



comentários **favorito (1)** **marcar como lido** **para impressão** **anotar**

SQL Magazine 135 - Índice

Mineração de dados na Prática – Parte 2

Veja nesse artigo um exemplo do uso da mineração de dados para análise do problema de evasão em cursos universitários. Será apresentada a aplicação de duas técnicas de mineração: agrupamento e árvore de decisão.



Gostei (0) (0)

Demais posts desta série:

[Mineração de dados na Prática – Parte 1](#)

Artigo no estilo **Curso**

Fique por dentro

A mineração de dados apoia a descoberta de informações úteis que normalmente estão ocultas em bases de dados com grande quantidade de registros. Neste artigo apresentaremos dois casos práticos do uso de técnicas de mineração para análise do problema de evasão em cursos universitários utilizando duas técnicas distintas: agrupamento e árvore de decisão. Esta discussão é útil pois mostra na prática como problemas reais podem ser tratados com o uso de técnicas de mineração.

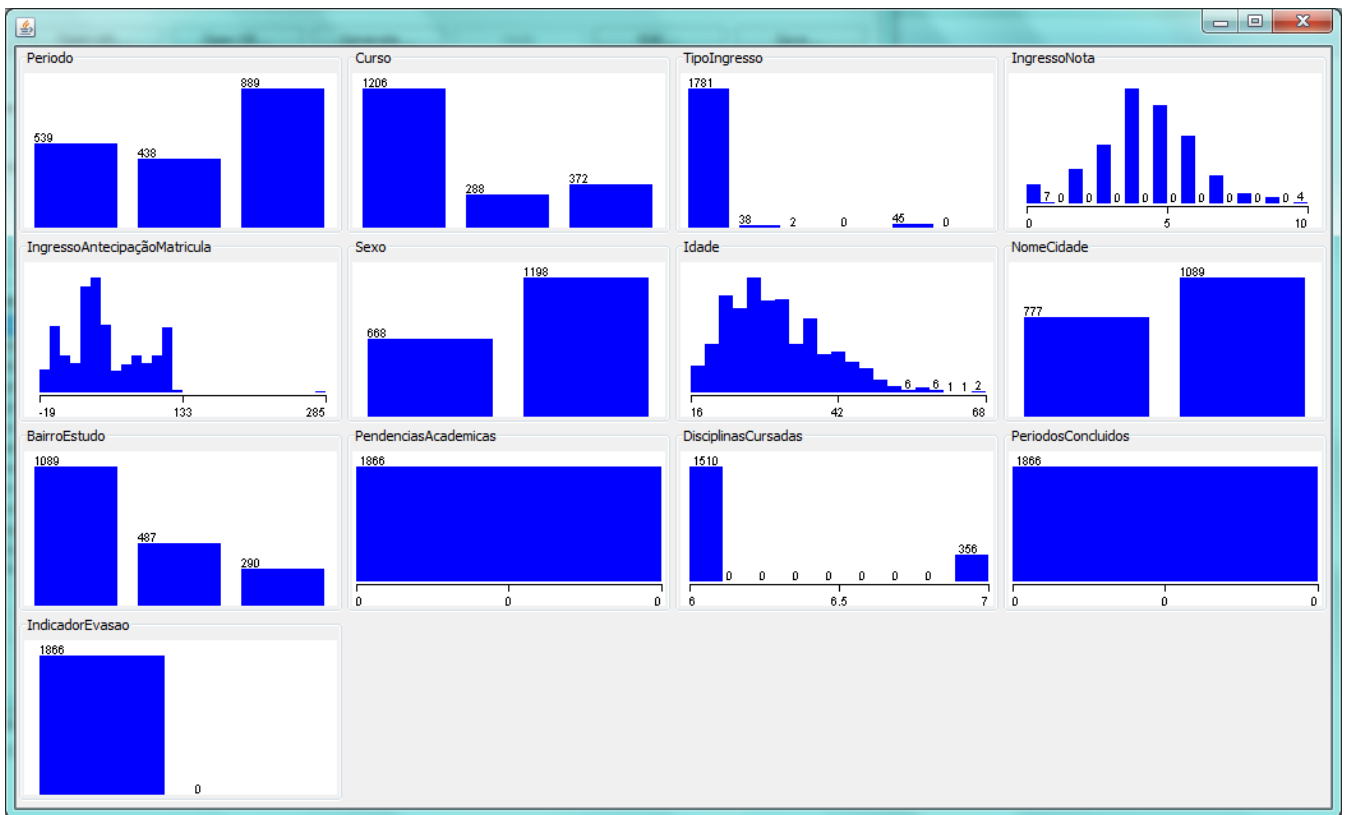
Autores: Péricles Magalhães e Rodrigo Oliveira Spínola

