UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Uso de *Machine Learning* para Classificação de Fornecedores no Contexto da *Data Science*

Laércio Almeida de Siqueira Junior

Itajubá
2021

UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Laércio Almeida de Siqueira Junior

Uso de *Machine Learning* para Classificação de Fornecedores no Contexto da *Data Science*

Dissertação submetida ao programa de Pós-Graduação em Engenharia de Produção como parte dos requisitos para obtenção do Título de Mestre em Ciências em Engenharia de Produção.

Área: Engenharia de Produção

Orientador: Prof. Dr. Alexandre Ferreira de Pinho

Itajubá

2021

DEDICATÓRIA

Dedico à minha mãe Rosimeire Almeida, falecida no decorrer deste estudo, e à minha namorada Camila Akemi que sempre me incentivaram à continuar persistindo.

AGRADECIMENTOS

Aos meus pais, Rosimeire Almeida de Siqueira e Laércio José Corrêa de Siqueira, por todo apoio e ensinamentos dados durante minha vida.

Em especial à minha mãe, que findou sua jornada na Terra, mas que sempre vai permanecer no meu coração e é a pessoa fundamental para me tornar o homem que sou. Obrigado por ter estado sempre ao meu lado. Mesmo quando lutava contra uma doença cruel nunca deixou de cuidar de mim.

Reitero o agradecimento mais que especial à minha querida e saudosa mãe. Farei de tudo para honrar seu nome e ser o homem que você me criou para ser.

À minha namorada, Camila Akemi Souza Kushiyama, por estar sempre ao meu lado mesmo nos momentos mais difíceis. Sem você não teria conseguido chegar até aqui.

Ao meu orientador, Alexandre Ferreira de Pinho, pela orientação deste trabalho, pelos debates sobre os conceitos mais complexos, por ter me mostrado o mundo de *Data Science* que tanto me fascinou e principalmente por não ter desistido de mim quando fiquei mal com os problemas pessoais que sofri nos últimos tempos.

A todos os professores e servidores da UNIFEI, em especial aos professores Carlos Melo, Pedro Paulo Balestrassi, José Henrique, Carlos Sanches, Fabiano Leal, José Antônio de Queiroz, José Arnaldo Barra Montevechi e todos os outros que pacientemente transmitiram seus conhecimentos.

Aos amigos da UNIFEI, em especial João Victor que me incentivou quando estava quase desistindo de defender meu trabalho.

Ao meu amigo Matheus Rocha.

Ao professor Renato Lima, coordenador do programa e que sempre esteve solicito e pronto a resolver qualquer eventualidade surgida nestes tempos complexos de uma pandemia mundial.

A CAPES, FAPEMIG e CNPq pelo apoio e incentivo à pesquisa, em especial a essa.

A Universidade Federal de Itajubá, a quem sempre terei uma dívida eterna de gratidão.

A todos vocês, meu muito obrigado!

EPÍGRAFE

“Você deve aproveitar ao máximo os pequenos desvios porque é aí que você encontrará as coisas mais importantes do que as que deseja.”

Yoshihiro Togashi

RESUMO

A tomada de decisão por grupos, públicos ou privados, é indispensável para o desenvolvimento das organizações e encontrar mecanismos que apoiem os gestores de forma mais assertiva é fundamental para essa finalidade. Saber utilizar dados brutos transformando-os em conhecimento permite que essas decisões sejam baseadas em dados além de puramente em intuição. Dentre as decisões importantes tomadas por qualquer organização, a classificação e seleção de fornecedores é uma prática importante para a engenharia de produção e a *Data Science* é um campo ascendente que estuda dados e como realizar essa transformação de dados brutos em conhecimento. Para essa pesquisa foram utilizados dados reais dos fornecedores de uma empresa do setor aeronáutico em suas análises. Então, esta pesquisa atuou entre *Data Science e* Classificação e Seleção de fornecedores e teve seu foco no problema conhecido como clusterização (ou agrupamento) que é a segmentação de dados em regiões o mais homogêneas possíveis quando não se apresentam categorias prévias e busca-se resolver este problema auxiliando no gerenciamento dos fornecedores. Isso acontece na prática utilizando-se de ferramentas de *Data Science* conhecidas como *Machine Learning* que consistem em algoritmos que podem ser utilizados na segmentação de grupos sem nenhuma classificação inicial. Para o desenvolvimento utilizou-se o procedimento CRISP-DM que permite solucionar problemas de análises de dado ajudando a estruturar o pensamento científico. Desta forma, por meio do auxílio desse procedimento, esta dissertação teve como objetivo geral utilizar a técnica de *Machine Learning* para auxiliar na classificação e seleção de fornecedores dessa organização. Tendo dois objetivos específicos, o primeiro consistindo na demonstração do funcionamento e comportamento dos algoritmos clássicos de clusterização na base de dados real das técnicas clássicas. E o segundo objetivo específico consistiu em analisar esses algoritmos de clusterização em busca do mais apropriado para a base de fornecedores em estudo culminando com a criação e sugestão de um *framework* que pode ser utilizado para análises futuras de clusterização. As modelagens das clusterizações foram realizadas e por meio de validação interna e de estabilidade a eficiência delas foi testada permitindo que os dados fossem separados em *clusters*. A utilização do CRISP-DM permitiu que o *framework* para clusterização fosse proposto.

Palavras-chave: *Data Science, Machine Learning*, Classificação e Seleção de Fornecedores, Clusterização, Tomada de decisão baseada em dados, *framework*.

ABSTRACT

Decision-making for groups, public or private, is indispensable to the development of organizations, and searching for mechanisms to support the managers more assertively is fundamental to this goal. Know how to use raw data transforming them into knowledge allows these decisions to be based on data besides purely on intuition. Between the important decisions taken by any organization, the classification and selection of suppliers are an important practice to industrial engineering and Data Science is an ascendant field that studies data and how to realize this transformation of raw data into knowledge. To this research were used real data from suppliers of an enterprise of the aeronautical sector in its analyses. So, this research acted between Data Science and Classification and Selection of suppliers and had the focus on a problem known as clusterization that is the segmentation of data in regions as homogeneous as possible when there´s no existence of previous categories and aim to solve this problem supporting in the supplier´s management. This happens in practice using Data Science tools known as Machine Learning that are algorithms that can be used in the segmentation of groups without an initial classification. To the development has been used the procedure CRISP-DM that allows elucidate analyses´ problems helping to structure the scientific thinking. That way, by using this procedure, this dissertation had its general objetive in the use of the technique of Machine Learning to help in the classification and selection of suppliers of that organization. Having two specific objectives, the first one consisted of an analysis of those algorithms in the demonstration of the operation and behavior of the classic algorithms of clusterization of the real database. The second one consisted in analyzing those clusterization algorithms in search of the most appropriate to the supplier´s base culminating with the creation and suggestion of a framework that can be used for future clustering analises. The clustering modelings were realized and through internal and stability validations had their efficiency tested allowing the data to be split into clusters. The use of CRISP-DM allowed that the clustering framework was proposed.

Keywords: *Data Science, Machine Learning, Classification and Selection of suppliers, Clustering, Data-driven decisions, framework.*

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Segundo Van der Aalst (2016) o interesse em *Data Science* está crescendo. Tanto que muitos consideram como a área profissional do futuro. Assim como ciência da computação emergiu como uma disciplina na década de 70, agora o mundo testemunha a rápida criação de centros de pesquisa, graduações e mestrados em *Data Science*.

Segundo Schutt e O'Neil (2013) *Data Science* compreende o estudo que busca transformar informações em novas formas de valor e vai ao encontro a Cao e Yu (2018) que baseiam seu uso em entender alguma área de conhecimento por meio do gerenciamento dos dados. Para a manipulação de dados são utilizados os *Machine Learning* e segundo Jordan e Mitchell (2015) *Machine Learning* progrediu drasticamente nas últimas duas décadas, de uma curiosidade de laboratório para uma tecnologia prática em amplo uso comercial.

Mehryar, Rostamizadeh e Talwalkar (2012, 2018) definem *Machine Learning* como métodos computacionais que além de melhorar o desempenho de um sistema, fazem predições com precisão com base em experiências passadas. E as experiências passadas disponíveis aos estudos geralmente constituem-se de dados eletrônicos coletados ou produzidos. Adicionalmente, Amershi *et al.* (2019) pontuam que os métodos de *Machine Learning* ajudam engenheiros a descobrir, coletar, obter, entender e transformar dados para depois treiná-los e implantá-los.

Machine Learning e *Data Science* apresentam estudos atuais em várias áreas de conhecimento. Lewin *et al.* (2018) mostram como geram oportunidades e desafios para novas ferramentas visualizarem, compararem e entenderem a conexão da sequência de genoma para evolução de ecossistemas. Kirchdoerfer e Ortiz (2016) aplicaram no contexto de computação mecânica, mais especificamente elástica quasistática, uma forma de se utilizar de padrões para entendimento de alguma coisa. Larson e Chang (2016) realizaram um estudo alinhando análise de dados, ciência de dados e um ponto forte na engenharia de produção atualmente, as chamadas metodologias ágeis, e para onde se direcionam esses assuntos.

O foco geral de *Machine Learning* no contexto de *Data Science* está na representação dos dados de entrada e na generalização dos padrões aprendidos para uso em dados futuros. (NAJAFABADI *et al.*, 2015). Donoho (2017) afirma que um dos maiores problemas atualmente da *Data Science* consiste em torná-la mais concreta e visível. Olson *et al.* (2016) entendem que enquanto o campo de *Data Science* continuar a crescer, vai existir uma demanda cada vez maior por ferramentas que tornem *Machine Learning* acessíveis para não-especialistas.

Segundo George *et al.* (2016), as técnicas de *Data Science* permitem que pesquisadores consigam resultados mais acurados e imediatos para testar teorias. Fazendo isso, espera-se a obtenção de estimativas mais precisas. Muitos aspectos de análises de negócios estão resultando em transformações reais devido à *Data Science* e *Machine Learning* e o termo análise de negócios está se tornando padrão para comunicar o ciclo de decisão direcionado à dados (BAESENS e*t al.*, 2016).

Para avaliar a relevância do tema para a engenharia de produção, pode-se partir do que disse Wing (2006), que o pensamento computacional é uma técnica fundamental para qualquer área, não apenas para cientistas da computação.

Tedre e Denning (2016) advogam que a visão de aprendizado de programação leva à outras habilidades cognitivas de ordem superior relacionadas. E vai ao encontro do pensamento de Lasi (2013), no qual entende que em diversas áreas da engenharia de produção tem-se o objetivo permanente de adquirirem-se conhecimentos.

1.1 Problematização da Pesquisa

De acordo com Botvinik-Nezer *et al*. (2020), os fluxos de trabalho de análise de dados em vários campos científicos tornaram-se cada vez mais complexos e flexíveis. Chen, Argentinis e Weber (2016) vão além, mostrando que os analistas de dados estão sob pressão para inovar cada vez mais rápido. Já que os volumes de informação estão cada vez maiores, mas apenas uma fração desses dados está sendo integrada, entendida e analisada. O desafio existe nos grandes volumes de dados, na integração desses dados de centenas de fontes distintas e no entendimento de vários formatos.

De acordo com Agarwal e Dhar (2014), existem questionamentos e oportunidades criados pela disponibilidade de dados e grandes avanços em *Machine Learning* permitindo que possam ser realizadas análises mais eficientes e que oferecem *insights* mais precisos dos dados. Waller e Fawcett (2013) corroboram essa informação, dizendo que existem inúmeras oportunidades para pesquisa no ponto de intersecção entre *Data Science* e classificação e seleção de fornecedores.

Segundo Hallikas *et al.* (2005), para que as práticas de gerenciamento de fornecedores passam ser planejadas e executadas é necessário uma classificação e seleção apropriadas. Adicionalmente, segundo George *et al.* (2016) na última década as teorias de gerenciamento começaram a enfatizar os tamanhos dos efeitos e isso pode ser aplicado com a classificação e seleção de fornecedores por meio da *Data Science*.

Este projeto de pesquisa se apropria de uma destas lacunas que consiste em encontrar formas de se melhorar decisões sobre fornecedores utilizando o conceito de *Data Science*, seja tentando entender quais parcerias precisam de um suporte em investimento maior ou quais os procedimentos podem ser utilizados para melhorar esse relacionamento na cadeia de suprimentos. Schoenherr e Speier-Pero (2015) notaram um relacionamento interessante entre a classificação e seleção de fornecedores, *Data Science* e *Machine Learning*, e se refere ao fato de como lidar com essa grande quantidade de dados e como aproveitar e aplicar análises de dados. Esse desafio é um resultado direto da facilidade com que os dados podem ser coletados via tecnologia de informação moderna, gerando volumes sem precedentes, variedade e velocidade de dados.

Utilizando este conjunto de ferramentas espera-se que o processo de *Data Science* permita que os conhecimentos sejam encontrados e possam ser úteis na tomada de decisão quanto aos fornecedores.

Mokadem (2017) analisou que as organizações estão sob constante pressão para serem competitivas nos seus mercados escolhidos. As condições de mercado existentes desafiam as empresas à fortalecerem e manterem suas capacidades de competir no mercado e terem um controle da sua cadeia de suprimentos, em especial dos fornecedores.

Segundo Restrepo e Villegas (2019), é fácil notar que com a importância crescente dos processos de compra, as decisões tomadas pertinentes à esse processo se tornaram mais relevantes porque as organizações estão mais dependentes de fornecedores e as consequências diretas ou indiretas de decisões errôneas são mais severas.

Então, ao se utilizar *Data Science* e *Machine Learning* tentando-se resolver um problema prático, como é o caso do problema de classificação e seleção de fornecedores, Karpatne *et al.* (2017) entendem que uma importante decisão quanto aos modelos de aprendizagem de *Data Science* está na escolha da família de modelos utilizada para representar os relacionamentos entre os dados de entrada e as variáveis de resposta. Em aplicações científicas, se o domínio do conhecimento sugere uma forma particular de relacionamento entre entradas e saídas, cuidados devem ser tomados para assegurar que a mesma forma de relacionamento seja usada no modelo de *Data Science*.

Segundo Govindan e Sivakumar (2016) a escolha dos fornecedores se tornou uma decisão importante na cadeia de suprimentos e utilizar *Data Science* pode configurar uma tratativa interessante.

1.2 Objetivos da dissertação

Esta pesquisa se baseia em duas áreas de conhecimento, uma delas é *Data Science* e a outra Classificação e Seleção de fornecedores. Será utilizado um banco de dados com informações de fornecedores de uma empresa aeronáutica brasileira em Minas Gerais objetivando encontrar conhecimentos que possam auxiliar no processo decisório da gestão de classificação e seleção de fornecedores da organização. Para que ocorra a transformação dessas informações de fornecedores em conhecimento é necessária à utilização de ferramentas de *Data Science* e é nesse ponto que as duas áreas do conhecimento se ligam nesta pesquisa.

A empresa busca tentar aprimorar o relacionamento com seus fornecedores tentando entender as carências e como pode agir ativamente buscando melhorar e assegurar que seus parceiros consigam atender suas demandas, utilizando por exemplo auditorias melhor programadas.

Esse projeto constitui-se de uma proposta para analisar qual o melhor algoritmo de clusterização, pertencente ao *Machine Learning*, para se classificar o conjunto de fornecedores de um banco de dados e otimizar uma métrica para a seleção destes. Basicamente, a pesquisa atua em engenharia de produção tentando resolver um problema de classificação e seleção de fornecedores possibilitando meios de auxiliar nas tomadas de decisão. Rodrigues *et al.* (2017) entendem que a clusterização, ou o problema da clusterização, pode ser informalmente definido como o desafio de separar dados em grupos. Objetos que estão no mesmo grupo geralmente são mais similares em comparação com aqueles em outros.

Desta forma, esta dissertação tem como objetivo geral utilizar a técnica de *Machine Learning* na classificação e seleção de fornecedores em uma empresa do setor aeronáutico. Esta pesquisa possui dois objetivos específicos. O primeiro é demonstrar o funcionamento e o comportamento dos algoritmos clássicos de clusterização em uma base de dados real. Já o segundo objetivo específico desta pesquisa é analisar diversos algoritmos de clusterização em busca do mais apropriado para a base de dados dos fornecedores criando um *framework* que pode ser seguido para realização desse tipo de análise em outros tipos de projetos.

1.3 Estrutura da dissertação

Esta dissertação aborda a análise de dados utilizando clusterização buscando classificar fornecedores de uma empresa privada. Dessa forma, essa pesquisa de mestrado é

composta por 7 (sete) capítulos. No Capítulo 1 é realizada uma breve introdução sobre o tema abordado nesta dissertação. Nele encontram-se a contextualização, justificativas e os objetivos do trabalho. O Capítulo 2 se trata de uma breve análise da literatura sobre o estado da arte de clusterização na literatura acadêmica e uma análise de um mapa de co-ocorrências sobre o tema, que explica pontos que serão tratados mais à frente, e que é fruto de uma análise bibliométrica. Em seguida, o Capítulo 3 traz o embasamento teórico para os temas de *Data Science*, *Machine Learning* e Classificação e Seleção de fornecedores. Dentro desses temas, serão tratados os conceitos e o relacionamento entre eles e como se estruturam na *Data Science*, além de um referencial para o R, linguagem computacional utilizada nos procedimentos práticos da pesquisa. O Capítulo 4 apresenta com detalhes a metodologia utilizada – o método CRISP-DM utilizado no capítulo seguinte deste trabalho, em conjunto com os conceitos abordados no capítulo anterior. No Capítulo 5, encontra-se a aplicação do método proposto no Capítulo 4 por meio dos dados obtidos de uma empresa do setor privado, bem como a aplicação dos conceitos do Capítulo 3. Em seguida, o Capítulo 6 finaliza este trabalho com o desenvolvimento de um *framework* e sua exibição, além de sugestões para trabalhos futuros e conclusões. Por fim encontram-se as referências usadas nessa dissertação.

2 ANÁLISE BIBLIOMÉTRICA

Este capítulo tem como objetivo utilizar análise bibliométrica para identificar o estado da arte da literatura acadêmica sobre clusterização e como se comportam os estudos sobre as suas validações. Garfield, Sher e Torpie (1964) entendem a análise bibliométrica como sendo uma aproximação quantitativa para compreender estudos publicados.

Segundo Valdez, Pickett e Goodson (2018) a análise bibliométrica trata-se de uma investigação sobre determinado assunto utilizando-se do poderio computacional para auxiliar na extração de temas de grandes quantidades de dados. Seus benefícios consistem na diminuição da subjetividade do pesquisador utilizando-se de algoritmos matemáticos e melhorando a eficiência analítica em comparação com os métodos de análise convencionais.

2.1 Metodologia da análise bibliométrica

É interessante definir uma metodologia para a análise bibliométrica e o método de Siqueira (2019) apresenta um modelo para sua efetiva realização. A metodologia proposta na Figura 2.1 foi desenvolvida por meio de um estudo de literaturas sobre o assunto e se trata de uma efetiva aplicação de bibliometria por meio de etapas denominadas definição, extração, conversão, transformação, análise e divulgação que acabam gerando um ciclo em busca de atualizações sobre qualquer área do conhecimento.

Etapas de análise bibliométrica:

- Definição: Etapa na qual é selecionada uma base de dados de periódicos para a pesquisa e sua definição, como as palavras-chave ou conjunto de operações lógicas que serão utilizados. Também é nessa etapa que acontece a realização dessa pesquisa na base ou bases de dados selecionadas;
- Extração: Etapa que abrange a escolha da ferramenta para a análise e a obtenção dos dados da pesquisa em um formato que seja aceito pela ferramenta selecionada, sendo os formatos mais comuns Mendeley, RefWorks, CSV, BibTex ou *plain text*;
- Conversão: Etapa de envio dos dados para a plataforma, programa ou ambiente de desenvolvimento utilizado para a análise e conversão de dados para o formato utilizado pelo meio de análise como bibliotecas da linguagem R ou *Python*;
- Transformação: Etapa de refinamento dos dados utilizando métodos computacionais. É nessa etapa que são removidos dados faltantes ou duplicados que prejudicam o resultado;

- Análise: realização da análise bibliométrica da pesquisa proposta;
- Divulgação: organização dos resultados obtidos da análise e eventual divulgação, sejam em meios acadêmicos ou de qualquer tipo.

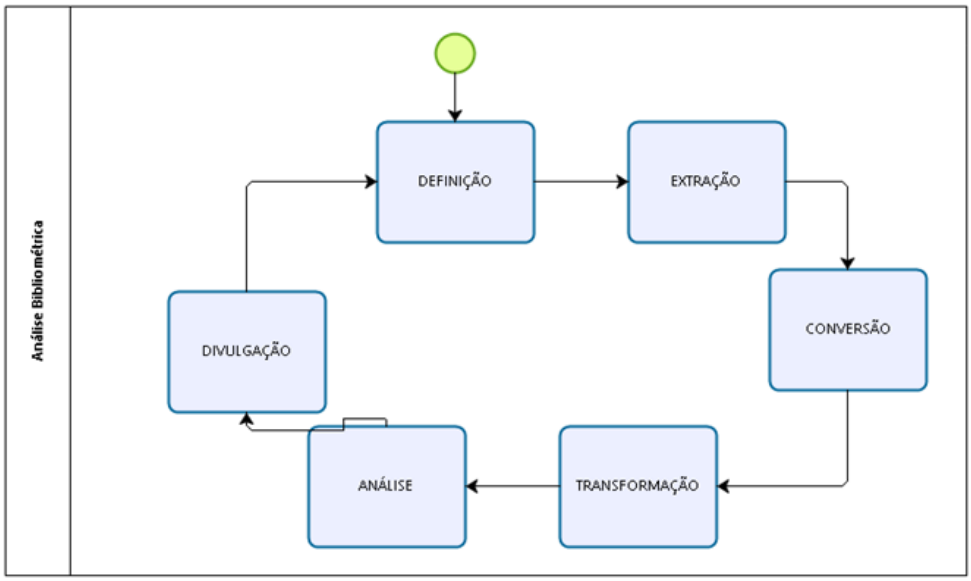

Figura 2.1: Fluxograma de Metodologia para Bibliometria
Fonte: Adaptado de Siqueira (2019)

2.1.1 Definição

Para a etapa de definição, primeiramente se faz necessário definir qual a base de dados dos artigos que vão fazer parte do acervo de dados da análise bibliométrica. Por base de dados considera-se ferramenta de pesquisa unificadora que permite adquirir, analisar e disseminar informações acadêmicas em tempo hábil.

Martín-Martín (2018) realizou um estudo comparando *Web of Science*, Scopus e *Google Scholar* chegando a conclusões interessantes de que apesar do *Google Scholar* ter uma quantidade maior de artigos para vários testes realizados ainda não existem métodos confiáveis e escaláveis para extração de dados e esta base de dados também apresenta uma correlação mais fraca para a área de engenharia, deixando as duas outras opções.

Além da escolha da base é importante também definir quais palavras-chaves são interessantes e fazem mais sentido para a pesquisa proposta, porque, segundo Madani (2016), para extrair os *papers* corretos da base de dados é importante aplicar palavras-chaves que se referem ao contexto estudado.

Primeiramente foi definido que as bases exploradas seriam *Web of Science* e Scopus, e uma busca preliminar foi realizada em ambas as ferramentas para definir qual opção escolher, utilizando-se apenas "*clustering algorithms"* como termo de pesquisa, resultando no extenso volume de 32693 artigos na Scopus e 3998 artigos na *Web of Science*, servindo como um indicador de interesse por parte da comunidade científica, sendo uma busca no texto inteiro.

Depois foi realizada uma nova pesquisa mais orientada pela junção de "*clustering algorithms"* + "*validation"* como termos de busca gerando um total de 814 documentos na Scopus e 191 na WoS, pertencentes apenas à artigos de revistas e congressos científicos.

O resultado da etapa de definição selecionou os termos para a busca de dados acadêmicos (conhecidos como metadados) e o volume mais extenso de artigos obtidos na Scopus em comparação com a WoS serviu como parâmetro para definição de qual base de dados utilizar culminando com a escolha da Scopus para realização dessa análise, já que a ferramenta utilizada não permitia que duas bases distintas fossem utilizadas simultaneamente.

2.1.2 Extração

A etapa seguinte é a de extração e o software ***R*** foi definido como ferramenta para a análise e os metadados obtidos da base Scopus foram exportados em sua totalidade para o formato *bibtex* que é aceito pela ferramenta selecionada.

R é uma linguagem livre para computação estatística e de geração de gráficos através do *RStudio*, um ambiente de desenvolvimento integrado que reúne características e ferramentas de apoio ao desenvolvimento de *R* com o objetivo de agilizar e facilitar o processo (CORE TEAM, 2018). O *R*, fundamentalmente estatístico, apresenta pacotes de bibliotecas que expandem suas funcionalidades básicas, dentre elas, a *Bibliometrix* que permite a realização de análise bibliométrica para dados obtidos dos principais bancos de dados de periódicos (ARIA; CUCCURULLO, 2017).

2.1.3 Conversão e Transformação

O pacote *Bibliometrix* do *R* foi utilizado para a etapa de conversão e os dados obtidos nas etapas anteriores foram explorados em busca de dados duplicados na etapa de transformação gerando dados lapidados e preparados para a pesquisa e consequente etapa de análise.

2.1.4 Análise

As informações obtidas dessa análise bibliométrica merecem uma atenção especial por parte dessa pesquisa. Os resultados obtidos serão divididos em uma análise básica que contém as informações gerais dos artigos analisados seguida de uma inquirição das palavras-chave dos autores.

2.1.4.1 Análise Básica

O pacote *Bibliometrix* do *R* permite uma análise inferencial dos conteúdos dos artigos, mostrando que os 814 documentos trabalhados estão divididos em 546 fontes de Revistas Científicas, livros e conferências. Geralmente existe uma média de 24 citações por documentos e 28061 referências no total.

A pesquisa trabalhou apenas com artigos e livros ou *papers* de conferências, tendo 542 e 272 (de artigos e livros) documentos de cada tipo respectivamente, ou seja, praticamente o dobro de artigos de periódicos científicos. Nota-se também que existe uma grande pluralidade de temas de estudos ao analisarem-se as palavras-chave dos autores que correspondem ao número 6432 para a quantia de 814 documentos. No que tange os autores existem 2721 autores pesquisando clusterização e suas validações de 2016 até 2020.

O tema apresenta também uma particularidade interessante, as pesquisas dessa área tendem a ser cooperativas tendo apenas 19 artigos desenvolvidos por pesquisadores individualmente (apenas 2,3% do total de artigos) e em média têm-se 3 autores por documento.

A produção científica anual sobre o tema apresenta um caráter bem constante, 149 documentos em 2016, 147 em 2017, 175 em 2018, 216 em 2019 e 127 após essa data até o momento da realização da análise bibliométrica.

2.1.4.2 Análise das palavras-chave dos autores

As revistas científicas e *papers* costumam utilizar-se de dois tipos de palavras-chave para organizarem seu acervo, o da própria revista e outro das selecionadas pelos próprios autores. As que foram escolhidas pela revista podem se utilizar, por exemplo, do termo "*man*" identificando que se trataria de um artigo para o público masculino. Em muitos casos são termos que não agregam valor a essa pesquisa em específico.

Como um dos objetivos de se realizar uma análise bibliométrica era entender o que efetivamente é considerado pela comunidade científica como sendo clusterização faz sentido utilizar as palavras-chave dos autores já que são os termos escolhidos pelos próprios pesquisadores para identificar suas pesquisas. Optou-se por trabalhar com as 50 palavras-chave mais usadas, mesmo que a relevância das últimas seja baixa, para que o resultado fosse o mais completo possível.

