UNIVERSIDADE DO EXTREMO SUL CATARINENSE - UNESC
CURSO DE CIÊNCIA DA COMPUTAÇÃO

ALINI MARANGONI EYNG

ANÁLISE DE AGRUPAMENTO PELOS MÉTODOS HIERÁRQUICO AGLOMERATIVO E PARTICIONAL *FUZZY* UTILIZADOS PARA *EDUCATIONAL DATA MINING* EM DADOS DE EDUCAÇÃO A DISTÂNCIA

CRICIÚMA
2019

ALINI MARANGONI EYNG

ANÁLISE DE AGRUPAMENTO PELOS MÉTODOS HIERÁRQUICO AGLOMERATIVO E PARTICIONAL *FUZZY* UTILIZADOS PARA *EDUCATIONAL DATA MINING* EM DADOS DE EDUCAÇÃO A DISTÂNCIA

Trabalho de Conclusão de Curso, apresentado para obtenção do grau de Bacharel no curso de Ciência da computação da Universidade do Extremo Sul Catarinense, UNESC.

Orientadora: Prof ª. Dra. Merisandra Côrtes de Mattos Garcia

CRICIÚMA

2019

ALINI MARANGONI EYNG

ANÁLISE DE AGRUPAMENTO PELOS MÉTODOS HIERÁRQUICO AGLOMERATIVO E PARTICIONAL FUZZY UTILIZADOS PARA EDUCATIONAL DATA MINING EM DADOS DE EDUCAÇÃO A DISTÂNCIA

Trabalho de Conclusão de Curso aprovado pela Banca Examinadora para obtenção do Grau de Bacharel, no Curso de Ciência da Computação da Universidade do Extremo Sul Catarinense, UNESC, com Linha de Pesquisa em Inteligência Artificial.

Criciúma, 28 de junho de 2019.

BANCA EXAMINADORA

Profa. Merisandra Côrtes de Mattos Garcia – Dra. – (UNESC) - Orientador

Profa. Graziela Fátima Giacomazzo – Dra. – (Programa de Pós-Graduação em Educação e Setor de Educação à Distância – UNESC)

Prof. Kristian Madeira – Dr.– (UNESC)

Prof. Luiz Ricardo Fieira – Me. – (ESUCRI)

A meus pais, que são meus exemplos de vida e inspiração.

AGRADECIMENTOS

Agradeço a meu pai Édio, por estar ao meu lado, incentivando e apoiando, independentemente da decisão a ser tomada. E a minha mãe Rosmeri, por sempre mostrar o caminho certo a ser percorrido, encorajando a nunca desistir, não importando a dificuldade. Nestes anos da graduação, vocês foram meus maiores incentivos.

Agradeço a minha vó Olivia, por ser uma excelente ouvinte, por passar os melhores ensinamentos e a minha irmã Eloisa, porque em qualquer momento da vida é quem me alegra e faz sorrir, vocês foram à alegria dos dias em que nada estava dando certo.

Agradeço ao meu esposo Diego, por pacientemente ter estado comigo em todos os momentos, foram dias até tarde, sábados e domingos ajudando a entender melhor o conteúdo, estudamos e aprendemos juntos. Sempre foi o meu maior companheiro e esteve presente em todas as etapas, durante estes anos.

Agradeço a Merisandra, pela sua dedicação como orientadora e professora, ensinando o melhor caminho, estando sempre presente e ajudando. Foi uma honra ser sua aluna e orientanda.

Agradeço imensamente a Laíne, minha colega de curso, que foi extremamente importante para esta pesquisa, compartilhando comigo todo seu conhecimento e ajudando naquelas em que foram as maiores dificuldades desta pesquisa.

Agradeço imensamente aos Setores de Educação a Distância (SEAD) e Tecnologia da Informação (TI), por terem nos disponibilizado a base e terem sido nossos parceiros.

Agradeço aos professores do curso de Ciência da Computação, por todo o conhecimento compartilhado.

E por fim, agradeço a todos os familiares e amigos que sempre torceram por mim e demonstraram todo o seu carinho.

“No meio da dificuldade encontra-se a oportunidade.”
Albert Einstein

RESUMO

O crescimento da tecnologia, faz com que a quantidade de dados em repositórios aumente, impossibilitando a análise por métodos tradicionais, surgindo à mineração de dados, aplicada por meio da descoberta de conhecimento. A educação gera dados relacionados a alunos, principalmente a educação à distância em que os dados são provenientes de um ambiente virtual de aprendizagem, se tornando uma área de interesse dos pesquisadores educacionais. Com isso, surge o *educational data mining*, que utiliza métodos da mineração de dados. Mediante as técnicas e tarefas de mineração, tem o agrupamento, que é dividido em agrupamento hierárquico aglomerativo e agrupamento particional. De modo que nesta pesquisa é realizada a comparação entre o algoritmo de AGNES para o agrupamento hierárquico aglomerativo e o algoritmo *fuzzy c-means* para o agrupamento particional, com o objetivo de identificar qual dos métodos possui melhor desempenho em dados educacionais. Os dados são provenientes da disciplina ministrada a distância de Introdução a Engenharia de Segurança do Trabalho, na Universidade do Extremo Sul Catarinense. A ferramenta R foi usada, por ser um software livre, para implementação dos algoritmos e métodos de validação. Ao iniciar a mineração, é necessário definir a distância da matriz de similaridade, em que é aplicado as distâncias *manhattan* e euclidiana em AGNES e *manhattan*, euclidiana, *correlattion* e *seuclidean* no *fuzzy* c-means*.* O algoritmo AGNES, precisa da identificação do método de conexão, para gerar os resultados, sendo aplicado teste com os métodos de *ward,* distância média, maior distância e menor distância. A verificação dos resultados apresentados pelos algoritmos é realizada por meio das medidas de qualidade, aplicando índices de validação. O modelo final definido para *fuzzy c-means*, foi o que aplica a matriz de similaridade *seuclidean* e para o AGNES o que tem a matriz de similaridade de *manhattan*, pelo método de conexão distância média. Comparando o resultado gerado pelo índice de *silhouette*, o agrupamento particional, foi definido como modelo final de agrupamento sobre os dados educacionais.

Palavras-chave: Educação a Distância, Educational Data Mining, Descoberta de Conhecimento, Agrupamento Hierárquico Aglomerativo, Agrupamento Particional.

ABSTRACT

The growth of technology makes the amount of data in repositories increase, making the analysis by traditional methods impossible and, as a consequence, data mining arises, applied through the discovery of knowledge. Education generates data related to students, especially distance education in which the data come from a virtual learning environment, becoming an area of interest for educational researchers. Consequently, educational data mining arises, using methods of data mining. Among mining techniques and tasks, there is the clustering, which is divided into agglomerative hierarchical clustering and partitional clustering. Therefore, the present research aims at making a comparison between AGNES algorithm for the agglomerative hierarchical clustering and the fuzzy c-means algorithm for the partitional clustering, in order to identify which of the methods has the best performance in relation to educational data. Data came from the subject Introduction to Occupational Safety Engineering taught in the distance format at the Universidade do Extremo Sul Catarinense (UNESC). As it is a free software the R tool was used for the implementation of algorithms and validation methods. When starting mining, it is necessary to define the distance of the similarity matrix, in which the Manhattan and Euclidean distances in AGNES are applied and the Manhattan, Euclidean, correlattion and seuclidean in the fuzzy c-means are applied. The AGNES algorithm needs identification of the connection method to generate the results and a test was applied with the methods of ward, average distance, longest distance and shortest distance. Results verification presented by the algorithms was carried out through quality measures, applying validation indexes of internal, external and relative criterion. The final model defined for fuzzy c-means was the one that applies the seuclidean similarity matrix and for the AGNES, the one that applies the manhattan similarity matrix by means of the medium distance connection method. Comparing the result generated by the silhouette index, the partitional clustering was defined as the final model of clustering in relation to educational data.

Key words: Distance Education, Educational Data Mining, Knowledge Discovery, Aglomerative Hierarchical Grouping, Partial Grouping.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS E SIGLAS

AGNES	Agglomerative Nested
AVA	Ambientes Virtuais de Aprendizagem
CPCC	Coeficiente de Correlação Cofenética
DBSCAN	Density Based Spatial Clustering of Applications with Noise
DIANA	DIvisive ANAlysis DIANA
EaD	Educação a Distância
EDM	Educacional Data Mining
FCM	Fuzzy C-Means
KDD	Knowledge Discovery in Databases
KS	Kolmogorov-Smirnov
PC	Partion coefficient
PE	Partion entropy
SEAD	Setor de Educação à distância
STING	STING
TI	Tecnologia da Informação
TIC	Tecnologias de Informação e Comunicação
UNESC	Universidade do Extremo Sul Catarinense

SUMÁRIO

1 INTRODUÇÃO

O crescimento no uso da *internet* para apoio de atividades educacionais tem gerado desafios para pesquisadores da área de Educação a Distância (EaD). Entre os desafios, está à forma em que deve ser realizada a manipulação dos dados, devido ao grande volume de informações sobre os alunos, que são armazenados nos repositórios de dados (GOTTARDO; KAESTNER; NORONHA, 2012).

Os recursos de EaD estão presentes em Ambientes Virtuais de Aprendizagem (AVA), sendo possível realizar o gerenciamento de um curso por meio do armazenamento de dados gerados pelo lançamento das notas, acompanhamento de frequência e interação entre o aluno, professor e os conteúdos ofertados (SOUZA et al., 2016). Neste contexto, o problema da EaD é associado a evasão e desempenho dos alunos, na qual a classificação é desafiadora, uma vez que o comportamento acadêmico depende de diversos fatores como variáveis pessoais, socioeconômicas, psicológicas e ambientais (SATYANARAYANA; NUCKOWSKI, 2016, tradução nossa).

O interesse na mineração de dados, para a área da educação, vem sendo ativamente desenvolvida nos últimos 20 anos (ATTO; KOTOVA, 2019). O *Educational Data Mining* (EDM), uma área proveniente da mineração de dados, foca-se no desenvolvimento de técnicas para coletar dados de uma base que seja procedente de um ambiente virtual de aprendizagem. A modalidade de EDM converte dados brutos, em informações úteis que podem ser utilizadas por professores, desenvolvedores, pesquisadores, entre outros (SILVA et al., 2015). Na área de EDM a tarefa de agrupamento, tem sido muito utilizada. Sendo aplicado para identificar grupos de alunos com características similares de aprendizado, com o propósito de acompanhar os alunos conforme suas características e competências (RAMOS et al., 2017).

O problema de agrupamento está voltado para a descoberta de grupos que são significativos no domínio de dados. O primeiro objetivo deste problema é encontrar instâncias que sejam semelhantes uma das outras. O segundo objetivo consiste em encontrar grupos separados espacialmente. Outra questão a respeito desse problema, é que cada grupo pode encontrar estruturas em diferentes níveis de refinação, dependendo dos valores de seus parâmetros (SOARES et al., 2008).

Desta forma, o agrupamento de dados é considerado um problema difícil, pois os *clusters* em dados podem ter tamanhos diferentes, além de não considerar exatamente quantos *clusters* devem ser formados (ZHENG; JIA, 2011, tradução nossa). O agrupamento é dividido em dois grupos principais, chamados de hierárquicos e de particionamento.

Para empregar o agrupamento hierárquico, é necessário que seu início se organize formando uma decomposição hierárquica em um conjunto de dados sendo representado por uma árvore, que o divide em subconjuntos menores, até que cada subconjunto seja formado por um objeto. Um dendrograma pode ser criado de duas formas, pela abordagem aglomerativa (*bottom-up*), no qual parte-se da folha de uma árvore para a raiz, ou pela abordagem divisiva (*top-down*), que parte da raiz de uma árvore para a folha (GOLDSCHMIDT; PASSOS; BEZERRA, 2015). No método de agrupamento hierárquico, o aglomerativo é o mais comum e usado para esse tipo de técnica (PANG-NING; STEINBACH; KUMAR, 2009, tradução nossa). O algoritmo hierárquico aglomerativo Agglomerative Nesting (AGNES), trabalha com o método de link único, na qual cada *cluster* pode ser representado por objetos de *cluster*. O objetivo do algoritmo é gerar repetidamente a fusão do *cluster* até que todo objeto se junte para formar um grupo (HUANG, 2015, tradução nossa).

O agrupamento não hierárquico ou de particionamento é definido como uma divisão de conjuntos de objetos de dados em subconjuntos, que não estão ligados e cada objeto de dados fica exatamente dentro do seu subconjunto (XU; WUNSCH, 2009, tradução nossa). Os algoritmos que são aplicados para agrupamento hierárquico, tentam minimizar determinados critérios de agrupamento (como uma função de erro quadrático), sendo tratados como problemas de aprimoramento (ZHENG; JIA, 2011, tradução nossa). Segundo Vengadeswaran e Balasundaram (2017, tradução nossa) vários algoritmos estão disponíveis no método de particionamento, sendo que os baseados na lógica *fuzzy* são úteis quando os grupos não estão bem separados e os limites são incertos, o que possibilita o descobrimento de relações mais sofisticadas entre um determinado objeto e os *clusters* apresentados (XU; WUNSCH, 2008, tradução nossa). Nos algoritmos de agrupamento *fuzzy*, o principal propósito é a divisão não exclusiva dos dados nos grupos. Os valores de dados são atribuídos com graus de pertinência a cada *cluster*. Esta técnica permite que o *cluster* cresça em sua forma natural, tendo-

se como um dos algoritmos o *fuzzy c-means* (MAHATME; BHOYAE, 2016, tradução nossa).

Esta pesquisa aplica a tarefa de agrupamento, por meio dos algoritmos AGNES para agrupamento hierárquico aglomerativo e *fuzzy c-means* para agrupamento particional, verificando qual tem melhores medidas de qualidade para o conjunto de dados da pesquisa na área de EDM, identificando qual o melhor modelo de agrupamento.

A aplicação da descoberta de conhecimento ocorre por meio de uma base de dados proveniente da disciplina de Introdução a Engenharia de Segurança do Trabalho, que é ministrada na modalidade à distância e ofertada pela Universidade do Extremo Sul Catarinense (UNESC).

1.1 OBJETIVO GERAL

Identificar um modelo de agrupamento em dados educacionais na modalidade à distância da disciplina de Introdução a Engenharia de Segurança do Trabalho da UNESC.

1.2 OBJETIVOS ESPECIFICOS

Os objetivos específicos desta pesquisa consistem em:

a) empregar o conceito de *Educational Data Mining*;
b) utilizar dados de interação do aluno com o ambiente de virtual de aprendizado na disciplina ministrada à distância de Introdução a Engenharia de Segurança do Trabalho;
c) comparar o agrupamento no conjunto de dados por meio dos métodos Hierárquicos e de particionamento;
d) aplicar os algoritmos AGNES (hierárquico aglomerativo) e *fuzzy c-means* (particional);
e) comparar por meio de medidas de qualidade em mineração de dados, grupos gerados por meio dos algoritmos de agrupamento hierárquico e de particionamento empregados na pesquisa.

1.3 JUSTIFICATIVA

É próspero o crescimento da utilização da tecnologia e seus meios, o que permite uma maior integração entre as instituições de ensino, discentes e docentes, pois os mesmos passaram a utilizar a tecnologia para produzir, compartilhar e administrar conteúdo didático, com o objetivo final de gerar e adquirir conhecimento, tanto na condição individual, como em grupo (SANTOS et al., 2016). Conforme dados do último censo da Educação Superior do MEC/Inep, o aumento do número de matrículas em cursos de graduação a distância, entre os anos de 2015 e 2016, foi o que mais cresceu no Brasil atingindo 20,0%, enquanto na educação presencial o valor é de 3,7%. Entretanto, a modalidade a distância para estudantes que concluíram o ensino superior no mesmo período diminuiu 1,3%, enquanto a educação presencial teve alta de 2,4%. Entre os motivos da evasão dos alunos, a falta de tempo tem sido a principal causa. Outros motivos apontados estão relacionados à questão financeira, adaptação na plataforma de EaD e a metodologia do curso (RAMOS et al., 2017). É muito importante para uma instituição, transformar a grande quantidade de dados, em conhecimento, pois pode ajudar professores e administradores na tomada de decisão. Além disso, também se pode constituir em avanço para a qualidade dos processos educacionais, fornecendo informações relevantes, para as diferentes partes interessadas (ASIF et al., 2017).

No entanto, na maioria dos sistemas de aprendizagem empregados pelas instituições, os sistemas são utilizados para a publicação e acesso de materiais em um determinado curso, não fornecendo aos educadores ferramentas para avaliar e rastrear as atividades realizadas pelos seus alunos e determinar de forma conveniente e eficaz o curso e o desenvolvimento dos seus discentes (DUTT; ISMAIL; HERAWAN, 2017, tradução nossa).

Empregar a *educational data mining* possibilita a definição de formas metodológicas que sirvam de modelos dos resultados de desempenho dos alunos, possibilitando ao professor e tutor a verificação da participação dos alunos, identificando os que estão em risco de reprovação ou evasão, e propor ações para recuperação (RAMOS, 2017). A maioria das técnicas tradicionais de mineração de dados, já foi utilizada com sucesso no campo educacional. Contudo, os sistemas educacionais têm características específicas que requerem um tratamento diferente

do problema de mineração de dados. Por consequência, fazendo com que a EDM seja um campo de pesquisa em desenvolvimento (ROMERO et al., 2010).

Dentre os fatores que levaram a escolha da aplicação de algoritmos de agrupamento, está o fato em que a análise de *cluster* tem sido empregada no contexto educacional, devido à necessidade dos pesquisadores da educação, em descobrir características comuns e distintas entre grupos de estudantes. A objeção é que existem vários algoritmos de agrupamento e poucas indicações sobre qual método escolher para análise dos dados. Embora existam estudos comparativos dos principais métodos de agrupamento, é perceptível a carência de estudos que relatem aplicações práticas que descrevam características de comparação entre um método hierárquico e não hierárquico (RAMOS et al., 2016).

Para aplicar o agrupamento hierárquico divisivo, é preciso considerar duas possíveis divisões de subconjuntos para um *cluster* com *N* pontos de dados, o que é computacionalmente limitado. Esta restrição torna a abordagem aglomerativa preferencial para utilização do agrupamento hierárquico (XU; WUNSCH, 2008, tradução nossa). A aplicação do agrupamento hierárquico é realizada pelo algoritmo AGNES, pois adota a estratégia aglomerativa para mesclar cluster (PANDE; SAMBARE; THAKRE, 2012, tradução nossa). Segundo Yamasari et al. (2016, tradução nossa), existem vários algoritmos que realizam o agrupamento do tipo particional. Os métodos básicos adotam a separação exclusiva do *cluster*, na qual cada objeto pertence somente a um grupo, a exceção são os algoritmos *fuzzy* (HAN; KAMBER; PEI, 2011, tradução nossa). Os dados para agrupamento envolvem o agrupamento de pontos de dados similares, entretanto um determinado cluster pode conter dados que pertencem a diferentes grupos (AL-SHAMMAA; ABBOD, 2015, tradução nossa). O agrupamento *fuzzy* será aplicado, pois enquanto os métodos convencionais reúnem cada ponto de dados em um único *cluster*, os algoritmos *fuzzy*, como o *fuzzy c-means*, pode agrupar um ponto de dados em dois ou mais clusters com diferentes graus de pertinência, gerando informações precisas (LIU; LUNG, 2011, tradução nossa).

A pesquisa será empregada na disciplina de Introdução a Engenharia de Segurança do Trabalho, pois é a primeira da área de conhecimento de ciências, engenharia e tecnologias ofertadas na modalidade à distância na UNESC.

1.4 ESTRUTURA DO TRABALHO

Esta pesquisa é composta em seis capítulos, sendo o Capítulo 1 a contextualização do tema proposto com a introdução, o objetivo geral e específico pretendido e a sua justificativa.

O Capítulo 2 aborda os principais conceitos relacionados à descoberta de conhecimento, tratando mais profundamente a mineração de dados e suas tarefas. Neste capitulo também é descrito as principais áreas de aplicação da mineração de dados.

A tarefa de agrupamento é tratada no Capítulo 3, mostrando os principais métodos, destacando o agrupamento particional e hierárquico.

Para o agrupamento particional é apresento o método *fuzzy*, bem como o funcionamento do algoritmo *fuzzy c-means* (FCM). Para o agrupamento hierárquico aglomerativo é apresentado o funcionamento do algoritmo *Agglomerative Nested* (AGNES). Ainda é demonstrado no capitulo as medidas de qualidades, que são aplicadas para realizar a validação do agrupamento.

O Capítulo 4 descreve a Educação a Distância (EaD), na qual é conceituado ambientes virtuais de aprendizagem (AVA), descrevendo o *educacional data mining* (EDM) como meio para tratar os dados.

O Capítulo 5 apresenta alguns trabalhos correlatos que usaram agrupamento, mineração de dados, *educational data mining* e os algoritmos *fuzzy c-means* e AGNES.

A base de dados, os processos para mineração pelos algoritmos AGNES e *fuzzy c-means,* e a apresentação dos resultados, mostrando o melhor modelo, conforme as medidas de validade, e discussão dos resultados, são apresentadas no Capítulo 6.

Finalizando, tem-se a conclusão da pesquisa e sugestão de trabalhos futuros.

2 DESCOBERTA DE CONHECIMENTO

O crescimento da tecnologia vem acontecendo de forma exponencial nos últimos anos, aumentando o volume das informações que são armazenadas nas bases de dados. A utilização de dados é importante, pois por meio deles é possível encontrar parâmetros importantes para diversas áreas e empresas. Grande parte das empresas nos dias atuais considera dados importantes para todas as decisões do negócio, por meio de bases com soluções encontradas a partir dos dados, existe a interferência nas estratégias de negócios (TALEB; SERHANI, 2017, tradução nossa). O uso disseminado da tecnologia e dos computadores vem resultando no acúmulo de grandes volumes de dados armazenados em sistema de nuvem, *data warehouses*, *data centers*, entre outros (KHOLOD, 2018, tradução nossa).

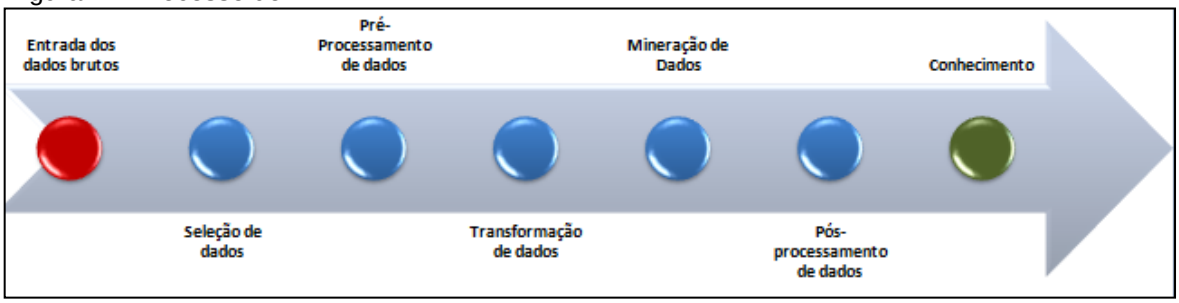
Em consequência, as ferramentas tradicionais usadas para o gerenciamento de dados, tornam-se insatisfatórias para análise dos dados (TAN; STEINBACH; KUMAR, 2009). Desta forma, surgem técnicas computacionais e ferramentas, que apoiam a extração do conhecimento útil em grandes bases de dados. Essas técnicas e ferramentas são dependentes do *Knowledge Discovery in Databases* (KDD) em base de dados e mineração de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, tradução nossa).

O KDD realiza a preparação dos dados para análise, separando em etapas iniciando do entendimento e preparação dos dados, partindo para a interpretação e utilização dos resultados do processamento de dados (LARA et al., 2014, tradução nossa)

O KDD aborda a metodologia geral da transformação de dados brutos em informações úteis, tendo como etapas do processo a preparação e seleção dos dados, aplicação dos algoritmos de mineração de dados, produzindo a leitura correta dos resultados (GOLDSCHMIDT; BEZERRA; PASSOS, 2015).

A descoberta de conhecimento, como processo, é representada na figura 1, inicia-se pela entrada dos dados brutos e as etapas são percorridas para que os dados se transformem em conhecimento útil.

Figura 1 – Processo do KDD.



Fonte: Do autor.

Os passos que estão expostos na figura 1 representam as técnicas do KDD e sua ordem. Os modelos são interativos, pois uma etapa depende do resultado da anterior, da mesma forma que são premissas para a seguinte. Entretanto, não é regra que todos os processos devam ser cumpridos, visto que os dados já podem estar com padrões necessários ou transformados, não necessitando, por exemplo, do pré-processamento (LIMA et al., 2017). Segue descrição das etapas do KDD, seguindo ordem iterativa do modelo:

a) **seleção de dados**: conhecida como redução de dados. É o processo que realiza a identificação de um subconjunto de dados que representam os dados iniciais sem perda de dados valiosos (LINCY; KUMAR, 2017, tradução nossa);

b) **pré-processamento de dados**: é uma etapa importante no processo de descoberta do conhecimento, visto que realiza a preparação dos dados para a mineração. Os dados são alterados para um formato que possibilite a utilização correta. Dados sem qualidade, não podem ser usados para avaliação, pois podem resultar em inconsistência e imprecisão (LINCY; KUMAR, 2017, tradução nossa). Os processos que fazem parte do pré-processamento são fusão e limpeza de dados, observações duplicadas e seleção de registros para mineração de dados (PANG-NING et. al, 2016, tradução nossa);

c) **transformação de dados**: é empregada a um conjunto de valores e dados (TAN; STEINBACH; KUMAR, 2009). Realiza o aprimoramento dos dados para destacar os que são interessantes para produzir conhecimento (LIMA et al., 2017, tradução nossa);

d) **mineração de dados**: consiste na aplicação de algoritmos, para descobrir informações que estão ocultas em meio a uma grande

quantidade de dados brutos. Mediante as técnicas, transforma dados em conhecimento útil (MADNI; ANWAR; SHAH, 2017, tradução nossa);

e) **pós-processamento**: esta etapa envolve a visualização, a análise e a interpretação do módulo de conhecimento gerado pela mineração de dados (GOLDSCHMIDT; BEZERRA; PASSOS, 2015). Os analistas verificam os resultados e indicam possibilidades novas para realizar a apuração dos dados.

2.1 MINERAÇÃO DE DADOS

A mineração de dados é fundamental nos dias atuais para pesquisa, realizando o processamento, análise e avaliação em bases de dados, estudando-as, para encontrar associações, classificações, agrupamento entre outros, para descobrir padrões, que transformam dados brutos em conhecimento útil (AROOJ; RIAZ; AKRAM, 2018, tradução nossa).

É uma etapa do processo de KDD, entretanto é constantemente empregada como sendo o processo de descoberta de conhecimento, isso porque é essencial para o reconhecimento de padrões (HAN; PEI; KAMBER, 2011, tradução nossa).

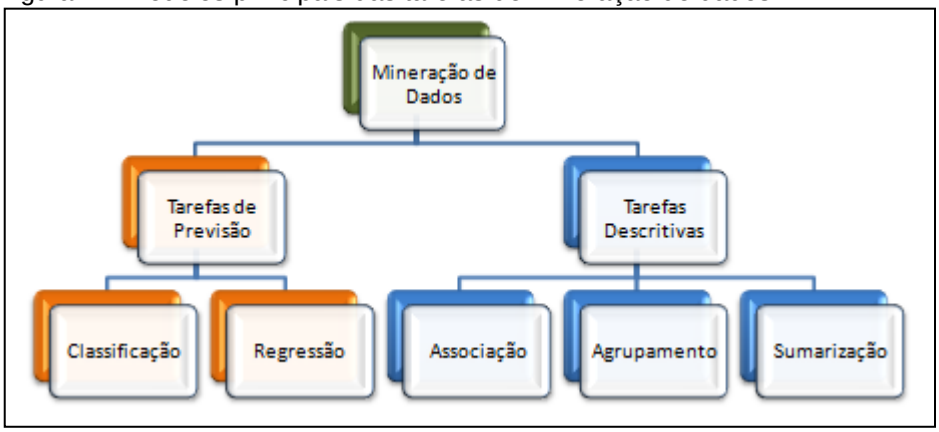
As técnicas de mineração de dados transformam dados em conhecimento útil (MADNI; ANWAR; SHAH, 2017, tradução nossa). É uma área multidisciplinar, que agrega sistemas de banco de dados com estatística, aprendizado de máquina e reconhecimento de padrões (ZAKI; MEIRA JUNIOR, 2014, tradução nossa).

Em decorrência do aumento dos dados nos últimos anos, a mineração de dados tem interessado a indústria da informação e a sociedade como um todo. O conhecimento adquirido pode ser usado para aplicações, tais como exploração de mercado, localização de fraudes e contenção de clientes, controle de produção e exploração científica. É visto como resultado da evolução natural da tecnologia da informação (HAN; KAMBER, 2006, tradução nossa).

A mineração de dados realiza a aplicação de algoritmos, na qual faz a extração de padrões de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, tradução nossa). Existem diversos métodos na mineração de dados que servem para minerar diferentes tipos de dados, sendo alguns deles regressão, classificação, agrupamento, entre outros (MOTTAGHI; KEYVANPOUR, 2017, tradução nossa).

Os métodos estão distribuídos em tarefas de mineração, que são divididas em dois modelos (figura 2). O primeiro modelo é referente às tarefas de previsão, na qual a proposta é presumir o valor de um determinado atributo, fundamentado nos valores de outros atributos. E o segundo modelo, são as tarefas descritivas, na qual o propósito é que os padrões sejam resultantes para que resumam os relacionamentos contidos dos dados (TAN; STEINBACH; KUMAR, 2009).

Figura 2 – Modelos principais das tarefas de mineração de dados.



Fonte: Do autor.

As tarefas preditivas são baseadas em classificação e regressão, enquanto as descritivas são as tarefas de associação, agrupamento e sumarização.

A classificação é uma das tarefas mais estudadas e pesquisadas em mineração de dados. Ela prevê o valor de um atributo tomando como base os valores de outros atributos (ROMERO et al., 2008). Possui o encargo de ordenar objetos em uma categoria, dentro de outras que são pré-selecionadas (TAN; STEINBACH; KUMAR, 2009). A tarefa de regressão é semelhante à classificação, sendo limitada a atributos contínuos. Engloba a busca por funções, que podem ou não ser linear, estruturando os registros de um conjunto de dados em padrões existentes (GOLDSCHMIDT; BEZERRA; PASSOS, 2015).

Nas tarefas descritivas, a associação “é útil para descobrir relacionamentos interessantes escondidos em conjuntos grandes de dados” (TAN; STEINBACH; KUMAR, 2009, p.389) A regra para associação prediz o atributo e a classe, possibilitando a autonomia de prever combinações de atributos (WITTEN; FRANK, 2005, tradução nossa).

O método de agrupamento é uma das áreas de mineração de dados mais relevantes e que está em rápido crescimento (BENDERSKAYA, 2017, tradução nossa). Realiza a análise de *cluster* para fazer a separação de objetos de dados em diferentes grupos, na qual os dados que são pertencentes ao mesmo grupo devem ser semelhantes entre si, diferentes dos objetos dos outros *clusters* (LI; LIU; 2018, tradução nossa).

A sumarização é conhecida como descrição de conceitos e apresenta as características que são comuns em um conjunto de dados, sendo que os mesmos devem apresentar ao menos uma característica em comum (GOLDSCHMIDT; BEZERRA; PASSOS, 2015).

2.1.1 Áreas de atuação da mineração de dados

A mineração de dados combina métodos tradicionais para extrair, analisar, transformar e modelar dados, com algoritmos que realizam o processamento em diversas bases de dados. A tecnologia que envolve a mineração é muito útil explorando dados e gerando conhecimento, o que facilita em tomadas de decisões (CAO; GUO, 2017, tradução nossa). Inúmeras áreas estão aplicando a mineração de dados, algumas em destaque:

a) **financeiro**: os bancos e instituições financeiras trabalham com um amplo conjunto de serviços, entre os quais, financeiro, seguros, investimentos e serviços de credito. Desta forma, os dados coletados são completos, confiáveis e de qualidade, o que facilita a análise e aplicação de mineração de dados (HAN; PEI; KAMBER, 2011, tradução nossa);
b) **medicina**: colabora na identificação de tratamentos eficientes e práticas que ajudam os pesquisadores a identificar características de doenças para diagnosticar e tratar (TAN et al., 2017, tradução nossa);
c) **indústria de varejo e telecomunicação**: o setor de varejo é uma área de aplicação em crescente crescimento, pois coleta e armazena dados sobre vendas, histórico de clientes, transporte, consumo e serviço. O crescimento dos dados coletados está expandindo rapidamente, devido ao crescimento de negócios realizados na *Web* em *e-commerce* (HAN; PEI; KAMBER, 2011, tradução nossa);

d) **educacional**: os dados são provenientes de bases de dados de escolas, universidades, entre outros. De forma que o EDM seja uma subcategoria da mineração de dados que lida com banco de dados educacionais e desenvolve técnicas para reconhecer padrões (JACOB et al., 2015, tradução nossa).

3 AGRUPAMENTO

A tarefa de agrupamento em mineração de dados tem como objetivo a criação de grupos (MARADITHAYA; HAREESHA, 2017, tradução nossa). É fundamental para a mineração de dados e aprendizado de máquina, estando presente em áreas como análise biológica, saúde, *marketing*, negócios e sistema de recomendação (LI; LIU, 2018, tradução nossa).

O agrupamento é considerado um problema de classificação não supervisionada, com o objetivo de partilhar ou dividir grupos, também chamados de *cluster* (PANDE; SAMBARE; THAKRE, 2012, tradução nossa). Os dados são divididos por meio de uma métrica de similaridade em *cluster*, na qual os dados que pertencem a um mesmo grupo possuem características comuns. Os algoritmos de agrupamento podem ter um desempenho diferente na aplicação, dependendo dos dados, por isso é necessário analisar e empregar o algoritmo que tem melhor desempenho para execução sobre uma base de dados, formando os grupos corretos e com maior similaridade. As medidas de similaridade para agrupamento determinam o grupo em que os objetos de dados pertencem, fazendo a comparação entre dois vetores, para determinar a igualdade e a diferença entre ambos. Geralmente é calculada tomando como base a medida de distância (PERES et al., 2012). Existem diversos métodos que realizam o cálculo para encontrar a medida de similaridade, dentre eles a euclidiana[1], *manhattan*[2], *maximum*[3], *minkowski*[4], *average*[5], *seuclidean* e *correlattion.*

O processo para o agrupamento inicia com a criação de um *cluster* ou grupo, sendo necessário ter no mínimo um objeto de dados em cada grupo. Cada

---

[1] É a distância mais utilizada, a euclidiana também conhecida como L2. Trabalha com dois pontos de dados *x* e *y* em um espaço bidimensional (XU; WUNSCH, 2008, tradução nossa).

[2] Além de chamada como distância de *manhattan*, também é conhecida como distância do bloco da cidade (GAN; M; WU, 2007, tradução nossa). Ela realiza a soma entre dois pontos mostrando a diferença absoluta de suas coordenadas cartesianas (ANYAIWE, 2017, tradução nossa).

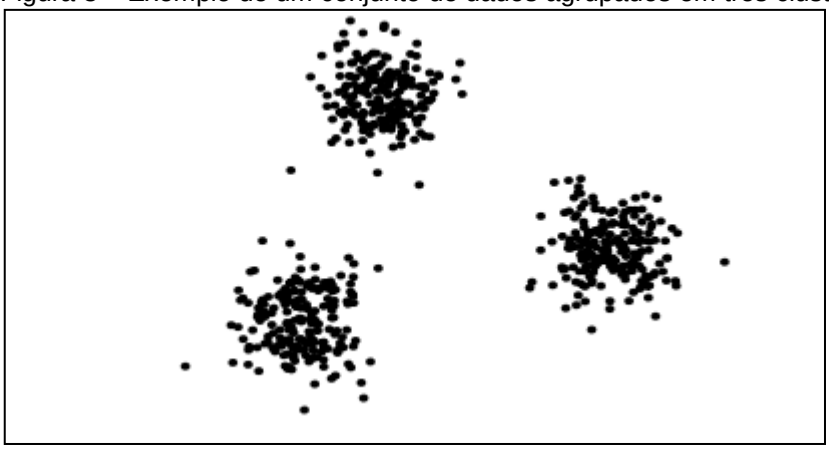
[3] É conhecida como distância *maximum*, devido ao cálculo que realiza para dois pontos de dados *x* e *y* em um espaço bidimensional, buscando o valor máximo entre as distâncias dos atributos (GAN; M; WU, 2007, tradução nossa).

[4] A distância *minkowski* é a soma das distâncias euclidiana, *manhattan* e *maximum*. É conhecida como o método geral, para realizar a medição das distâncias, realizando o cálculo bidimensional entre dois *clusters* (YANG, 2007, tradução nossa).

[5] A distância de *average* é baseada na distância euclidiana, sendo criada para suprir as duas desvantagens que ela possui, na qual dois pontos de dados sem valor, podem ter a mesma distância de um ponto de dados menor (GAN; M; WU, 2007, tradução nossa).

vez que um *cluster* é criado, o elemento com maior número de conexões é procurado e incluído nele. O processo é realizado conforme a maior conectividade de *cluster* em formação. A restrição na criação dos grupos, é que o tamanho do *cluster* não ultrapasse o valor especificado de *N*, que significa a quantidade de grupos. O algoritmo termina de realizar o processo, quando não existem mais elementos para agrupar (GAVRILOV et al., 2018, tradução nossa). A figura 3 mostra um conjunto de dados, agrupado em três *clusters*, formando os grupos conforme as características em comum dos dados.

Figura 3 – Exemplo de um conjunto de dados agrupados em três clusters



Fonte: Gana, Ma e Wu (2007).