Neste artigo, os arquivos gerados para a mineração de dados (apresentados na primeira parte) serão utilizados em estudos de caso em uma aplicação de algoritmo de clustering e em uma aplicação de algoritmo de classificação.

Caso 1 – Aplicação de Algoritmo de Clustering

O agrupamento ou clustering identifica similaridades entre os valores dos atributos analisados e, a partir dessa análise, particiona a base de dados em grupos. Para a execução da técnica, no estudo de caso, foi selecionado o algoritmo SimpleKMeans que, a partir da indicação da quantidade (k) de clusters desejada, divide a base de dados de forma que a similaridade dos elementos de cada cluster seja alta e, entre os clusters seja baixa.

O arquivo de entrada de dados gerado para essa aplicação, descrito no artigo anterior, foi carregado no WEKA onde algumas análises e considerações foram realizadas sobre a distribuição dos valores dos atributos e seu impacto na atividade. A **Figura 1** apresenta as distribuições dos valores de cada atributo da base de dados carregada. Como pode ser observado, os atributos PendenciasAcademicas, PeriodosConcluidos e IndicadorEvasao apresentam apenas um valor, cada, em toda a base utilizada. Dessa forma, não possuem nenhuma interferência na criação dos agrupamentos.



[abrir imagem em nova janela](#)

Figura 1. Representação gráfica da distribuição dos valores do arquivo de entrada para o caso 1

O algoritmo simpleKmeans apresenta algumas variáveis de configuração para a sua execução:

- displayStdDevs: indica a exibição de desvios padrão dos atributos numéricos e contagens de atributos nominais. Seu valor padrão é false;
- distanceFunction: determina a função de distância a ser usada para comparação das instâncias. Como padrão, é utilizada a weka.core.EuclideanDistance;
- dontReplaceMissingValues: indica se os valores faltantes devem ser substituídos pela média ou moda. A indicação padrão para esse parâmetro é false, permitindo a substituição dos valores ausentes;
- fastDistanceCalc: indica a utilização de “pontos de corte” para acelerar o cálculo da distância. Possui valor inicial false;
- initializeUsingKMeansPlusPlusMethod: determina a detecção dos centros dos clusters através do método probabilístico k-means++. O valor padrão para o parâmetro é false;

- `maxIterations`: determina o número máximo de iterações. Sugere-se como valor padrão 500 iterações;
- `numClusters`: determina o número de clusters a ser gerado. A indicação inicial aponta a geração de apenas dois agrupamentos;
- `preserveInstancesOrder`: indica se a ordem original das instâncias deve ser preservada. Por padrão, o valor `false` indica que a ordem das instâncias pode ser modificada;
- `seed`: referência para a utilização na geração de valores aleatórios.

Desses parâmetros, alteramos somente a indicação da quantidade de clusters a serem gerados (`numClusters`). Para determinar o melhor valor para o indicador, foram realizados alguns testes, alternando valores e observando resultados, sobretudo a dispersão dos clusters gerados. A utilização de dois clusters como sugere o valor padrão do Weka resulta numa distribuição dos registros na ordem de 42% e 58%. Por outro lado, ao utilizarmos 10 clusters obtemos valores percentuais da distribuição entre 4% e 18%, com uma variação entre 6 e 8 pontos percentuais do valor médio. A utilização de 15 clusters foi selecionada por evidenciar 3 grupos com o dobro do percentual médio da base, conforme as **Figuras 2 e 3**, que apresentam o resultado gerado pelo Weka a partir da aplicação do algoritmo `simpleKMeans` na base de dados do Caso 1.

