AVALIAÇÃO DE PERFIS DE ALUNOS E CURSOS COM MINERAÇÃO DE DADOS

Carlos Henrique Alves Silva (IC) e Leandro Augusto da Silva (Orientador)

Apoio:PIVIC Mackenzie

RESUMO

Este artigo apresenta uma metodologia de mineração de dados com técnica de agrupamento para categorizar disciplinas de um curso de Sistemas de Informação. Os dados utilizados foram conseguidos por meio de pesquisa aos alunos que estão no 6º. Semestre do curso, onde os mesmos preencheram uma planilha com suas médias finais em cada disciplina. Todo o processo para descobrir os grupos foi discutido e os resultados encontrados permitem analisar o perfil dos alunos, o perfil do curso e relacionar disciplinas de cada grupo descoberto. Os resultados apresentados podem ter utilidade como construção de indicadores, planejamento de ações e, também, subsidiar trabalhos futuros sobre o desempenho de cursos em instituições de ensino.

Palavras-chave: Mineração de dados educacionais, análise de agrupamento, análise de cursos.

ABSTRACT

This paper addresses a data mining methodology with clustering analysis to categorize subjects from Information Systems courses. The database used was obtained by the students that are in the 6th semester, where they completed a spreadsheet with the final grades for each course. The process used to find the groups was discussed and the results allow us to analyse the student profile, the outline course and to list the subjects in each group discovered. The results presented can be useful as indicators parameter, action planning and support future work on the courses performance in educational institutions

Keywords: Educational Data mining, Clustering Analysis, Courses Analysis.

1. INTRODUÇÃO

Um banco de dados constitui de uma estrutura para armazenamento de fatos que se relacionam de forma a fazer sentido. A chegada dos Sistemas Gerenciadores de Banco de Dados (SGBDs) trouxe, entre outras coisas, um armazenamento relacionado para eliminar redundâncias e aumentar a capacidade de armazenamento, uma coleção de restrições para garantir consistência no armazenamento dos dados e um conjunto de álgebras que permitiu a construção de linguagens para consultas padronizadas. Podemos atribuir a esse fato, ocorrido por volta dos anos 80, como um dos principais propulsores para o desenvolvimento dos Sistemas de Informação (SI) (Silberchatz, Korth e Sudarshan, 2006).

A área profissional que atua em um SI é chamada de Tecnologia da Informação (TI) e por muito tempo tinha como principal desafio a sistematização de processos e uso de SGBDs para o armazenamento de dados. Com a maturidade de linguagens de consulta estruturada ou SQL (*Structured Query Language*), baseada em álgebras relacionais e teorias de conjuntos, os dados passam a ser manipulados de tal forma a permitir a construção de relatórios estatísticos e os gestores passaram a vislumbrar esses resultados como apoio a tomadas de decisões.

Em paralelo a tudo isso, o hardware que antes era de dimensão em bytes reduzida e muito caro, evolui para um cenário totalmente oposto. E assim, começou a gerar grande volume de dados e o interesse por uma segunda geração de análise de dados, conhecida por Mineração de Dados ou DM (*Data Mining*). O DM é parte de um processo que tem como entrada uma base de dados e como saída um conhecimento que estava intrínseco aos dados. Esse processo ficou conhecido por Descoberta de Conhecimento em Banco de Dados ou KDD (*Knowledge Discovery Database*) (Fayyad et. al., 1996), (Witten etl al. 2011).

Não diferente das grandes corporações empresarias, as Instituições de Ensino, em geral, nas quais se incluem as de Ensino Superior (IES), tem grande parte das transações diárias armazenada em banco de dados. A cada dia estes dados são atualizados, como, por exemplo, no preenchimento das listas de presença, lançamento de notas de provas parciais, pagamentos de mensalidade, transferência de alunos etc. Com o uso da linguagem SQL as IES também geram uma série de relatórios estatísticos, como, por exemplo, levantamento sobre ausências de alunos, médias parciais do semestre, levantamentos de aprovação de alunos, disciplinas com maior número de

reprovação e muitas outras informações que auxiliam os gestores e coordenadores no acompanhamento de disciplinas, alunos e infraestrutura de uma IES.

