Análise de Agrupamento Hierárquico aplicada aos microdados do ENADE do curso de graduação em Ciência da Computação

Nicolas P. B. Vista[1], João B. Barasuol[1], Michele F. Figueiró[1,2], Patrícia M. M. Chicon[1], Angela P. Ansuj[2]

[1]Centro de Ciências Humanas e Sociais – Universidade de Cruz Alta (UNICRUZ)
Cruz Alta – RS – Brasil

[2]Departamento de Estatística – Universidade Federal de Santa Maria (UFSM)
Santa Maria – RS - Brasil

nicolaspastorio@yhaoo.com.br,joaobrenobarasuol@hotmail.com,
{mfigueiro, pmozzaquatro}@unicruz.edu.br, angelaansuj@yahoo.com

Abstract. *Seeking to improve the quality of teaching is something that is desired more and more, because the better able the student leaves the university, the greater his chance of success in the job market. But how to know which aspects should be improved by institution? The article aims at applying the Data Mining technique called Hierarchical Grouping Analysis applied to ENADE microdata to extract information regarding the quality of the Computer Science course of Higher Education Institutions (HEIs) of the Consortium of Gauchas Community Universities (COMUNG). According to the results obtained, four distinct groups were formed following the performance of the HEIs of the COMUNG group in the general, general formation and specific knowledge in the ENADE grades.*

Resumo. *Buscar melhorar a qualidade de ensino é algo que se deseja cada vez mais, pois quanto mais capacitado o aluno sai da universidade, maior será sua chance de sucesso no mercado de trabalho. Mas como saber quais aspectos devem ser melhorados pelas instituições? O artigo tem como objetivo a aplicação da técnica pertencente à Mineração de Dados chamada de Análise de Agrupamentos Hierárquicos aplicada aos microdados do ENADE para extrair informações referentes à qualidade do curso de Ciência da Computação das Instituições de Ensino Superior (IES) do grupo gaúcho Consórcio das Universidades Comunitárias Gaúchas (COMUNG). Com base nos resultados obtidos, foram formados quatro grupos distintos de acordo com o desempenho das IES do grupo COMUNG nas notas brutas geral, de formação geral e de conhecimento específico no ENADE.*

1. Introdução

A Mineração de Dados (Data Mining em inglês, ou MD), segundo Amo (2004), simplesmente, trata-se de extrair ou minerar conhecimento de grandes volumes de dados. Em outras palavras, é o processo de analisar dados de diferentes perspectivas,

categorizar estes dados e identificar relações entre estes, com a intenção de extrair informações úteis.

Durante anos, empresas de várias áreas tem utilizado a MD para a extração de conhecimento. Pode-se citar um exemplo de detecção de padrões (compras, pagamentos, etc.) de um cliente em um supermercado, visando melhorar a interação da empresa com o cliente. A continua evolução no poder computacional e dos softwares de estatística que vem acontecendo durante os anos, vem aumentando consideravelmente a precisão dos resultados e a diminuição dos custos.

A MD prove um método automático para descobrir padrões em dados, sem a tendenciosidade e a limitação de uma análise baseada meramente na intuição humana (BRAGA, 2005). Basicamente o processo de encontrar relações ou padrões em grandes bases de dados relacionadas.

A MD é dividida, basicamente em três diferentes hierarquias de aprendizados: Aprendizado Intuitivo, Aprendizado Supervisionado e Aprendizado Não-supervisionado, este último sendo o método utilizados neste artigo.

O estudo apresentado neste artigo aplica uma técnica de mineração de dados chamada de clusterização hierárquica aos microdados do ENADE do ano de 2014 para formar grupos das Instituições de Ensino Superior (IES) do Consórcio das Universidades Comunitárias Gaúchas (COMUNG) que oferecem o curso de graduação em Ciência da Computação.

2. Trabalhos correlatos

A pesquisa realizada por Renan Belazari Bento (2015) intitulada Validação da Técnica de Clusterização Hierarchical Clustering: Comparação Entre os Métodos Median, Single e Ward Integrantes da Função Linkage teve como objetivo validar a técnica de clusterização Hierarchical Clustering, integrante da área de mineração de dados.

O autor Jean Metz (2006) desenvolveu a pesquisa intitulada Interpretação de clusters gerados por algoritmos de clustering hierarquico, que teve como objetivo propor e desenvolver um modulo de aprendizado não-supervisionado, agregando algoritmos de clustering hierarquico e ferramentas para análise de clusters com o intuito de ajudar auxiliar o especialista de domínio na interpretação dos resultados do clustering.

A pesquisa de Gustavo Figueiredo Araújo (2007) apresentou um estudo de Codificação e Clustering de Proteinas, analisando a aplicação do método Sequence Coding by Windows (SCSW) e em seguida o método K-Means para verificar a eficácia do método anterior.