Number of iterations: 35
 Within cluster sum of squared errors: 857.4328575259917
 Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (1866)	Cluster#							
		0 (268)	1 (88)	2 (116)	3 (74)	4 (115)	5 (147)	6 (226)	
Curso	BACHARELADO	BACHARELADO	BACHARELADO	BACHARELADO	BACHARELADO	BACHARELADO	BACHARELADO	LICENCIATURA	BACHARELADO
TipoIngresso	VESTIBULAR	VESTIBULAR	VESTIBULAR	VESTIBULAR	VESTIBULAR	VESTIBULAR	VESTIBULAR	VESTIBULAR	VESTIBULAR
IngressoNota	4.418	4.5746	4.6818	5.1552	4.2838	4.0957	4.9184	5.2301	5.2301
IngressoAntecipaçãoMatricula	49.9496	50.5485	28	104.9741	103	61.487	51.6871	43.4425	43.4425
Sexo	2	1	1	2	1	2	2	2	2
Idade	31.0531	30.306	24.6023	25.4138	27.0946	40.9043	31.1293	27.3451	27.3451
NomeCidade	INTERIOR	CAPITAL	INTERIOR	INTERIOR	INTERIOR	INTERIOR	INTERIOR	INTERIOR	CAPITAL
BairroEstudo	CENTRO	IGUATEMI	CENTRO	CENTRO	CENTRO	CENTRO	CENTRO	CENTRO	IGUATEMI

Time taken to build model (full training data) : 0.23 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	268 (14%)
1	88 (5%)
2	116 (6%)
3	74 (4%)
4	115 (6%)
5	147 (8%)
6	226 (12%)
7	112 (6%)
8	80 (4%)
9	161 (9%)
10	63 (3%)
11	59 (3%)
12	230 (12%)
13	48 (3%)
14	79 (4%)

[abrir imagem em nova janela](#)

Figura 2. Resultado da aplicação do algoritmo de agrupamento na base de dados do Caso 1

	7 (112)	8 (80)	9 (161)	10 (63)	11 (59)	12 (230)	13 (48)	14 (79)
BACHARELADO	BACHARELADO	TECNOLOGICO	BACHARELADO	LICENCIATURA	BACHARELADO	BACHARELADO	BACHARELADO	BACHARELADO
VESTIBULAR	VESTIBULAR	VESTIBULAR	VESTIBULAR	VESTIBULAR	VESTIBULAR	VESTIBULAR	VESTIBULAR	VESTIBULAR
3.3214	3.225	5.0807	1.9524	1.661	5.1304	6.5417	1.9241	1.9241
57.7054	47.0125	40.7453	30.4762	39.3898	27.2174	57.6875	36.1013	36.1013
2	1	2	1	2	2	1	2	2
40.125	42.4125	38.4037	25.8889	32.4237	25.5478	33.6667	27.0759	27.0759
CAPITAL	INTERIOR	CAPITAL	INTERIOR	INTERIOR	INTERIOR	INTERIOR	INTERIOR	INTERIOR
IGUATEMI	CENTRO	IGUATEMI	CENTRO	CENTRO	CENTRO	CENTRO	CENTRO	CENTRO

Figura 3. Resultado da aplicação do algoritmo de Agrupamento na base de dados do Caso 1 – continuação

O cluster 0, que abrange 14% dos dados, aponta evadidos de cursos de bacharelado da capital do sexo feminino com média de idade de 30 anos que ingressaram por vestibular com nota média de 4,5 e anteciparam sua matrícula em cerca de dois meses. Já o cluster 6, com 12% de ocorrências, apresenta uma pequena variação em relação ao primeiro - evadidos de cursos de bacharelado da

capital do sexo masculino com média de idade de 27 anos que ingressaram por vestibular com nota média de 5,2 e anteciparam sua matrícula em cerca de um mês e meio.

