AGRUPAMENTO DE DADOS PARA CARACTERIZAÇÃO DOS ESTUDANTES DE EDUCAÇÃO PROFISSIONAL DE NÍVEL MÉDIO DO INSTITUTO FEDERAL DE SÃO PAULO

HUGO LEONARDO DOS SANTOS[1], FÁBIO JOSÉ JUSTO DOS SANTOS[2], CRISTIANE AKEMI YAGUINUMA[3], CÍNTIA MAGNO BRAZOROTTO[4]

1 Discente, Bolsista PIBIFSP, IFSP, Câmpus Araraquara, hugo.leonardo@aluno.ifsp.edu.br
2 Docente, IFSP, Câmpus Araraquara, fabiojjs@ifsp.edu.br
3 Docente, IFSP, Câmpus Araraquara, cristiane.yaguinuma@ifsp.edu.br
4 Técnico administrativo, IFSP, Câmpus Araraquara, cbrazorotto@ifsp.edu.br
Área de conhecimento (Tabela CNPq): 1.00.00.00-3 Ciências Exatas e da Terra

Apresentado no
10° Congresso de Inovação, Ciência e Tecnologia do IFSP
27 e 28 de novembro de 2019- Sorocaba-SP, Brasil

RESUMO: Gestores e educadores atualmente dispõem de conjuntos de dados extensos, necessitando de técnicas para facilitar o acesso a dados relevantes em uma vasta quantidade de informação. Neste contexto podem ser aplicadas técnicas para agrupamento de dados, agrupando objetos de acordo com sua relevância para encontrar padrões, para auxiliar na tomada de decisões estratégicas. Para se chegar a tais padrões, as técnicas utilizadas neste projeto foram: limpeza de dados; integração de dados; seleção de dados; transformação de dados; mineração de dados; avaliação de padrões; apresentação do conhecimento. Os experimentos foram realizados utilizando dados de candidatos inscritos em cursos do Ensino Médio Integrado do IFSP.

PALAVRAS-CHAVE: mineração de dados, agrupamento de dados, algoritmo k-modes, dados educacionais.

DATA CLUSTERING ANALYSIS TO DESCRIBE HIGH SCHOOL PROFESSIONAL STUDENTS OF FEDERAL INSTITUTE OF SÃO PAULO

ABSTRACT: Managers and educators nowadays have extensive data sets, requiring techniques to facilitate access to relevant data in a vast amount of information. In this context, data clustering techniques can be applied for grouping objects according to their relevance to find patterns, helping in making strategic decisions. To reach such patterns, the techniques used in this research project were: data cleaning; data integration; selection of data; data transformation; data mining; evaluation of standards; presentation of knowledge. The experiments were performed using database of candidates subscribed to High School Professional Courses of IFSP.

KEYWORDS: Data mining, data clustering, k-modes algorithm, educational data.

INTRODUÇÃO

O processamento manual de grandes conjuntos de dados é oneroso, lento e não confiável. Para auxiliar esse processo, faz-se necessário o uso de técnicas para estruturar, organizar e analisar os dados

disponíveis a fim de identificar padrões interessantes que apresentem informações não disponíveis em uma primeira análise. Neste sentido, a área de mineração de dados (do inglês, *Data Mining*, ou DM) pode ser vista como um resultado da evolução natural da tecnologia de informação, permitindo extrair conhecimento a partir do processamento de volumes abundantes de dados brutos (HAN; KAMBER; PEI, 2011).

De acordo com Han, Kamber e Pei (2011), as tarefas de mineração de dados são classificadas em duas categorias gerais: descritivas e preditivas. As tarefas descritivas caracterizam propriedades gerais dos dados, enquanto tarefas preditivas permitem realizar inferências sobre os dados. Dentre as tarefas descritivas, o agrupamento de dados é amplamente utilizado na literatura por permitir a descoberta de grupos de objetos significativos que muitas vezes não são identificados em análises superficiais.