3. Análise de Agrupamentos Hierárquicos

A Análise de Agrupamento (Clusterização ou Clustering em inglês) é o termo designado para técnicas computacionais as quais tem por objetivo separar dados em grupos, com base nas características destes dados, ou seja, agrupar dados semelhantes em um mesmo grupo (ou cluster) baseando em um critério dado por uma função de similaridade, ou dissimilaridade, funções estas que serão abordadas posteriormente neste trabalho.

O propósito de identificar clusters é obter a partição de um banco de dados de registros tal que os registros possuam similaridade entre si. Com isso, permite obter características de cada clusters formados (BENTO, 2015).

Segundo Metz (2006), a técnica de clusterização é frequentemente utilizada em tarefas de exploração de dados e padrões, e uma de suas principais utilizações é na área da bioinformática detectando características e segmentações em imagens.

A clusterização segue alguns passos, que Bento (2015) cita como sendo o pré-processamento: onde se prepara e transforma os dados de acordo com sua similaridade; a seleção da medida de similaridade: onde se analisa o conjunto de dados e se escolhe a medida para o cálculo da similaridade; e a avaliação de clusters: onde os padrões gerados pelos clusters formados de acordo com a medida selecionada são avaliados quantitativamente utilizando índices estatísticos.

O Agrupamento Hierárquico é baseado em um simples conceito: agrupar variáveis (observações, elementos, dado ou objeto) passo a passo, de maneira hierárquica. Segundo Bento (2015), o algoritmo Hierarchical Clustering foi desenvolvido por King na década de 60, porém somente na mesma década que se tornou conhecido através de Johnson.

A principal vantagem desta técnica consiste no fato de que não se faz necessário qualquer parâmetro de entrada para o início do processo. Linden (2009), cita que os algoritmos hierárquicos criam uma hierarquia de relacionamentos entre os elementos, e que existem duas versões, a aglomerativa e a divisiva.

Adicionam-se variáveis aos clusters de acordo com a maneira em que a diferença entre os clusters mudam, conforme estes crescem. Isto é conhecido como Agrupamento Hierárquico Aglomerativo, também conhecido como Bottom-Up. Quando se está agrupando clusters hierarquicamente, as variáveis não podem ser movidas ou trocadas entre clusters, uma vez que esta variável foi adicionada em um cluster aglomerativo, ela permanece neste até o processo estar completo.

A Figura 1 mostra um Diagrama de atividades com os passos básicos presentes no funcionamento de um agrupamento hierárquico com aglomeração.

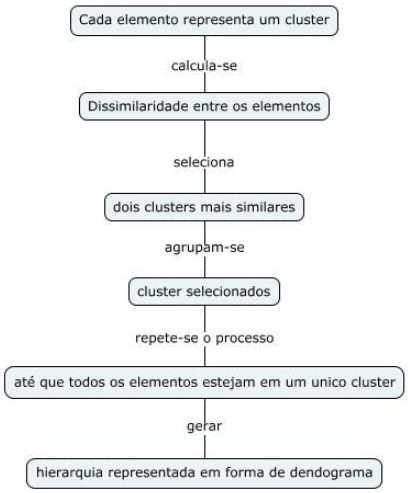

Figura 1. Diagrama de atividades de um Agrupamento Hierárquico Aglomerativo.

Este processo pode ser executado em ordem reversa, onde se inicia com somente um grande cluster que contém todas as variáveis, e então se divide os clusters, removendo as variáveis de acordo com a mudança nas distâncias entre os clusters. Este processo é conhecido como Agrupamento Hierárquico Divisivo ou também como Top-Down. Na Figura 2 é possível observar um diagrama de atividades com os passos básicos presentes no funcionamento de um agrupamento hierárquico com divisão.

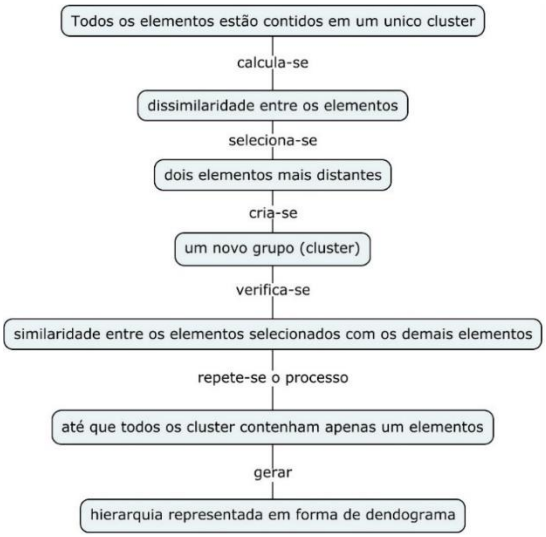

Figura 2. Diagrama de atividades de um Agrupamento Hierárquico Divisivo.

