Instituto Federal de Educação, Ciência e Tecnologia da Paraíba - IFPB
Programa de Pós Graduação em Tecnologia da Informação
Mestrado Profissional em Tecnologia da Informação

# Relatório Técnico

João Pessoa, 10 de março de 2020.

## Avaliação de técnica de clusterização para a identificação de grupos de discentes

Johnny Yuri Solano Marinho, IFPB
Damires Yluska Souza Fernandes, IFPB

***Resumo.*** *Diante do atual panorama educacional brasileiro e da utilização de abordagens ainda tradicionais de ensino, da larga presença da pedagogia bancária e de abordagens pedagógicas e didáticas uniformes em turmas heterogêneas, este trabalho realiza um estudo sobre a aplicação de técnica de clusterização para mineração de dados educacionais na tentativa de prover o ensino mais personalizado. Nessa perspectiva avalia-se o algoritmo K-Means de uma maneira experimental. Os resultados evidenciam que o agrupamento baseado neste algoritmo pode fornecer informações valiosas e favorecer suporte docente no planejamento pedagógico e didático na busca por um ensino mais personalizado.*

## 1. Introdução

Em nossa contemporaneidade, ainda é perceptível a adoção do modelo tradicional de ensino conforme descrito por Blikstein (2012), que relata um cenário alarmante com estudantes desmotivados somado a uma escola despreparada no tocante ao processo de ensino e aprendizagem. Esse modelo, historicamente adotado no Brasil, caracteriza-se pela educação de um para muitos onde, segundo Freire (1987), o professor detém o saber e centraliza o processo de ensino e aprendizagem. Nesse modelo, o conteúdo é abordado uniformemente em sala de aula. Tal condução do ensino, muitas vezes, ocorre sem o devido preparo e planejamento didático necessário. Outro fator que agrava tal cenário é o problema da heterogeneidade discente, em que constatam-se turmas que possuem estudantes com diferentes níveis acadêmicos, cognitivos, profissionais e etários. A heterogeneidade discente encontrada revela-se um dos grandes desafios enfrentados por muitos professores da modalidade que, muitas vezes, persistem no emprego das abordagens didáticas tradicionais. Deste modo a utilização de abordagens tradicionais em turmas diversificadas contrapõe ao modelo proposto por Perrenoud (2000a) que considera que essa heterogeneidade exige métodos complementares e, portanto, uma forma de inventividade didática e organizacional. Nesse sentido, consideram-se relevantes propostas administrativas, pedagógicas e didáticas, além de estratégias, que dinamizam as aulas a fim de prover auxílio às atividades docentes que promovam ensino mais personalizado e busquem melhores resultados acadêmicos.

Nessa perspectiva, o uso da Mineração de Dados Educacionais, também denominada como EDM (do inglês *Educational Data Mining*) é definida como subárea da mineração de dados, apresentando foco principal no desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais. Conforme Banker e Isotani (2011) a EDM pode proporcionar maior compreensão do cenário no qual os estudante estão inseridos, bem como a forma em que a aprendizagem ocorre. Tais informações

poderão oferecer suporte no planejamento docente, por exemplo, ao prover auxílio na tomada de decisões administrativas, pedagógicas e didáticas. Nesse panorama, este artigo apresenta uma avaliação experimental em uma turma de nível técnico em uma instituição pública de ensino superior, mediante aplicação de um método de clusterização de dados.

O relatório está estruturado como descrito a seguir. Na Seção 2 são introduzidos os conceitos e alguns trabalhos relacionados à temática. Na Seção 3 é abordada a metodologia utilizada. Descreve-se o experimento realizado na Seção 4, bem como uma análise de seus resultados na secção 5. Por fim, a Seção 6 apresenta as considerações deste trabalho e indica trabalhos futuros.

## 2. Fundamentação teórica e trabalhos relacionados

Nesta seção são apresentados alguns conceitos básicos e trabalhos relacionados à temática do experimento.

### 2.1 Conceitos básicos

Alguns conceitos norteadores são abordados sucintamente nesta seção. Tais conceitos compreendem a temática da mineração de dados educacionais, taxonomia Bloom e o processo de descoberta de conhecimento.

#### 2.1.1 Mineração de dados educacionais

Segundo Fayyad, Piatetsky-shapiro e Smyth (1996a), existem muitos dados presentes em bancos de dados espalhados pelo mundo, contudo, muitos estão na forma bruta e são considerados de pouco valor. Tais dados povoam

diversas áreas e contextos: governamentais, negócios, medicina, ciência, market, entre outras.

Nesse panorama possibilita-se a extração de informações úteis através da mineração de dados. Tais informações são consideradas de grande valor conforme Fayyad, Piatetsky-shapiro e Smyth (1996a). Nessa perspectiva os dados disponíveis em ambientes educacionais também são recursos em potencial que, quando aproveitados, podem gerar benefícios. Para Baker e Isotani (2011) a utilização de técnicas de mineração de dados educacionais visa mitigar problemas oriundos da educação. Romero e Ventura (2013) citam algumas aplicações da EDM em ambientes educacionais, conforme descritas na Tabela 1.