Chama a atenção também o cluster 12. Este também possui 12% de incidência, que aponta para evadidos do interior do Estado do sexo masculino e média de idade de 25 anos, oriundos de cursos de bacharelado que ingressaram por vestibular com nota média de 5,1 e antecipação de cerca de 1 mês.

A separação dos clusters pela aplicação aponta para a predominância, entre os evadidos, de estudantes de cursos de bacharelado, que ingressaram por vestibular com notas relativamente baixas (entre 4 e 5). Predominam nos grupos, também, estudantes do sexo masculino, do interior do estado e com idade média de 20 anos.

Caso 2 – Aplicação de Algoritmo de Classificação

A geração de árvores de decisão através de algoritmos de classificação na base de dados do Caso 2 permite identificar, hierarquicamente, os atributos que mais contribuem para a evasão, assim como quais as relações entre atributos e seus valores têm maior possibilidade de determinar a permanência dos estudantes. Foi utilizado para a técnica o algoritmo J48, que procura formar a árvore mais adequada sobre o conjunto de dados através da poda de regras, mantendo as que melhoram a sua eficiência.

A configuração da execução do algoritmo no Weka utiliza os seguintes parâmetros:

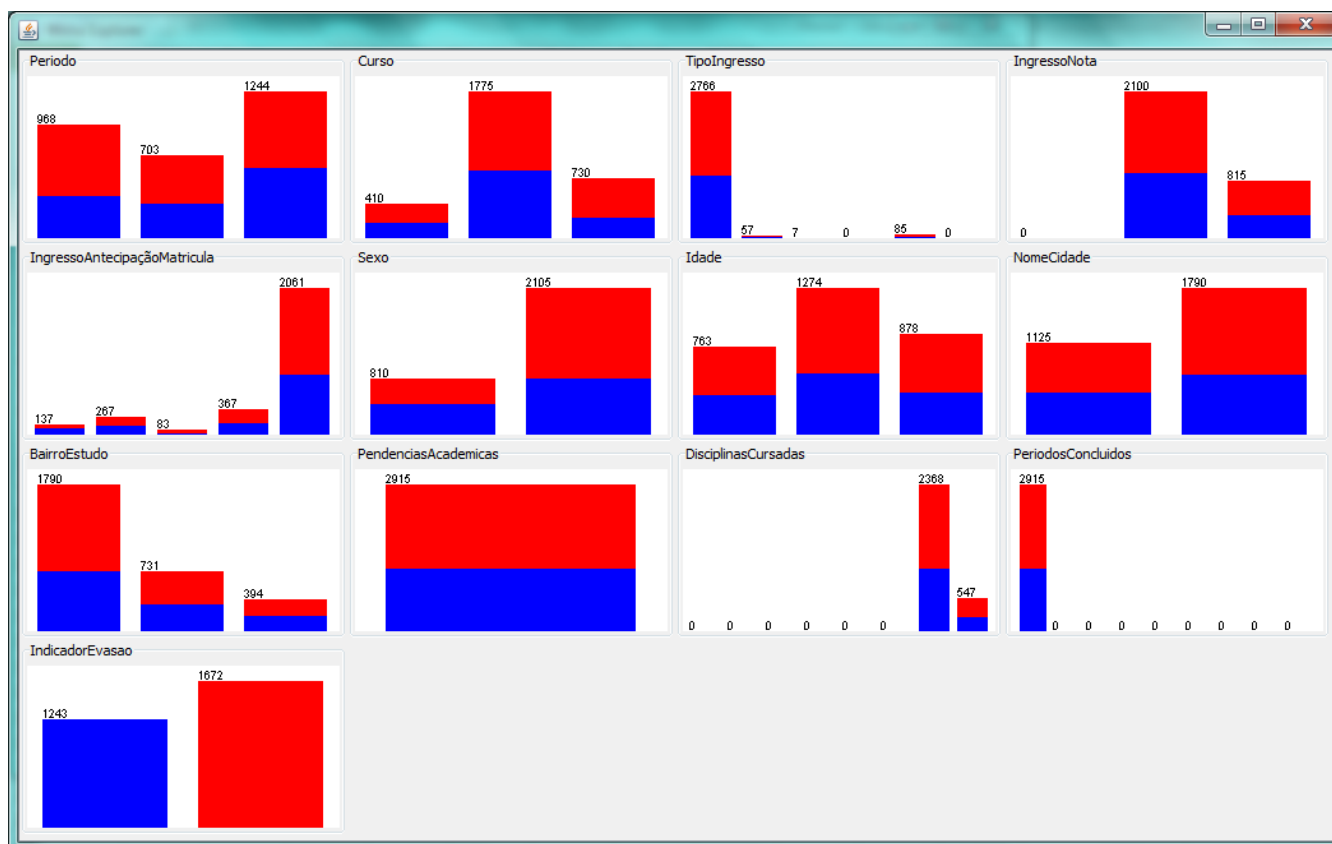
- binarySplits: indica a utilização de divisões binárias nos atributos nominais (valor padrão=false);
- collapseTree: indica se as partes que não diminuem erros de treinamento devem ser removidas (valor padrão=true);
- confidenceFactor: determina o fator de confiança utilizado para o comprimento dos galhos ou “poda” (valor padrão=0.25. Valores menores incorrem em podas maiores e galhos menores);
- debug: indica se a ferramenta deve exibir informações adicionais para o console (valor padrão=false);