A Figura 3 mostra um dendograma que exemplifica a diferença na ordem do funcionamento e na representação dos resultados entre o agrupamento hierárquico com aglomeração e com divisão.

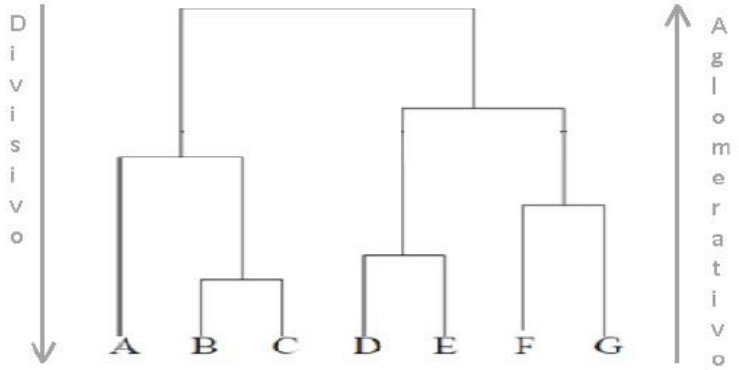

Figura 3. Dendograma de métodos Hierárquicos Aglomerativos versus Divisivos

Destas duas formas de agrupamento, a mais comum e utilizada é o Agrupamento Hierárquico Aglomerativo, e é também o método utilizado neste artigo.

4. Exame Nacional de Desempenho dos Estudantes (ENADE)

O ENADE (Exame Nacional de Desempenho de Estudantes) é um dos métodos de avaliação do SINAES (Sistema Nacional de Avaliação da Educação Superior) que é

realizado pelo INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), o qual é uma autarquia federal vinculada ao MEC (Ministério da Educação), com base nas diretrizes definidas pelo CONAES (Comissão Nacional de Avaliação da Educação Superior), órgão colegiado de coordenação e supervisão do SINAES.

O ENADE tem como objetivo, acompanhar e avaliar o rendimento dos alunos concluintes, de cursos de graduação, levando em consideração o conteúdo previsto nas diretrizes curriculares do respectivo curso em que o aluno está matriculado. Os estudantes aptos a realizar o ENADE são concluintes do ano e dos cursos que serão avaliados neste mesmo ano. Estes concluintes são aqueles que estão finalizando o curso de graduação e com mais de oitenta por cento da carga horário do curso realizada até a data de realização do exame. A inscrição para a realização do ENADE é de responsabilidade da instituição em que o aluno está matriculado, os alunos aptos, e que não forem inscritos pela instituição, não poderão participar do exame. A instituição que não realizar a inscrição dos estudantes aptos a realização do exame dentro dos prazos estipulados, poderá sofrer suspensão temporária da abertura de processo seletivo para os cursos em questão.

Dentre todos estes dados disponíveis na base de dados do ENADE referente ao ano de 2014, foram filtradas somente as instituições presentes no Rio Grande do Sul e que ofertam o curso de Ciência da Computação, após estas terem sido filtradas, foram selecionadas apenas as variáveis relevantes para gerar o agrupamento referente ao desempenho na prova do ENADE, e estas variáveis são as seguintes: Inscritos, Participantes, Nota Bruta da Formação Geral, Nota Bruta Conhecimento Especifico, Nota Bruta Geral e Conceito ENADE (contínuo).

5. Consórcio das universidades comunitárias gaúchas (COMUNG)

O grupo gaúcho COMUNG foi fundado em março de 1993, com o intuito de integrar as universidades para que fosse possível o fortalecimento individual das instituições e da comunidade universitária rio-grandense bem como da sociedade gaúcha. O Consórcio das Universidades Comunitárias Gaúchas (COMUNG) reúne as instituições comunitárias de ensino superior do estado do Rio Grande do Sul, que, diferente das universidades públicas e privadas, as universidades comunitárias não possuem fins lucrativos e têm um forte vínculo com suas comunidades, sendo assim, autênticas instituições públicas não estatais (COMUNG, 2017).

6. Procedimentos Metodológicos

Neste estudo, foi utilizada a técnica de agrupamento hierárquico com aglomeração com o objetivo de formar grupos das IES do grupo COMUNG que apresentam similaridade nos microdados do ENADE referentes ao ano de 2014 do curso de graduação em Ciência da Computação. Todas as analise estatísticas foram realizadas no software estatístico R, juntamente com sua interface RStudio. Para a análise de comparação das médias dos conceitos contínuos ENADE, IGC e CPC, foi utilizado o teste ANOVA one-way com teste post-hoc de Tukey. Foi considerado significativo os testes estatísticos com $p \leq 5\%$.