Na análise de *cluster*, os objetos agrupados devem ser divididos em grupos *k* sem intersecção, obtendo características semelhantes entre si, de acordo com a métrica escolhida. Existem várias métricas e formas para que elas sejam calculadas (BENDERSKAYA, 2017, tradução nossa). O processo básico para iniciar a análise de *cluster* envolve quatro fases, representação de dados, modelagem, otimização e validação (GAN; MA; WU, 2007, tradução nossa). Quais são descritas como:

a) **representação de dados**: determina a estrutura de *cluster* que pode ser descoberta nos dados (GAN; MA; WU, 2007, tradução nossa). Não existem indicações na teoria que mostre modelos que são apropriados para a representação de dados. O processo é constituído em reunir fatos e hipóteses sobre os dados (JAIN; MURTY; FLYNN, 1999, tradução nossa);
b) **modelagem**: na etapa de KDD, é a parte que realiza a mineração dos dados. Consiste na escolha e aplicação das técnicas sobre os dados

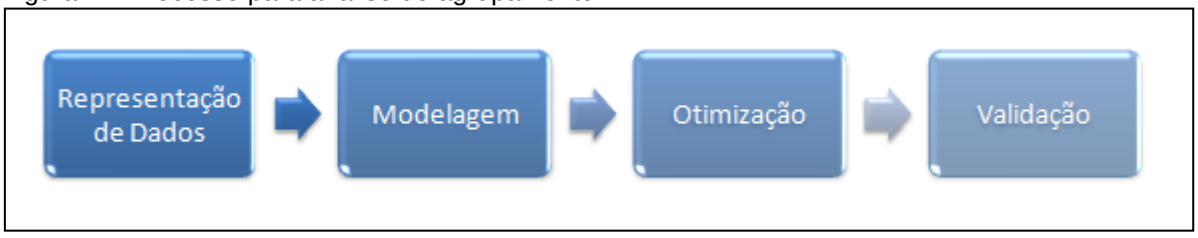
em que o algoritmo está sendo aplicado, baseado em testes e comparação com mais de um modelo de mineração (GOLDSCHMIDT; BEZERRA; PASSOS, 2015);

c) **otimização**: consiste no melhoramento de recursos limitados, potencializando as variáveis de resultado. Visto que a medida de proximidade é determinada, o agrupamento pode ser interpretado como um problema de otimização com uma função específica de critério (XU; WUNSCH, 2005, tradução nossa);

d) **validação**: consiste na avaliação dos *clusters,* atestando a qualidade dos mesmos, tendo como objetivo fornecer para o usuário um resultado de confiança. Os algoritmos de agrupamento produzem partições, sem considerar a existência de dados no *cluster*. Ainda, diversas abordagens de agrupamento levam para diferentes *clusters* de dados, a seleção de um parâmetro pode levar a um resultado final diferente, mesmo usando um mesmo algoritmo. Portanto, os critérios de avaliação são importantes para o resultado de agrupamento. (XU; WUNSCH, 2005, tradução nossa).

A figura 4 representa a ordem do processo para análise de *cluster*, iniciando com a representação dos dados, seguindo para modelagem e otimização, encerrando o processo com a validação.

Figura 4 – Processo para análise de agrupamento



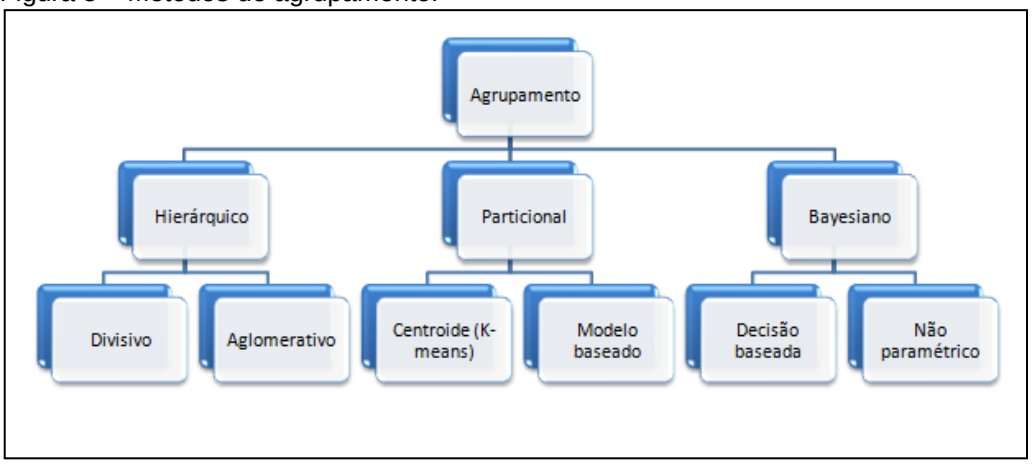
Fonte: Do autor.

Os algoritmos de agrupamento são divididos em *hard*, em que cada objeto de dados deve estar dentro de um cluster e *soft*, que representa a possibilidade de um objeto de dados pertencerem a diferentes *clusters* (GEIGER; AMJAD, 2017, tradução nossa).

Para realizar a técnica de agrupamento é preciso identificar o método de *cluster* apropriado (figura 5). Vários algoritmos estão disponíveis em cada método

para agrupar os conjuntos de dados, criando padrões para mineração dos dados (VENGADESWARAN; BALASUNDARAM, 2017, tradução nossa).

Figura 5 – Métodos de agrupamento.



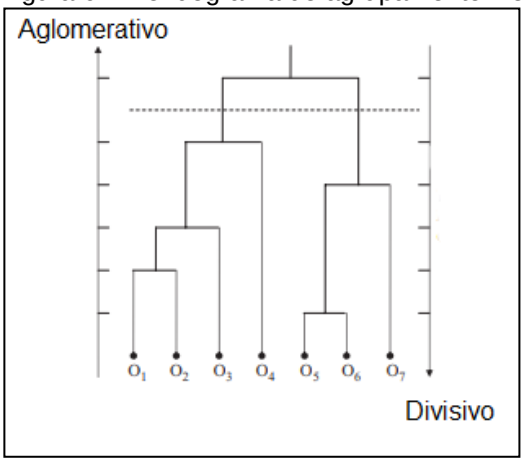
Fonte: Beyene (2017, tradução nossa).

O agrupamento pode ser classificado em diversos métodos, tendo como os mais conhecidos o particionamento, o hierárquico, os baseados em densidade, em grade e baseado em modelo (ARYAL; WANG, 2017, tradução nossa).

O método hierárquico agrupa os dados, formando uma decomposição hierárquica representada por um dendograma (figura 6). Cada bloco de dados é considerado, a princípio, um só *cluster*, que combina os blocos mais semelhantes para formar um maior. O algoritmo termina o processo, quando todos os blocos são mesclados em um só *cluster* (VENGADESWARAN; BALASUNDARAM, 2017, tradução nossa). Um dendograma pode ser formado pelo agrupamento hierárquico aglomerativo ou divisivo (GOLDSCHMIDT; BEZERRA; PASSOS, 2015). O algoritmo DIvisive ANAlysis[6] (DIANA) é um exemplo baseado no método hierárquico.

[6] O algoritmo DIANA segue por meio de uma série de divisões consecutivas, na qual o maior aglomerado é dividido até que no passo *n-1*, cada dado esteja em um único grupo. Ele é um algoritmo hierárquico divisivo (GAN; MA; WU, 2007, tradução nossa).

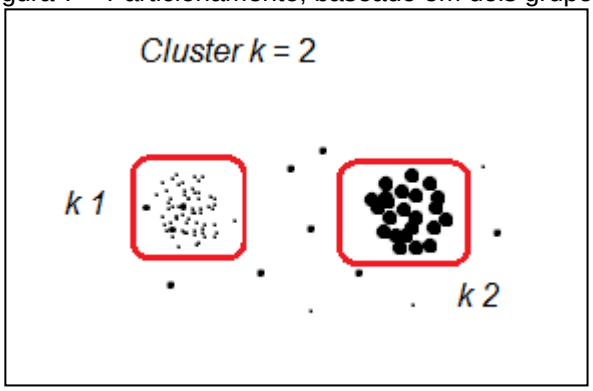
Figura 6 – Dendograma de agrupamento hierárquico



Fonte: Xu e Wunsch (2008, tradução nossa).

Os algoritmos de particionamento também são conhecidos como não hierárquicos (RAMOS, 2016). De acordo com Tan, Steinbach e Kumar (2009, p. 587) "é simplesmente uma divisão do conjunto de objetos de dados em um subconjunto (grupos) não interseccionados". O agrupamento de algoritmos de particionamento atribui em um conjunto de pontos de dados em *k clusters*, sem nenhum tipo de hierarquia (XU; WUNSCH, 2008, tradução nossa). O *K-modes*[7] é um dos algoritmos baseado no método de particionamento. A figura 7 mostra uma partição de *k*=2, na qual os dados são divididos em dois grupos.

Figura 7 – Particionamento, baseado em dois grupos



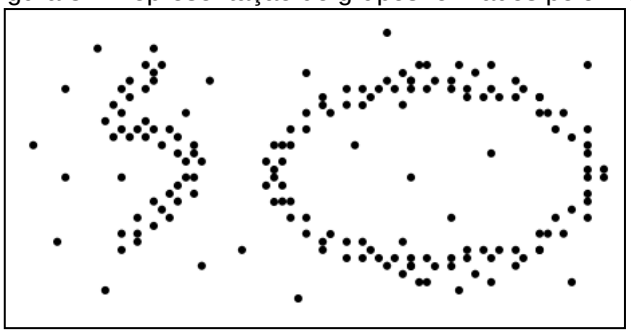
Fonte: Do autor.

Os algoritmos de densidade são indicados para formar grupos em forma arbitrária (figura 8) (HAN; PEI; KAMBER; 2011, tradução nossa), agrupando objetos de acordo com funções objetivas de densidade. Neste modelo, o *cluster* continua

[7] É uma modificação do *K-means*, usado para dados com recursos nominal ou ordinal, pois o algoritmo *K-means* é propicio apenas para tipos de dados numéricos (FAHMI et al., 2016, tradução nossa).

crescendo desde que o número de objetos de *cluster* vizinhos exceda o parâmetro definido (MOTTAGHI; KEYVANPOUR, 2017, tradução nossa). O *DensityBased Spatial Clustering of Applications with Noise*[8] (DBSCAN) é um dos algoritmos baseado no método de densidade.

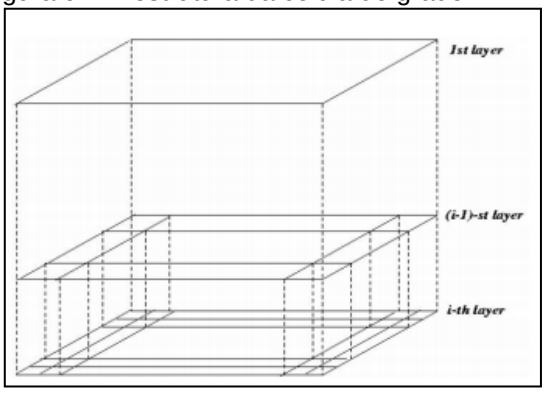
Figura 8 – Representação de grupos formados pelo método de densidade.



Fonte: Han, Pei e Kamber (2011).

O agrupamento por grade trabalha com uma estrutura de multirresolução, gerando uma estrutura de grades, que realiza toda a estrutura do agrupamento (HAN, PEI, KAMBER, 2011, tradução nossa). É baseado em estrutura de grades multinível (figura 9), de modo que o nível superior armazena as informações do próximo nível (PANDE; THAKRE; SAMBARE, 2012, tradução nossa). O *STatistical INformation Grid*[9] (STING) é um dos algoritmos baseado no método de grades.

Figura 9 – A estrutura da célula de grade.



Fonte: Pande, Thakre e Sambare (2012)

[8] O DBSCAN descobre *clusters* de forma arbitrária, sendo necessário apenas um parâmetro de entrada, o usuário pode determinar também o valor apropriado para este parâmetro de entrada (GAN; MA; WU, 2007, tradução nossa).

[9] No algoritmo de STING, serve para facilitar tipos de consultas espaciais, na qual a área espacial é dividida em células retangulares, representada por uma estrutura hierárquica. No STING cada celular tem quatro filhos e cada um corresponde a um quadrante da célula pai. Apenas o espaço bidimensional é considerado (GAN; MA; WU, 2007, tradução nossa).

No método de agrupamento baseado em modelo, os dados são provenientes de um conjunto de probabilidade e distribuição, de forma que cada conjunto representa um *cluster* diferente. Os dados são gerados por uma combinação de distribuição de probabilidade e distribuição, na qual cada parte representa um cluster que é diferente de outro (GAN; MA; WU, 2007, tradução nossa). O STUCCO[10] é um dos algoritmos baseado no método de grades.

Os métodos principais de agrupamento são os hierárquicos e de particionamento (VENGADESWARAN; BALASUNDARAM, 2017, tradução nossa).

Nesta pesquisa, são aplicados algoritmos de agrupamento hierárquico e de particionamento, de modo que seja necessário um entendimento mais específico para ambos os métodos nos subcapítulos subsequentes.

3.1 AGRUPAMENTO PARTICIONAL

Os algoritmos de particionamento geram divisões após a aplicação de critérios. São igualmente conhecidos como algoritmos de agrupamento não hierárquico, de modo que apenas um grupo de dados forma a saída de um algoritmo de agrupamento. Neste caso, o usuário informa o número de grupos ou *clusters* que devem ser formados (BEYENE, 2017, tradução nossa). Cada partição corresponde a um *cluster* específico (ONAN, 2017, tradução nossa). O k corresponde às partições de dados e n aos grupos, na qual $k \leq n$. O método de particionamento classifica os dados em k grupos, seguindo os seguintes critérios: cada grupo precisa ter no mínimo um objeto de dados e cada objeto deve pertencer a apenas um grupo (HAN; KAMBER, 2006, tradução nossa).

A análise de *cluster* busca uma partição em que os dados sejam altamente semelhantes, apresentando a mesma estrutura enquanto são devidamente separados de outros grupos. A homogeneidade e a separação são classificadas por meio de funções de critérios, sendo que a função de critério mais conhecida e usada é a soma do erro quadrado (XU; WUNSCH, 2012, tradução nossa).

---

[10] No STUCCO um novo conceito de conjunto de contraste é definido, para isso um método de busca é usado para calcular todas as combinações possíveis dos valores de um atributo, em seguida pós-processa esse conjunto de dados, selecionando um subconjunto.

Porém, existem circunstâncias na qual os *clusters* não estão bem separados e os limites são ambíguos, para estes casos são empregados algoritmos de agrupamento *fuzzy* (XU; WUNSCH, 2008, tradução nossa).

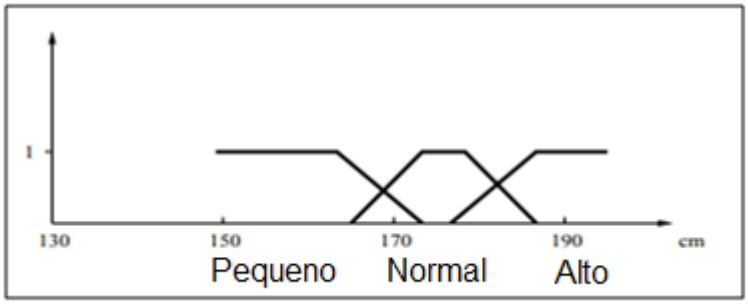
3.1.1 Agrupamento método fuzzy

Os algoritmos de agrupamento *fuzzy*, realizam o particionamento não exclusivo de dados na formação do *cluster*. A lógica do algoritmo fuzzy consiste no fato de que o *cluster* evolua em sua forma natural (MAHATME; BHOYAR, 2016, tradução nossa). As técnicas para agrupamento são definidas como *hard*, correspondendo ao método em que cada objeto de dados é atribuído a apenas um *cluster*. Entretanto a exceção está em algoritmos de agrupamento *fuzzy*, na qual um objeto de dados pode pertencer a todos os *clusters* que tem semelhança. Sendo bastante útil, quando os limites entre os *clusters* são incertos e não estão claramente separados (XU; WUNSCH, 2008, tradução nossa).

O agrupamento é representado por uma função de associação $z = (z_i)$, na qual *i* pertencente ao conjunto *I* em que z_i $(0 \leq z_i \leq 1)$ é interpretado como o grau de associação de *i* para o *cluster*. A associação z_i é restrita apenas aos valores 1 ou 0. O conjunto de *cluster* de um agrupamento *fuzzy* $(zk = (zik), k = 1, ..., K)$, forma uma partição de dados na qual o grau total de membros para uma entidade *i* pertencente a *I* é 1, $P \sum k\ zik = 1$. Em um método particional *fuzzy,* o grau de filiação de uma entidade é 1, pode pertencer a diferentes *cluster*. Ao dividir um grupo de pessoas pela sua altura pode envolver conjuntos *fuzzy* (MIRKIN, 2005, tradução nossa). A figura 10 apresenta um conjunto *fuzzy* realizando a função de associação para os conceitos pequeno, normal e alto, na altura de um grupo de pessoas.

.

Figura 10 – Conjuntos Fuzzy.



Fonte: Mirkin (2005, tradução nossa).

O grau de pertinência varia entre o valor mínimo representado por 0 e o valor máximo, representado por 1. Desta forma, os objetos de dados que estão no centro do *cluster* atingem o valor máximo, enquanto que para os objetos de dados que estão afastados do centro do *cluster* o valor vai reduzindo, quanto maior a distância menor será o valor da associação (MAHATME; BHOYAR, 2016, tradução nossa).

3.1.2 Algoritmo fuzzy c-means

O *fuzzy c-means* (FCM) é empregado em áreas como análise de imagem, medicina, astronomia, química, entre outros. O cálculo do algoritmo é dividido em duas partes, a primeira calcula os centros do *cluster* e a segunda atribui pontos de dados aos *clusters* por meio de um cálculo que é utilizado para medida de distância. O FCM é parecido com o *K-means*, considerando que dados devem ser separados em *k cluster,* conforme a similaridade dos dados, o que difere é que ele incorpora o conceito *fuzzy*, em que um dado pode pertencer a mais de um grupo (MAHATME; BHOYAR, 2016, tradução).

Para iniciar o algoritmo, é necessário possuir a pseudoparticipação difusa, que é definido por meio de um conjunto de dados, *x = {x1, x2. ..., xm},* de maneira que para cada xi, tenha um ponto de n dimensões, *i.e., xi = {xi1,...,xin}*. Um conjunto de grupos difusos, *C1, C2, C3, ... Ck*, seja um subconjunto de todos os conjuntos difusos previsto de *x*. Ou seja, para os pesos de pertinência (w_{ij}), são atribuídos valores que estão entre 0 e 1, para cada ponto *xi* e grupo Cj (TAN; STEINBACH, KUMAR, 2009).

De modo que, o FCM garante que um objeto de dados possa pertencer a todos os *cluster* com adesão entre 0 e 1 (KAPOOR; SINGHAL, 2017, tradução nossa). As etapas de atualização do algoritmo FCM, são apresentadas na tabela 1.

Tabela 1 – Etapas para cálculo dos grupos pelo algoritmo fuzzy c-means.

Algoritmo *Fuzzy C-means Difuso Básico*	
Etapa	**Descrição**
1	Selecionar uma pseudoparticipação *fuzzy* inicial, i.e., atribua valores a todos os W_{ij}
2	Repita
3	Calcule o centroide de cada grupo usando a pseudoparticipação *fuzzy.*
4	Recalcule a pseudoparticipação *fuzzy* inicial, i.e., e os W_{ij}.
5	Até que os centróides não mudem.
(Condições alternativas de parada são “se a mudança no erro estiver abaixo de um limite especificado” ou “se a mudança absoluta em algum $\boldsymbol{W_{ij}}$estiver abaixo de um determinado limite”).	

Fonte: Tan, Steinback e Kumar (2009).

A inicialização do algoritmo é aleatória, visto que o número de entradas de grupos pode ser incerto. O passo seguinte é o cálculo do centroide de cada grupo, de modo que para cada grupo *Cj*, o centroide condizente (*Cj*) é proposto pela fórmula a seguir:

$$cj = \sum_{i=1}^{m} w_{ij}^{p}\ x_i / \sum_{i=1}^{m} w_{ij}^{p} \qquad (1)$$

Para a definição do centroide no *fuzzy c-means,* todos os pontos são considerados, ou seja, o ponto pode pertencer a mais de um grupo. A tabela 2 mostra os elementos da equação, para cálculo do centroide.

Tabela 2 - Elementos da equação para cálculo dos centroides.

Elemento	**Definição**
$\boldsymbol{cj}$	Centroide correspondente para determinado grupo.
$\boldsymbol{m}$	Parâmetro que controla o quanto o *cluster* será difuso.
$\boldsymbol{p}$	Valor inserido na equação, que por parâmetro é iniciado com 2, porém quanto maior mais difuso se torna o *cluster*.
$\boldsymbol{x_{ij}}$	Ponte de dados.
$\boldsymbol{w_{ij}}$	Pesos de pertinência, de modo que os valores atribuídos devem ficar entre 0 e 1.

Fonte: Tan, Steinback e Kumar (2009).

A atualização da pseudoparticipação é apresentada a seguir, considerando que p = 2.

$$w_{ij} = 1/dist\ (xi, cj)^2 / \sum_{q=1}^{k} 1/dist\ (xi, cq)^2 \quad (2)$$

O peso de w_{ij} indica que a participação do ponto x_i, no grupo c_j, que apresenta valor alto se x_i estiver próximo ao centroide de c_j e baixa se estiver longe. De modo que, para a atualização da pseudoparticipação é indicado que os dados sejam normatizados, porque a soma do peso da participação do ponto não terá a soma final como um (TAN; STEINBACH, KUMAR, 2009). Os elementos da equação para calculo de atualização das etapas de atualização do pseudoparticipação estão descritas na tabela 3.

Tabela 3 - Elementos do cálculo de atualização dos pseudoparticipação.

Elemento	Definição
$\boldsymbol{w_{ij}}$	Indica o grau de participação do ponto x_i, no grupo c_j
$(\boldsymbol{xi, cj})$	Ponte de dados x_i e grupos c_j.
$\boldsymbol{dist}$	Distância entre x_i e c_j.
$\sum_{\boldsymbol{q=1}}^{\boldsymbol{k}} \mathbf{1}$	Soma dos pesos X1, que devem somar 1.

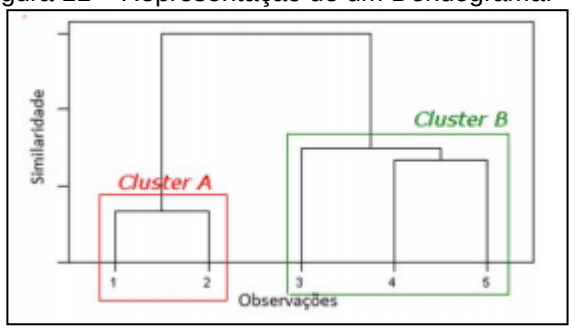
Fonte: Tan, Steinback e Kumar (2009).

O cálculo para encontrar o grau de associação, depende da medida de distância escolhida (FUNG, 2001, tradução nossa). Para a pesquisa, foram selecionadas as distâncias *euclidiana*, *manhattan*, *seuclidean* e *correlattion*.

3.2 AGRUPAMENTO HIERÁRQUICO

Os métodos de agrupamento hierárquico consistem em uma sequência de agrupamentos ou divisões de elementos consecutivas, tendo em vista que os elementos são agregados ou desagregados, objetivando construir uma hierarquia de *cluster*. Desta forma, o agrupamento hierárquico gera uma árvore de *cluster* para mostrar o resultado, conhecido como de dendograma (figura 11) (RAMOS et al., 2016).

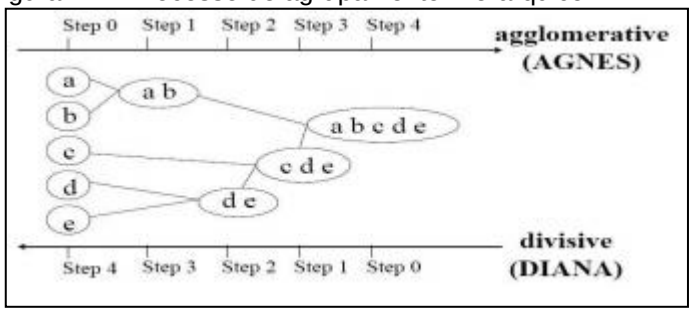
Figura 11 – Representação de um Dendograma.



Fonte: Ramos et al. (2016).

A separação hierárquica é produzida por meio do dendograma, iterativamente uma árvore divide um grupo de dados em subgrupos com quantidade menor comparada a inicial, repetindo o processo até que reste no último subconjunto um objeto de dados (GOLDSCHMIDT; BEZERRA; PASSOS, 2015). Pode ser classificado em duas categorias conforme o processo de *cluster* de seus algoritmos, as abordagens são aglomerativa e a divisiva (PANDE; SAMBARE; THAKRE, 2012, tradução nossa). Os exemplos de algoritmos das abordagens são: o algoritmo AGNES que pertence à abordagem aglomerativa e o algoritmo DIANA que pertence à abordagem divisiva (figura 12).

Figura 12 – Processo de agrupamento Hierárquico.



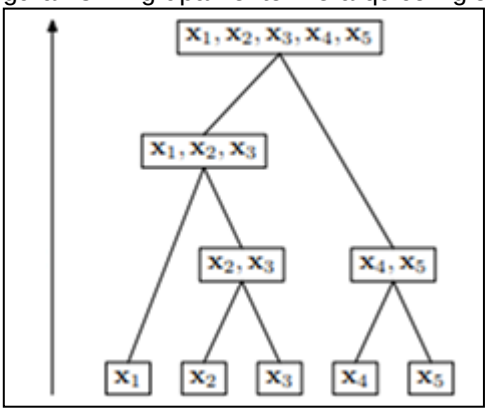
Fonte: Pande, Sambare e Thakre (2012).

Partindo da folha para a raiz da árvore, o método aglomerativo usa a estratégia *bottom-up.* Permite que cada objeto de dados forme o seu próprio *cluster*, unindo iterativamente os *clusters*, transformando a cada iteração em grupos maiores, até que todos os objetos de dados formem um único *cluster*, tornando-se a raiz da hierarquia. Ao contrário do método aglomerativo, a abordagem divisiva trabalha com a estratégia *top-down* na qual parte da raiz para a folha. Os objetos iniciam todos em um mesmo *cluster*, na raiz da hierarquia. Em seguida é realizada a divisão em vários *clusters* menores, assim seguindo até que reste apenas um objeto dentro de cada *cluster* (HAN; PEI; KAMBER, 2011, tradução nossa).

3.2.1 Método aglomerativo

O método aglomerativo é uma categoria de algoritmos que pertencem ao agrupamento hierárquico. Seu processo inicia com todos os dados dentro do seu próprio *cluster*. O processo seguinte realiza a divisão dos objetos, de acordo com funções de similaridades buscando a fusão entre os pares mais próximos de *cluster*. O processo é repetido, até que os dados estejam dentro de um mesmo *cluster*. As desvantagens da aglomeração são dadas devido ao agrupamento incorreto no estágio inicial da realocação e as diferentes medidas de similaridade que podem levar a resultados diversos (GAN; MA; WU, 2007, tradução nossa). A figura 13 apresenta o agrupamento dos dados no agrupamento hierárquico aglomerativo.

Figura 13 – Agrupamento Hierárquico Aglomerativo



Fonte: Gan, Ma e Wu (2007).

O agrupamento aglomerativo inicia com *N cluster* e cada um inclui exatamente um ponto de dados, na qual uma série de fusão é realizada forçando todos os objetos a irem para um mesmo grupo. É resumido por quatro etapas, (1) na

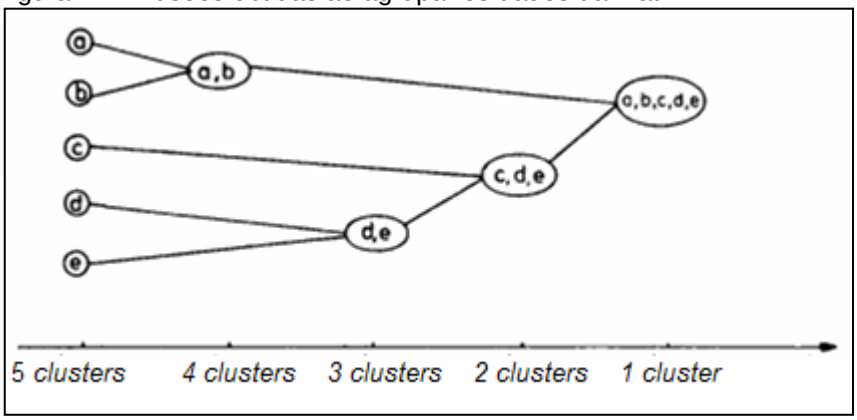
qual o processo inicia com *N cluster*, calculando a sua matriz de proximidade, (2) seguindo para a etapa na qual busca a distância *D* (C_i, C_j), em que *D* combina o *cluster* C_i e C_j para formar um novo cluster C_{ij}, o passo seguinte (3) é a atualização da matriz de proximidade, calculando as distâncias entre os *clusters C ij* e os outros *clusters*, (4) o processo é repetido, até que o agrupamento esteja concluído (XU; WUNSCH, 2008, tradução nossa).

3.2.2 Algoritmo AGNES

O *Agglomerative Nested* (AGNES) é um algoritmo que realiza agrupamento. É baseado no método de agrupamento hierárquico aglomerativo. Seguindo as métricas do método aglomerativo, o algoritmo trabalha com a abordagem de baixo para cima em uma hierarquia, na qual inicialmente cada objeto pertence a um *cluster*. No processo seguinte é realizado o cálculo da medida da distância mais próxima, para que seja realizada a fusão dos *clusters.* Existe um início de dissimilaridade que o algoritmo AGNES precisa alcançar. O agrupamento de AGNES insere os objetos de dados em *cluster* diferentes, até que a dissimilaridade seja alcançada para terminar o processo (SHARAF et al., 2016, tradução nossa).

Considerado um algoritmo de *link* único, cada *cluster* é representado por todos os objetos do *cluster*, tendo a semelhança calculada pela menor distância entre os dados em um grupo. O processo é repetido até que a fusão de todos os *clusters* esteja completa (figura 14), sendo que o usuário é quem define a quantidade de *cluster* como condição final (HUANG et al., 2015, tradução nossa).

Figura 14 – Fusões obtidas ao agrupar os dados da matriz



Fonte: Kaufman e Rousseeuw (2009, tradução nossa).

As etapas de atualização do algoritmo AGNES, são apresentadas na tabela 4.

Tabela 4 – Etapas para cálculo dos grupos pelo algoritmo *fuzzy c-means.*

Algoritmo *Fuzzy C-means Difuso Básico*	
Etapa	**Descrição**
1	Calcular a matriz de similaridade.
2	Alocar cada exemplar dos dados em um grupo, criando os nós folhas de arvore D.
3	Enquanto existir a possibilidade, de fusão de grupos, o algoritmo vai calcular.
4	O algoritmo verifica a distância entre todos os pares de grupos, usando a matriz de similaridade.
5	O algoritmo encontra o par de grupos mais similares e os transforma em um único grupo, criando um nó inteiro na hierarquia da arvore d.

Fonte: Da Silva, Peres e Boscarioli (2017).

A definição de AGNES, parte da existem de dois grupos (Ω_R e Ω_Q) e os números de objetos (|*R*| e |*Q*|), em que a dissimilaridade *d*(Ω_R, Ω_Q) é definida em forma de média para todas *d*(*i,j*). Desta forma *i* é objeto distinto e *j* um objeto qualquer (RODRIGUES, 2016).

$$d(R,Q) = \frac{1}{|R||Q|} \sum_{\substack{i \in \cap R \\ j \in \cap Q}} d\ (i,j) \tag{3}$$

A soma da equação para fazer a fusão de dois grupos. Supondo que existem dois *clusters* chamados A e B, sendo a seguinte etapa a junção de ambos, para formar um novo *cluster* R. É preciso aplicar a dissimilaridade de *d* (*R*, *Q*) de *R* para qualquer outro *cluster* de *Q* (KAUFMAN, ROUSSEEUW; 2009, tradução nossa).

$$d(R,Q) = \frac{A}{|R|}\ d\ (A,Q) + \frac{B}{|R|}\ d\ (B,Q) \tag{4}$$

A tabela 5 mostra os elementos e definições da equação de agrupamento do algoritmo AGNES.

Tabela 5 – Elementos da equação do algoritmo AGNES.

Elemento	Definição
$(\boldsymbol{\Omega_R}$ **e** $\boldsymbol{\Omega_Q})$	Representação de dois grupos.
(|*R*| e |*Q*|)	Número de objetos.
d $(\boldsymbol{\Omega_R, \Omega_Q})$	Dissimilaridade entre dois grupos.
A e B	Divisão dos *cluster*s.
d (*i, j*)	Média das dissimilaridades.

Fonte: do autor.

O processo para atualização do algoritmo inicia com *n* objetos de dados, finalizando com o valor de *k cluster*. A etapa seguinte faz a visualização dos objetos como um grupo, seguindo para encontrar os dois grupos mais próximos conforme a menor distância e realizando a fusão entre ambos. Por fim, após mesclar os grupos próximos, um novo *cluster* é gerado e assim o processo se repete, até que o número final definido seja alcançado (HUANG et al., 2015, tradução nossa). Para cálculo da medida de similaridade, o algoritmo AGNES, foi aplicado por meio da distância *euclidiana* e *manhattan.*

3.3 MEDIDAS DE QUALIDADE

Independente do conjunto de dados e seu tamanho existem diversas formas para a aplicação dos algoritmos de agrupamento, desempenhando o objetivo de formar grupos. O número de *cluster* a ser usado é um problema, pois nem sempre a divisão é correta (KIRKLAND; DE LA IGLESIA, 2013, tradução nossa). O problema ocorre ao gerar a divisão do conjunto de dados o algoritmo de agrupamento não realiza a verificação da estrutura, não conferindo se ela realmente existe. É fundamental conferir se os grupos estavam ocultos ou apenas foram gerados pelo algoritmo, não apresentando nenhum significado (BOSCARIOLI, 2008).

Desta forma, surgem dois questionamentos em relação ao *cluster*, o primeiro relacionado à validade, em como é avaliada a saída de um algoritmo de agrupamento, e a segunda questão é relacionada ao resultado, em como saber o que caracteriza um resultado de ser bom ou ruim. Assim sendo, a primeira etapa

para análise é baseado na avaliação do domínio de dados e não do algoritmo, na qual os dados que não são considerados para a formação de *clusters*, não devem entrar no processo para agrupamento, em vista que alguns algoritmos tem melhor desempenho que outros (KANTARDZIC, 2011, tradução nossa). Segunda etapa de análise é baseada na necessidade de encontrar uma forma de validar a seriedade das partições após o agrupamento (LIU, 2013, tradução nossa). O processo é realizado por meio dos índices de validação, na qual é testada a medida de qualidade dos *clusters* que são formados pelo agrupamento (ZERABI; MESHOUL, 2017, tradução nossa). Os índices de validação comparam o resultado de diferentes agrupamentos, assim como os resultados do mesmo algoritmo com diversos valores de *cluster* (CEBECI; KAVLAK; YILDIZ, 2017, tradução nossa).

O processo que determina a melhor partição, e o número ideal de *cluster* em um conjunto de dados, usando as medidas de validação é realizado em quatro etapas, partindo da (1) inicialização de uma lista de algoritmos de agrupamento, para aplicar em um determinado conjunto de dados, (2) na qual para cada algoritmo é idealizado a utilização de diferentes combinações de parâmetro, para que sejam obtidos diferentes resultados, (3) seguindo com o cálculo de validação correspondente ao índice de cada partição obtida, (4) escolhendo assim a melhor partição e o *cluster* ideal (CHEN; LI; LI, 2011, tradução nossa). O agrupamento pode ser validado, seguindo as seguintes regras:

a) determinando a tendência de agrupamento;
b) aplicar índices de critérios internos, externo e relativos;
c) comparar dois conjuntos de dados para determinar o melhor.

3.3.1 Tendência de agrupamento

A tendência de agrupamento determina se um conjunto de dados possui grupos significativos. Quando recebem dados, os algoritmos de agrupamento costumam formar grupos, não verificando se existem dados nos grupos e se foram formados com qualidade. Para resolver este problema, o pesquisador pode aplicar os dados em múltiplos algoritmos e avaliar a qualidade dos grupos resultantes, verificando se os resultados são igualmente fracos, significa que não existe grupo nos dados. Outra forma de avaliar a tendência de agrupamento é por meio de testes

estatísticos (TAN; STEINBACH; KUMAR, 2009). Uma das abordagens conhecidas, é por meio da estatística de *Hopkins*.

A estatística de *Hopkins* é baseada na distância de um ponto real do vizinho mais próximo, com a distância de um ponto escolhido aleatoriamente. Suponha que a distância mais próxima marcada para o seu vizinho é calculado W_i, enquanto U_i, seja à distância dos pontos da amostra do conjunto de dados original (LAWSON; JURS, 1990, tradução nossa). A estatística de *Hopkins* é definida pela equação:

$$H = \sum U_i / (\sum U_i + \sum W_i) \quad (5)$$

O teste estatístico de *Hopkins* pode ser conduzido de forma iterativa, considerando valor >0,5, como valor para rejeitar a hipótese alternativa (KASSAMBARA, 2017, tradução nossa). Valores próximos a zero (0), indicam que os dados estão altamente agrupados.

3.3.2 Índices de validação pelos critérios internos, externo e relativo

Existem várias medidas de validade e estatísticas que realizam a validação de qualidade, na qual pode ser dividido em três tipos principais, critério interno, externo e relativo (ZAKI; MEIRA JUNIOR, 2014, tradução nossa).

As abordagens critério externo e interno envolvem testes estatísticos na aplicação, o que faz com que sejam métodos computacionalmente caros. O critério relativo, não trabalha com testes estatísticos (GAN; MA; WU, 2007, tradução nossa). Os testes para critério externo e interno são os que chamam a atenção dos pesquisadores. Em contrapartida, na maioria das vezes é usado o critério relativo para encontrar um bom valor de *k*, tendo em vista que a validação de *cluster* é uma etapa absoluta no processo de agrupamento em situações não supervisionadas (BAARSCH; CELEBI, 2012, tradução nossa). A figura 15 mostra a separação dos critérios de validação em estatístico e não estatístico.

Figura 15 – Testes estatísticos e não estáticos de validação.

Fonte: Do autor.

O critério externo determina se uma estrutura é internamente apropriada para os dados (KANTARDZIC, 2011, tradução nossa). O critério interno é aplicado em duas situações, na qual são selecionadas conforme a estrutura de agrupamento, sendo a primeira para a hierarquia de *cluster* e a segunda para *cluster* único (HALKIDI; BATISTAKIS; VAZIRGIANNIS, tradução nossa, 2001). O critério mede a qualidade da estrutura de agrupamento, determinando o quão relacionado estão os objetos dos *clusters* e o quanto está separado um grupo de outros (TAN; STEINBACH; KUMAR, 2009). Dentre os índices de validação pelo critério externo, estão os índices de *dunn* e *silhouette.*

O índice de *dunn* realiza a comparação entre a distância dos grupos, com o grupo mais distante, com o objetivo de encontrar a relação mínima entre os grupos, quanto maior a distância de um grupo para outro, melhor é a separabilidade entre os grupos e de formar grupos. O índice de *silhouette* verifica o quão bem um exemplar pertence a um grupo, com valores entre 1 e -1, quanto mais próximo de um (1), melhor é o resultado. A tabela 6 mostra as equações dos índices de *dunn* e *silhouette*:

Tabela 6 – Equações dos indices de validação do critério externo.

Índice Validação	**Equação**
Índice de Dunn	$I_{Dunn} = min_{1 \le p \le k} \left\{ min_{1 \le q \le k, p \ne q} \left\{ \frac{dist\ (G_p, G_q)}{\max disp\ (G_k)} \right\} \right\}$
Índice Silhouette	$I_{Sil} = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$

Fonte: Do autor.

A tabela 7 descreve os elementos das equações dos índices de critérios internos de *Dunn* e *silhouette*.

Tabela 7 – Descrição das equações dos indices de validação do critério externo.