| | |
|---|---|
| Criação de alertas | Como forma de comunicação aos interessados no processo de aprendizagem, auxiliando administradores e educadores na tomada de decisão. |
| Manutenção e melhoria dos cursos | Envolve tarefas de pesquisa científica, construção de material didático, planejamento e programação, onde se deve analisar como ocorre a aprendizagem do estudante, verificando por exemplo, o que foi e como foi utilizado. |
| Geração de recomendação | Verifica as necessidades do estudante em um dado momento e gera uma recomendação, que pode lhe proporcionar aprofundamento em um determinado domínio ou auxiliá-lo em uma dúvida. |
| Previsão de notas e resultados de aprendizagem | Consiste em utilizar os dados de atividades do curso para prever as notas finais do estudante ou algum outro tipo de resultado de aprendizagem, como uma possível evasão ou futura capacidade de aprender algo. |
| Criação de perfis de estudantes (ou grupos de estudantes) | Utilizada para detectar o estado e as características dos estudantes, como a satisfação, motivação, progresso de aprendizagem, estilo de aprendizagem, preferências e assim por diante. |
| Análise da estrutura de domínio | Realizada para determinar a qualidade do conteúdo apresentado e a sequência em que ele foi dado, através da previsão do desempenho dos estudantes, descrevendo o domínio de instrução em termos de conceitos, habilidades, itens de aprendizagem e suas inter-relações. |

**Tabela 1 - Aplicações EDM**

Nesse contexto, pode-se aplicar e coexistir mais de uma aplicação de EDM para mitigar um ou mais de um problema educacional, por exemplo, em um ambiente virtual de aprendizagem pode-se: (1) criar alertas para os estudantes e (2) recomendar auxílio discente a grupos de estudo. Outro fator

importante remete ao cliente/usuário ou beneficiário da aplicação, que pode ser direcionada a(os):

- Estudante(s),
- Professore(s),
- Gestor(diretores, coordenadores e outros técnicos em assuntos educacionais).

Segundo Romero e Ventura (2013), a EDM está inserida em três principais áreas: Ciência da Computação, Educação e Estatística. Existe, também, o relacionamento de intersecção com algumas subáreas como: educação baseada em computador, aprendizado de máquina e análise da aprendizagem.

### 2.1.2 Taxonomia Bloom

A taxonomia Bloom foi criada na década de 50 por Benjamin Bloom. Ela compreende a classificação de 6 níveis cognitivos de acordo com Bloom et al. (1956). Descrevem-se os seis níveis a seguir, respectivamente, do nível mais superficial (conhecimento) de cognição ao nível mais complexo (avaliação):

- **Conhecimento:** Primeiro nível, mais superficial dentre os seis níveis. Neste nível, por exemplo, realizam-se ações docentes para testar se um estudante obteve informações específicas da lição: lembrar a informação, memorizar algo, rotular, reconhecer.
- **Compreensão:** Neste nível verifica-se se o estudante não somente capta a informação mas também compreende essa informação. Exemplos: entender um conceito, desenvolver sobre um conceito, exemplificar tal conceito, interpretar um fato.
- **Aplicação:** Na aplicação espera-se que os estudantes coloquem em prática algo, quando precisam aplicar ou usar o conhecimento que aprenderam. Exemplos: usar uma ferramenta, fazer algo, construir ou criar.

- **Análise:** Este nível apura a capacidade do estudante de investigar diante de um problema ou um fato no qual existem algumas variáveis . Exemplos: analisar tomadas de decisão, explicar um fato, investigar um problema.
- **Síntese**: Esse nível examina se o estudante é capaz de criar novas teorias ou fazer previsões. Exemplos: inventar algo, imaginar algo, criar alguma coisa, compor algo.
- **Avaliação:** Para Bloom et al. (1956) a avaliação é o nível mais abstrato, verifica-se se o estudante é capaz de realizar julgamentos. Exemplos: selecionar, julgar, debater, recomendar.

Cada nível cognitivo pode aparecer em uma avaliação de diversos modos através de sinônimos e verbos de significado semântico semelhantes que remetem ao um propósito avaliativo cognitivo. Porém avaliações comumente remetem aos níveis mais superficiais inferiores de cognição conforme descrito por Bloom et al. (1956).

### 2.1.3 KDD, Clusterização e k-means

Existem diversas metodologias para a utilização da mineração de dados, contudo, comumente, utiliza-se o processo de descoberta de conhecimento em bancos de dados, também conhecido pelo acrônimo KDD (do inglês *Knowledge Discovery in Databases*) Fayyad et al. (1996a). O KDD pode ser entendido como um processo de descoberta de conhecimento útil diante de uma base de dados bruta.

O KDD é constituído e fragmentado em algumas etapas (seleção de dados, pré-processamento, transformação, mineração e interpretação/avaliação) e organiza-se sequencialmente, porém comporta-se de forma interativa e iterativamente segundo Fayyad et al. (1996a).

Esse processo é descrito mais detalhadamente por Fayyad et al. (1996b). Primeiramente ocorre a seleção de dados, onde, nessa etapa

almeja-se compreender o domínio e selecionar os dados consistentes, válidos e adequados diante do objetivo proposto. Em seguida, o pré-processamento consiste na limpeza (eliminação dos ruídos) e tratamento dos dados ausentes ou incompatíveis. Na transformação ocorre a formatação ou conversão dos dados para a aplicação do algoritmo de mineração. A mineração de dados é a etapa na qual aplica-se alguma tarefa de mineração de dados (classificação, agrupamento, regras de associação, entre outras técnicas). Por fim é realizada a interpretação/avaliação dos resultados gerados a partir da etapa de mineração dos dados. Ressalta-se a dinamicidade e iteratividade do processo, no qual, pode-se voltar em fases anteriores para, por exemplo, redimensionar a base de dados selecionada ou aplicar outro algoritmo de mineração de dados.