No entanto, estes levantamentos típicos são extrações triviais de dados que nem sempre permitem uma compreensão mais aprofundada e fundamentada para tomada de decisões. Além do mais, acredita-se que muitas outras informações estão escondidas nos grandes volumes de dados o que, consequentemente, poderiam ser úteis para apoiarem gestores em diversas esferas estratégicas na tomada de decisão. E nesse momento que se faz necessário o uso da DM. Através de seu uso, seria possível enxergar relações entre dados que hoje não são levadas em consideração na preparação de aulas por um professor, afinidades para algum grupo de disciplina, dificuldade em algum tema específico, exploração de atividades mais efetivas nos alunos.

Diante o contexto previamente apresentado, este trabalho tem como dois principais objetivos. O primeiro é contextualizar o processo de KDD em um problema real, discutindo a importância de cada fase do processo, servindo assim como um estudo de caso em EDM na classe de Educação presencial. O segundo e mais importante objetivo é usar notas de alunos cursando o 6º. Semestre em Sistemas de Informação e, através da técnica de mineração de dados de análise de grupos, descobrir três diferentes padrões de disciplinas. Cada padrão será chamado de: difícil, regular e fácil. O algoritmo de agrupamento utilizado será o k-Médias, que de forma iterativa, particiona a base de dados em k distintos grupos (Witten et. al. 2011). Para a realização dos experimentos será utilizada a ferramenta Rapid Miner, disponível para download sem a necessidade de licença (Mierswa et. al. 2006). A principal novidade deste trabalho é que ao invés de avaliar o perfil do aluno, como em outros trabalhos da literatura (Romero e Ventura 2008; Romero e Ventura 2010), queremos aqui avaliar o perfil das disciplinas. Portanto, uma base de dados que, geralmente é organizada em exemplos (linhas) com alunos e atributos (colunas) com disciplinas, tendo como valor as notas tiradas, para permitir o estudo proposto neste trabalho, ela deverá passar por um rotacionamento, trocando as linhas por colunas.

2. REFERENCIAL TEÓRICO

DM é uma área interdisciplinar que envolve, basicamente, banco de dados, estatística e inteligência computacional e se aplica em várias áreas do conhecimento para descoberta de informações escondidas em bancos de dados (Fayyad et. al., 1996).

Na área de educação, há na literatura muitas propostas de trabalhos usando Mineração de Dados (Pimentel e Omar, 2006; Dias et. al., 2008; Romero e Ventura 2007; Romero e Ventura 2008; Marinho, Dermeval e Ferreira, 2009; Romero e Ventura 2010; Malvezzi, Mourão e Bressan, 2010). Esta é uma área de interesse conhecida como Mineração de Dados Educacionais ou EDM (*Educational Data Mining*) (Romero e Ventura, 2007; Romero e Ventura, 2010). Segundo Romero e Ventura (2010) os tipos de estudos em EDM são classificados em:

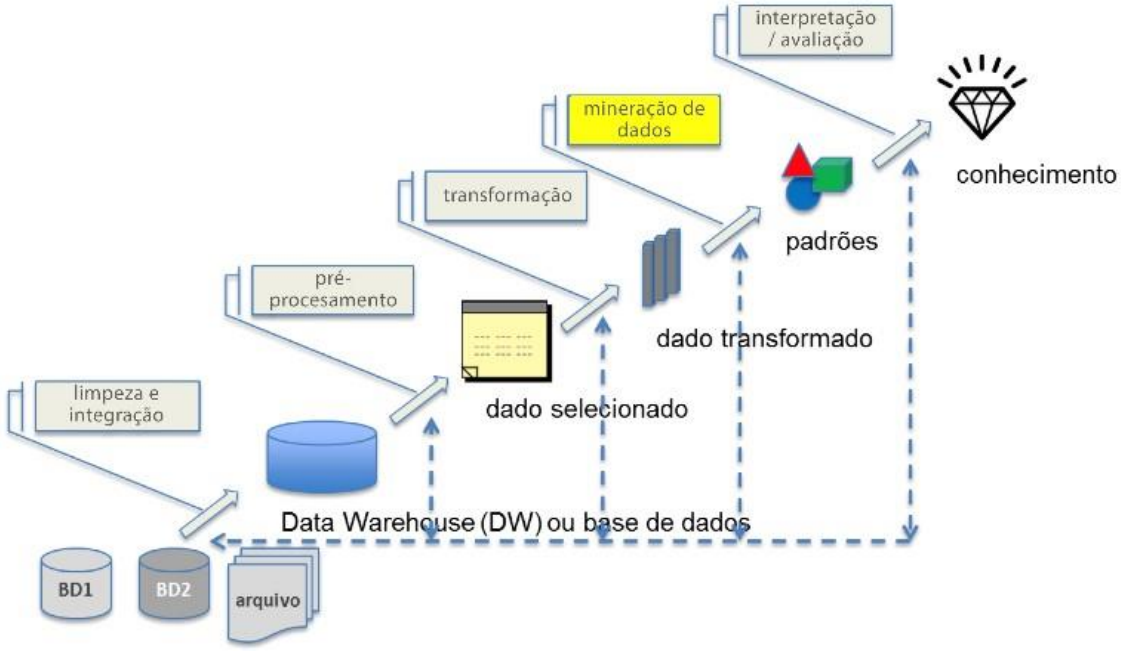
(i) Educação presencial (*Offline education*): analisa os alunos com dados adquiridos a partir do seu comportamento em sala de aula, desempenho nas avaliações, curriculum etc. Pode ser considerado como exemplo deste tipo de estudo, quando se deseja saber com antecedência o conhecimento do aluno em uma disciplina, a partir de uma avaliação prévia (Pimentel e Omar, 2006; Romero e Ventura 2007; Romero e Ventura 2010).