Índice Validação	Elemento	Descrição
Índice de Dunn	$dist\ (G_p, G_q)$	Medida de distância entre grupos.
	$disp\ (G_k)$	Medida de dispersão em um grupo.
	$p, q = 1 \ldots K$	É a quantidade de grupos.
Índice Silhouette	$a(i)$	Distância média do exemplar i, a todos os demais exemplares do grupo.
	$b(i)$	Distância média do exemplar i, de todos os exemplares do grupo mais próximo do seu.

Fonte: Do autor.

De acordo com Silva, Peres e Boscarioli (2016), a validação pelo critério interno é baseado em qualquer tipo de informação extra sobre o conjunto de dados sob análise, sendo unicamente dependente da distribuição dos exemplares entre os grupos resultantes do algoritmo de agrupamento.

O critério externo aborda o resultado das estruturas dos algoritmos de agrupamento, baseado em uma estrutura pré-definida e determinada por um conjunto de dados, refletindo no conjunto de dados a sua estrutura intuitiva. É dividido em duas abordagens, na qual (a) compara uma estrutura de agrupamento decorrente de C com uma partição independente dos dados P e (b) compara a matriz de similaridade do grupo G, que é definido por meio do número de grupos, com uma partição P (GAN; MA; WU, 2007, tradução nossa).

Diversos índices comparam as organizações G e P, de modo que a soma para medir o grau de similaridade é descrita na tabela 8, nomeados como A, B, C e D.

Tabela 8 – Descrição das etapas da soma para medir o grau de similaridade.

Soma	Descrição
Soma A	Quantidade de pares de exemplares que pertencem ao mesmo grupo em G e a mesma partição em P.
Soma B	Quantidade de pares de exemplares, que pertencem ao mesmo grupo G e a partições P diferentes.
Soma C	Quantidade de pares de exemplares que pertencem a grupos G diferentes e a mesma partição P.
Soma D	Quantidade de pares de exemplares, que pertencem a grupos G diferentes e a partições P diferentes.

Fonte: Silva, Peres e Boscarioli (2016).

Alguns índices são capazes de medir o grau de similaridade entre G e P. A tabela 9 montra os principais índices de validação externo.

Tabela 9 – Equações dos índices de validação do critério externo.

Índice	Equação
Índice de Rand	$I_{Rand} = (A + D)/(A + B + C + D)$
Índice de Jaccard	$I_{Jaccard} = A/(A + B + C)$
Índice de Folkes Mallows	$I_{FM} = \sqrt{\frac{A}{A+B} \times \frac{A}{A+C}}$

Fonte: Silva, Peres e Boscarioli (2016).

O valor um (1) é considerado como máximo e indica alto grau de similaridade entre grupos e partições, de modo que quanto mais próximo o resultado do índice for para um (1), melhor é considerado o agrupamento (SILVA; PERES; BOSCARIOLI, 2016).

À distância cofenética é um critério interno, e um método utilizado para agrupamento hierárquico, pois realiza a validação entre dois objetos de dados, inserindo em um mesmo grupo. Um dos usos mais comuns é utilizando o Coeficiente de Correlação Cofenética (CPCC), verificando o quão bom um agrupamento hierárquico se adapta aos dados. Geralmente o CPCC é usado, para avaliar qual

método de conexão é bom para determinado tipo de dados em um agrupamento hierárquico (TAN; STEINBACH; KUMAR, 2009).

O critério relativo realiza a comparação de uma estrutura de agrupamento *C*, com outras estruturas, que são obtidas após a aplicação de diferentes algoritmos ou do mesmo algoritmo (XU; WUNSCH, 2008, tradução nossa). Esta abordagem mede o mérito relativo do cluster, para aplicar é necessário resolver a seleção de estruturas para seleção (KANTARDZIC, 2011, tradução nossa). Os índices de critério externo e *Dunn* e *Silhouette*, também são considerados critério relativo, visto que comparam resultados diferentes, provenientes de diferentes execuções de algoritmos de agrupamento (SILVA; PERES; BOSCARIOLI, 2016).

Por último, por meio da validação de *fuzzy* é possível definir um índice de qualidade para agrupamento *fuzzy*, com o objetivo é encontrar técnicas de agrupamento, na qual os pontos em um conjunto de dados mostrem um alto grau de participação em um cluster. Normalmente existem duas categorias para validação fuzzy, sendo uma envolvendo valores de associação e a segunda envolvendo valores de associação e valores de dados (GAN; MA; WU, 2007, tradução nossa). O coeficiente de partição envolve valores de associação e o coeficiente de entropia é o segundo coeficiente da validação *fuzzy* (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001, tradução nossa). É dividido pelas equações do índice do coeficiente de partição (PC) e índice do coeficiente de entropia (PE).

4 EDUCAÇÃO A DISTÂNCIA

A tecnologia está presente no cotidiano das pessoas, alterando a sociedade em todos os aspectos, dentre eles as áreas social, política, econômica, médica e educacional (RODRIGUES, 2014).

Acompanhando o desenvolvimento tecnológico, está o crescente uso da internet, permitindo maior acesso à informação e trazendo mudança na educação, em que o uso das Tecnologias de Informação e Comunicação (TIC), se torna fundamental. Este fato permite a integração entre instituições de ensino, docentes e discentes, que passa a utilizar a tecnologia para preparar e gerir conteúdos e administrar disciplinas e cursos presenciais, semipresenciais e a distância (SANTOS et al., 2016).

De acordo com Vilaça (2011), uma das modalidades que relacionam a educação e tecnologia e que está em constante crescimento, fazendo com que as instituições de ensino e educadores estejam gradativamente ligadas a atividades e projetos relacionados ao mesmo, é chamada de Educação a Distância (EaD).

Nesta modalidade, os alunos e professores estão separados fisicamente no espaço e tempo, necessitando de meios tecnológicos para repassar e obter a informação e o conhecimento, podendo ter momentos presenciais (ALVES, 2018).

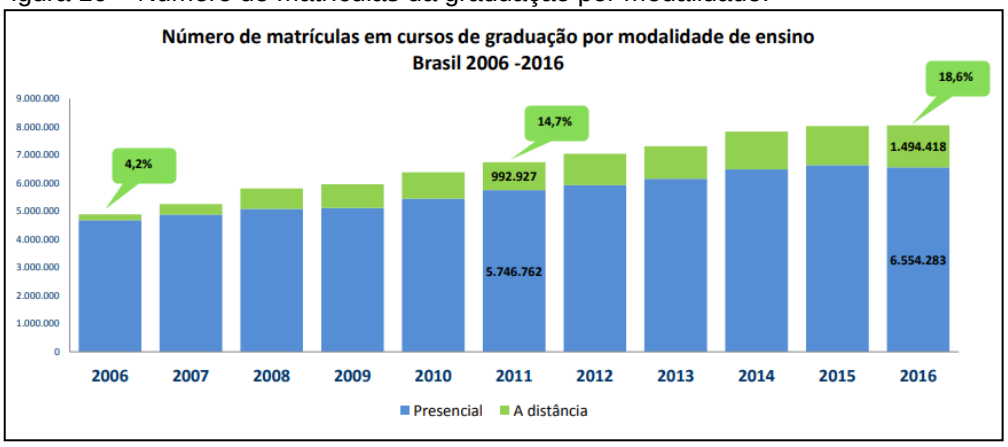
O EaD surgiu em 1728, quando Caleb Philips criou um novo método de ensino, ministrando aulas por meio de correspondências. No decorrer do tempo, novos meios foram surgindo como o rádio e entre as décadas de 60 e 70 a TV educacional, no entanto eram meios de comunicação que poderiam causar desmotivação, pois são meios que não permitem dinamismo entre tutores e alunos (SOUZA et al., 2016). Na década de 90, teve início a integração de redes de computador com estações de trabalho multimídia, gerando para o EaD novas expectativas, possibilitando o acesso sistematizado a informação e trazendo a interação entre os alunos e professores (FARIA; SALVADORI, 2010).

No Brasil, atualmente o EaD aparece como uma possibilidade de inclusão social, popularizando e melhorando o processo educacional, pois elimina problemas tradicionais como distanciamento geográfico dos centros de ensino e aprendizagem, conflitos de horários, impossibilidade de combinar o trabalho com os estudos e custos adicionais como transporte (LESSA, 2018). O EaD vem crescendo no Brasil, entre os anos de 2015 e 2016 o número de ingressantes na graduação foi de 2,2%,

isso ocorreu, pois, a modalidade de EaD teve um aumento de 20%, enquanto o ensino presencial teve um decréscimo no número de ingressantes de 3,7% (INEP, 2018).

A participação da educação à distância no número de matrículas da graduação por modalidade (figura 16) cresceu entre os anos de 2006 e 2016. Os cursos à distância representavam 4,2% das matrículas de graduação em 2006, passando em 2016 para uma participação de 18,6%.

Figura 16 – Número de matrículas da graduação por modalidade.



Fonte: Inep (2018).

O aumento significativo no número de matrículas demonstra a importância que a educação a distância vem tendo no Brasil, tendo como consequência o índice de crescimento de pessoas que tem acesso ao ensino superior, realizando uma mudança social (BIELSCHOWSKY, 2018).

Outro ponto importante para o crescimento do EaD é a facilidade para acesso do aluno, pois é auxiliado por Ambientes Virtuais de Aprendizagem (AVA), que ajudam na administração e gerenciamento de um ou mais cursos, apoiando nas matrículas, lançamento de notas, interação entre os alunos e professores, não necessitando do deslocamento do aluno ate um local físico para a aplicação da aula(SOUZA et al., 2016).

O EaD tem como problema a evasão dos alunos, onde pode ser ocasionado por vários fatores, entre eles fatores sociais, econômicos, qualidade do curso ofertado, dificuldades com relação à ambiente, entre outros. Neste contexto, surge o *Educational Data Mining*, voltado à análise de dados educacionais e para auxiliar instituições e tutores na tomada de decisão, previsão de desempenho e

evasão dos alunos, entre outros. Os dados usados para esse tipo de trabalhos são provenientes de AVAs.

Por meio dos subcapítulos seguintes, este trabalho apresenta a definição de EDM e AVA, mostrando as áreas de aplicação para as tarefas de agrupamento do *educational data mining*.

4.1 AMBIENTES VIRTUAIS DE APRENDIZAGEM

Os Ambientes Virtuais de Aprendizagem podem ser usados tanto para o ensino presencial como para o a distância. Permitindo a comunicação por meio de fóruns e *chats*, postagem de materiais didáticos ou trabalhos, compartilhamento e gerenciamento de informação (PEREIRA; FRANÇA, 2017).

Os AVAs permitem o acesso ao conteúdo de diferentes maneiras como vídeo, texto, imagens e sons, pois são *softwares* usados para o ensino e aprendizagem, possibilitando a interatividade em diferentes ferramentas. Sendo formado por meios disponíveis em uma plataforma, que organiza a troca de informação. Destacando o *Moodle* como sendo a plataforma mais popular, ela é gratuita e usada em mais de 237 países, possuindo mais de 84.524 registros de sites de cursos (BARBOZA; SILVA, 2016). , existem diversas plataformas que realizam o gerenciamento de aprendizagem, tais como *Blackboard*, *Sakai*, *Desire2Learn*, *TelEduc* e outras (EMREDE**,** 2015).

O uso de ambientes virtuais e de sistemas ligados à educação, geram grande quantidade de dados, proveniente das interações e registros contínuos dos professores, alunos, gestores e demais pessoas ligadas à plataforma. Os dados se tornam extremamente importantes, pois podem gerar muito mais que relatórios com informações para gerenciamento ou desempenho (RODRIGUES et al., 2014). Por meio dos dados, é possível prever e medir o desempenho dos alunos, prever o risco de evasão e formar grupos que contenham o mesmo comportamento para trabalhar de forma adequada, entre muitas outras possibilidades (RAMOS et al., 2014).

Apesar do grande número de dados armazenado em banco de dados, provenientes de AVA, nem sempre eles são processados da maneira exata, não sendo considerada uma informação útil para apoiar os educadores e gestores educacionais na tomada de decisões. Encontrar informação relevante em uma base de dados, não é uma tarefa simples para ser realizado sem a ajuda de ferramentas

computacionais (GOTTARDO; KAESTNER; NORONHA, 2012). Com o auxílio da mineração de dados é possível descobrir conhecimento relevante, que pode ser uteis para pesquisadores educacionais, educadores e as instituições, existindo uma área especifica pra trabalhar com os dados educacionais, chama de e*ducational data mining* (RODRIGUES et al., 2014).

4.2 EDUCACIONAL DATA MINING

A mineração de dados é uma área que unifica o aprendizado de máquina, estatística, banco de dados, reconhecimento de padrões, na qual por meio dos resultados facilita na tomada de decisão das diversas áreas que pode ser empregada (PRISTYANTO; PRATAMA; NUGRAHA, 2018, tradução nossa).

A solidificação mundial da rede vem possibilitando que a educação a distância tenha grande aumento de usuários, sendo eles alunos, professores e pessoas relacionadas às instituições de educação (RAMOS et al., 2016, tradução nossa). A internet passou a ser usada na educação, gerando um grande repositório de dados com as informações de alunos. Os dados são importantes para especialista em educação, que utilizam para explorar e formar conhecimento útil, para entender o comportamento dos alunos e como eles apreendem (ROMERO; VENTURA, 2013, tradução nossa).

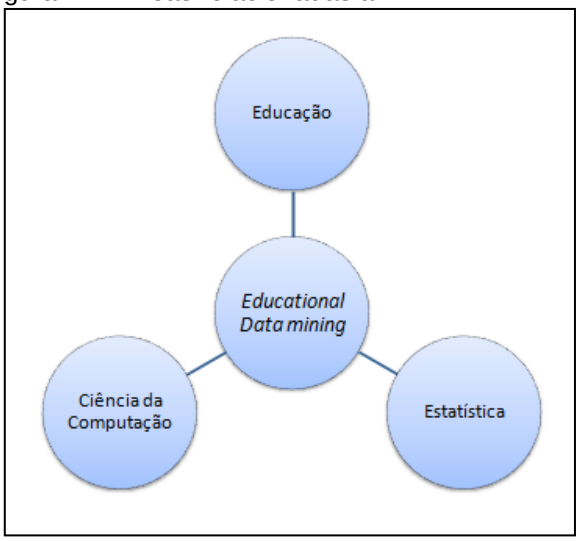
Nos últimos tempos o interesse na educação à distância, fez com que surgisse um novo conceito de mineração de dados, especialmente focado na educação e no conhecimento gerado a partir dos dados provenientes de escolas e universidades, o conceito é conhecido como *educational data mining (EDM)* (STEFANOVA; KABAKCHIEVA, 2017, tradução nossa).

O EDM utiliza as taxonomias de *mineração de dados* como predição (classificação, regressão, estimação de densidade), mineração de regras de associação, descoberta de modelos e agrupamento, buscando padrões que possam melhorar os sistemas educacionais (ZHENG; JIA, 2011, tradução nossa).

Os dados não são restritos apenas a iteração dos alunos com os sistemas educacionais, é possível analisar dados administrativos, demográficos, *chats*, entre outros. A EDM trabalha com três áreas principais, sendo a educação, estatística e ciência da computação (ROMERO; VENTURA, 2013, tradução nossa). A figura 17

mostra as áreas que o *educational mineração de dados* aborda, estando no centro como conceito.

Figura 17 – Áreas relacionadas à EDM.



Fonte: Do autor.

Existem várias direções de pesquisa da EDM, que são baseados nos algoritmos de mineração de dados. Os principais algoritmos de interesse da área são os de predição, agrupamento e mineração de relações. Na predição existem três formas para realizar a mineração de dados educacionais, sendo a classificação, regressão e estimação de dados (BAKER; ISOTANI; CARVALHO, 2011). Em geral, as aplicações de EDM são agrupadas conforme propósitos educacionais, um algoritmo pode ser aplicado para diferentes propósitos **(**GOLDSCHMIDT; BEZERRA; PASSOS, 2015). A tabela 10 mostra a classificação das tarefas de EDM, que foram sugeridas por Romero e Ventura (2010), indicando os algoritmos que são aplicados para cada caso em específico e demonstrando que o EDM é usado para diversas áreas da educação, sendo aplicadas para alunos, instrutores, instituições, entre outros.

Tabela 10 – Classificação das tarefas de EDM.

Aplicação	Descrição	Algoritmos
Análise e visualização de dados, voltada a instrutores.	Destacar informações que são uteis para apoiar na tomada de decisão. Pode ajudar educadores e administradores dos cursos a ter uma visão geral do aprendizado dos alunos.	Medidas Estatísticas Informação de visualização
Fornecendo feedback para instrutores de apoio, voltada a instrutores.	Fornece o feedback para apoiar na tomada de decisão, permitindo que sejam tomadas medidas para escolha de abordagens pedagógicas.	Associação, Classificação, Agrupamento e Sequências.
Recomendações para estudantes, voltada para estudantes.	Recomendações diretas para alunos, baseado nas suas atividades, podendo adaptar conteúdos para cada aluno em particular.	Associação, Agrupamento e Sequencia.
Prever o desempenho do aluno, voltado para características dos alunos.	Encontrar um valor desconhecido de uma variável que descreve um aluno.	Regressão e Classificação.
Modelagem de estudantes, voltado para características dos alunos.	Construir modelos de conhecimento relacionado aos usuários, estudantes.	Agrupamento, Classificação e Sumarização.
Agrupando alunos, voltado para características dos alunos.	Criar grupos para identificar características em comum e construir um sistema de aprendizado personalizado.	Agrupamento e Classificação.
Análise de redes sociais, voltado para características dos alunos.	Analisar a relação entre indivíduos Que estão conectados por relações sociais em redes sociais e ambientes virtuais.	Classificação e Agrupamento.
Desenvolvendo mapas conceituais, voltado para o conhecimento da relação entre estudantes e conceitos.	Ajuda educadores no processo de desenvolvimento de mapas conceituais, para mostrar relação entre conceitos.	Associação.
Construindo material didático, voltado para desenvolvimento e planejamento dos cursos.	Ajuda no desenvolvimento materiais didáticos automaticamente.	Classificação e Agrupamento.
Planejamento e agendamento, voltado para desenvolvimento e planejamento dos cursos.	Melhora o processo de educação tradicional, planejando cursos futuros.	Regressão, Classificação e Associação.

Fonte: Romero e Ventura (2010).

5 TRABALHOS CORRELATOS

Nos subcapítulos a seguir são abordados cinco casos de uso das técnicas de agrupamento, em trabalhos que estão relacionados ao tema da pesquisa, sendo do âmbito educacional ou utilização dos métodos de agrupamento hierárquico aglomerativo ou particional. Os algoritmos apresentados são fuzzy *c-means* e AGNES.

5.1 UMA BREVE ANÁLISE DAS PRINCIPAIS TECNOLOGIAS E APLICAÇÕES DA MINERAÇÃO DE DADOS EDUCACIONAIS NA PLATAFORMA DE APRENDIZAGEM ON-LINE

A pesquisa é escrita por Wei Zhang e Shiming Qin e foi publicada no *Institute of Electrical and Electronic Engineers* (IEEE), no IEEE *3rd International Conference on Big Data Analysis* em 2018, na cidade de Shanghai, China. O objetivo do estudo é permitir que pesquisadores ou educadores adquirissem maior compreensão em como usar mineração de dados para educação, indicando as melhores técnicas de mineração, que podem ajudar na melhoria do ensino de alunos (ZHANG; QIN, 2018, tradução nossa).

Os autores Zhang e Qin (2018) propõem uma análise sobre a mineração de dados educacionais, para que os educadores tenham entendimento sobre a relação e desenvolvimento dos conceitos da EDM, visto que é uma área que está em crescimento, em que a tomada de decisão sobre o ensino é assertiva baseado na análise dos dados que são provenientes da aplicação do EDM.

Os autores Zhang e Qin (2018), destacam que o processo deve ser dividido em três partes: separação dos dados ou pré-processamento, aplicação da mineração e interpretação do resultado. A primeira etapa é a preparação dos dados, visto que a qualidade dos dados é essencial para um resultado eficiente. A etapa seguinte é a EDM, sendo necessário ao pesquisador conhecer as principais técnicas de mineração, os algoritmos e como aplicar na área educacional, baseado no tipo de análise, buscando atingir o objetivo do estudo. Conforme os autores, saber decidir como usar esses algoritmos nos dados, aumenta a eficiência da pesquisa e melhora o valor do conhecimento. Classificação, agrupamento, associação, regressão e predição são as principais técnicas de EDM.

Os autores Zhang e Qin (2018), destacam que a EDM para tomada de decisão na educação, conduz os estudos a uma análise mais científica e com maior conhecimento, para a educação a distância. Entretanto, é preciso ter cuidado para abordar os aspectos gerenciais, éticos e técnicos.

5.2 UM ESTUDO COMPARATIVO ENTRE *MÉTODOS DE AGRUPAMENTO EM MINERAÇÃO DE DADOS EDUCACIONAIS*

Este artigo foi desenvolvido por Jorge Luis Cavalcanti Ramos, Ricardo Euller Dantas e Silva, João Carlos Sedraz Silva, Rodrigo Lins Rodrigues e Alex Sandro Gomes, sendo apoiado pela Fundação de Amparo à Ciência e Tecnologia de Pernambuco (FACEPE) e publicado em 2016 no IEEE *Latin America Transactions*.

Neste artigo Ramos et al. (2012), propuseram encontrar por meio da EDM novos métodos, técnicas e procedimentos para melhorar a aplicação da descoberta de conhecimento em dados educacionais, a fim de obter alternativas que indiquem dificuldades em estruturas educacionais. Para isso, foi realizado um estudo comparativo entre os métodos de agrupamento hierárquico e particional.

Os dados utilizados na pesquisa são de alunos do Curso de Bacharelado em Administração Pública, na modalidade a distância da Universidade Federal do Vale do São Francisco - Brasil (UNIVASF). Possuindo 206 alunos, que estão divididos em seis polos de cidades da região. Os dados para a aplicação da pesquisa são provenientes da disciplina de Introdução a Educação a Distância, que é ofertada por meio de um Ambiente Virtual de Aprendizagem (AVA) chamado *Moodle* (RAMOS et al., 2016).

A primeira etapa do processo foi a coleta e tratamento dos dados, sendo dividida em três partes principais, conhecidas como preparação dos dados, pós-processamento e mineração de dados. Na preparação de dados, foram realizadas três etapas: seleção, pré-processamento e transformação. Na seleção, os dados das interações dos alunos foram coletados a partir de consultas de Structured Query Language (SQL) na base de dados. A fase seguinte referiu-se ao pré-processamento dos dados, na qual foi realizada a limpeza dos dados, com o propósito de conservar a qualidade e precisão. A preparação finalizou a etapa de seleção, normalizando os dados aos algoritmos selecionados, conforme a aplicação utilizada para a mineração de dados (RAMOS et al., 2016).

Na etapa de mineração de dados foi empregado para o agrupamento hierárquico técnicas de análise de *clusters*, a partir dos métodos de Ward e a distância Euclidiana, enquanto para agrupamento particional foi utilizado o algoritmo *K-means*. Os autores escolheram quatro (*k=4*) grupos ambos os métodos, como o número ideal de grupos (RAMOS et al., 2016).

Os resultados foram obtidos com ajuda de dois softwares o pacote estático R para agrupamento hierárquico e o *RapidMiner* para a agrupamento não hierárquica. O resultado principal apresentou que ambos os métodos apresentados indicam resultados semelhantes, visto que os autores montaram uma matriz de semelhança identificando que alunos de determinado grupo hierárquico, estavam em presentes nos mesmos grupos particionais, apresentando semelhança na divisão, características e tamanhos dos grupos, indicando que ambos os métodos obtém bom resultado para o agrupamento dos dados (RAMOS et al., 2016).

5.3 RESUMO DE DOCUMENTOS PELA ABORDAGEM DE AGRUPAMENTO ANINHADO

Este artigo tem como autores Aakanksha Sharaff, Hari Shrawgi, Priyank Arora e Anshul Verma, publicado em dezembro de 2016 na Conferência Internacional IEEE 2016 sobre os avanços em eletrônica, comunicação e tecnologia de computadores (*International Conference on Advances in Electronics, Communication and Computer Technology* - ICAECCT), apoiado pela Faculdade de engenharia de Rajarshi Shahu, em Pune, Índia de Pune.

Os autores propõem um modelo extrativo de documentos baseado em agrupamento aglomerativo por meio do algoritmo AGNES. Uma interpretação precisa de um documento requer dedução de frases importantes do texto. Por meio do agrupamento é possível reunir frases semelhantes em um conjunto de dados. Frases fortemente classificadas são as que fazem sentido para o texto. Portanto, um resumo gerado é um conjunto destas frases altamente classificadas (SHARAFF et al., 2016, tradução nossa).

Os autores coletaram diversos dados e armazenaram em uma base de dados, aplicando o pré-processamento e em seguida aplicando a mineração de dados, pelo algoritmo AGNES, gerando grupos. As frases contidas dentro dos

grupos foram classificadas, formando um resumo a partir de frases de alto escalão de cada grupo (SHARAFF et al., 2016, tradução nossa).

A avaliação de desempenho do método proposto foi feita por meio de um resumo escrito manualmente por um humano, em que a pessoa deveria escrever baseado no texto original. O resumo automático do método proposto foi aplicado no mesmo texto e os autores verificaram pelas medidas de similaridade, que em 10 testes, 80% dos documentos, o método proposto apresentava um resumo 70% mais parecido com o texto do documento original em relação ao escrito pelo humano, sendo mais eficaz (SHARAFF et al., 2016, tradução nossa).

5.4 MINERAÇÃO DE DADOS DO SISTEMA ACADÊMICO DO INSTITUTO FEDERAL DO SUDESTE DE MINAS GERAIS – CAMPUS JUIZ DE FORA

O artigo de Aluísio Cardoso Silva e Ricardo Costa Pinto e Santos foi publica no ano de 2018, na revista Seminários de Trabalhos de Conclusão de Curso do Bacharelado em Sistemas de Informação, do instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais - Campus Juiz de Fora.

No artigo, os autores aplicam o processo de mineração de dados para apurar a influência do perfil dos alunos dos cursos técnicos do IF Sudeste MG, do campus de Juiz de Fora. Os dados são provenientes do Sistema Integrado de Gestão Acadêmica (SIGA) e foram coletados os registros das disciplinas ministradas entre 2011 e 2015. O processo compreende a etapa de pré-processamento, incluindo seleção e a aglomeração das informações disponíveis em tabelas SQL e seguindo com as etapas de processamento e análise dos dados (SILVA; SANTOS, 2018).

Os autores buscaram por meio de métodos estatísticos e análise de *cluster* extrair informações relevantes e encontrar padrões existentes nos dados. Para a aplicação da mineração de dados foi utilizado o algoritmo FCM, por ser considerado o mais popular em estudos da área de mineração de dados educacionais. O Matlab foi o *software* escolhido para a obtenção dos resultados, seguindo com o cálculo de validação de dados para definir a quantidade exata de *cluster* e realizar a aplicação do FCM. Após a separação em dois (K=2) *clusters*, os autores realizaram o estudo estatístico das características dos *cluster*s, buscando encontrar as características de cada um deles (SILVA; SANTOS, 2018).

Os resultados encontrados com a análise de *clusters* mostram que a distância percorrida pelos alunos para estudarem na instituição pode afetar no desempenho, devido ao desgaste físico, visto que os alunos que moram perto da instituição possuem um melhor desempenho. Outro ponto é relacionado à faixa de renda, pois os alunos que dispõem de mais recursos financeiros têm acesso facilitado à *internet* e materiais que são uteis para a aprendizagem (SILVA; SANTOS, 2018).

Os autores Silva e Santos (2018) não encontraram as causas para os resultados, pois não era o objetivo do trabalho, identificaram que estudos nesse sentido podem direcionar em futuras análises em dados educacionais.

5.5 MODELO DE ANÁLISE DE AGRUPAMENTO PARA A INFERTILIDADE MASCULINA EM DADOS BIOMÉDICOS DE UMA CLÍNICA DO EXTREMO SUL CATARINENSE

A pesquisa foi desenvolvida por Rodrigo Búrigo Esmeraldino, sendo apresentada como trabalho de conclusão de curso na Universidade do Extremo Sul Catarinense, em junho de 2017, na cidade de Criciúma em Santa Catarina. O autor por meio dos algoritmos de agrupamento AGNES e DIANA, realiza a mineração de dados de uma base sobre infertilidade masculina, coletado em uma clínica de produção assistida, entre os anos de 2012 e 2014, no sul de Santa Catarina (ESMERALDINO, 2017).

O objetivo da pesquisa, de acordo com Esmeraldino (2017), é comparar os resultados do agrupamento hierárquico aglomerativo pelo algoritmo AGNES e agrupamento hierárquico divisivo, pelo algoritmo DIANA, para identificar o modelo final que atenda a natureza dos dados, validando por meio de medidas de qualidade a formação dos grupos.

Para aplicar a mineração de dados, o autor precisou seguir os processos da descoberta de conhecimento, para garantir que os dados obtivessem qualidade ao serem minerados. As etapas do pré-processamento aplicadas, foram a normalização, para evitar discrepância entre os dados e a conferência pelo mento estatístico de *Hopkins* se o conjunto de dados era bom para aplicar agrupamento, com variáveis numéricas ou alfanumérico.

Ao iniciar a mineração dos dados, os métodos de *manhattan* e *euclidiana* foram utilizados para calcular a matriz de distância dos dados, etapa necessária para ambos os algoritmos na formação do agrupamento.

O algoritmo AGNES utiliza medidas de conexão para aglomerar os dados em grupos, enquanto o algoritmo DIANA usa a distância entre os dados para fazer a divisão em grupos. O autor aplicou os algoritmos com divisões de grupos de dois (*k=2*) a dez (*k=10*), para identificar qual é a ideal para aplicar no conjunto de dados.

Para a validação dos resultados, obtidos pelos algoritmos, Esmeraldino (2017), utiliza certeiros de validação externo e interno, de modo que os melhores valores devem ser próximos ou iguais a um (1), para o critério externo, sendo aplicado *Jaccard*, *Rand*, *Folkes Mallows* e *Hubert*. Para o interno os valores máximos, são os que possuem melhores resultados para os índices de *Dunn*, *Silhouette* e menores valores para *Davies-Bouldin*, *C-Index* e *Xie Beni*. Por meio do coeficiente de correlação cofenética, foi realizado a validação da hierarquia, formado pelas distâncias de *manhattan* e *euclidiana*, e os métodos de conexão.

De acordo com Esmeraldino (2017), para o algoritmo AGNES à distância *euclidiana,* aplicado com o método de conexão média de grupos, obteve melhores resultados, enquanto para o algoritmo DIANA, a distância escolhida foi a *manhattan*. O modelo final definido foi o com o algoritmo AGNES, pela fácil visualização da hierarquia e resultado adequado ao objetivo proposto.

6 ANÁLISE DE AGRUPAMENTO PELOS MÉTODOS HIERÁRQUICO AGLOMERATIVO E PARTICIONAL *FUZZY* UTILIZADOS PARA *EDUCATIONAL DATA MINING* EM DADOS DE EDUCAÇÃO A DISTÂNCIA

Esta pesquisa tem como objetivo propor um modelo de análise de agrupamento hierárquico aglomerativo e particional *fuzzy*, utilizando-se o conceito de *educational data mining*, nos dados provenientes da disciplina de Introdução a Engenharia de Segurança do Trabalho, que é ministrada na modalidade à distância na Universidade do Extremo Sul Catarinense. A descoberta de conhecimento, por meio das técnicas de KDD, foi aplicada para realizar a identificação do modelo final de agrupamento para os dados educacionais, com o uso de medidas de qualidade, fazendo a validação por meio de critérios internos e externos nos resultados do agrupamento hierárquico aglomerativo com o algoritmo *AGNES* e o agrupamento *fuzzy*, pelo algoritmo *fuzzy c-means*.

6.1 BASE DE DADOS

A base de dados empregada no desenvolvimento desta pesquisa foi disponibilizada pelo Setor de Educação à distância (SEAD), juntamente com o departamento de Tecnologia da Informação (TI) da Universidade do Extremo Sul Catarinense (UNESC), de acordo com o anexo A. A pesquisa foi submetida ao Comitê de Ética em Pesquisa (CEP), da UNESC, em agosto de 2018, sendo aprovado conforme o parecer de 2.857.694 (anexo B).

Os dados empregados na pesquisa são referentes à disciplina de Introdução a Engenharia de Segurança do Trabalho, ministrada na UNESC na modalidade à distância. A disciplina iniciou nesta modalidade, no segundo semestre de 2017, como parte da grade do curso presencial de Engenharia Civil da UNESC, estando dividida em duas turmas, uma para os cursos das engenharias que acontecem no período matutino e outra com os alunos dos cursos noturnos, totalizando em três semestres 315 alunos (tabela 11).

Tabela 11 – Quantidade de alunos por semestre.

Semestre	Noturno	Matutino	Total
2017/2	93	30	123
2018/1	94	20	114
2018/2	51	27	78
		TOTAL	**315**

Fonte: Do autor.

A disciplina tem 32 créditos, de modo que as atividades estão divididas em 9 semanas/aulas (tabela 12). Na maioria das aulas, os alunos têm uma atividade avaliativa para postar e no final da disciplina uma avaliação presencial. As atividades juntamente com a avaliação formam a nota final do aluno, que define a aprovação ou reprovação.

Tabela 12 – Cronograma da disciplina, separado pelas semanas.

Semana	Atividade Avaliativa	Descrição
Semana 1	Atividade 1	
Semana 2	Atividade 2	Quiz
Semana 3	Atividade 3	Fechamento da AD1 (atividade 1 + atividade 2 + atividade 3)
Semana 4	Atividade 4	
Semana 5	Atividade 5	
Semana 6	Atividade 6	Fechamento da AD2 (atividade 4 + atividade 5 + atividade 6)
Semana 7	-	Sem atividades, tutoria a distância.
Semana 8	Avaliação	Avaliação presencial
Semana 9	Avaliação recuperação	Avaliação presencial de recuperação da avaliação da semana 8, para alunos que não atingiram a média 6.

Fonte: Do autor.

As atividades semanais formam duas notas conhecidas como AD1 e AD2, equivalendo a 20% (peso 2) cada, na nota final do curso. A soma das atividades das semanas 1, 2 e 3 formam a primeira nota AD1 e as atividades das semanas 4, 5 e 6, formam a segunda nota AD2. A avaliação presencial acontece no final do curso, equivalendo a 60% (peso 6) da nota do aluno na disciplina. No caso do aluno não comparecer ou não atingir a média esperada, ele tem a oportunidade de refazer a avaliação de recuperação, para chegar à média final da disciplina (tabela 13).

Tabela 13 – Divisão das notas para média final.

Nota Final	Avaliação	Peso	Divisão da Nota (%)
Peso nota final é 10	AD1	Peso 2	20%
	AD3	Peso 2	20%
	AP	Peso 6	60%

Fonte: Do autor.

O cálculo é baseado na média ponderada, de modo que a nota máxima é 10, precisando o aluno ter média maior ou igual a 6 para ser considerado aprovado.

$$Nota\ Final = \frac{(AD1 * peso\ 2) + (AD2 * peso\ 2) + (AP * peso\ 6)}{(peso\ 2 + peso\ 2 + peso\ 6)} \quad (6)$$

O aluno que possuir nota menor que seis (6) é considerado reprovado e precisa refazer a disciplina. Na base de dados, no universo de 315 alunos, 35 foram reprovados por não atingirem a média e alguns estão presentes em mais de um semestre.

Para que os dados pudessem ser processados, foi necessário realizar o estudo das tabelas do ambiente virtual de aprendizagem *Moodle*, versão 3.6.2+ (Build: 20190215), para solicitar ao setor de TI apenas as que fossem úteis para o desenvolvimento da pesquisa. Após a entrega das tabelas, foi aplicado *scripts* para extrair os dados do banco. Os *scripts* foram feitos no *Mysql Workbench*, uma ferramenta usada por arquitetos de banco de dados e desenvolvedores, pois fornece ferramentas visuais para criar, executar e aperfeiçoar consultas *SQL*, permitindo o gerenciamento das conexões de bancos de dados padrões, modelagem de dados, desenvolvimento *SQL*, entre outras funções (MYSQL..., 2019).

Após a execução de cada *script*, os dados foram exportados e adicionados a uma mesma tabela de Excel, de forma que cada atributo foi relacionado pelo código do aluno e disciplina.

Tabela 14 – Atributos selecionados para aplicação do KDD.

Atributo	**Descrição**
Id_aluno	Código de identificação do aluno no Moodle
id_disciplina	Código de identificação da disciplina no Moodle
semestre	Semestre da disciplina
nunca_acessou	Identificação para alunos que nunca acessaram o Moodle (0 = nunca acessou, 1= acessou)
qtde_acessos	Quantidade de acessos na disciplina do Moodle
qtde_posts_foruns	Quantidade de posts que o aluno fez no fórum
qtde_discuss_forum_abertas	Quantidade de discussões abertas em fóruns
qtde_forum_assina	Quantidade de fóruns que o aluno assina
nota_quiz	Nota de cada quiz realizado
tentativa	Quantidade de tentativas que o aluno fez
nota_atividade	Nota de cada atividade realizada
nota_final	Nota final do aluno na disciplina
qtde_msgs_env_aluno_monitor_lidas	Quantidade de mensagens enviadas lidas de um aluno para monitores
qtde_msgs_env_aluno_prof_lidas	Quantidade de mensagens enviadas lidas de um aluno para professores
qtde_msgs_env_monitor_aluno_lidas	Quantidade de mensagens enviadas lidas de monitores para um aluno
qtde_msgs_env_prof_aluno_lidas	Quantidade de mensagens enviadas lidas de professores para um aluno
qtde_msgs_env_aluno_monitor_naolidas	Quantidade de mensagens enviadas não lidas de um aluno para monitores
qtde_msgs_env_aluno_prof_naolidas	Quantidade de mensagens enviadas não lidas de um aluno para professores
qtde_msgs_env_monitor_aluno_naolidas	Quantidade de mensagens enviadas não lidas de monitores para um aluno
qtde_msgs_env_prof_aluno_naolidas	Quantidade de mensagens enviadas não lidas de professores para um aluno
posts_forum_ref_post_prof	Quantidades de posts de um aluno em fóruns referenciando posts de professores
qtde_msgs_env_aluno_aluno_lidas	Quantidade de mensagens enviadas lidas de um aluno para alunos

qtde_msgs_env_aluno_aluno_naolidas	Quantidade de mensagens enviadas não lidas de um aluno para alunos
aprovado	Resultado obtido por meio da nota final do aluno. Representa se o aluno foi aprovado (1) ou não (2)

Fonte: Do autor.

Com os *scripts* concluídos e os dados compilados em uma mesma tabela, foi possível iniciar a aplicação da descoberta de conhecimento.

6.2 METODOLOGIA

As etapas metodológicas empregadas no desenvolvimento foram as seguintes: seleção da base de dados da disciplina a distância de Introdução Engenharia de Segurança do Trabalho para aplicação dos algoritmos de agrupamento, pré-processamento da base de dados selecionada, levantamento bibliográfico, aplicação do método de agrupamento particional por meio do algoritmo *fuzzy c-means*, aplicação do método de agrupamento hierárquico aglomerativo por meio do algoritmo AGNES, análise dos modelos obtidos por meio de medidas de qualidade para agrupamento em mineração de dados.