Para a etapa de mineração de dados, são consideradas duas categorias de tarefas Fayyad et al. (1996a): (1) supervisionada e a (2)não supervisionada. As tarefas que utilizam dados previamente conhecidos estão enquadradas nas tarefas supervisionadas que possibilitam classificar dados ou predizer resultados. Neste segmento existem a classificação e regressão. Para o não supervisionado, existem as tarefas de associação e clusterização que processam os dados mesmo sem conhecimento prévio e sem rótulos predefinidos. Na clusterização busca-se a divisão de dados observando-se suas similaridades. Já na associação utiliza-se de correlações e regras para encontrar padrões. Existem diversos algoritmos associados às tarefas de mineração de dados para processar os dados e efetuar a mineração propriamente dita.

O k-means é um dos algoritmos de clusterização mais utilizado na mineração de dados educacionais conforme Dutt et al. (2017). O algoritmo k-means funciona da seguinte forma: (1) primeiramente define-se previamente o número de clusters que será utilizado no processamento do algoritmo; a partir do número de clusters é possível (2) distribuir randomicamente os locais onde os centróides de cada cluster serão alocados. Esses centróides são utilizados como ponto de referência para alocação dos demais elementos dos clusters que serão incluídos através da proximidade e da distância euclidiana

comum em cálculos realizados em cada iteração. (3) O terceiro passo do algoritmo é recalcular os centróides e reorganizar os elementos pertencentes aos clusters até chegar em uma condição de parada com base no cálculo da média das distâncias.

## 2.2 Trabalhos Relacionados

Diversos estudos sobre mineração de dados educacionais estão disponíveis em Dutt et al. (2017), trata-se de uma revisão sistemática referente a diversos estudos de EDM, em que, evidencia-se a notável utilização da clusterização e do algoritmo k-means, em especial, em pesquisas ao longo de três últimas décadas, o que fornece um panorama do que foi pesquisado nos últimos anos.

Estudos realizados por Pimentel et al. (2003), Correia e Pimentel (2012) e de França e do Amaral (2013) recorrem a EDM e utilizam como tarefa de mineração de dados a clusterização. Esses trabalhos, utilizaram o algoritmo k-means a fim de agrupar estudantes com características similares em ambientes escolares.

Pimentel et al. (2003) realizaram um estudo de caso e avaliaram mediante dados provenientes de questionários o conhecimento prévio de estudantes no qual os resultados desse estudo revelaram que técnicas de clusterização são bastante úteis para a formação de grupos homogêneos de aprendizes. Correia ae Pimentel (2012) utilizaram clusterização para a formação de turmas mais homogêneas para a aplicação da recuperação paralela com o intuito de atender à diversidade de ritmos de aprendizagem dos estudantes. Ao fazer uso da EDM em um estudo de caso, de França and do Amaral (2013) possibilitaram agrupar estudantes de uma turma a partir de dificuldades de aprendizagem encontradas nas atividades e avaliações ao longo de um semestre em uma turma de programação, para tal utilizou o mapeamento cognitivo através da taxonomia de bloom descrita por Wilson (2016).

Nesse sentido, esse estudo utiliza EDM para o agrupamento de diferentes grupos presentes em sala de aula, análogo aos estudos de (Pimentel *et al.* (2003), Correia e Pimentel (2012) e de França e do Amaral (2013). Contudo, neste estudo, ressalta-se como diferencial a utilização de dados acadêmicos, profissionais, cognitivos e etários a fim de verificar a eficácia do algoritmo k-means num processo de agrupamento dos estudantes em turmas diversificadas. Portanto, objetiva-se obter agrupamentos que propiciem informações valiosas para o planejamento docente, em concordância com Perrenoud (2000a). Vale ressaltar que este estudo integra uma das etapas iniciais de uma pesquisa que visa proporcionar recomendações pedagógicas ativas com base no perfil discente da turma.

## 3. Desenvolvimento/metodologia

Nesse estudo, optou-se pela utilização do processo KDD. As cinco etapas desse processo estão postas conforme a Figura 1. Primeiramente ocorre a definição e seleção dos dados, onde, nesse momento atenta-se à relevância dos dados, o que é abordado mais detalhadamente na *subseção 3.1,* em seguida geram-se dados sintéticos (*subseção 3.2*).