Em seguida, foi desenvolvido um mapa de co-ocorrências utilizando as palavras-chave, esse é um mapa que exibe o relacionamento de todas essas palavras (que são representadas pelos vértices) e também destaca as densidades delas, basicamente, quanto maior o tamanho de um vértice maior a relevância dessa palavra para o assunto pesquisado. Esse mapa se encontra na Figura 2.2

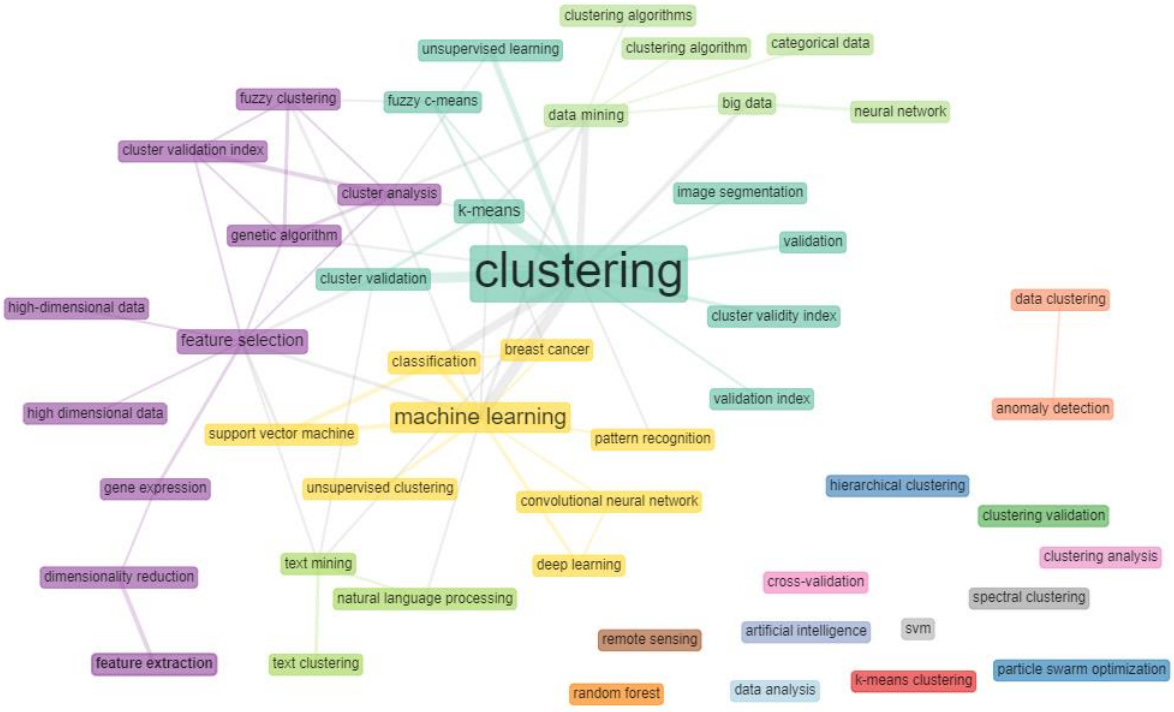

Figura 2.2: Mapa de co-ocorrência de palavras-chave dos autores

Fonte: Base de Dados da Scopus

As palavras-chave são os vértices e as ligações entre os vértices indicam uma relação entre os tópicos, como já foi dito. O tamanho dos vértices mostra quais são as palavras-chave mais utilizadas e o espaçamento das ligações, a força do relacionamento, quanto mais afastadas menos utilizadas pelos autores essas palavras são.

A própria análise bibliométrica sobre clusterização trata-se de uma clusterização propriamente dita, já que o mapa de co-ocorrências de palavras chaves criou *clusters* para as palavras com mais relação entre si que são definidos pela sua cor.

É interessante observar que as duas palavras-chave que representam os maiores vértices correspondem a *clustering* e *machine learning*, ou seja, já se nota a relação existente entre os algoritmos de aprendizagem de máquina e a clusterização, já que o segundo termo corresponde a uma forma de atuar sobre um problema específico quando se consideram estudos sobre aprendizagem de máquina, ambos os termos serão estudados em detalhes no Capítulo 3 desta dissertação.

O agrupamento que engloba *clustering* declara a relação que apresenta com *unsupervised learning* que é uma forma de se catalogar aprendizagem de máquina e reitera que a clusterização trabalha com dados que não apresentam categorias anteriores a criação dos *clusters*. Novamente, são termos que serão mais bem trabalhados nessa pesquisa.

Esse *cluster* de cor verde água, que é o mais alinhado com a pesquisa proposta, trabalha com índices de validação de clusterização que correspondem a métodos para se validar a criação desses grupos por meio dos algoritmos selecionados e são alguns dos tópicos chave do Capítulo 3.

O *cluster* de cor roxa mostra que existe um interesse dos trabalhos no campo da medicina e o *cluster* que apresenta *data mining* e *big data* mostram uma relação com mineração de dados em grandes volumes de dados.

Novamente dentro do *cluster* de cor roxa tem-se *dimensionality reduction* que é um procedimento que se faz necessário ao realizar clusterização com dados que apresentam muitas dimensões. Também pertencendo a esse *cluster* estão os dados conhecidos como *high dimensional data*, ou seja, que contêm muitas colunas de informação.

Os *clusters* mostram alguns algoritmos representativos como "*fuzzy c-means*", "*k-means*", algoritmos hierárquicos, dentre outros, mostrando que as pesquisas se interessam por alguns em específico. Vários serão testados futuramente nessa pesquisa.

Ela também mostra alguns aquecimentos como redes neurais e inteligência artificial, que não fazem parte do escopo desse estudo. Todavia de maneira geral, nota-se que o resultado obtido nesse pré-estudo se encaixa bastante com o direcionamento que essa dissertação de mestrado segue demonstrando que existe um alinhamento com o que os pesquisadores vêm trabalhando atualmente.

Existem vários outros termos utilizados, mas consistem em sua maioridade por sinônimos de termos mais comuns, formas de se escrever sobre a mesma coisa ou tópicos que se encaixam nos tópicos apresentados.

De forma simples e concreta consegue-se ter uma visão mais clara de onde se encaixa a clusterização dentro do universo de *machine learning* e esses conceitos serão mais bem definidos no decorrer desse trabalho.

E também pode-se perceber que a pesquisa se encaixa nos *clusters* de cor roxa, azul claro e amarelo, que são os mais significativos para o assunto.

3 REVISÃO BIBLIOGRÁFICA

Neste capítulo serão apresentados os conceitos principais para a pesquisa, consistindo de análise baseada em dados, *Machine Learning*, *Data Science*, clusterização e seus algoritmos, redimensionamento de dados, classificação e seleção de fornecedores e a linguagem R.

Ressalta-se que o embasamento científico presente nesse capítulo constitui a base na qual foi construída essa dissertação e que suporta os resultados e conclusões obtidos com esse trabalho e nota-se um relacionamento forte com o mapa de co-ocorrência de palavras-chave dos autores apresentado no Capítulo 2.

A fundamentação teórica abrangeu, principalmente, periódicos nacionais e internacionais, artigos de congressos, livros, teses e dissertações com temas correlatos.

3.1 Análise baseada em dados - *Data Driven Analytics*

Provost e Fawcett (2013) consideram a tomada de decisão usando dados como sendo a prática de se basear decisões em análises de dados ao invés de puramente em intuição. Já Carillo *et al.* (2019) enfatizam a necessidade de encorajar atitudes positivas com relação à análise de dados para desenvolver futuros tomadores de decisão mais preparados.

Segundo Ballou, Heitger e Stoel (2018) a demanda crescente por informações em tempo real requer desenvolvimento de análise de dados e a habilidade de usar grandes quantidades de dado como uma vantagem. Mehryar, Rostamizadeh e Talwalkar (2018) corroboram dizendo que a quantidade e qualidade dos dados são importantes para uma análise satisfatória.

Para Wang, Kung e Byrd (2018) a utilização de ferramentas de análise de dados em grandes quantidades de dados geram informativos detalhados e identificam tendências gerando pensamento criativo e acelerando novas ideias de negócio. E essa geração de novas ideias não é só necessária para inovação organizacional como também pode levar à mudanças nos processos produtivos que vão aumentar produtividade e construir vantagens competitivas.

Pensando nas organizações, segundo Rose (2016), uma corporação orientada a dados apresenta seus analistas de dados com três grandes áreas de responsabilidade:

- Coletar, acessar e reportar dados: Processar dados brutos em algo que todo mundo consiga entender;
- Realizar bons questionamentos;

- Tornar os dados demandáveis: É responsabilidade dos membros de equipe divulgar o que foi aprendido com os dados e entender como pode ser aplicado em prol da organização.

3.2 Ciência dos Dados - *Data Science*

Segundo Schutt e O'Neil (2013) *Data Science* compreende o estudo que busca transformar informações em novas formas de valor e vai ao encontro a Cao e Yu (2018) que baseiam seu uso em entender alguma área de conhecimento por meio do gerenciamento dos dados.

Diez-Olivan *et al.* (2019) apontam que a crescente quantidade de dados nas indústrias motiva a fusão dos mesmos com métodos de aprendizagem de máquina para atingir necessidades e objetivos industriais específicos culminando com o que conhecemos como *Data Science*. Segundo Amirian, Lang e Loggerenberg (2017) esse aumento significativo na quantidade de dados disponíveis permite reduzir custos e complexidade de testes, assim como foco e qualidade dos dados são trocados por quantidade e variedade.

Para Provost e Fawcett (2013) existe uma confusão sobre como definir exatamente *Data Science* e existem boas razões para não se apontar exatamente o que representa o conceito, a mais discutida é que está intrinsecamente inter-relacionada com outras áreas do conhecimento de importância crescente como mineração de dados e *Machine Learning*.

Segundo Schutt e O'Neil, (2013) existe a modelagem estatística proveniente da estatística e os algoritmos de aprendizagem de máquina provenientes da ciência da computação. Certos métodos e técnicas são considerados partes de ambos e não é útil para a *Data Science* discutir de que setor veio a técnica.

Apesar de estar conectada mais fortemente com áreas como de sistemas de informação e ciência da computação, muitas técnicas diferentes (incluindo técnicas não-matemáticas) são necessárias permitindo que se relacione com qualquer área do conhecimento (JAMES *et al.*, 2017).

Para Provost e Fawcett (2013) os conceitos fundamentais da *Data Science* são retirados de muitos campos de estudos na área de análise de dados e são enumerados a seguir:

- Extração de conhecimento útil de dados para resolver problemas de negócios que pode ser tratado por um processo com estágios razoavelmente bem definidos;
- Tecnologia da informação que pode ser usada para encontrar atributos informativos das entidades de interesse de grandes quantidades de dados;

- Se uma análise profunda de um conjunto de dados acontecer, será encontrado algo útil;
- Formular a análise de dados e avaliar os resultados envolve pensar cuidadosamente sobre os conceitos que serão utilizados.

Basicamente, *Data Science* serve para transformar informação em novas formas de valor buscando entender algum domínio gerenciando dados e descobrindo conhecimento (CAO e YU, 2018; SCHUTT e O'NEIL, 2013). De forma geral, *Data Science* também pode ser vista como a área que utiliza-se de métodos computacionais para solucionar problemas reais e atua entre ciências da computação, estatística e demais áreas preocupadas com melhorias e auxílio à tomada de decisão utilizando dados (JORDAN e MITCHELL, 2015)

Como todo processo, *Data Science* gera produtos e as saídas são produtos de dados, conhecimento, inteligência e sabedoria (CAO; YU, 2018) que foram ordenados por Schutt e O'Neil (2013):

- Análise exploratória dos dados;
- *Dashboards* e métricas;
- Encontrar entendimento de negócios;
- Tomada de decisão baseada em dados;
- Engenharia de dados e *Big Data*;
- Obter os próprios dados;
- Escrita de patentes;
- Trabalhos de investigação;
- Prever comportamentos futuros;
- Escrever artigos acadêmicos e apresentações;
- Programação (em diversas linguagens);
- Otimizações;
- Fazer inferências com os dados;
- Construir produtos de dados;
- Encontrar formas de processar e analisar dados;
- Desenvolver e analisar experimentos;
- Encontrar correlação de dados.

3.3 Aprendizagem de Máquina - *Machine Learning*

Segundo Alpaydin (2010), *Machine Learning* representa o ato de programar computadores para melhorar o desempenho de determinado sistema usando dados ou experiência passada para essa finalidade. Sua aplicação se dá quando a *expertise* humana não existe, quando são incapazes de explicar sua *expertise*, se a solução muda com o tempo ou se precisa ser adaptada para casos particulares (ALPAYDIN, 2014).

Mehryar, Rostamizadeh e Talwalkar (2012, 2018) corroboram dizendo que *Machine Learning* são métodos computacionais que além de melhorar o desempenho de um sistema, fazem predições com precisão. E as experiências passadas disponíveis aos estudos geralmente constituem-se de dados eletrônicos coletados ou produzidos.

Alpaydin (2016) explica que, no *machine learning*, estatística e computação tem papéis claros nos estudos, o da estatística é encontrar inferências de amostras e o da computação é utilizar algoritmos eficientes para resolver algum problema, representar ou avaliar um modelo e uma utilização eficiente dos dois campos de aprendizagem gera resultados atrativos.

Os algoritmos de *Machine Learning* se dividem em dois cenários principais:

- Cenários supervisionados: São direcionados a um objetivo específico representado pelo valor de uma característica particular, que apresentam um conjunto de dados rotulados com as respostas corretas, onde o objetivo é aprender com um mapeamento das entradas e saídas e encontrar os valores corretos providos pelo supervisor (AGGARWAL, 2015; ALPAYDIN, 2016; GRUS, 2015);
- Cenários não-supervisionados tem a ausência de uma base de dados especial para "ensinar" o modelo sobre a noção de um agrupamento apropriado, com suposições sobre as propriedades estruturais dos dados e não apresentam um supervisor, apenas dados de entrada (AGGARWAL, 2015; ALPAYDIN, 2016; JORDAN; MITCHELL, 2015).

  Eles são utilizados para resolver problemas de aprendizagem segundo Jordan e Mitchell (2015).

Os problemas mais comuns são:

- Classificação: o problema da classificação é aprender a estrutura dos conjunto de dados de exemplos, já particionados em grupos, chamados de categorias ou classes (AGGARWAL, 2015);

- Clusterização: Para Mehryar, Rostamizadeh e Talwalkar (2018), é o problema de particionar itens em regiões homogêneas sem a figura de um supervisor;
- Associação: Segundo Cao e Yu (2018), quando se fala em aprendizagem de associação, geralmente trata-se de relações de dependência (incluindo relacionamentos escondidos), associação, correlação e co-ocorrência.

Para finalizar, as melhorias nas áreas de *Machine Learning* estão focadas em superar desafios estatísticos e melhorar segurança dos dados, segundo Yang *et al.* (2019).

3.4 Clusterização - *Clustering*

Como explicou Deng *et al.* (2016) a clusterização é uma técnica fundamental da ciência de dados porque pode ser usada para segmentar bases de dados. Rodrigues *et al.* (2017) entende que a clusterização, ou o problema da clusterização, pode ser informalmente definido como o desafio de separar dados em grupos. Objetos que estão no mesmo grupo geralmente são mais similares em comparação com aqueles em outros. Essa definição assume que exista alguma medida de qualidade responsável por controlar similaridade ou dissimilaridade entre dados.

Para De Morsier *et al.* (2015) o problema de se separar conjuntos de dado em diferentes grupos é bastante estudado em vários campos onde a informação semântica dos dados não está disponível. Portanto, clusterização é uma técnica de aprendizagem não supervisionada que apresenta diferentes visões para a estrutura inerente de um conjunto de dados dividindo-o em várias quantidades de sobreposições ou grupos disjuntos (JAYARAM REDDY *et al.*, 2018).

Kassambara (2017) segue a mesma linha e aponta que na literatura a clusterização é referida como “reconhecimento de padrões” ou “aprendizagem de máquina não supervisionada” e o “não supervisionada” é devido ao fato de não ser guiada por ideias a priori de quais variáveis ou amostras pertencem a cada cluster e “aprendizagem” porque o algoritmo de máquina “aprende” como clusterizar.

O resultado de uma clusterização é chamado de *cluster*. Um *cluster* é uma coleção de objetos similares que são diferentes de objetos de outros *clusters* (OMRANI, SHAFAAT e EMROUZNEJAD, 2018).

Segundo Liu *et al.* (2016), a clusterização pode ser alcançada por meio de vários algoritmos que diferem significantemente nas suas noções sobre o que constitui um *cluster* e como encontrar um *cluster* de forma eficiente.

De acordo com Rodriguez *et al.* (2019) essa ferramenta para análise de dados é fortemente relacionada com a missão de criar um modelo dos dados, que é definir um conjunto de propriedades simplificado que possa prover explicação intuitiva sobre aspectos relevantes do conjunto de dados. Métodos de clusterização são geralmente mais exigentes do que aproximações supervisionadas, porém proporcionam mais conhecimento sobre dados complexos.

As aplicações de clusterização na literatura estão bem atualizadas como o estudo de Chen *et al.* (2020) que fez uma análise de casos virais do novo corona vírus em Wuhan na China no ano de 2019 chegando em conclusões de que a doença é mais propensa a infectar homens com doenças consideradas como comorbidades e que nesses casos pode resultar em doenças respiratórias severas e até mesmo fatais. Também existe o estudo de Lin, Liu e Peng (2017) que procurou relações entre o preço da terra em Taichung *City* no Taiwan e avanços na infraestrutura e desenvolvimento da cidade.

3.5 Algoritmos de Clusterização

Neste subtítulo serão definidos os algoritmos de clusterização utilizados no desenvolvimento dessa pesquisa, foram utilizados algoritmos mais clássicos para a composição dessa lista. Os considerados clássicos são os mais difundidos na literatura e amplamente explorados e utilizados para clusterização.

Para Nivio (2007), algoritmo é uma sequência finita de ações executáveis que objetivam a obtenção de uma solução para determinado tipo de problema, são como uma receita de bolo que pode ser seguida. Construir modelos e trabalhar com dados não é de valor neutro. O pesquisador escolhe os problemas que vai tentar solucionar e faz suposições naqueles modelos e escolhe as métricas e os algoritmos que vai utilizar (SCHUTT e O’NEIL, 2013).

3.5.1 Clusterização de Particionamento

Segundo Kassambara (2017), clusterização de particionamento corresponde a métodos usados para classificar objetivos em múltiplos grupos baseados na similaridade entre eles, dentro de um conjunto de dados.

Eles também são conhecidos como algoritmos *representative-based* (baseados em representante) e são os mais simples de todos os algoritmos de clusterização porque eles dependem diretamente de noções intuitivas de distância (ou similaridade) para clusterizar pontos. Os *clusters* são criados de uma vez e relacionamentos hierárquicos não existem entre diferentes *clusters* (AGGARWAL, 2015).

3.5.1.1 *K-Means*

Para Deng *et al.* (2016) o método baseado em *k-means* primeiro conduz um agrupamento em todas as amostras várias vezes e, em seguida, usa os centros do *cluster* como as amostras de referência.

Kassambara (2017) sumarizou o algoritmo K-*means* em cinco etapas:

1. Especificar a quantidade de *clusters* (K) a ser criada (pelo analista);
2. Selecionar aleatoriamente k objetos do conjunto de dados como os centros de *cluster* ou médias iniciais deles;
3. Atribuir cada observação ao seu centroide mais próximo baseado na distância euclidiana entre o objeto e o centroide;
4. Para cada um dos k *clusters* atualize o centroide do *cluster* calculando o novo valor da média de todos os pontos de dados no *cluster*. O centroide do K$_{ésimo}$ *cluster* é um vetor de tamanho p contendo as médias de todas as variáveis para as observações no K$_{ésimo}$ *cluster* e p é a quantidade de variáveis;
5. Iterativamente reduza a soma de quadrados total. Isso é, iterativamente repita os passos 3 e 4 até que as atribuições do *cluster* parem de mudar ou a quantidade máxima de iterações seja alcançada.

Segundo Hartigan e Wong (1979), o objetivo do algoritmo *k-means* é dividir M pontos em N dimensões e K *clusters* de forma com que a soma dos quadrados totais seja minimizada. Nunes (2016) ponderou sobre a convergência, mostrando que o *k-means* converge para uma solução ótima parcial de um problema em um número finito de iterações e Chouhan e Purohit

(2018) concordam que o *k-means* é um método de clusterização de fácil convergência mas sofre com o problema da inicialização de centroides.

Complementando sobre o problema, Capó, Pérez e Lozano (2017), entendem que apesar da dependência da configuração inicial e da grande quantidade de distância computacional que é requerida para a convergência, o *k-means* permanece como um dos métodos de clusterização mais populares até para grandes quantidades de dados.

E essa popularidade se deve ao fato de que pessoas preferem um algoritmo cujas limitações são conhecidas do que outros melhores, mas cujas limitações não são bem conhecidas e também porque o *k-means* pode ser utilizado em conjunto com outros algoritmos (FRÄNTI, 2000).

3.5.1.2 *K-Medoids: Partitioning around medoids* (PAM)

O PAM é um algoritmo de clusterização com o método de particionamento que produz um conjunto de *clusters* como uma saída e objetiva minimizar a dissimilaridade média dos objetos para os objetos centrais de cada *cluster* (LI, WANG e HE, 2017).

Basicamente, de acordo com Kaufman e Rousseeuw (1990), o *Partitioning around medoids* (PAM) tem como objeto representante do *cluster* um medóide , que é definido como o objeto de um *cluster* no qual a dissimilaridade média para todos os objetos do *cluster* é mínima (o objeto mais próximo aos outros do *cluster*). Como o objetivo é encontrar os tais k objetos, ele recebe o nome de método *k-medoid*.

Kassambara (2017) descreveu as etapas do algoritmo como se segue:

1. Selecionar k objetos para se tornarem medóides, ou se já foram selecionados com antecedência usá-los como medóides;
2. Calcular a matriz de dissimilaridade se não for oferecida previamente;
3. Atribuir cada objeto ao seu medóide mais próximo;
4. Para cada cluster, investigar se quaisquer dos objetos diminui o coeficiente médio de dissimilaridade. Se algum deles mudá-lo, selecionar a entidade que diminuiu mais o coeficiente como o novo medóide do *cluster*;
5. Se pelo menos um medóide mudou, retorne para o Passo 3. Se não, encerre o algoritmo.

Segundo Zerzucha e Walczak (2016), para calcular a matriz de dissimilaridade, uma matriz N x N contendo a distância entre N objetos, pode-se utilizar algumas métricas, como as

distâncias euclidianas que são a raiz da soma de quadrados das diferenças entre elas ou a distância Manhattan que é a soma das distâncias absolutas.

O PAM pode ter um bom desempenho em aplicações práticas, mas consome muito tempo de processamento devido a sua estrutura. Entretanto, ele tem uma boa atuação com uma pequena quantidade de dados (KHATAMI *et al.*, 2017).

Yu *et al.* (2018) complementam que apesar de ser utilizada em muitas aplicações práticas, a clusterização por *k-medoids* sofre de algumas desvantagens. Pode ficar presa na busca pelo local ótimo, consome bastante tempo e é sensitiva à inicialização e a *outliers*, que são causados pela seleção aleatória dos k medóides e também pela troca de todas os pares de medóides e não-medóides.

3.5.1.3 *Clustering Large Applications* (CLARA)

Em várias aplicações, o tamanho das bases de dados pode ser muito grande. Nesses casos, os dados podem não ser armazenados na memória principal, o que impõe um desafio significativo para os algoritmos (SCHUTT e O'NEIL, 2013).

De acordo com Aboubi, Drias e Kamel (2016) o CLARA produz relativamente boas soluções com qualidade em um tempo computacional razoável para grandes conjuntos de dados, mas é menos eficiente se utilizar amostras e não o conjunto inteiro de dados.

Isso se deve ao que discutiram Song, Lee e Han (2017) que o CLARA basicamente aplica o PAM em algumas amostras para encontrar medoides, já que os resultados são bastante afetados pela amostra no segundo algoritmo, então o CLARA repete esse procedimento múltiplas vezes selecionando o melhor conjunto como o resultado final.

O procedimento do algoritmo CLARA foi descrito por Kassambara (2017):

1. Separar aleatoriamente os conjuntos de dados em múltiplos subconjuntos com tamanhos fixos;
2. Computar o algoritmo PAM em cada subconjunto;
3. Calcular a média (ou a soma) das dissimilaridades das observações para seu medoide mais próximo (isso é usado como uma medida da eficiência do algoritmo);
4. Reter o subconjunto de dados para o qual a média (ou soma) é mínima. Então, uma análise mais completa é efetuada na partição final.

3.5.2 Clusterização Hierárquica

Para Kaufman e Rousseeuw (1990b) os algoritmos de clusterização hierárquica constroem uma série de partições de um conjunto de padrões na mesma execução. Basicamente, algoritmos hierárquicos produzem *clusters* hierárquicos aninhados. Um processo de clusterização hierárquica pode ser representado por uma estrutura em árvore, conhecida como dendrograma, que permite a inspeção de vários resultados obtidos pelo algoritmo depois de finalizado.

Os algoritmos hierárquicos costumam agrupar os dados com distâncias, mas o uso de funções de distância não é compulsório. Muitos algoritmos hierárquicos usam outros métodos de clusterização, como aqueles baseados em densidade ou grafos como uma sub-rotina para construir a hierarquia.

Existem dois tipos de algoritmos hierárquicos, dependendo da árvore hierárquica na qual os *clusters* são construídos:

1. *Bottom-up* (aglomerativos): Os pontos individuais são sucessivamente aglomerados em *clusters* de nível mais alto. A variação principal entre esses métodos é a escolha da função objetivo usada para decidir a união de *clusters*;
2. *Top-down* (divisivos): Essa abordagem é usada para particionar os dados em uma estrutura de árvore separando os nós mais pesados em quantidades menores (AGGARWAL, 2015).

3.5.2.1 *Agglomerative Nesting* (AGNES)

Segundo Zhao *et al.* (2018), o AGNES (*AGglomerative NESting*) é um algoritmo hierárquico. Primeiramente, objetos são entradas e cada uma constitui um *cluster* inicial por si só. Então, os dois *clusters* com a menor distância são continuamente unidos em um único até que o número atinja o valor k satisfazendo a condição de parada.

Os Passos para o AGNES foram adaptados da definição de Kassambara (2017):

1. Computar as informações de dissimilaridade entre cada par de objetos no conjunto de dados;
2. Usar função de ligação para agrupar objetos em uma árvore de *cluster* hierárquica, baseada na informação de distância utilizada no Passo 1. Objetos e *clusters* que estão próximos são unidos usando a função de ligação.

   As funções de ligação mais comuns são:

- Ligação Máxima ou completa: A distância entre dois *clusters* é definida como o valor máximo de todas as distâncias em pares entre os elementos no *cluster* 1 e no *cluster* 2;
- Ligação mínima ou individual: A distância entre dois *clusters* é definida como o menor valor de todas as distâncias em pares entre os elementos no *cluster* 1 e no *cluster* 2;
- Média ou ligação média: A distância entre dois *clusters* é definida como a distância média entre os elementos no *cluster* 1 e no *cluster* 2;
- Ligação Centroide: A distância entre dois *clusters* é definida como a distância entre o centroide do *cluster* 1 (um vetor médio de tamanho p variáveis) e o centroide do *cluster* 2;
- Método da Variância mínima de Ward: Minimiza a variância total interna do *cluster*. A cada Passo o par de *clusters* com distância mínima é unido.

3. Determinar onde cortar a árvore hierárquica em *clusters*. Isso cria uma partição dos dados.

Camargos e Do Carmo Nicoletti (2017), entendem que o AGNES é bastante afetado pela distância intra-*cluster*, entretanto o algoritmo é bastante robusto. E de acordo com Wang *et al.* (2020) quando comparado com amostras agregadas com o método convencional de centroides, o AGNES é mais acessível independente da seleção inicial de valores, e livre do formato de distribuição das amostras. E ele também pode agregar todas as amostras juntas.

3.5.2.2 *Divisive Analysis* (DIANA)

Segundo Orsi (2017), DIANA trabalha na direção oposta ao AGNES. Começa com todos os dados em um único grupo e então vai quebrando os grupos grandes em qualquer quantidade de *clusters* menores.

O procedimento do algoritmo foi enunciado por Patnaik, Bhuyan e Krishna Rao (2016):

1. O algoritmo DIANA segue a clusterização aglomerativa hierárquica até o *cluster* conter todos os objetos, então o *Divisive Analysis Clustering* segue uma aproximação *top-down* assumindo que o *cluster* individual tem nível L (0) = *n* e sequência numérica *m* = 0.

2. O par de *clusters* mais dissimilar no *cluster* atual é descoberto, sendo r e s no qual d [r, s] = min d [i, j], onde min representa os pares de *clusters* completos.
3. O número da sequência é incrementado de maneira que $m = m + 1$. O *cluster* é quebrado em *clusters* r e s para formar o próximo *cluster* a fazer o nível de clusterização: L (m_1) = d [r] e L (m_2) = d [s].
4. A matriz distância (D) é atualizada pela adição de linhas e colunas correspondendo aos *clusters* r e s. Se todos os objetos são *clusters* distintos então encerrar o algoritmo. Se não, retornar para o Passo 2.