(ii) Educação à distância e Sistema para gerenciamento de aprendizagem ou LMS (Learning Management System) constitui-se da forma mais simples de fazer mineração de dados, pois o ambiente permite armazenar muitos parâmetros dos alunos como tempo de: estudo, leitura, resolução de exercícios etc; desempenho em: exercícios, avaliações etc; entre outros (Dias et. al., 2008; Romero e Ventura 2008; Romero e Ventura 2010; Malvezzi, Mourão e Bressan, 2010).

(iii) Sistemas tutores inteligentes ou ITS (*Intelligent Tutoring System*) e Sistemas hipermídia educacionais adaptativos ou AEHS (*Adaptive Educational Hypermedia System*) onde a mineração de dados é utilizada para adaptar os sistemas de ensino para cada perfil de estudante, usando como fonte de dados os arquivos de log do usuário, modelo de alunos, os quais permitem fazer sugestões de estudos, ajuste de conteúdos etc (Romero e Ventura 2007; Marinho, Dermeval e Ferreira, 2009; Romero e Ventura 2010).

Como se pode observar a partir da Figura 1, Mineração de Dados é parte de um processo maior chamado Descoberta de Conhecimento em Base de Dados ou KDD (Knowledge Discovery Database).

Figura 1 - Processo de KDD no qual se inclui a Mineração de Dados



Autor: (Fayyad et. al., 1996).

No processo de KDD, cada fase deve ser avaliada em relação à necessidade, a partir do problema a ser resolvido. E também, cada fase pode ser repensada, em caso no passo seguinte do processo uma tarefa anterior seja necessária. Ou seja, o processo é iterativo e interativo. De forma breve, cada etapa do processo é definida como:

(i) Limpeza e integração: os dados, oriundos de uma ou mais bases, são integrados em um repositório único. Nesse processo, valores faltantes (*missing values*) e ou fora de um padrão (*outlier*) pode surgir na base de dados.

(ii) Pré-processamento: alguns exemplos da base (linhas) ou atributos (colunas) são mais representativos que outros no processo de KDD e deve se pensar em manter ou em eliminar para a redução de complexidade na etapa do DM.

(iii) Transformação: para que os dados estejam em um único formato, muitas vezes é preciso que se faça a normalização dos dados para que os valores fiquem em uma única escala de valor.

(iv) Mineração de Dados ou DM: decide-se pela técnica de mineração a ser aplicada, de acordo com o problema em análise. Estas técnicas, de forma breve, podem ser definidas como:

Predição: consiste em mapear um exemplo desconhecido da base de dados, também chamada de base de treinamento, em um atributo especial chamado de rótulo

(ou label, em inglês). Pode ser dividida em classificação (mapeamento em rótulos com valores discretos) e estimação (mapeamento em rótulos com valores contínuos).

Agrupamento: consiste em agrupar os exemplos da base de dados que são similares entre si e dissimilares entre os grupos.

Associação de dados: consiste em encontrar relações entre atributos da base de dados.

(v) Interpretação ou avaliação dos resultados: os resultados de todo o processo precisam ser avaliados, em termos de desempenho, para que seja feito uma análise qualitativa (interpretação) ou quantitativa (avaliação) dos resultados obtidos.