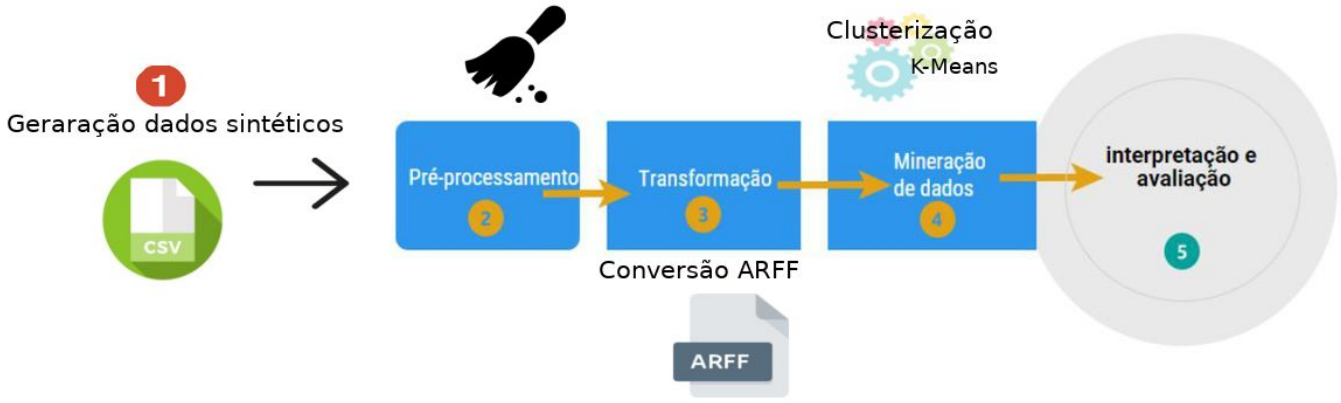


**Figura 1 - Processo de descoberta do conhecimento**

A etapa dois destina-se ao pré-processamento, descreve-se esse processo na *subsecção 3.3*. A partir da etapa dois os dados foram submetidos ao pré-processamento: categorização e normalização dos dados. Na etapa 3 ocorreu a

transformação dos dados(*subseção 3.4*). Na etapa 4 remete a tarefa de clusterização(*subseção 3.5*) para agrupar os dados a fim de encontrar padrões, por fim, na etapa 5, a interpretação e avaliação dos resultados que será abordada na secção 4.

## 3.1 Atributos relevantes

A Tabela 2 exibem-se os atributos: etaria, academica, profissionalArea, profissionalStatus, notaIngresso e na Tabela 3 exibem-se os atributos: q1, q2, q3, q4, q5 e q6. Tais atributos foram definidos e selecionados na fase inicial do processo KDD. No tocante à importância dos atributos para preparação e planejamento pedagógico docente destacam-se algumas considerações:

- **Lidar com níveis etários:** Proporciona ao docente selecionar e elaborar materiais com conteúdos mais adequados, no tocante à capacidade cognitiva descrita por Ausubel (2003), que ressalta os impactos no processo de ensino e aprendizagem que a inobservância deste item pode ocasionar. Enquadra-se nesse item o atributo *etaria*.

- **Atentar aos níveis acadêmicos (escolaridade) e área de atuação profissional:** Conhecimento prévio dos discentes são valiosos de acordo com Vygotsky (2000). Toda aprendizagem se processa de acordo com o contexto social em que o indivíduo está inserido. Nesse panorama a ideia de interação social e de mediação é ponto central do processo educativo para o autor. Tais ideias oportunizam o docente agrupar estudantes com mais experiências acadêmicas e profissionais com estudantes desprovidos ou imaturos nas competências de temáticas abordadas, a fim de proporcionar troca de conhecimento e benefícios no processo de ensino e aprendizagem. Enquadram-se nesse item os atributos: academica e profissionalArea.

- **Sensibilidade a ocupação ou trabalho:** A respeito do trabalhador-estudante ou estudante-trabalhador, eles comumente estudam em turnos noturnos e não podem frequentar a escola em outro turno pois estão exercendo alguma atividade profissional. Essa realidade é apresentada por de Paula e Vargas (2013) que relata as dificuldades e desafios vivenciados pelos trabalhador-estudante ou estudante-trabalhador. Nesse sentido o planejamento pedagógico no tocante à avaliação e na metodologia abordada em sala de aula revela-se um fator pertinente. Enquadra-se nesse item o atributo: profissionalStatus.

- **Visualizar níveis cognitivos em habilidades e conhecimentos:** Visualizar os estudantes considerados de alto nível ou defasados pode proporciona, por exemplo, a preparação e planejamento de conteúdos mais direcionados, elaboração de materiais extras para reforço, condução de estudantes considerados defasados para reforço extraclasse, adoção de abordagens metodológicas mais flexíveis, com vistas aos diferentes ritmos de aprendizagens existentes entre outras medidas. Nesse item, as habilidades e conhecimentos avaliados foram relacionados a: informática, produção textual, interpretação de texto e pesquisa. Enquadram-se nesse item os atributos: notaIngresso, q1, q2, q3, q4, q5 e q6.

Nesse sentido a partir de verbos cognitivos classificados pela taxonomia de Bloom(apresentada na seção 2.1.2), a fim de obter informações prévias dos discente no tocante à cognitividade em algumas competências.

Os atributos *q1, q2, q3, q4, q5 e q6* referem-se a competências relacionadas a cada estudante. Estão relacionadas da seguinte forma:

q1) Utilizar processadores de texto(Word, Writer, outros).

q2) Utilizar o computador para acessar sites, verificar emails.

q3) Utilizar planilhas eletrônicas (Excel, Calc, outros).

q4) Redigir um relatório das atividade desenvolvidas.

q5) Entender um texto acadêmico(Artigo, relatório).

q6) Pesquisar sobre um tema abordado em sala de aula.

### 3.2 Dataset

A fim de preencher os dados necessários para o estudo, gerou-se sinteticamente os dados, com base no estudo de atributos relevantes listados na *subsecção 3.1*. Posteriormente armazenou-se o conjunto de dados em formato CSV. Tais dados foram gerados de forma manual, visto que tratava-se de um dataset com apenas 30 instâncias e onze atributos.