O levantamento bibliográfico levou em consideração a fundamentação e o entendimento de todos os temas envolvidos na pesquisa, tais como o processo de descoberta de conhecimento em bases de dados, mineração de dados, agrupamento, agrupamento hierárquico, agrupamento aglomerativo, método *fuzzy*, algoritmo AGNES, o algoritmo *fuzzy c-means*, medidas de qualidade e desempenho para algoritmos de agrupamento de dados, educação à distância, ambientes virtuais de aprendizagem e *educational data mining*.

6.2.1 Pré-processamento

O pré-processamento aborda diversas etapas, que devem ser aplicadas para tornar os dados apropriados para a mineração, melhorando e corrigindo erros e inconsistências para a aplicação do algoritmo (TAN; STEINBACH; KUMAR, 2009).

A preparação dos dados envolveu três ferramentas: *Excel*[11], *Weka*[12] e *R Studio*[13].

O processo iniciou pela seleção de dados, que é a etapa que identifica as informações que são relevantes para o KDD e presentes na base de dados aplicada na pesquisa (GOLDSCHMIDT; BEZERRA; PASSOS, 2015). Todos os atributos apresentados na tabela 14 foram compilados após os *scripts*, analisados e alguns selecionados para a pesquisa. Para a seleção dos dados, levou-se em consideração o cronograma da disciplina (tabela 12), ambiguidade de atributos, interação dos alunos e atividades da disciplina.

Os atributos relacionados a fórum e *posts* foram descartados, pois não eram atividades e iterações propostas na disciplina, não possuindo dados relevantes. As mensagens não foram consideradas, pois o que o *Moodle* arquiva, são as mensagens totais por usuário, não sendo possível identificar quais eram relacionadas a alunos ou professores da disciplina. A tentativa do *Quiz* foi descartada, pois na disciplina era permitida ao aluno somente uma tentativa.

O *id* do aluno foi desconsiderado, pois é o código de identificação do *Moodle* e conforme a plataforma Brasil, o participante da pesquisa não deve ser identificado.

Após análise dos dados e entendimento do funcionamento da disciplina, os atributos necessários para aplicar a descoberta de conhecimento foram definidos, conforme descrito na tabela 15.

---

[11] *Excel* é um aplicativo de criação de planilhas eletrônicas, com a licença fornecida pela UNESC.
[12] *WEKA* é uma ferramenta de *mineração de dados* em Java, *open source* e licenciado pela *General Public License*. Disponível em: http://www.cs.waikato.ac.nz/ml/weka.
[13] R Studio é um software livre, usado para estatística e mineração de dados gratuita. Disponível em: www.r-project.org.

Tabela 15 – Atributos selecionados para aplicação do KDD.

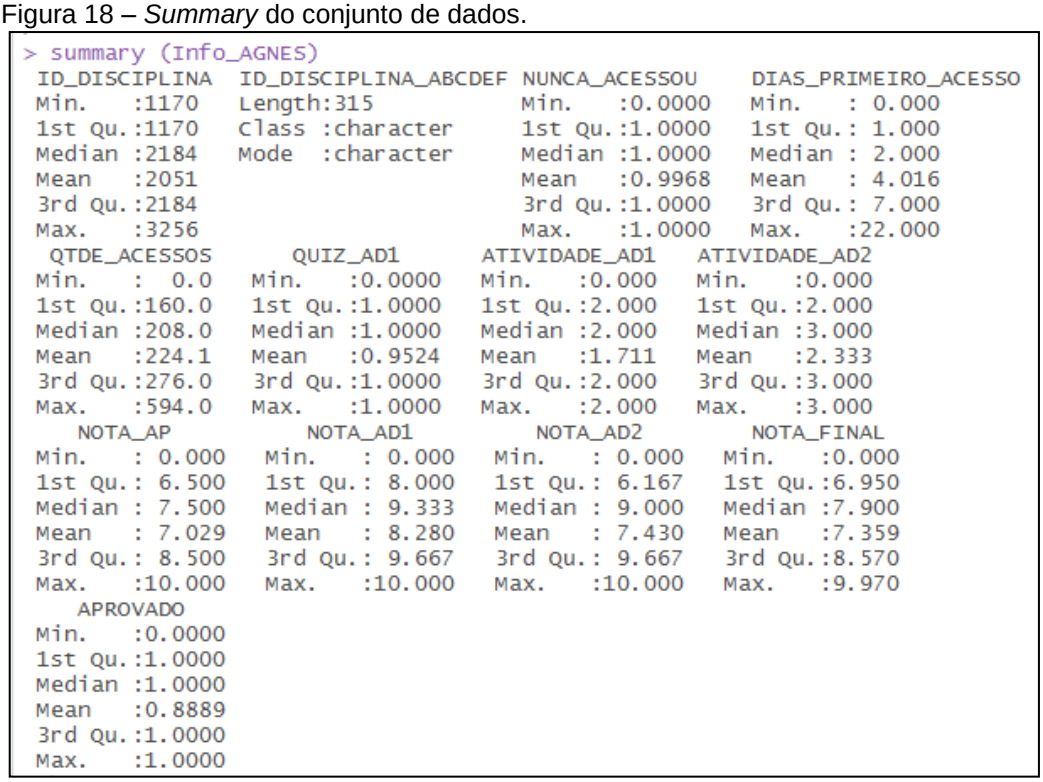
Atributos	Descrição	Valor
ID_DISCIPLINA	Número de identificação da disciplina.	int
NUNCA_ACESSOU	Identificação dos alunos que acessaram a aula, durante o período.	int
DIAS_PRIMEIRO_ACESSO	Tempo em que o estudante levou entre o início da disciplina e o primeiro acesso.	num
QTDE_ACESSOS	Quantidade de acessos realizados por cada aluno.	num
QUIZ_AD1	Quantidade de Quiz feito, por cada estudante.	int
ATIVIDADE_AD1	Quantidade de atividades da AD1 realizada por aluno.	num
ATIVIDADE_AD2	Quantidade de atividades da AD2 realizada por aluno.	num
NOTA_AP	Nota da avaliação presencial.	num
NOTA_AD1	Nota da atividade AD1.	num
NOTA_AD2	Nota da atividade AD2.	num
NOTA_FINAL	Nota final da disciplina, por aluno.	num
APROVADO	Identificações do aluno, se ele foi aprovado ou reprovado.	int

Fonte: Do autor.

Posteriormente, realizou-se a limpeza de dados que tem como objetivo de abrandar dois problemas: valores ausentes e valores ruidosos. O problema de valores ausentes ocorre quando os atributos de um conjunto de dados, não apresentam valores para alguns exemplares ou algum atributo de interesse, inviabilizando o processo de análise pelo algoritmo, por não poder lidar com a ausência de valores. O pesquisador tem três opções, para resolver o problema, a primeira é a remoção do exemplar em que ocorre a falta do valor a segunda opção é o preenchimento manual dos valores ausentes, e por último, a terceira opção é o preenchimento automático por meio de um valor constante, médio ou mediano (SILVA; PERES; BOSCARIOLI, 2017).

Nos dados da pesquisa, foram identificados valores ruidosos no atributo *dias primeiro acesso*, sendo preenchido com a mediana do mesmo, o valor 2. O valor foi identificado no R e com a aplicação da função *summary*, que ajuda na inspeção do conjunto de dados, apresentando o resumo como resultado (figura 18).

Figura 18 – *Summary* do conjunto de dados.



```
> summary (Info_AGNES)
 ID_DISCIPLINA  ID_DISCIPLINA_ABCDEF NUNCA_ACESSOU    DIAS_PRIMEIRO_ACESSO
 Min.   :1170   Length:315          Min.   :0.0000   Min.   : 0.000
 1st Qu.:1170   Class :character    1st Qu.:1.0000   1st Qu.: 1.000
 Median :2184   Mode  :character    Median :1.0000   Median : 2.000
 Mean   :2051                       Mean   :0.9968   Mean   : 4.016
 3rd Qu.:2184                       3rd Qu.:1.0000   3rd Qu.: 7.000
 Max.   :3256                       Max.   :1.0000   Max.   :22.000
  QTDE_ACESSOS      QUIZ_AD1      ATIVIDADE_AD1   ATIVIDADE_AD2
 Min.   :  0.0   Min.   :0.0000   Min.   :0.000   Min.   :0.000
 1st Qu.:160.0   1st Qu.:1.0000   1st Qu.:2.000   1st Qu.:2.000
 Median :208.0   Median :1.0000   Median :2.000   Median :3.000
 Mean   :224.1   Mean   :0.9524   Mean   :1.711   Mean   :2.333
 3rd Qu.:276.0   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:3.000
 Max.   :594.0   Max.   :1.0000   Max.   :2.000   Max.   :3.000
    NOTA_AP          NOTA_AD1         NOTA_AD2        NOTA_FINAL
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   :0.000
 1st Qu.: 6.500   1st Qu.: 8.000   1st Qu.: 6.167   1st Qu.:6.950
 Median : 7.500   Median : 9.333   Median : 9.000   Median :7.900
 Mean   : 7.029   Mean   : 8.280   Mean   : 7.430   Mean   :7.359
 3rd Qu.: 8.500   3rd Qu.: 9.667   3rd Qu.: 9.667   3rd Qu.:8.570
 Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :9.970
    APROVADO
 Min.   :0.0000
 1st Qu.:1.0000
 Median :1.0000
 Mean   :0.8889
 3rd Qu.:1.0000
 Max.   :1.0000
```

Fonte: Do autor.

Na ferramenta R, foi aplicado o teste de *Kolmogorov-Smirnov* (KS), para verificar a necessidade de normalizar os dados. O teste KS fornece o valor de prova (*valor-p*, *p-value*), que pode ser interpretado como a medida de grau de concordância entre os dados e a hipótese. Quanto menor o valor de *p-value*, menor é a consistência entre a hipótese e os dados (DE MESQUITA LOPES; BRANCO; SOARES, 2013). Se o teste é significativo, com o valor < 0,05, significa que os dados precisam ser normalizados, pois diferem de uma normalização normal, em que o teste significativo deve ser maior que 0,05 (FIELD, 2009).

O teste KS foi aplicado no R para cada atributo, conforme tabela 16, constatando-se a necessidade da normalização.

Tabela 16 – Teste de *Kolmogorov-Smirnov.*

ATRIBUTO	***P-VALUE***
DIAS_PRIMEIRO_ACESSO	0,0000000325700000
QTDE_ACESSOS	0,0102800000000000
QUIZ_AD1	0,0000000000000002
ATIVIDADE_AD1	0,0000000000000002
ATIVIDADE_AD2	0,0000000000000002
NOTA_AP	0,0000000000108800
NOTA_AD1	0,0000000000000002
NOTA_AD2	0,0000000000000002
NOTA_FINAL	0,0000042030000000
APROVADO	0,0000000000000002

Fonte: Do autor.

O objetivo da normalização é fazer com que o conjunto de dados tenha uma determinada propriedade, para isso, as variáveis que são diferentes devem ser combinadas de alguma forma, para que valores grandes não sejam dominantes no resultado (TAN; STEINBACH; KUMAR, 2009). Na pesquisa, conforme figura 18, o atributo *quantidade de acesso* que tem o valor máximo de 594, enquanto o atributo *dias primeiro acesso*, tem o valor máximo de 22, de modo que para mineração de dados o atributo *quantidade de acesso* torna se dominante no resultado.

Com a normalização, os valores devem ficar no intervalo [0,1], fazendo com que os dados tenham a mesma ordem de grandeza. Não constando mais valores dominantes, conforme a figura 19, o atributo quantidade de acesso possui valor máximo sendo 1.

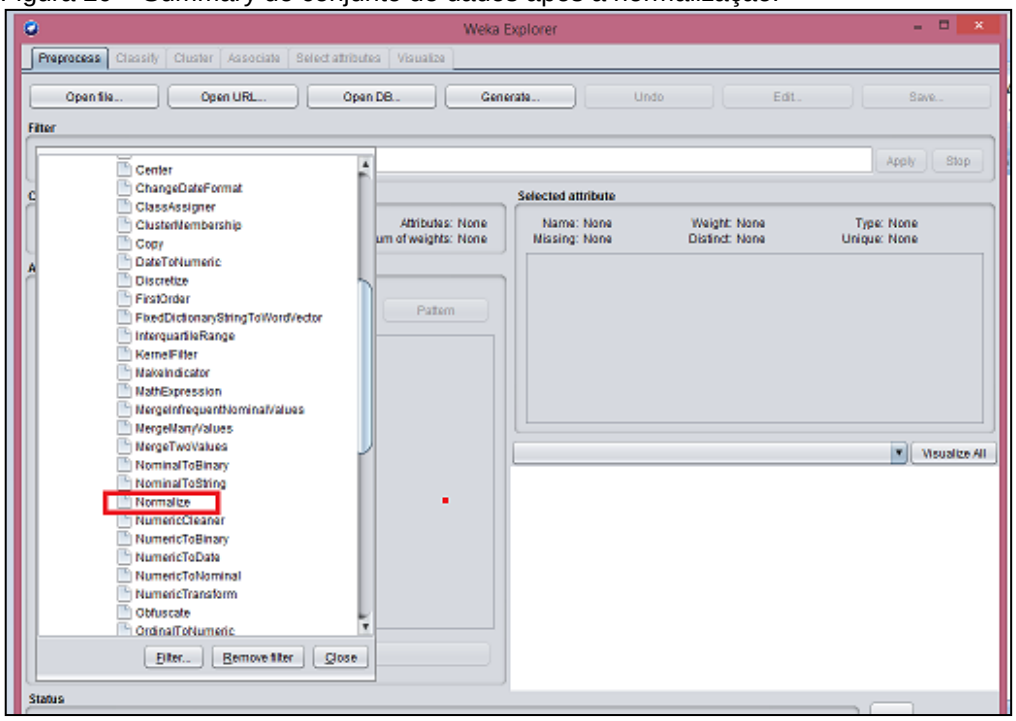
Figura 19 – Summary do conjunto de dados após a normalização.

```
> #Resumo da tabela
> summary(Info_AGNES)
 ID_DISCIPLINA    NUNCA_ACESSOU     DIAS_PRIMEIRO_ACESSO  QTDE_ACESSOS
 Min.    :1.000   Min.    :0.0000   Min.    :0.00000     Min.    :0.0000
 1st Qu.:1.000   1st Qu.:1.0000   1st Qu.:0.04546     1st Qu.:0.2694
 Median :3.000   Median :1.0000   Median :0.09091     Median :0.3502
 Mean   :2.959   Mean   :0.9968   Mean   :0.18254     Mean   :0.3773
 3rd Qu.:4.000   3rd Qu.:1.0000   3rd Qu.:0.31818     3rd Qu.:0.4646
 Max.    :6.000   Max.    :1.0000   Max.    :1.00000     Max.    :1.0000
    QUIZ_AD1       ATIVIDADE_AD1    ATIVIDADE_AD2       NOTA_AP
 Min.    :0.0000   Min.    :0.0000   Min.    :0.0000   Min.    :0.0000
 1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.6667   1st Qu.:0.6500
 Median :1.0000   Median :1.0000   Median :1.0000   Median :0.7500
 Mean   :0.9524   Mean   :0.8556   Mean   :0.7778   Mean   :0.7029
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.8500
 Max.    :1.0000   Max.    :1.0000   Max.    :1.0000   Max.    :1.0000
    NOTA_AD1         NOTA_AD2       NOTA_FINAL        APROVADO
 Min.    :0.0000   Min.    :0.0000   Min.    :0.0000   Min.    :0.0000
 1st Qu.:0.8000   1st Qu.:0.6167   1st Qu.:0.6971   1st Qu.:1.0000
 Median :0.9333   Median :0.9000   Median :0.7924   Median :1.0000
 Mean   :0.8280   Mean   :0.7430   Mean   :0.7382   Mean   :0.8889
 3rd Qu.:0.9667   3rd Qu.:0.9667   3rd Qu.:0.8596   3rd Qu.:1.0000
 Max.    :1.0000   Max.    :1.0000   Max.    :1.0000   Max.    :1.0000
```

Fonte: Do autor.

Para os atributos *QTDE_ACESSOS*, *QUIZ_AD1* e *APROVADO*, não foram necessários aplicar a normalização, visto que os valores já estão no intervalo [0,1], apresentando a mesma grandeza. A normalização foi realizada na ferramenta *Weka* por meio do filtro *normalize* (figura 20).

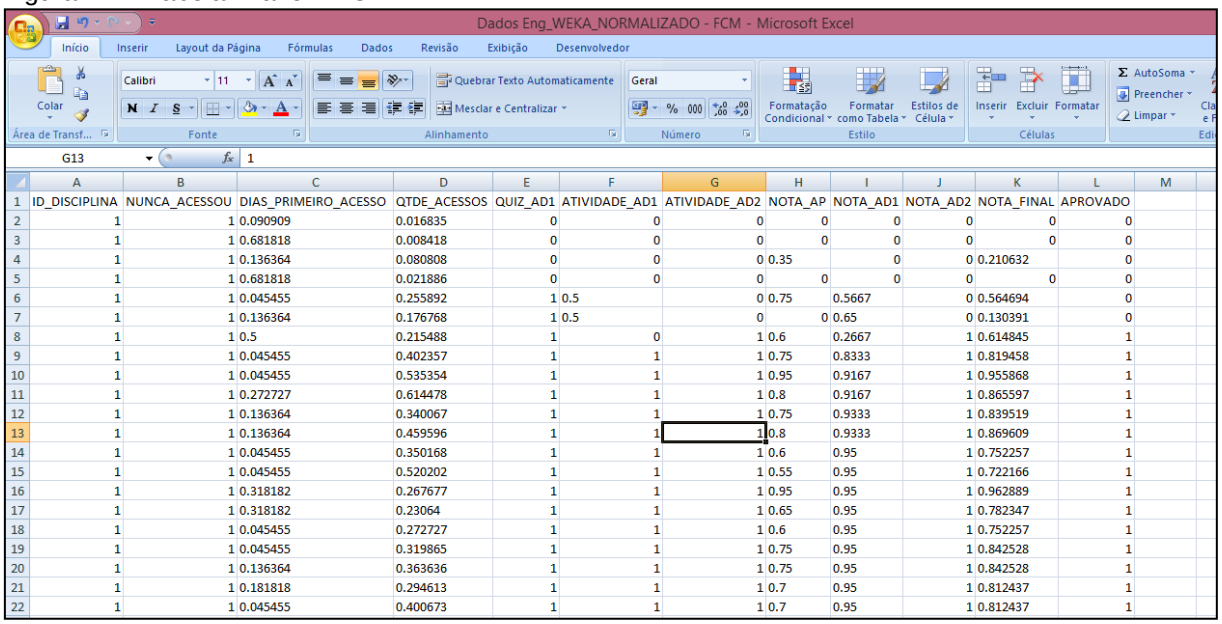
Figura 20 – Summary do conjunto de dados após a normalização.



Fonte: Do autor.

Após o pré-processamento, organizou-se a tabela final no formato *xlsx*, ficando mais claro, em relação aos dos dados originais e com os atributos normalizados.

Figura 21 – Tabela final em xlsx.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ID_DISCIPLINA	NUNCA_ACESSOU	DIAS_PRIMEIRO_ACESSO	QTDE_ACESSOS	QUIZ_AD1	ATIVIDADE_AD1	ATIVIDADE_AD2	NOTA_AP	NOTA_AD1	NOTA_AD2	NOTA_FINAL	APROVADO	
2	1	1	0.090909	0.016835	0	0	0	0	0	0	0	0	
3	1	1	0.681818	0.008418	0	0	0	0	0	0	0	0	
4	1	1	0.136364	0.080808	0	0	0	0.35	0	0	0.210632	0	
5	1	1	0.681818	0.021886	0	0	0	0	0	0	0	0	
6	1	1	0.045455	0.255892	1	0.5	0	0.75	0.5667	0	0.564694	0	
7	1	1	0.136364	0.176768	1	0.5	0	0	0.65	0	0.130391	0	
8	1	1	0.5	0.215488	1	0	1	0.6	0.2667	1	0.614845	1	
9	1	1	0.045455	0.402357	1	1	1	0.75	0.8333	1	0.819458	1	
10	1	1	0.045455	0.535354	1	1	1	0.95	0.9167	1	0.955868	1	
11	1	1	0.272727	0.614478	1	1	1	0.8	0.9167	1	0.865597	1	
12	1	1	0.136364	0.340067	1	1	1	0.75	0.9333	1	0.839519	1	
13	1	1	0.136364	0.459596	1	1	1	0.8	0.9333	1	0.869609	1	
14	1	1	0.045455	0.350168	1	1	1	0.6	0.95	1	0.752257	1	
15	1	1	0.045455	0.520202	1	1	1	0.55	0.95	1	0.722166	1	
16	1	1	0.318182	0.267677	1	1	1	0.95	0.95	1	0.962889	1	
17	1	1	0.318182	0.23064	1	1	1	0.65	0.95	1	0.782347	1	
18	1	1	0.045455	0.272727	1	1	1	0.6	0.95	1	0.752257	1	
19	1	1	0.045455	0.319865	1	1	1	0.75	0.95	1	0.842528	1	
20	1	1	0.136364	0.363636	1	1	1	0.75	0.95	1	0.842528	1	
21	1	1	0.181818	0.294613	1	1	1	0.7	0.95	1	0.812437	1	
22	1	1	0.045455	0.400673	1	1	1	0.7	0.95	1	0.812437	1	

Fonte: Do autor.

6.2.2 Mineração de dados

A ferramenta escolhida para a execução da mineração de dados foi o R por ter uma ampla variedade de métodos da mineração de dados, medidas de qualidade e por ser um *software* livre. Para isso foi realizado um estudo sobre a ferramenta, a utilização dos algoritmos AGNES e *fuzzy c-means,* juntamente com as medidas de qualidade para validação do agrupamento.

Antes de aplicar a mineração pelos algoritmos, é necessário realizar a validação da tendência de agrupamento, para que fosse verificado se os dados possuem grupos significativos para a aplicação da mineração de dados, que pode ser feito visualmente ou por métodos estatísticos, na pesquisa foi aplicado o método de *Hopkins*.

A mineração inicia com a definição das chamadas matriz de dados e matriz de similaridade, que são testadas com a aplicação das matrizes *euclidiana*, *manhattan*, *correlattion* e *seuclidian*.

Definido a matriz de similaridade, os algoritmos são executados, de modo que para o agrupamento hierárquico, deve ser definido o método de conexão, na pesquisa são testados os métodos de *ward*, distância média, menor distância e maior distância.

Os resultados apresentados pelos algoritmos são validados por meio de índices de medidas de qualidade, separadas em critério interno, externo e relativo.

Para definição da divisão de grupos, foram realizados teste com variação de grupos, identificando o valor ideal, comparando as medidas de qualidade.

A tabela 17 mostra os métodos por etapa, para aplicação da mineração de dados.

Tabela 17 – Métodos da mineração de dados aplicados na pesquisa.

Agoritimo	**Matriz de Similaridade**	**Métodos de conexão**	**Critérios**	**Índices de validação**	**Grupos**
AGNES	Euclidiano Manhattan	Ward Distância média Maior distância Menor distância	Externo Interno	Rand Jaccard Russel Folk Mal *Dunn* *Silhouette*	*K=2* *K=2 até k=10*
Fuzzy C-means	Correlation Seuclidean Manhattan Euclidiano	-	Interno *Fuzzy*	*Silhouette* Coefficient Entropy	*k=2 até k=10*

Fonte: Do autor.

A descrição das etapas de mineração de dados aplicada na pesquisa está descrida nos próximos capítulos.

6.2.2.1 Ferramenta Estatística R

O R é um conjunto integrado de recursos de *software*, para manipulação de dados, computação estatística e manipulação de gráficos. A ferramenta está disponível como *software* livre, sobre os termos do GNU *General Public da Free Software Foundation*. Compila e executa em diversos sistemas operacionais, tais como *UNIX*, *Windows*, *Linux* e *MacOS* (R, 2018).

O *software* está disponível gratuitamente no site do R (www.r-project.org). O usuário deve clicar em *CRAN*, para ser direcionado para uma lista de sites que hospedam os arquivos de instalação do R.

Normalmente, os dados ficam armazenados em planilhas, arquivos de texto ou em base de dados. Sendo possível a importação dos mesmos para o R basicamente de duas maneiras, por meio da leitura de arquivos ou da base de dados, sendo mais comuns por meio de uma *Structured Query Language* (*SQL*) e *Comma Separated Values* (*CSV*). Para esta pesquisa, os dados foram importados para o R por meio de um arquivo *CSV*.

Posteriormente a importação, o usuário pode trabalhar com funções matemáticas, estatísticas, vetores, matrizes, data frames, listas, arrays, funções e gráficos, conforme a necessidade da análise a ser realizada.

As funções do R estão organizadas em pacotes, conhecidos em inglês como *packages*, que ficam disponíveis ao usuário a partir do momento em que o pacote é carregado (SILVA; PERES; BOSCARIOLI, 2016).

Nesta pesquisa, foram empregados os pacotes *stats* para mineração e representação gráfica do dendograma, gerado pelo algoritmo AGNES, *fpc* e *clv* para validação dos critérios internos e externos, respectivamente.

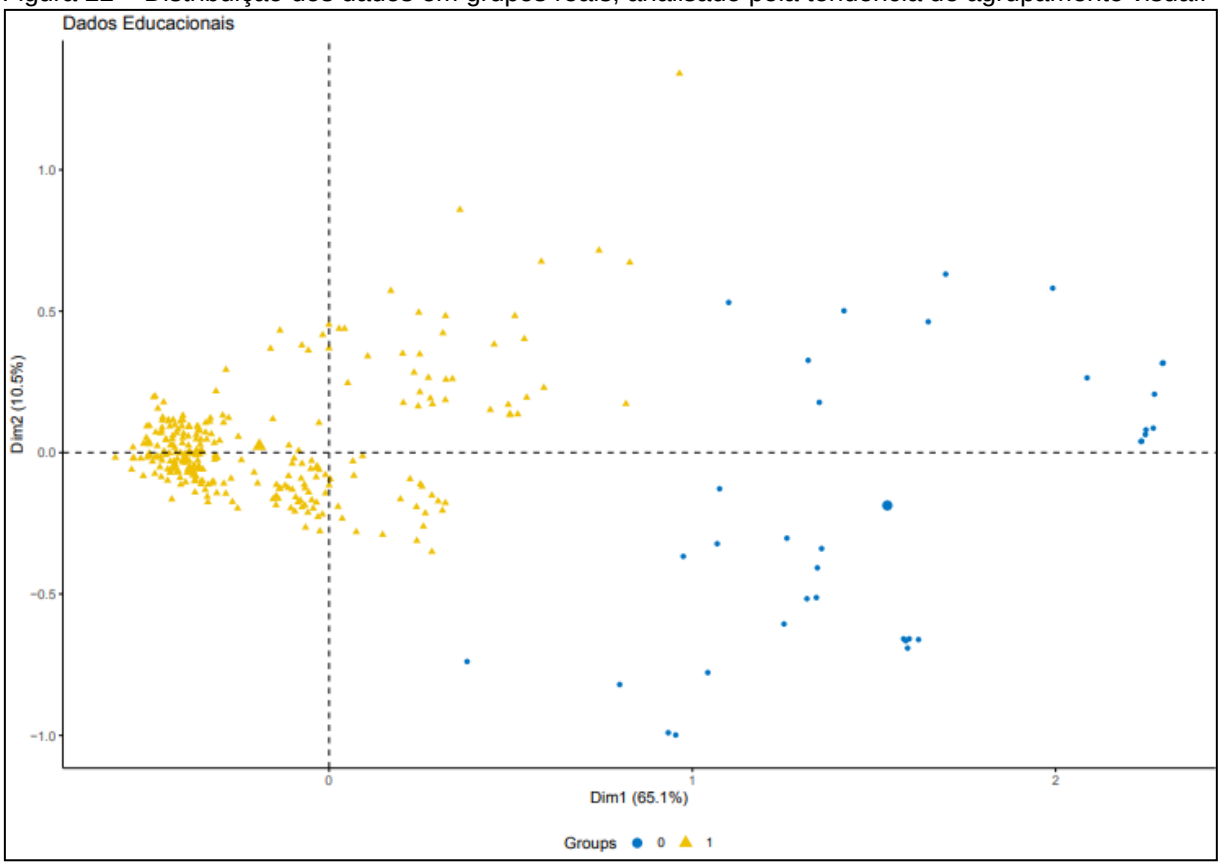
Na mineração de dados por meio do algoritmo *fuzzy c-means*, foi empregado o pacote *e1071*, o *fclust* para validação e o *ppclust* para a representação gráfica da separação dos grupos.

6.2.2.2 Avaliando a tendência de agrupamento

Para aplicação dos algoritmos de agrupamento, é importante verificar se o conjunto de dados possui grupos significativos. Para isso, é necessário avaliar a tendência de agrupamento ou a possibilidade da análise do agrupamento. Um dos problemas na análise, é que os métodos podem retornar grupos, mesmo se não contiverem dados, de modo que o algoritmo deve dividir os grupos, independente se tiver dados, pois é o que ele deve fazer (KASSAMBARA, 2017, tradução nossa).

A análise da distribuição dos dados possibilita a verificação da tendência de agrupamento visual, mostrando se existe similaridade nos dados para gerar os grupos. A figura 22 mostra a distribuição da base antes da aplicação da mineração de dados, pelos algoritmos AGNES e *fuzzy c-means*, identificando que o conjunto de dados possui dois grupos reais, de modo que os dados são bons para serem divididos em grupos. Considerando o atributo *situação*, que indica se o aluno é aprovado (1) ou reprovado (0).

Figura 22 – Distribuição dos dados em grupos reais, analisado pela tendência de agrupamento visual.



Fonte: Do autor.

Porém, a análise por tendência de agrupamento visual, pode gerar dúvida, visto que depende da interpretação do pesquisador. Desta forma, existem métodos estatísticos que ajudam a verificar se existe tendência, sobre o conjunto de dados. Nesta pesquisa, foi aplicado o método de *Hopkins*, que testa a aleatoriedade espacial, comprovando se a estrutura está uniformemente distribuída. A tabela 18 apresenta os resultados obtidos pelo método de *Hopkins* sobre a base de dados.

Tabela 18 – Resultado do método de Hopkins.

Arquivo de dados	**Hopkins**
Dados para aplicar o algoritmo AGNES	0,1080572
Dados para aplicar algoritmo *fuzzy c-means*	0,1135723

Fonte: Do autor.

Valores de dados próximos a um (1) indicam que os dados estão distribuídos regularmente e não contem grupos significativos, enquanto próximo à zero (0), indicam que os dados estão altamente agrupados formando grupos significativos (TAN; STEINBACH; KUMAR, 2009, tradução nossa). O resultado do método de *Hopkins* para a base de dados dos algoritmos AGNES (0,1080572) e *fuzzy c-means* (0,1135723), apresenta valores próximos à zero (0), mostrando que os dados tem significância para formar grupos.

6.2.2.3 Matrizes de distância dos dados

Os algoritmos de agrupamento precisam que os dados estejam armazenados em uma estrutura que seja capaz de manter os objetos para serem processados, chamadas de matriz de dados e matriz de similaridade. A matriz de dados possui linhas que representam os objetos a serem agrupados e as colunas são os seus atributos. Enquanto a matriz de similaridade representa a distância entre cada par de objetos (GOLDSCHMIDT; BEZERRA; PASSOS, 2015). A figura 23 mostra a matriz de dados da pesquisa, a linha 1 possui os dados do aluno, enquanto nas colunas estão os atributos com as características dos dados.

Figura 23 - Matriz de dados.

	NUNCA_ACESSOU	DIAS_PRIMEIRO_ACESSO	QTDE_ACESSOS	QUIZ_AD1	ATIVIDADE_AD1
1	1	0.090909	0.016835	0	0.0
2	1	0.681818	0.008418	0	0.0
3	1	0.136364	0.080808	0	0.0
4	1	0.681818	0.021886	0	0.0
5	1	0.045455	0.255892	1	0.5
6	1	0.136364	0.176768	1	0.5
7	1	0.500000	0.215488	1	0.0
8	1	0.045455	0.402357	1	1.0
9	1	0.045455	0.535354	1	1.0
10	1	0.272727	0.614478	1	1.0
11	1	0.136364	0.340067	1	1.0
12	1	0.136364	0.459596	1	1.0

Fonte: Do autor.

A figura 24 representa a matriz de similaridade pelo método euclidiano, um dos que foram aplicados na pesquisa. É empregado o cálculo na matriz de tamanho *n x n*, em que *n* é número de exemplares no conjunto, para que a distância entre os pares de objetos seja encontrada, ou seja, entre as coordenadas (*x*, *y*). Desta forma, o tamanho da matriz de dados da pesquisa é 315 x 315, de modo que a distância entre as coordenadas (2,1) é 0,59. Quando as coordenadas são iguais, não existe a necessidade de armazenar a distância entre os objetos, visto que o valor é zero conforme representado pela coordenada (3,3), em que a diagonal da matriz sempre é zero.

Figura 24 – Matriz de similaridade pelo método euclidiano.

	1	2	3	4	5	6	7	8	9	10	11	12
1	0.00	0.59	0.42	0.59	1.58	1.31	2.24	2.66	2.82	2.76	2.69	2.73
2	0.59	0.00	0.69	0.01	1.71	1.42	2.21	2.74	2.89	2.79	2.75	2.79
3	0.42	0.69	0.00	0.68	1.38	1.35	2.11	2.52	2.64	2.60	2.55	2.58
4	0.59	0.01	0.68	0.00	1.71	1.42	2.21	2.73	2.89	2.78	2.75	2.79
5	1.58	1.71	1.38	1.71	0.00	0.88	1.89	1.85	1.91	1.91	1.86	1.88
6	1.31	1.42	1.35	1.42	0.88	0.00	2.03	2.09	2.25	2.17	2.10	2.14
7	2.24	2.21	2.11	2.21	1.89	2.03	0.00	1.28	1.40	1.32	1.29	1.32
8	2.66	2.74	2.52	2.73	1.85	2.09	1.28	0.00	0.29	0.33	0.15	0.16
9	2.82	2.89	2.64	2.89	1.91	2.25	1.40	0.29	0.00	0.30	0.32	0.21
10	2.76	2.79	2.60	2.78	1.91	2.17	1.32	0.33	0.30	0.00	0.31	0.21
11	2.69	2.75	2.55	2.75	1.86	2.10	1.29	0.15	0.32	0.31	0.00	0.13
12	2.73	2.79	2.58	2.79	1.88	2.14	1.32	0.16	0.21	0.21	0.13	0.00

Fonte: Do autor.

No agrupamento hierárquico, a pesquisa consiste na aplicação de duas matrizes de similaridade, *manhattan* e *euclidiano* (tabela 19). De acordo com Kassambara (2017, tradução nossa), eles são métodos clássicos para agrupamento.

Tabela 19 – Medidas usadas no agrupamento hierárquico.

Agrupamento	**Matriz de Distância**
Hierárquico – pelo algoritmo AGNES	Manhattan
	Euclidiana

Fonte: Do autor.

Para o agrupamento particional *fuzzy*, a pesquisa consiste na aplicação de quatro matrizes de similaridade, as que já são usadas no agrupamento hierárquico *manhattan*, *euclidiano* e *seuclidean* e *correlattion,* que são usadas com frequência para o agrupamento do algoritmo *fuzzy c-means* (tabela 20).

Tabela 20 – Medidas usadas no agrupamento particional.

Agrupamento	**Matriz de Distância**
Particionamento – algoritimo *fuzzy c-means*	Manhattan
	Euclidiana
	Seuclidean
	Correlattion

Fonte: Do autor.

No decorrer a execução do algoritmo, é essencial que a primeira etapa seja a transformação da matriz de dados em uma matriz de similaridade. Quanto

mais próximas de zero, for o cálculo de cada distância, mais similares são os objetos (GOLDSCHMIDT; BEZERRA; PASSOS, 2015).

Após o cálculo da matriz de similaridade, foi possível dar continuidade a pesquisa, aplicando os algoritmos AGENS e *fuzzy c-means*.

6.2.2.4 Aplicação do algoritmo AGNES

O agrupamento hierárquico aglomerativo é produzido, por meio da criação de uma estrutura de árvore. De modo que a estrutura é representada graficamente por um dendograma, na qual o conjunto de dados é definido como os nós das folhas e os nós internos revelam uma organização baseada na similaridade entre os exemplares. Em cada nível da árvore, tem-se a organização dos dados em grupos. Com o algoritmo AGNES, a construção da árvore é iniciada pelos nós das folhas, ocorrendo à iteração até que os grupos estejam fundidos em um mesmo grupo, partindo do critério de similaridade entre eles (SILVA; PEREZ, 2017). Diante disto, há necessidade do cálculo da matriz de similaridade antes de se aplicar o algoritmo.

A fim de aplicar o algoritmo hierárquico, é preciso definir a proximidade entre os grupos, denominado método de conexão. Para isso, o cálculo da proximidade é realizado entre dois grupos (TAN; STEINBACH; KUMAR, 2019).

Na aplicação dos métodos de conexão, existem quatro abordagens possíveis e que são aplicadas na pesquisa, conforme mostrado na tabela 21.

Tabela 21 – Experimentos do algoritmo AGNES.

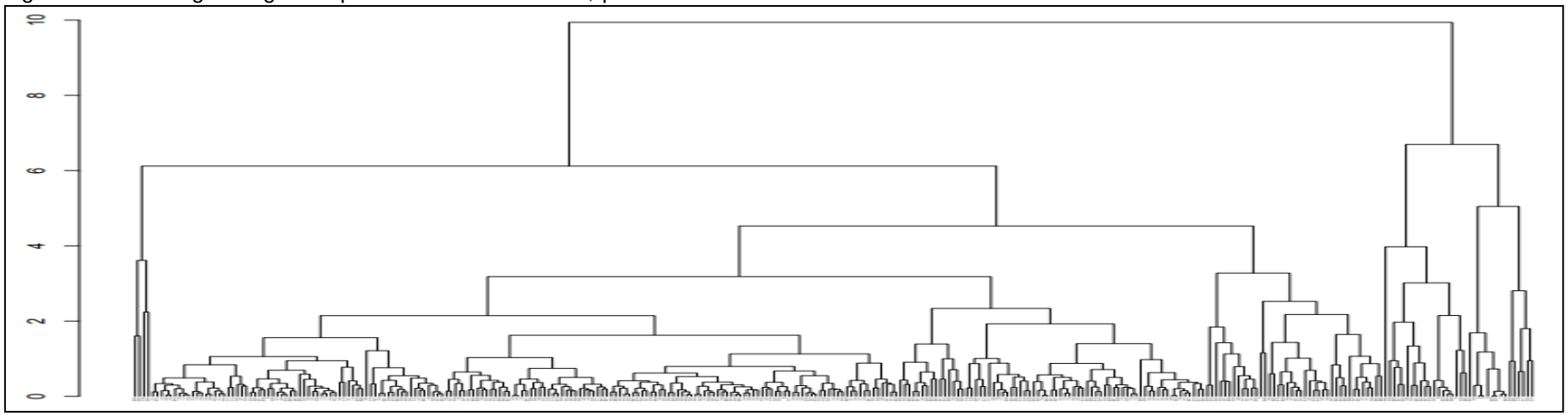
Experimento	Distância	Método de conexão
1	Euclidiana	Menor distância
2		Distância média
3		Maior distância
4		Ward
5	Manhattan	Menor distância
6		Distância média
7		Maior distância
8		Ward

Fonte: Do autor.