6.1. Resultados e Discussões

Após a finalização da etapa teórica, se deu início ao processo de modelagem, a qual foi desenvolvida em linguagem UML para construir e visualizar atributos de um sistema complexo, onde foram desenvolvidos os Diagramas de atividades que definem a proposta do trabalho, os passos para o desenvolvimento como um todo e da aplicação da técnica utilizada no trabalho (Figura 4).

Figura 4. Diagrama de atividades **da proposta do trabalho.**

6.1.1 Etapa I

O Desenvolvimento prático deste trabalho, se inicia com a obtenção da base de dados do ENADE, que para este trabalho foi utilizada a base de 2014. Esta base de dados está disponível no portal no INEP (INEP, 2017).

Após a obtenção dos dados, foi necessária a preparação destes para a aplicação das técnicas de mineração de dados, selecionando apenas o curso de graduação em Ciência da Computação das IES do grupo COMUNG.

Com os dados já preparados, foi feito o download do software para a aplicação das técnicas de mineração de dados. O software utilizado foi o software estatístico R, juntamente com a IDE que adiciona algumas funcionalidades e melhora a interface e utilização RStudio.

Os dados preparados do ENADE são importados para o Software R e a técnica de Agrupamento Hierárquico com Aglomeração é aplicada através de códigos, digitados e executados através do console do programa (Figura 5). São aplicadas as medidas de distância, a função de ligação Ward, e o coeficiente de correlação cofenética, para avaliar os agrupamentos gerados por cada medida de distância.

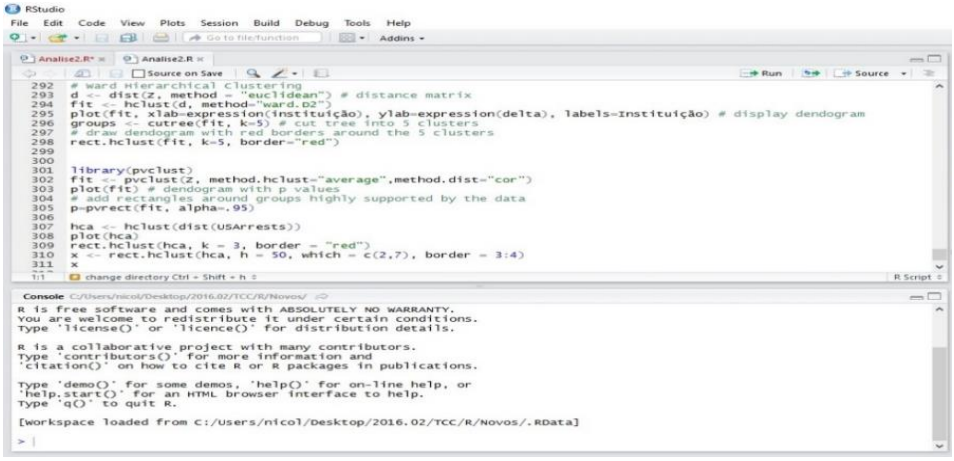

Figura 5. Tela Software R

6.1.2 Etapa II

Para a etapa 2, a base de dados utilizada para a obtenção dos resultados foi aumentada, acrescentando-se à base de dados do ENADE 2014, os conceitos contínuos do IGC (Índice Geral de Cursos) e do CPC (Conceito Preliminar de Cursos), ambos também referentes ao ano de 2014, os quais formam os Indicadores de qualidade do MEC e cujas bases de dados estão disponíveis para download através do portal INEP (INEP2, 2017).

Após a obtenção das bases de dados dos Indicadores de Qualidade, foi realizada a filtragem e seleção dos dados, selecionado as variáveis necessárias referentes ao desempenho dos acadêmicos e da instituição. Todos estes dados foram então agrupados em uma base de dados única, a qual foi utilizada para a realização dos testes e obtenção dos resultados.

A análise foi realizada em duas partes, primeiramente separando somente as instituições de ensino superior gaúchas integrantes do grupo comunitário COMUNG e posteriormente utilizando-se todas as Instituição do Rio Grande do Sul que ofertam o curso de Ciência de Computação. Portanto foram-se utilizadas duas tabelas, somente com o número de instituições diferentes, conforme pode ser observado na Tabela 5, referente aos dados dos Indicadores de Qualidade das IES gaúchas, onde aquelas destacadas na cor vermelha, são as integrantes do grupo COMUNG.

Os passos para a obtenção dos resultados, apresentados a seguir, foram executados somente para as IES do grupo COMUNG.