- minNumObj: número mínimo de instâncias por folha (valor padrão=2);
- reducedErrorPruning: indica a poda a ser utilizada. O valor true indica a utilização da “poda de erros reduzidos”, enquanto que false (padrão) aponta para a utilização da “poda C.4.5”;
- numFolds: indica a quantidade de dados utilizada para a “poda de erros reduzidos” (valor padrão=3);
- saveInstanceData: indica se os dados de treinamento gerados devem ser salvos para futura visualização (valor padrão=false);
- seed: determina o valor base para a randomização de dados ao utilizar a “poda de erros reduzidos” (valor padrão=1);
- subtreeRaising: indica, se verdadeiro (padrão), se a poda deve considerar o crescimento de subárvores;
- unpruned: o valor false (padrão) determina que a poda deve ser realizada;
- useLaplace: determina que a contagem de folhas deve ser suavizada com base em Laplace (valor padrão=false). A suavização de Laplace (ou Laplace smoothing) é uma técnica geralmente utilizada para se evitar que cálculos probabilísticos resultem em zero por um devido fator nulo dentro de uma série;
- useMDLcorrection: o valor true (padrão) indica que a correção MDL (*Minimum Description Length* - este método mede o tamanho de uma árvore de decisão por meio do número de bits necessários para codificar a árvore e determina árvores codificadas com menor quantidade de bits) deve ser utilizada ao identificar divisões em atributos numéricos.

Para a aplicação da técnica, a base de dados gerada foi separada em dois arquivos distintos: o primeiro contendo 2/3 dos dados selecionados aleatoriamente, foi utilizado para o treinamento e geração da árvore modelo; o segundo, contendo os demais dados, com o objetivo de testar a exatidão da árvore modelo gerada. A realização das duas etapas se justifica pela necessidade de utilização do modelo em valores desconhecidos e não apenas nos dados de que dispomos. Ao aplicar o modelo no conjunto de dados de teste, garantimos que a sua exatidão permanece a

mesma para qualquer conjunto de dados.

Composta por 2.915 instâncias (2/3 dos 4.372 registros), o conjunto de treinamento foi carregado no Weka. A distribuição dos dados pode ser verificada na **Figura 4**.