Os dendogramas foram gerados com as matrizes de similaridade *Manhattan e* euclidiana, aplicando os quatro métodos de conexão: menor distância, distância média, maior distância e ward. A figura 25 mostra o dendograma gerado aplicando da distância de similaridade euclidian, pelo método de conexão maior

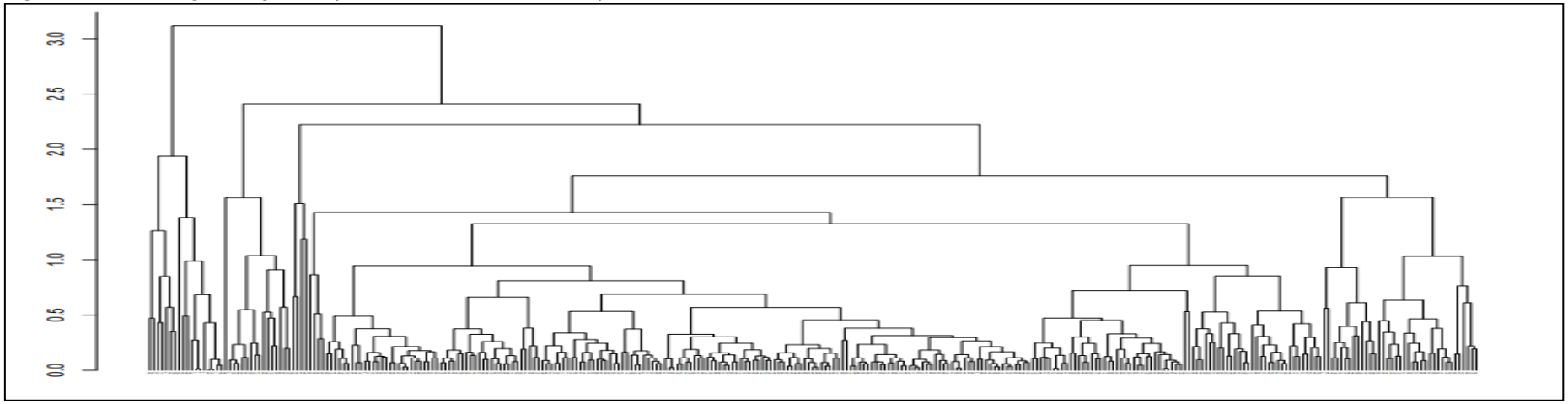
distância, enquanto a figura 26, apresenta o dendograma gerado pela distância de *manhattan*, com o método de conexão maior distância. Os dendogramas gerados pelos métodos distância média, maior distância e ward, estão no anexo da pesquisa.

Figura 25 – Dendograma gerado pela distância *euclidiana,* pelo método maior distância.



Fonte: Do autor.

Figura 26 – Dendograma gerado pela distância *manhattan*, pelo método maior distância.



Fonte: Do autor.

A partir da aplicação da matriz de similaridade e medidas de distância entre grupos. O algoritmo é executado, gerando a divisão dos dados em grupos, apresentando o resultado por meio do dendograma.

Após a aplicação dos dois métodos, foram realizados testes envolvendo a variação dos números de grupos (*k*). A variação aconteceu entre os valores de grupos *k=2* e *k=10*, testando ambos os métodos *manhattan* e *euclidiana*, para verificar qual apresenta o melhor resultado, dentro da natureza dos dados.

6.2.2.5 Aplicação das medidas de qualidade para o algoritmo AGNES

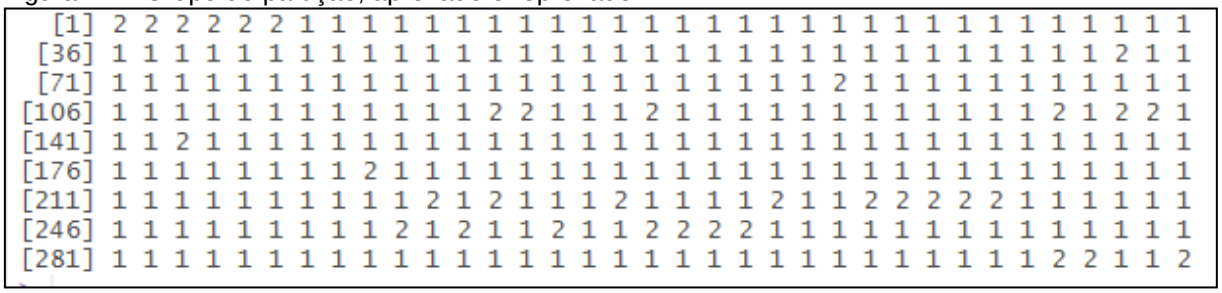
Com a execução do algoritmo concluída, aplicaram-se as medidas de qualidade, para atestar se os valores gerados pelo algoritmo têm significância sobre os dados.

O primeiro passo da pesquisa é a validação da estrutura da hierarquia, para isso realiza-se o cálculo da distância cofenética entre dois objetos, para verificar se os dados se adaptam ao método de conexão proposto. A medida é considerada com resultados satisfatórios, quando o coeficiente de correlação cofenética (CPCC) é próximo a um (1) (HAMZAH; RIJAL; NOOR, 2017, tradução nossa).

A validação do agrupamento, baseado em critério externo depende do conhecimento sobre como os dados estão organizados, pois a validação compara o grupo gerado pelo algoritmo e a estrutura natural dos dados. O conjunto gerado antes da aplicação do algoritmo é conhecido como grupo de partição P, enquanto o gerado pelo algoritmo é conhecido como G (KASSAMBARA, 2017).

O grupo de dados da pesquisa possui a partição para o resultado do aluno na disciplina, de modo que ele pode ser aprovado ou reprovado, dividindo a base em dois grupos. A figura 27 mostra a separação dos alunos em partições, em que um (1) são considerados os alunos aprovados e dois (2) os alunos reprovados.

Figura 27 – Grupo de partição, aprovado e reprovado.



Fonte: Do autor.

Verificado o número de partições, ficou determinado o número de *k*, para aplicar o critério externo como *k = 2*, grupo de alunos aprovado e reprovado.

Após a definição do grupo de partição, é realizado o cálculo com quatro índices: *Rand*, *Jaccard* e *Russel* para que seja medido o grau de similaridade entre P e G (tabela 22).

Tabela 22 – Medidas de qualidade externa.

Medida de Qualidade Externa				
Experimento	Distância	Medida de conexão	Índice	Grupos (K)
1	Euclidiana	Menor distância	Rand, Jaccard e Russel	K = 2
2		Distância média		
3		Maior distância		
4		Ward		
5	Manhattan	Menor distância		
6		Distância média		
7		Maior distância		
8		Ward		

Fonte: Do autor.

Estando os valores entre o intervalo de [0,1], quanto maiores os resultados apresentados pelos índices, mais semelhantes são os valores entre P e G (XU; WUNSCH, 2008, tradução nossa).

Nesta pesquisa foram realizados experimentos com as quatro medidas de conexão, aplicadas para as distâncias *manhattan* e *euclidiana*, procurando a identificação do melhor modelo para agrupamento pelo algoritmo AGNES.

Para o critério interno, não é preciso saber qualquer tipo de informação sobre o conjunto de dados que está sobre análise. Sendo dependente dos grupos de dados gerados pelo algoritmo, e avaliados por índices que sejam capazes de verificar a capacidade de separabilidade dos grupos. Para validação, por meio do

critério interno, foram aplicados dois índices para análise: *Dunn* e *Silhouette* (KASSAMBARA, 2017).

O critério interno não possui regras para ser aplicado, então foram realizados testes com diversas quantidades de grupos definidos entre *k=2* e *k=10*, para identificar o melhor agrupamento pelo AGNES, aplicando as quatro medidas de conexão, pelas distâncias *euclidiana* e *manhattan* (tabela 23).

Tabela 23 – Medidas de qualidade interna.

Medida de Qualidade Interna				
Experimento	Distância	Medida de conexão	Índice	Grupos (K)
1	Euclidiana	Menor distância	Dunn e Silhouette	K = 2 a K = 10
2		Distância média		
3		Maior distância		
4		Ward		
5	Manhattan	Menor distância		
6		Distância média		
7		Maior distância		
8		Ward		

Fonte: Do autor.

O critério de *Dunn* é baseado na premissa que bons conjuntos de *cluster* são compactos e bem separados uns dos outros. Os valores possíveis estão entre zero (0) e infinito (∞), tendo o valor máximo como melhor resultado (BHADANA; SINGH, 2017, tradução nossa). Enquanto para o índice *silhouette*, resulta em um número entre [-1,1], sendo que o valor próximo de 1, indica que o agrupamento está melhor definido (KASSAMBARA, 2017).

Tabela 24 – Valores dos critérios para validação de AGNES.

Critério	**Índice**	**Valor Min**	**Valor Max**
Externo	Rand	0	1
	Jaccard	0	1
	Russel	0	1
Interno	Dunn	0	∞
	Silhouette	-1	1

Fonte: Do autor.

A tabela 24 indica os valores máximos e mínimos para os índices de critério externo e interno. O melhor resultado é o que está mais próximo do valor máximo dos critérios, sendo um (1) para os índices externo (*Rand*, *Jaccard* e

Russel) e o índice interno de *silhouette*, enquanto para *dunn* o valor máximo é infinito (∞).

6.2.2.6 Aplicação do algoritmo fuzzy c-means

A lógica *fuzzy* cria classificações intermediárias, sendo útil quando os limites entre os grupos não são bem separados. Permitindo que um objeto pertença a um conjunto de pertinência entre zero (0) e um (1) (XU; WUNSCH, 2008, tradução nossa).

O *fuzzy c-means* é iniciado, quando um número inteiro é especificado como o número de *cluster*, como não existe uma especificação, foram realizados experimentos na pesquisa com valores entre *k=2* e *k=10*, para identificar o agrupamento gerado pelo algoritmo, que satisfaz a análise dos dados.

Identificado os números para *k*, é necessário escolher a matriz de similaridade, para que os algoritmos possam rodar. Para isso, foram escolhidas quatro: *euclidiana*, *manhattan*, *seuclidean* e *correlation* (tabela 25).

Tabela 25 – Distância de similaridade empregadas no *fuzzy c-means*.

Experimento	Distância	Grupos (K)
1	Euclidiana	
2	Manhattan	
3	Seuclidean	
4	Correlattion	*k=2* a *k=10*

Fonte: Do autor.

Após a inicialização, o *fuzzy c-means* calcula repetidamente os centróides de cada grupo, até que a partição não mude, gerando o grupo.

A figura 28 mostra o vetor com a divisão dos dados pelo algoritmo *fuzzy c-means*, de modo que a matriz de similaridade é calculada e os dados foram divididos em dois grupos (*k=2*), o algoritmo foi aplicado nos dados dos 315 alunos. De modo que a primeira linha representa o aluno e a segunda o grupo pertencente, após a aplicação do algoritmo.

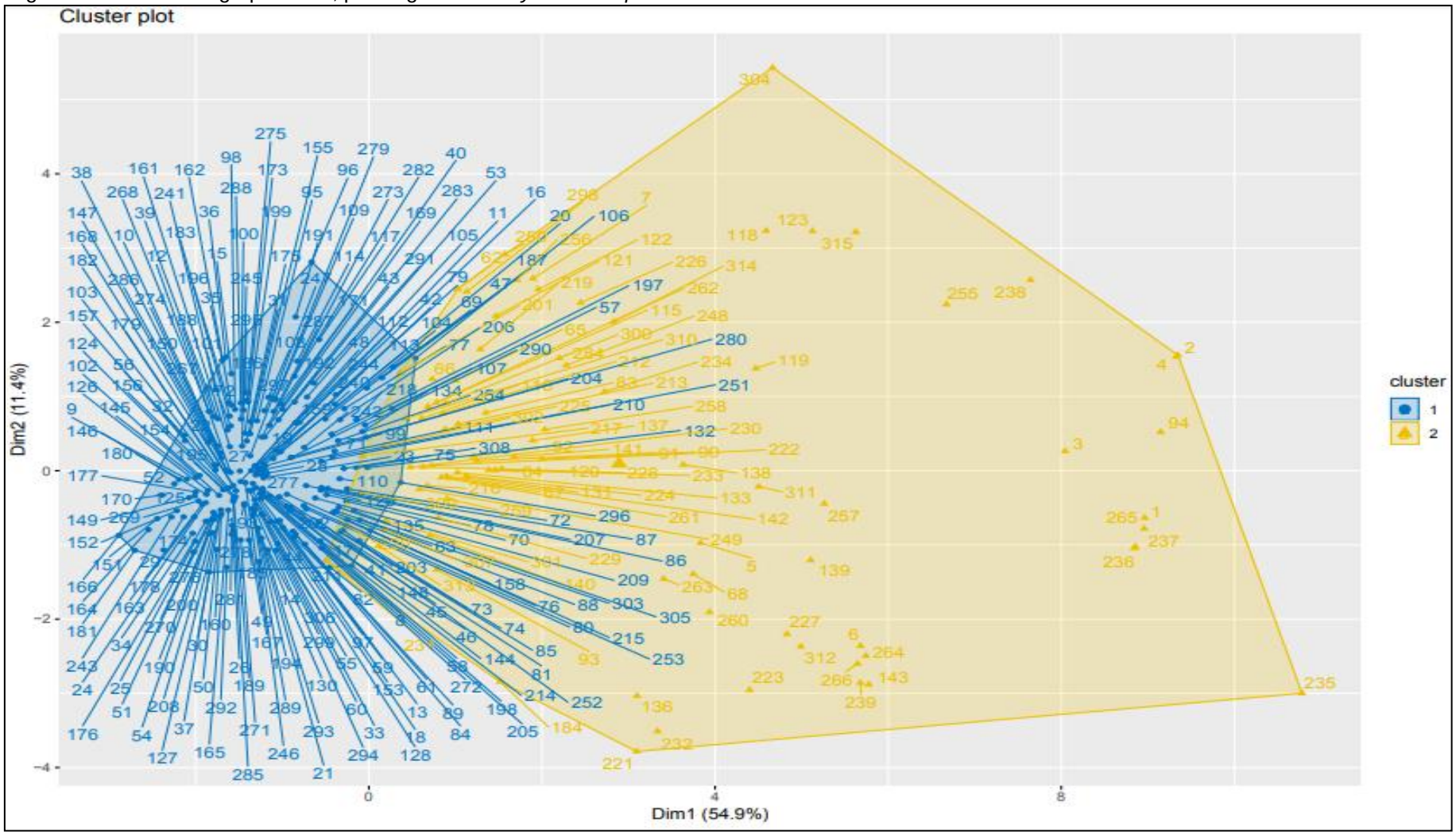
Figura 28 – Vetor de agrupamento, pelo algoritmo *fuzzy c-means.*

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
2	2	2	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1
20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76
1	1	1	1	2	1	2	2	2	2	2	1	1	1	1	1	1	1	1
77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95
1	1	1	1	1	1	2	1	1	1	1	1	1	2	2	2	2	2	1
96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133
2	2	1	2	2	2	2	2	2	1	1	1	1	1	1	1	2	1	2
134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152
1	1	2	2	2	2	2	2	2	2	1	1	1	1	1	1	1	1	1
153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228
1	1	2	2	1	1	2	1	1	2	1	2	2	2	2	1	2	2	2
229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247
1	1	1	2	2	2	2	2	2	2	2	1	1	1	1	1	1	1	1
248	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266
2	2	2	1	1	1	1	2	2	2	2	2	2	2	1	2	2	2	2
267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1
286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304
1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	2
305	306	307	308	309	310	311	312	313	314	315								
1	1	1	1	1	2	2	2	1	2	2								

Fonte: Do autor.

A figura 29 representa graficamente o agrupamento pelo *fuzzy c-means*, com a divisão de dois grupos (*k=2*), em que a cor azul representa o grupo um (1) e a cor amarela o dois (2). Os números indicam o *id* do aluno, e a imagem ilustra a divisão de grupos, representada no vetor de grupamento da figura 28.

Figura 29 – Vetor de agrupamento, pelo algoritmo *fuzzy c-means pela distância euclidiana.*



Fonte: Do autor.

Após a aplicação dos quatro métodos, *manhattan*, *euclidiana*, *seuclidean* e *correlattion*, foram realizados testes envolvendo a variação dos números de grupos (*k*), para identificar qual apresenta melhor resultado dentro da natureza dos dados.

6.2.2.7 Aplicação das medidas de qualidade pelo algoritmo *fuzzy c-means*

Os índices de validação para o *fuzzy c-means* buscam esquemas de agrupamento em que no conjunto de dados, os pontos exibam um alto grau de participação em um *cluster* (GAN; MA; WU, 2007, tradução nossa).

Para a pesquisa, foram determinados dois índices para validação do agrupamento gerado pelo *fuzzy c-means*: *partion coefficient* (PC) e *partion entropy* (PE). Ambos os índices foram aplicados para as quatro distâncias: *manhattan*, euclidiana, *seuclidean* e *correlattion*, verificando a qualidade de grupos de *k* com valores entre *k=2* e *k=10*.

Os índices PE e PC envolvem valores de associação. De modo que, para o PC, quanto mais próximo o valor de um (1), mais difuso é o agrupamento. Enquanto o índice de PE possui valores próximos ao limite superior, indicando a ausência de qualquer estrutura de *clustering* característico ao conjunto de dados (YANG; WU, 2001, tradução nossa).

O índice *fuzzy silhouette* é o mesmo que foi utilizado na validação do agrupamento do algoritmo de AGNES. De modo que o valor do índice, para ser considerado satisfatório, deve ser o mais próximo possível de zero. A tabela 26 mostra os índices de validação para agrupamento *fuzzy c-means*,

Tabela 26 – Índice de validação para agrupamento gerado por *fuzzy c-means*.

Experiência	Distância	Índice	Grupos (K)
1	Manhattan	Partion Coefficient, Partion Entropy e Fuzzy Silhouette	K = 2 a K = 10
2	Euclidiana		
3	Seuclidean		
4	Correlattion		

Fonte: Do autor.

A tabela 27 mostra os valores máximos e mínimos para os índices de validação do algoritmo *fuzzy c-means*.

Tabela 27 - Valores dos critérios para validação do *fuzzy c-means*.

Índice	Valor Min	Valor Max
Partion Coefficient	0	1
Partion Entropy	0	Log c
Silhouette	-1	1

Fonte: Do autor.

6.3 RESULTADOS OBTIDOS

A aplicabilidade das medidas de qualidade foram descritas na etapa anterior, gerando os resultados para análise e avalição das medidas, para que fosse possvivel identificar o modelo final mais adequado para a base de dados.

6.3.1 Agrupamento gerado pelo AGNES

A validação da hierarquia dos modelos gerados pelo AGENS foi realizada pela aplicação do Coeficiente de Correlação Cofenética (CPCC), para avaliar que tipo de agrupamento é mais adequado para base de dados da pesquisa. Estes valores são mostrados na tabela 28, para as distâncias de *euclidiana* e *manhattan*, sendo aplicado nos quatro métodos de conexão de agrupamento hierárquico aglomerativo.

Tabela 28 – Coeficiente de correlação cofenética do AGNES.

Experimento	Método	Manhattan	Euclidiana
1	Menor distância	0,87715040	0,91649220
2	Distância média	0,89515900	0,92535580
3	Maior distância	0,88728420	0,90759120
4	Ward	0,86048150	0,86492380

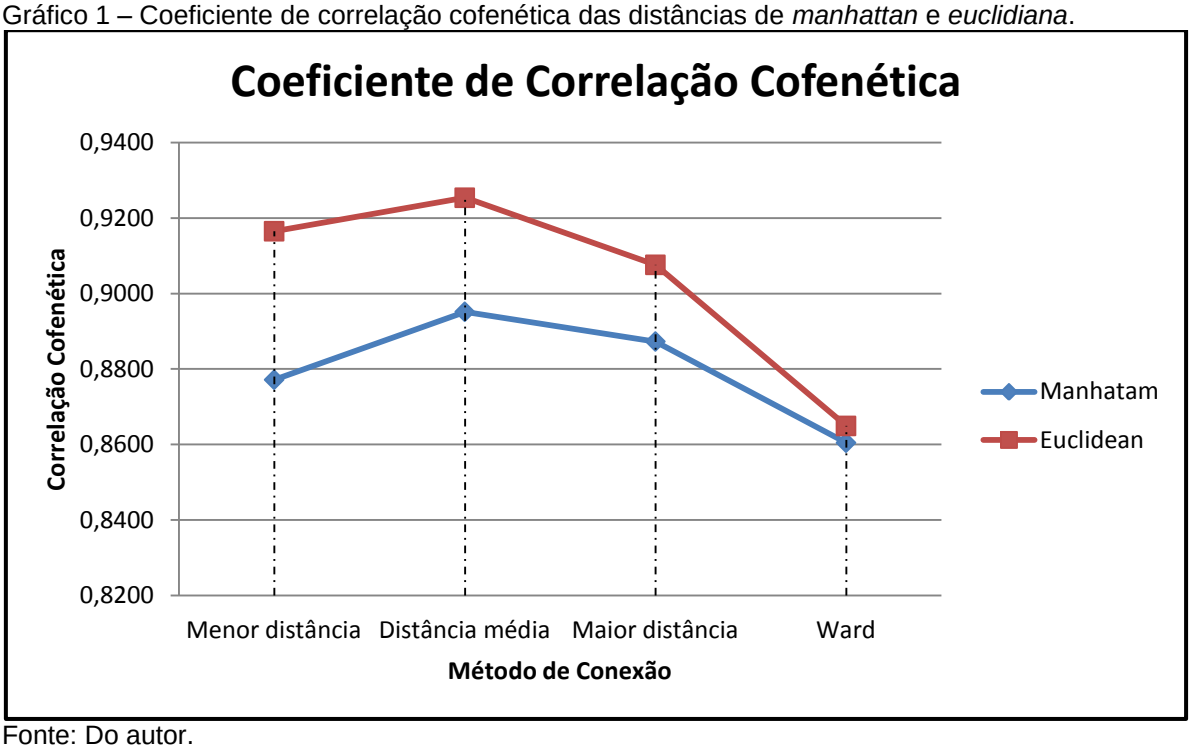
Fonte: Do autor.

O agrupamento hierárquico produzido pelo método da distância média, está em evidência para ambas as distâncias de similaridade, apresentando os melhores resultados, com valores próximos a 1 (*euclidiana* = 92535580 e *manhattan* = 0,89515900). O método *Ward,* não se adapta aos dados como os agrupamentos produzidos pela maior distância, distância média e menor distância apresentando os

menores resultados para a distância euclidiana (0,86492380) e *manhattan (*0,86048150).

Ambas as distâncias, obtiveram resultados próximos a 1, o que indica que todas as conexões se adaptam bem aos dados que estão sendo analisados. Porém, na aplicação de todos os métodos, a distância *euclidiana* obteve valores maiores que a distância *manhattan,* com média de 0,9035907 e 0,88001878, respectivamente. O gráfico 1, mostra a distância euclidiana com valores superiores, de modo que para a menor distância, distância média e maior distância os valores foram superiores a 0,9.

Gráfico 1 – Coeficiente de correlação cofenética das distâncias de *manhattan* e *euclidiana*.



Fonte: Do autor.

Após a validação da hierarquia, foram realizadas as avaliações das medidas externas comparando com a partição *k=2*, para o grupo de partição que considera os alunos como aprovados e reprovados.

A tabela 29, mostra os experimentos do algoritmo AGNES, para as distâncias *euclidiana* e *manhattan* e suas conexões, com os valores das medidas externas de *Rand, Jaccard, Russel* e *Folk.mal,* para seus respectivos experimentos

De acordo com a tabela 29, alguns índices mostraram valores superiores a 0,99, estando próximo a 1 e atingindo valores satisfatórios, com destaque para os métodos de conexão distância média (experimento 1) e de ward (experimento 2),

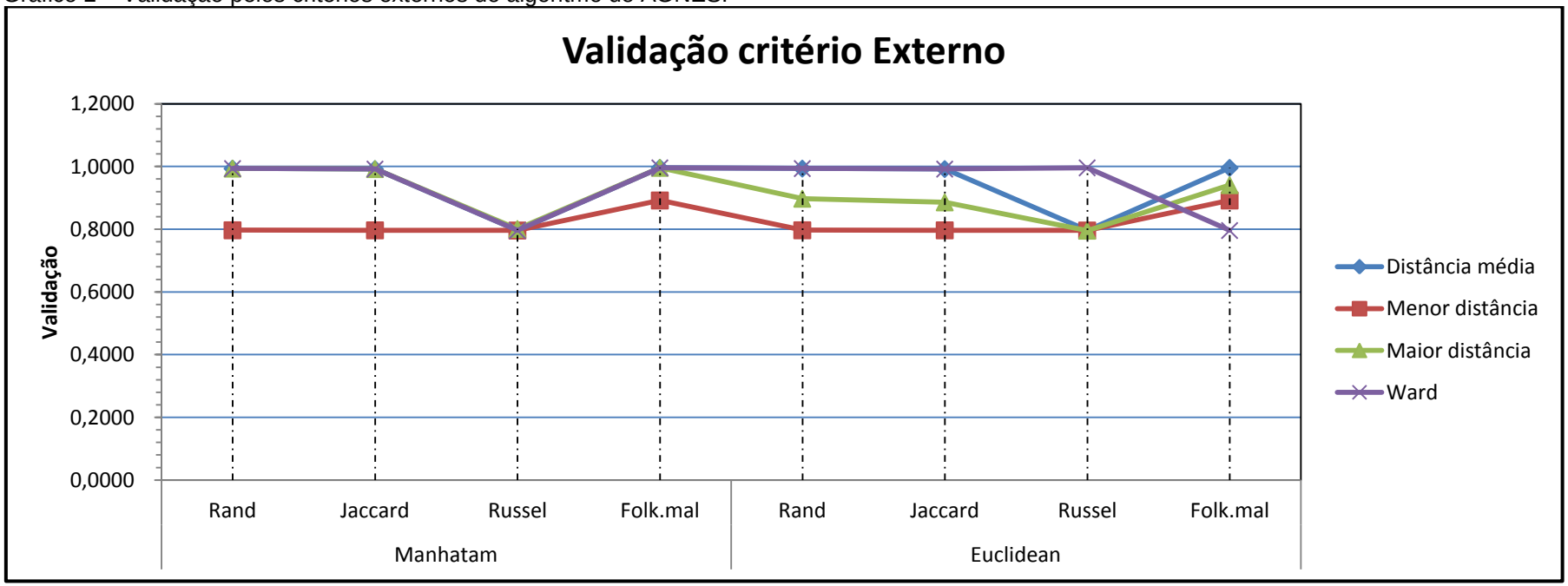
apresentando para cada, seis valores satisfatório no calculo das medidas de validação.

Tabela 29 – Validação dos resultados do algoritmo AGNES.

		Manhattan				Euclidiana			
Experimento	**Método de conexão**	**Rand**	**Jaccard**	**Russel**	**Folk.mal**	**Rand**	**Jaccard**	**Russel**	**Folk.mal**
1	**Distância média**	0,9936508	0,9920887	0,7961986	0,9960334	0,9936508	0,9920887	0,7961986	0,9960334
2	**Menor distância**	0,7969063	0,7967624	0,7961986	0,8919913	0,7969063	0,7967624	0,7961986	0,8919913
3	**Maior distância**	0,9936508	0,9921372	0,8011526	0,9960579	0,8975634	0,8859394	0,7956526	0,9408516
4	**Ward**	0,9936508	0,9920887	0,7961986	0,9960334	0,9936508	0,9920887	0,9960334	0,7961986

Fonte: Do autor.

Gráfico 2 – Validação pelos critérios externos do algoritmo de AGNES.



Fonte: Do autor.

O gráfico 2, mostra que para ambas as distâncias de *manhattan e euclidiana,* o método de conexão menor distância, experimento 2 da tabela 27, é inferior para todos os resultados apresentados pelas validações de critério externo.

O método de conexão menor, apresentado como experimento 3, da tabela 27, obteve valores satisfatórios para a distância de *manhattan*, porém não fio satisfatório para a distância *euclidiana*.

Após a validação da hierarquia e a análise das medidas externas, analisou-se a qualidade interna do agrupamento, aplicando as medidas internas de *dunn* e *Silhouette*, para cada distância e seus respectivos métodos de conexão.

A tabela 30 mostra os valores em que os resultados devem estar contemplados, de modo que quanto mais próximo do valor máximo, melhor é o resultado apresentado pelo índice de validação do critério interno.

Tabela 30 – Valores dos critérios para validação interna de AGNES.

Critério	Índice	Valor Min	Valor Max
Interno	Dunn	0	∞
	Silhouette	-1	1

Fonte: Do autor.

As medidas são testadas com número de grupos de *k=2* até *k=10* para as distâncias *Manhattan* e euclidiana*,* com os melhores valores estando em destaque, para cada medida de conexão e índice de validação interno, dividido em quatro experimentos.

A tabela 31, que demonstra o resultado de validação do algoritmo na base de dados pelo método de *manhattan*, o experimento 1 (método de conexão), mostra que o melhor número de grupos é de quatro (*k=4*), para ambos os métodos de validação. Enquanto que, para os experimentos 2 (menor distância) e 3 (método de *Ward*), o melhor número de grupos é dois (*k=2*)*.* O experimento 3 (maior distância) apresenta para o índice de *dunn*, o número dois (*k=2*) como resultado satisfatório, enquanto do silhouette apresenta sete (*k=7*), como bom para a divisão dos grupos.

Tabela 31 – Validação interna da distância de *manhattan do* algoritmo de AGNES.

Manhattan

Experimento	Método de conexão	Índice	Número de grupos								
			k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
1	Distância média	Silhouette	0,537425	0,658653	0,659229	0,650570	0,515940	0,513201	0,512554	0,509645	0,506912
		Dunn	0,225448	0,221425	0,264550	0,225378	0,227913	0,227913	0,156963	0,190481	0,190481
2	Menor distância	Silhouette	0,706520	0,651288	0,617094	0,619183	0,512941	0,504844	0,507132	0,457560	0,486230
		Dunn	0,221425	0,168624	0,204467	0,207677	0,156963	0,172723	0,190481	0,196173	0,181093
3	Maior distância	Silhouette	0,707240	0,511277	0,584891	0,650375	0,483839	0,461985	0,428968	0,407290	0,389606
		Dunn	0,210856	0,220072	0,225477	0,264550	0,259983	0,250554	0,277853	0,235733	0,222810
4	Ward	Silhouette	0,707240	0,434035	0,443991	0,471014	0,479953	0,489541	0,492773	0,478127	0,477843
		Dunn	0,210856	0,077656	0,077656	0,077656	0,099782	0,099782	0,158606	0,158606	0,161001

Fonte: Do autor.

Tabela 32 – Validação interna da distância de *euclidiana do* algoritmo de AGNES.

Euclidiano

Experimento	Método de conexão	Índice	Número de grupos								
			k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
1	Distância média	Silhouette	0,416296	0,353162	0,397803	0,316353	0,316353	0,316353	0,244935	0,267752	0,267752
		Dunn	0,646172	0,613429	0,611764	0,537006	0,502142	0,497468	0,463613	0,470068	0,466931
2	Menor distância	Silhouette	0,268382	0,290976	0,333868	0,368351	0,206037	0,206295	0,213622	0,225502	0,233039
		Dunn	0,622777	0,600908	0,575378	0,580732	0,439142	0,384116	0,379415	0,381907	0,387962
3	Maior distância	Silhouette	0,381972	0,423565	0,468918	0,468294	0,455944	0,397803	0,304746	0,368351	0,343136
		Dunn	0,517187	0,561249	0,606335	0,510691	0,505253	0,505409	0,477650	0,478580	0,473830
4	Ward	Silhouette	0,416296	0,142169	0,165947	0,165947	0,172222	0,172222	0,224878	0,060498	0,060498
		Dunn	0,646172	0,388491	0,391338	0,415888	0,424696	0,437466	0,438812	0,295663	0,313335

Fonte: Do autor.

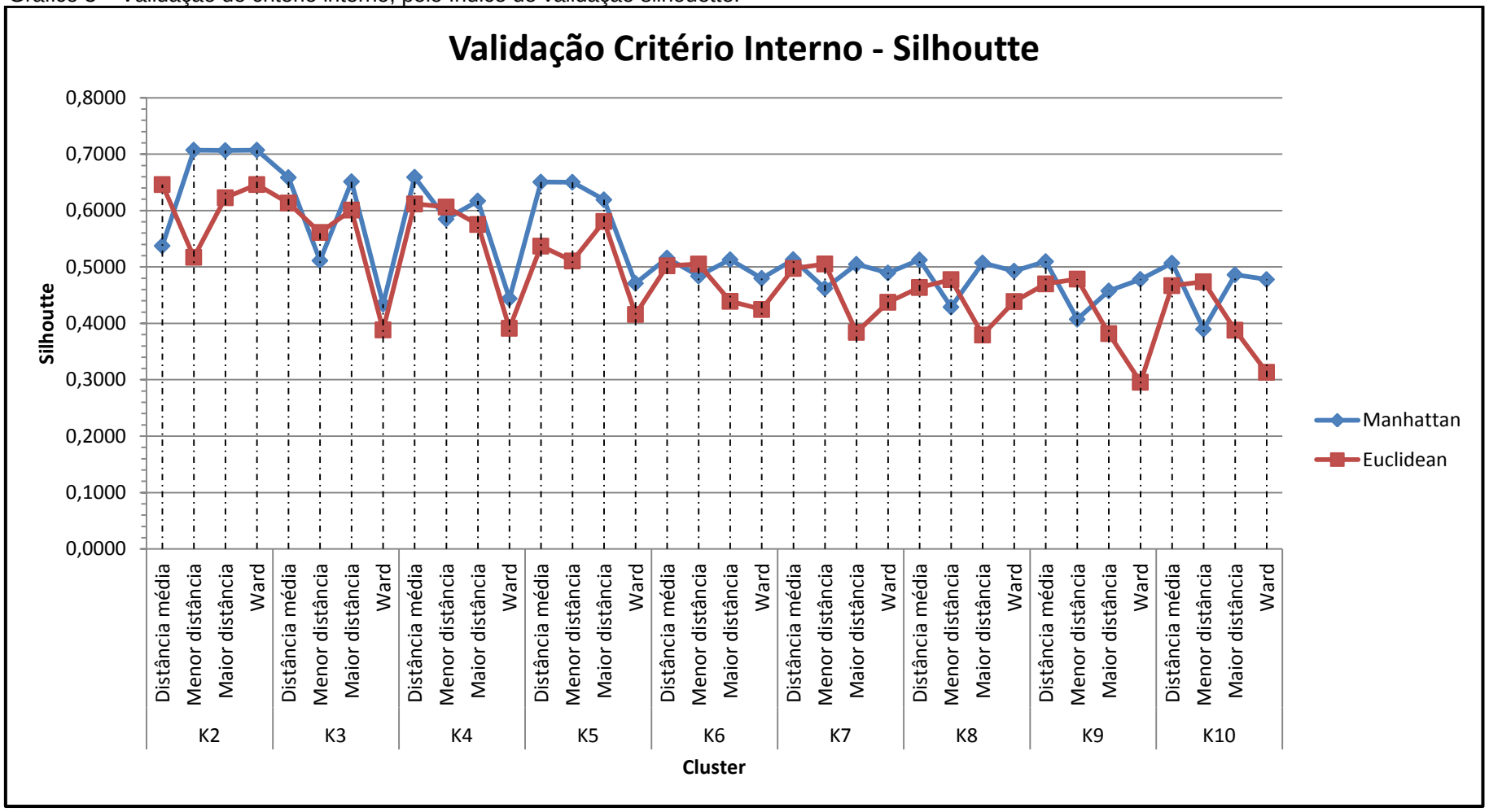
A tabela 32 demonstra o resultado de validação do algoritmo na base de dados pelo método euclidiano*. Os experimentos* 1 (distância média) e 4 (método de *Ward*)*, apresentam dois (k=2),* para divisão de grupos, enquanto o experimento 2 (menor distância), em ambos critérios de validação, o valor quatro (*k=4*) apresenta bons para formar grupos. O experimento 2 (menor distância), para o índice de validação de *dunn* apresenta dois (*k=2*) como melhor resultado, enquanto o índice de validação *silhouette* apresenta para melhor números de grupo o valor cinco (*k=5*).

De acordo com o gráfico 3, pelo índice de *Silhouette*, os valores gerados para a distância de *manhattan,* possuem valores mais altos (próximos a um) que a distância *euclidiana*, de modo que os maiores valores para número de grupos estão em dois (*k=2*), para todas as distâncias.

Os valores gerados, para validação dos grupos pelo índice de *dunn*, estão representados no gráfico 4, de modo que a distância euclidiana é superior em grande parte dos resultados, relacionado a distância de *manhattan*, porém existe variação entre os métodos de conexão e aos resultados individuais, não possuindo uma tendência para poder identificar o melhor resultado.

