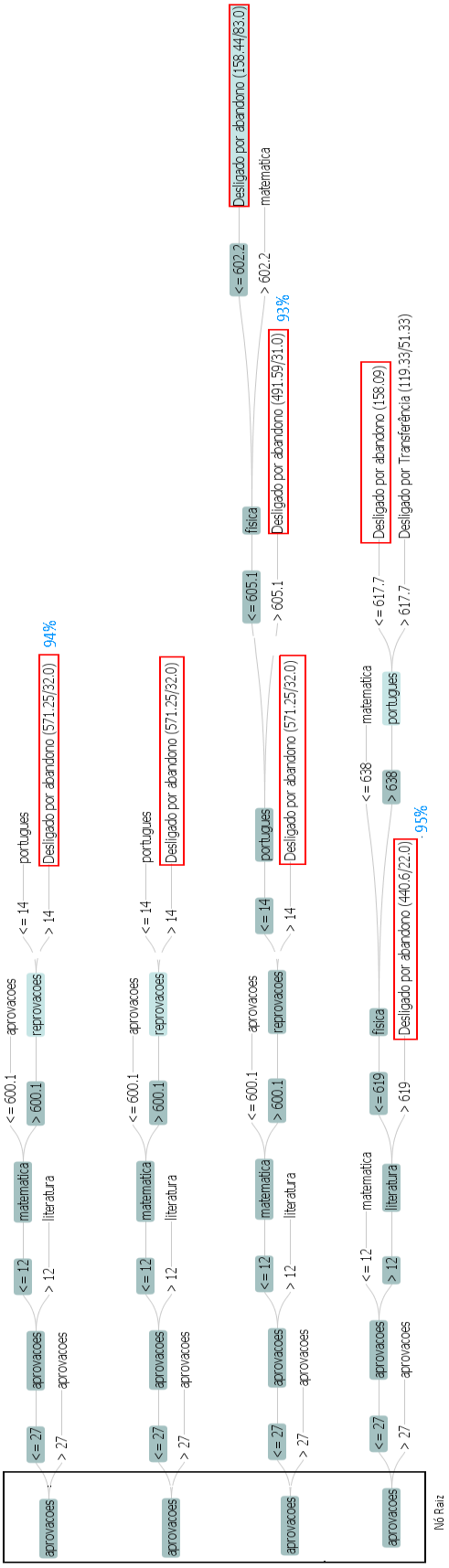
```
=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
  2550    47    72     0     5    57    72     0     0 |     a = Desligado por abandono
     0 13857    27     0     0     0     0     0     0 |     b = Formado
   229     0   167     0    12     1    12     0     0 |     c = Desligado por Transferência
    81     0    27     0    22     0     4     0     0 |     d = Desligado mudança de curso
     0     0     0     0    59     0    14     0     0 |     e = Mobilidade Acadêmica CSF
   186     0     4     0     0   108     1     0     0 |     f = Desligado a Pedido
    16    67    55     0    25    30   772     0     0 |     g = Matriculado
     0    55     0     0     0     0     0     0     0 |     h = Garantia de Vaga
     9     0     0     0     0     0     0     0     0 |     i = Afastado Temporariamente
```

Fonte: Elaborada pelos autores.

O modelo de classificação baseado em árvore gerado pelo algoritmo J48 é apresentado na fig.4. Estão destacados em vermelho os nós folhas, que representam registros classificados como Desligado por Abandono. Enquanto os classificados corretamente estão destacados em azul.

As regras de classificação são bem distintas, variando muito o número de aprovações e reprovações e as notas das provas do vestibular. Entretanto, destaca-se a regra que classifica estudantes bons em literatura, mesmo que já tenham cursado e aprovados em várias disciplinas acabam por abandonar o curso em 95% das vezes. Outro ponto a ser salientado, é o fato do aluno que contém um número de reprovações maior que o de aprovações, tender a deixar o curso.

Figura 4 - Árvore de decisão gerada pelo C4.5



Fonte: Elaborada pelos autores.