A análise de agrupamento, como brevemente definido anteriormente, é uma técnica de mineração de dados usada para segmentar uma base de dados em grupos. Na literatura encontramos várias abordagens para descoberta de grupos em base de dados, como por exemplo, particionais, hierárquicas ou por grafo. Neste trabalho a escolha foi pela abordagem particional, que é comumente utilizada. Dentro dessa, a técnica utilizada foi a popularmente conhecida como k-Médias (k-Means) (Witten et. al. 2011).

O algoritmo do k-Médias opera para gerar um particionamento da base de dados em k inteiros grupos da seguinte maneira:

Escolha k grupos aleatoriamente

Calcule a centroide para cada grupo

Repita

Atribua cada exemplo da base ao centroide mais próximo

Recalcule o centroide para cada grupo

Até estabilidade

O valor de k e a métrica de distância são dois parâmetros que devem ser fornecidos à priori para o algoritmo k-Médias. A escolha do valor de k pode ser um grande desafio para a aplicação do algoritmo. Em contextos onde se sabe previamente o número de grupos desejado, como neste trabalho, a escolha do parâmetro k é um problema trivial. Por outro lado, quando não se sabe o número de grupos ou, então,

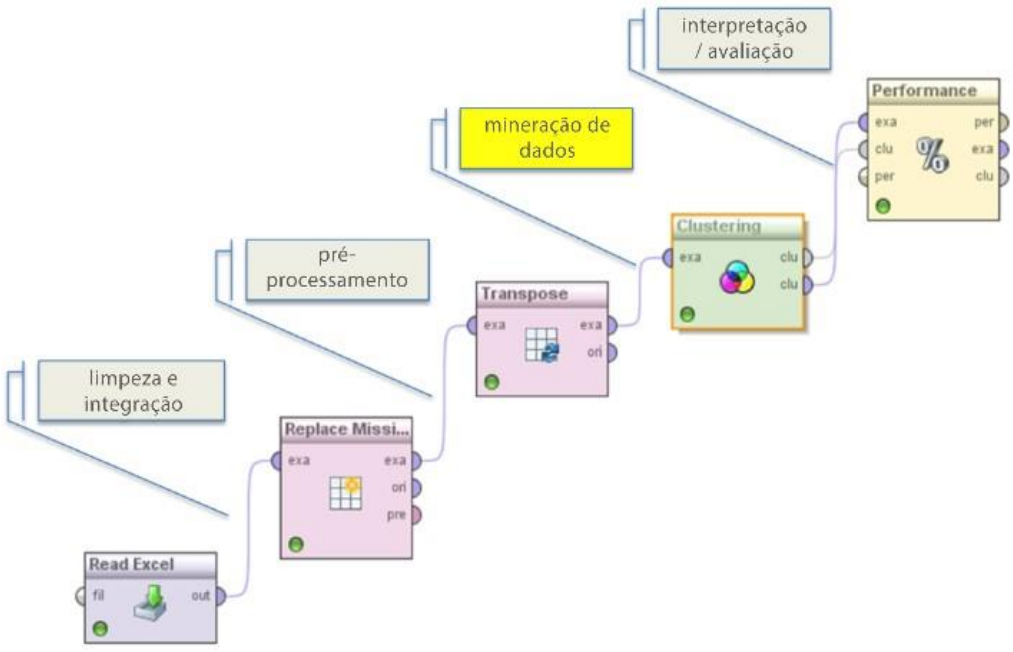
quando se deseja descobrir o número de grupos, é preciso considerar a variação do parâmetro e o uso de medidas quantitativas para que se faça a melhor escolha.

Entretanto, a escolha do outro parâmetro do algoritmo, métrica de distância, é um problema de menor proporção. Por consenso, é comum que se adote a distância Euclidiana. Contudo, há várias outras métricas que se pode utilizar como distância de Manhattan, distância do cosseno dentre outras (Witten et. al. 2011). O Rapid Miner permite a escolha dessas métricas de distância e de muitas outras (Mierswa et. al. 2006).