O processo começa com a importação da base de dados para o *software* R, a qual dá-se através do comando HC = READ.XLSX, o qual recebe o nome do arquivo onde estão armazenados os dados e o nome da planilha, conforme a Figura 6.

```
8
9
10  ### Hierarchical Analysis ###
11  hc = read.xlsx("DADOS.xlsx", sheetName = "DADOS2")
12
13
```

Figura 6. Importação da base de dados no Software R.

Após a importação da base de dados, são obtidas as medidas descritivas mínimo, máximo, média, mediana, 1º Quartil, 3º Quartil, desvio padrão, e coeficiente de variação.

A Figura 7 mostra a padronização da variáveis, através do comando SCALE(), onde cada variável é tratada e armazenada em uma matriz, que será utilizada para a obtenção dos agrupamentos. Esta etapa é necessária para a obtenção de um agrupamento mais eficaz.

```
## standarlize variables
require(base)
z1=scale(Inscritos, center=TRUE, scale=TRUE)
z2=scale(Participantes, center=TRUE, scale=TRUE)
z3=scale(NBFG, center=TRUE, scale=TRUE)
z4=scale(NBCE, center=TRUE, scale=TRUE)
z5=scale(NBG, center=TRUE, scale=TRUE)
z6=scale(ConceitoENADE, center=TRUE, scale=TRUE)
z7=scale(IGC, center=TRUE, scale=TRUE)
z8=scale(CMG, center=TRUE, scale=TRUE)
z9=scale(ND, center=TRUE, scale=TRUE)
z10=scale(NM, center=TRUE, scale=TRUE)
z11=scale(NBM, center=TRUE, scale=TRUE)
z12=scale(NBD, center=TRUE, scale=TRUE)
z13=scale(NBRT, center=TRUE, scale=TRUE)
z14=scale(NBODP, center=TRUE, scale=TRUE)
z15=scale(NBIIF, center=TRUE, scale=TRUE)
z16=scale(NBOAF, center=TRUE, scale=TRUE)
z17=scale(CPC, center=TRUE, scale=TRUE)
# vector
Z = cbind(z1, z2, z3, z4, z5, z6, z7, z8, z9, z10, z11, z12, z13, z14, z15, z16, z17)
```

Figura 7. Padronização das Variáveis.

Com todas as variáveis padronizadas e armazenadas em uma matriz, foram aplicadas as medidas de distância através do comando DIST() conforme a Figura 8, que recebe a matriz com as variáveis e a distância utilizada. Estas medidas formam as matrizes de dissimilaridade, de onde serão obtidos os agrupamentos.

```


####################################################
## III) METHOD Maximum "D3" = CHEBYCHEV
####################################################
D3 = dist(Z, method='maximum', diag=TRUE) ### Maximum Distance
D3


############################
## IV) METHOD Manhattan "D4"
############################
D4 = dist(Z, method='manhattan', diag=TRUE) ### Manhattan Distance
D4

```

Figura 8. Aplicação da Medida de Distância.

Para a obtenção do Agrupamento Hierárquico Aglomerativo, apresentado na Figura 9, é utilizado o comando HCLUST(), o qual recebe a matriz de dissimilaridade e a função de ligação WARD. A função de ligação Ward, também conhecido como ‘Mínima Variância’, foi originalmente proposta por Joe H. Ward (1963), e basea-se no pressuposto que o cluster deve ser o mais homogêneo possível.

Nesta função de ligação, a distância entre dois clusters é dada através da soma dos quadrados entre os dois clusters feita sobre todos os elementos, onde em cada iteração, são combinados os cluster que apresentarem o menor aumento na soma dos quadrados dentro do cluster. A função Ward ainda tende a resultar em clusters com tamanhos mais similares, por minimizar a variação interna dos elementos. Este método é considerado muito eficiente, porém, tende a criar clusters de tamanho pequeno.

O coeficiente de correlação cofenética foi calculado utilizando o comando COPHENETIC(), o qual recebe o agrupamento gerado anteriormente. O coeficiente de correlação cofenética tem como objetivo principal, avaliar os agrupamentos gerados por uma técnica de mineração de dados.