### 3.3 Pré-processamento

Apresenta-se, na Tabela 2, o atributo idAluno que identifica unicamente um estudante. O atributo *academica* está associado à escolaridade (ensino médio, técnico, graduação, licenciatura, especialização e mestrado) do estudante. O atributo numérico *etaria* define a idade do estudante, que compreende um valor do tipo inteiro. A nota alcançada no processo seletivo de ingresso na instituição está atribuída à variável *notaIngresso* que comporta valores de 0 (zero) a 100 (cem). Os atributos *profissionalArea* e *profissionalStatus* estão associados, respectivamente, à área de atuação profissional (tecnologia da informação; administração; Engenharia; Comunicação; Saúde; Ciências Sociais e humanas; Arte e Design; outra;

nenhuma) e à situação profissional (empregado ou desempregado). Ressalta-se a utilização de valores nominais e numéricos durante o processo KDD. Na fase de transformação e formatação dos dados normalizou-se o atributo *profissionalStatus* que recebeu valores 0 para desempregado e 1 para empregado. Outro atributo que recebeu categorização foi o *academica* que categorizou-se para valores nominais: med, tec, lic, gra, mes. Também foi categorizado os valores associados à área profissional do estudante no atributo *profissionalArea* que receberam valores categóricos: TI, ADM, ENG, COM, SAU, CSH, ART, OUT, NEN conforme apresentado na Tabela 2.

| Atributos | Descrição | Tipo de dados | Domínio |
|---|---|---|---|
| idAluno | identificador do aluno | numerico | [1,n] |
| academica | escolaridade do aluno | Nominal | MED-Ens.Médio<br>TEC-Técnico<br>GRAD-Graduação<br>LIC-Licenciatura<br>ESP-Especialização<br>MET-Mestrado<br>OUT-OUTRO<br>?-EM BRANCO |
| etaria | idade do aluno | numerico | [1,n] |
| notaIngresso | nota da avaliação de ingresso | Nominal | [0,100] |
| pofissionalArea | Área de atuação profissional | Nominal | TI-tecnologia da informação<br>ADM-administração<br>ENG-Engenharia<br>COM-Comunicação<br>SAU-saúde<br>CSH-Ciências<br>Sociais e humanas<br>ART-Arte e Design<br>OUT-outra,<br>NEN-nenhuma |
| profissionalStatus | empregado(sim) desempre-gado(não) | nominal | 1 -SIM<br>0 -NÃO |

**Tabela 2 - Atributos relevantes**

Na Tabela 3 exibem-se atributos: q1, q2, q3, q4, q5 e q6 que foram gerados sinteticamente conforme descrito na subseção *3.2*. Tais atributos Identificam o nível da capacidade dos estudante em competências pré definidas na *subseção 3.1*.

O domínio desses atributos refere-se aos conceitos definidos e representam: A (excelente), B (ótimo), C(bom), D(insuficiente), E(baixo), F(ausência de conhecimento ou habilidade).

| Atributos | Descrição | Tipo de dados | Domínio |
|---|---|---|---|
| q1 | Nível cognitivo do item avaliado:aplicar (utilizar processador de texto) | nominal | A,B,C,D,E,F,? |
| q2 | Nível cognitivo do item avaliado:aplicar (acesso a internet) | nominal | A,B,C,D,E,F,? |
| q3 | Nível cognitivo do item avaliado:aplicar (utilizar planilhas eletrônicas) | nominal | A,B,C,D,E,F,? |
| q4 | Nível cognitivo do item avaliado:síntese (redigir um texto) | nominal | A,B,C,D,E,F,? |
| q5 | Nível cognitivo do item avaliado:compreensão (avaliar , interpretar) | nominal | A,B,C,D,E,F,? |
| q6 | Nível cognitivo do item avaliado:síntese (pesquisar temas) | nominal | A,B,C,D,E,F,? |

**Tabela 3 - Atributos mapeamento cognitivo mediante Taxonomia Bloom**

### 3.4 Transformação dos dados

Na etapa de transformação dos dados devido à utilização da ferramenta de mineração de dados *Weka (abordada na próxima secção)* e problemas de importação dos dados no arquivo no formato CSV(comma-separated-values), este foi convertido manualmente para o formato ARFF(Attribute-Relation File Format) conforme Figura 2, o que possibilitou a importação dos dados sem problemas de compatibilidade.

```
1
2   @relation edm
3
4
5   @attribute idAluno NUMERIC
6   @attribute academico { med,tec,gra,lic,esp,met,out,?}
7   @attribute etario NUMERIC
8   @attribute notaIngresso NUMERIC
9   @attribute profissionalArea {TI,ADM,ENG,COM,SAU,CSH,ART,OUT,NEN}
10  @attribute profissionalStatus {0,1}
11  @attribute q1 {A,B,C,D,E,F,?}
12  @attribute q2 {A,B,C,D,E,F,?}
13  @attribute q3 {A,B,C,D,E,F,?}
14  @attribute q4 {A,B,C,D,E,F,?}
15  @attribute q5 {A,B,C,D,E,F,?}
16  @attribute q6 {A,B,C,D,E,F,?}
17
18
19  @data
20  1,med,24,55,TI,1,A,A,A,B,B,C
21  2,med,40,50,ADM,0,B,A,B,A,A,B
22  3,met,30,70,ADM,0,A,A,A,A,A,A
23  4,med,19,90,NEN,0,A,C,A,B,A,B
24  5,gra,27,95,ADM,1,A,B,B,A,A,A
25  6,med,18,50,NEN,0,A,A,C,B,B,C
26  7,med,28,65,OUT,0,A,B,A,B,A,B
27  8,med,30,70,OUT,0,C,E,A,C,B,C
28  9,med,18,30,NEN,0,D,C,C,B,B,C
29  10,med,34,50,NEN,0,A,B,C,D,E,F
30  11,med,28,66,NEN,0,F,F,A,A,B,C
31  12,gra,28,90,OUT,1,B,B,A,A,C,C
32  13,lic,33,78,ART,0,A,B,C,C,C,F
33  14,med,21,80,NEN,0,A,A,A,B,B,C
34  15,med,29,68,OUT,0,B,C,D,A,A,F
35  16,med,19,56,ADM,1,B,A,A,C,A,A
36  17,med,20,55,?,0,F,C,F,C,C,F
```