3. METODOLOGIA

Como introduzido previamente, os experimentos deste trabalho foram feitos com o uso da ferramenta de mineração de dados chamada Rapid Miner (Mierswa et. al. 2006). Trata-se de uma ferramenta de aprendizado simples, onde todas as fases da descoberta de conhecimento em bases de dados podem ser preparadas na forma de um processo, veja Figura 2. Note que, de todas as etapas do processo de KDD apresentado na Figura 1, apenas a transformação de dados não se fez necessário para o objetivo deste trabalho. Isso devido ao fato de todas as notas estarem em um intervalo de valores de 0 a 10.

Figura 2. Processo empregado no trabalho para a segmentação da base de dados.



Autor: (Fayyad et. al., 1996).

Para a formação da base de dados, cada um dos 50 alunos cursando o 6º. Semestre em Sistemas de Informação recebeu uma planilha, exemplificada na Tabela 1. O aluno, sem se identificar, preencheu sua média final em todas as disciplinas cursadas até o corrente período. Como resultado de preenchimento da planilha gerouse uma base de dados com 50 exemplos (linhas) representando cada aluno e 39 atributos (colunas) representando as disciplinas oferecidas nos cinco primeiros semestres do curso.

Tabela 1 - Exemplo da planilha que cada aluno recebeu para preenchimento

	1º. Sem.			***2º. Sem.***			
	Ing. Téc. I.	**Mat.**	...	**Est. de Dados**	**Tec. Web**	...	
1	7,0	5,5	...	6,0	4,8	...	
...	...	...	...	...	...	...	
50	6,5	5,0	...	4,5	6,0	...	

Os alunos que tinham dependência foram orientados a preencher a planilha com a média final que tirou na primeira vez que cursou a disciplina.

Depois de finalizada a etapa de integração de todas as planilhas, observou-se que algumas notas não foram preenchidas. Isso se deve ao fato que na IES onde a pesquisa foi realizada, o aluno pode optar em montar a sua grade de disciplinas. Portanto, a base apresenta valores ausentes, em razão do aluno ainda não ter feito à disciplina, e que devem ser tratados pela limpeza.

Para a realização da limpeza de preenchimento de valores ausentes, a ferramenta Rapid Miner oferece as seguintes opções: mínimo, média, máximo e nulo. A opção escolhida no projeto foi pelo valor mínimo, isto significa que as notas faltantes em alguma disciplina foram consultadas e escolheu-se para preenchimento àquela de menor valor. Esta etapa foi necessária para 1,5% de todos os valores da base de dados.

Como o objetivo do trabalho é identificar o perfil de cada disciplina, ou seja, verificar as disciplinas que os alunos têm maior dificuldade em aprovar, e a base de

dados está estruturada com linha para alunos e colunas para disciplinas, houve a necessidade, ainda no pré-processamento, de inverter a estrutura da base de dados.

No Rapid Miner isso é possível com o uso do operador “Transpose” do módulo de préprocessamento. Assim, a base terá, após este processo, as linhas com disciplinas e as colunas com os alunos. Para a mineração de dados, como adiantado anteriormente, escolheu-se o algoritmo k-Médias. Os parâmetros escolhidos para que o algoritmo segmentasse a base de dados foi k=3, pelas razões também discutidas anteriormente (três níveis de dificuldade), e medida de distância Euclidiana.

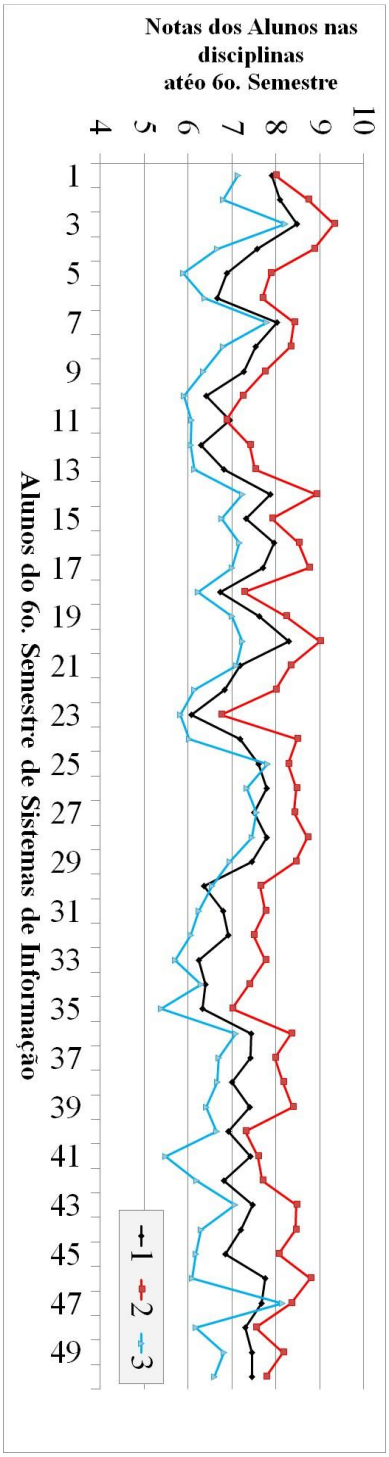
Na próxima seção, os resultados obtidos são apresentados e, com isso, se apresenta também a avaliação e a interpretação dos experimentos, que é a última fase do processo de KDD (Figura 1).