Segundo Pires (2010), o crescimento dos Institutos Federais (IF) representou um avanço para a educação dos trabalhadores, principalmente, pela retomada da oferta dos cursos de Ensino Médio Integrado ao Técnico (EMI). Com o crescimento dos IF, o número de alunos atendidos também aumentou significativamente. Inúmeros trabalhos são encontrados na literatura com foco na análise do perfil de estudantes nos mais diversos níveis (PEREIRA; VIEIRA, 2013) (STEIMBACH, 2012) (FIGUEIRA, 2014). Entretanto, há uma carência de pesquisas aplicadas à análise de dados com foco no perfil de alunos matriculados para o processo seletivo de ensino médio, comparado ao perfil dos alunos aprovados nesse processo de seleção.

Esse trabalho tem como proposta aplicar agrupamento de dados para realizar uma análise do perfil de alunos inscritos em cursos do EMI do IFSP (Instituto Federal de São Paulo) versus alunos que foram aprovados no processo seletivo. Os atributos considerados na análise foram “Raça”, “Gênero”, “Renda” e “Origem”, sendo o último atributo utilizado para identificar se o aluno cursou o ensino básico em rede pública, particular ou mista. Os principais conceitos envolvidos, bem como a metodologia, os resultados obtidos e as conclusões são apresentados na sequência.

MATERIAL E MÉTODOS

AGRUPAMENTO DE DADOS

Agrupamento de Dados é a tarefa de agrupar os dados em classes ou *clusters* (conjuntos), onde os objetos dentro do *cluster* são semelhantes entre si, e diferentes em relação aos objetos de outro *cluster*. Dentre os principais algoritmos de agrupamento de dados disponíveis na literatura, é possível destacar o K-means (MACQUEEN, 1967), o K-medoid (KAUFMAN; ROUSSEEUW,1987) e o Kmedians (JAIN E DUBES, 1988).

Independente do algoritmo utilizado, os princípios básicos são aplicados à todas as técnicas. Dado um conjunto de n objetos, e k o número de grupos, os objetos são organizados em k partições, cada uma representando um *cluster*. Para isso, usa-se uma função de similaridade ou dissimilaridade baseada na distância, onde objetos do mesmo cluster são “semelhantes” e os de outro cluster são “dissimilares” (HAN; KAMBER; PEI, 2011).

Atualmente é possível encontrar na literatura uma grande variedade de algoritmos de agrupamento. Um dos principais algoritmos capaz de manipular dados categóricos na análise para definição dos clusters aos dados utilizados foi o K-modes (HUANG, 1998).

ALGORITMO K-MODES

O algoritmo de agrupamento K-modes (HUANG, 1998) é uma extensão do K-means. O Kmodes utiliza como centro do cluster as modas, que são os valores mais frequentes, sendo sempre atualizadas de modo a buscar a melhor representação dos dados. Na sequência será apresentada uma breve descrição das etapas do algoritmo. Mais detalhes sobre o funcionamento do algoritmo podem ser encontrados em Huang (1998).

1. Selecione k modas iniciais, uma para cada cluster.
2. Cada objeto é associado pelo algoritmo a um cluster cuja moda seja a mais semelhante.
3. Após cada associação de todos objetos a um cluster, as modas são atualizadas.

4. A dissimilaridade é calculada novamente de cada objeto para todos os clusters. Se um objeto for encontrado de modo que sua moda mais próxima pertença a outro cluster em vez de seu atual, realoca o objeto para esse cluster e atualiza as modas de ambos os clusters.
5. Repita as etapas 3 e 4 até que nenhum objeto tenha mudado de cluster após um teste de ciclo completo de todo o conjunto de dados, ou até que o número máximo de iterações seja alcançado.

Como os dados utilizados nos experimentos possuem atributos como "Raça", "Gênero", "Renda" e "Origem", que envolvem valores não numéricos, ou seja, categóricos, a etapa de mineração de dados considerou o algoritmo K-modes. Os resultados são discutidos na próxima seção.

RESULTADOS E DISCUSSÃO

Os dados utilizados nos experimentos são do processo seletivo do EMI aplicado no IFSP no ano de 2016. Os atributos considerados na análise foram "Raça", "Gênero", "Renda" e "Origem". Para a obtenção dos resultados foi utilizada a ferramenta R na execução do algoritmo K-modes. A Tabela 1 apresenta uma comparação entre inscritos no vestibular do EMI e os aprovados no IFSP no ano de 2016.

TABELA 1. Perfil dos Alunos Inscritos no Processo Seletivo vs. Aprovados no IFSP Ano 2016.