[abrir imagem em nova janela](#)

Figura 4. Representação gráfica da distribuição dos valores do arquivo de entrada para o caso 2

Ao executar o algoritmo de classificação J48 na base carregada, com a indicação de utilizar o conjunto de treinamento (Use training test), o modelo da árvore foi gerado indicando uma exatidão de aproximadamente 65% conforme a **Figura 5**.

=== Summary ===

Correctly Classified Instances	1880	64.494 %
Incorrectly Classified Instances	1035	35.506 %
Kappa statistic	0.2285	
Mean absolute error	0.4493	
Root mean squared error	0.474	
Relative absolute error	91.8422 %	
Root relative squared error	95.8351 %	
Coverage of cases (0.95 level)	100 %	
Mean rel. region size (0.95 level)	99.2281 %	
Total Number of Instances	2915	

Figura 5. Sumário do resultado do algoritmo J48 no conjunto de treinamento

A execução do algoritmo no conjunto de testes, com base no modelo criado, apresentou uma exatidão de 61,5%, similar à do conjunto de treinamento, confirmando a operação. A **Figura 6** apresenta o sumário dessa execução.

=== Summary ===

Correctly Classified Instances	897	61.5649 %
Incorrectly Classified Instances	560	38.4351 %
Kappa statistic	0.1682	
Mean absolute error	0.4696	
Root mean squared error	0.4963	
Relative absolute error	95.9596 %	
Root relative squared error	100.3113 %	
Coverage of cases (0.95 level)	98.8332 %	
Mean rel. region size (0.95 level)	98.8675 %	
Total Number of Instances	1457	

Figura 6. Sumário do resultado do algoritmo J48 no conjunto de testes

Um grau de exatidão entre 60 e 65% indica uma árvore de decisão com necessidades de melhoria para a produção de resultados mais conclusivos sobre o tema analisado. A introdução de novos dados ou a utilização de um conjunto de atributos diferente, com refinamentos sucessivos do processo, pode levar a melhores níveis de exatidão, mas, para efeito do estudo de caso atual, entendemos que o grau de exatidão obtido é suficiente para as análises desejadas.

A **Tabela 1** apresenta uma representação simplificada da árvore de decisão gerada pelo algoritmo J48 a partir dos dados fornecidos. Dispostas em duas colunas, separadas pelo atributo-raiz (Sexo), estão apresentados somente os ramos e folhas que indicam a condição de evasão (SIM).

Sexo = 1	Sexo = 2
Período = 20111	AntecipaçãoMatricula = SemAntecipacao
DisciplinasCursadas = 6	Período = 20111
NomeCidade = CAPITAL: SIM (66.0/26.0)	Curso = TECNOLÓGICO: SIM (5.0/1.0)
Período = 20112	Curso = BACHARELADO: SIM (5.0/2.0)
AntecipaçãoMatricula = SemAntecipacao	Período = 20112
BairroEstudo = CENTRO: SIM (9.0/2.0)	BairroEstudo = IGUATEMI
BairroEstudo = PARALELA: SIM (1.0)	Curso = TECNOLÓGICO: SIM (2.0)
AntecipaçãoMatricula = AtéUmaSemana: SIM (77.0/23.0)	Curso = BACHARELADO: SIM (1.0)
AntecipaçãoMatricula = AtéDuasSemanas	BairroEstudo = PARALELA: SIM (4.0)
BairroEstudo = PARALELA: SIM (5.0/1.0)	Período = 20121: SIM (41.0/12.0)
AntecipaçãoMatricula = AtéUmMês	AntecipaçãoMatricula = AtéUmaSemana
BairroEstudo = IGUATEMI: SIM (12.0/2.0)	NomeCidade = CAPITAL
BairroEstudo = PARALELA	Idade = Até25anos
Idade = Até25anos: SIM (0.0)	Curso = TECNOLÓGICO: SIM (1.0)
Idade = de25a35anos: SIM (3.0)	Curso = LICENCIATURA
Idade = Maisque35anos	BairroEstudo = PARALELA: SIM (2.0)
IngressoNota = >=6: SIM (3.0/1.0)	Idade = de25a35anos
AntecipaçãoMatricula = MaisDeUmMês	IngressoNota = SemNota: SIM (0.0)
Idade = Até25anos	IngressoNota = <6: SIM (17.0/5.0)
TipologiaIngresso = VEST	IngressoNota = >=6
IngressoNota = <6: SIM (3.0/1.0)	Curso = TECNOLÓGICO: SIM (6.0/3.0)
TipologiaIngresso = ENEM: SIM (3.0)	Curso = BACHARELADO: SIM (8.0/2.0)
TipologiaIngresso = ME: SIM (0.0)	Idade = Maisque35anos
TipologiaIngresso = TE: SIM (0.0)	BairroEstudo = PARALELA
TipologiaIngresso = TI: SIM (0.0)	Curso = TECNOLÓGICO: SIM (2.0)
TipologiaIngresso = PROUNI: SIM (0.0)	Curso = BACHARELADO: SIM (3.0/1.0)
	AntecipaçãoMatricula = AtéUmMês