3.5.2.3 *Heatmap*

Segundo Wilkinson e Friendly (2009) um *heatmap* de *cluster* é uma forma gráfica popular para visualizar dados de alta dimensionalidade. Nele, uma tabela de números é escalada e desenvolvida como uma matriz ladrilhada de células coloridas. As linhas e colunas da matriz são ordenadas para realçar padrões e são frequentemente acompanhadas por dendrogramas e colunas extras de anotações categóricas.

Adaptado de Kassambara (2017):

1. Clusterização hierárquica é feita nas colunas e linhas da matriz de dados;
2. As colunas e linhas são reordenadas de acordo com o resultado da clusterização hierárquica colocando resultados similares próximos uns aos outros;
3. Um esquema de cores é aplicado para a visualização e a matriz de dados é exibida.

Os clustergramas, ou *heatmaps*, são representantes de uma técnica que visualiza os dados diretamente sem necessidade de redução de dimensionalidade. Além disso, são fáceis de interpretar e são amplamente utilizados em visualizações de dados em publicações científicas (FERNANDEZ *et al.*, 2017).

Bujack *et al.* (2018) pontuou a importância do esquema de seleção de cores para o *heatmap*, que deve ter uma diferença entre as cores perceptíveis ao olho humano e que facilitem a compreensão dos resultados.

Para Khomtchouk, Hennessy e Wahlestedt (2017), o advento de *heatmaps* sofisticados e interativos e o surgimento de grandes volumes de dados alinhou com um interesse crescente da comunidade científica em examinar esses dados de maneira interativa.

Gu *et al.* (2018) aponta os problemas que existem na técnica. A visualização de um *heatmap* tem dificuldade em configurar de forma apropriada a ordenação de linhas para

descobrir padrões em vários *heatmaps* simultaneamente. A maioria das técnicas simplesmente ordena linhas baseadas na média das linhas de uma matriz normalizada.

3.5.3 *Soft Clustering*

Segundo Aggarwal (2015), a maioria dos algoritmos são *hard* (pesados), o que significa que cada dado é deterministicamente atribuído a um *cluster* particular. Modelos probabilísticos são algoritmos *soft* (leves), aos quais cada ponto pode ter uma atribuição não zerada de probabilidade para vários (tipicamente todos) os *clusters*.

De acordo com Roul (2018), *soft clustering* acontece quando cada dado tem um valor de filiação em mais de um *cluster*.

3.5.3.1 *Fuzzy Clustering*

Fuzzy Clustering é um método que permite que um objeto pertença a múltiplos *clusters* simultaneamente e é baseado em partições *fuzzy* da matemática *fuzzy* fazendo com que os dados sejam classificados como pertencendo a cada classe com diferentes graus de filiação. Isso coincide com o fato de que a maioria dos dados na realidade é incerta (NI *et al*., 2017).

Fuzzy é usado para reconhecimento de padrões e tarefas de clusterização (OMRANI, SHAFAAT e EMROUZNEJAD, 2018). Segundo Son e Thong (2017), os modelos de lógica *fuzzy* tentam modelar tomada de decisão e são capazes de prover resultados mais precisos baseados em incertezas e informações vagas.

Para Yang *et al.* (2019a), o propósito da *fuzzy clustering* é dividir os dados amostrais em alguns conjuntos *fuzzy*, onde a filiação parcial é permitida e os elementos podem pertencer a vários conjuntos com diferentes graus. Os algoritmos de *Fuzzy Clustering* geralmente tratam dados como componentes característicos com a mesma importância (YANG e NATALIANI, 2018).

Kassambara (2017) explica que o *fuzzy c-means* (FCM) é diferente do *k-means* e do *k-medoids* onde cada objeto é afetado exatamente por um *cluster*. Nele, pontos próximos ao centro de um *cluster* tem uma chance maior de estar no *cluster* do que pontos na borda de um *cluster*. A chance de um elemento pertencer a um *cluster* é um valor numérico entre 0 e 1 e o centroide de um *cluster* é calculado como a média de todos os pontos, com peso pelo seu grau de pertencer a um *cluster*.

Geralmente com a lógica *Fuzzy*, cada modelo é representado como “se antecedência, então consequência”. O antecedente é o mesmo para todos os modelos *fuzzy*, mas a consequência tem diferentes formas para cada modelo (SHOKOUHIFAR e JALALI, 2017).

3.5.4 Clusterização avançada

Os demais algoritmos se encontram nessa categoria.

3.5.4.1 *Hierarchical K-Means Clustering* (HKM)

O problema do *k-Means* é que ele seleciona a sementes iniciais aleatoriamente e fica difícil evitar escolher ruídos ou pontos muito próximos como sementes (QI *et al*., 2017).

De acordo com Nguyen *et al.* (2019), clusterização hierárquica auxilia o *Hierarchical K-Means* à definir os centroides iniciais para o *k-means*.

Kassambara (2017) sumariza o algoritmo a seguir:

1. Computar clusterização hierárquica e cortar a árvore em k *clusters*;
2. Computar o centro (média) de cada *cluster*;
3. Computar *k-means* usando o conjunto de centros de *clusters* (definido no Passo 2) como os centros de *clusters* iniciais.

Segundo Liao *et al.* (2017), o *hierarchical k-means* é muito eficiente em análises de dados de larga escala, todavia a velocidade e precisão do HKM ainda apresentam espaço para melhorias.

3.5.4.2 *Model-Based Clustering*

Segundo Mcnicholas (2016), *model-based clustering* se refere ao uso de (finitos) modelos de mistura para desempenhar clusterização. Fop e Murphy (2018) relatam que é um método estabelecido e bem popular no qual a clusterização é formulada em um *framework* de modelagem.

Modelos de finita mistura assumem que os dados surgem de um número finito de *clusters* homogêneos e são utilizados quando não se tem como identificar a qual população pertence cada elemento da amostra, sendo uma combinação de duas ou mais funções de densidade de probabilidade (MCPARLAND e GORMLEY, 2016)

A conexão entre *model-based clustering* e a modelagem das finitas misturas, de acordo com Zhu e Melnykov (2015), é estabelecida pelas médias da regra de decisão de Bayes, que atribui cada observação para classes com as probabilidades posteriores estimadas mais altas.

De acordo com Zhu (2018), o *model-based clustering* goza de interpretabilidade atraente e potenciais promissores.

A seleção de variável ou recurso é de particular importância em situações onde apenas um subconjunto das variáveis fornece informações para a clusterização. Isso permite a seleção de um modelo mais parcimonioso, produzindo estimativas mais eficientes, uma interpretação mais clara e, frequentemente, partições de *clusters* aprimoradas (Scrucca e Raftery, 2018).

Kassambara (2017) enuncia que cada componente k desse tipo de clusterização é modelado pela distribuição normal ou gaussiana que é caracterizada pelos parâmetros: Vetor das médias, matriz de covariância e uma probabilidade associada na mistura. Cada ponto tem uma probabilidade de pertencer a cada *cluster* de acordo com o algoritmo.

Os parâmetros do modelo podem ser estimados com um algoritmo chamado *Expectation-Maximization* (EM) iniciado por um *model-based clustering* hierárquico no qual cada *cluster* k é centrado no vetor das médias com densidade aumentada para pontos próximos à média e as características geométricas (forma, volume e orientação) de cada *cluster* são determinadas pela matriz de covariância.

3.6 *Principal Component Analysis* (PCA)

Nos anos recentes, armazenar e modelar dados multidimensionais se tornou bastante comum. Potenciais conjuntos de dados incluem diferentes atributos de potenciais clientes, múltiplas taxas cambiais de várias moedas por dia e até imagens vetoriais. Apesar desses dados geralmente se encontrarem em variedades não lineares, uma variedade linear ou uma combinação de variedades lineares pode frequentemente prover uma aproximação prática e adequadamente precisa (GUPTA e BARBU, 2018).

Os dados que são utilizados costumam ser multivariados e uma classe de visualizações conhecida como multidimensional tem sido utilizada para dar suporte ao entendimento das instâncias de relacionamento dos conjuntos de dados. Essas técnicas miram em diminuir a dimensionalidade de um espaço multidimensional mantendo a similaridade e o relacionamento entre instâncias de dados tanto quanto possível. Então, o espaço pode ser

visualizado em uma tela de computador, permitindo melhor entendimento dos relacionamentos e do espaço multidimensional (ELER *et al.*, 2015).

PCA consiste em projetar dados de entrada N-dimensionais em um subespaço k-dimensional linear que diminua o erro de reconstrução, que consiste na soma de quadrados L2-distâncias do dado original e o dado projeto. Então, o algoritmo PCA é definido pela solução matriz P* ortogonal do seguinte problema de minimização da Eq. (1) (MEHRYAR, ROSTAMIZADEH e TALWALKAR, 2012):

$$\min_{P \in P_k} \|PX - X\|_F^2 \tag{1}$$

De forma simplificada, Cui, Li e Zhang (2019) definem que o objetivo do Principal Component Analysis (PCA) é transformar um conjunto de variáveis correlacionadas em um conjunto de variáveis minimamente correlacionadas.

O PCA é utilizado nesse trabalho para redução de dimensionalidade para visualização dos gráficos.

3.7 Validação de Clusterização

Depois que um algoritmo de clusterização processa os dados e os particiona, um dos problemas chave é o posterior teste desses resultados, e essa adversidade pode ser chamada de validação de clusterização. A complicação consiste em configurar uma função de um índice de validação, rodar os algoritmos e encerrar o processo de busca da quantidade otimizada de *clusters* dinamicamente enquanto avalia os resultados (HALKIDI, BATISTAKIS e VAZIRGIANNIS, 2001).

A validação de *clusters* é frequentemente difícil com conjuntos de dados da vida real porque os problemas são definidos de uma maneira não supervisionada e não existe um critério de validação externo disponível para avaliar a clusterização, então critérios de validação internas têm de ser definidos para validar a qualidade da clusterização (AGGARWAL, 2015).

Segundo Hämäläinen, Jauhiainen e Kärkkäinen, (2017) os índices de validação medem quão bem os objetivos gerais da clusterização são atingidos. Objetivos que consistem em atingir alta similaridade de informação dentro do mesmo *cluster* e alta dissimilaridade entre *clusters*. Esses índices são considerados como medidas de separação internas de cluster (Intra) e de separação entre cluster (inter), sendo que valores baixos são melhores para o primeiro

caso e valores maiores para o segundo são melhores. Normalmente, uma divisão entre intra e inter é feita e o valor ótimo é o mínimo ou o máximo dependendo da ordem da divisão.

Segundo Liu *et al.* (2013), como um método não supervisionado a clusterização precisa de uma validação da qualidade de suas partições ou seria difícil de se comparar diferentes resultados. Dessa forma, os índices de validação têm um papel importante nas análises de clusterização e são usados para avaliar e verificar os resultados.

Para De Morsier *et al.* (2015), os índices de validação podem ser classificados em duas categorias: índices de validação externa e índices de validação interna. Os índices de validação externa focam em comparar um resultado de clusterização com um resultado pré-determinado que é utilizado como referência enquanto os índices de validação interna são geralmente aplicados para selecionar o melhor algoritmo de clusterização e a quantidade ótima de *clusters* sem nenhuma informação adicional.

Para Kassambara (2017), a validação de *clusters* consiste em medir a eficiência dos resultados da clusterização. Antes de aplicar qualquer algoritmo de clusterização a um conjunto de dados, a primeira coisa a se fazer é acessar a tendência de clusterização, ou seja, se a aplicação de clusterização é adequada para os dados. Se a resposta for sim, então quantos *clusters* serão usados. Depois, pode-se realizar o tipo de clusterização escolhido e finalmente aplicar uma série de índices para medir a eficiência dos resultados obtidos.

Para a validação interna, Liu *et al.* (2013), consideram que as validações geralmente se baseiam em dois critérios:

- Compacidade: Medida de quão bem estão relacionados os objetos em um *cluster*. Avalia a compacidade baseada na variância. Baixa variância indica melhor compacidade. Em adição, existem inúmeras medidas que estimam a compacidade do *cluster* baseada na distância (como máxima ou média de distância de pares e distância média baseada no centro);
- Separação: Mede quão distintos ou bem separado um *cluster* está dos outros *clusters*. Por exemplo, a distância de pares entre o centro de *clusters* ou distância mínima de pares entre objetos em diferentes *clusters* são bastante utilizadas como medidas de separação. Medidas baseadas em densidade também compõem os índices.

  Kassambara (2017) adiciona um terceiro critério:

- Conectividade: Corresponde a quais medidas os itens são colocados no mesmo *cluster* como seus vizinhos mais próximos no espaço dos dados. A conectividade tem um valor entre 0 e infinito e deve ser minimizada.

Segundo Patil e Baidari (2019), na literatura, muitos índices internos foram propostos para analisar os resultados e determinar a quantidade otimizada de *clusters*.

Seghier (2018) reitera a existência de uma grande quantidade de índices para validação da quantidade otimizada de *clusters* em um processo de validação de clusterização, provavelmente mais de 50 índices, e que foge de o escopo estudar cada um deles em detalhes.

3.7.1 Coeficiente *Silhouettte*

O coeficiente *Silhouette* é um índice de validação interna que mede quanto o resultado D é similar aos seus próprios *clusters* (compacidade) comparado com outros *clusters* (separação) (ROUL, 2018).

Ele mede quão bem uma observação é clusterizada e estima a distância média entre *clusters*. Ele mostra a medida de quão próximos cada ponto está dos *clusters* vizinhos.

Para cada observação i, o *Silhouette* tem largura s_i e é calculado como se segue:

1. Para cada observação i, calcula-se a média de dissimilaridade a_i, entre i e todos os pontos do *cluster* a que ele pertence;
2. Para todos os outros *clusters* C, aos quais i não pertence, calcule a dissimilaridade média d (i, C) de i para todas as observações de C. A menor dessas d (i, C) é definida como $b_i = \min_c d$ (i, C). Os valores de b_i podem ser vistos como a dissimilaridade entre o ponto e o *cluster* "vizinho", aquele que é mais próximo e que ele não pertence.
3. Finalmente, a largura do *Silhouette* da observação i é definida pela Eq. (2):

$$S_I = \frac{(b_i - a_i)}{\max(a_i, b_i)} \quad (2)$$

Um grande S_i (mais próximo de 1) mostra dados muito bem clusterizados, um S_i pequeno (próximo de 0) significa que a observação habita entre dois *clusters* e um S_i negativo que estão provavelmente no *cluster* errado (KASSAMBARA, 2017).

3.7.2 Índice Dunn

O índice Dunn é outra medida de validação interna de clusterização que pode ser computada como se segue:

1. Para cada *cluster*, compute a distância entre cada um dos objetos no *cluster* e os objetos nos outros *clusters*;
2. Use a mínima distância emparelhada como a separação inter-*cluster* (min.separação);
3. Para cada *cluster*, compute a distância entre os objetos no mesmo *cluster*;
4. Use a máxima distância intra-*cluster* (max.diâmetro) como a compacidade intra-*cluster*;
5. Calcule o índice Dunn como se segue na Eq. (3):

$$D = \frac{min.separação}{max.diâmetro} \tag{3}$$

Se o conjunto de dados é compacto e apresenta *clusters* bem separados, o diâmetro do *cluster* tende a ser pequeno e a distância entre *clusters* tende a ser grande. Então o índice Dunn é maximizado (KASSAMBARA, 2017).

3.8 Classificação e Seleção de Fornecedores

Fornecedores são uma peça muito importante nas operações de negócios. Os fornecedores asseguram o fornecimento de materiais, matérias-primas e commodities em quantidade suficiente, com qualidade, estabilidade e precisão em atender os requerimentos da produção e de negócios, com baixo custo e entregas *on-time*. Portanto, classificar, selecionar e gerenciar bons fornecedores é um pré-requisito para organizar a produção de produtos de qualidade, de acordo com os prazos, com preços razoáveis e competitivos no mercado (Wang *et al.*, 2018).

Para Cao e Zhang (2011), uma empresa precisa alavancar seus relacionamentos na cadeia de suprimentos para ser mais responsiva em mudar condições e aumentar demandas de clientes. As literaturas sobre gestão de fornecedores indicam um crescimento do interesse das empresas em desenvolver relacionamentos cooperativos e mutuamente benéficos com seus fornecedores (TAN; LYMAN; WISNER, 2002). Dessa forma, avaliar e selecionar fornecedores emergiu como uma área central nas execuções de uma gestão de fornecedores (SALAM e KHAN, 2018).

Para Memari *et al.* (2019), no contexto de uma cadeia de suprimentos sustentável, selecionar os melhores fornecedores que são capazes de adquirir materiais e componentes é uma decisão desafiadora. Consequentemente, oferecer assistência na tomada de decisão é benéfico para os movimentos produtivos organizacionais buscando implementar fontes

operacionais sustentáveis. Esta decisão resulta em redução de custos de compra e competitividade organizacional. (MOHEB-ALIZADEH, MAHMOUDI e BAGHERI, 2017).

De acordo com Govindan e Sivakumar (2016) a escolha dos fornecedores se tornou uma decisão importante na cadeia de suprimentos e envolve avaliar e selecionar os melhores fornecedores para requisição de abastecimento e atribuição de volumes de produção e essa forma de seleção em algumas circunstâncias é oriunda da limitação de capacidade dos mais adequados.

Segundo Sabbagh, Ameri e Yoder (2018), a classificação e seleção de fornecedores é um passo analítico necessário nos estágios iniciais da formação de uma cadeia de suprimentos e existem diferentes abordagens técnicas que podem ser utilizadas incluindo classificação, clusterização e modelagem tópica. Corroborado por Zhang *et al.* (2019) que entendem que quando empresas selecionam fornecedores, existem muitos tipos de fornecedores para serem escolhidos e os tipos de alternativas selecionadas pelos humanos vão se sobrepor. Para reduzir essa sobreposição, os algoritmos de clusterização podem classificar fornecedores de acordo com valores característicos de análise.

Nigel, Stuart e Robert (2002) explicam que existem critérios conhecidos na seleção de fornecedores que são válidos para todos os insumos e serviços adquiridos e são definidos a seguir:

- Preço correto;
- Entrega no momento certo;
- Produtos e serviços da qualidade correta;
- Na quantidade correta;
- Da fonte correta.

Para Ferreira *et al.* (2019), além desses critérios padrões, fornecedores devem ser classificados e selecionado de acordo com alguns requerimentos de especificação que atendam às necessidades particulares da companhia. Mesmo que critérios padrões não sejam usados, esta avaliação não deve depender exclusivamente dos custos associados aos produtos. Vários critérios podem ser utilizados para garantir a efetividade a avaliação dos parceiros. Enfim, cada organização deve escolher seus critérios e indicadores que melhor se adequem às suas políticas de gerenciamento.

Segundo Banaeian *et al.* (2018) além dos critérios tradicionais de avaliação de fornecedores a incorporação de um critério ambiental na seleção de fornecedores está com

importância crescente e o desenvolvimento e disponibilidade de novas ferramentas analíticas e modelos de seleção de fornecedores podem ajudar a resolver muitos desafios enfrentados pela gestão de compras e suprimentos.

Pelissari, Ben-Amor e De oliveira (2019) realizaram uma análise de artigos e notaram que existe um grande problema na gestão de fornecedores que corresponde a seleção dos fornecedores e a maioria das técnicas aplicadas para solucionar os problemas se baseiam em métodos que atribuem *ranking*.

Dessa forma, avaliar e selecionar fornecedores emergiu com um foco renovado nas execuções de uma gestão de fornecedores (KUMAR; ROUTROY, 2017).

3.9 Linguagem R

Para Lin, Liu e Peng (2017) o R auxilia na análise de dados com poderosos recursos de operações matriciais e ferramentas gráficas, como é um s*oftware* gratuito e *open source*, apresenta vasta quantidade de pacotes escritos por usuários R que podem ser encontrados em seu site oficial. Além de que, nos anos recentes se tornou muito popular e usado por vários profissionais como analistas de riscos, pesquisadores e estatísticos e sua rápida popularização é devido a suas capacidades orientadas a objetos, habilidade de executar funções e pacotes definidos por usuários, flexibilidade na sintaxe e facilidade de editar características.

O R é um *software* livre para computação estatística e de geração de gráficos que compila e executa em uma ampla variedade de plataformas Unix, Windows e Mac (CORE TEAM, 2018).

O R é fundamentalmente estatístico, entretanto sua equipe de desenvolvimento e a comunidade ativa de autores de pacotes investiram bastante tempo e esforço para expandir sua usabilidade e a linguagem apresenta pacotes de bibliotecas que expandem suas funcionalidades básicas.

Para facilitar a utilização dessa linguagem foi criado o *RStudio*, um ambiente de desenvolvimento integrado reunindo características e ferramentas de apoio ao desenvolvimento de R com o objetivo de agilizar o processo.

Existem pesquisas que relacionam quais são as linguagens mais utilizadas por profissionais de *Data Science*, uma delas, chamada de “2017 *Data Science Salary Survey*” contou com a participação de cerca de 800 profissionais de 69 países para definir melhor quais ferramentas estão sendo usadas, para onde a indústria se direciona e uma noção do salário da comunidade (SUDA, 2018)

Quando os participantes foram perguntados sobre as linguagens de programação, SQL lidera com 64% dos correspondentes, seguidos de Python com 63% e R com 54%. O R está entre as mais utilizadas para trabalhar com *Data Science* mostrando a relevância da linguagem.

De acordo com uma pesquisa realizada em 2020 pela maior organização profissional do mundo dedicada à engenharia e às ciências aplicadas – IEEE – a linguagem R está entre as dez principais linguagens de programação de código aberto enunciadas na Figura 3.2.

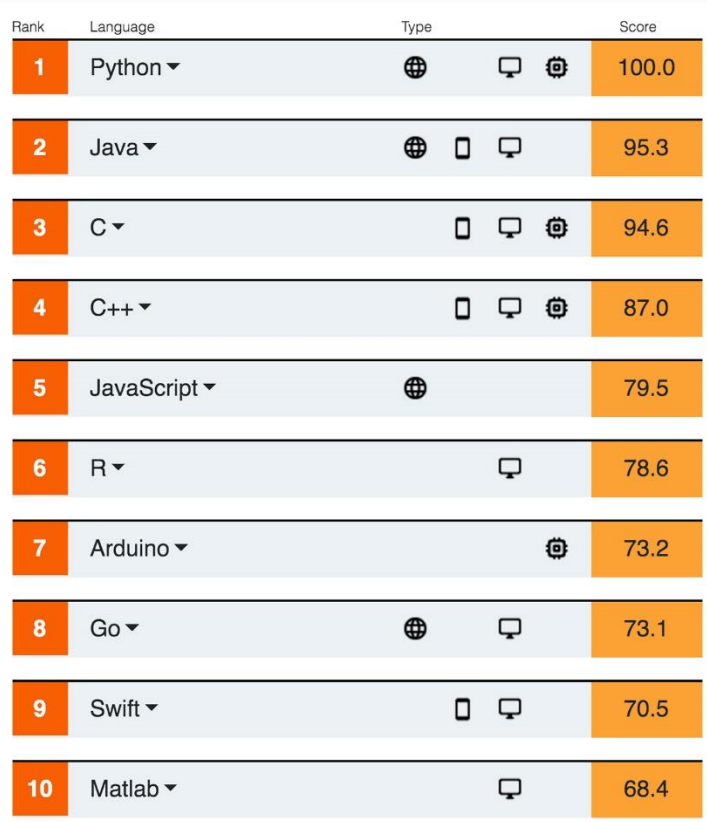

Figura 3.1: Comparação de Linguagens de Programação

Fonte: Adaptado de IEEE Spectrum

4 METODO DE PESQUISA

4.1 Classificação da Pesquisa

A Figura 4.1 apresenta a classificação da metodologia de pesquisa conforme o modelo proposto por Cauchick Miguel *et al.* (2018).

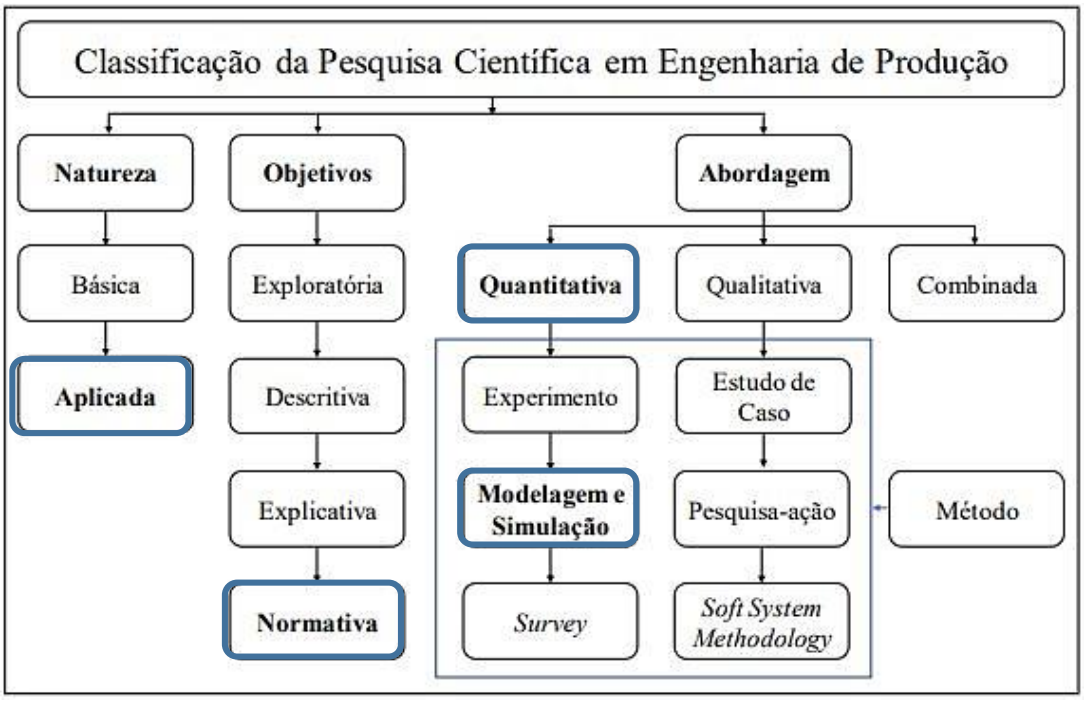

Figura 4.1: Classificação da Pesquisa

Fonte: Adaptado de Cauchick Miguel *et al.* (2018)

Turrioni e Mello (2012) e Cauchick Miguel *et al.* (2018) classificam as pesquisas cientificas na área de engenharia de produção em quatro dimensões básicas: quanto à natureza, objetivos, abordagem e método. Esta pesquisa pode ser classificada como de natureza aplicada, pois segundo Turrioni e Mello (2012, p.80) uma pesquisa aplicada "caracteriza-se por seu interesse prático, isto é, que os resultados sejam aplicados ou utilizados imediatamente na solução de problemas que ocorrem na realidade".

Quanto aos objetivos da pesquisa, pode ser classificado como exploratório, descritivo, explicativo ou normativo. De acordo com Bertrand e Fransoo (2002), uma pesquisa normativa é aquela que busca encontrar soluções otimizadas para novas definições ou está interessada no desenvolvimento de políticas e estratégias para aperfeiçoar resultados encontrados na literatura. Logo, esta pesquisa possui objetivos normativos, uma vez que objetiva encontrar um *framework* que possa ser utilizado para se agrupar dados reais sem supervisão.

Além disto, esse trabalho apresenta uma abordagem quantitativa, a qual parte da teoria para se formular hipóteses, coletar dados através de observações, analisá-los e gerar resultados (CAUCHICK MIGUEL *et al.*, 2018). Dentro desta abordagem existem diversos métodos e para esta pesquisa o método de modelagem e simulação mostrou-se o mais adequado, uma vez que, segundo Cauchick Miguel *et al.* (2018), representa a manipulação de variáveis em um modelo representativo da realidade, todavia sem afetar o ambiente real durante esta manipulação.

4.2 *Cross-Industry Industrial Standard Process for Data Mining* (CRISP-DM)

Extrair conhecimento útil de dados para solucionar problemas de negócios pode ser tratado sistematicamente utilizando métodos de pesquisa (PROVOST e FAWCETT, 2013). Segundo Rose (2016), *Data Science* não é direcionada à objetivos, é exploratória e usa um método científico, não é sobre quão bem uma organização opera mas sobre ganhar conhecimento útil para negócios.