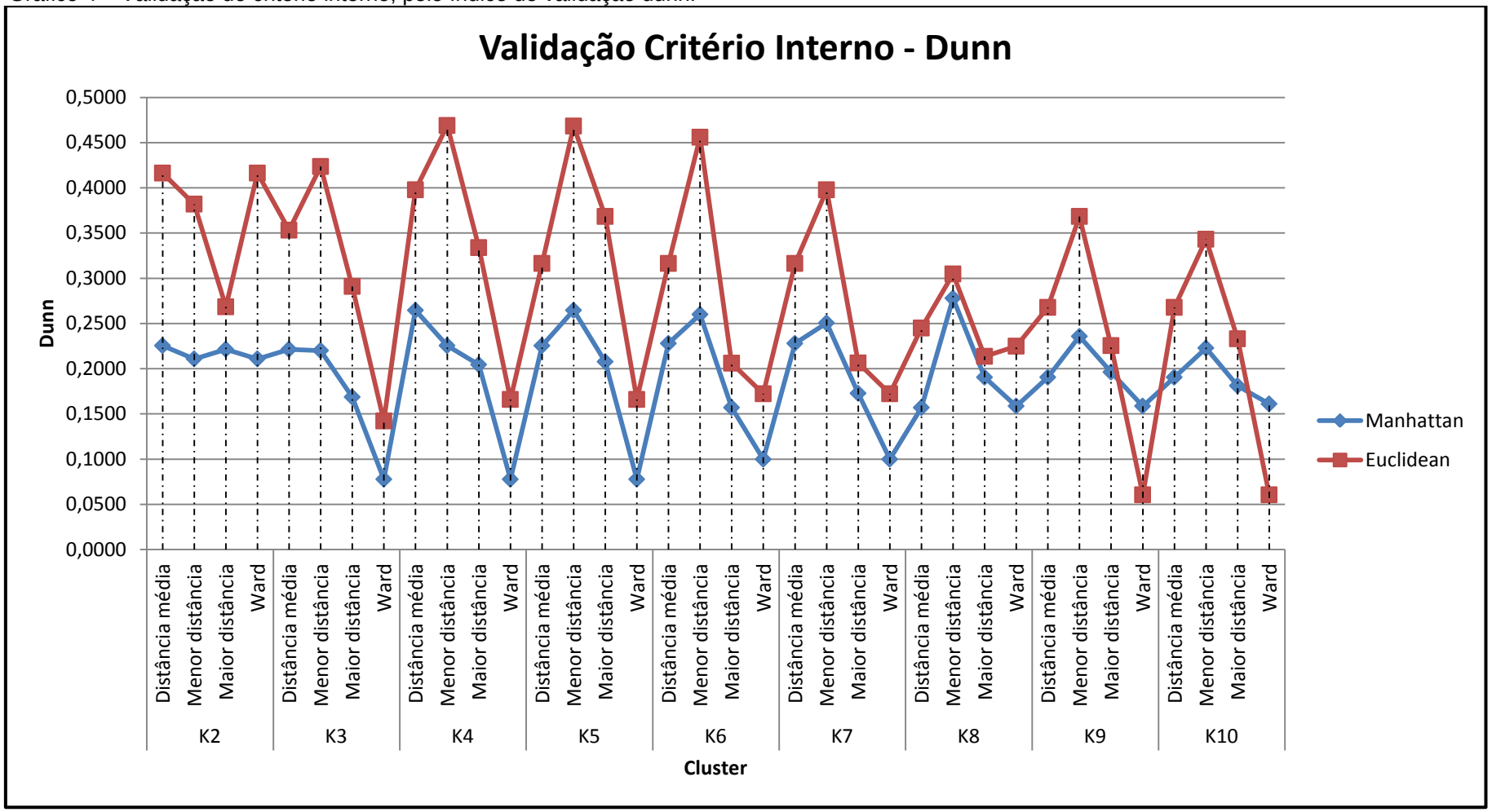
Considerando a distância média, que obteve o melhor resultado para ambas as distâncias (*manhattan e euclidiana*), o número dois (*k=2)* como melhor número para gerar grupos, teve maior destaque nas medidas de qualidade. A distância *manhattan* obteve os melhores resultados em grande parte dos índices de validação. De acordo com a análise, verficou-se que aplicando distância *manhattan,* com o método de conexão distância média, com o número de *k=2* para quantidade de grupo, se tem um bom resultado sobre a base de dados da pesquisa

Gráfico 3 – Validação do critério interno, pelo índice de validação silhouette.



Fonte: Do autor.

Gráfico 4 – Validação do critério interno, pelo índice de validação dunn.



Fonte: Do autor.

6.3.2 Agrupamento gerado pelo *fuzzy c-means*

O algoritmo *fuzzy c-means* foi aplicado com quatro distâncias de similaridade: *manhattan, euclidiana, seuclidean e correlattion*, utilizando o número de grupos entre dois e dez (*k=2* até *k=10*), para identificar o melhor modelo, para ser aplicada na base de dados.

Desta forma, foram aplicados índices de validação, que buscam agrupamento que tem alto grau de participação no grupo, sendo eles o *partion coefficient* e *partion entropy* e o índice de validação *silhouette.*

De acordo com os resultados apresentados para os 15 experimentos, o número dois (*k=2)* apresenta os melhores valores, conforme as validações dos critérios internos para geração de grupos.

Os experimentos de 1 a 3 foram gerados com a medida de distância *euclidiana*, de modo que os grupos gerados com *k=6*, *k=8* e *k=7*, não ficaram entre os valores do índice de *silhouette*, que devem compreender entre -1 e 1. Verificando o agrupamento formado, foi possível identificar que os grupos foram formados com dados ausentes.

Os experimentos de 4 a 15 apresentam valores entre os que são indicados pelos índices de qualidade. Porém, os experimentos 9, 10 e 11, pertencentes ao agrupamento pela distância Seuclidean, apresentam os melhores resultados. Conforme a regra do índice de *silhouette* e *partion coefficient*, quanto mais próximo de 1, mais qualidade tem o agrupamento. Os experimentos 9 (0,8665779) e 10 (0,8395023) apresentaram os melhores resultados.

Por meio do índice *partion entropy*, o melhor resultado é considerado o valor que estiver mais próximo de zero, desta forma o experimento 10 apresentou o valor mais adequado para o agrupamento (0,2747515).

A tabela 33 mostra os resultados das validações gerados para os métodos do algoritmo *fuzzy c-means*.

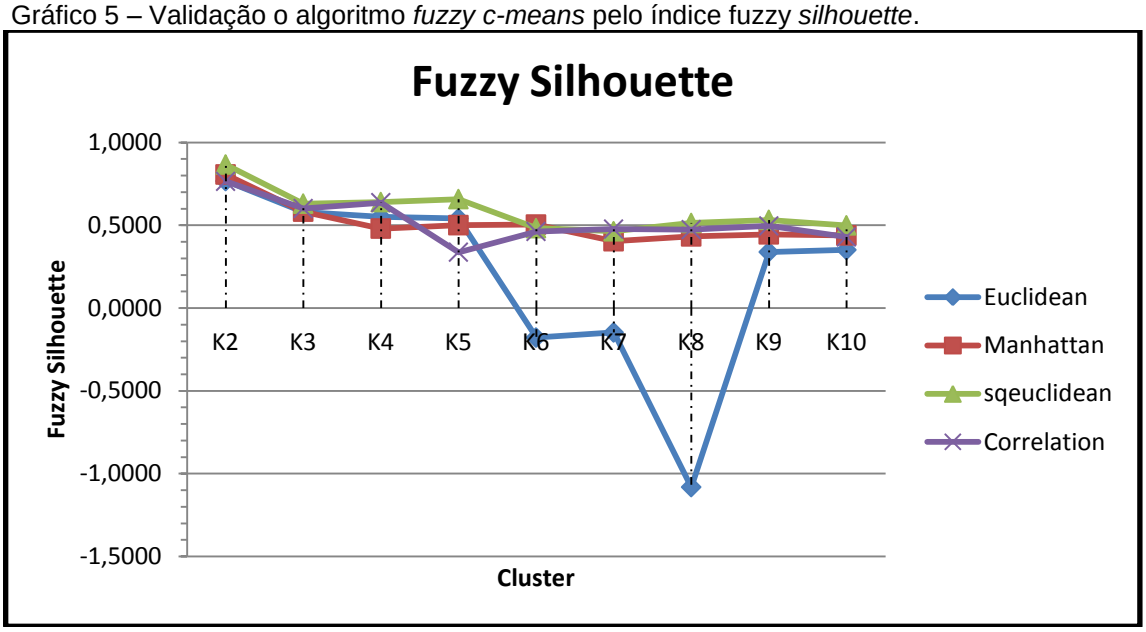
Tabela 33 - Validação do algoritmo *fuzzy c-means*.

			Número de Grupos								
Experimento	**Distância**	**Índice**	**K2**	**K3**	**K4**	**K5**	**K6**	**K7**	**K8**	**K9**	**K10**
1		Fuzzy Silhouette	0,7664356	0,5804837	0,5508999	0,5405697	-0,1785182	-0,1474001	-1,081329	0,3383781	0,3510517
2	Euclidiana	Partion Coefficient	0,5294093	0,3589939	0,2703487	0,215754	0,1795108	0,1543612	0,1351588	0,1216174	0,1093013
3		Partion Entropy	0,6632066	1,0617550	1,3482350	1,5728210	1,7544890	1,9077050	2,0413460	2,1476450	2,2531900
5		Fuzzy Silhouette	0,8063261	0,5821361	0,4799316	0,5005658	0,5036933	0,4036238	0,4327396	0,4452087	0,4387599
6	Manhattan	Partion Coefficient	0,5501451	0,3743513	0,2836158	0,2267151	0,1882312	0,1619152	0,1418846	0,126084	0,1133706
7		Partion Entropy	0,6415017	1,0346060	1,3132160	1,5363310	1,7220860	1,8742590	2,0089560	2,1286060	2,2335180
9		Fuzzy Silhouette	0,8665779	0,6305414	0,6397812	0,6579769	0,482417	0,4636427	0,5136069	0,5324904	0,4992071
10	sqeuclidean	Partion Coefficient	0,8395023	0,6663458	0,5868016	0,5815421	0,4659548	0,4054801	0,4313077	0,4003433	0,3994248
11		Partion Entropy	0,2747515	0,5795259	0,7789124	0,8283938	1,0569400	1,2344500	1,2141890	1,3441980	1,3507670
13		Fuzzy Silhouette	0,7646514	0,6010011	0,6366042	0,3371978	0,4621942	0,4768401	0,473563	0,4959004	0,4289134
14	Correlattion	Partion Coefficient	0,7132836	0,6359723	0,5833402	0,487497	0,4788609	0,4570537	0,4483417	0,4310007	0,3954093
15		Partion Entropy	0,4386597	0,6400774	0,8036058	0,9951785	1,0740280	1,1674580	1,2200060	1,2982830	1,3998010

Fonte: Do autor.

O gráfico 5, mostra a disparidade do agrupamento gerado pela distância *manhattan*, identificando como bom resultado até *k=5*, e decaindo para números maiores de grupos.

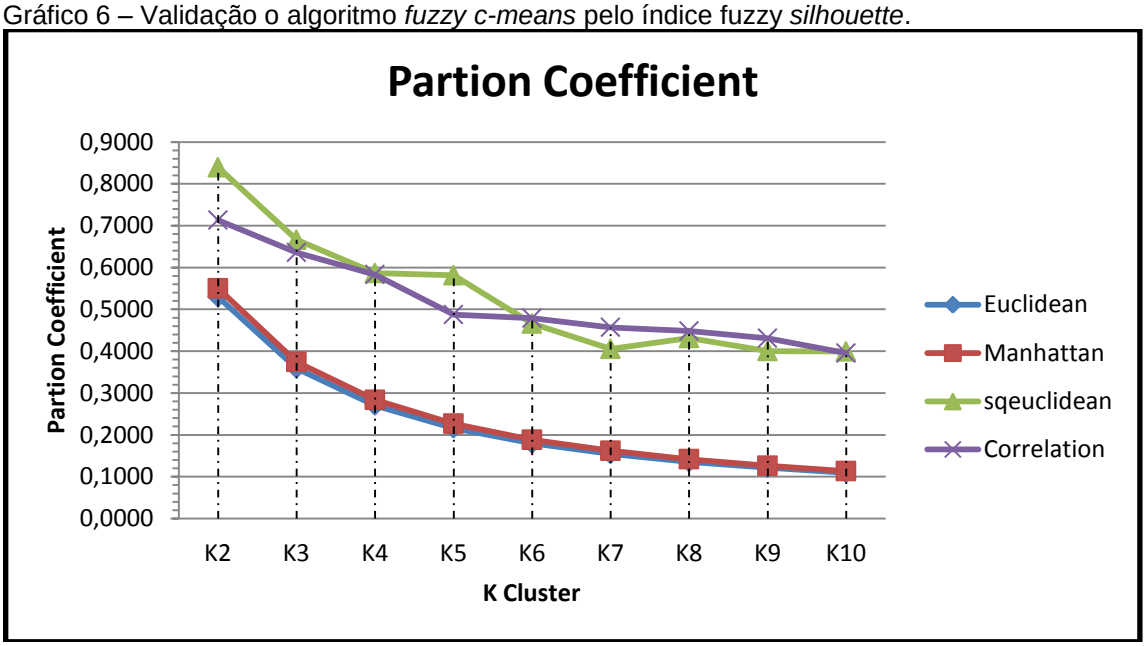
Gráfico 5 – Validação o algoritmo *fuzzy c-means* pelo índice fuzzy *silhouette*.



Fonte: Do autor.

Nos índices de *partion coefficient* (gráfico 6) e *entropy* (gráfico 7), os melhores valores, foram formados pelas distâncias *seuclidean* e *correlattion*, indicando que para a natureza dos dados, além da distância *manhattan,* a *euclidiana* também não tem boa representação.

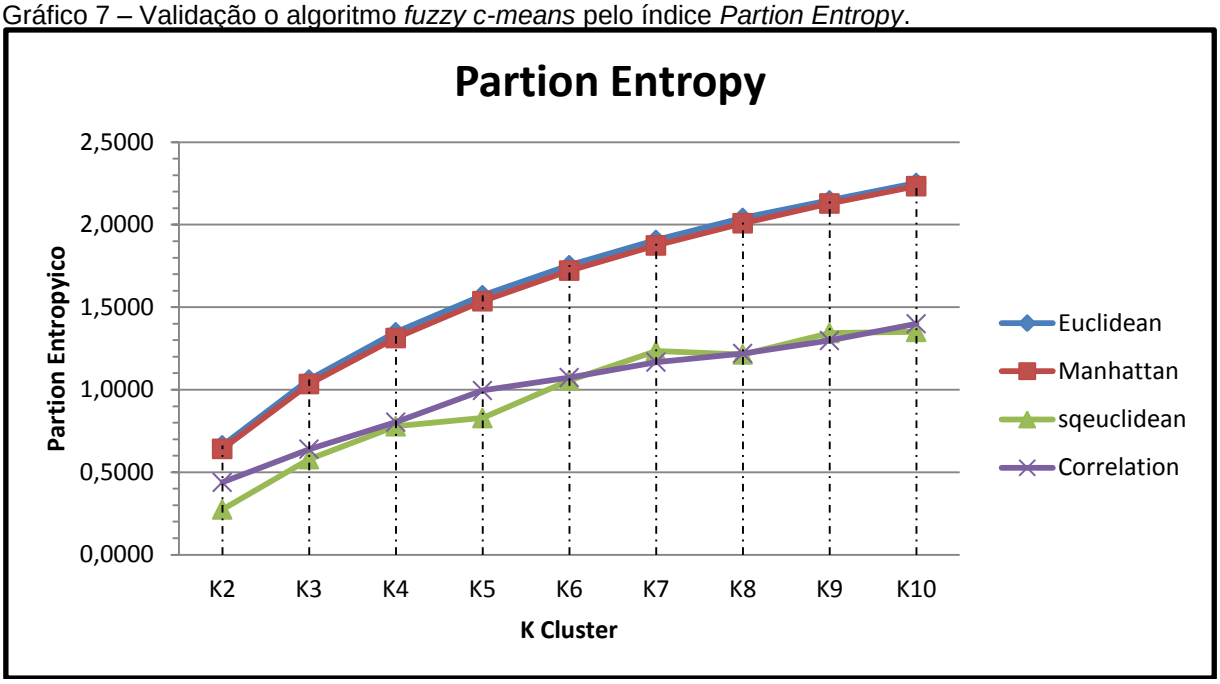
Gráfico 6 – Validação o algoritmo *fuzzy c-means* pelo índice fuzzy *silhouette*.



Fonte: Do autor.

O gráfico 7, mostra os resultados apresentados pela validação do índice *partion entropy*, indicando que o índice Entropy gera resultados que deve ser contrário a *partion coeficiente.*

Gráfico 7 – Validação o algoritmo *fuzzy c-means* pelo índice *Partion Entropy*.



Fonte: Do autor.

Com a finalização dos experimentos, foi possível identificar que para o algoritmo *fuzzy c-means,* o melhor agrupamento apresentado para os dados educacionais, são os que aplicam a distância *seuclidean*, apresentando o número dois (*k=2*), como melhor número para geração de grupos sobre os dados.

6.3.3 Identificação do Modelo Final

A pesquisa foi realizada por meio da aplicação da descoberta de conhecimento, em dados educacionais de uma disciplina ministrada a distância.

As analises consistem em identificar os melhores modelos gerados pelo método de agrupamento particional, pelo algoritmo de AGNES e pelo agrupamento particional, pelo algoritmo *fuzzy c-means*.

Para o método de agrupamento hierárquico aglomerativo, por meio do coeficiente da correlação cofenética, para validação da hierarquia, o métodos de conexão distância média, obteve os melhores resultados, pois os valores foram próximo a um (1), apresentando o valor 0,92535580 para a distância *euclidiana* e

0,89515900 para o método *manhattan*, mostrando que é o melhor modelo para ambas as distâncias.

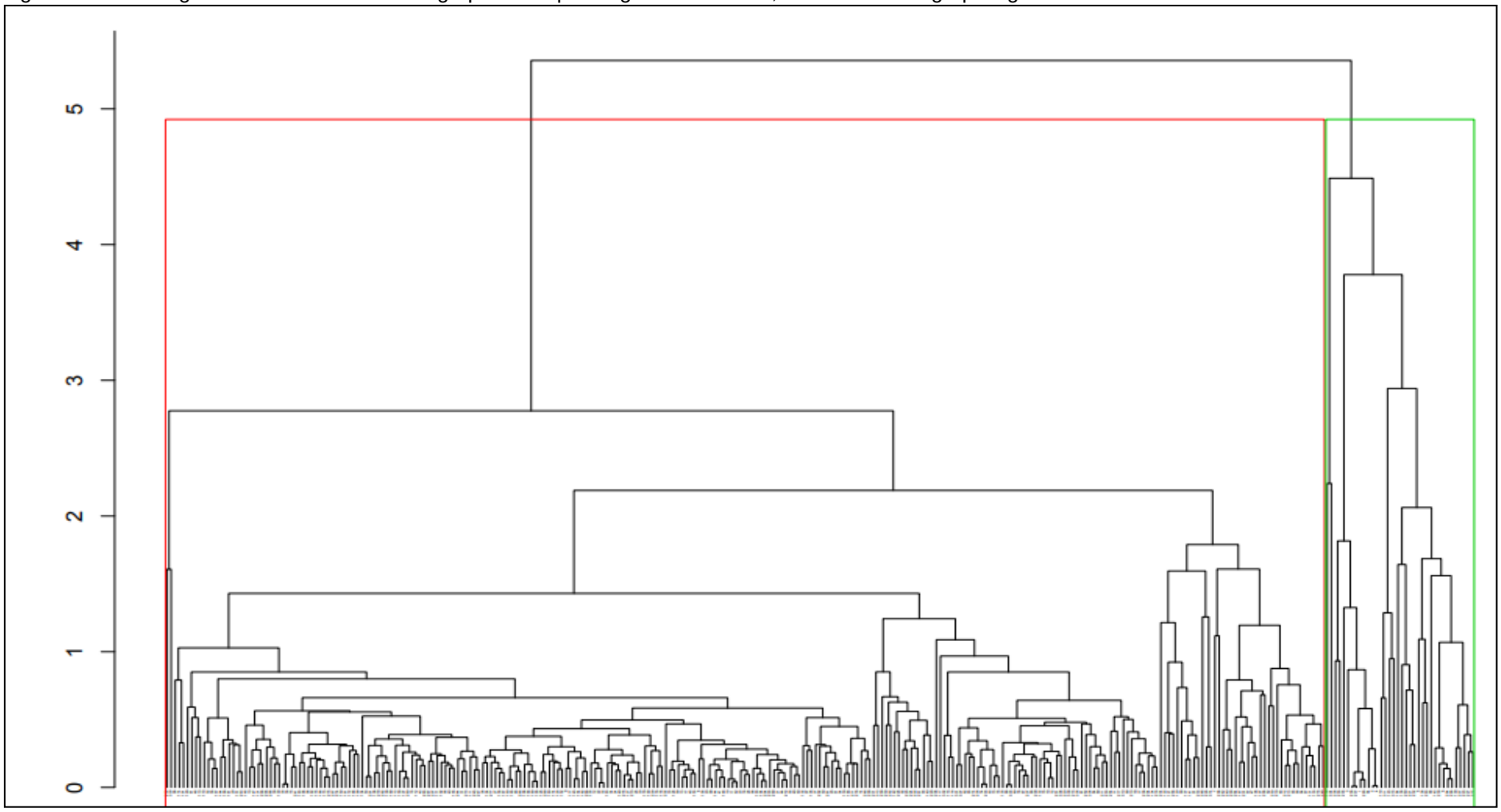
Pelo critério de validação externa, a distância média também obteve os melhores resultados, com valores bons para a distância *manhattan* nos índices *Rand* (0,9936508), *Jaccard* (0,9920887) e *Folk Mal* (0,9960334) e também nos mesmos índices, *Rand* (0,9936508), *Jaccard* (0,9920887) e *Folk Mal* (0,9960334) para a distância *euclidiana.* Em geral, pelo critério de validação externo, a distância de *manhattan* obteve nove (9) experimentos com valores bons pelos índices de *Rand*, *Jaccard* e *Folk mal,* próximos a um (1), enquanto a distância *euclidiana* obteve seis (6) resultados pelos índices de *Rand*, *Jaccard*, Russel e *Folk Mal* com valores bons, próximos a um (1).

O índice de validação interno foi decisivo para a definição de grupos, para agrupamento hierárquico aglomerativo, de modo que obteve os melhores valores para os índices de validação quando a divisão dos grupos é dois (*k=2*), independente da distância (*manhattan* e *euclidiana*).

Na validação da hierarquia à distância *euclidiana* obteve os melhores resultados, porém na aplicação das medidas de qualidade pela validação externa e interna, a distância de *manhattan*, obteve os melhores resultados, para a maior parte da aplicação de grupos.

Desta forma, o modelo identificado para agrupamento hierárquico aglomerativo, pelo algoritmo de AGNES, é formado por dois (*k=2*) grupos, usando o método da distância média, por meio da distância de *manhattan*, a figura 30 mostra o dendograma do modelo final, com os grupos destacados, para visualizar o número de grupos (*k=2*), de modo que cada nó representa um aluno.

Figura 30 – Dendograma do modelo final de agrupamento pelo algoritmo *AGNES*, com divisão de grupos igual a dois.



Fonte: Do autor.

Para o método de agrupamento particional, para os quatro métodos de validação aplicada, o número dois (*k=2*) como divisão de grupos, obteve os melhores resultados, com os números próximos aos critérios dos métodos de validação. A figura 31 demonstra a divisão dos dados em dois grupos (*k=2*), de modo que existem alunos que estão entre os dois, grupos, pois tem dados semelhantes com ambos os grupos.

A tabela 34 mostra a divisão dos grupos gerados pelos algoritmos AGNES e *fuzzy c-means*, levando em consideração os modelos finais para ambos os agrupamentos.

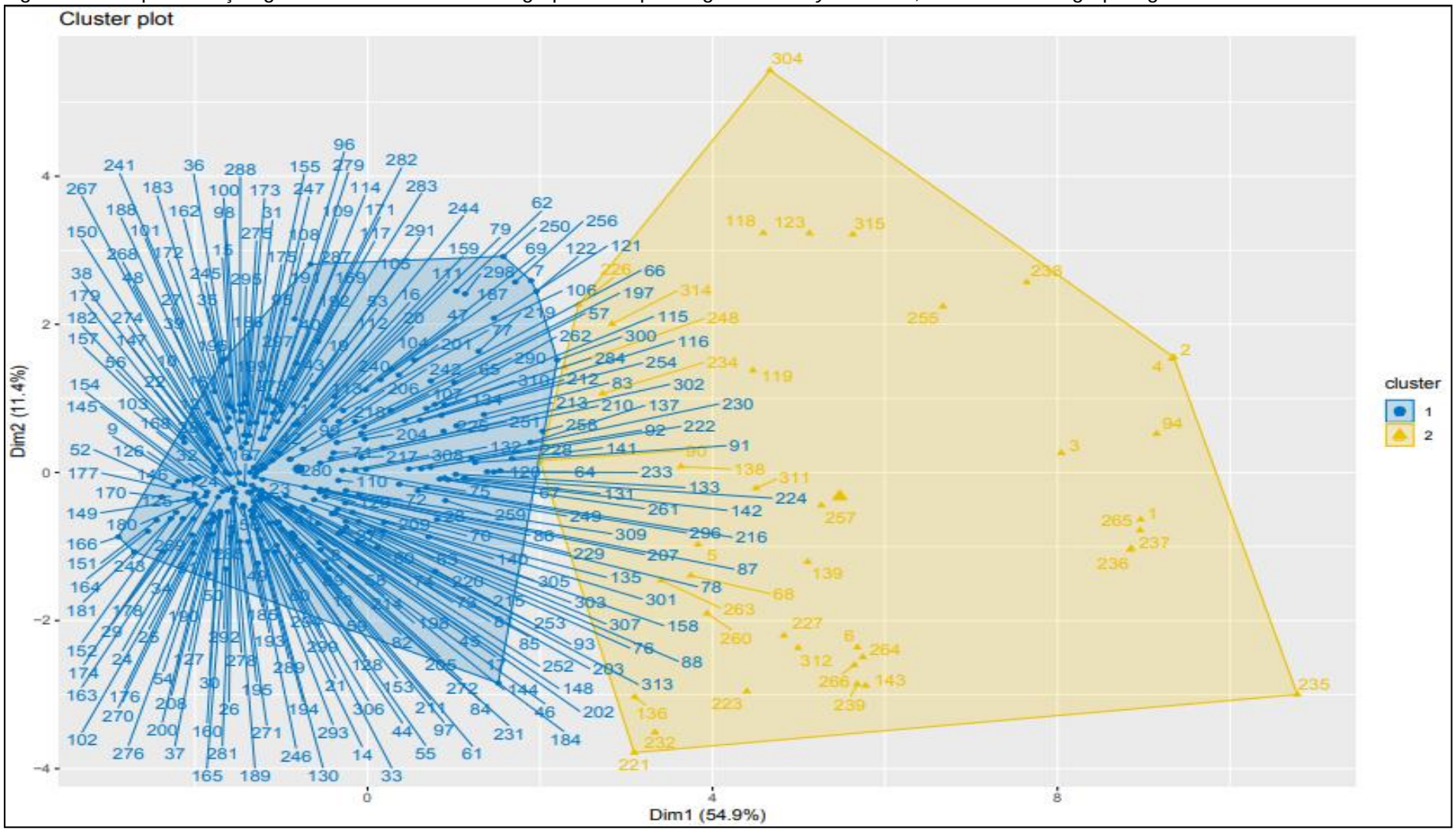
Tabela 34 - Divisão dos dados em *k = 2* grupos.

GRUPOS	QUANTIDADE DE ALUNOS POR GRUPO
AGNES - GRUPO 1	36
AGNES - GRUPO 2	279
FUZZY - GRUPO 1	40
FUZZY - GRUPO 2	275

Fonte: Do autor.

O agrupamento foi gerado para agrupamento hierárquico aglomerativo e particional, considerando dois (*k=2*) como número ideal para divisão dos grupos. Percebe-se que os grupos, ficaram com quantidade de dados próximos, comparando o grupo 1 gerados pelos algoritmos AGNES e *fuzzy c-means* e o grupo 2, também gerado por ambos.

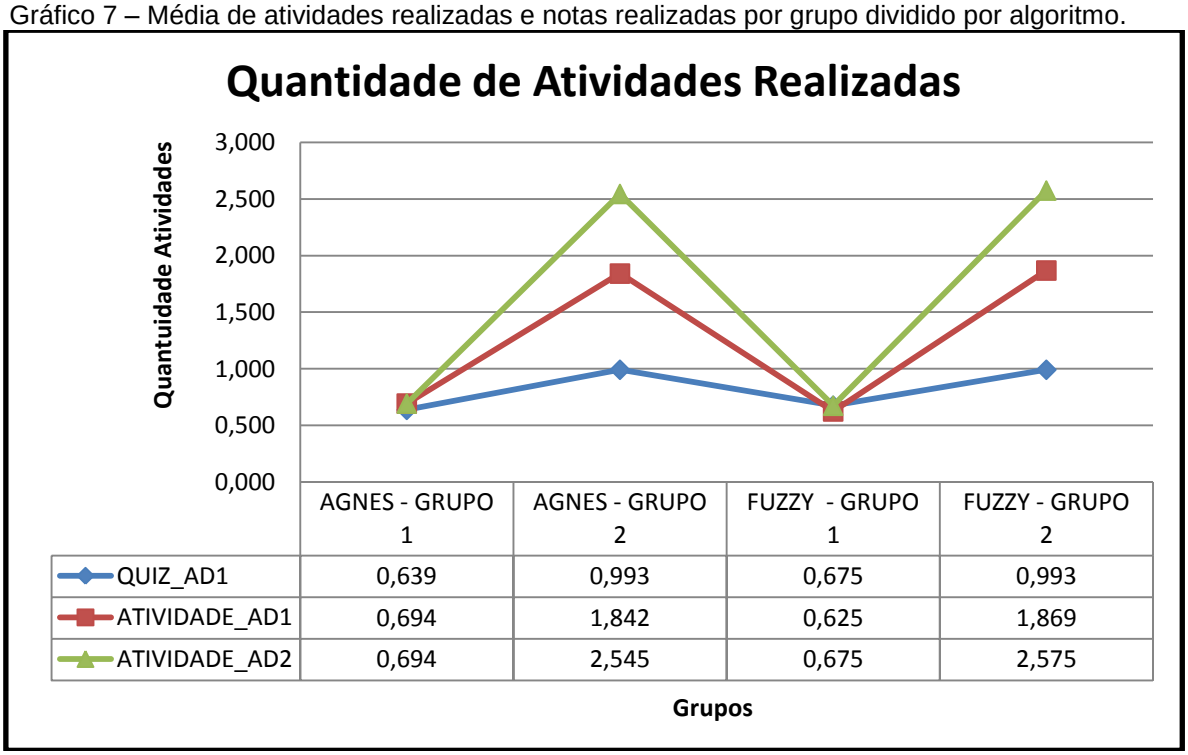
Figura 31 – Representação gráfica do modelo final de agrupamento pelo algoritmo *fuzzy c-means*, com divisão de grupos igual a dois.



Fonte: Do autor.

Ao analisar as divisões dos grupos por meio das médias por atributo, é possível identificar que ambos os agrupamentos, hierárquico e particional, realizaram a separação dos grupos de forma semelhante. Verificando o gráfico 7, referente à média de atividade realizadas pelo grupo é possível identificar que o valores de AGNES para o grupo 1 e *fuzzy* para o grupo 1, ficam abaixo de um, indicando que os alunos que ficaram em ambos os grupos não concluem as atividades.

Gráfico 7 – Média de atividades realizadas e notas realizadas por grupo dividido por algoritmo.

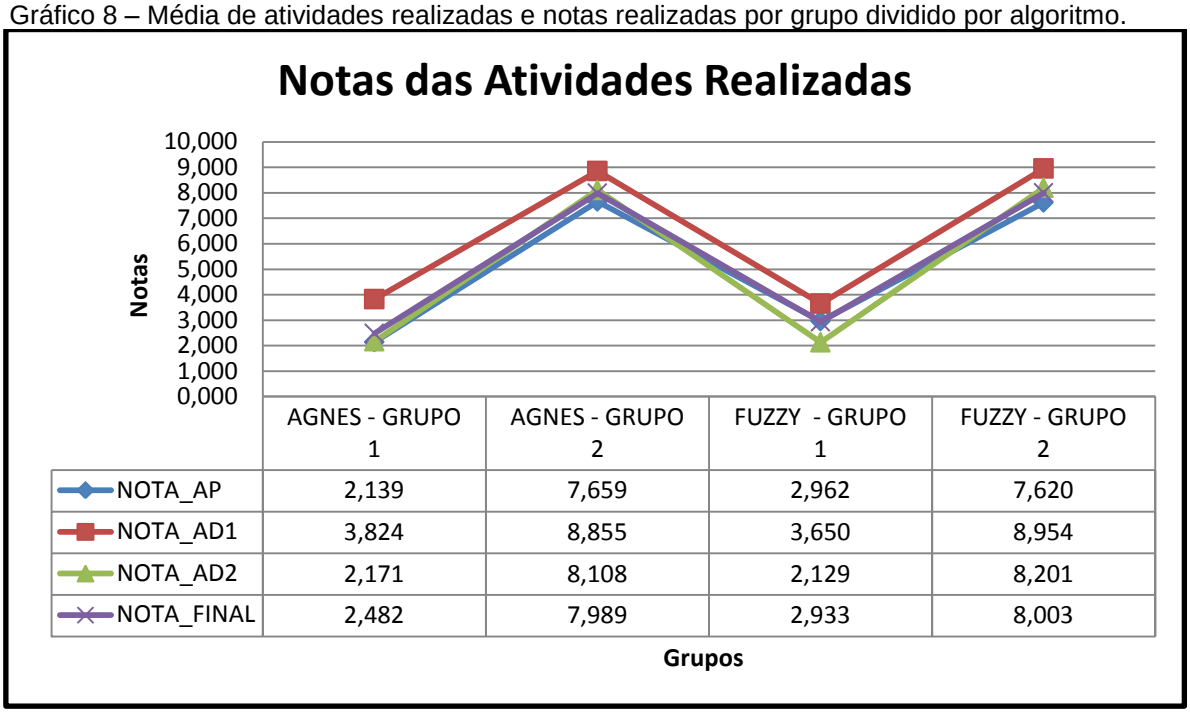


	AGNES - GRUPO 1	AGNES - GRUPO 2	FUZZY - GRUPO 1	FUZZY - GRUPO 2
QUIZ_AD1	0,639	0,993	0,675	0,993
ATIVIDADE_AD1	0,694	1,842	0,625	1,869
ATIVIDADE_AD2	0,694	2,545	0,675	2,575

Fonte: Do autor.

O gráfico 8, demonstra a média das notas por grupos, seguindo a mesma relação entre os agrupamentos gerados pelos algoritmos de AGNES e *fuzzy c-means*, de modo que os grupos um gerados pelos dois algoritmos, possuem médias parecidas e abaixo da nota seis, enquanto no grupo dois em ambos os modelos, a média é superior a sete, indicando bons resultados.

Gráfico 8 – Média de atividades realizadas e notas realizadas por grupo dividido por algoritmo.

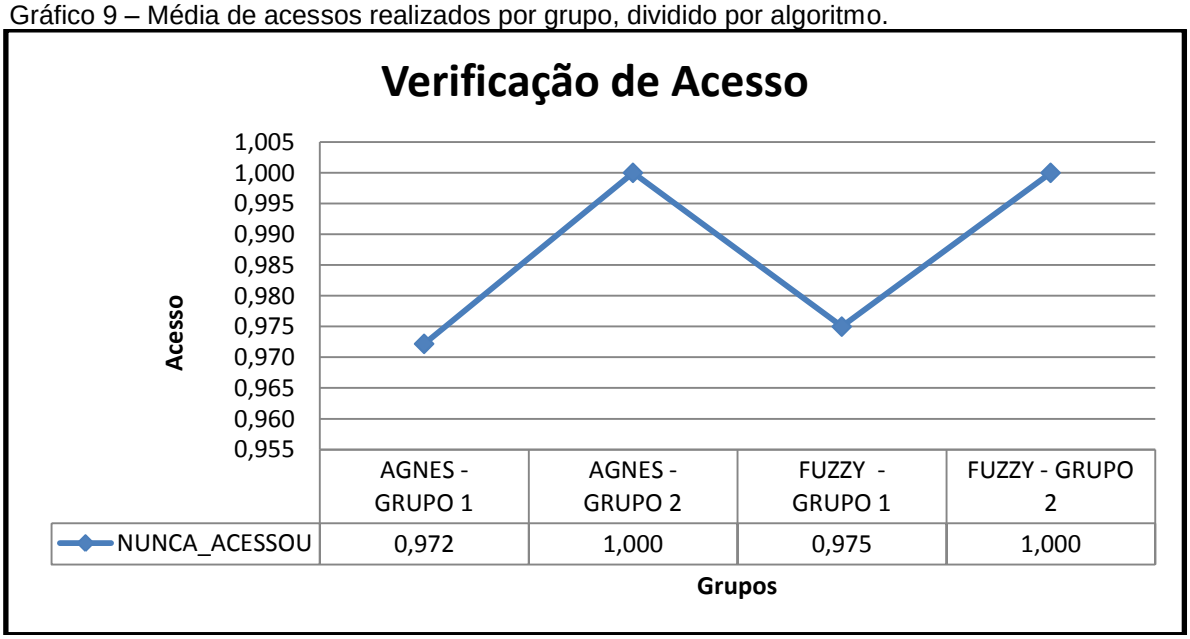


	AGNES - GRUPO 1	AGNES - GRUPO 2	FUZZY - GRUPO 1	FUZZY - GRUPO 2
NOTA_AP	2,139	7,659	2,962	7,620
NOTA_AD1	3,824	8,855	3,650	8,954
NOTA_AD2	2,171	8,108	2,129	8,201
NOTA_FINAL	2,482	7,989	2,933	8,003

Fonte: Do autor.

Considerando como zero (0), para alunos que não acessaram e um (1) alunos que acessaram o *Moodle*, o grupo dois gerado pelo AGNES e o *fuzzy c-means*, mostra a média de acesso como um (1), indicando que todos os alunos de ambos os grupos acessaram ao *Moodle*. O gráfico 9, mostra a média de acesso por grupos, separados pelos algoritmos AGNES e *fuzzy c-means*.

Gráfico 9 – Média de acessos realizados por grupo, dividido por algoritmo.

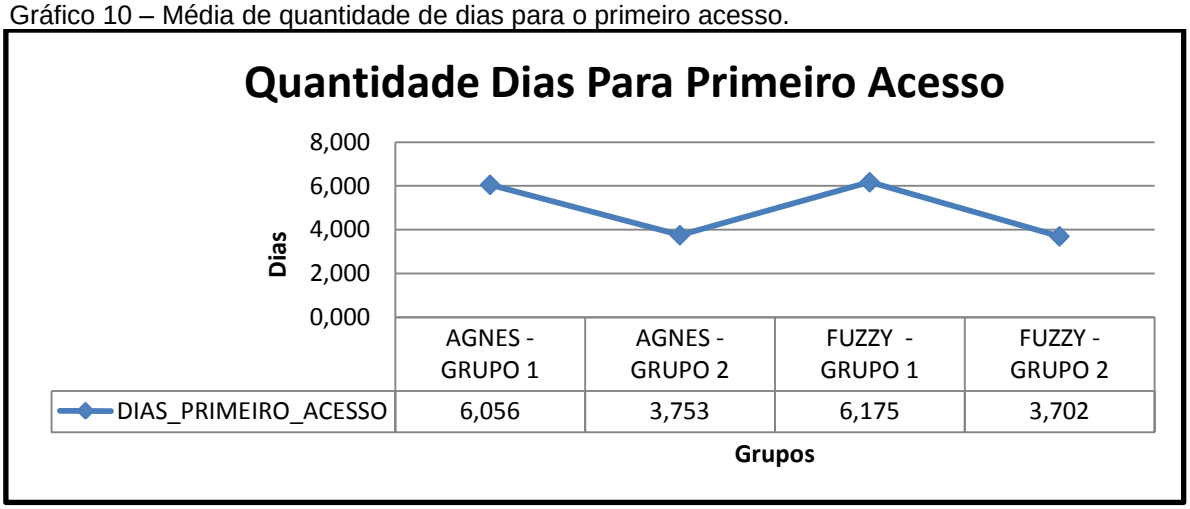


	AGNES - GRUPO 1	AGNES - GRUPO 2	FUZZY - GRUPO 1	FUZZY - GRUPO 2
NUNCA_ACESSOU	0,972	1,000	0,975	1,000

Fonte: Do autor.