Idade = Maisque35anos	Período = 20112
BairroEstudo = CENTRO	TipoIngresso = ENEM: SIM (5.0/1.0)
IngressoNota = <6: SIM (3.0/1.0)	Período = 20121: SIM (25.0/8.0)
BairroEstudo = IGUATEMI: SIM (4.0/1.0)	
Período = 20121: SIM (354.0/138.0)	

Tabela 1. Representação simplificada da árvore de decisão gerada no Caso 2

Analisando a árvore gerada, podemos inferir algumas situações de evasão indicadas.

Identificamos, por exemplo, que estudantes do sexo feminino que ingressaram em 2011.2 com seis disciplinas na capital são, potencialmente, um grupo de evasão, uma vez que das 66 instâncias identificadas segundo essas características, apenas 26 não evadiram, ou seja, possuem 71,7% de ocorrência de evadidos.

Os atributos Período e AntecipacaoMatricula estão mais próximos da raiz, respectivamente, para instâncias do sexo feminino (1) e masculino (2), enquanto que Idade, TipoIngresso e IngressoNota se apresentam, predominantemente, mais próximos das folhas da árvore.

Chama a atenção também o fato de 81% (46 instâncias, de um total de 60) dos estudantes do sexo masculino que anteciparam suas matrículas em até duas semanas persistirem nos seus cursos, indicando um baixo índice de evasão nessa categoria.

Análise dos resultados

A utilização do modelo proposto em técnicas de mineração de dados propicia a identificação de indícios de evasão em estudantes. A sua formulação indica que deve haver um conjunto de indícios comuns, detectados, tanto através da utilização de técnicas de mineração com o modelo de dados proposto, quanto através da observação estatística direta nos dados utilizados.

Um dos indícios de evasão identificados através da análise direta da base de dados encontra-se em cursos de Graduação Tecnológica. Mesmo que discretamente, há uma pequena tendência a evadir em cursos dessa categoria do que em Bacharelados ou Licenciaturas. A análise da árvore de decisão gerada aponta cinco conjuntos de instâncias que contêm dados dessa categoria, todas com um alto índice de ocorrências de acertos nas regras, como pode ser notado na **Tabela 2**.

NÍVEL 1 (SEXO)	NÍVEL 2 (ANTECIPAÇÃO DA MATRÍCULA)	NÍVEL 3	NÍVEL 4	NÍVEL 5	OCORRÊNCIAS	% ACERTO
MASC	SemAntecipacao	Periodo = 20111			(5.0/1.0)	83,30%
MASC	SemAntecipacao	Periodo = 20112	BairroEstudo = IGUATEMI		(2.0)	100,00%
MASC	AtéUmaSemana	Cidade = CAPITAL	Idade = Até25anos		(1.0)	100,00%
MASC	AtéUmaSemana	Cidade = CAPITAL	Idade = de25a35anos	IngressoNota = >=6	(6.0/3.0)	66,70%
MASC	AtéUmaSemana	Cidade = CAPITAL	Idade = Maisque35anos	BairroEstudo = PARALELA	(2.0)	100,00%