Por fim a função PLOT() exibe o agrupamento formado no formato de dendogramas.

```
232
233 ###################
234 ### Ward with Quadratic Euclidian Distance "hc21"
235 ###################
236 hc21 = hclust(D2, method='ward.D2')
237 d21 <- cophenetic(hc21)
238 cor(D2, d21)
239 #####
240 ### coeficiente de correlação cofenética = 0,43 < 0,7 método utilizado não foi
241 ### adequado para resumir a informação ao conjunto de dados.
242 #####
243 plot(hclust(D, method='ward.D2'), xlab=expression(instituição), ylab=expression(delta), labels=Instituição)
244
245 ###################
246 ### Ward with Maximum Distance "hc22"
247 ###################
248 hc22 = hclust(D3, method='ward.D2')
249 d22 <- cophenetic(hc22)
250 cor(D3, d22)
251 #####
252 ### coeficiente de correlação cofenética = 0,69 < 0,7 método utilizado não foi
253 ### adequado para resumir a informação ao conjunto de dados.
254 #####
255 plot(hclust(D3, method='ward.D2'), xlab=expression(instituição), ylab=expression(delta), labels=Instituição)
256
257
```

Figura 9. Aplicação da função de ligação Ward e do coeficiente de correlação cofenética

De acordo com as medidas descritivas obtidas, tem-se:

INSCRITOS: Os valores mínimo e máximo foram de 8 (UCPEL) e 86 (PUCRS), respectivamente. A média foi de 37,12 com desvio padrão 20,04. 25% do número de inscritos das IES foi inferior a 24,25, 50% inferior a 32,50, enquanto que 75% inferior a 48 inscritos.

PARTICIPANTES: Os valores mínimo e máximo foram de 6 (UCPEL) e 77 (PUCRS), respectivamente. A média foi de 32.75 com desvio padrão 17, 00.

NBFG (Nota Bruta da Formação Geral): Os valores mínimo e máximo foram de 30,80 (UCPEL) e 70,41 (UNISINOS), respectivamente. A média foi de 59,81 com desvio padrão 8,87.

NBCE (Nota Bruta do Conhecimento Especifico): Os valores mínimo e máximo foram de 24,50 (UCPEL) e 49,00 (UNISINOS), respectivamente. A média foi de 37,33 com desvio padrão 7,58.

NBG (Nota Bruta Geral): Os valores mínimo e máximo foram de 26,10 (UCPEL) e 54,40 (UNISINOS), respectivamente. A média foi de 42,98 com desvio padrão 7,25.

CONCEITO ENADE (contínuo): Os valores mínimo e máximo foram de 0,43 (UCPEL) e 3,60 (UNISINOS), respectivamente. A média foi de 2,23 com desvio padrão 0,84.

CONCEITO IGC (Índice Geral de Cursos) (contínuo): Os valores mínimo e máximo foram de 2,75 (URI – FREDERICO WESTPHALEN) e 3,58 (UNISINOS), respectivamente. A média foi de 2,99 com desvio padrão 0,27.

CONCEITO CMG (Conceito Médio da Graduação) (contínuo): Os valores mínimo e máximo foram de 2,70 (URI – FREDERICO WESTPHALEN) e 3,30 (UNISINOS), respectivamente. A média foi de 2,86 com desvio padrão 0,18.

ND (Número de Docentes): Os valores mínimo e máximo foram de 15 (URI – SANTIAGO) e 60 (UNISINOS), respectivamente. A média foi de 29,75 com desvio padrão 0,63.

NM (Número de Matrículas): Os valores mínimo e máximo foram de 11 (UCPEL) e 507 (PUCRS), respectivamente. A média foi de 185,44 com desvio padrão 12,75.

NBM (Nota Bruta - Mestre): Os valores mínimo e máximo foram de 0,67 (URI – SANTIAGO) e 1 (PUCRS), respectivamente. A média foi de 0,92 com desvio padrão 131,48.

NBD (Nota Bruta - Doutores): Os valores mínimo e máximo foram de 0 (URI – ERECHIN) e 0,70 (PUCRS), respectivamente. A média foi de 0,29 com desvio padrão 0,11.

NBRT (Nota Bruta - Regime de Trabalho): Os valores mínimo e máximo foram de 0,39 (UNILASSALE) e 1 (FEEVALE), respectivamente. A média foi de 0,76 com desvio padrão 0,20.

NBODP (Nota Bruta - Organização Didático-Pedagógica): Os valores mínimo e máximo foram de 4,51 (UNISC) e 5,58 (UNIJUI – SANTA ROSA), respectivamente. A média foi de 4,96 com desvio padrão 0,19.

NBIIF (Nota Bruta - Infraestrutura e Instalações Físicas): Os valores mínimo e máximo foram de 4,42 (UNCRUZ) e 5,70 (UNIFRA), respectivamente. A média foi de 5,22 com desvio padrão 0,33.

NBOAF (Nota Bruta - Oportunidades de Ampliação da Formação): Os valores mínimo e máximo foram de 3,98 (UCPEL) e 5,61 (UNIJUI - SANTA ROSA), respectivamente. A média foi de 4,85 com desvio padrão 0,33.