Por fim, é importante ressaltar que a fig.4 necessitou de ajustes para apresentação neste trabalho, mas em nenhum momento suas regras foram alteradas, apenas em alguns casos ficaram ocultas, para uma melhor apresentação da árvore e suas principais regras.

Visando potencializar os resultados, num segundo experimento foi utilizada outra técnica de classificação conhecida por *Bagging*. Este algoritmo possibilita criar múltiplas amostras do conjunto de dados, treinar um classificador para cada amostra que é combinada para dar a predição final.

O algoritmo *Bagging* foi configurado utilizando 10 iterações do algoritmo J48Graft (WEBB, 1999), este uma derivação do algoritmo C4.5. Também foi utilizada validação cruzada com 10 partições. Como se pode observar nas figs.5,6, foi obtido um pequeno ganho com a escolha desses algoritmos.

Figura 5 - Avaliação da classificação utilizando o algoritmo Bagging

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       17712               95.0062 %
Incorrectly Classified Instances       931                4.9938 %
Kappa statistic                          0.8785
Mean absolute error                      0.0203
Root mean squared error                  0.0927
Relative absolute error                 21.8238 %
Root relative squared error             42.9568 %
Total Number of Instances            18643

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.955     0.037      0.822     0.955      0.884      0.989    Desligado por abandono
                 0.999     0.029      0.99      0.999      0.995      0.998    Formado
                 0.409     0.003      0.778     0.409      0.536      0.99     Desligado por Transferência
                 0.037     0          1         0.037      0.072      0.984    Desligado mudança de curso
                 0.233     0          1         0.233      0.378      1        Mobilidade Acadêmica CSF
                 0.324     0.001      0.808     0.324      0.463      0.987    Desligado a Pedido
                 0.909     0.008      0.859     0.909      0.883      0.998    Matriculado
                 0         0          0         0          0          0.936    Garantia de Vaga
                 0         0          0         0          0          0.997    Afastado Temporariamente
Weighted Avg.    0.95      0.027      0.947     0.95       0.941      0.996
```

Fonte: Elaborada pelos autores.

Figura 6 - Matriz de confusão da classificação utilizando o algoritmo Bagging

```
=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
  2677    47     7     0     0     5    67     0     0 |     a = Desligado por abandono
     4 13867    13     0     0     0     0     0     0 |     b = Formado
   247     0   172     0     0     2     0     0     0 |     c = Desligado por Transferência
    86     0    22     5     0     0    21     0     0 |     d = Desligado mudança de curso
     0     0     0     0    17     0    56     0     0 |     e = Mobilidade Acadêmica CSF
   202     0     0     0     0    97     0     0     0 |     f = Desligado a Pedido
    31    34     7     0     0    16   877     0     0 |     g = Matriculado
     0    55     0     0     0     0     0     0     0 |     h = Garantia de Vaga
     9     0     0     0     0     0     0     0     0 |     i = Afastado Temporariamente
```

Fonte: Elaborada pelos autores.

Nas regras encontradas destacam-se alunos que possuem médias altas nas áreas das ciências e humanas e tendem a abandonar o curso. A fig.7 apresenta algumas das regras geradas neste experimento.

Figura 7 - Algumas regras de classificação utilizando o algoritmo Bagging.