4. RESULTADO E DISCUSSÃO

Os resultados apresentados a seguir mostram os grupos de disciplinas que, a partir do cálculo da média das notas dos alunos, serão classificados em difícil, regular e fácil. A partir do agrupamento, apresenta-se também como resultado as disciplinas segmentadas, permitindo uma análise quanto ao perfil de cada uma delas, ou seja, se as disciplinas são da área de programação, tecnológicas ou humanísticas. Estes resultados possibilitam, entre outras analises, a identificação de relações entre as disciplinas.

O resultado apresentado na Figura 3 relaciona a média final (eixo y) para cada aluno (eixo x). O desempenho do aluno em cada disciplina é visto no gráfico de dispersão, onde cada curva está rotulada por cores, de acordo com o grupo descoberto pelo algoritmo k-Médias.

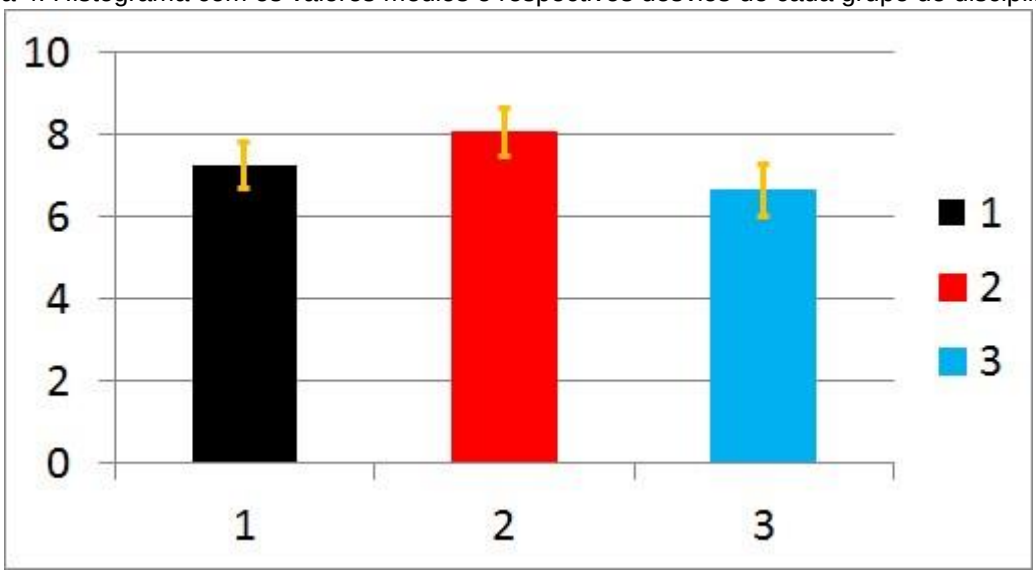
Figura 3 - Série com as médias dos 50 alunos segmentadas em 3 grupos.



A partir do resultado da Figura 3, deseja-se saber a média dos alunos em cada grupo de disciplina, junto com o desvio padrão. Esses resultados são apresentados na Figura 4, cujo resultado já permite classificar cada grupo de disciplinas, a partir da média. Ou seja, o grupo 3 é tido como aquele com disciplinas de menor média, 6,6 ± 0,65, e pode ser classificado como o grupo de disciplinas difíceis. O grupo 1, por outro lado, é

o segundo de menor média, 7,2 ± 0,57, e pode ser classificado como disciplinas regulares. Por conseguinte, o grupo 2 tem disciplinas com média 8,0 ± 0,58, sendo classificada de disciplinas fáceis.