**Figura 2 - Formato ARFF**

ARFF é um formato padrão utilizado no WEKA, segundo De Oliveira et. al, (2004) o ARFF possui a seguinte sintaxe: nome do dataset é setado em *@relation*, a definição do nome e do tipo dos atributos do dataset configura-se

em @*atribute*, a partir do @*data* armazena-se os dados(para cada linha uma instância) separa-se os atributos por vírgulas.

### 3.5 Metodo de aprendizado de máquina

Dentre os métodos de aprendizado de máquina existentes, optou-se pela utilização da clusterização. O algoritmo escolhido foi o K-means. A escolha da utilização do K-means levou em consideração os resultados gerados pelos trabalhos relacionados na Seção 2.2 e a revisão sistemática de Dutt et al. (2017) que aponta o k-means como algoritmo de clusterização mais utilizado no tocante a estudo realizados com EDM.

## 4. Resultados e discussão

Para a análise e exploração dos dados utilizou-se o software WEKA. Segundo Hall et al. (2009) essa ferramenta é largamente utilizada e está presente em diversos estudos de mineração de dados. A partir dos dados gerados sinteticamente experimentou-se dois cenários como demonstra-se a Tabela 4.

| Experimento | Propósito | Variáveis utilizadas |
|---|---|---|
| **Experimento 1** | Avaliar agrupamentos gerados, bem como a eficiência do algoritmo k-means no agrupamento. | etaria,profissionalArea, profissionalStatus,academica ,notaIngresso, q1, q2, q3, q4, q5 e q6 |
| **Experimento 2** | Avaliar o resultado da clusterização, depois de realizada a redução das variáveis. | etaria,profissionalArea, profissionalStatus,academica ,notaIngresso |

**Tabela 4. Experimentos realizados**

## 4.1 Experimento 1

No experimento realizou-se a variação no número de clusters, que foi configurado e testado de 2 até 10 clusters. Os resultados obtidos com 5 clusters se aproximaram da proposta do estudo (identificação de grupos heterogêneos).

Identificaram-se aspectos similares e distintos entre os clusters conforme Figura 3. A seguir alguma considerações acerca dos agrupamentos realizados pelo algoritmo *K-means*.

Cluster centroids:

| | | Cluster# | | | | |
|---|---|---|---|---|---|---|
| Attribute | Full Data | 0 | 1 | 2 | 3 | 4 |
| | (30) | (7) | (6) | (6) | (6) | (5) |
| academico | med | med | med | med | med | gra |
| etario | 27.6 | 21.2857 | 32.3333 | 30.8333 | 29 | 25.2 |
| notaIngresso | 68.4333 | 66 | 74.5 | 62.5 | 63.3333 | 77.8 |
| profissionalArea | ADM | NEN | ADM | TI | ADM | ADM |
| profissionalStatus | 0 | 0 | 0 | 0 | 0 | 1 |
| q1 | A | A | C | A | A | A |
| q2 | A | A | F | B | A | A |
| q3 | A | A | A | C | B | A |
| q4 | A | B | A | C | A | A |
| q5 | B | B | B | C | A | A |
| q6 | C | C | B | F | B | A |

**Figura 3 - Experimento um, utilizando k-means configurado para 5 clusters.**

Os cinco clusters foram formados, a princípio notam-se similaridades das médias etárias entre os clusters, que varia do cluster 0 de menor idade 21,3 até o cluster 1 que possui 32,3. Os demais clusters ficaram dentro desse escopo: cluster 2 com média etária 30.8, cluster 3 com 29 e cluster 4 com 25.2. Tais dados apontam, dentre os clusters analisados, para um aspecto etário mais homogêneo da turma.

Com vistas à análise dos aspectos profissionais, o atributo *profissionalArea* destacou dois clusters diferenciados. O cluster 0 setado como sem área de atuação (NEM) e o cluster 2 setado para tecnologia da informação (TI). Os demais clusters foram setados na área administrativa (ADM). Outro atributo vinculado ao escopo profissional, o *profissionalStatus*, obteve apenas o

cluster 4 setado como 1 (um), ou seja, estudantes encontram-se com alguma ocupação profissional. Portanto podemos concluir que os estudante pertencentes aos demais clusters estão definidos como desempregados.