Existem alguns procedimentos para se trabalhar com *Data Science* como o "*Data Science Process*" (Schutt, O´Neil, 2014), o "Processamento de dados" (Aggarwal, 2015) e o "*Cross-industry Industrial Standard Process for Data Mining*" (CRISP-DM) de Pete *et al.* (2000). Para essa pesquisa optou-se por trabalhar com o CRISP-DM, que, como pontua Wirth e Hipp (2000), consiste em um procedimento desenvolvido nos anos 90 por um grupo de cinco empresas: SPSS, *TeraData*, Daimler AG, NCR e OHRA.

O procedimento escolhido e os demais são bem semelhantes entre si quando analisadas quanto aos seus fundamentos para entender o processo de *Data Science* e basicamente representam o mesmo processo científico de se estudar dados com pequenas particularidades.

Utilizar o procedimento CRISP-DM ajuda a estruturar o pensamento científico para solucionar problemas de análises de dados. Ter esse pensamento estruturado sobre os dados enfatiza aspectos desconhecidos anteriormente para a tomada de decisão usando dados (PROVOST e FAWCETT, 2013).

Segundo Abbasi, Sarker e Chiang (2016) o procedimento CRISP-DM é amplamente considerado como o princípio orientador mais relevante e compreensivo para realização de projetos analíticos. Esse procedimento e suas fases são enunciadas na Figura 4.2.

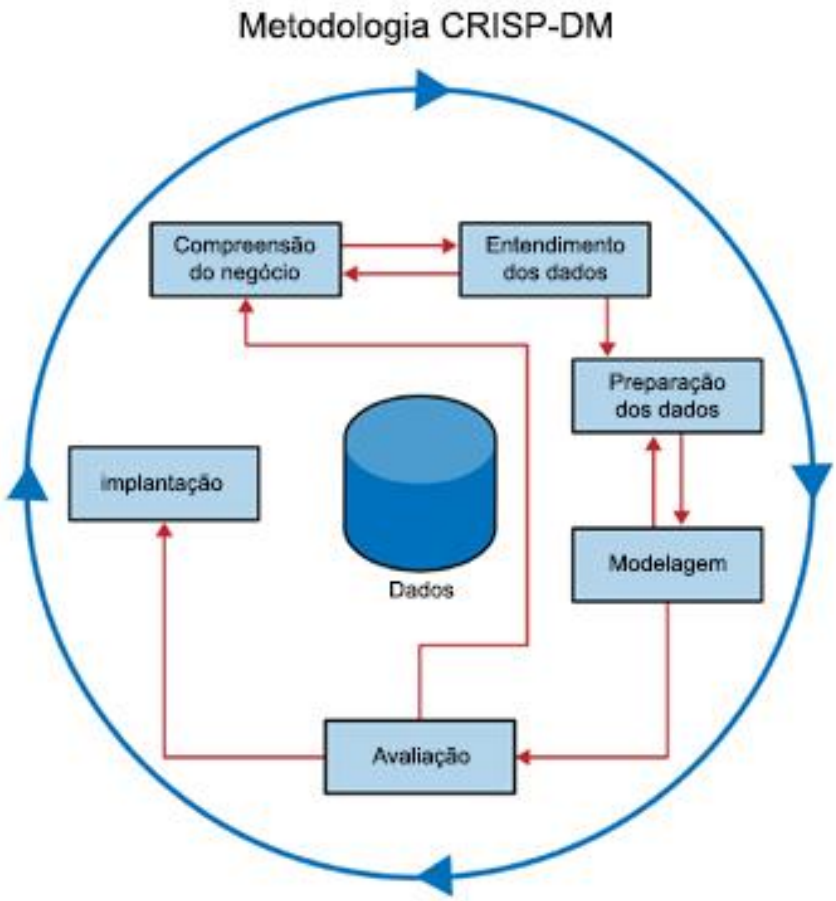

Figura 4.2: Procedimento CRISP-DM

Fonte: Traduzido de Provost e Fawcett (2013)

As fases do procedimento são detalhadas abaixo:

- Compreensão do Negócio (*Business Understanding*): Envolve a compreensão do problema a ser resolvido através da análise do cenário dos dados (o contexto) e da definição dos objetivos;
- Compreensão dos Dados (*Data Understanding*): Procura o entendimento profundo dos dados, seus pontos fortes e suas limitações. Por meio da organização, descrição e análise da qualidade dos dados;
- Preparação dos Dados (*Data Preparation*): Para que se encontre informações relevantes dos dados pode ser necessário que se realizem manipulações neles para que passem por algum processo posterior;
- Modelagem (*Modeling*): Nessa fase assume-se que os dados se encaixam em uma suposição que pode ser capturada por um modelo;

- Avaliação (*Evaluation*): Tem como finalidade garantir que o modelo gere resultados válidos e confiáveis que satisfaçam os objetivos definidos na primeira fase [PROVOST e FAWCETT, 2013],
- Implantação (*Deployment*): Segundo Olson e Delen (2008), nessa fase final, o conhecimento adquirido por meio da mineração de dados deve estar relacionado aos objetivos iniciais do projeto e pode, então, ser aplicado no ambiente de negócios. Esse conhecimento é útil no planejamento e tomada de decisão, mas é importante que os resultados sejam monitorados e o modelo seja adaptado sempre que necessário.

Segundo Provost e Fawcett (2013), o processo é útil como um *framework* para analisar um projeto ou proposta. Ele provê uma organização sistemática incluindo um conjunto de questões que podem ser inquiridas para ajudar a entender se o projeto é bem desenvolvido ou fundamentalmente imperfeito.

5 DESENVOLVIMENTO

Os títulos primários seguintes consistem na aplicação das 6 etapas do procedimento CRISP-DM seguindo o detalhamento proposto no Capítulo 4. Começando com a compreensão do negócio no contexto da empresa aeronáutica estudada, seguido por uma compreensão e preparação mais profunda dos dados, além da validação da base obtida quanto à tendência de clusterização.

Com essa tendência comprovada, a modelagem dos métodos para resolução do problema da clusterização foi realizada, em seguida, ocorreu a etapa de implantação, na qual houve a validação dos algoritmos de clusterização, confirmando quais os mais eficientes para a base de dados em estudo. Dentro dessa mesma etapa, foi realizada a modelagem desses melhores algoritmos encontrados com seus parâmetros mais assertivos.

5.1 Compreensão do Negócio

Essa pesquisa utilizou-se de dados dos fornecedores de uma empresa do setor aeronáutico que atua em Minas Gerais contando com aproximadamente 560 fornecedores e que não será identificada por questões de sigilo. Para o gerenciamento desses fornecedores, a empresa usa um software corporativo de *Enterprise Resource Planning* (ERP) e o Microsoft Excel.

ERP abrange todas as funções e departamentos em um sistema integrado e consiste de uma base empresarial na qual todas as transações comerciais são adicionadas, armazenadas, processadas, monitoradas e relatadas (UMBLE, HAFT e UMBLE, 2003).

A primeira experiência para começar a compreender o negócio é entender qual o problema de pesquisa. E esse problema se encontra nas estratégias de classificação e seleção de fornecedores. Para que uma empresa as defina é imprescindível que haja uma boa métrica para a seleção de seus fornecedores, e essa seleção de parceiros necessita de uma classificação eficiente. Portanto, para que as práticas de gerenciamento de fornecedores possam ser planejadas é necessário uma classificação e seleção apropriadas. É nesse ponto que a pesquisa age.

A Figura 5.1 representa o problema estudado sob dois pontos de vista, o computacional e o do contexto da empresa (real).

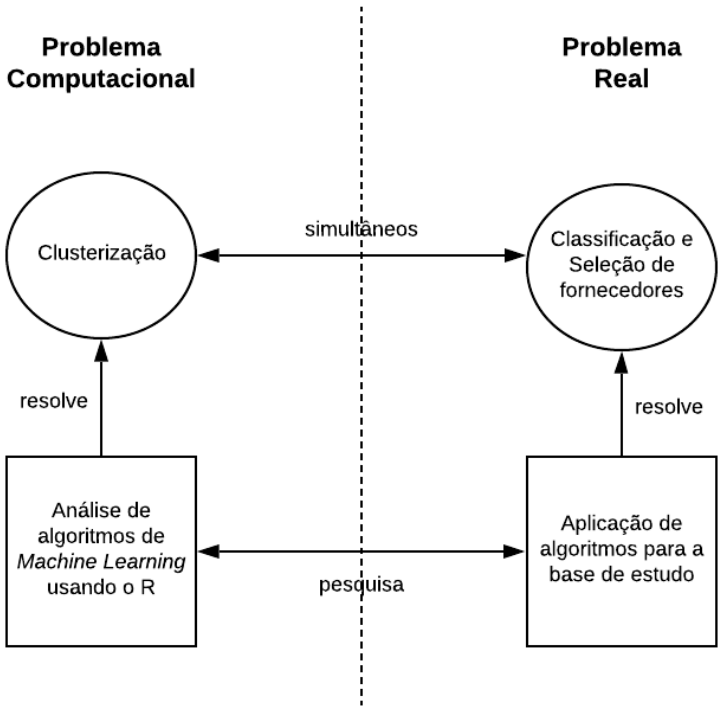

Figura 5.1: Problema de Pesquisa

Fonte: Autoria Própria

Nota-se pela Figura 1 que, computacionalmente, existem problemas que podem ser solucionados utilizando *Machine Learning*, um deles é o problema da clusterização, também conhecido como agrupamento. Segundo Mohri, Rostamizadeh e Talwalkar (2012) clusterização representa particionar itens em regiões homogêneas, quando não se tem categorias previamente definidas. Para se trabalhar com *Machine Learning* é necessário utilizar alguma ferramenta computacional que permita estas manipulações, para esta pesquisa será utilizada a linguagem computacional R. Além disso, para se resolver o problema de clusterização é necessário estudar os algoritmos com essa finalidade em busca de um entendimento sobre eles e quando e como podem ser aplicados.

Do outro lado, têm-se o problema real da empresa de classificação e seleção de fornecedores que é análogo ao problema computacional de clusterização. Dessa forma, busca-se por meio do uso de *Data Science* melhorar esse processo decisório de fornecedores, resolvendo um problema computacional e real simultaneamente. Isso se encaixa no problema da clusterização que segmenta uma população de indivíduos por suas similaridades sendo seu resultado utilizado pela empresa como uma entrada no processo de tomada de decisão

focando em investigar quais grupos de fornecedores podem ser mais interessantes baseados em suas características e nas necessidades da empresa.

É interessante notar que o problema da clusterização de fornecedores apresenta uma gama de ferramentas que permitem alcançar sua solução e para que se possam definir os melhores modelos, métricas precisam ser utilizadas para suas validações. De forma sintetizada, espera-se organizar em *clusters* os fornecedores da empresa buscando encontrar os melhores algoritmos que realizem essa segmentação auxiliando os tomadores de decisão da organização em uma melhor tratativa para a classificação e seleção de seus parceiros fornecedores.

5.2 Compreensão dos Dados

Como já foi citado, a empresa utiliza-se de um ERP para o controle de seus fornecedores e os dados utilizados para a realização dessa pesquisa acadêmica foram coletados dessa ferramenta. Para Rose (2016) um dos maiores desafios da ciência dos dados é ter acesso aos dados organizacionais. Foi um processo lento dada a capacidade de resposta dos computadores utilizados, que culminou em dados de compras de janeiro a novembro de 2019.

Sendo os dados difíceis de se obter dada a complexidade de requisição de acesso interno à eles, optou-se por levar mais tempo nessa obtenção de dados, mas já coletando todos os dados brutos disponíveis no ano do estudo para oferecer maior qualidade às análises realizadas e evitar a burocracia de uma aquisição por partes.

Os dados obtidos consistiam em planilhas do Microsoft Excel e antes de enviá-los para o R para as análises, eles passaram por uma estruturação e as informações foram organizadas e unidas em uma mesma planilha.

Segundo Rose (2016), os dados podem ser estruturados, tendo um formato específico em uma estrutura definida, recebendo o nome de modelo de dados. Também podem ser semiestruturados e apresentarem certa estrutura, mas com flexibilidade para mudar nomes de campos ou para criar novos valores. Também existem os dados desestruturados que não seguem um esquema definido.

Os dados obtidos para a pesquisa podem ser considerados estruturados por consistirem de planilhas com formato e estrutura definidos consistindo em linhas e colunas no formato matricial. Os dados coletados do sistema ERP foram organizados em uma planilha de Excel contendo como campos: Código identificador exclusivo utilizado pelo ERP, nome do

fornecedor, a localidade e o país em que opera, a soma da quantidade de itens dos pedidos, a soma do valor líquido dos pedidos, a soma de preço médio móvel e os conjuntos de atividade utilizados para gerenciar essa compra

Os campos dos conjuntos de atividades utilizadas para gerenciar as compras consistem em auditorias, avaliação de relevância dos suprimentos, avaliação da demanda da companhia para os produtos requeridos, análise das não-conformidades geradas e as avaliações de desempenho periódicas e suas datas relevantes.

Essa planilha foi utilizada como base de dados para dar continuidade ao CRISP-DM. Um adendo importante faz-se necessário, o campo nome do fornecedor foi removido e não utilizado por segurança da informação. Para a pesquisa, o código identificador do ERP era suficiente para a identificação dos dados e para os tomadores de decisão é fácil restaurar essa informação. Finalmente, essa planilha foi preparada para iniciar as análises.

5.3 Preparação dos Dados

Dos campos da planilha, efetivou-se uma análise da qualidade das informações quantitativas e foram excluídos todos os resultados que apresentavam dados faltantes ou que estavam duplicados, atingindo o número final de 65 fornecedores no ano da pesquisa.

O próximo passo foi organizar os dados que continham informações numéricas em faixas de valores para que houvesse algum significado ao realizar a clusterização e para que pudessem ser comparáveis. Por meio de uma análise das médias globais dos campos foi considerado o grupo de 7 faixas de valores apresentado na Tabela 5.1 que representa um exemplo dessa tratativa de dados.

Tabela 5.1: Transformação de valores

Faixa (Início)	Valor
0	7
1000	6
10000	5
50000	4
100000	3
500000	2
1000000	1

Fonte: Autoria Própria

Assim, o mesmo foi feito com os dados de auditoria, mas com o objetivo de separá-los em categorias atingindo 17 grupos auditáveis. Por fim, atingiu-se uma maturidade maior para os dados serem trabalhados no R e a base foi enviada para o *RStudio*, uma interface de desenvolvimento do R, e foi convertida para um formato aceito pela linguagem.

Essa etapa apresentou um grande problema, os dados ficaram com os formatos ‘tbl_df’, ‘tbl’ e '*data.frame*' quando convertidos para R, ou seja, uma mescla de formatos de dado que não seriam aceitos por nenhum algoritmo utilizado posteriormente e que devido a isso, acabou ocasionando um primeiro retorno do ciclo do CRISP-DM para a etapa de preparação dos dados para conversão desse formato. A própria metodologia entende que essas situações poderiam ocorrer e apresenta ligações de ida e volta entre as duas etapas. Todos os códigos desenvolvidos estão no Anexo A – Código R para Clusterização.

Com essa informação do formato necessário para a análise, manipulações computacionais nos dados foram efetuadas para convertê-los para apenas o formato ‘*data-frame*’. Dessa forma, o problema foi solucionado e a pesquisa pôde ter continuidade.

Como se trata de uma possível modelagem de clusterização, antes da aplicação de qualquer técnica em busca de uma solução para o problema de pesquisa se faz necessário avaliar a natureza dos próprios dados e se é possível validar a aplicação de tais ferramentas. Para isso, é necessário realizar a tendência de clusterização nessa base estudada.

Outra necessidade intrínseca é a padronização dos dados antes das medições de dissimilaridades entre eles para fazer com que as variáveis sejam comparáveis. O R apresenta uma função chamada daisy(), baseada no método de Kaufman e Rousseeuw (1990b), que se encontra no pacote *cluster* e que consegue trabalhar com dados de diferentes formatos permitindo que possam ser comparáveis.

Optou-se por realizar uma nova compreensão dos dados dentro da etapa de preparação dos dados agora que eles estavam em um formato possível de se comparar e avaliar. Antes da aplicação de um modelo de clusterização efetivamente ser desenvolvido com base neles e isso aconteceu por meio de dois método, o método da Estatística de Hopkins e o método chamado avaliação visual de tendência de *cluster* que se utiliza de *heatmap*.

Existe um método estatístico conhecido como Estatística de Hopkins que basicamente consiste de um teste de hipótese cuja hipótese nula assume que os dados são uniformemente variados e sua alternativa representa que os dados não são uniformemente variados, ou seja, contêm *clusters* significativos.

Para valores resultantes no teste menores que 0,5 pode-se rejeitar a hipótese nula e aceitar que os dados são clusterizáveis.

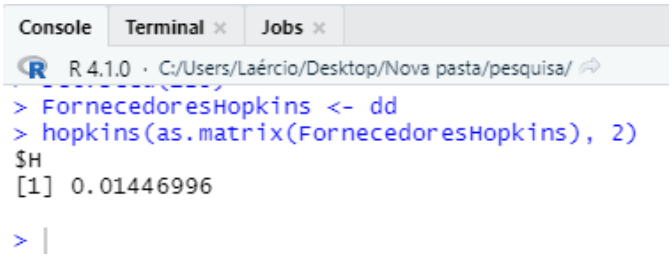

Figura 5.2: Resultado da Estatística de Hopkins no R

Fonte: Autoria Própria

A Figura 5.2 mostra o resultado desse índice, a estatística de Hopkins apresenta valor de 0,01447 e pode-se considerar que a base analisada apresenta *clusters* significativos. Consequentemente, a escolha dos dados para a realização da clusterização incluídos nessa pesquisa pôde ser comprovada com tendência para a clusterização.

Além do método estatístico, existe o método visual conhecido como avaliação visual da tendência do *cluster* cujo resultado se encontra na Figura 5.3.

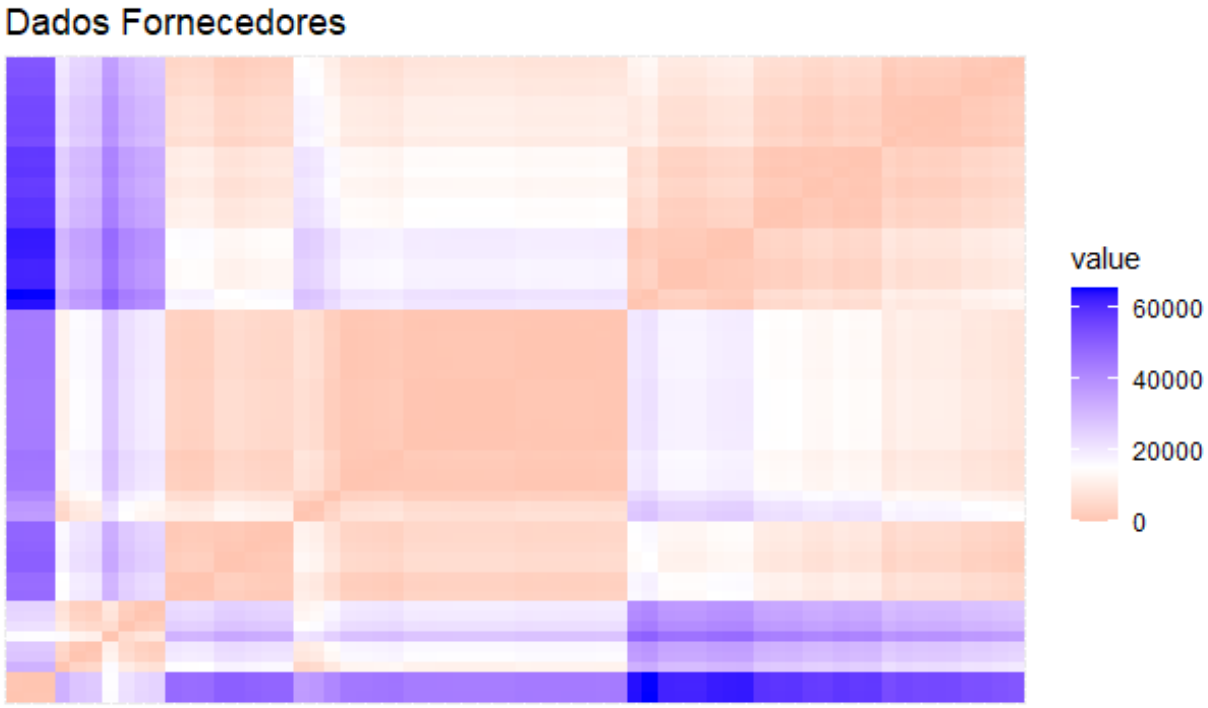

Figura 5.3: Método Visual para avaliar dados

Fonte: Autoria Própria

Nesse método, dados em vermelho representam uma similaridade alta entre as observações e dados azuis uma similaridade baixa entre eles. Com base nisso, pode-se

visualizar que os dados têm uma característica bastante interessante para a clusterização pela densa população vermelha.

Além disso, a coloração da diagonal principal também mostra a tendência para clusterização, se for vermelho escuro os dados não são interessantes para o estudo, se forem de uma tonalidade bem clara tem um comportamento que vale a pena ser estudado. De acordo com esse teste, os dados também se mostraram alinhados com a pesquisa.

Realizada a preparação de dados, conversão para um formato tabular, remoção de valores ausentes, verificação de *outliers* e conversão e validação da base, a etapa de preparação dos dados foi realizada.

5.4 Modelagem

A Modelagem é uma etapa crítica do CRISP-DM e encontrar formas de representar os dados desenvolvidos na pesquisa em um modelo pré-existente e que essa representação seja plausível e tenha utilidade para os tomadores de decisão exige que se encontrem métricas que permitam avaliar essas construções.

O primeiro passo é encontrar quais mecanismos da ciência dos dados permitem que se estudem os problemas em análise, começando em um entendimento deles já realizado na etapa de compreensão e atribuindo o contexto do negócio nessa análise.

Notou-se a inexistência de uma medida de classificação concreta por parte da empresa, o que representa que não existe nenhum parâmetro de comparação com dados verdadeiros de uma classificação e na falta de uma medida comparativa faz-se necessário utilizar mecanismos não supervisionados, ou seja, que independam de quaisquer dados externos aos obtidos na pesquisa para realizar a classificação.

Por medida de classificação está sendo considerado um resultado externo que permitisse uma comparação dos resultados obtidos de uma classificação com essa base, ocasionando utilizar técnicas de classificação em oposição aos interesses da clusterização. Por exemplo, poderia existir uma lista de fornecedores classificando-os como péssimos, ruins, bons e ótimos. Caso existisse essa informação poderia existir uma comparação com essa base rotulada.

Portanto, em posse dessa avaliação fica mais claro que os objetivos se encaixam na clusterização que é a técnica de classificação sem esse comparativo, com base no entendimento dos próprios dados que permitem que se definam grupos (conhecidos como

clusters), dados no mesmo grupo são mais similares entre si e de grupos diferentes o mais heterogêneos quanto possível.

Com o conhecimento da natureza dos dados, do funcionamento do processo de análise e do problema que deve ser combatido pôde-se dar continuidade na seleção da opção mais eficiente de ferramenta a ser utilizada já que os dados foram validados para uma clusterização na etapa anterior e a clusterização foi confirmada nessa etapa.

Para realização de uma clusterização podem ser utilizadas técnicas de particionamento, hierárquicas, *soft clustering* ou avançadas. Os dados utilizados nesta pesquisa têm caráter multivariado e necessitam de certas manipulações para uma melhor visualização dos resultados na maioria dos algoritmos utilizados. Consequentemente, eles precisavam ter um redimensionamento apropriado para a obtenção dos gráficos dos resultados dos algoritmos.

Para a produção de tais gráficos dos modelos, o método *Principal Component Analysis* (PCA) foi utilizado. O PCA utiliza os dados originais como argumentos quando existem mais de 2 variáveis na base multidimensional que ao final do processo originam uma redução para apenas duas novas variáveis que melhor representam os dados e podem ser visualizadas no eixo cartesiano minimizando as perdas de suas características iniciais. Então nos gráficos, Dim1 e Dim2 representam essas novas variáveis e a nova escala utilizada se baseia nos valores obtidos pelos métodos PCA para que os dados fiquem mais bem visualizados nos gráficos, com exceção do *Heatmap* que não necessita de redução de dimensionalidade e consiste de uma forma visual de clusterização por natureza.

Para visualização dos exemplos de aplicação foram utilizados parâmetros para a quantidade de 3 *clusters*. Todas as distâncias utilizadas nos gráficos tratam-se de distâncias euclidianas para medidas de dissimilaridade entre objetos. Outra questão de destaque foi com relação as visualizações dos gráficos, optou-se por utilizar manipulações computacionais no R que impedem sobreposição de rótulos, ou seja, os gráficos ficaram com todos os dados de todos os *clusters* a mostra, sem esconder nenhum.

Uma consideração importante é que o desenvolvimento de todos os algoritmos como se segue foi realizado para alinhar com o ***primeiro objetivo específico*** de demonstrar o funcionamento e o comportamento dos algoritmos clássicos de clusterização em uma base de dados real. Posteriormente nessa pesquisa será efetivada uma validação para a escolha dos melhores algoritmos e dos melhores parâmetros e esses modelos também serão desenvolvidos.

As modelagens desenvolvidas nessa pesquisa estão enunciadas a seguir, de acordo a ordem definida no Capítulo 3.

5.4.1 *K-Means*

A primeira modelagem gerada para a pesquisa é o *k-means*, representante da clusterização de particionamento. No R, uma das possíveis soluções dessa modelagem se encontra no pacote *stats* por meio da função *kmeans()* e foi o método utilizado nessa pesquisa. Essa biblioteca utiliza o algoritmo clássico de Hartigan e Wong (1979).

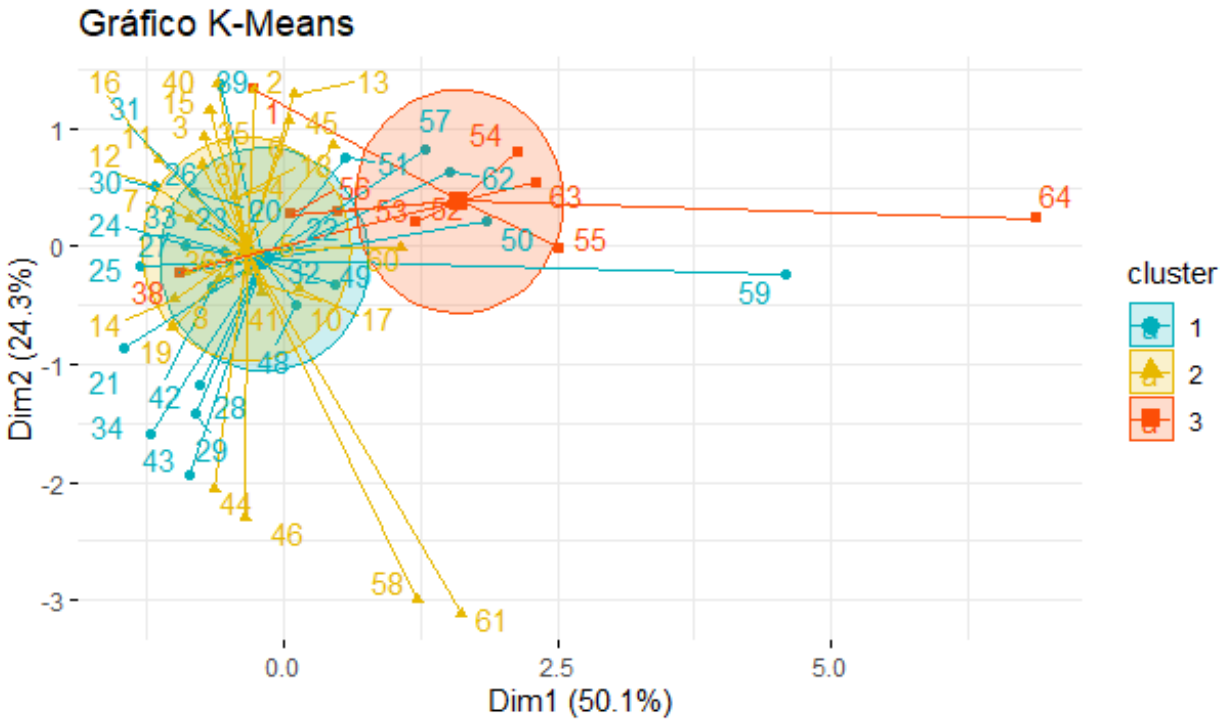

Figura 5.4: Gráfico de Dispersão dos resultados da clusterização pelo método *k-Means* (3 *Clusters*)

Fonte: Autoria própria

O Gráfico da Figura 5.4 mostra o resultado pós dimensionamento PCA, e para sua criação foi utilizado o pacote *factoextra* do R e a função fviz_cluster(). Esse gráfico representa o resultado dessa clusterização englobando todos os dados como representações pontuais e a que *cluster* pertence cada dado de entrada. Os três *clusters* obtidos são representados pelas cores azul, amarela e vermelha.