Ao validar a quantidade de dias em que os alunos levaram para acessar ao *Moodle* (gráfico 10), os alunos que estão no grupo um (1), levaram seis dias para realizar o primeiro acesso, enquanto os alunos do grupo 2, acessaram ao *Moodle* no período de três dias.

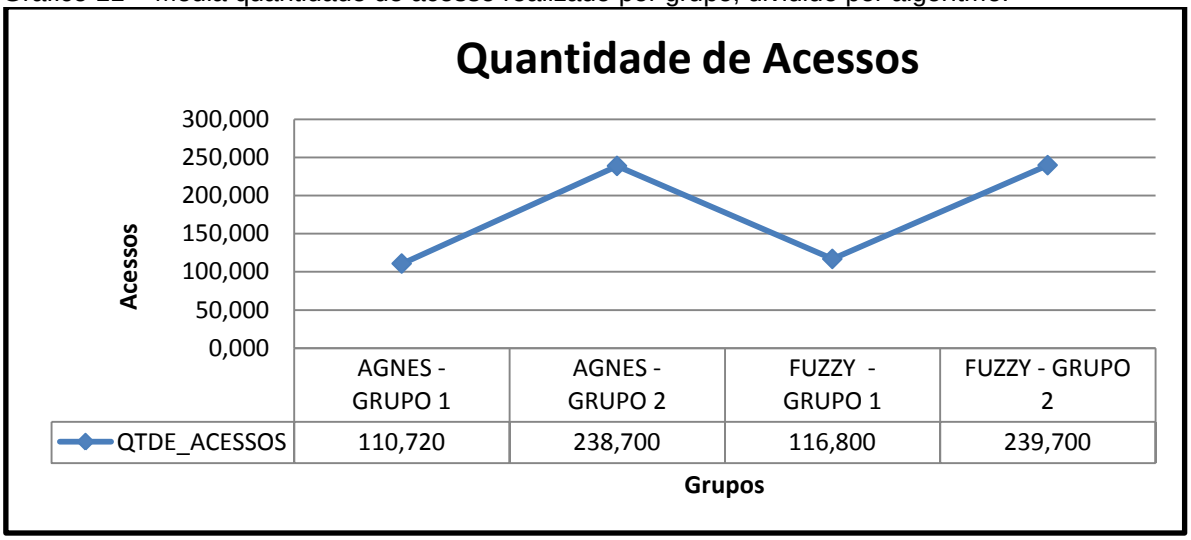
Gráfico 10 – Média de quantidade de dias para o primeiro acesso.



Fonte: Do autor.

O gráfico 11, mostra a média da quantidade de acesso por grupos, verificando novamente que os algoritmos AGNES e *fuzzy c-means* realizaram de forma semelhante à divisão dos dados em grupos, visto que a média é próxima entre o grupo 1 e grupo 2, gerado por AGNES e *fuzzy c-means*.

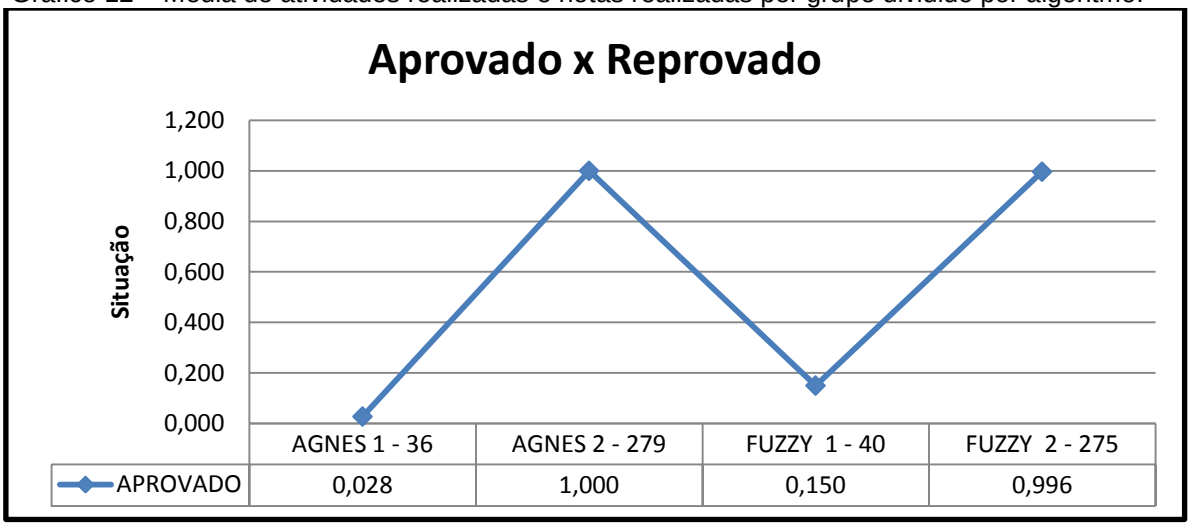
Gráfico 11 – Média quantidade de acesso realizado por grupo, dividido por algoritmo.



Fonte: Do autor.

Ao analisar a média de grupo, pelo resultado final da disciplina, ou seja, se os alunos foram aprovados ou reprovados do gráfico 12, percebe-se que o grupo dois gerado por ambos os algoritmos, possui em sua maioria, alunos aprovados enquanto o grupo dois foi formado por alunos que reprovaram na disciplina. Considerando que para zero (0), são os alunos considerados reprovados e para um (1), são os alunos considerados aprovados.

Gráfico 12 – Média de atividades realizadas e notas realizadas por grupo dividido por algoritmo.



Fonte: Do autor.

Ao analisar as médias dos grupos gerados pelo algoritmo AGNES, com a distância *euclidiana* e o método de conexão distância média e o algoritmo *fuzzy c-means*, com a distância seuclidean, que foram identificados por meio das medidas de qualidade e considerados ideais para a natureza dos dados, é possível identificar que o agrupamento hierárquico aglomerativo e o agrupamento particional, são bons para agrupamento de dados educacionais, visto que ambos apresentam resultados semelhantes no pós-processamento dos dados, gerando conhecimento útil sobre os mesmos.

O índice *silhouette* foi aplicado em comum, para validação dos resultados obtidos pelos algoritmos AGNES e *fuzzy c-means.* A tabela 35 apresenta os valores gerados pelo índice para os modelos finais do agrupamento hierárquico aglomerativo e particional.

Tabela 35 – Índice de validação silhouette.

	Algoritmo	
Grupo	**Fuzzy C-means**	**Agnes**
K=2	0,8665779	0,5374249

Fonte: Do autor.

O fuzzy c-means, apresentou valor superior que o gerado para validação do algoritmo AGNES, indicando que o método de agrupamento particional *fuzzy*, possui valores adequados em relação à natureza da base de dados empregada.

6.3.4 Discussão dos resultados

Em relação ao agrupamento com dados educacionais, notou-se que quando a quantidade de grupos aponta para dois (*k=2*), os grupos são formados considerando o atributo *situação*, sendo divididos em zero (0), para alunos reprovados e um (1), para alunos aprovados, levando em consideração as características semelhantes que possibilitam a verificação de qual tratamento seria aplicado a determinado grupo.

Quando aplicada a validação nos resultados obtidos pela distância de *manhattan* e *euclidiana*, para o algoritmo *fuzzy c-means*, percebe-se que quanto maior o número de grupos, inferiores são os valores apresentados pelos índices de validação, mostrando que essas não são matrizes de distâncias boas para aplicar o algoritmo sobre a natureza dos dados.

A validação dos resultados gerados pelo algoritmo AGNES, mostram que as medidas de distância *manhattan* e *euclidiana,* apresentam bons resultados ao aplicar na natureza dos dados, porém ao avaliar a hierarquia pelo coeficiente de correlação cofenética, a distância *euclidiana* possui melhor solução.

Ao analisar os grupos gerados pelo agrupamento hierárquico aglomerativo e agrupamento particional, foram perceptíveis as semelhanças entre as divisões de grupos.

A pesquisa apresenta por Zhang e Qin (2018), é uma pesquisa teórica, que destaca as principais etapas para a mineração de dados educacionais. Os autores indicam que para minerar dados educacionais, o processo deve iniciar pelo pré-processamento, seguindo para a mineração e a interpretação dos resultados. Conforme os autores a utilização correta dos algoritmos, melhora a eficiência da pesquisa e valor do conhecimento.

Os autores Ramos et al. (2016), utilizam o algoritmo *k-means* e o método de *ward*, para identificar o melhor modelo para analise de dados educacionais pelos métodos de agrupamento hierárquico e particional. Os autores identificaram quatro (*k=4*) como o número ideal de *clusters*, usando a matriz de similaridade *euclidiana*. Por meio de uma matriz para verificar a semelhança entre os grupos gerados pelos algoritmos, os resultados apresentados demonstraram que os métodos hierárquico e particional são bons para a análise de dados educacionais, o que se assemelha ao ocorrido nesta pesquisa, mesmo usando algoritmos diferentes e dois (*k=2*) como

número ideal de grupos, verificando a semelhanças entre os grupos, formados por ambas as abordagens, mostram eficiência para aplicação em dados educacionais. Diferentemente desta pesquisa, os autores Ramos et al. (2012), não usaram as medidas de qualidade para identificar pelos índices, o melhor método de agrupamento. Nesta pesquisa, pelo índice de *silhouette*, foi possível identificar o método de agrupamento hierárquico, como modelo final.

Na pesquisa de Silva e Santos (2018), buscaram por meio de métodos estatísticos e analise de *cluster*, identificar padrões existentes nos dados. Assim como nesta pesquisa, os autores usaram o teste de *Kolmogorov-Smirnov*, para a verificação da normalidade das amostras o que atestou que não era necessária a aplicação da mesma, enquanto que nesta pesquisa, os dados foram normalizados. A eficiência da divisão de *cluster* foi testada, aplicando de 2 até 10 grupos*.* Os autores utilizaram o algoritmo *fuzzy c-means* para minerar os dados e o índice PBM, para validação. Este índice não foi aplicado nesta pesquisa, porém a divisão dois (*k=2)* foi considerada como melhor resultado assim como o resultado apresentado pelos índices desta pesquisa.

A pesquisa apresentada por Esmeraldino (2017), não utiliza dados educacionais, porém verifica quais métodos de agrupamento hierárquico possui consistência sobre a base de dados. Assim como nesta pesquisa, o autor usa o algoritmo AGNES para mineração de dados, comprando as distâncias de similaridade de *Manhattan* e euclidiana*,* com os métodos de conexão *ward*, distância média, menor distância e maior distancia, utilizando índices de critérios interno e externo para validar os resultados, e o coeficiente de correlação cofenética para validar a hierarquia. A distância euclidiana, assim como nesta pesquisa teve melhor resultado na distância cofenética, porém os índices de validade desta pesquisa indicam a distância *manhattan* com o método de conexão distância média como modelo final, e Esmeraldino (2017), aponta como modelo final o agrupamento com a distância *euclidiana* e método de conexão distância média. Esta diferença pode ocorrer, pois as bases de dados da pesquisa são de áreas distintas, indicando para a pesquisa de Esmeraldino (2017) a divisão de grupos como oito (*k=8*), enquanto para está pesquisa foi identificado o número dois (*k=2*).

Finalmente, após as buscas de pesquisas relacionadas, foi possível analisar a abrangência do cenário que envolve a análise de agrupamento.

Facilitando a elaboração da conclusão deste trabalho, visto que as pesquisas relacionadas deram melhor concepção sobre o assunto.

7 CONCLUSÃO

A mineração de dados é um conceito que ganha força com o aumento do volume de dados, principalmente na área educacional, com o surgimento da educação a distância, em que a interação do aluno é realizada por meio de ambientes virtuais de aprendizagem.

Desta forma, baseado na mineração de dados, surge o conceito do *educational data mining* (EDM), que aplica as técnicas de mineração de dados em bases educacionais.

Com a possibilidade de formar grupos, em que os dados são divididos conforme a similaridade entre si, o método de agrupamento é uma das principais tarefas da mineração de dados e *educational data mining*.

O método de agrupamento hierárquico aglomerativo, divide os dados em uma dendograma, separando em grupos a facilitando a visualização da hierarquia de dados.

O método de agrupamento particional *fuzzy*, divide os dados em grupos, possibilitando que um dado possa pertencer a mais de um grupo.

Esta pesquisa baseou se na aplicação da descoberta de conhecimento, para realizar a comparação entre os dois métodos de agrupamento, para que por meio de medidas de qualidade, fosse possível identificar qual método de agrupamento e algoritmo tem melhor qualidade para o conjunto de dados educacional, identificando qual o melhor modelo de agrupamento.

As dificuldades encontradas foram em relação à definição de grupos, visto que os resultados apresentados pela validação dos resultados dos algoritmos podem ser próximos, dificultando identificar o melhor número de grupos para dividir a natureza dos dados, que ficou melhor para visualização quando plotados em um gráfico. A escolha da ferramenta para aplicar as técnicas de mineração de dados, também foi uma decisão difícil, visto que existem inúmeras e foi necessário encontrar uma que atendesse todas as etapas presentes na pesquisa.

Para obter os atributos para aplicar os dados, fez necessário o estudo da plataforma *moodle*, visto que foi necessário o conhecimento da mesma e as tabelas do banco para aplicar as *SQL* e montar a tabela final para mineração dos dados.

Apesar das dificuldades, os resultados obtidos pela pesquisa, foram considerados satisfatórios para que fosse cumprido o objetivo desta pesquisa,

possibilitando a identificação dos modelos finais, definidos por meio das medidas de qualidade e análise.

Ambos os métodos de agrupamento aglomerativo hierárquico, pelo algoritmo AGNES e agrupamento particional, pelo algoritmo *fuzzy c-means,* foram considerados bons para análise de dados educacionais, baseado na base de dados da disciplina de engenharia de segurança do trabalho, visto que os resultados pós-processamento foram semelhantes para os dois modelos de agrupamento.

O algoritmo AGNES, por meio das medidas de qualidade, obteve melhores resultados com a medida de similaridade *euclidiana*, utilizando o método de conexão, distância média. O algoritmo *fuzzy c-means*, obteve melhores resultados utilizando para cálculo da matriz de similaridade a distância de *seuclidean*. Ao comparar o índice de *silhouette*, que foi aplicado para validação dos dois algoritmos, o agrupamento particional, gerado por meio do *fuzzy c-means* obteve melhor resultado.

Com o conhecimento adquiro para esta pesquisa, decorrente dos estudos realizados para a aplicação da pesquisa, foram deixadas sugestões para trabalhos futuros:

a) Aplicar mais de um algoritmo no mesmo método de agrupamento, para identificar qual obtém melhor resultado. Exemplo: PAM e K-means para agrupamento particional;
b) Realizar o cálculo de outras matrizes de distância. Exemplo distância de *Maximum* e *Minkowskii;*
c) Aplicar os algoritmos AGNES e *fuzzy c-means* em bases de dados educacionais, de outra disciplina ministrada à distância;
d) Aplicar a tarefa de agrupamento, para minerar dados de interação do aluno com professores, monitoria, coordenação. Exemplo: mensagens trocadas;
e) Analisar os motivos de desempenho dos alunos, após a mineração de dados, por atributos que envolvam. Exemplo faixa de renda, local em que mora, período da disciplina, para identificar padrões que ajudem pesquisadores educacionais a identificas problemas que possam ser tratados para melhorar o desempenho do aluno.

REFERÊNCIAS

AL-SHAMMAA, Mohammed; ABBOD, Maysam F. Automatic generation of fuzzy classification rules using granulation-based adaptive clustering. In: Systems Conference (SysCon), 2015 9th Annual IEEE International. IEEE, 2015. p. 653-659.

ALVES, Lucineia. Educação a distância: conceitos e história no Brasil e no mundo. **Revista Brasileira de Aprendizagem Aberta e a Distância**, v. 10, 2018. Disponível em: http://seer.abed.net.br/index.php/RBAAD/article/view/235. Acesso em: 03 de junho de 2018.

AROOJ, Ansif; RIAZ, Mohsin; AKRAM, Malik Naeem. **Evaluation of predictive mineração de dados algorithms in soil data classification for optimized crop recommendation**. In: Advancements in Computational Sciences (ICACS), 2018 International Conference on. IEEE, 2018. p. 1-6.

ARYAL, Amar Mani; WANG, Sujing. Discovery of patterns in spatio-temporal data using agrupamento techniques. In: **Image, Vision and Computing (ICIVC), 2017 2nd International Conference on**. IEEE, 2017. p. 990-995.

ATTO, Karim; KOTOVA, Elena E. Data Mining Agents as Means of Communicating Users in an E-Learning Environment. In: **2019 Communication Strategies in Digital Society Workshop (ComSDS)**. IEEE, 2019. p. 39-42.

BAARSCH, Jonathan; CELEBI, M. Emre. Investigation of internal validity measures for K-means clustering. In: **Proceedings of the international multiconference of engineers and computer scientists**. 2012. p. 14-16.

BAKER, Ryan Shaun Joazeiro de; ISOTANI, Seiji; CARVALHO, Adriana Maria Joazeiro Baker de. **Mineração de Dados Educacionais: Oportunidades para o Brasil**. **Revista Brasileira de Informática na Educação,** [s.l], v. 19, n. 2, p.2-13, ago. 2011. Disponível em: <http://br-ie.org/pub/index.php/rbie/article/view/1301>. Acesso em: 01 out. 2017.

BARBOZA, Esdras Jorge Santos; SILVA, Marcia Terra da. COMPARAÇÃO ENTRE OS PRINCIPAIS AVA'S QUANTO A INTERATIVIDADE. In: XIV INTERNATIONAL CONFERENCE ON ENGINEERING AND TECHNOLOGY EDUCATION, 9., 2016, Salvador. Proceedings... . Salvador: Copec, 2016. p. 96 - 100. Disponível em: <http://copec.eu/intertech2016/proc/works/21.pdf>. Acesso em: 16 jun. 2018.

BENDERSKAYA, E. N. Cluster analysis problems and bio-inspired clustering methods. In: **Soft Computing and Measurements (SCM), 2017 XX IEEE International Conference on**. IEEE, 2017. p. 162-164.

BEYENE, Wendemagengnehu Tsegaye. Reduced-Order Modeling of high-Speed Channels Using Machine Learning Techniques: Partitional and Hierarchical Clusterings. Disponivél em: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8329767. Acesso em: 22 de abril de 2018.

BHADANA, Amit; SINGH, Manoj. Fusão do Algoritmo K-Means com o Índice de Dunn para Clustering Melhorado. In: **2017 2ª Conferência Internacional sobre Sistemas Computacionais e Tecnologia da Informação para Solução Sustentável (CSITSS)** . IEEE, 2017. p. 1-5.

BIELSCHOWSKY, Carlos Eduardo. Qualidade na Educação Superior a Distância no Brasil: Na qual Estamos, para Na qual Vamos?. **EAD em FOCO**, v. 8, n. 1, 2018.

BOSCARIOLI, Clodis. **Análise de agrupamentos baseada na topologia dos dados e em mapas auto-organizáveis**. 2008. Tese de Doutorado. Universidade de São Paulo. Disponível em: http://www.teses.usp.br/teses/disponiveis/3/3142/tde-11082008-132720/en.php. Acesso em: 01 de abril de 2018.

CAO, Mengmeng; GUO, Chaoyou. Research on the Improvement of Association Rule Algorithm for Power Monitoring Data Mining. In: **Computational Intelligence and Design (ISCID), 2017 10th International Symposium on**. IEEE, 2017. p. 112-115.

CEBECI, Zeynel; KAVLAK, Alper Tuna; YILDIZ, Figen. Validation of fuzzy and possibilistic agrupamento results. In: **Artificial Intelligence and Data Processing Symposium (IDAP), 2017 International**. IEEE, 2017. p. 1-7.

CHEN, Cai; LI, Yan; LI, Tie-Song. KM-A∗ pathfinding algorithm based on hierarchical agrupamento and strengthened DB Index criteria. In: **Machine Learning and Cybernetics (ICMLC), 2011 International Conference on**. IEEE, 2011. p. 1571-1576.

DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017.

DE MESQUITA LOPES, Manuela; BRANCO, Verônica Teixeira Franco Castelo; SOARES, Jorge Barbosa. Utilização dos testes estatísticos de Kolmogorov-Smirnov e Shapiro-Wilk para verificação da normalidade para materiais de pavimentação. **Transportes**, v. 21, n. 1, p. 59-66, 2013.

DUTT, Ashish; ISMAIL, Maizatul Akmar; HERAWAN, Tutut. **A Systematic Review on Education Data Mining**. IEEE Access, 2017.

EMREDE: Revista de Educação a Distância. [s.l.]: Unirede, v. 2, n. 2, 2015. Disponível em: <file:///C:/Users/Alini1/Desktop/TCC - arquivos/ARTIGOS/EAD - BRASILEIROS/67-389-1-PB.pdf>. Acesso em: 19 jun. 2018.

ESMERALDINO, Rodrigo Burigo. **Modelo de Análise de Agrupamento para a Infertilidade Masculina em Dados Biomédicos de Uma Clinica do Extremo Sul Catarinense.** 2017. 115 f. TCC (Graduação) - Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense, Criciúma, 2017. Cap. 6.

FARIA[1], Adriano Antonio; SALVADORI, Angela. A educação a distância e seu movimento histórico no Brasil. **Revista das Faculdades Santa Cruz**, v. 8, n. 1, 2010.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, v. 39, n. 11, p. 27-34, 1996.

FIELD, Andy. **Descobrindo a estatística usando o SPSS-2**. Bookman Editora, 2009.

FUNG, Glenn. A comprehensive overview of basic clustering algorithms. 2001. Disponível em: https://pdfs.semanticscholar.org/b7cd/0a4e0a129b55b76b7a0faf8ec73a5ebd1645.pdf. Acesso em: 02 de maio de 2018.

GAN, Guojun; MA, Chaoqun; WU, Jianhong. **Data clustering**: theory, algorithms, and applications. Society for Industrial and Applied Mathematics, 2007.

GAVRILOV, Sergey et al. Agrupamento optimization based on simulated annealing algorithm for reconfigurable systems-on-chip. In: **Young Researchers in Electrical and Electronic Engineering (ElConRus), 2018 IEEE Conference of Russian**. IEEE, 2018. p. 1492-1495.

GEIGER, Bernhard C.; AMJAD, Rana Ali. Mutual information-based clustering: Hard or soft?. In: **SCC 2017; 11th International ITG Conference on Systems, Communications and Coding; Proceedings of**. VDE, 2017. p. 1-6.

GOLDSCHMIDT, Ronaldo; BEZERRA, Eduardo; PASSOS, E. Data mining: conceitos, técnicas, algoritmos, orientações e aplicações. **Rio de Janeiro-RJ: Elsevier**, p. 56-60, 2015.

GOTTARDO, Ernani; KAESTNER, Celso; NORONHA, Robinson Vida. Avaliação de desempenho de estudantes em cursos de educação a distância utilizando mineração de dados. In: **Anais do Workshop de Desafios da Computação Aplicada à Educação**. 2012. p. 30-39.

HAMZAH, Mohd Faizal Mohd; RIJAL, Omar Mohd; NOOR, Norliza Mohd. Validation of clusters for M-foot shape measurement: Malaysian women foot shape. In: **TENCON 2017-2017 IEEE Region 10 Conference**. IEEE, 2017. p. 906-911.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. **Data mining**: concepts and techniques. Elsevier, 2011.

HUANG, Wei-qing et al. an efficient cluster mining algorithm for the internal motion target path based on the enhanced AGNES. In: **Trustcom / BigDataSE / ISPA, 2015 IEEE** . IEEE, 2015. p. 1318-1323.

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (n.d.) Microdados do Censo de Educação Superior 2005-2016 e Microdados do Enade. Disponível em: http://portal.inep.gov.br/web/guest/microdados. Acesso em: 19 de junho de 2018.

JACOB, John et al. Educational Mineração de dados techniques and their applications. In: **Green Computing and Internet of Things (ICGCIoT), 2015 International Conference on**. IEEE, 2015. p. 1344-1348.

JAIN, Anil K.; MURTY, M. Narasimha; FLYNN, Patrick J. Data Clustering: a review. **ACM computing surveys (CSUR)**, v. 31, n. 3, p. 264-323, 1999.

KANTARDZIC, Mehmed. **Data mining: concepts, models, methods, and algorithms**. John Wiley & Sons, 2011.

KAPOOR, Akanksha; SINGHAL, Abhishek. A comparative study of K-Means, K-Means++ and Fuzzy C-Means clustering algorithms. In: **Computational Intelligence & Communication Technology (CICT), 2017 3rd International Conference on**. IEEE, 2017. p. 1-6.

KASSAMBARA, Alboukadel. **Practical guide to cluster analysis in R: unsupervised machine learning**. STHDA, 2017.

KAUFMAN, Leonard; ROUSSEEUW, Peter J. **Finding groups in data: an introduction to cluster analysis**. John Wiley & Sons, 2009.

KHOLOD, Ivan I. Conditions for parallel execution of functions in data mining algorithm. In: **Young Researchers in Electrical and Electronic Engineering (ElConRus), 2018 IEEE Conference of Russian**. IEEE, 2018. p. 308-312.

KIRKLAND, Oliver; DE LA IGLESIA, Beatriz. Experimental evaluation of cluster quality measures. In: **Computational Intelligence (UKCI), 2013 13th UK Workshop on**. IEEE, 2013. p. 236-243.

LARA, Juan A. et al. Data preparation for KDD through automatic reasoning based on description logic. **Information systems**, v. 44, p. 54-72, 2014.

LAWSON, Richard G.; JURS, Peter C. New index for clustering tendency and its application to chemical problems. **Journal of chemical information and computer sciences**, v. 30, n. 1, p. 36-41, 1990.

LESSA, Shara Christina Ferreira. Os reflexos da legislação de educação a distância no Brasil. **Revista Brasileira de Aprendizagem Aberta e a Distância**, v. 10, 2018.

LI, Xue; LIU, Hongfu. Greedy optimization for K-means-based consensus agrupamento. **Tsinghua Science and Technology**, v. 23, n. 2, p. 184-194, 2018.

LIMA, Thiago G. et al. KDD processes in non-relational data: The case of the MineraMongo tool. In: **Information Systems and Technologies (CISTI), 2017 12th Iberian Conference on**. IEEE, 2017. p. 1-6.

LINCY, SS Blessy Trencia; KUMAR, N. Suresh. An enhanced pre-processing model for big data processing: A quality framework. In: **Innovations in Green Energy and Healthcare Technologies (IGEHT), 2017 International Conference on**. IEEE, 2017. p. 1-7.

LIU, Duo; LUNG, Chung-Horng. **P2P traffic identification and optimization using fuzzy c-means clustering**. In: Fuzzy Systems (FUZZ), 2011 IEEE International Conference on. IEEE, 2011. p. 2245-2252.

LIU, Yanchi et al. Understanding and enhancement of internal clustering validation measures. IEEE transactions on cybernetics, v. 43, n. 3, p. 982-994, 2013.

MADNI, Hussain Ahmad; ANWAR, Zahid; SHAH, Munam Ali. Data mining techniques and applications—A decade review. In: **Automation and Computing (ICAC), 2017 23rd International Conference on**. IEEE, 2017. p. 1-7.

MAHATME, V. P.; BHOYAR, K. K. Data Mining with Fuzzy Method Towards Intelligent Questions Categorization in E-Learning. In: **Computational Intelligence and Communication Networks (CICN), 2016 8th International Conference on**. IEEE, 2016. p. 682-687.

MARADITHAYA, Sumana; HAREESHA, K. S. Secure model for clustering distributed data. In: **Advances in Computing, Communications and Informatics (ICACCI), 2017 International Conference on**. IEEE, 2017. p. 1256-1260.

MIRKIN, Boris. **Clustering for data mining: a data recovery approach**. Chapman and Hall/CRC, 2005.

MOTTAGHI, Nasrin; KEYVANPOUR, Mohammad Reza. Test suite reduction using data mining techniques: A review article. In: **Computer Science and Software Engineering Conference (CSSE), 2017 International Symposium on**. IEEE, 2017. p. 61-66.

MYSQL Workbench. Disponível em: <https://www.mysql.com/products/workbench/>. Acesso em: 22 maio 2019.

ONAN, Aytuğ. A K-medoids based clustering scheme with an application to document clustering. In: **Computer Science and Engineering (UBMK), 2017 International Conference on**. IEEE, 2017. p. 354-359.

PANDE, S. R.; SAMBARE, S. S.; THAKRE, V. M. Data clustering using data mining techniques. International Journal of Advanced Research in Computer and Communication Engineering, v. 1, n. 8, p. 494-9, 2012.

PANG-NING, Tan; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao "Data Mining" -Mineração de Dados**. Rio de Janeiro: Editora Ciência Moderna, 2009.

PATHAN, Asraful Alam et al. Educational data mining: A mining model for developing students' programming skills. In: **The 8th International Conference on Software,**

Knowledge, Information Management and Applications (SKIMA 2014). IEEE, 2014. p. 1-5.

PEREIRA, Lourivaldo dos Santos Souza Aragão; FRANÇA, George. Os ambientes virtuais de aprendizagem (AVA): um estudo do moodle no curso de pedagogia da UFT. **InterSciencePlace**, v. 1, n. 25, 2015.

PERES, Sarajane Marques et al. Tutorial sobre Fuzzy-c-Means e Fuzzy Learning Vector Quantization: Abordagens Híbridas para Tarefas de Agrupamento e Classificação. **Revista de Informática Teórica e Aplicada**, v. 19, n. 1, p. 120-163, 2012.

PIATETSKY-SHAPIRO, Gregory. **Advances in knowledge discovery and mineração de dados**. Menlo Park: AAAI press, 1996.

PRISTYANTO, Yoga; PRATAMA, Irfan; NUGRAHA, Anggit Ferdita. Data Level Approach for Imbalanced Class Handling on Educational Data Mining Multiclass Classification. Disponível em: https://bit.ly/2ZpW66P0792. Acesso em: 13 de maio de 2018.

R:SOFTWARE Development Life Cycle: A Description of R's Development, Testing, Release and Maintenance Processes. A Description of R's Development, Testing, Release and Maintenance Processes. 2018. The R Foundation for Statistical Computing c/o Institute for Statistics and Mathematics. Disponível em: <https://www.r-project.org/doc/R-SDLC.pdf>. Acesso em: 18 maio 2019.

RAMOS, Jorge Luis Cavalcanti et al**.** A Comparative Study between Clustering Methods in Educational Data Mining**. IEEE Latin America Transactions**, v. 14, n. 8, p. 3755-3761, 2016.

RAMOS, Jorge Luis Cavalcanti et al. Analisando Fatores que afetam o Desempenho de Estudantes Iniciantes em um Curso a Distância. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. 2014. p. 99.

RAMOS, Jorge Luis Cavalcanti et al. **Um Modelo Preditivo da Evasão dos Alunos na EAD a Partir dos Construtos da Teoria da Distância Transacional**. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). 2017. p. 1227.

RAMOS, Thaiza et al. **Use of educational Data Mining to identify distance learning students' profiles and patterns of participation**. In: Information Systems and Technologies (CISTI), 2017 12th Iberian Conference on. IEEE, 2017. p. 1-6.

RODRIGUES, Rodrigo Lins et al. A literatura brasileira sobre mineração de dados educacionais. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. 2014. p. 621.

ROMERO, Cristobal et al. (Ed.). **Handbook of education Data Mining**. CRC press, 2010.

ROMERO, Cristóbal et al. Data Mining algorithms to classify students. In: **Educational Data Mining 2008**. 2008.

ROMERO, Cristóbal; VENTURA, Sebastián; GARCÍA, Enrique. **Data Mining in course management systems**: Moodle case study and tutorial. Computers & Education, v. 51, n. 1, p. 368-384, 2008.

SANTOS, Rodrigo et al. Análise de Trabalhos Sobre a Aplicação de Técnicas de Mineração de Dados Educacionais na Previsão de Desempenho Acadêmico. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. 2016. p. 960.

SATYANARAYANA, Ashwin; NUCKOWSKI, Mariusz. **Data Mining using Ensemble Classifiers for Improved Prediction of Student Academic Performance**. 2016.

SHARAFF, Aakanksha et al. Document Summarization by Agglomerative nested agrupamento approach. In: **Advances in Electronics, Communication and Computer Technology (ICAECCT), 2016 IEEE International Conference on**. IEEE, 2016. p. 187-191.

SILVA, Aluisio Cardoso et al. Mineração de Dados do Sistema Acadêmico do Instituto Federal do Sudeste de Minas Gerais-Campus Juiz de Fora. Seminários de Trabalho de Conclusão de Curso do Bacharelado em Sistemas de Informação, v. 2, n. 1, 2018.

SILVA, João C. Sedraz et al. **An EDM Approach to the Analysis of Students' Engagement in Online Courses from Constructs of the Transactional Distance**. In: Advanced Learning Technologies (ICALT), 2016 IEEE 16th International Conference on. IEEE, 2016. p. 230-231.

SILVA, Leandro Augusto da; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução a Mineração de Dados:** Com Aplicação em R. Rio de Janeiro: Elsevier, 2016. 376 p.

SOARES, Rodrigo GF et al. **An evolutionary approach for the clustering data problem**. In: Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on. IEEE, 2008. p. 1945-1950.

SOUZA, Rafael et al. Um Ambiente Inteligente de Avaliação de Comportamentos de Tutores e Turmas no Ambiente Virtual de Aprendizagem Moodle. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. 2016. p. 417.

STEFANOVA, Kamelia; KABAKCHIEVA, Dorina. Educational data mining perspectives within university big data environment. In: **Engineering, Technology and Innovation (ICE/ITMC), 2017 International Conference on**. IEEE, 2017. p. 264-270.

TALEB, Ikbal; SERHANI, Mohamed Adel. Big Data Pre-Processing: Closing the Data Quality Enforcement Loop. In: **Big Data (BigData Congress), 2017 IEEE International Congress on**. IEEE, 2017. p. 498-501.

TAN, Junyan et al. A data mining approach to study risk factors of hyperuricemia. In: **Computer and Communications (ICCC), 2017 3rd IEEE International Conference on**. IEEE, 2017. p. 2699-2703.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao datamining: mineração de dados**. Ciência Moderna, 2009.

VENGADESWARAN, S.; BALASUNDARAM, S. R. Significance of hierarchical and partitioning based clustering in grouping aware data placement for data intensive applications. In: **Parallel Computing Technologies (PARCOMPTECH), 2017 National Conference on**. IEEE, 2017. p. 1-5.

VILAÇA, Márcio Luiz Corrêa. Educação a Distância e Tecnologias: conceitos, termos e um pouco de história. **Revista Magistro**, v. 2, n. 1, 2011.

WITTEN, Ian H.; FRANK, Eibe. **Data Mining:** Practical Machine Learning Tools and Techniques. 2. ed. [s. L.]: Elsevier, 2005. 558 p.

XU, Rui; WUNSCH, Don. **Clustering**. John Wiley & Sons, 2008.

XU, Rui; WUNSCH, Donald. **Survey of clustering algorithms**. IEEE Transactions on neural networks, v. 16, n. 3, 2005.

YAMASARI, Y. et al. Improving the cluster validity on student's psychomotor domain using feature selection. In: **2018 International Conference on Information and Communications Technology (ICOIACT)**. IEEE, 2018. p. 460-465.

YAMASARI, Yuni et al. **Features extraction to improve performance of clustering process on student achievement**. In: Computer Science and Engineering Conference (ICSEC), 2016 International. IEEE, 2016. p. 1-5.

YANG, Chunmei et al. Arguing the Validation of Dunn's Index in Gene Clustering. In: Biomedical Engineering and Informatics, 2009. BMEI'09. 2nd International Conference on. IEEE, 2009. p. 1-4.

YANG, Ming-Der; HSU, Chan-Hsiang; SU, Tung-Ching. Optimal cluster numbers of unsupervised classification in Minkowski spaces. In: Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International. IEEE, 2007. p. 2044-2047. só tem yang 2001

ZAKI, Mohammed J.; MEIRA JR, Wagner; MEIRA, Wagner. **Data Mining and analysis: fundamental concepts and algorithms**. Cambridge University Press, 2014.

ZERABI, Soumeya; MESHOUL, Souham. External clustering validation in big data context. In: **Cloud Computing Technologies and Applications (CloudTech), 2017 3rd International Conference of**. IEEE, 2017. p. 1-6.

ZHANG, Wei; QIN, Shiming. A brief analysis of the key technologies and applications of educational data mining on online learning platform. In: **Big Data Analysis (ICBDA), 2018 IEEE 3rd International Conference on**. IEEE, 2018. p. 83-86.

ZHENG, Xiangwei; JIA, Yuanjiang. A study on educational data clustering approach based on improved particle swarm optimizer. In: **IT in Medicine and Education (ITME), 2011 International Symposium on**. IEEE, 2011. p. 442-445.

APÊNDICE(S)

ANÁLISE DE AGRUPAMENTO PELOS MÉTODOS HIERÁRQUICO AGLOMERATIVO E PARTICIONAL *FUZZY* UTILIZADOS PARA *EDUCATIONAL DATA MINING* EM DADOS DE EDUCAÇÃO A DISTÂNCIA

Alini Marangoni Eyng[1] Merisandra Cortês de Mattos Garcia[1]

[1]Curso de Bacharelado em Ciência da Computação – Universidade do Extremo Sul Catarinense (UNESC) – Criciúma – SC – Brazil

alinimarangoni.eyng@gmail.com, mem@unesc.net

Abstract. *This article describes the methods of KDD to apply a new data base on educational data from the area of Distance Education. The data comes from a Virtual Learning Environment. Data mining is one of KDD's methods, which seeks to find patterns to identify subset of data by discovering information that is hidden in a large volume of raw data. Through the techniques and tasks, data mining transformed data into useful knowledge. For this research, data mining is applied through clustering task algorithms so that a comparison between the AGNES algorithm for agglomerative hierarchical grouping and the Fuzzy method is used for the partitional grouping. The dosage of the methods is improved through quality measurement.*

Resumo. *Este artigo descreve os métodos de KDD para aplicar a descoberta de conhecimento sobre dados de uma base de dados educacional, da área de Educação a Distância. Os dados são provenientes de um Ambiente Virtual de Aprendizagem. Data mining é um dos métodos do KDD, que busca encontrar padrões para identificar subconjunto de dados, descobrindo informações que estão ocultas em meio a uma grande quantidade de dados brutos. Mediante as técnicas e tarefas, data mining transforma dados em conhecimento útil. Para esta pesquisa, é aplicado o data mining, por meio de algoritmos da tarefa de agrupamento, de modo que é realizada a comparação entre o algoritmo de AGNES para o agrupamento hierárquico aglomerativo e o método Fuzzy são empregados para o agrupamento particional. A verificação dos métodos com melhor desempenho é realizado através das medidas de qualidade.*

1. Introdução

O uso disseminado da tecnologia e dos computadores vem resultando no acúmulo de dados. Em consequência, as ferramentas tradicionais usadas para o gerenciamento de dados, tornam-se insatisfatórias para análise (TAN; STEINBACH; KUMAR, 2009). Desta forma, surgem técnicas computacionais e ferramentas, que apoiam a extração do conhecimento útil em grandes bases de dados. Essas técnicas e ferramentas são baseadas no Knowledge Discovery in Databases (KDD) e mineração de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, tradução nossa). O KDD aborda a metodologia geral da transformação de dados brutos em informações úteis, tendo como etapas do processo a preparação e seleção dos dados, aplicação dos algoritmos de mineração de dados, produzindo a leitura correta

dos resultados (GOLDSCHMIDT; BEZERRA; PASSOS, 2015). A mineração de dados é uma etapa do processo de KDD, entretanto é constantemente empregado como sendo o processo de descoberta de conhecimento, isso porque é essencial para o reconhecimento de padrões (HAN; PEI; KAMBER, 2011, tradução nossa).