No tocante ao atributo cognitivo *notaIngresso*, existiram similaridades entre os grupos, com exceção do cluster 4 e do cluster 1. A seguir algumas considerações sobre os resultados para cada cluster em relação aos atributos *q1, q2, q3, q4, q5 e q6* :

- Os estudantes pertencentes ao cluster 0 demonstraram competências com conceito A (excelente) para a utilização de Ferramentas processador de texto(q1), planilhas(q2) e navegação na Internet(q3). No tocante à síntese na produção textual(q4) e compreensão de texto(q5) os estudantes agrupados nesse cluster obtiveram conceito B e foram considerados ótimos. A análise de informações para um trabalho(q6) foi avaliada como bom pelos estudans desse grupo com conceito C.

- No tocante ao cluster 1 obteve conceito bom na utilização das ferramentas do tipo processador de texto(q1). Porém nunca utilizou a Internet para acessar sites e verificar emails(q2).A respeito da utilização de planilhas(q3) e síntese na produção de textual(q4), foram classificadas como Excelente. Compreensão de texto(q5) e Análise de informações foram avaliadas como ótimos(q6).

- A respeito do estudantes presentes no cluster 2, revelou-se excelência na utilização de ferramentas de processamento de texto(q1). Para a utilização da Internet(q2) foi considerada ótima. Utilização de planilhas(q3), síntese na Produção(q4) e interpretação de texto(q5) foram considerados bons. Outro ponto observado foi a análise de informações para um trabalho(q6), para os estudants desse cluster foi avaliada como uma tarefa nunca realizada.

- No cluster 3 obteve-se resultado excelente na utilização de ferramentas processaedor de texto(q1), navegação na Internet(q2). Os resultados relacionados à síntese na Produção de um textual(q4) e compreensão na interpretação de texto(q5) foram excelentes. Utilização de planilhas(q3) foi considerado ótimo. Análise de informações para um trabalho(q6) foi avaliada como ótimo pelos estudantes desse grupo.

- Por fim no cluster 4, que agrupou 5 estudantes (17 por cento do total), os resultados obtidos denotam um grupo diferenciado na turma, quando se analisa a porcentagem de trabalhadores (setado para 1, o atributo *profissionalStatus*, ou seja, encontra-se trabalhando), notas de *notasIngresso*(7.8) e distinção no nível acadêmico (grad) e cognitivo(A,A,A,A,A,A).

### 4.2 Experimento 2

No experimento dois, retiraram-se as variáveis *q1, q2, q3, q4, q5 e q6.* Conforme descrito na Tabela 3 essas variáveis estão relacionadas ao aspecto cognitivo dos estudantes em algumas competências. Objetiva-se com a redução do número de variáveis encontrar novos padrões nos agrupamentos gerados ou diminuição da taxas de erros. Os clusters foram formados conforme a *Figura 4.*

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 19.7462487337837
Missing values globally replaced with mean/mode