5.4.2 *Partitioning around medoids* (PAM)

Outro representante da clusterização de particionamento é o método PAM, uma alternativa ao *k-means* que o deixa menos sensitivo a *outliers*. No R, utiliza-se o pacote *cluster* com a função pam() e utiliza o algoritmo criado por Kaufman e Rousseeuw (1990).

A Figura 5.5 representa um gráfico de dispersão colorido pelos números de *clusters*. As cores azuis ficaram relativamente parecidas, todavia os *clusters* 1 e 2 estão bem definidos na imagem, o terceiro *cluster* é representado pela cor amarela. Para desenvolvimento do gráfico utilizou-se o pacote *factoextra* no R e a função fviz_cluster().

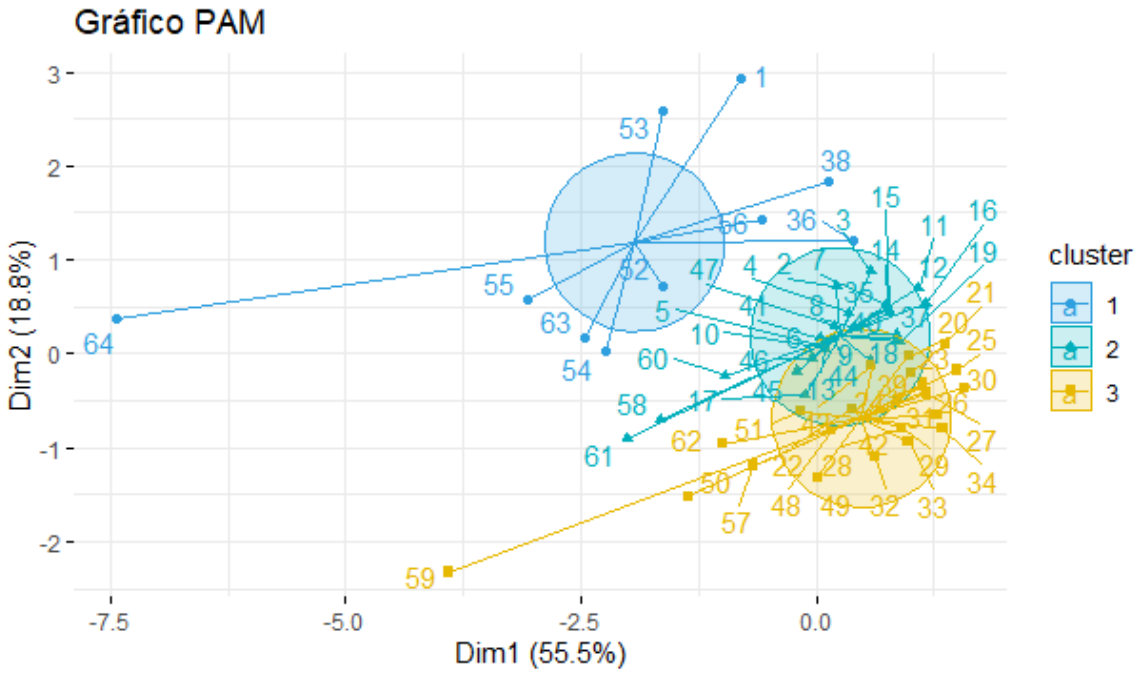

Figura 5.5: Gráfico de Dispersão dos resultados da clusterização pelo método PAM (3 *Clusters*)

Fonte: Autoria Própria

5.4.3 *Clustering Large Applications* (CLARA)

O PAM apresenta uma variante que tem um desempenho melhor ao alimentá-la com uma grande quantidade de dados e se trata do método CLARA (*Clustering Large Applications*), outro representante da clusterização de particionamento. No R, utilizou-se o mesmo pacote do algoritmo anterior, o pacote *cluster* e a função clara() que utiliza o algoritmo de Kaufman e Rousseeuw (1990).

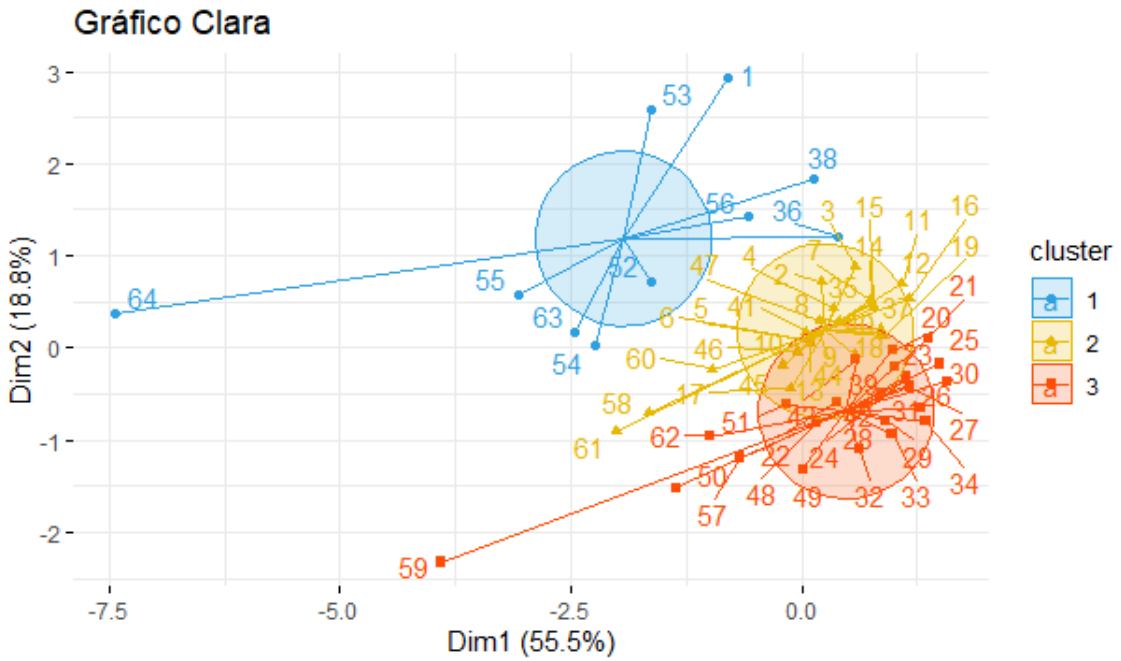

Figura 5.6: Gráfico de Dispersão dos resultados da clusterização pelo método PAM (3 *Clusters*)

Fonte: Autoria Própria

O Gráfico da Figura 5.6 mostra os resultados obtidos ao realizar a ingestão de dados no algoritmo CLARA com os dados do banco de dados dos fornecedores da empresa estudada. O mesmo pacote foi utilizado para criação desse gráfico de dispersão, o *factoextra*, e a função fviz_cluster() que pertence a esse pacote.

Têm-se três *clusters* representados nas cores azul, amarelo e vermelho. Apesar de se utilizar uma cor mais distinguível ele tem o mesmo comportamento do PAM exibido na Figura 5.5. O CLARA é um PAM para grandes quantidades de dados então faz sentido os dois resultados apresentados serem bem semelhantes. A ideia dele era reduzir o trabalho computacional para grandes bases de dados, mas não havia essa necessidade para a pesquisa por se tratar de uma base pequena.

5.4.4 *Agglomerative Nesting* (AGNES)

Como representante da clusterização hierárquica o algoritmo AGNES (*Agglomerative Nesting*) foi desenvolvido no R. Os métodos hierárquicos criam informações visuais interessantes por meio de dendrogramas.

Para se utilizar esse tipo de clusterização pode-se utilizar a função agnes() do pacote *cluster*, que já realiza todos os procedimentos necessários à base de dados para o desenvolvimento do modelo no R. O algoritmo do R é descrito em Kaufman e Rousseeuw

(1990). Para se compilar esse algoritmo, os dados de entrada precisavam ser inseridos com as linhas contendo informações e as colunas sendo as variáveis, um formato típico de um banco de dados relacional aos quais se encaixam os dados estudados.

O AGNES trata individualmente as observações e vai unindo até formar a árvore exibida no dendrograma. É "*bottom-up*", método que consiste em unir *clusters* formando *clusters* maiores, de baixo para cima. No dendrograma, cada folha corresponde à um objeto, enquanto se move para cima na árvore, objetos que são similares são combinados em ramos, que são fundidos nos níveis mais altos. A altura da fusão, indicada pelo eixo vertical, indica a dissimilaridade entre os objetos/*clusters*. Quanto mais alta for essa altura, menos similares são os objetos. Essa altura é conhecida como "*cophenetic distance*" entre dois objetos.

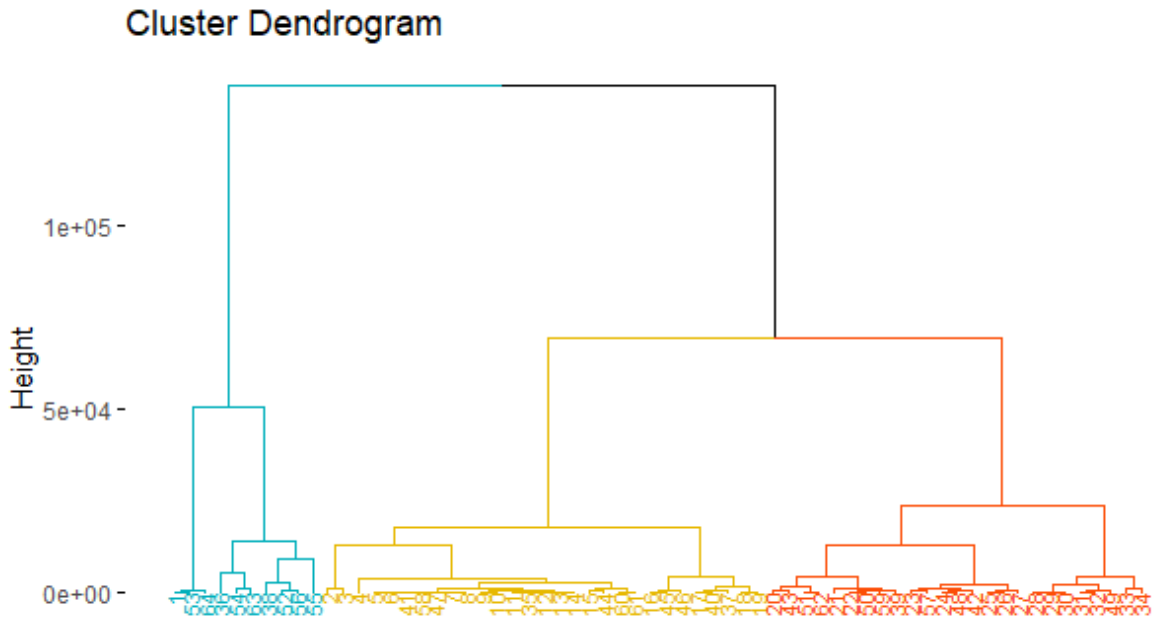

Figura 5.7: Dendrograma dos resultados da clusterização pelo método AGNES (3 *Clusters*)

Fonte: Autoria Própria

O dendrograma da Figura 5.7 mostra os resultados obtidos pela clusterização usando o AGNES nesse formato intuitivo para visualização de *clusters*. Um dendrograma corresponde a representação gráfica de uma árvore hierárquica gerada pela função fviz_dend() do pacote *factoextra*.

Os três *clusters* são exibidos nas cores azul, amarela e vermelha, e a estrutura de árvore do dendrograma permite que se observem as similaridades maiores mesmo dentro dos objetos do mesmo *cluster*. O Gráfico da Figura 5.8 representa o resultado da clusterização

usando o algoritmo AGNES com o processo PCA. Este gráfico foi gerado com o suporte do pacote *factoextra* do pacote fviz_cluster(). Os *clusters* são representados pelas cores azul, amarela e vermelha e representam os mesmos *clusters* exibidos na Figura 5.7 em outro formato visual.

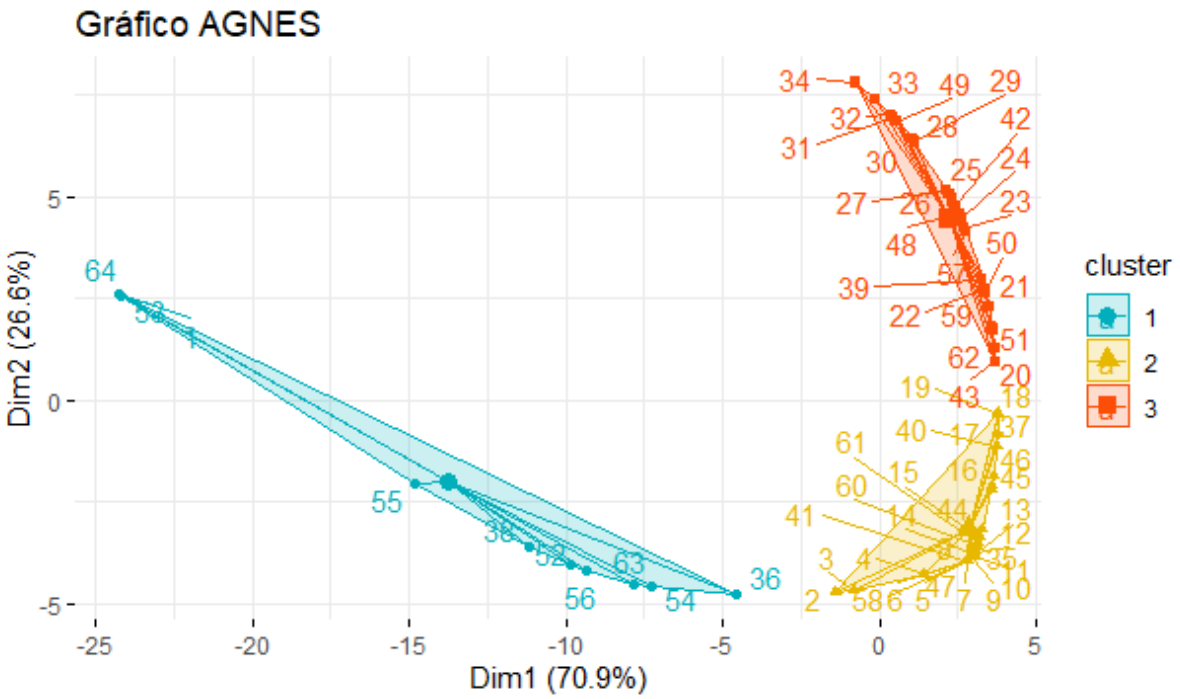

Figura 5.8: Gráfico de Dispersão dos resultados da clusterização pelo método AGNES (3 *Clusters*)

Fonte: Autoria Própria

5.4.5 *Divisive Analysis* (DIANA)

DIANA é um método de clusterização hierárquica assim como o AGNES. Para desenvolvê-lo no R utilizou-se a função diana(), também pertencente ao pacote *cluster* que é descrito em Kaufman e Rousseeuw (1990). A Figura 5.9 mostra o dendrograma da clusterização usando o método DIANA usando a função fviz_dend() do pacote *factoextra*.

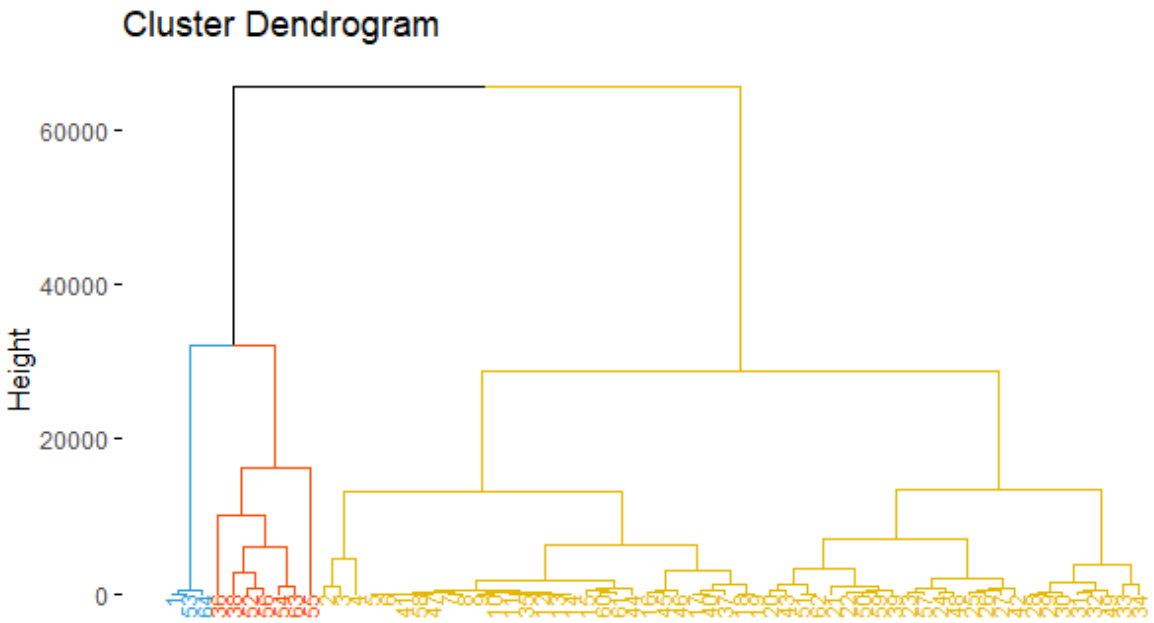

Figura 5.9: Dendrograma dos resultados da Clusterização pelo método DIANA (3 *Clusters*)

Fonte: Autoria Própria

Na Figura 5.9, o dendrograma apresenta as cores azul, amarela e vermelha representando os três clusters obtidos.

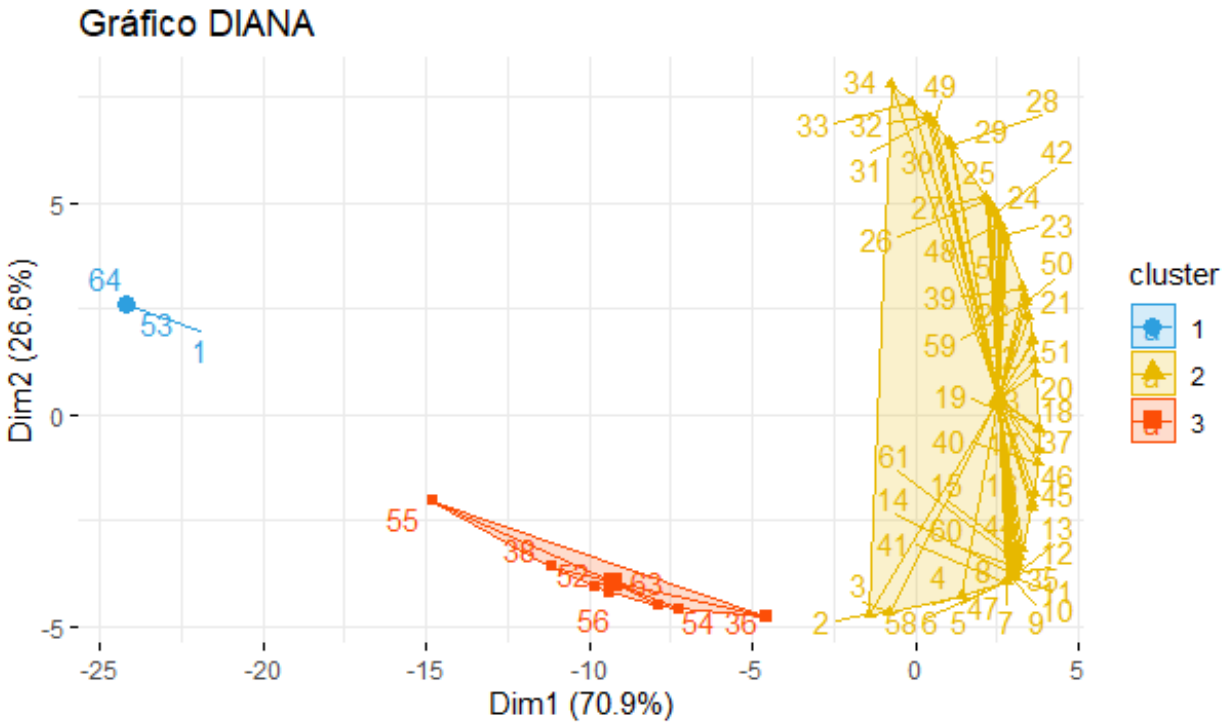

Figura 5.10: Gráfico de Dispersão dos resultados da clusterização pelo método DIANA (3 *Clusters*)

Fonte: Autoria Própria

O gráfico da Figura 5.10 mostra o resultado da clusterização usando o algoritmo DIANA para 3 *clusters* feito pela fviz_cluster() no R que são representados pelas cores azul, amarela e vermelha e estão de acordo com a Figura 5.9.

5.4.6 *HEATMAP*

Foi utilizada a função pheatmap() no R que pertence ao pacote pheatmap para representar o algoritmo *heatmap* de clusterização hierárquica.

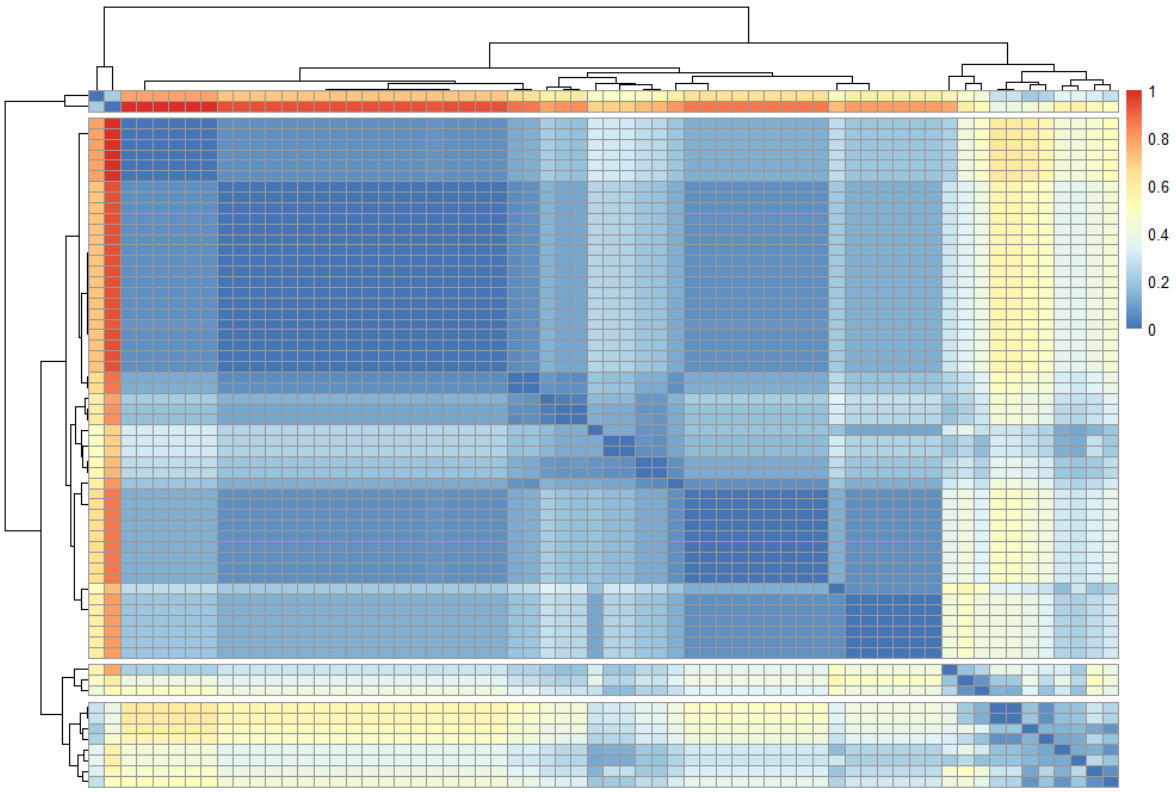

Figura 5.11: Gráfico dos resultados da clusterização pelo método Heatmap

Fonte: Autoria Própria

O Gráfico exibido na Figura 5.11 é um *heatmap* desenvolvido com a função do R pheatmap() para os dados analisados nessa pesquisa. O Heatmap é outra forma hierárquica para visualização e dados em um formato que classifique um conjunto de dados sem supervisão. Onde os dados são transformados em escala de cor.

5.4.7 *FUZZY*

No R, pode-se utilizar a biblioteca fanny() do pacote *cluster* e a versão do algoritmo utilizada é a de Kaufman e Rousseeuw (1990b) para esse representante de soft *clustering*. Na lógica *Fuzzy* cada elemento tem uma probabilidade de pertencer a cada *cluster*, sendo diferente do *k-means* e do *k-medoids*, por exemplo, onde cada objeto pertence a exatamente um *cluster*. O Gráfico exibido na Figura 5.12 mostra o resultado obtido ao se analisar os dados pelo algoritmo *Fuzzy* usando a função do R fviz_cluster().

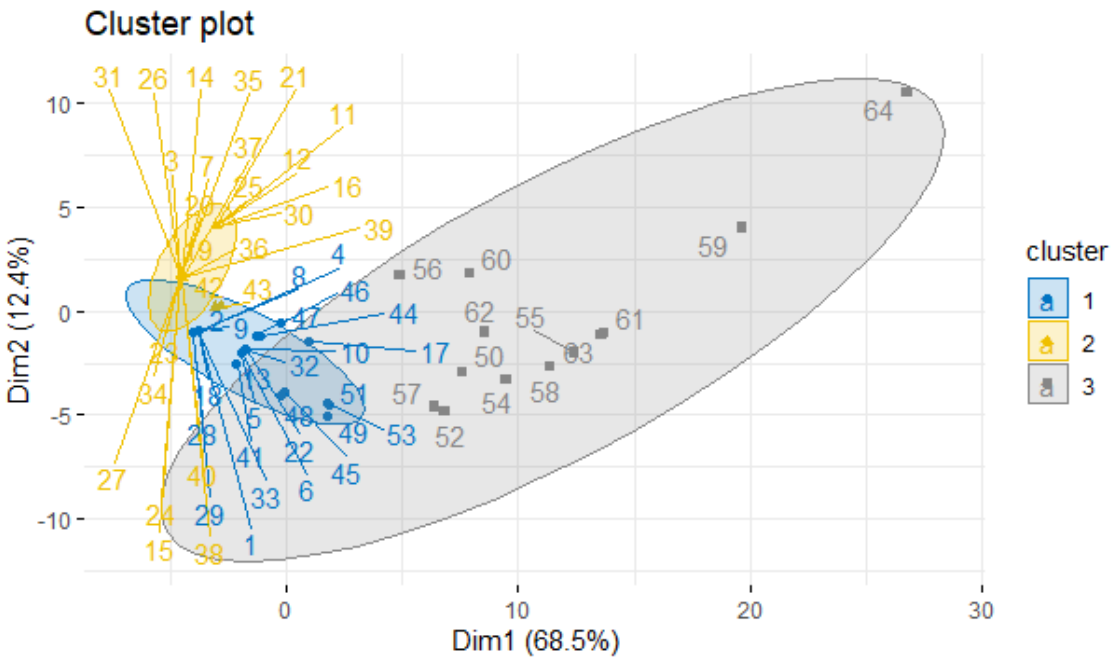

Figura 5.12: Gráfico de Dispersão dos resultados da clusterização pelo método Lógica Fuzzy (3 *Clusters*)
Fonte: Autoria Própria

Os três *clusters* são os exibidos nas cores azul, amarelo e roxo e cada dado apresenta uma chance de pertencer a qualquer um dos *clusters*, mas são identificados no de maior chance.