As atividades educacionais e a utilização de banco de dados estão em fase crescente, devido à evolução da internet. Surge na área educacional a Educação a Distância (EaD), na qual é possível realizar um curso sem estar presente em uma sala de aula, bastando ao aluno ter acesso a internet. Os problemas da EaD são o de evasão e desempenho dos alunos, na qual o estudo e predição são difíceis, visto que o comportamento acadêmico depende de diversos fatores como variáveis pessoais (SATYANARAYANA; NUCKOWSKI, 2016, tradução nossa). O Education Data Mining (EDM), uma área proveniente da mineração de dados, com foco no desenvolvimento de técnicas para coletar dados de uma base educacional. Na área do EDM a tarefa de agrupamento é aplicada para identificar grupos de alunos com características similares de aprendizado (RAMOS et al., 2017).

O problema de agrupamento está voltado para a descoberta de grupos que são significativos na extração dos dados. O primeiro objetivo deste problema é encontrar grupos que sejam semelhantes. O segundo objetivo consiste no fato dos dados serem semelhantes no mesmo grupo, porém diferentes dos outros grupos (SOARES et al., 2008). O agrupamento é dividido em dois métodos: agrupamento hierárquico e agrupamento de particionamento. Para empregar o agrupamento hierárquico, é necessário que seu início se organize formando uma decomposição hierárquica em um conjunto de dados sendo representado por uma árvore, que o divide em subconjuntos menores, até que cada subconjunto seja formado por um objeto (GOLDSCHMIDT; PASSOS; BEZERRA, 2015). O algoritmo Hierárquico Aglomerativo AGNES (Agglomerative Nesting), trabalha com o método de link único, na qual cada cluster pode ser representado por objetos de cluster. O objetivo do algoritmo é gerar repetidamente a fusão do cluster até que todo objeto se juntam para formar um grupo (HUANG, 2015, tradução nossa). O agrupamento de particionamento é definido como uma divisão de conjuntos de objetos de dados em subconjuntos, que não estão ligados e cada objeto de dados fica exatamente dentro do seu subconjunto. Existem diversos algoritmos de agrupamento particional, um dos mais conhecidos é o fuzzy c-means, baseados na lógica fuzzy é usado quando

os grupos não estão bem separados, o que possibilita o descobrimento de dados que podem pertencer a mais de um grupo (XU; WUNSCH, 2008, tradução nossa).

Esta pesquisa aplica a tarefa de agrupamento, por meio dos algoritmos AGNES para agrupamento hierárquico aglomerativo e fuzzy c-means para agrupamento particional, verificando qual tem melhores medidas de qualidade para o conjunto de dados da pesquisa na área de EDM, identificando qual o melhor modelo de agrupamento. A aplicação da descoberta de conhecimento ocorre por meio de uma base de dados proveniente da disciplina de Introdução a Engenharia de Segurança do Trabalho, que é ministrada na modalidade à distância e ofertada pela Universidade do Extremo Sul Catarinense (UNESC).

2. Base de Dados

Os dados empregados na pesquisa são referentes à disciplina de Introdução a Engenharia de Segurança do Trabalho, ministrada na UNESC na modalidade à distância. A disciplina iniciou nesta modalidade, no segundo semestre de 2017, como parte da grade das engenharias da UNESC, estando dividida em duas turmas, uma para os cursos das engenharias que acontecem no período matutino e outra com os alunos dos cursos noturnos, totalizando em três semestres 315 alunos. A disciplina tem 36 créditos, de modo que as atividades estão divididas em 9 semanas/aulas. Na maioria das aulas, os alunos têm uma atividade avaliativa para postar e no final da disciplina uma avaliação presencial. As atividades juntamente com a avaliação formam a nota final do aluno, que define a aprovação ou reprovação.

3. Metodologia

As etapas metodológicas empregadas no desenvolvimento foram as seguintes: seleção da base de dados da disciplina a distância de Introdução Engenharia de Segurança do Trabalho para aplicação dos algoritmos de agrupamento, pré-processamento da base de dados selecionada, levantamento bibliográfico, aplicação do método de agrupamento particional por meio do algoritmo fuzzy c-means, aplicação do método de agrupamento hierárquico aglomerativo por meio do algoritmo AGNES, análise dos modelos obtidos por meio de medidas de qualidade para agrupamento em mineração de dados.

4 Pré-Processamento

O pré-processamento aborda diversas etapas, que devem ser aplicadas para tornar os dados apropriados para a mineração, melhorando e corrigindo erros e

inconsistências para a aplicação do algoritmo (TAN; STEINBACH; KUMAR, 2009). A preparação dos dados envolveu três ferramentas: *Excel*, *Weka* e *R Studio*. O processo iniciou pela seleção de dados, que é a etapa que identifica as informações que são relevantes para o KDD e presentes na base de dados aplicada na pesquisa (GOLDSCHMIDT; BEZERRA; PASSOS, 2015). Os atributos relacionados a fórum e *posts* foram descartados, pois não eram atividades e iterações propostas na disciplina, não possuindo dados relevantes. As mensagens não foram consideradas, pois o que o *Moodle* arquiva, são as mensagens totais por usuário, não sendo possível identificar quais eram relacionadas a alunos ou professores da disciplina. A tentativa do *Quiz* foi descartada, pois na disciplina era permitida ao aluno somente uma tentativa.

Posteriormente, realizou-se a limpeza de dados que tem como objetivo de abrandar dois problemas: valores ausentes e valores ruidosos. O problema de valores ausentes ocorre quando os atributos de um conjunto de dados, não apresentam valores para alguns exemplares ou algum atributo de interesse, inviabilizando o processo de análise pelo algoritmo, por não poder lidar com a ausência de valores. (SILVA; PERES; BOSCARIOLI, 2017). Nos dados da pesquisa, foram identificados valores ruidosos no atributo *dias primeiro acesso*, sendo preenchido com a mediana do mesmo, o valor 2.

Na ferramenta R, foi aplicado o teste de *Kolmogorov-Smirnov* (KS), para verificar a necessidade de normalizar os dados. O teste KS fornece o valor de prova (*valor-p*, *p-value*), que pode ser interpretado como a medida de grau de concordância entre os dados e a hipótese. Quanto menor o valor de *p-value*, menor é a consistência entre a hipótese e os dados (MESQUITA LOPES; BRANCO; SOARES, 2013). Se o teste é significativo, com o valor < 0,5, significa que os dados precisam ser normalizados, pois diferem de uma normalização normal, em que o teste significativo deve ser maior que 0,5 (FIELD, 2009).

5 Mineração de dados

A ferramenta escolhida para a execução da mineração de dados foi o R por ter uma ampla variedade de métodos da mineração de dados, medidas de qualidade e por ser um *software* livre. Para isso foi realizado um estudo sobre a ferramenta, a utilização dos algoritmos AGNES e *fuzzy c-means,* juntamente com as medidas de qualidade para validação do agrupamento.

Antes de aplicar a mineração pelos algoritmos, é necessário realizar a validação da tendência de agrupamento, para que fosse verificado se os dados possuem grupos significativos para a aplicação da mineração de dados, que pode ser feito visualmente ou por métodos estatísticos, na pesquisa foi aplicado o método de *Hopkins*. A mineração inicia com a definição das chamadas matriz de dados e matriz de similaridade, que são testadas com a aplicação das matrizes *euclidiana*, *manhattan*, *correlattion* e *seuclidian*. Definido a matriz de similaridade, os algoritmos são executados, de modo que para o agrupamento hierárquico, deve ser definido o método de conexão, na pesquisa são testados os métodos de *ward*, distância média, menor distância e maior distância. Os resultados apresentados pelos algoritmos são validados por meio de índices de medidas de qualidade, separadas em critério interno, externo e relativo. Para definição da divisão de grupos, foram realizados teste com variação de grupos, identificando o valor ideal, comparando as medidas de qualidade. A tabela 1 mostra os métodos por etapa, para aplicação da mineração de dados.

Tabela 1 – Métodos da mineração de dados aplicados na pesquisa.

Agoritimo	**Matriz de Similaridade**	**Métodos de conexão**	**Critérios**	**Índices de validação**	**Grupos**
AGNES	Euclidiano	Ward	Externo	Rand	*K=2*
				Jaccard	
	Manhattan	Distância média		Russel	
				Folk Mal	
		Maior distância			
		Menor distância	Interno	*Dunn*	*K=2 até*
				Silhouette	*k=10*
Fuzzy C-means	Correlation	-	Interno	*Silhouette*	*k=2 até k=10*
	Seuclidean				
			Fuzzy	Coefficient	
	Manhattan				
				Entropy	
	Euclidiano				

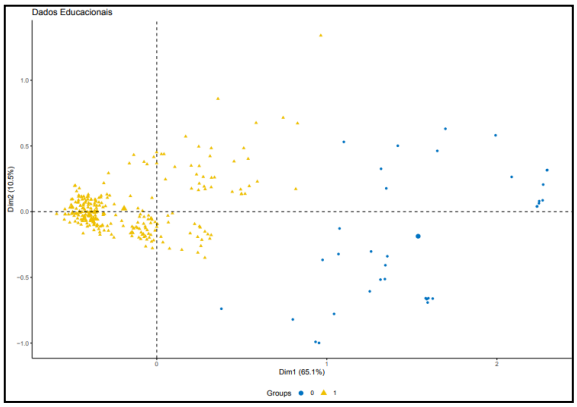
Fonte: Do autor.

6 Avaliando a tendência de agrupamento

A análise da distribuição dos dados possibilita a verificação da tendência de agrupamento visual, mostrando se existe similaridade nos dados para gerar os grupos. A figura 1 mostra a distribuição da base antes da aplicação da mineração de

dados, pelos algoritmos AGNES e *fuzzy c-means*, identificando que o conjunto de dados possui dois grupos reais, de modo que os dados são bons para serem divididos em grupos. Considerando o atributo *situação*, que indica se o aluno é aprovado (1) ou reprovado (0).

Figura 1 – Distribuição dos dados em grupos reais.



Fonte: Do autor.

Porém, a análise por tendência de agrupamento visual, pode gerar dúvida, visto que depende da interpretação do pesquisador. Desta forma, foi aplicado o método de *Hopkins*, que testa a aleatoriedade espacial, comprovando se a estrutura está uniformemente distribuída. A tabela 2 apresenta os resultados obtidos pelo método de *Hopkins* sobre a base de dados, com valores próximos à zero (0), indicando que os dados formam grupos significativos.

Tabela 2 – Resultado do método de Hopkins.

Arquivo de dados	Hopkins
Dados para aplicar o algoritmo AGNES	0,1080572
Dados para aplicar algoritmo *fuzzy c-means*	0,1135723

Fonte: Do autor.

7 RESULTADOS OBTIDOS

A aplicabilidade das medidas de qualidade foram descritas na etapa anterior, gerando os resultados para análise e avalição das medidas, para que fosse possvivel identificar o modelo final mais adequado para a base de dados.

7.1 Agrupamentos gerados pelo AGNES

A validação da hierarquia dos modelos gerados pelo AGENS foi realizada pela aplicação do Coeficiente de Correlação Cofenética (CPCC), para avaliar que tipo de agrupamento é mais adequado para base de dados da pesquisa. Indicando o método distância media, como melhor resultado para ambas as matrizes de similaridade (tabela 3).

Tabela 3 – Coeficiente de correlação cofenética do AGNES.

Experimento	Método	Manhattan	Euclidiana
1	Menor distância	0,87715040	0,91649220
2	Distância média	0,89515900	0,92535580
3	Maior distância	0,88728420	0,90759120
4	Ward	0,86048150	0,86492380

Fonte: Do autor.

Após a validação da hierarquia, foram realizadas as avaliações das medidas externas comparando com a partição *k=2*, para o grupo de partição que considera os alunos como aprovados e reprovados. A tabela 4, mostra os experimentos do algoritmo AGNES, para as distâncias *euclidiana* e *manhattan* e suas conexões, com os valores das medidas externas de *Rand, Jaccard, Russel* e *Folk.mal,* para seus respectivos experimentos, em que os melhores resultados são os que possuem valores próximos a um (1).

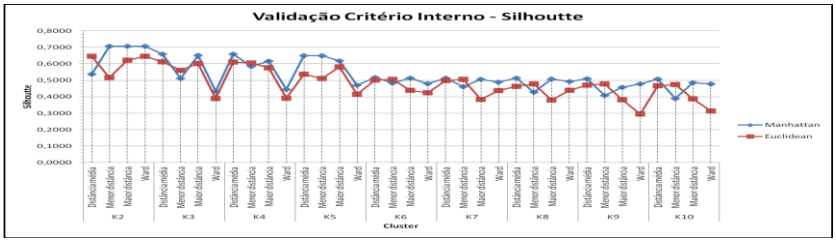
Tabela 4 – Validação dos resultados do algoritmo AGNES.

	Manhattan				Euclidiana			
Método de conexão	**Rand**	**Jaccard**	**Russel**	**Folk.mal**	**Rand**	**Jaccard**	**Russel**	**Folk.mal**
Distância média	0,9936508	0,9920887	0,7961986	0,9960334	0,9936508	0,9920887	0,7961986	0,9960334
Menor distância	0,7969063	0,7967624	0,7961986	0,8919913	0,7969063	0,7967624	0,7961986	0,8919913
Maior distância	0,9936508	0,9921372	0,8011526	0,9960579	0,8975634	0,8859394	0,7956526	0,9408516
Ward	0,9936508	0,9920887	0,7961986	0,9960334	0,9936508	0,9920887	0,9960334	0,7961986

Fonte: Do autor.

Após a validação da hierarquia e a análise das medidas externas, analisou-se a qualidade interna do agrupamento, aplicando as medidas internas de *dunn* e *Silhouette*, para cada distância e seus respectivos métodos de conexão. As medidas são testadas com número de grupos de *k=2* até *k=10* para as distâncias *Manhattan* e *euclidiana,* com os melhores valores estando em destaque, para cada medida de conexão e índice de validação interno, dividido em quatro experimentos.

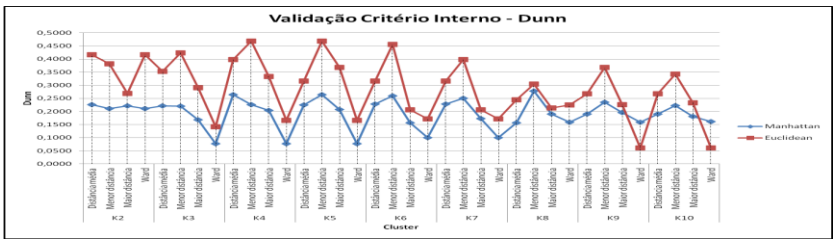
Gráfico 1 – Validação índice de silhouette.



Fonte: Do autor.

De acordo com o gráfico 1, pelo índice de *Silhouette*, os valores gerados para a distância de *manhattan,* possuem valores mais altos (próximos a um) que a distância *euclidiana*, de modo que os maiores valores para número de grupos estão em dois (*k=2*), para todas as distâncias.

Gráfico 2 – Validação índice de dunn.



Fonte: Do autor.

Os valores gerados, para validação dos grupos pelo índice de *dunn*, estão representados no gráfico 2, de modo que a distância euclidiana é superior em grande parte dos resultados, relacionado a distância de *manhattan*, porém existe variação entre os métodos de conexão e aos resultados individuais, não possuindo uma tendência para poder identificar o melhor resultado. Considerando a distância média, que obteve o melhor resultado para ambas as distâncias (*manhattan e euclidiana*), o número dois (*k=2)* como melhor número para gerar grupos, teve maior destaque nas medidas de qualidade. A distância *manhattan* obteve os melhores resultados em grande parte dos índices de validação. De acordo com a análise, verficou-se que aplicando distância *manhattan,* com o método de conexão distância média, com o número de *k=2* para quantidade de grupo, se tem um bom resultado sobre a base de dados da pesquisa.

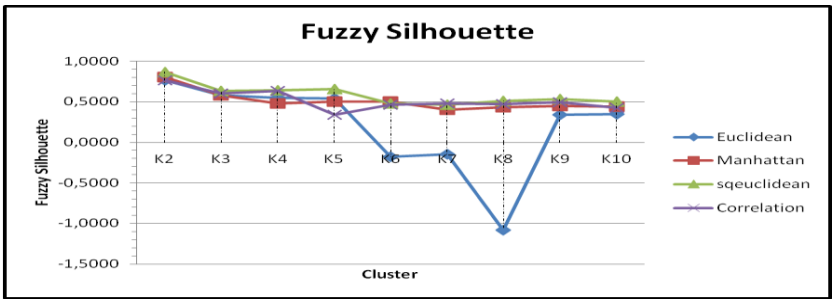
7.2 Agrupamento gerado pelo *fuzzy c-means*

O algoritmo *fuzzy c-means* foi aplicado com quatro distâncias de similaridade: *manhattan, euclidiana, seuclidean e correlattion*, utilizando o número de grupos entre dois e dez (*k=2* até *k=10*), para identificar o melhor modelo, para ser aplicada na base de dados. Desta forma, foram aplicados índices de validação, que buscam

agrupamento que tem alto grau de participação no grupo, sendo eles o *partion coefficient* e *partion entropy* e o índice de validação *silhouette.*

De acordo com os resultados apresentados para os 15 experimentos, o número dois (*k=2)* apresenta os melhores valores, conforme as validações dos critérios internos para geração de grupos. O gráfico 3, mostra a disparidade do agrupamento gerado pela distância *manhattan*, identificando como bom resultado até *k=5*, e decaindo para números maiores de grupos.

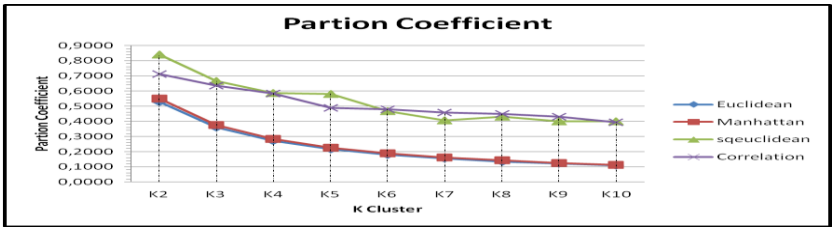
Gráfico 3 – Validação o algoritmo fuzzy c-means pelo índice fuzzy silhouette.



Fonte: Do autor.

Nos índices de *partion coefficient* (gráfico 4) e *entropy* (gráfico 5), os melhores valores, foram formados pelas distâncias *seuclidean* e *correlattion*, indicando que para a natureza dos dados, além da distância *manhattan,* a *euclidiana* também não tem boa representação.

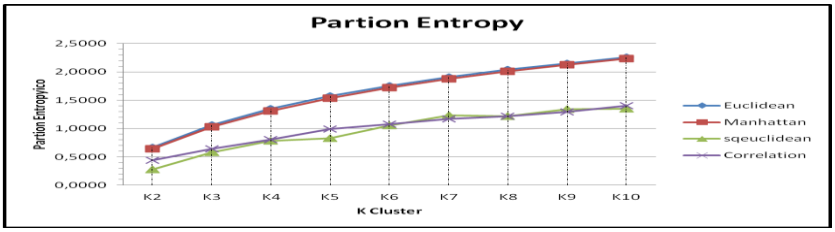
Gráfico 4 – Validação o algoritmo fuzzy c-means pelo índice fuzzy silhouette.



Fonte: Do autor.

O gráfico 5, mostra os resultados apresentados pela validação do índice *partion entropy*, indicando que o índice Entropy gera resultados que deve ser contrário a *partion coeficiente.*

Gráfico 5 – Validação do fuzzy c-means pelo índice Partion entropy.



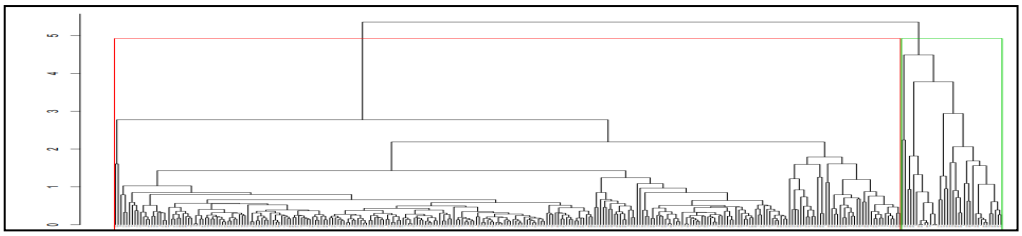
Fonte: Do autor.

Com a finalização dos experimentos, foi possível identificar que para o algoritmo *fuzzy c-means,* o melhor agrupamento apresentado para os dados educacionais, são os que aplicam a distância *seuclidean*, apresentando o número dois (*k=2*), como melhor número para geração de grupos sobre os dados.

7.3 Identificação do Modelo Final

Desta forma, o modelo identificado para agrupamento hierárquico aglomerativo, pelo algoritmo de AGNES, é formado por dois (*k=2*) grupos, usando o método da distância média, por meio da distância de *manhattan*, a figura 2 mostra o dendograma do modelo final, com os grupos destacados, para visualizar o número de grupos (*k=2*), de modo que cada nó representa um aluno.

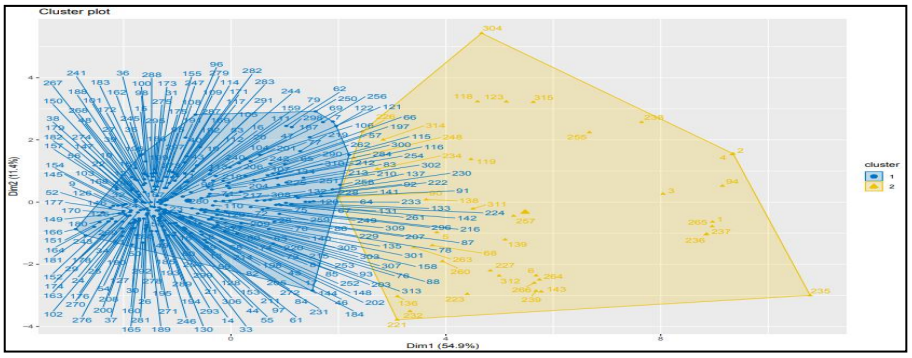
Figura 2 – Dendograma do modelo final de agrupamento pelo algoritmo AGNES.



Fonte: Do autor.

Para o método de agrupamento particional, para os quatro métodos de validação aplicada, o número dois (*k=2*) como divisão de grupos, obteve os melhores resultados, com os números próximos aos critérios dos métodos de validação. A figura 3 demonstra a divisão dos dados em dois grupos (*k=2*), de modo que existem alunos que estão entre os dois, grupos, pois tem dados semelhantes com ambos os grupos.

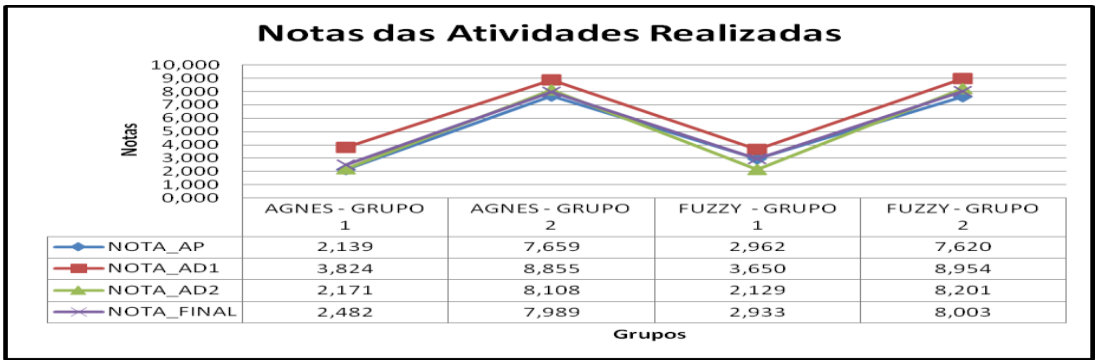
Figura 3 – Representação gráfica do modelo final gerado pelo algoritmo fuzzy c-means.



Fonte: Do autor.

Ao analisar as médias dos grupos gerados pelo algoritmo AGNES, com a distância *euclidiana* e o método de conexão distância média e o algoritmo *fuzzy c-means*, com a distância seuclidean, que foram identificados por meio das medidas de qualidade e considerados ideais para a natureza dos dados, é possível identificar que o agrupamento hierárquico aglomerativo e o agrupamento particional, são bons para agrupamento de dados educacionais, visto que ambos apresentam resultados semelhantes no pós-processamento dos dados, gerando conhecimento útil sobre os mesmos.

Gráfico 6 – Média de atividades realizadas e notas realizadas por grupo dividido por algoritmo.



	AGNES - GRUPO 1	AGNES - GRUPO 2	FUZZY - GRUPO 1	FUZZY - GRUPO 2
NOTA_AP	2,139	7,659	2,962	7,620
NOTA_AD1	3,824	8,855	3,650	8,954
NOTA_AD2	2,171	8,108	2,129	8,201
NOTA_FINAL	2,482	7,989	2,933	8,003

Fonte: Do autor.

O índice *silhouette* foi aplicado em comum, para validação dos resultados obtidos pelos algoritmos AGNES e *fuzzy c-means.* A tabela 5 apresenta os valores gerados pelo índice para os modelos finais do agrupamento hierárquico aglomerativo e particional.

Tabela 5 – Índice de validação silhouette.

	Algoritmo	
Grupo	Fuzzy C-means	Agnes
K=2	0,8665779	0,5374249

Fonte: Do autor.

O fuzzy c-means, apresentou valor superior que o gerado para validação do algoritmo AGNES, indicando que o método de agrupamento particional *fuzzy*, possui valores adequados em relação à natureza da base de dados empregada.

8 Discussão dos Resultados

Em relação ao agrupamento com dados educacionais, notou-se que quando a quantidade de grupos aponta para dois (*k=2*), os grupos são formados considerando o atributo *situação*, sendo divididos em zero (0), para alunos reprovados e um (1), para alunos aprovados, levando em consideração as características semelhantes que possibilitam a verificação de qual tratamento seria aplicado a determinado grupo. Quando aplicada a validação nos resultados obtidos pela distância de *manhattan* e *euclidiana*, para o algoritmo *fuzzy c-means*, percebe-se que quanto maior o número de grupos, inferiores são os valores apresentados pelos índices de validação, mostrando que essas não são matrizes de distâncias boas para aplicar o algoritmo sobre a natureza dos dados. A validação dos resultados gerados pelo algoritmo AGNES, mostram que as medidas de distância *manhattan* e *euclidiana,* apresentam bons resultados ao aplicar na natureza dos dados, porém ao avaliar a hierarquia pelo coeficiente de correlação cofenética, a distância *euclidiana* possui melhor solução.

A pesquisa apresenta por Zhang e Qin (2018), é uma pesquisa teórica, que destaca as principais etapas para a mineração de dados educacionais. Os autores indicam que para minerar dados educacionais, o processo deve iniciar pelo pré-processamento, seguindo para a mineração e a interpretação dos resultados. Conforme os autores a utilização correta dos algoritmos, melhora a eficiência da pesquisa e valor do conhecimento.

Os autores Ramos et al. (2016), utilizam o algoritmo *k-means* e o método de *ward*, para identificar o melhor modelo para analise de dados educacionais pelos métodos de agrupamento hierárquico e particional. Os autores identificaram quatro (*k=4*) como o número ideal de *clusters*, usando a matriz de similaridade *euclidiana*. Por meio de uma matriz para verificar a semelhança entre os grupos gerados pelos

algoritmos, os resultados apresentados demonstraram que os métodos hierárquicos e particional são bons para a análise de dados educacionais, o que se assemelha ao ocorrido nesta pesquisa, mesmo usando algoritmos diferentes e dois (*k=2*) como número ideal de grupos, verificando a semelhanças entre os grupos, formados por ambas as abordagens, mostram eficiência para aplicação em dados educacionais. Diferentemente desta pesquisa, os autores Ramos et al. (2012), não usaram as medidas de qualidade para identificar pelos índices, o melhor método de agrupamento. Nesta pesquisa, pelo índice de *silhouette*, foi possível identificar o método de agrupamento hierárquico, como modelo final.

Na pesquisa de Silva e Santos (2018), buscaram por meio de métodos estatísticos e analise de *cluster*, identificar padrões existentes nos dados. Assim como nesta pesquisa, os autores usaram o teste de *Kolmogorov-Smirnov*, para a verificação da normalidade das amostras o que atestou que não era necessária a aplicação da mesma, enquanto que nesta pesquisa, os dados foram normalizados. A eficiência da divisão de *cluster* foi testada, aplicando de 2 até 10 grupos*.* Os autores utilizaram o algoritmo *fuzzy c-means* para minerar os dados e o índice PBM, para validação. Este índice não foi aplicado nesta pesquisa, porém a divisão dois (*k=2)* foi considerada como melhor resultado assim como o resultado apresentado pelos índices desta pesquisa.

A pesquisa apresentada por Esmeraldino (2017), não utiliza dados educacionais, porém verifica quais métodos de agrupamento hierárquico possui consistência sobre a base de dados. Assim como nesta pesquisa, o autor usa o algoritmo AGNES para mineração de dados, comprando as distâncias de similaridade de *Manhattan* e euclidiana*,* com os métodos de conexão *ward*, distância média, menor distância e maior distancia, utilizando índices de critérios interno e externo para validar os resultados, e o coeficiente de correlação cofenética para validar a hierarquia. A distância euclidiana, assim como nesta pesquisa teve melhor resultado na distância cofenética, porém os índices de validade desta pesquisa indicam a distância *manhattan* com o método de conexão distância média como modelo final, e Esmeraldino (2017), aponta como modelo final o agrupamento com a distância *euclidiana* e método de conexão distância média. Esta diferença pode ocorrer, pois as bases de dados da pesquisa são de áreas distintas, indicando para a pesquisa de Esmeraldino (2017) a divisão de grupos como oito (*k=8*), enquanto para está pesquisa foi identificado o número dois (*k=2*).

9 CONCLUSÃO

A mineração de dados é um conceito que ganha força com o aumento do volume de dados, principalmente na área educacional, principalmente com o surgimento da educação a distância, em que a interação do aluno é realizada por meio de ambientes virtuais de aprendizagem. Desta forma, baseado na mineração de dados, surge o conceito do *educational data mining* (EDM), que aplica as técnicas de mineração de dados em bases educacionais. Com a possibilidade de formar grupos, em que os dados são divididos conforme a similaridade entre si, o método de agrupamento é uma das principais tarefas da mineração de dados e *educational data mining*.

O método de agrupamento hierárquico aglomerativo, divide os dados em uma dendograma, separando em grupos a facilitando a visualização da hierarquia de dados. O método de agrupamento particional *fuzzy*, divide os dados em grupos, possibilitando que um dado possa pertencer a mais de um grupo.

As dificuldades encontradas foram em relação à definição de grupos, visto que os resultados apresentados pela validação dos resultados dos algoritmos podem ser próximos, dificultando identificar o melhor número de grupos para dividir a natureza dos dados, que ficou melhor para visualização quando plotados em um gráfico. A escolha da ferramenta para aplicar as técnicas de mineração de dados, também foi uma decisão difícil, visto que existem inúmeras e foi necessário encontrar uma que atendesse todas as etapas presentes na pesquisa. Para obter os atributos para aplicar os dados, fez necessário o estudo da plataforma *moodle*, visto que foi necessário o conhecimento da mesma e as tabelas do banco para aplicar as *SQL* e montar a tabela final para mineração dos dados.

Apesar das dificuldades, os resultados obtidos pela pesquisa, foram considerados satisfatórios para o comprimento do objetivo desta pesquisa, possibilitando a identificação dos modelos finais, definidos por meio das medidas de qualidade e análise. Ambos os métodos de agrupamento aglomerativo hierárquico, pelo algoritmo AGNES e agrupamento particional, pelo algoritmo *fuzzy c-means,* foram considerados bons para análise de dados educacionais, baseado na base de dados da disciplina de engenharia de segurança do trabalho, visto que os resultados pós-processamento foram semelhantes para os dois modelos de agrupamento.

O algoritmo AGNES, por meio das medidas de qualidade, obteve melhores resultados com a medida de similaridade *euclidiana*, utilizando o método

de conexão, distância média. O algoritmo *fuzzy c-means*, obteve melhores resultados utilizando para cálculo da matriz de similaridade a distância de *seuclidean*. Ao comparar o índice de *silhouette*, que foi aplicado para validação dos dois algoritmos, o agrupamento particional, gerado por meio do *fuzzy c-means* obteve melhor resultado.

10 Referências

ESMERALDINO, Rodrigo Burigo. **Modelo de Análise de Agrupamento para a Infertilidade Masculina em Dados Biomédicos de Uma Clinica do Extremo Sul Catarinense.** 2017. 115 f. TCC (Graduação) - Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense, Criciúma, 2017. Cap. 6.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, v. 39, n. 11, p. 27-34, 1996.

FIELD, Andy. **Descobrindo a estatística usando o SPSS-2**. Bookman Editora, 2009.

GOLDSCHMIDT, Ronaldo; BEZERRA, Eduardo; PASSOS, E. Data mining: conceitos, técnicas, algoritmos, orientações e aplicações. **Rio de Janeiro-RJ: Elsevier**, p. 56-60, 2015.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. **Data mining**: concepts and techniques. Elsevier, 2011.

HUANG, Wei-qing et al. an efficient cluster mining algorithm for the internal motion target path based on the enhanced AGNES. In: **Trustcom / BigDataSE / ISPA, 2015 IEEE** . IEEE, 2015. p. 1318-1323.

MESQUITA LOPES, Manuela; BRANCO, Verônica Teixeira Franco Castelo; SOARES, Jorge Barbosa. Utilização dos testes estatísticos de Kolmogorov-Smirnov e Shapiro-Wilk para verificação da normalidade para materiais de pavimentação. **Transportes**, v. 21, n. 1, p. 59-66, 2013.

RAMOS, Jorge Luis Cavalcanti et al**.** A Comparative Study between Clustering Methods in Educational Data Mining**. IEEE Latin America Transactions**, v. 14, n. 8, p. 3755-3761, 2016.

RAMOS, Jorge Luis Cavalcanti et al. **Um Modelo Preditivo da Evasão dos Alunos na EAD a Partir dos Construtos da Teoria da Distância Transacional**. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). 2017. p. 1227.

SATYANARAYANA, Ashwin; NUCKOWSKI, Mariusz. Data Mining using Ensemble Classifiers for Improved Prediction of Student Academic Performance. 2016

SILVA, Leandro Augusto da; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução a Mineração de Dados:** Com Aplicação em R. Rio de Janeiro: Elsevier, 2016. 376 p.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao datamining: mineração de dados**. Ciência Moderna, 2009.

XU, Rui; WUNSCH, Don. Clustering. John Wiley & Sons, 2008

ZHANG, Wei; QIN, Shiming. A brief analysis of the key technologies and applications of educational data mining on online learning platform. In: **Big Data Analysis (ICBDA), 2018 IEEE 3rd International Conference on**. IEEE, 2018. p. 83-86.

ANEXO(S)

ANEXO A – CARTA DE ACEITE SETOR DE TECNOLOGIA E INFORMAÇÃO UNESC E SETOR DE EDUCAÇÃO A DISTÂNCIA UNESC

unesc
Universidade
do Extremo
Sul Catarinense

CARTA DE ACEITE

Declaramos, para os devidos fins que se fizerem necessários, que concordamos em disponibilizar a base de dados (sem a identificação dos indivíduos) das disciplinas ofertadas na modalidade a distância na Instituição Universidade do Extremo Sul Catarinense, localizada na Av. Universitária, nº 1105 – Bairro Universitário, Criciúma/ SC – 88806-000, para o desenvolvimento da pesquisa intitulada "Educacional Big Data em Educação a Distância" sob a responsabilidade da professora responsável Merisandra Côrtes de Mattos Garcia e pesquisadoras Alini Manrangoni Eyng, Ana Claudia Fontana Medeiros, Laíne Dimer e Stefany Mendes do Nascimento do Curso de Ciência da Computação da Universidade do Extremo Sul Catarinense – UNESC, pelo período de execução previsto no referido projeto.

Rogério Antônio Casagrande
Gerente de TI

Prof. Me. Rogério Antônio Casagrande
Gerente do Setor de Tecnologia da Informação da UNESC

unesc
Universidade
do Extremo
Sul Catarinense

CARTA DE ACEITE

Declaramos, para os devidos fins que se fizerem necessários, que concordamos em disponibilizar a base de dados (sem a identificação dos indivíduos) das disciplinas ofertadas na modalidade a distância na Instituição Universidade do Extremo Sul Catarinense, localizada na Av. Universitária, nº 1105 – Bairro Universitário, Criciúma/ SC – 88806-000, para o desenvolvimento da pesquisa intitulada "Educacional Big Data em Educação a Distância" sob a responsabilidade da professora responsável Merisandra Côrtes de Mattos Garcia e pesquisadoras Alini Manrangoni Eyng, Ana Claudia Fontana Medeiros, Laíne Dimer e Stefany Mendes do Nascimento do Curso de Ciência da Computação da Universidade do Extremo Sul Catarinense – UNESC, pelo período de execução previsto no referido projeto.

Profa. Dra. Graziela Fatima Giacomazzo
Coordenadora do Setor de Educação a Distância da UNESC

Profª Dra. Graziela Fátima Giacomazzo
Coordenadora Sead UNESC

ANEXO B – Carta de aprovação do Comitê de Ética em Pesquisa

RESOLUÇÃO

O Comitê de Ética em Pesquisa UNESC, reconhecido pela Comissão Nacional de Ética em Pesquisa (CONEP) / Ministério da Saúde analisou o projeto abaixo.

Parecer nº: 2.857.694 **CAAE:** 96550518.5.0000.0119

Pesquisador (a) Responsável: MERISANDRA CÔRTES DE MATTOS GARCIA

Pesquisador (a): ALINI MARANGONI EYNG
ANA CLAUDIA FONTANA MEDEIROS
LAINE DIMER
STEFANY MENDES DO NASCIMENTO

Título: "EDUCATIONAL BIG DATA EM EDUCAÇÃO A DISTÂNCIA".

Este projeto foi **Aprovado** em seus aspectos éticos e metodológicos, de acordo com as Diretrizes e Normas Internacionais e Nacionais. Toda e qualquer alteração do Projeto deverá ser comunicada ao CEP. Os membros do CEP não participaram do processo de avaliação dos projetos onde constam como pesquisadores.

Criciúma, 30 de agosto de 2018.

Renan Antônio Ceretta
Coordenador do CEP