Figura 4. Histograma com os valores médios e respectivos desvios de cada grupo de disciplinas.

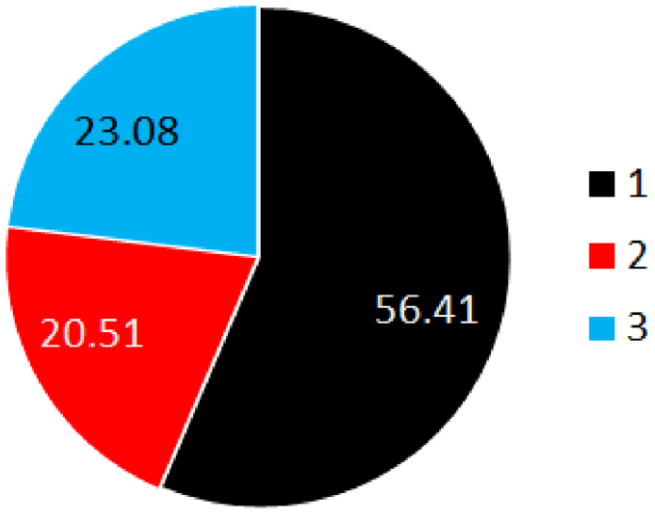


As disciplinas de cada grupo, agora já classificadas, são apresentadas na Tabela 2. Estas informações são importantes para saber a quantidade de disciplinas mapeada em cada um dos grupos e, também, para conhecer a relação entre elas, bem como o perfil das mesmas. Avaliando as disciplinas da Tabela 2 é possível apresentar a porcentagem de disciplinas em cada grupo, Figura 5.

Tabela 2 – Relação das disciplinas mapeadas em cada grupo

1	2	3
Éti. E Cida. I	Ing. Técn. I	Mat. para S.I. II
Mat. Para SI I	Fun. S.I.	Arq. de Comp.
Amb. Oper.	Fun. De S.I. II	Mat. Para S.I. III
Int. Prog.	Éti. E Cida. II	Est. de Dados II
Com. E Escr. Técn.	Ing. Técn. II	Ling. de Progr II
Est. de Dados I	Aná. De Proc. S.I.	Sis. Oper.
Ling. de Prog I	Téc. Prog. Apli. II	Mod. De Sist.
Tec. Web	Int. Econ.	Sis. Dist.
Adiministração		Des. de BD
Eng. de Soft I		
Redes		
Pro. E Esta.		
Aná, de Sist. I		
Téc. Prog. Apli. I		
Int. a BD		
Pro. Inte. I		
Int Cont.		
Mat. Fina.		
Pes. Ope.		
MTC		
Eng. de Soft. II		
Direito		

Figura 5 - Distribuição das disciplinas mapeadas em cada grupo.



Com base em todas as informações descobertas, a partir do processo de agrupamento de dados, é importante que todos os resultados sejam sumarizados, com o objetivo principal de apresentar um conhecimento dos dados analisados. Estes resultados são apresentados na Tabela 3 e permitem um relacionamento das informações de classificação de cada grupo de disciplina, com os respectivos valores médios; um conhecimento sobre a porcentagem de disciplina no grupo; um conhecimento do perfil do grupo de disciplinas. Com esse último, consequentemente, é possível inferir o perfil de cada aluno no curso de Sistemas de Informação.

Diante os resultados apresentados, principalmente os que estão sumarizados na Tabela 3, é possível tirar alguns conhecimentos do perfil do curso, disciplina e alunos. Por exemplo:

(i) Matérias difíceis e de grande reprovação se encontram no grupo 3. Estas matérias representam 23,08% das disciplinas do curso e tem o perfil, basicamente, de programação. Com esse resultado se pode constatar a dificuldade dos alunos neste tipo de disciplina.

(ii) A maior parte das disciplinas, 56,41%, se encontra no grupo 2, com média de 8,0 ± 0,58. Levando-se em conta que a média para aprovação na IES onde a pesquisa foi realizada é 6,0, pode-se afirmar que na maioria das disciplinas até o 6º. semestre, os alunos não têm dificuldade de aprovação.

(iii) Ainda analisando as disciplinas do grupo 2, percebe-se que elas estão associadas, na maioria dos casos, com disciplinas humanísticas.

(iv) O grupo 1 concentra disciplinas de Tecnologia e, também de Programação, e tem média final de 7,2 ± 0,57, acima da média para aprovação que é de 6,0.