5.4.8 *Hierarchical k-means clustering* - HKM

A clusterização hierárquica *K-Means* é uma versão mais eficiente do famoso *k-means* e no R foi desenvolvida utilizando a função hkmeans() na biblioteca *factoextra*. A Figura 5.13 representa um dendrograma obtido pela execução do HKMeans nos dados estudados utilizando a função fviz_dend() no R. A Figura 5.14 apresenta os resultados da clusterização

por meio do método PCA para 3 *clusters* do algoritmo HKMEANS. Gráfico desenvolvido com a função fviz_cluster() do pacote factoextra.

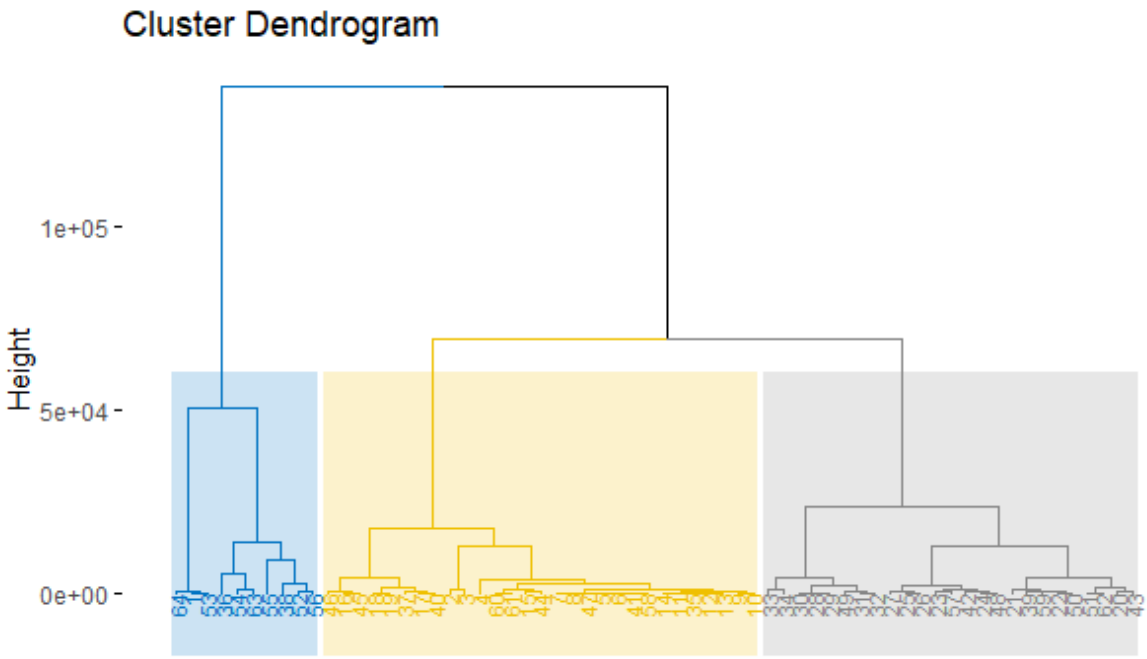

Figura 5.13: Dendrograma dos resultados da Clusterização pelo método HKMEANS (3 Clusters)

Fonte: Autoria Própria

Observa-se que as cores roxa, azul e amarela representam os três *cluster*s encontrados, mas que se pode notar maiores similaridades entre determinados dados pertencentes ao mesmo *cluster* em detrimento de outros.

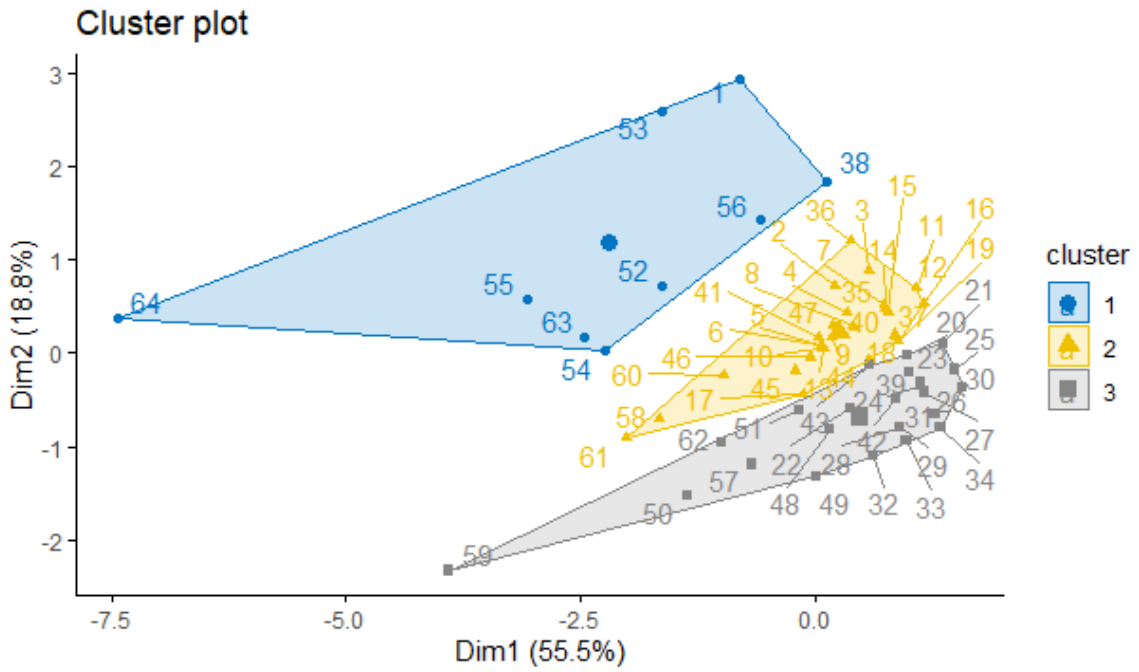

Figura 5.14: Gráfico de Dispersão dos resultados da clusterização pelo método HKMEANS (3 *Clusters*)

Fonte: Autoria Própria

A Figura 5.14 exibe os 3 *clusters* obtidos e eles apresentam aspecto análogo ao dendrograma da Figura 5.13.

5.4.9 *Model-Based Clustering*

Para desenvolver o MBC utilizou-se no R a biblioteca Mclust() no pacote mclust. O Gráfico da Figura 5.15 mostra as estimativas dos parâmetros utilizando um algoritmo chamado *Expectation-Maximization* (EM), que retorna uma sigla com três letras, a primeira letra se refere ao volume, a segunda a forma e a terceira a orientação dos *clusters*. Essa letra pode ser E quando representar igualdade, V quando for variável e I se forem eixos coordenados.

O melhor modelo é escolhido usando *Bayesian Information Criterion* (BIC), uma pontuação BIC alta indica forte evidência para o modelo correspondente. No caso do resultado obtido VII, representa um modelo que o volume de todos os *clusters* é variável, a forma e a orientação é a identidade (I) ou eixos coordenados. Esse é o único caso do exemplo que o próprio algoritmo já considera a melhor cotação de quantidade de clusters otimizadas e no caso encontrou 9 de acordo com o método BIC.

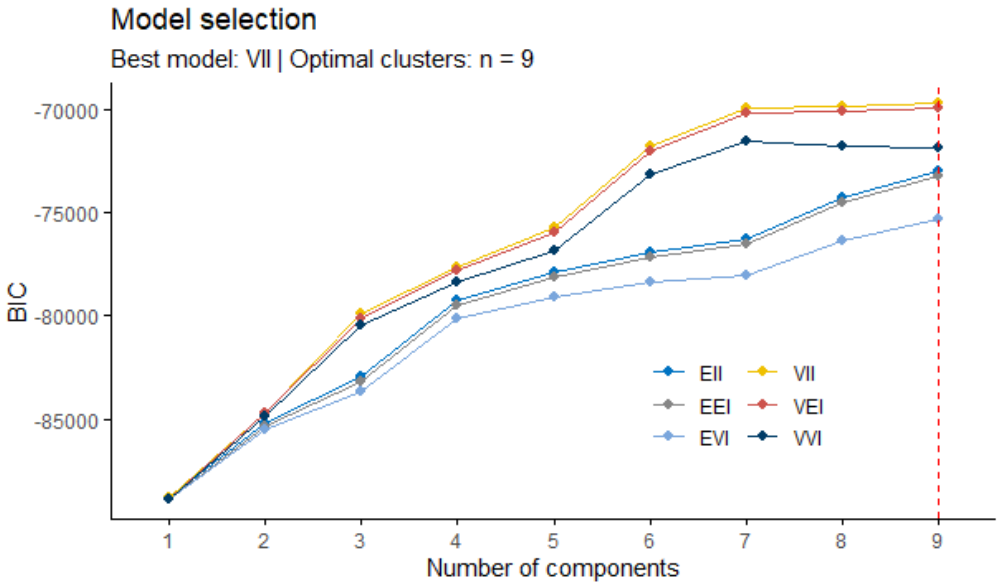

Figura 5.15: Valores BIC usados para encontrar o número de *clusters*

Fonte: Autoria Própria

A Figura 5.16 representa um gráfico de dispersão da clusterização definida anteriormente. Como esse algoritmo foi destoante dos demais e gerou 9 *clusters*, eles são representados de forma diferente, além da utilização de cores para representação de *clusters*, também foi utilizado um símbolo para representar dados de cada um deles. Infelizmente o algoritmo não permite a impressão de nomes no gráfico gerado.

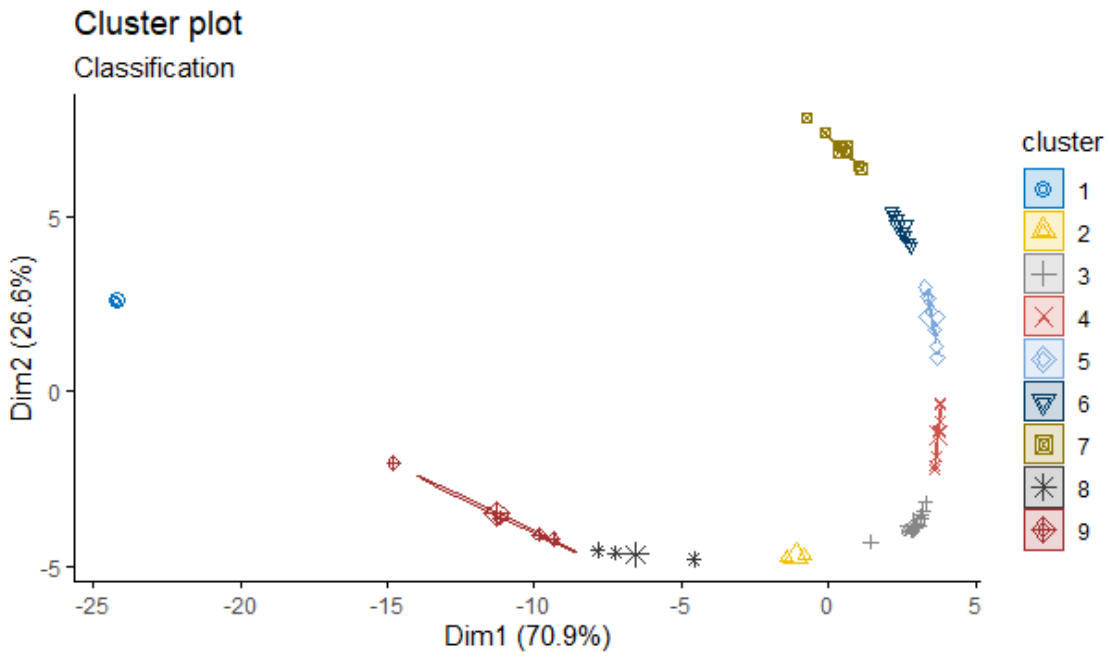

Figura 5.16: Gráfico de Dispersão dos resultados da clusterização pelo método MBC (9 *Clusters*)

Fonte: Autoria Própria

Após tudo isso percebe-se que a intenção de conclusão do primeiro objetivo específico de se analisar o comportamento dos algoritmos clássicos de clusterização quando recebem como entradas os dados reais dos fornecedores pôde ser realizada.

5.5 Implantação

Seguindo com as etapas de CRISP-DM, tem-se a etapa de implantação. Nesta etapa acontece a validação dos resultados obtidos na pesquisa, e para a validação da clusterização podem ser utilizados índices de validação interna e de estabilidade para definir qual dos modelos propostos mais se encaixa nos dados experimentais e qual a quantidade ideal de *clusters*.

5.5.1 Validações da clusterização

O R apresenta um pacote chamado clValid com a função clValid(), que compara os algoritmos propostos e define com base em indicadores quais as melhores opções. Essa função permite que sejam adicionados os métodos de clusterização que se tem interesse em testar e também uma quantidade inteira de *clusters* para servirem de opções.

As validações são realizadas em duas etapas, a primeira é chamada de validação interna e se utiliza dos índices *Silhouette*, Dunn e conectividade. As saídas do processo consistem dos três algoritmos que tiveram melhor desempenho e as quantidades otimizadas de *clusters*, um para cada índice. O algoritmo e a quantidade de *clusters* pode se repetir ou não para os diferentes índices.

Na segunda, tem-se uma validação de estabilidade utilizando-se dos índices *Average Proportion of non-overlap* (APN), *Average Distance between means(*ADM) e *Figure of Merit* (FOM) que variam de 0 a 1. Resultados mais próximos de zero tornam mais consistente a clusterização. Também se utiliza o índice *Average Distance* (AD) que varia de 0 a infinito com os melhores resultados sendo os mais próximos de zero.

De acordo com a documentação da biblioteca, APN mede a proporção média das observações não colocadas no mesmo *cluster* pela clusterização baseada nos dados totais e também na clusterização com uma coluna removida. ADM mede a distância média entre os centros de *cluster* para observações colocadas no mesmo *cluster* com o conjunto de dados totais e com a remoção de uma coluna. FOM mede a variância intra-*cluster* média da coluna deletada, onde a clusterização é baseada nas colunas não deletadas. AD mede a distância

média entre as observações colocadas no mesmo *cluster* com o conjunto de dados totais e com a remoção de uma coluna.

Para as validações internas e de estabilidade testaram-se de 2 a 6 *clusters*. A Tabela 5.2 apresenta os resultados obtidos pelos algoritmos estudados para as validações internas e a Tabela 5.3 as validações de estabilidade.

Tabela 5.2: Validações internas das clusterizações

VALIDAÇÃO INTERNA

Métodos de Clusterização:
Hierarchical, kmeans, pam, diana, fuzzy, model based, clara e agnes

Tamanhos de clusters:
2 3 4 5 6

Validações:	Método	2	3	4	5	6
hierarchical	Conectividade	4,3579	8,2159	9,7159	15,1734	20,6821
	Dunn	0,6222	0,5165	0,5165	0,297	0,2722
	Silhouette	0,6791	0,5111	0,4956	0,4356	0,4175
k-means	Conectividade	4,3579	14,1111	22,6044	24,1044	26,4897
	Dunn	0,6222	0,0546	0,1597	0,1597	0,2141
	Silhouette	0,6791	0,3532	0,4208	0,4093	0,3952
pam	Conectividade	17,1246	29,0056	30,3333	34,5663	37,1425
	Dunn	0,0773	0,029	0,0291	0,0559	0,0602
	Silhouette	0,4234	0,2473	0,2337	0,2793	0,325
diana	Conectividade	4,3579	14,1111	24,9933	27,5694	31,2651
	Dunn	0,6222	0,0546	0,0573	0,0644	0,0689
	Silhouette	0,6791	0,3532	0,3982	0,3802	0,3759
fuzzy	Conectividade	24,9988	NA	NA	NA	NA
	Dunn	0,0589	NA	NA	NA	NA
	Silhouette	0,2829	NA	NA	NA	NA
model based	Conectividade	19,2028	30,9044	40,3071	46,8409	53,7409
	Dunn	0,0588	0,0628	0,0582	0,0291	0,029
	Silhouette	0,3215	0,2824	0,1044	0,1137	0,1802
clara	Conectividade	21,2218	19,879	21,3909	30,0619	36,1944
	Dunn	0,0773	0,0291	0,0523	0,0559	0,1204
	Silhouette	0,3935	0,2393	0,2676	0,2852	0,3263
agnes	Conectividade	4,3579	8,2159	9,7159	15,1734	20,6821
	Dunn	0,6222	0,5165	0,5165	0,297	0,2722
	Silhouette	0,6791	0,5111	0,4956	0,4356	0,4175

Fonte: Autoria Própria

Tabela 5.3: Validações de estabilidade das clusterizações

VALIDAÇÃO DE ESTABILIDADE

Métodos de Clusterização:
Hierarchical kmeans pam diana fuzzy model based clara agnes

Tamanhos de clusters:
2 3 4 5 6

Validações:	Método	2	3	4	5	6
hierarchical	APN	0,0151	0,037	0,1306	0,2255	0,174
	AD	2,079	1,9541	1,8916	1,7254	1,5038
	ADM	0,1256	0,2693	0,4255	0,6078	0,5994
	FOM	0,9574	0,9084	0,9086	0,8789	0,8505
k-means	APN	0,0655	0,1889	0,122	0,1686	0,2206
	AD	2,0785	1,9198	1,6484	1,5439	1,4753
	ADM	0,2077	0,6905	0,5554	0,5329	0,6728
	FOM	0,9572	0,9267	0,8953	0,8901	0,8694
pam	APN	0,1375	0,1844	0,2111	0,2901	0,3398
	AD	1,9504	1,7375	1,6196	1,481	1,3953
	ADM	0,3666	0,4583	0,5527	0,6639	0,6848
	FOM	0,9228	0,9165	0,9099	0,8908	0,8809
diana	APN	0,0707	0,1318	0,105	0,1072	0,1764
	AD	2,0783	1,8512	1,5881	1,5251	1,4173
	ADM	0,2129	0,4612	0,4668	0,5429	0,5362
	FOM	0,9571	0,9362	0,8985	0,8853	0,8531
fuzzy	APN	0,2536	NA	NA	NA	NA
	AD	2,074	NA	NA	NA	NA
	ADM	0,6132	NA	NA	NA	NA
	FOM	0,9663	NA	NA	NA	NA
model based	APN	0,0609	0,2915	0,3315	0,4075	0,3797
	AD	1,9557	1,9136	1,8475	1,8561	1,6911
	ADM	0,226	0,784	0,8208	0,977	0,9506
	FOM	0,9178	0,9068	0,8974	0,8873	0,8757
clara	APN	0,1739	0,2393	0,3412	0,3035	0,4012
	AD	1,9843	1,803	1,6715	1,4972	1,4154
	ADM	0,4443	0,6195	0,8366	0,6784	0,7455
	FOM	0,9313	0,9336	0,8925	0,8945	0,8752
agnes	APN	0,0151	0,037	0,1306	0,2255	0,174
	AD	2,079	1,9541	1,8916	1,7254	1,5038
	ADM	0,1256	0,2693	0,4255	0,6078	0,5994
	FOM	0,9574	0,9084	0,9086	0,8789	0,8505

Fonte: Autoria Própria

A Tabela 5.4 apresenta os melhores algoritmos selecionados pela função do R para a validação interna e a Tabela 5.5 é análoga, mas para as validações de estabilidade.

Tabela 5.4: Melhores algoritmos para as validações internas

	Pontuação	Método	Clusters
Conectividade	4,3579	hierarchical	2
Dunn	0,6222	hierarchical	2
Silhouette	0,6791	hierarchical	2

Fonte: Autoria Própria

Tabela 5.5: Melhores algoritmos para as validações de estabilidade

	Pontuação	Método	Clusters
APN	0,0151	hierachical	2
AD	1,3953	pam	6
ADM	0,1256	hierachical	2
FOM	0,8505	hierachical	6

Fonte: Autoria Própria

O resultado dos três índices para a validação interna, tanto para o dunn, conectividade e *silhouette*, foi o algoritmo hierárquico com 2 *clusters*. Todos esses índices foram estudados e demonstrados no Capítulo 3.

Para uma reflexão sobre o assunto, se o índice *silhouette* corresponder a um valor mais alto, melhor é o resultado da clusterização, se for menor do que zero provavelmente apresenta dados nos *clusters* errados e se for muito próximo de zero apresenta dados que estão dentro de dois *clusters*. Para o *hierarchical k-means* com 2 *clusters* o índice *silhouette* foi de 0,6222, um valor alto mostrando uma clusterização bem estabelecida. Para o índice Dunn, quanto maior o resultado melhor é a clusterização e teve resultado 0,6791 mostrando que para esse quesito o algoritmo selecionado também foi eficiente. E para a conectividade, o índice varia de 0 a infinito então o resultado 4,3579 foi um resultado interessante. Para os índices de estabilidade, quanto mais próximos de zero melhor é o resultado da clusterização.

Concluiu-se para a validação de estabilidade, que para a base de dados estudada, o melhores resultados são do algoritmo *Hierarchical K-Means* com 2 *clusters* que apresentou o índice APN 0,0151 e ADM 0,1256, do algoritmo *hierarchical K-means* com 6 *clusters* de índice FOM 0,8505 e para o índice AD 1,3953 o modelo PAM com 6 *clusters*.

Em uma análise geral dos resultados, o algoritmo *hierarchical k-means* com 2 *clusters* foi o mais atrativo para a base de dados estudada porque nas análises de validações internas obteve o melhor valor para todos os índices e nas validações externas o melhor para dois dos quatro índices, ou seja, o único algoritmo aprovado em ambas as validações mostradas nas Tabelas 5.4 e 5.5.

5.5.2 Melhores algoritmos

Estas clusterizações encontradas nos resultados das validações foram desenvolvidas no R usando as mesmas bibliotecas e funções utilizadas ao se explicar o resultado dos modelos para 3 *clusters* e portanto não serão reiteradas essas informações técnicas em detalhes. E no caso do *hierarchical k-mean*, apresentando também seus respectivos dendrogramas gerados pelas mesmas funções.

5.5.2.1 *Hierarchical K-Means* com 2 *clusters*

O *Hierchical K-means* com 2 *clusters* foi a opção definida como a melhor para todos os índices de validação interna e também pelos índices APN e ADM de validação de estabilidade. As Figuras 5.17 e 5.18 representam a clusterização e o dendrograma respectivamente.

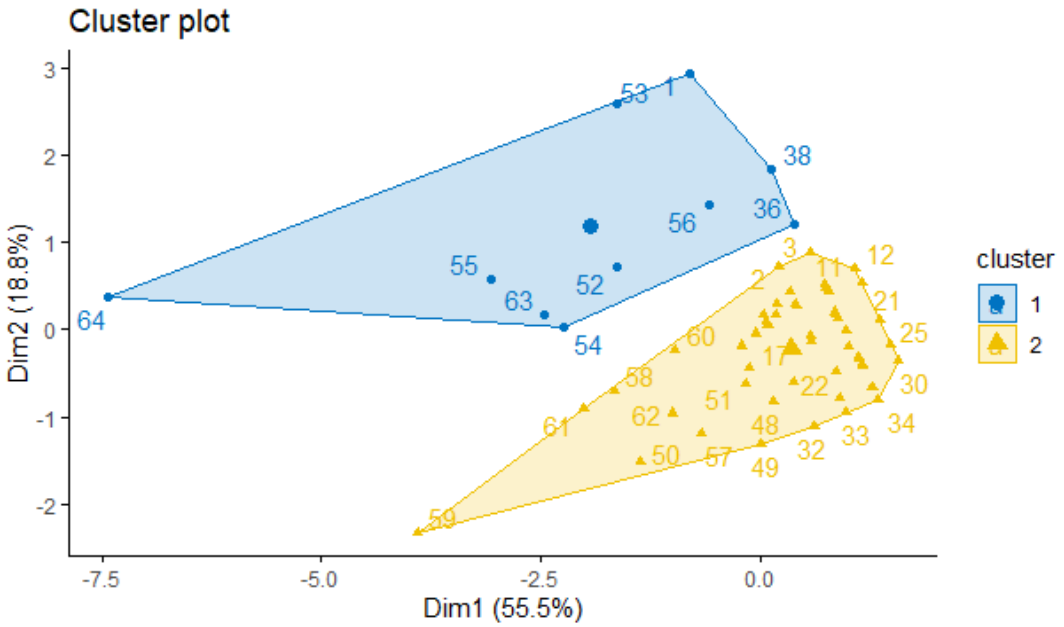

Figura 5.17: Gráfico de Dispersão dos resultados da clusterização pelo método HKMEANS (2 *Clusters*)

Fonte: Autoria Própria

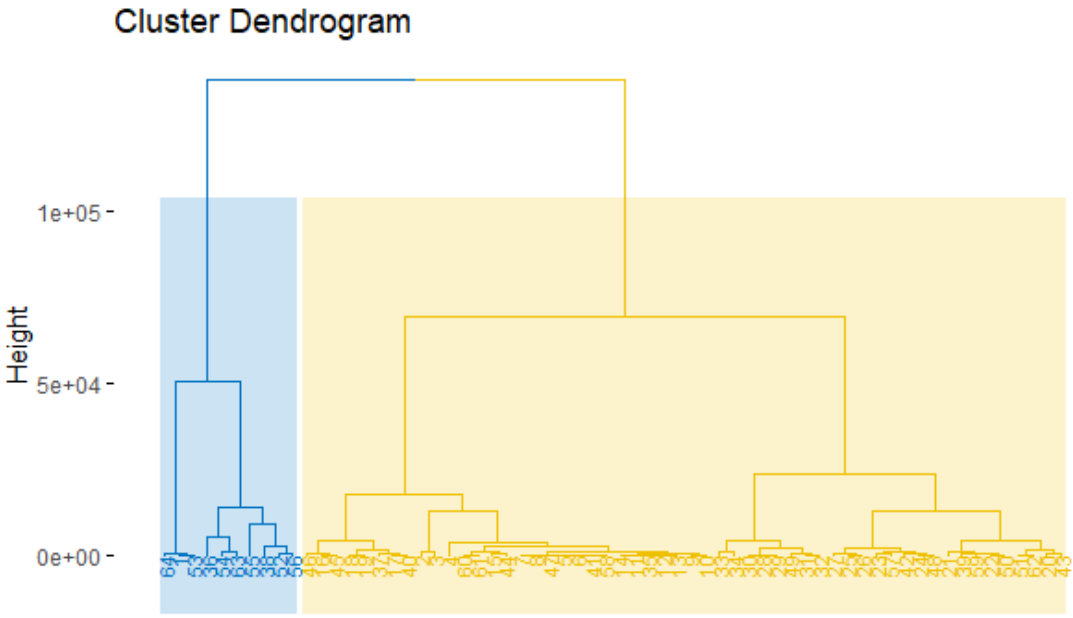

Figura 5.18: Dendrograma dos resultados da Clusterização pelo método HKMEANS (2 *Clusters*)

Fonte: Autoria Própria

5.5.2.2 *Hierarchical K-means* com 6 *clusters*

O Hierarchical *K-Means* com 6 *clusters* foi a melhor opção considerada pelo índice de validação de estabilidade FOM e o resultado está representado na Figura 5.19.

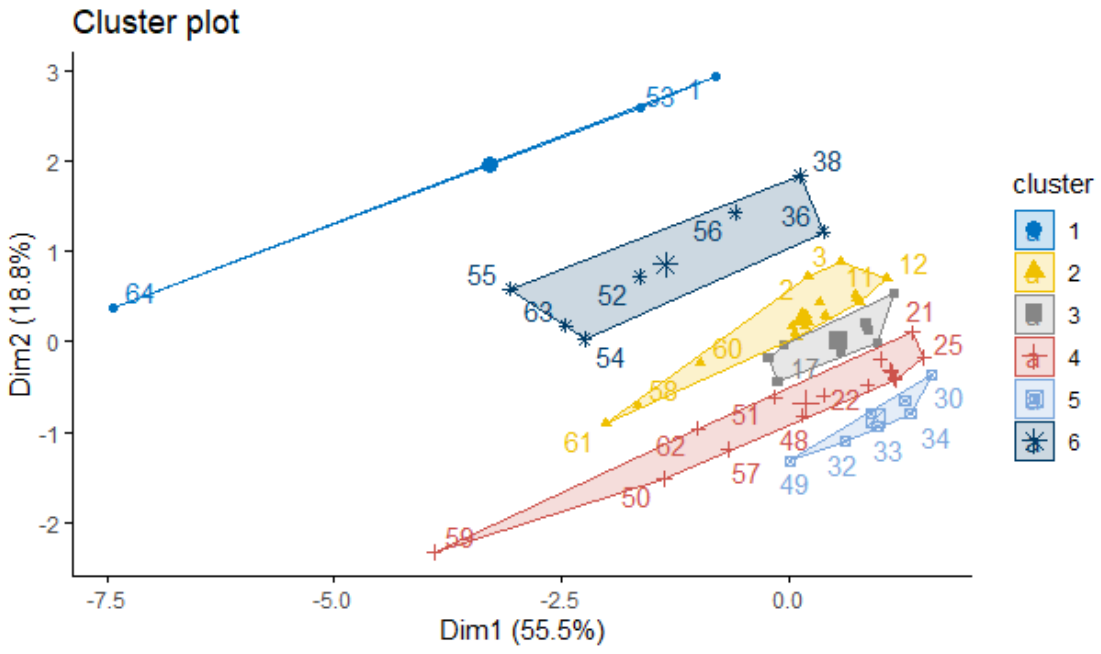

Figura 5.19: Gráfico de Dispersão dos resultados da clusterização pelo método HKMEANS (6 *Clusters*)

Fonte: Autoria Própria

O dendrograma dessa clusterização está representado na Figura 5.20.