CONCEITO CPC (Conceito Preliminar de Cursos) (contínuo): Os valores mínimo e máximo foram de 1,47 (UCPEL) e 3,63 (UNISINOS), respectivamente. A média foi de 2,56 com desvio padrão 0,44.

A variável INSCRITOS foi a que apresentou maior variabilidade, ou seja, seus valores estão mais afastados da média, representando assim o grupo mais heterogêneo (CV=53,97%), enquanto que a variável NBM (Nota Bruta - Mestre) foi a que apresentou menor variabilidade, ou seja, seus dados estão mais próximos da média, representando, assim, o grupo mais homogêneo (CV = 1,43%).

Estas comparações foram feitas utilizando-se apenas a função de ligação (linkage) Ward, pelo fato de, após realizada uma pesquisa bibliográfica, foi constatado que a grande maioria dos artigos na literatura especifica, vêm fazendo o uso desta função, e também pelo fato desta gerar clusters da maneira mais homogênea possível.

Além disso, a função de ligação Ward leva em consideração dados que apresentam grande variabilidade, como é o caso da base de dados dos Indicadores de Qualidade de 2014 utilizada. Por estes motivos, não foram testadas outras medidas para este trabalho. Para indicar o melhor agrupamento, foram extraídos os coeficientes de correlação cofenética obtidos na comparação entre as medidas de distância feita acima, e os valores gerados foram de (Tabela 1):

Tabela 1. Coeficiente de correlação cofenética IES do grupo COMUNG.

Medida de distância de dissimilaridade	Coeficiente de Correlação Cofenética
Euclidian	0,5887905
Quadratic	**0,743888**
Manhattan	0,6738376

Minkowsky	0,5887905
Canberra	0,2599587
Maximum	0,6207279

Através dos valores obtidos, pode-se definir que o melhor agrupamento foi aquele que utilizou a medida de distância quadrática, pois conforme define o coeficiente de correlação cofenética, quando mais próximo a 1 (um) o valor, melhor foi o agrupamento. A Figura 10 apresenta o dendograma gerado utilizando-se estas medidas.

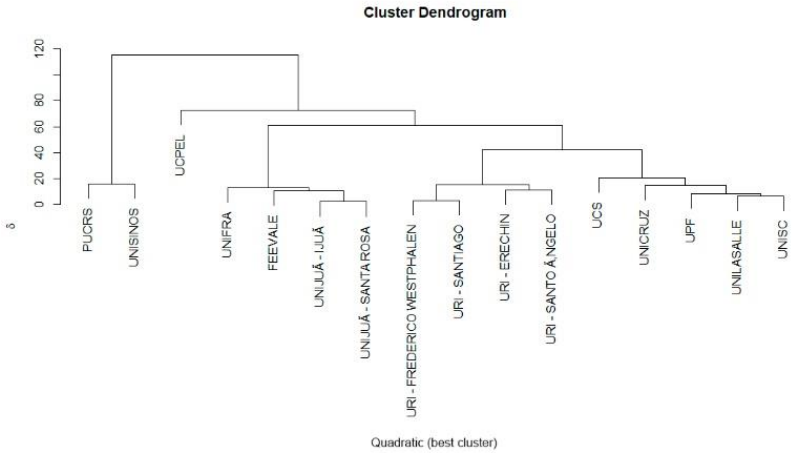

Figura 10 - Dendograma das IES do grupo COMUNG.

A medida de distância Quadrática foi selecionada como a que apresentou o melhor agrupamento pelo fato de ter obtido o maior valor do coeficiente de correlação cofenética entre as medidas testadas. Analisando o dendograma gerado com a utilização da medida de distância Quadrática com a função de ligação Ward, pode-se identificar a formação de quatro grupos, conforme apresentado pela Figura 11.

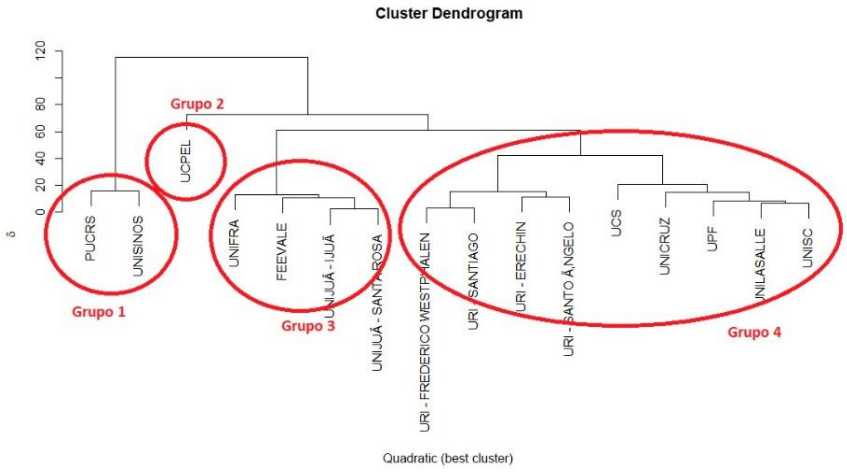

Figura 11 - Grupos das IES do grupo COMUNG.