Tabela 2. Ramos da árvore de decisão que incluem estudantes de Graduação Tecnológica

Por outro lado, nas informações sobre evasão, extraídas a partir do resultado da aplicação do algoritmo de agrupamento, no caso 1, não foi encontrado nenhum destaque a respeito de cursos de Graduação Tecnológica e, pelo contrário, um grupo se evidenciou dos demais pelo seu percentual de incidência relacionado a cursos de Bacharelado. Apesar de, aparentemente contraditórios, esses resultados não são incompatíveis, uma vez que é possível, de acordo com a classificação, identificar situações com maior potencial de evasão para cursos de determinada categoria e, com a aplicação de outras técnicas, serem encontrados resultados que destacam outras situações.

A análise estatística direta apontou uma maior tendência a evadir em estudantes oriundos de processos de matrícula especial (portadores de diploma superior) e transferência externa, que apresentou uma proporção no grupo de evadidos (8,9%) maior do que na totalidade na amostra (0,2%). A aplicação do algoritmo de agrupamento, por outro lado, evidenciou que ingressantes por vestibular possuem uma maior propensão a evadir. Isso ocorre porque, diferentemente da análise estatística, realizada atributo a atributo, a técnica do agrupamento busca o comportamento dos

dados a partir do conjunto dos atributos utilizados.

Ao analisar o impacto da antecipação das matrículas dos estudantes no fenômeno da evasão, porém, as três análises convergem quando apontam que estudantes que se matriculam com muita antecedência tendem a evadir mais que aqueles que o fazem em períodos mais próximos do encerramento do prazo. A análise direta indica que esse grupo, dentre os evadidos, tem uma incidência de 73,2%, diferentemente da incidência no universo de matriculados, com 57,1%. Os clusters de evadidos gerados na aplicação do algoritmo do caso 1 apontam uma média de 40 dias de antecipação e, na aplicação do algoritmo de classificação, no caso 2, o atributo aparece próximo do topo da árvore, determinando a evasão em 45 ocorrências, com incidência média de 70%.

Diversas inferências podem ser realizadas sobre potenciais fatores de evasão de estudantes a partir dos resultados das aplicações dos algoritmos dos casos 1 e 2 e, como descrito, alguns desses indícios convergem com as análises realizadas a partir dos dados estatísticos dos dados, antes da mineração. Futuras análises, com mais detalhamento sobre esses resultados e com refinamentos na execução dos algoritmos nas bases de dados construídas podem trazer mais e melhores conclusões sobre os impactos desses atributos nas decisões dos estudantes em evadir ou não de seus cursos.





DevMedia

A DevMedia é um portal para analistas, desenvolvedores de sistemas, gerentes e DBAs com milhares de artigos, dicas, cursos e videoaulas gratuitos e exclusivos para assinantes.

Publicado em 2015

O que você achou deste post?

 [Gostei \(0\)](#)  (0)

[+ Mais conteúdo sobre SQL](#)

Não há comentários

[Postar dúvida / Comentário](#)

[Meus comentarios](#)

Publicidade



Mais posts

Video aula

[Agrupando registros com Group by - Curso Completo MySQL - Aula 39](#)

Video aula

[Busca redundante em uma tabela - Curso Completo MySQL - Aula 38](#)

Video aula

[Obtendo dados de mais de uma tabela - Curso Completo MySQL - Aula 37](#)

Video aula

[Introdução ao Comando Select - Curso Completo MySQL - Aula 36](#)

Video aula

[Entendendo os tipos de dados Data no MySQL - Curso Completo MySQL - Aula 35](#)

Aplicativo com fontes

Código Fonte - Curso Dominando XML com SQL Server

Listar mais conteúdo



Anuncie | Loja | Publique | Assine | Fale conosco



DevMedia

Curtir Página

81 mil curtidas

Seja o primeiro de seus amigos a curtir isso.



Hospedagem web por Porta 80 Web Hosting