```
Size of the tree :  279


J48graft pruned tree
------------------

aprovacoes <= 27
|   aprovacoes <= 12
|   |   reprovacoes <= 17
|   |   |   fisica <= 549.2
|   |   |   |   aprovacoes <= 7
|   |   |   |   |   matematica <= 521: Desligado por abandono (119.28/38.0)
|   |   |   |   |   matematica > 521
|   |   |   |   |   |   media_podre <= 251.05: Desligado por abandono (0.0|52.0)
|   |   |   |   |   |   media_podre > 251.05
|   |   |   |   |   |   |   quimica <= 405.4: Desligado por abandono (0.0|23.0)
|   |   |   |   |   |   |   quimica > 405.4
|   |   |   |   |   |   |   |   literatura <= 731.6
|   |   |   |   |   |   |   |   |   portugues <= 729
|   |   |   |   |   |   |   |   |   |   matematica <= 809.5
|   |   |   |   |   |   |   |   |   |   |   portugues <= 443.5: Desligado a Pedido (0.0|17.0)
|   |   |   |   |   |   |   |   |   |   |   portugues > 443.5
|   |   |   |   |   |   |   |   |   |   |   |   biologia <= 688.2
|   |   |   |   |   |   |   |   |   |   |   |   |   quimica <= 680.25: Desligado a Pedido (101.0/30.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   quimica > 680.25: Desligado por abandono (0.0|171.0/21.0) 87.7 %
|   |   |   |   |   |   |   |   |   |   |   |   biologia > 688.2: Desligado por abandono (0.0|139.0/13.0) 90.6%
|   |   |   |   |   |   |   |   |   |   |   matematica > 809.5: Desligado por abandono (0.0|17.0)
|   |   |   |   |   |   |   |   |   portugues > 729: Desligado por abandono (0.0|18.0)
|   |   |   |   |   |   |   |   literatura > 731.6: Desligado por abandono (0.0|19.0)
|   |   |   |   aprovacoes > 7
|   |   |   |   |   media_podre <= 344.2: Desligado por abandono (0.0|120.0)
|   |   |   |   |   media_podre > 344.2
|   |   |   |   |   |   biologia <= 680.3
|   |   |   |   |   |   |   portugues <= 702
|   |   |   |   |   |   |   |   quimica <= 713
|   |   |   |   |   |   |   |   |   matematica <= 429.45: Desligado por Transferência (0.0|36.0)
|   |   |   |   |   |   |   |   |   matematica > 429.45
```

Fonte: Elaborada pelos autores.

3 Conclusão

O problema da evasão estudantil não é recente e há anos vem sendo estudado e debatido nas Universidades, de posse deste tema, o presente artigo buscou encontrar as possíveis causas que poderiam levar a este acontecimento.

Nos resultados obtidos dos estudos de caso 1 e 2, podemos identificar o perfil do aluno que abandona o curso, no qual o desempenho do discente, associado as notas obtidas no vestibular, tal como, o número de aprovações e reprovações, pode ser determinante para sua continuidade na graduação.

Nas regras apresentadas na árvore de decisão da fig.4, é observado algumas informações interessantes, como: o perfil dos alunos que abandonam o curso terem relativamente notas altas, em disciplinas das ciências humanas, desta forma, pode-se predizer, que o perfil deste aluno não pertence a área de exatas e a possibilidade de ter escolhido o curso tenha sido de maneira equivocada.

Por fim, com a identificação destes perfis, torna-se mais fácil criar estratégias para manter o aluno ou inclusive que o mesmo possa seguir em outros cursos na própria universidade, portanto cabe aos responsáveis analisarem qual a melhor decisão para este cenário.

Referências

FACELI, K. **Inteligência Artificial: Uma abordagem de aprendizado de máquina.** GrupoGen-LTC, 2011.

FAYYAD, U. M., PIATETSKY-SHAPIRO, G., SMYTH, P., and Uthurusamy, R., editors. **Advancesin Knowledge Discovery and Data Mining**. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.

QUINLAN, R. **C4.5**: **Programs for Machine Learning. Morgan Kaufmann Publishers**, SanMateo, CA, 1996.

SILVA FILHO, R. L. L. et al. **A evasão no ensino superior brasileiro**. Cadernos de pesquisa, 37(132):641–659, 2007.

TAN, P.-N., STEINBACH, M., KUMAR, V. **Introdução ao data mining: mineração de dados**. Ciencia Moderna, 2009.

WEBB, G. I. **Decision tree grafting from the all-tests-but-one partition**. InProceedings ofthe 16th International Joint Conference on Artificial Intelligence - Volume 2, pages 702–707, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc, 1999.

WITTEN, I. H., FRANK, E. **Data Mining: Practical machine learning tools and techniques**. Morgan Kaufmann, 2005.