Os 4 grupos identificados no dendograma apresentado pela Figura 11, são compostos pelas seguintes instituições do grupo COMUNG:

GRUPO 1: PUCRS e UNISINOS; as médias dos conceitos (contínuo) ENADE, IGC e CPC foram de 3,37, 3,55 e 3,47, respectivamente.

GRUPO 2: UCPEL; as médias dos conceitos (contínuo) ENADE, IGC e CPC foram 0,43, 3,07 e 1,47, respectivamente.

GRUPO 3: UNIFRA, FEEVALE, UNIJUI – IJUI e UNIJUI – SANTA ROSA; as médias dos conceitos (contínuo) ENADE, IGC e CPC foram de 2,58, 3,09 e 2,96, respectivamente.

GRUPO 4: URI – FREDERICO WESTPHALEN, URI – SANTO ANGELO, URI ERECHIN, URI – SANTO ANGELO, UCS, UNICRUZ, UPF, UNILASSALE e UNISC; as médias dos conceitos (contínuo) ENADE, IGC e CPC foram de 2,01, 2,81 e 2,29, respectivamente.

Para a análise de comparação das médias dos conceitos ENADE, IGC e CPC entre os grupos 1, 3 e 4, foi realizado o teste ANOVA one-way. De acordo com o teste F, α=5%, há uma diferença estatisticamente significativa dos conceitos (contínuo) ENADE ($p=0,003$), IGC ($p<0,001$) e CPC ($p<0,001$) entre os GRUPOS 1, 3 e 4.

7. Considerações Finais

O presente trabalho teve como objetivo realizar uma análise de agrupamento com base nos dados do índice de qualidade ENADE das Instituições de Ensino Superior do grupo COMUNG, que ofertam o curso em graduação de Ciência da Computação. Este estudo possibilitou uma análise exploratória dos dados dos Indicadores de Qualidade referentes ao ano de 2014 do curso de Ciência da Computação.

Este trabalho ainda abre a possibilidade de realização de várias outras pesquisas nesta área, como a aplicação das técnicas utilizadas na base de dados de indicadores de qualidade de anos anteriores ou posteriores a 2014, possibilitando assim a observação do desempenho de uma certa IES ao longo dos anos. Pode-se também expandir a pesquisa para outros cursos de graduação, não somente ao curso de Ciência da Computação que é o foco do presente trabalho.

Referências

Araújo, Gustavo Figueiredo. Codificação e Clustering de Proteinas. Monografia de graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras – UFLA, 2007

Amo, Sandra de. “Técnicas de Mineração de Dados”, Curso de Faculdade de Computação, Universidade Federal de Uberlândia, Uberlândia, 2004.

Bento, Renan Delazari. “Validação da técnica de clusterização hierarchical clustering: comparação entre os metodos median, single e ward integrantes da função linkage”. 205 f. TCC (Graduação) - Curso de Ciência da Computação, Universidade de Cruz Alta – Unicruz, Cruz Alta- RS, 2015.

BRAGA, Luis Paulo Vieira. “Introdução à Mineração de Dados”.2ª edição revisada e ampliada. Rio de Janeiro: E-papers Serviços Editoriais, 2005. 211 p.

COMUNG. Disponível em <http://www.comung.org.br>. Acesso em: 20 out. 2017.

INEP. Disponível em <http://portal.inep.gov.br/basica-levantamentos-acessar>. Acesso

em: 20 out. 2017.

INEP2. Disponível em < http://portal.inep.gov.br/indicadores-de-qualidade>. Acesso em: 20 out. 2017.

Linden, Ricardo. “Técnicas de Agrupamento”. Revista de Sistemas de Informação da Fsma**,** Macaé, v. 4, n. 4, p.18-36, jan. 2009.

Metz, Jean. Interpretação de clusters gerados por algoritmos de clustering hierárquico. 126 f. Dissertação (Mestrado) - Curso de Ciência da Computação, Instituto de Ciências Matemáticas e de Computação - ICMC, Universidade de São Paulo, São Paulo, 2006.

R. Disponível em < www.r-project.org>. Acesso em: 20 out. 2017.

RStudio. Disponível em <www.rstudio.com>. Acesso em: 20 out. 2017.

Ward, Joe H. “Hierarchical Grouping to Optimize na Objective Function”. Journal of the American Statistical Association**,** v. 58, n. 4, p.236-244, mar. 1963.