	Inscritos				**Aprovados**			
Cluster	**Raça**	**Gênero**	**Renda**	**Origem**	**Raça**	**Gênero**	**Renda**	**Origem**
C1	Branca	Masculino	1 a 2 SM	Escola Pública	Parda	Masculino	5 a 10 SM	Escola Pública
C2	Branca	Feminino	3 a 5 SM	Escola Pública	Branca	Masculino	3 a 5 SM	Escola Particular
C3	Branca	Masculino	2 a 3 SM	Escola Pública	Branca	Masculino	2 a 3 SM	Escola Pública
C4	Branca	Feminino	1 a 2 SM	Escola Pública	Branca	Feminino	3 a 5 SM	Escola Pública
C5	Parda	Feminino	1 a 2 SM	Escola Pública	Parda	Feminino	2 a 3 SM	Escola Pública
C6	Preta	Feminino	2 a 3 SM	Escola Pública	Parda	Masculino	2 a 3 SM	Escola Particular

Muitas análises são possíveis a partir dos resultados apresentados nas Tabela 1. Entretanto, as que merecem destaque inicial neste estudo são a com foco em Renda e Origem. É possível identificar que a procura pelo EMI é majoritariamente de pessoas com um perfil de renda baixo e com origem de escola pública. Entretanto, os alunos aprovados no processo seletivo são os de maior renda sendo, uma parcela significativa, de alunos com origem de escolas particulares.

CONCLUSÕES

As técnicas de agrupamento de dados aplicadas a dados educacionais permitem identificar grupos para descrever o perfil dos estudantes. Os resultados obtidos nos experimentos realizados neste trabalho indicam que os alunos de classes menos favorecidas financeiramente possuem uma formação básica de menor qualidade, que resulta na não aprovação no processo seletivo e também implicará em todo o processo de aprendizado educacional futuro destes alunos.

A expansão da rede técnica federal tem como parte do seu objetivo realizar inserção social de cidadãos pertencentes a classes menos favorecidas por meio de uma formação integradora e da produção do conhecimento. Entretanto, o estudo apresentado comprova que maioria pessoas, majoritariamente, de classes mais favorecidas possuem acesso à educação pública de qualidade, em especial EMI, mais facilmente que os de classes menos favorecidas.

Como trabalhos futuros serão realizados novos experimentos a partir de técnicas de extração de regras de associação. O objetivo é conseguir realizar o cruzamento dos resultados obtidos em ambas as técnicas de mineração de dados.

AGRADECIMENTOS

Agradecimentos ao Programa Institucional de Bolsas de Iniciação Científica e Tecnológica do IFSP (PIBIFSP) - Edital nº 011/2016 - por financiar esta pesquisa.

REFERÊNCIAS

FIGUEIRA, R. (2014) “Abordagem temática e a introdução de conteúdos de física moderna e contemporânea no ensino médio: uma primeira aproximação” Dissertação de Mestrado, UFSCar. São Carlos, 135 f.

HAN, J., KAMBER, M., PEI, J. (2011) "Data Mining: Concepts and Techniques" 3ª. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

HUANG, Z. (1998) "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values" Data Mining and Knowledge Discovery, n. 2, p. 283-304.

JAIN, A. K., DUBES, R. C. (1988) "Algorithms for clustering data mining". Prentice-Hall, Inc.

KAUFMAN, L., ROUSSEEUW, P. (1987) "Clustering by means of medoids" Statistical Data Analysis Based on the L1 Norm and Related Methods p. 405-416.

MACQUEEN, J. (1967) "Some methods for classification and analysis of multivariate observations" Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1 p. 281297.

PEREIRA, E. J., VIEIRA, N. J. (2013) "Os Estilos de Aprendizagem no Ensino Médio a partir do Novo ILS e a Sua Influência na Disciplina de Matemática", Revista de Educação em Ciência e Tecnologia, v.6, n.3, p.173-190.

PIRES, L. L. A. (2010) "Ensino médio e educação profissional: a consolidação nos Institutos Federais" Revista Retratos da Escola. Brasília, v. 4, n.7, p. 353-365.

STEIMBACH, A. A. (2012) "Juventude, escola e trabalho: razões da permanência e do abandono no Curso Técnico em Agropecuária Integrado" Dissertação de Mestrado UFPR, Curitiba, 127f.