(v) Resultados coerentes aparecem na classe regular, com 20,51% das disciplinas de programação e tecnologia.

(vi) O aluno do curso de Sistemas de Informação da IES analisada, de modo razoável, tem dificuldades em disciplinas que envolvem programação.

Analisando as classes regular e fácil, juntas elas representam 76,92% das disciplinas. Em ambos os casos, o índice de aprovação é alto, lembrando que a média final no curso é de 6,0. Por outro lado, o nível de reprovação está abaixo dos 23,08%, o que pode ser considerado normal, pois trata-se de disciplinas de programação, o que é um problema nacionalmente conhecido em cursos de computação.

5. CONSIDERAÇÕES FINAIS

O trabalho teve como objetivos a contextualização do processo de KDD em um problema real, aqui foi levado em consideração avalição do perfil de disciplinas do curso presencial de Sistemas de Informação de uma IES, expondo a necessidade e importância de cada etapa desse processo. Utilizando como base para técnica de análise de grupos as notas de alunos do 6º semestre do curso, as disciplinas foram separadas em três padrões: fácil, regular e difícil.

Através da classificação dessas disciplinas, é possível observar uma série de possibilidades na tomada de decisão de professores e coordenadores de curso, na hora de endereçar preparação de grade curricular, entender os pontos nos quais a instituição está indo bem em determinado grupo de matérias.

Como trabalho futuro, é proposto relacionar os dados dos grupos de disciplinas e os estilos de ensino e modelos de aula, e descobrir se é efetivo utilizar técnicas de ensino similares em cada grupo de disciplina.

6. REFERÊNCIAS

Dias, M. M., Silva-Filho, L. A., Lino, A. D. P., Favero, E. L e Ramos, E. M. L. S., "*Aplicação de Técnicas de Mineração de Dados no Processo de Aprendizagem na Educação a Distância"*. In: SBIE - Simpósio Brasileiro de Informática na Educação, p. 105-114, 2008

Fayyad, U., Piatetsky-Shapiro, G., Smith, P., "*From Data Mining to KDD*". In: AI Magazine v. 17, n. 3, p. 37-54, 1996

Malvezzi, W. R., Mourão, A. B. e BRESSAN, G., "*Uma Ferramenta Baseada em Teoria Fuzzy para o Acompanhamento de Alunos Aplicado ao Modelo de Educação Presencial Mediado por Tecnologia"* In: SBIE - Simpósio Brasileiro de Informática na Educação, v. Único, p. 9, 2010

Marinho, T., Dermeval, D., Ferreira, R., Braz, L. M., Bittencourt, I. I., Costa, E. B. e Luna, H. P., "*Um Framework para Mineração de Dados Educacionais Baseado em Serviços Semânticos"*. In: WIE - Workshop de Informática na Escola, p. 2368-2373, 2009

Mierswa, I., Wurst, M., Klinkenberg R., Scholz, M. and Euler, E. (20 "Yale: Rapid prototyping for complex data mining tasks" In Lyle Ungar, Mark Craven, Dimitrios Gunopulos, and Tina Eliassi-Rad, editors, KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, p. 935-940, 2006

Pimentel, E. e Omar, N., "*Descobrindo Conhecimentos em Dados de Avaliação da Aprendizagem com Técnicas de Mineração de Dados*". In: WIE - Workshop de Informática na Escola, p. 147-155, 2006

Romero, C., & Ventura, S., Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 33(1), 135-146, 2007

Romero, C., Ventura, S. e García, H. "*Data mining in course management systems: Moodle case study and tutorial*". In: Computers & Education 51, pages 368-384, 2008
Romero, C., & Ventura, S. *Educational data mining: a review of the state of the art*. Systems, Man, and Cybernetics, Part C: IEEE Transactions on Applications and Reviews, 40(6), 601-618, 2010

Silberschatz, A., Korth, H. F., & Sudarshan, S., *"Sistema de banco de dados*". Campus. 3a. edição, 2006

Silva, L.A., Peres, S.M. and Boscarioli, C. Introdução à Mineração de Dados com Aplicações, R. Elsevier, 1ª. edição, 2016

Witten, I.H., Frank, E. and Hall, M.A., "Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann", 3rd edition, 2011

Contatos: contato@carloshsilva.com leandroaugusto.silva@mackenzie.br