Cluster centroids:
                              Cluster#
Attribute            Full Data        0         1         2         3         4
                        (30)        (6)       (7)       (4)       (7)       (6)
=============================================================================================
academico                 med        med       med       med       med       gra
etario                   27.6    20.8333   29.2857     34.75   26.5714   28.8333
notaIngresso          68.4333       65.5   75.1429      60.5   56.1429   83.1667
profissionalArea          ADM        NEN       OUT       NEN       ADM       ADM
profissionalStatus          0          0         0         0         0         1
```

**Figura 4 - Experimento dois, utilizando k-means configurado para 5 clusters.**

Agrupou-se os estudantes em cinco clusters. O cluster 0 e o cluster 4 com 6 estudantes, o cluster 1 e o cluster 3 com 7 estudantes, por fim o cluster 2 com 4 estudantes. A princípio notam-se similaridades das médias etárias e quanto ao aspecto acadêmico nos resultados expostos no *Experimento 1.*

Notam-se similaridades com o Experimento 1. Uma das similaridades observadas, refere-se ao Cluster 4, que possui características diferenciadas dos demais grupos. Esse cluster contém estudantes com escolaridade de nível superior (GRA), que encontram-se empregados e com nota de ingresso elevada.Apesar das similaridades encontradas nos agrupamentos efetuados em relação ao Experimento 1, destaca-se o SSE, que é uma das métricas mais importantes ao agrupar elementos através da tarefa de clusterização. TAN et al. (2013) discorrem que o Sum of Squared Error (SSE), mediante cálculos matemáticos, fornece um panorama dos erros do conjunto de dados agrupados. Nesse contexto verificou-se que o SSE obtido no Experimento dois ( por volta de 19), em comparação ao Experimento um, que foi de 106, percebe-se uma relevante diferença que conforme TAN et al. (2013), quanto menor o valor do SSE mais eficiente realizou-se a clusterização. Contudo deve-se ter cuidado ao retirar ou adicionar atributos, apesar de ser uma ação comum em um processo interativo e iterativo (característica do KDD), tal ação pode mudar os objetivos traçados nas primeiras fase do processo KDD.

## 5. Considerações e trabalhos futuros

Diante dos problemas educacionais brasileiros apresentado por Blikstein (2012) e Freire (1987) e da heterogeneidade discente comumente presente em ambientes escolares conforme Perrenoud (2000a), busca-se um tratamento mais personalizado e aulas mais atrativas. Nesse sentido este estudo revela que a mineração de dados pode promover suporte docente na identificação de grupos diferenciados em sala de aula.

Esse trabalho apresentou um estudo com a técnica de clusterização baseada no algoritmo k-means, que permitiu agrupar discentes a partir de atributos cognitivos, etários, acadêmicos e profissionais. Ressalta-se a descoberta de um grupo diferenciado dos demais nos dois experimentos, o que possibilita o professor tomar decisões que favoreçam o processo de ensino e aprendizagem. Por exemplo, pode-se convidar estudantes mais especialistas para auxiliá-los como monitores em sala de aula conforme ideas de Vygotsky (2000) ou usar atividades avaliativas e desafios diferenciados e mais adequados aos níveis dos estudantes, utilizando, por exemplo, abordagens metodológicas ativas como a gamificação e sala de aula invertida. Não é possível concluir ainda se os resultados obtidos através da técnica do aprendizado de máquina não supervisionado utilizada neste estudo (clusterização) com k-means são significativos. Contudo, esses resultados podem ser utilizados para apoiar no planejamento docente individualmente, nas reuniões pedagógicas, tomada de decisão na formação de grupos mais homogêneos e no ensino mais personalizado, recomendação de suporte e reforço escolar a grupos considerados defasados ou desnivelados.

Como trabalhos futuros pretende-se: analisar outros algoritmos de clusterização, analisar outras tarefas de mineração de dados e por fim realizar recomendação de estratégias pedagógicas e didáticas diferenciadas e mais adequadas para um cenário escolar heterogêneo, dentro do contexto de metodologias ativas.

## 6. Referências

AUSUBEL, David P. Aquisição e retenção de conhecimentos: uma perspectiva cognitiva. **Lisboa: Plátano**, v. 1, 2003.

BAKER, Ryan; ISOTANI, Seiji; CARVALHO, Adriana. Mineração de dados educacionais: Oportunidades para o brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 02, p. 03, 2011.

BLIKSTEIN, Paulo. O mito do mau aluno e porque o Brasil pode ser o líder mundial de uma revolução educacional. 2012.

BLOOM, Benjamin S. et al. Taxonomy of educational objectives. Vol. 1: Cognitive domain. **New York: McKay**, p. 20-24, 1956.

CORREIA, Creusa Fernandes; PIMENTEL, Edson Pinheiro. Mineração de dados na formação de turmas para a recuperação paralela na educação básica. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. 2012.

DE FRANÇA, Rozelma Soares; DO AMARAL, Haroldo José Costa. Mineraçao de dados na identificaçao de grupos de estudantes com dificuldades de aprendizagemno ensino de programaçao. **RENOTE-Revista Novas Tecnologias na Educação**, v. 11, n. 1, 2013.

VARGAS, Hustana Maria; PAULA, Maria de Fátima Costa de. A inclusão do estudante-trabalhador e do trabalhador-estudante na educação superior: desafio público a ser enfrentado. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, v. 18, n. 2, p. 459-485, 2013.

DUTT, Ashish; ISMAIL, Maizatul Akmar; HERAWAN, Tutut. A systematic review on educational data mining. **Ieee Access**, 2017.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, 1996.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37-37, 1996.

FREIRE, Paulo. Pedagogia do oprimido. Rio de Janeiro: Paz e terra, 1987. _. **Pedagogia da autonomia**, 1974.

HALL, Mark et al. The WEKA data mining software: an update. **ACM SIGKDD explorations newsletter**, 2009.

NETO, Cipolla; BARRETO, Luis Silveira Menna; AFECHE, Solange Castro. A formação social da mente Vygotski, LS 153.65-V631 Psicologia e Pedagogia O desenvolvimento dos processos psicológicos superiores. **Psicologia**, 1998.

PERRENOUD, Philippe. A formação dos professores no século XXI. **PERRENOUD, Philippe; THURLER, Monica Gather, et al. As competências para ensinar no século XXI: a formação dos professores e o desafio da avaliação. Porto Alegre: Artmed. 2002a**, 2000.

PERRENOUD, Philippe. Pedagogia diferenciada. **Porto Alegre: Artmed**, 2000.

PIMENTEL, Edson P.; DE FRANÇA, Vilma F.; OMAR, Nizam. A identificação de grupos de aprendizes no ensino presencial utilizando técnicas de clusterização. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. 2003. p. 495-504.

ROMERO, Cristobal; VENTURA, Sebastian. Data mining in education. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 3, n. 1, p. 12-27, 2013.

VYGOTSKY, Lev Semenovich. A formação social da mente. brasileira. **São Paulo, Martins**, 1988.

WILSON, Leslie Owen. Anderson and Krathwohl–Bloom's taxonomy revised. **Understanding the New Version of Bloom's Taxonomy**, 2016.

ANTONENKO, Pavlo D.; TOY, Serkan; NIEDERHAUSER, Dale S. Using cluster analysis for data mining in educational technology research. **Educational Technology Research and Development**, v. 60, n. 3, p. 383-398, 2012.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. Data mining cluster analysis: basic concepts and algorithms. **Introduction to data mining**, p. 487-533, 2013.

DE OLIVEIRA, Aracele Garcia; GARCIA, Denise Ferreira. Mineração de Base de Dados de um Processo Seletivo Universitário. **INFOCOMP Journal of Computer Science**, v. 3, n. 2, p. 38-43, 2004.