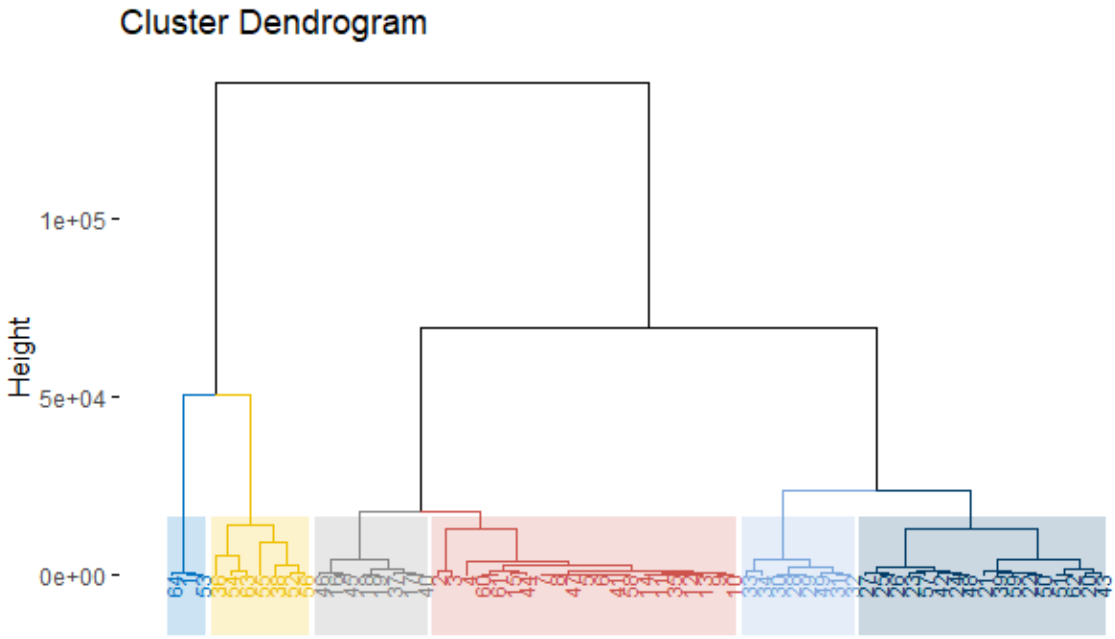

Figura 5.20: Dendrograma dos resultados da Clusterização pelo método HKMEANS (2 *Clusters*)

Fonte: Autoria Própria

5.5.2.3 PAM com 2 *clusters*

O PAM com dois *clusters* foi a melhor opção encontrada pelo índice de estabilidade AD e está representado na Figura 5.21.

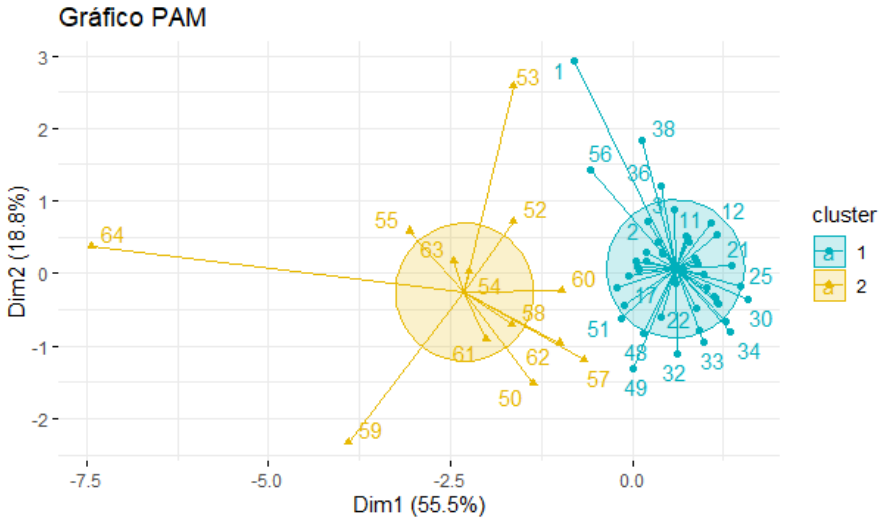

Figura 5.21: Gráfico de Dispersão dos resultados da clusterização pelo método PAM (2 *Clusters*)

Fonte: Autoria Própria

5.6 Análise dos resultados

Com as análises expostas acima demonstra-se como a aplicação do procedimento CRISP-DM foi utilizada nessa pesquisa. A primeira etapa da compreensão do negócio possibilitou que uma exploração do problema de pesquisa fosse contextualizada aos negócios da empresa e permitiu encontrar como a ciência dos dados poderia ser utilizada para procurar soluções para os problemas de classificação e seleção dos fornecedores.

A compreensão dos dados possibilitou um entendimento melhor das fontes de dados utilizadas, o formato dos dados (e posteriormente por tentativa e erro em quais formatos os dados deveriam estar) e a remoção dos nomes dos fornecedores para manutenção do sigilo dos dados da empresa. Em seguida, a preparação dos dados permitiu entender como devem se comportar os dados para que possam ser utilizados nas ferramentas de clusterização, além de um entendimento da tendência de clusterização que culmina com um entendimento da viabilidade ou não dos dados estudados para a clusterização.

Em seguida, a etapa de modelagem permitiu realizar a construção dos modelos estudados na teoria e verificar como se aplicam aos dados reais culminando com a etapa de implantação que possibilitou uma validação dos modelos e quais algoritmos apresentam melhor eficiência para os fornecedores. Vale destacar que essa pesquisa objetivava aplicar *Data Science* para auxiliar aos tomadores de decisão e entra no escopo dessa pesquisa explicar de quais formas isso é possível.

Por meio de uma análise de *clusters* pode-se encontrar fornecedores que apresentam um alto valor de inconformidades e atuam em um país específico por exemplo.

De um ponto de vista prático, com essas ferramentas o gestor pode tomar decisões de forma mais fácil porque pode decidir em qual grupo:

- Investir recursos para auditar in loco ou não;
- Atuar para reduzir não conformidades;
- Trabalhar para melhorar o relacionamento com aqueles que são mais representativos, etc.

Em outras palavras, a empresa pode atuar, não apenas empiricamente, direcionando adequadamente os esforços para cada grupo. Tendo esse tipo de informação a empresa pode aplicar técnicas de desenvolvimento de fornecedores para algum grupo, como avaliar as dificuldades que os fornecedores estão enfrentando, realizar auditorias in loco ou visitas para

conhecer as bases operacionais de seus parceiros, investir em operações dos fornecedores, entre outras abordagens.

Da mesma forma que pode permitir que menos recurso sejam direcionado para parceiros específicos que apresentam baixo índice de não-conformidades ou que a companhia apresente pouco contato comercial. Também pode-se procurar indícios que apresentam alguma concentração territorial em comum para fornecedores que precisam de atenção, por exemplo, otimizando seus recursos na decisão de auditorias para o mesmo período.

De maneira geral, pode contribuir com *insights* relevantes para os tomadores de decisão que poderão atuar buscando melhorar o gerenciamento de seus fornecedores da melhor forma possível.

6 PROPOSTA DO *FRAMEWORK*

Depois de efetuadas todas as etapas do CRISP-DM pôde-se considerar o ***segundo objetivo específico*** desta pesquisa que consiste na apresentação de um *framework* para auxiliar futuras pesquisas com uso de clusterização. E, com base nas experiências adquiridas no desenvolvimento do estudo, foi proposto um método para clusterização que pode ser utilizado para futuras análises de situações não supervisionadas quando for necessário realizar alguma segmentação em grupos.

Para sua utilização é necessário certa experiência em linguagem R. É essencial um conhecimento prévio de funções, importações de bibliotecas, codificação, atribuições de variáveis e conhecimento de tipos de variáveis e conversões de formato. Se espera também discernimento de inserção de bancos de dados no *RStudio*.

A aplicação do CRISP-DM permitiu que conclusões fossem tiradas e que essa nova metodologia fosse criada. Caso uma nova pesquisa venha a ser desenvolvida sobre clusterização, ela pode ser utilizada de forma direta sem a necessidade de um procedimento genérico de *Data Science* se o foco for direcionado na aplicação e validação de clusterização.

Para sua realização são necessárias as etapas definidas na Figura 6.1. Basicamente, as etapas são definidas pelo fluxograma que também indica as possíveis transições entre elas. Lembrando que o processo é cíclico, afinal, os dados, em sua grande maioria, não deixam de ser atualizados e uma nova clusterização deve acontecer eventualmente.

Uma descrição do que deve ser feito em cada uma delas se encontra a seguir:

- Escolha dos dados: Com o alto volume de dados que existe atualmente faz-se necessário selecionar primeiramente quais serão utilizados para a realização da clusterização. Cabe perguntar se são dados que tem significado para as entidades estudadas. Essa escolha não fica presa a uma única fonte de dados, caso exista essa possibilidade. Além disso, nessa etapa é feita uma contextualização dos dados escolhidos com o problema estudado no contexto da organização/instituição para que antes de se realizar a codificação em si tenha-se uma ideia prévia da viabilidade;

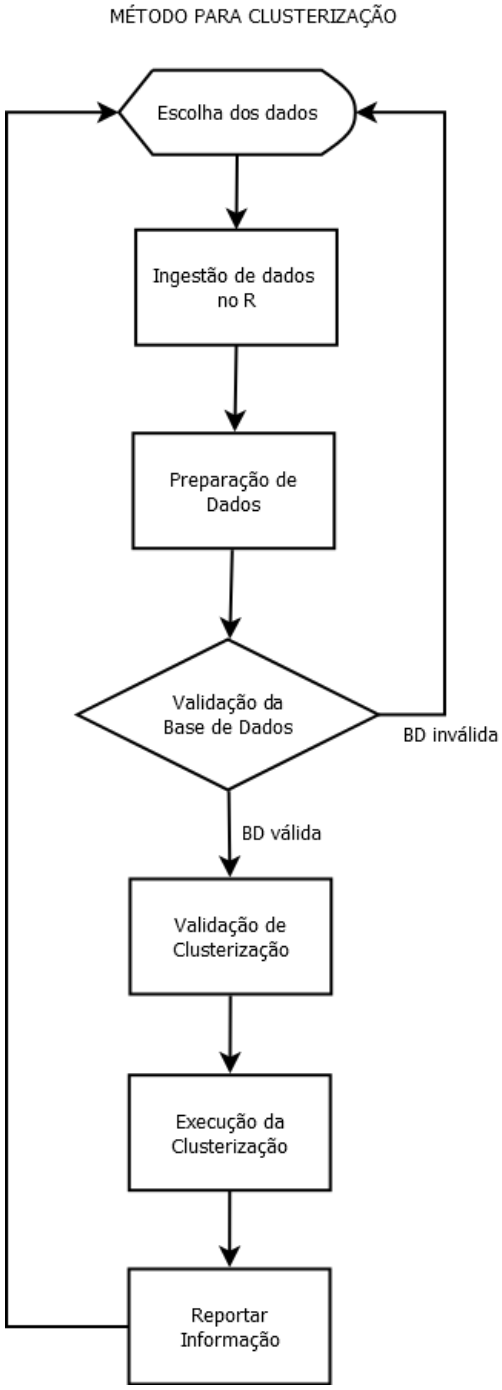

Figura 6.1: Método para clusterização

Fonte: Autoria Própria

- Ingestão de dados no R: Os dados selecionados, sejam de uma única fonte ou de fontes variadas, precisam ser enviados ao *RStudio* para que possa ser realizada a clusterização. Tudo que foi definido anteriormente irá compor a base de dados utilizada para o estudo;

- Preparação de Dados: Em posse de todos os dados, é necessário manipulações básicas para evitar dados faltantes, remoção dos pontos fora da curva e remoção de dados duplicados. É nessa etapa que também deve acontecer a normalização dos dados, ou seja, colocá-los em uma escala comum para que possam ser comparáveis;
- Validação da Base de Dados: Essa etapa consiste em utilizar a estatística de Hopkins e/ou o método de análise visual de tendência de clusterização para validar se os dados selecionados e refinados são uma escolha interessante para a realização de uma clusterização. O pacote clustertend e a função Hopkins() do R que pertence a esse pacote realizam essa validação dos dados selecionados e também a função dist() do pacote factoextra para o método visual. Se for confirmado que existe um potencial para a clusterização segue-se para a validação de clusterização, caso contrário, faz-se necessário retornar para a escolha de uma nova fonte de dados ou avaliação das fontes de dados atuais e recomeçar o ciclo;
- Validação de Clusterização: Para essa etapa utiliza-se o pacote do R NbClust e a função NbClust() para realizar as validações internas e de estabilidade que vão permitir a escolha dos algoritmos mais eficientes para a clusterização;
- Execução da Clusterização: Nessa etapa, utilizam-se as informações obtidas na etapa anterior e realiza-se a clusterização com esses parâmetros obtidos dos algoritmos considerados mais eficientes;
- Reportar informações: Nessa etapa, as clusterizações obtidas já foram realizadas e precisam ser apresentadas em formato visual para que sejam úteis para prover auxílio aos tomadores de decisão. Esses formatos obtidos podem ser gráficos, dendrogramas ou *heatmaps*.

O método (framework) foi construído se baseando nas literaturas estudadas nessa pesquisa e permite que a clusterização com dados reais seja mais intuitiva e direcionada pelos estudiosos e cientistas de dados. Esta proposta de método é resultado da experiência final da aplicação da metodologia CRISP-DM e sua criação se desenvolveu após várias tentativas de erros e acertos.

Espera-se que a prática de clusterização pelos novos pesquisadores seja mais difundida e que contribua para o desenvolvimento científico no país e no mundo.

7 CONCLUSÕES

O objetivo geral da pesquisa consistia na utilização da técnica de *Machine Learning* no contexto da *Data Science* na classificação e seleção de fornecedores da empresa estudada e foram utilizados os algoritmos de clusterização que fazem parte de *Machine Learning* para segmentação dos fornecedores da organização. Essa aplicação, em convergência com o que a pesquisa se propunha a realizar, culminou com *reports* visuais e a expectativa de que possam ser úteis aos tomadores de decisão.

Em posse dessas informações os tomadores de decisão vão poder procurar informações que estavam escondidas, tais como fornecedores que acabaram sendo sobrepostos pela visão humana e foram deixados de lado em detrimento de outros que pareciam melhores num primeiro momento.

A utilização de clusterização no gerenciamento da classificação e seleção de fornecedores não objetiva a exclusão da *expertise* humana de forma alguma, muito pelo contrário, recomenda-se que sejam utilizadas em conjunto para oferecerem possíveis novos *insights*. Espera-se que essa aliança de conhecimentos contribua positivamente para um bom gerenciamento da organização e que a tecnologia seja aplicada de forma positiva em prol da organização.

O primeiro objetivo específico consistia em uma análise do funcionamento e do comportamento dos algoritmos clássicos de clusterização na base de dados real e o procedimento CRISP-DM ajudou na conclusão desse objetivo.

Pôde-se concluir que para a base de dados em estudo a aplicação dos algoritmos é possível já que utilizando-se de dois métodos para validação da tendência para clusterização apresentaram resultados satisfatórios nesse quesito. Por meio das etapas do procedimento pôde-se encontrar formas de se aplicar em um contexto científico as validações necessárias para comparação entre os algoritmos clássicos culminando com *Hierarchical K*-Means com 2 *clusters* sendo o mais eficiente computacionalmente pela análise das validações internas e de estabilidade.

Dessa forma, partindo-se dessa confirmação encontrou-se o funcionamento e como se comportam cada um dos algoritmos quando aplicados aos dados dos fornecedores e como atuam quando comparados entre si, ocasionando na percepção que o segundo objetivo específico também era passível de conclusão.

O segundo objetivo específico da pesquisa era analisar os algoritmos de clusterização em busca do mais apropriado para a base de dados dos fornecedores criando um *framework* que pudesse ser seguido para realização desse tipo de análise.

Com base na literatura estudada, foram definidas etapas em um procedimento que pode ser replicado utilizando linguagem R e *RStudio*, e que permite comparação entre os algoritmos de clusterização independente do foco da pesquisa, ou estudo, que tenha como objetivo separar em *clusters* algum conjunto de dados. Pode-se concluir que existiu a viabilidade da criação desse *framework* e que ele pode ser replicado desde que os pesquisadores atendam às exigências de conhecimento necessárias para seu desenvolvimento.

Aplicando o processo da pesquisa, a empresa que foi objeto de estudo pode apresentar uma considerável redução de custo e tempo na cadeia de suprimentos se encontrar focos de insight como demonstrado na análise de resultados do Capítulo 5. Os custos relacionados às horas de trabalho para gerenciar os processos dos fornecedores e de executar auditorias pode ser reduzido, além da redução de tempo investido em fornecedores que não precisam de tanto investimento. Além do mais, o processo foi melhorado por se mostrar mais simplificado. Dessa forma, acaba sendo uma contribuição de grande valia para a engenharia de produção.

Conclui-se que essa contribuição pode agregar bastante para futuros estudos e análises práticas de empresas que podem seguir o *framework* definido buscando encontrar qual a melhor forma de realizar a clusterização de seus dados. Dessa forma, conclui-se que os objetivos propostos foram atingidos de forma satisfatória. A engenharia de produção está cada vez mais focada na indústria 4.0 e pode-se perceber que a pesquisa e suas possíveis replicações em trabalhos futuros se encaixam bastante nos novos métodos e necessidades desse novo contexto mundial e a utilização da *Data Science* para resolver problemas existentes atualmente se mostrou possível com os resultados atingidos e abrem mais espaço para a tomada de decisão baseada em dados.

7.1 Sugestões para Pesquisas Futuras

Uma sugestão interessante para uma nova pesquisa utilizando-se o *framework* criado consiste no desenvolvimento de uma clusterização em grandes quantidades de dados para uma verificação do comportamento do método quando se depara com essas particularidades do *Big Data* e que também permita uma visualização prática das diferenças entre os métodos PAM e CLARA.

Outra sugestão é que o princípio utilizado aqui pode ser replicado para outros problemas enfrentados pela ciência de dados, como a classificação ou previsão. Dessa forma, outros *frameworks* podem ser criados e também contribuírem com o avanço científico da *Data Science*.

Também é possível adaptar essa pesquisa que teve seu foco nos algoritmos mais clássicos do assunto e investigar novas formas que permitam que novos métodos possam ser comparados também e que deixem a ferramenta ainda mais robusta, afinal, o assunto é bastante difundido e trabalhado atualmente.

Existe outra possibilidade que também segue essa linha, a seleção de um dos algoritmos clássicos e a realização de um estudo direcionado em busca de oportunidades de melhoria. O assunto é de interesse da comunidade acadêmica e podem surgir muitas contribuições que agregariam conhecimento por meio de uma pesquisa bem direcionada.

Outra possível pesquisa futura consiste na utilização das ferramentas criadas nessa pesquisa indo além da aplicação, buscando efetivamente agregar conhecimento por meio da tomada de decisão. Os resultados da clusterização podem ser utilizados em parceria com alguma técnica estruturada para organizar e analisar decisões complexas buscando obter conhecimento. Uma pesquisa assim pode mostrar a validação prática da aplicação do *framework.*

Os pontos destacados derivam apenas de um conjunto de *insights* observados pelo autor no decorrer desta pesquisa em meio a uma gama de oportunidades e desafios presentes na literatura.

REFERÊNCIAS BIBLIOGRÁFICAS

ABBASI, A.; SARKER, S.; CHIANG, R. H. L. Big data research in information systems: Toward an inclusive research agenda. **Journal of the Association for Information Systems**, v. 17, n. 2, p. 1–32, 2016.

ABOUBI, Y.; DRIAS, H.; KAMEL, N. BAT-CLARA: BAT-inspired algorithm for Clustering LARge Applications. **IFAC-PapersOnLine**, v. 49, n. 12, p. 243–248, 2016.

AGARWAL, R.; DHAR, V. Big data, data science, and analytics: The opportunity and challenge for IS research. **Information Systems Research**, v. 25, n. 3, p. 443–448, 2014.

AGGARWAL, C. **Data Mining: The Textbook**. 1. ed. New York, USA: Springer International Publishing Switzerland, 2015. v. 14

ALPAYDIN, E. **Introduction to Machine Learning Second Edition**. 2. ed. Cambridge, Massachusetts: The MIT Press, 2010. v. 1107

ALPAYDIN, E. **Introduction to machine learning**. 3. ed. Cambridge, Mass.: MIT Press, 2014.

ALPAYDIN, E. **Machine learning : the new AI**. Cambridge, Mass.: Massachusetts Institute of Technology, 2016.

AMERSHI, S. et al. **Software Engineering for Machine Learning: A Case Study**. Proceedings - 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice, ICSE-SEIP 2019. **Anais**...IEEE, 2019

AMIRIAN, P.; LANG, T.; LOGGERENBERG, F. **Big data in healthcare : extracting knowledge from point-of-care machines**. 1. ed. Cham, Switzerland: SpringerBriefs in Pharmaceutical Science & Drug Development, 2017.

ARIA, M.; CUCCURULLO, C. bibliometrix: An R-tool for comprehensive science mapping analysis. **Journal of Informetrics**, v. 11, n. 4, p. 959–975, 2017.

BAESENS, B. et al. Transformation Issues of Big Data and Analytics in Networked Business. **Mis Quartely**, v. 40, n. 4, p. 807–818, 2016.

BALLOU, B.; HEITGER, D. L.; STOEL, D. Data-driven decision-making and its impact on accounting undergraduate curriculum. **Journal of Accounting Education**, v. 44, n. July 2017, p. 14–24, 2018.

BANAEIAN, N. et al. Green supplier selection using fuzzy group decision making methods: A case study from the agri-food industry. **Computers and Operations Research**, v. 89, p. 337–347, 2018.

BERTRAND, J. W. M.; FRANSOO, J. C. Operations management research methodologies using quantitative modeling. **International Journal of Operations and Production Management**, v. 22, n. 2, p. 241–264, 2002.

BOTVINIK-NEZER, R. et al. Variability in the analysis of a single neuroimaging dataset by many teams. **Nature**, v. 582, n. 7810, p. 84–88, 2020.

BUJACK, R. et al. The Good, the Bad, and the Ugly: A Theoretical Framework for the Assessment of Continuous Colormaps. **IEEE Transactions on Visualization and Computer Graphics**, v. 24, n. 1, p. 923–933, 2018.

CAMARGOS, R. C.; DO CARMO NICOLETTI, M. Three case studies using agglomerative clustering. **Advances in Intelligent Systems and Computing**, v. 557, p. 67–76, 2017.

CAO, L.; YU, P. S. **Data Analytics Series editors**. 1. ed. NSW, Australia: Springer International Publishing, 2018.

CAO, M.; ZHANG, Q. Supply chain collaboration: Impact on collaborative advantage and firm performance. **Journal of Operations Management**, v. 29, n. 3, p. 163–180, 2011.

CAPÓ, M.; PÉREZ, A.; LOZANO, J. A. An efficient approximation to the K-means clustering for massive data. **Knowledge-Based Systems**, v. 117, p. 56–69, 2017.

CARILLO, K. D. A. et al. How to turn managers into data-driven decision makers: Measuring attitudes towards business analytics. **Business Process Management Journal**, v. 25, n. 3, p. 553–578, 2019.

CAUCHICK MIGUEL, P. A. et al. **Metodologia de Pesquisa em Engenharia de Produção e Gestão**

de Operações. 3. ed. Rio de Janeiro: Elsevier, 2018.

CHEN, N. et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. **The Lancet**, v. 395, n. 10223, p. 507–513, 2020.

CHEN, Y.; ARGENTINIS, E.; WEBER, G. IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research. **Clinical Therapeutics**, v. 38, n. 4, p. 688–701, 2016.

CHOUHAN, R.; PUROHIT, A. An approach for document clustering using PSO and K-means algorithm. **Proceedings of the 2nd International Conference on Inventive Systems and Control, ICISC 2018**, n. Icisc, p. 1380–1384, 2018.

CORE TEAM, R. D. A Language and Environment for Statistical Computing. **R Foundation for Statistical Computing**, v. 2, p. https://www.R-project.org, 2018.

CUI, Z.; LI, F.; ZHANG, W. Bat algorithm with principal component analysis. **International Journal of Machine Learning and Cybernetics**, v. 10, n. 3, p. 603–622, 2019.

DE MORSIER, F. et al. Cluster validity measure and merging system for hierarchical clustering considering outliers. **Pattern Recognition**, v. 48, n. 4, p. 1478–1489, 2015.

DENG, Z. et al. Efficient kNN classification algorithm for big data. **Neurocomputing**, v. 195, p. 143–148, 2016.

DIEZ-OLIVAN, A. et al. Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0. **Information Fusion**, v. 50, p. 92–111, 2019.

DONOHO, D. 50 Years of Data Science. **Journal of Computational and Graphical Statistics**, v. 26, n. 4, p. 745–766, 2017.

ELER, D. M. et al. **Simplified stress and simplified silhouette coefficient to a faster quality evaluation of multidimensional projection techniques and feature spaces**. Proceedings of the International Conference on Information Visualisation. **Anais**...2015

FERNANDEZ, N. F. et al. Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. **Scientific Data**, v. 4, p. 1–12, 2017.

FERREIRA, V. et al. A comprehensive supplier classification model for SME outsourcing. **Procedia Manufacturing**, v. 38, n. 2019, p. 1461–1472, 2019.

FOP, M.; MURPHY, T. B. Variable selection methods for model-based clustering. **Statistics Surveys**, v. 12, p. 18–65, 2018.

FRÄNTI, P. Genetic algorithm with deterministic crossover for vector quantization. **Pattern Recognition Letters**, v. 21, n. 1, p. 61–68, 2000.

GARFIELD, E.; SHER, I.; TORPIE, R. Use of citation data in writing the history of science. **Library of Congress Catalog Card Number**, n. 64, 1964.

GEORGE, G. et al. From the editors: Big data and data science methods for management research. **Academy of Management Journal**, v. 59, n. 5, p. 1493–1507, 2016.

GOVINDAN, K.; SIVAKUMAR, R. Green supplier selection and order allocation in a low-carbon paper industry: integrated multi-criteria heterogeneous decision-making and multi-objective linear programming approaches. **Annals of Operations Research**, v. 238, n. 1–2, p. 243–276, 2016.

GRUS, J. **Data Science from Scratch**. 1. ed. Cambridge, Mass.: O'Reilly Media, 2015. v. 1

GU, Z. et al. EnrichedHeatmap: An R/Bioconductor package for comprehensive visualization of genomic signal associations. **BMC Genomics**, v. 19, n. 1, p. 1–7, 2018.

GUPTA, A.; BARBU, A. Parameterized principal component analysis. **Pattern Recognition**, v. 78, p. 215–227, 2018.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. **Journal of Intelligent Information Systems**, v. 17, n. 2–3, p. 107–145, 2001.

HALLIKAS, J. et al. Risk-based classification of supplier relationships. **Journal of Purchasing and Supply Management**, v. 11, n. 2–3, p. 72–82, 2005.

HÄMÄLÄINEN, J.; JAUHIAINEN, S.; KÄRKKÄINEN, T. Comparison of internal clustering validation

indices for prototype-based clustering. **Algorithms**, v. 10, n. 3, 2017.

HARTIGAN, J.; WONG, M. Algorithm AS 136 A K-Means Clustering Algorithm. **Journal of the Royal Statistical Society Series B Methodological**, v. 28, n. 1, p. 100–108, 1979.

JAMES, G. et al. **An Introduction to Statistical Learning**. 8. ed. New York, USA: Springer New York, 2017.

JAYARAM REDDY, A. et al. Performance Analysis of Clustering Algorithm in Data Mining in R Language. **Communications in Computer and Information Science**, v. 837, n. November, p. 364–372, 2018.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, v. 349, n. 6245, p. 255–260, 2015.

KARPATNE, A. et al. Theory-guided data science: A new paradigm for scientific discovery from data. **IEEE Transactions on Knowledge and Data Engineering**, v. 29, n. 10, p. 2318–2331, 2017.

KASSAMBARA, A. **Multivariate Analysis 1: Practical Guide To Cluster Analysis in R: Unsupervised machine learning**. 1. ed. [s.l.] STHDA, 2017.

KAUFMAN, L.; ROUSSEEUW, P. Partitioning Around Medoids (Program PAM). p. 68–125, 1990a.

KAUFMAN, L.; ROUSSEEUW, P. **Finding Groups in Data: An Introduction to Cluster Analysis**. 1. ed. Wiley, Nova York: [s.n.].

KHATAMI, A. et al. A new PSO-based approach to fire flame detection using K-Medoids clustering. **Expert Systems with Applications**, v. 68, p. 69–80, 2017.

KHOMTCHOUK, B. B.; HENNESSY, J. R.; WAHLESTEDT, C. Shinyheatmap: Ultra fast low memory heatmap web interface for big data genomics. **PLoS ONE**, v. 12, n. 5, p. 1–9, 2017.

KIRCHDOERFER, T.; ORTIZ, M. Data-driven computational mechanics. **Computer Methods in Applied Mechanics and Engineering**, v. 304, p. 81–101, 2016.

KUANG, L.; ZHANG, L. **A scheduling algorithm based on Clara clustering**. AIP Conference Proceedings. **Anais**...2017

KUMAR, S.; ROUTROY, S. Performance analysis of supplier development programs. **Benchmarking: An International Journal**, v. 24, n. 2, p. 488–510, 2017.

LARSON, D.; CHANG, V. A review and future direction of agile, business intelligence, analytics and data science. **International Journal of Information Management**, v. 36, n. 5, p. 700–710, 2016.

LASI, H. **Industrial intelligence - A business intelligence-based approach to enhance manufacturing engineering in industrial companies**. Procedia CIRP. **Anais**...Elsevier B.V., 2013Disponível em: <http://dx.doi.org/10.1016/j.procir.2013.09.066>

LEWIN, H. A. et al. Earth BioGenome Project: Sequencing life for the future of life. **Proceedings of the National Academy of Sciences of the United States of America**, v. 115, n. 17, p. 4325–4333, 2018.

LI, Z.; WANG, G.; HE, G. Milling tool wear state recognition based on partitioning around medoids (PAM) clustering. **International Journal of Advanced Manufacturing Technology**, v. 88, n. 5–8, p. 1203–1213, 2017.

LIAO, K. et al. Parallel N-Path Quantification Hierarchical K-Means Clustering Algorithm for Video Retrieval. **International Journal of Pattern Recognition and Artificial Intelligence**, v. 31, n. 9, p. 1–26, 2017.

LIN, C. H.; LIU, J. C.; PENG, T. C. Performance evaluation of cluster algorithms for Big Data analysis on cloud. **Proceedings of the 2017 IEEE International Conference on Applied System Innovation: Applied System Innovation for Modern Technology, ICASI 2017**, p. 1434–1437, 2017.

LIU, Q. et al. Distributed k-means algorithm for sensor networks based on multi-agent consensus theory. **Proceedings of the IEEE International Conference on Industrial Technology**, v. 2016-May, p. 2114–2119, 2016.

LIU, Y. et al. Understanding and enhancement of internal clustering validation measures. **IEEE Transactions on Cybernetics**, v. 43, n. 3, p. 982–994, 2013.

MADANI, F.; WEBER, C. The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis. **World Patent Information**, v. 46, p. 32–48, 2016.

MARTÍN-MARTÍN, A. et al. Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. **Journal of Informetrics**, v. 12, n. 4, p. 1160–1177, 2018.

MCNICHOLAS, P. D. Model-Based Clustering. **Journal of Classification**, v. 373, n. November, p. 331–373, 2016.

MCPARLAND, D.; GORMLEY, I. C. Model based clustering for mixed data: clustMD. **Advances in Data Analysis and Classification**, v. 10, n. 2, p. 155–169, 2016.

MEHRYAR, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of Machine Learning**. The MIT Pr ed. Cambridge, Mass.: The MIT Press, 2012.

MEHRYAR, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Machine learning**. 2. ed. [s.l.] The Adaptive Computations and Machine Learning, 2018. v. 0

MEMARI, A. et al. Sustainable supplier selection: A multi-criteria intuitionistic fuzzy TOPSIS method. **Journal of Manufacturing Systems**, v. 50, n. September 2018, p. 9–24, 2019.

MOHEB-ALIZADEH, H.; MAHMOUDI, M.; BAGHERI, R. Supplier selection and order allocation using a stochastic multi-objective programming model and genetic algorithm. **International Journal of Integrated Supply Management**, v. 11, n. 4, p. 291–315, 2017.

MOKADEM, M. EL. The classification of supplier selection criteria with respect to Lean or Agile manufacturing strategies. **Journal of Manufacturing Technology Management**, v. 28, n. 2, p. 1–27, 2017.

NAJAFABADI, M. M. et al. Deep learning applications and challenges in big data analytics. **Journal of Big Data**, v. 2, n. 1, p. 1–21, 2015.

NGUYEN, H. et al. A new soft computing model for estimating and controlling blast-produced ground vibration based on Hierarchical K-means clustering and Cubist algorithms. **Applied Soft Computing**, v. 77, p. 376–386, 2019.

NI, Q. et al. A Novel Cluster Head Selection Algorithm Based on Fuzzy Clustering and Particle Swarm Optimization. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, v. 14, n. 1, p. 76–84, 2017.

NIGEL, S.; STUART, C.; ROBERT, J. **Administração da Produção**. São Paulo: Atlas, 2002.

NIVIO, Z. **Projeto de algoritmos: com implementações em Java e C++**. 1. ed. São Paulo: Cengage Learning, 2007.

NUNES, D. H. F. **Um Estudo Sobre O Algoritmo K-Means**. [s.l.] Universidade de Coimbra, 2016.

OLSON, R. S. et al. Evaluation of a tree-based pipeline optimization tool for automating data science. **GECCO 2016 - Proceedings of the 2016 Genetic and Evolutionary Computation Conference**, p. 485–492, 2016.

OMRANI, H.; SHAFAAT, K.; EMROUZNEJAD, A. **An integrated fuzzy clustering cooperative game data envelopment analysis model with application in hospital efficiency**. [s.l.] Elsevier Ltd, 2018. v. 114

ORSI, R. Use of multiple cluster analysis methods to explore the validity of a community outcomes concept map. **Evaluation and Program Planning**, v. 60, p. 277–283, 2017.

PATIL, C.; BAIDARI, I. Estimating the Optimal Number of Clusters k in a Dataset Using Data Depth. **Data Science and Engineering**, v. 4, n. 2, p. 132–140, 2019.

PATNAIK, A. K.; BHUYAN, P. K.; KRISHNA RAO, K. V. Divisive Analysis (DIANA) of hierarchical clustering and GPS data for level of service criteria of urban streets. **Alexandria Engineering Journal**, v. 55, n. 1, p. 407–418, 2016.

PELISSARI, R.; BEN-AMOR, S.; DE OLIVEIRA, M. C. **A Multiple-Criteria Decision Sorting Model for Pharmaceutical Suppliers Classification Under Multiple Uncertainties**. [s.l.] Springer International Publishing, 2019.

PETE, C. et al. Crisp-Dm 1.0. **CRISP-DM Consortium**, p. 76, 2000.

PROVOST, F.; FAWCETT, T. **Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking**. 1. ed. Sebastopol: O'Reilly Media, Inc., 2013.

QI, J. et al. An effective and efficient hierarchical K-means clustering algorithm. **International Journal of Distributed Sensor Networks**, v. 13, n. 8, p. 1–17, 2017.

RESTREPO, R.; VILLEGAS, J. G. Supplier evaluation and classification in a Colombian motorcycle assembly company using data envelopment analysis. **Academia Revista Latinoamericana de Administracion**, v. 32, n. 2, p. 159–180, 2019.

RODRIGUES, É. O. et al. k-MS: A novel clustering algorithm based on morphological reconstruction. **Pattern Recognition**, v. 66, p. 392–403, 2017.

RODRIGUEZ, M. Z. et al. **Clustering algorithms: A comparative approach**. [s.l: s.n.]. v. 14

ROSE, D. **Data Science: Create Teams that ask the right question and deliver real value**. 1. ed. Atlanta, Georgia: Apress, 2016.

ROUL, R. K. An effective approach for semantic-based clustering and topic-based ranking of web documents. **International Journal of Data Science and Analytics**, v. 5, n. 4, p. 269–284, 2018.

SABBAGH, R.; AMERI, F.; YODER, R. Thesaurus-guided text analytics technique for capability-based classification of manufacturing suppliers. **Journal of Computing and Information Science in Engineering**, v. 18, n. 3, p. 1–14, 2018.

SALAM, M. A.; KHAN, S. A. Achieving supply chain excellence through supplier management: A case study of fast moving consumer goods. **Benchmarking**, v. 25, n. 9, p. 4084–4102, 2018.

SCHOENHERR, T.; SPEIER-PERO, C. Data science, predictive analytics, and big data in supply chain management: Current state and future potential. **Journal of Business Logistics**, v. 36, n. 1, p. 120–132, 2015.

SCHUTT, R.; O'NEIL, C. **Doing Data Science: Straight Talk**. [s.l: s.n.].

SCRUCCA, L.; RAFTERY, A. E. clustvarsel: A Package Implementing Variable Selection for Gaussian Model-Based Clustering in R. **J Stat Softw.**, v. 84, n. 1, p. 139–148, 2018.

SEGHIER, M. L. Clustering of fMRI data: The elusive optimal number of clusters. **PeerJ**, v. 2018, n. 10, 2018.

SHOKOUHIFAR, M.; JALALI, A. Optimized sugeno fuzzy clustering algorithm for wireless sensor networks. **Engineering Applications of Artificial Intelligence**, v. 60, n. October 2016, p. 16–25, 2017.

SIQUEIRA, L. **Análise Bibliométrica de E-commerce no contexto de Supply Chain**. [s.l.] FAI, 2019.

SON, L. H.; THONG, P. H. Some novel hybrid forecast methods based on picture fuzzy clustering for weather nowcasting from satellite image sequences. **Applied Intelligence**, v. 46, n. 1, 2017.

SONG, H.; LEE, J.-G.; HAN, W.-S. **Pamae: Parallel k-Medoids Clustering with High Accuracy and Efficiency**. International Conference on Knowledge Discovery and Data Mining - KDD. **Anais**...2017

SUDA, B. **Data Science Salary Survey**. 1. ed. Canadá: O'Reilly Media, Inc, 2018.

TAN, K. C.; LYMAN, S. B.; WISNER, J. D. Supply chain management: A strategic perspective. **International Journal of Operations and Production Management**, v. 22, n. 5–6, p. 614–631, 2002.

TEDRE, M.; DENNING, P. J. **The long quest for computational thinking**. ACM International Conference Proceeding Series. **Anais**...2016

TURRIONI, J. B.; MELLO, C. H. P. **Metodologia de pesquisa em engenharia de produção**. [s.l: s.n.].

UMBLE, E. J.; HAFT, R. R.; UMBLE, M. M. Enterprise resource planning: Implementation procedures and critical success factors. **European Journal of Operational Research**, v. 146, n. 2, p. 241–257, 2003.

VALDEZ, D.; PICKETT, A. C.; GOODSON, P. Topic Modeling: Latent Semantic Analysis for the Social Sciences. **Social Science Quarterly**, v. 99, n. 5, p. 1665–1679, 2018.

VAN DER AALST, W. **Process mining: Data science in action**. [s.l: s.n.].

WALLER, M. A.; FAWCETT, S. E. Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. **Journal of Business Logistics**, v. 34, n. 2, p. 77–84, 2013.

WANG, C. N. et al. A Multi-Criteria Decision-Making (MCDM) Approach Using Hybrid SCOR Metrics, AHP, and TOPSIS for supplier evaluation and selection in the gas and oil industry. **Processes**, v. 6, n. 252, p. 2–12, 2018.

WANG, X. et al. AGNES-SMOTE: An Oversampling Algorithm Based on Hierarchical Clustering and Improved SMOTE. **Scientific Programming**, v. 2020, p. 1–9, 2020.

WANG, Y.; KUNG, L. A.; BYRD, T. A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. **Technological Forecasting and Social Change**, v. 126, p. 3–13, 2018.

WILKINSON, L.; FRIENDLY, M. History corner: The history of the cluster heat map. **American Statistician**, v. 63, n. 2, p. 179–184, 2009.

WING, J. Computational Thinking. **COMMUNICATIONS OF THE ACM**, v. 49, n. 3, p. 68-1-68–18, 2006.

WIRTH, R.; HIPP, J. **Towards a standard process model for data mining**. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining. **Anais**...Manchester, United Kingdon: 2000

YANG, H. et al. Resource Assignment Based on Dynamic Fuzzy Clustering in Elastic Optical Networks with Multi-Core Fibers. **IEEE Transactions on Communications**, v. 67, n. 5, p. 3457–3469, 2019a.

YANG, M. S.; NATALIANI, Y. A Feature-Reduction Fuzzy Clustering Algorithm Based on Feature-Weighted Entropy. **IEEE Transactions on Fuzzy Systems**, v. 26, n. 2, p. 817–835, 2018.

YANG, Q. et al. Understanding of internal clustering validation measures. **ACM Transactions on Intelligent Systems and Technology**, v. 10, n. 2, p. 1–19, 2019b.

YU, D. et al. An improved K-medoids algorithm based on step increasing and optimizing medoids. **Expert Systems with Applications**, v. 92, p. 464–473, 2018.

ZERZUCHA, P.; WALCZAK, B. Concept of (dis)similarity in data analysis. **Trends in Analytical Chemistry**, v. 38, n. 2019, p. 1–14, 2016.

ZHANG, Y. et al. Intelligent logistics supplier selection based on improved agglomerative hierarchical clustering algorithm. **IEEE International Conference on Industrial Informatics (INDIN)**, v. 2019-July, p. 1309–1314, 2019.

ZHAO, Z. et al. An energy-efficient clustering routing protocol for wireless sensor networks based on AGNES with balanced energy consumption optimization. **Sensors (Switzerland)**, v. 18, n. 11, 2018.

ZHU, X. Variable diagnostics in model-based clustering through variation partition. **Journal of Applied Statistics**, v. 45, n. 16, p. 2888–2905, 2018.

ZHU, X.; MELNYKOV, V. Probabilistic assessment of model-based clustering. **Advances in Data Analysis and Classification**, v. 9, n. 4, p. 395–422, 2015.

ANEXO A – CÓDIGO R PARA CLUSTERIZAÇÃO

```
----------#Validação de BD----------
require("factoextra")
require("clustertend")

#a base Fornecedores foi incluida via Import Dataset -> From Excel
FornecedoresTeste <- Fornecedores

#preparação dos dados para virarem data.frame
df <- data.frame(FornecedoresTeste[,1])
for (column in c(2:ncol(FornecedoresTeste))){
  print(column)
  df <- cbind(df, FornecedoresTeste[,column])
}

dd <- daisy(df) #deixando os dados comparáveis

#testando o MÉTODO DE HOPKINS

library(clustertend)

#calcular hopkins para a base de fornecedores
set.seed(123)
FornecedoresHopkins <- dd
hopkins(as.matrix(FornecedoresHopkins), 2)

#hipótese nula = conjunto de dados é uniformemente distribuído ( clusters irrelevantes)
#hipótese alternativa = conjunto de dados não é uniformemente distribuído (contém clusters com sentido)

#limite = 0.5  clusterizável
#mais próximos de zero mais tendência à clusterização

#testando o MÉTODO VISUAL ASSESSMENT OF CLUSTER TENDENCY (VAT)
fviz_dist(dist(dd), show_labels = FALSE) + labs(title = "Dados Fornecedores")

#observar a diagonal, se os pontos forem vermelho escuro não é clusterizável
#dissimilaridade entre objetos
#vermelho puro se dist(xi,xj) = 0
```

```
#vermelho puro se dist(xi,xj) = 1
#objetos pertencentes ao mesmo cluster são mostrados em sequência

library(factoextra)
library(fpc)
library(NbClust)

----------------------#K-MEANS----------------------------
#K-MEANS
#partitioning clustering
#numérico e lógico
require("datasets")
require("dplyr")
library(factoextra)

set.seed(123) #para maior reprodutividade dos resultados
#nstart testa x tarefas aleatórias iniciais diferentes e seleciona o melhor
#resultado correspondendo aquele com menor variação dentro do cluster (maior estabilidade)
km.res <- kmeans (dd,  3, nstart = 25)
#escrever os resultados na tela
print(km.res)

#RESULTADOS K-MEANS

#mostra a que cluster pertence cada observação:
km.res$cluster
#o head() mostra valores das primeiras colunas:
head(km.res$cluster, 4)
#mostra a quantidade de observações em cada cluster:
km.res$size

#VISUALIZAÇÃO K-MEANS

fviz_cluster(km.res, data = dd,
        palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
        ellipse.type = "euclid", #elipse de concentração
        star.plot = TRUE, #Adiciona segmentos dos centroides para itens
        repel = TRUE, #Evita label overplotting
        main="Gráfico K-Means",
        ggtheme = theme_minimal()
```

```
        )


#---------------------------K-MEDOIDS (PAM) -------------------------------------
#K-MEDOIDS (PAM)
#partitioning clustering
#pacotes pam e fpc
#demanda bastante processamentos pra bds muito grandes

#Medoid a dissimilaridade entre esse objeto e todos os outros membros do cluster
#é mínima e corresponde ao ponto mais centralmente localizado do cluster
#tem um medoid por cluster e ele é um exemplo representativo desse cluster

library(cluster)
library(factoextra)

#DESENVOLVENDO PAM

#PAM deixa os dados comparáveis no algoritmo
pam.res <-pam(df, 3)

#Visualizando PAM
print (pam.res)
fviz_cluster(pam.res,
        palette = c( "#00AFBB", "#E7B800", "#FC4E07"),
        ellipse.type = "euclid", #elipse de concentração
        star.plot = TRUE, #Adiciona segmentos dos centroides para itens
        repel = TRUE, #Evita label overplotting
        main="Gráfico PAM",
        ggtheme = theme_minimal()
)

#mostra os medoids (objetos que representam clusters, um por cluster):
pam.res$medoids

#clustering é um vetor que contém o número do cluster de cada observação
head(pam.res$clustering)
pam.res$clustering
```

```
#--------------------------CLARA-------------------------------------
#CLARA
#partitioning clustering
#feito com amostragens
#precisa preparação de dados

#é o método k-medoid para grandes quantidades de dados (mais de vários milhares)
#para reduzir o tempo computacional e problema de armazenamento de RAM
#Isso se deve ao uso de amostragem
library(cluster)
library(factoextra)
set.seed(1234)

#DESENVOLVENDO CLARA

#computar CLARA com 2 clusters com quantidade x de amostras
clara.res <- clara(df, 2, samples = 50, pamLike = TRUE)

#RESULTADOS CLARA

#escreve um overview dos resultados
print(clara.res)
#retorna os medoids encontrados pelo algoritmo
clara.res$medoids
#retorna um vetor que contém os clusters de cada observação
head(clara.res$clustering, 10)

#VISUALIZAÇÃO CLARA

fviz_cluster(clara.res, data = dd,
        palette = c("#2E9FDF", "#E7B800", "#FC4E07"),
        ellipse.type = "euclid", #elipse de concentração
        star.plot = TRUE, #Adiciona segmentos dos centroides para itens
        repel = TRUE, #Evita label overplotting
        main="Gráfico Clara",
        ggtheme = theme_minimal()
)

#---------------------------------AGNES-----------------------------
#AGNES
```

```
#hierarchical clustering
#Agglomerative
#Agglomerative Nesting ( Clusterização Hierarquica) AGNES
#função que faz o trabalho bruto

#trata todos os objetos como clusters unitários, em seguida pares de clustes
#são fundidos até todos os clusters serem fundidos em um grande cluster contendo
#todos os objetos, gerando uma represetanção baseada em árvore chamada dendograma

#é bottom-up (inicialmente considerado como um cluster de um elemento - folha)
#dois clusters mais similares são combinados em um maior - nó
#processo repetido até todos serem membros de um único cluster - raiz

#obs: agnes é bom em identificar clusters pequenos e Diana clusters grandes

library("cluster")
library("factoextra")  #para plotar dendogramas

res.agnes <- agnes (x = dd, #matriz dos dados
                    diss = TRUE,
                    stand = TRUE, # padronização dos dados
                    method = "ward" #metodo de linkagem
                    )

#VISUALIZAÇÃO AGNES

grp <- cutree(res.agnes, k = 3)
fviz_cluster(list(data=dd, cluster = grp),
             palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
             ellipse.type = "convex", #elipse de concentração
             star.plot = TRUE, #Adiciona segmentos dos centroides para itens
             repel = TRUE, #Evita label overplotting
             main="Gráfico AGNES",
             shoW.clust.cent = FALSE,
             ggtheme = theme_minimal()
)
#dendrograma
fviz_dend(res.agnes, cex = 0.6, k = 4)
```

```
#-----------------------------------DIANA---------------------------------------
#DIANA
#hierarchical clustering
#Divisive
#DIvisive ANAlytics Clustering DIANA

#é top-down (começa na raiz), todos dentro de um unico cluster
#em cada etapa, os mais heterogeneos são divididos em dois
#até chegar ao cluster individual

#obs: agnes é bom em identificar clusters pequenos e Diana clusters grandes

library("cluster")
#DIvisive ANAlytics Clustering DIANA

#desenvolvendo Diana
res.diana <- diana (x = dd, #matriz dos dados
                    diss = TRUE,
                    stand = TRUE, # padronização dos dados
                    metric = "euclidian", #metrica para a matriz de distancia
)

#VISUALIZAÇÃO Diana

grp <- cutree(res.diana, k = 3)
fviz_cluster(list(data=dd, cluster = grp),
             palette = c("#2E9FDF", "#E7B800", "#FC4E07"),
             ellipse.type = "convex", #elipse de concentração
             star.plot = TRUE, #Adiciona segmentos dos centroides para itens
             repel = TRUE, #Evita label overplotting
             main="Gráfico DIANA",
             shoW.clust.cent = FALSE,
             ggtheme = theme_minimal()
)

res.diana <- diana (x = FornecedoresDIANA, #matriz dos dados
                    stand = TRUE, # padronização dos dados
                    metric = "euclidian", #metrica para a matriz de distancia
)
```

```
#Dendrograma Diana

fviz_dend(res.diana, cex = 0.6, k = 4)

#-------------------------HEATMAP-------------------------------------
#heatmap
#hierarchical clustering

#heat map permite visualizar clusters de amostras e caracteristicas
#é feito com colunas e linhas de uma matriz de dados
#as colunas/linhas são reordenadas de acordo com o resultado da clusterização
#hierarquica, colocando observações similares próximas umas das outras
#os blocos de valores "altos" e "baixos" são adjacentes em uma matriz de dados
#finalmente um esquema de cor é aplicado para visualização e a matriz de
#dados é mostrada ajudando a encontrar variáveis que parecem ser caracteristicas
#para cada cluster de amostra

library("pheatmap")
#cutree_rows: num de clusters que as linhas são dividas, baseadas na clusterização
#hierárquica usando cutree (cut a tree into groups of data)
#cutree: corta uma árvore (clusterizacao hierarquica) em vários grupos especificando
#a quantidade desejada de grupos ou altura do corte
pheatmap(dd, cutree_rows = 4, legend = TRUE, annotation_legend = TRUE, show_colnames = TRUE )

#---------------------HIERARCHICAL K-MEANS-----------------------------------------
#hierarchical k-means
#hibrido entre hierarchical clustering e partitioning clustering

#passo 1:faz uma clusterização hierarquica e corta uma arvore em k-clusters
#passo 2:computa o centro (a média) de cada cluster
#passo 3:computa k-means usando um conjunto de centro de clusters (do passo 2)
#como um conjunto de centro de clusters iniciais

#o algoritmo k-means vai melhorar o particionamento inicial gerado no passo 2
library(factoextra)
require("datasets")

#DESENVOLVENDO HIERARCHICAL K-MEANS

#matriz e número de clusters
```

```
res.hk <- hkmeans(df, 6)

#RESULTADOS DO HIERARCHICAL K-MEANS
res.hk #escreve na tela todos os resultados

#VISUALIZAÇÃO DO HIERARCHICAL K-MEANS

#visualizar a árvore (o dendograma)
fviz_dend(res.hk, cex = 0.6, palette = "jco",
        rect = TRUE, rect_border = "jco", rect_fill = TRUE)
#visualizar os clusters finais do hkmeans
fviz_cluster(res.hk, palette = "jco", repel = TRUE,
           ggtheme = theme_classic())

#----------------------------FUZZY C-MEANS-----------------------------------------
#fuzzy c-means (FCM)
#soft clustering
#fanny(x, k, metric = "euclidean", stand = FALSE)

#cada elemento tem uma probabilidade de pertencer a cada cluster
#ou seja, cada elemento tem um conjunto de coeficientes de assinaturas
#corresponde ao grau de pertencer a dado cluster

#ele é diferente do k-means e do k-medoids onde cada objeto é influenciado
#por exatamente um cluster

#na clusterização fuzzy pontos próximos do centro do cluster podem estar em
#um cluster com um grau maior do que clusters na borda do cluster.
#o grau, no qual um element pertence a um cluster é um número de 0 a 1

#No fuzzy c-means (FCM) o centroide de um cluster é calculado como a média
#de todos os pontos pesado pelo grau deles de pertencer ao cluster

library(cluster)
require("datasets")

#DESENVOLVENDO FUZZY C-MEANS
#fuzzy com k = 3
res.fanny <- fanny(FornecedoresFUZZY, 3, diss = FALSE, stand = FALSE)
#coeficientes de assinatura dos 3 primeiros, chance de pertencer a cada um dos 3 clusters
```

```
#RESULTADOS FUZZY C-MEANS

head(res.fanny$membership, 3)
#coeficiente de partição de dunn
#Esse coef é a soma de todos os coeficientes de assinatura ao quadrado
#dividido pela quantidade de observações
#um valor baixo indica um agrupamento muito fuzzy (muito confuso)
#e um valor próximo de 1 indica uma clusterização near-crisp (muito nítida)
res.fanny$coeff #coeficiente de partição de dunn
#clusters dos primeiros valores
head(res.fanny$clustering)

#VISUALIZAÇÃO DO FUZZY C-MEANS

#visualizar clusters
fviz_cluster(res.fanny, ellipse.type = "norm", palette = "jco", repel = TRUE,
         ggtheme = theme_minimal(), legend = "right")

#------------------------------------------------------------------
#model based clustering

library(mclust)
mc<- Mclust(dd)


library(factoextra)
#seleção do modelo
fviz_mclust(mc ,"BIC", palette = "jco") #valores BIC usados para escolher quantidade de clusters
fviz_mclust(mc, "classification", geom = "point",
        pointsize = 1.5, palette = "jco") #classification: plotagem mostrando os clusters
#visualizando mbc
fviz_mclust(mc, data = FornecedoresMBC, "uncertainty", palette = "jco"
        ) #classification: uncertaint

#xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
#xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
#encontrando o melhor algoritmo para os dados
```

```
library(clValid)
require("datasets")

#método clvalid para comparar algoritmos com dados da base
#validacao interna
clmethods <- c("hierarchical", "kmeans", "fanny", "pam", "clara", "agnes", "diana")
intern <- clValid (dd, nClust = 2:6,
              clMethods = clmethods, validation = "internal")
summary(intern)
#validacao estabilidade
clmethods <- c("hierarchical", "kmeans", "fanny", "pam", "clara", "agnes", "diana")
stab <- clValid (dd, nClust = 2:6,
            clMethods = clmethods, validation = "stability")
summary(stab)
```