UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

ANA PAULA BARBOSA DE MORAIS

TÉCNICAS DE CLUSTERIZAÇÃO E COMPARAÇÃO DE GRUPOS BASEADOS NAS CARACTERÍSTICAS DE INOVAÇÃO EMPRESAS DE LONDRINA E REGIÃO

LONDRINA

2022

ANA PAULA BARBOSA DE MORAIS

TÉCNICAS DE CLUSTERIZAÇÃO E COMPARAÇÃO DE GRUPOS BASEADOS NAS CARACTERÍSTICAS DE INOVAÇÃO EM EMPRESAS DE LONDRINA E REGIÃO

Clustering techniques and comparison of groups based on innovation characteristics of companies in Londrina and region

Trabalho de conclusão de curso de graduação apresentada como requisito para obtenção do título de Bacharel em Engenharia de Produção da Universidade Tecnológica Federal do Paraná (UTFPR).
Orientador(a): Dr. Bruno Samways dos Santos.

LONDRINA

2022

ANA PAULA BARBOSA DE MORAIS

TÉCNICAS DE CLUSTERIZAÇÃO E COMPARAÇÃO DE GRUPOS BASEADOS NAS CARACTERÍSTICAS DE INOVAÇÃO EMPRESAS DE LONDRINA E REGIÃO

Trabalho de conclusão de curso de graduação apresentada como requisito para obtenção do título de Bacharel em Engenharia de Produção da Universidade Tecnológica Federal do Paraná (UTFPR).

Data de aprovação: 23 de novembro de 2022

_____

Bruno Samways dos Santos
Doutor
Universidade Tecnológica Federal do Paraná

_____

Pedro Rochavetz de Lara Andrade
Doutor
Universidade Tecnológica Federal do Paraná

_____

Rafael Henrique Palma Lima
Doutor
Universidade Tecnológica Federal do Paraná

LONDRINA
2022

AGRADECIMENTOS

Agradeço primeiramente à minha família e meus amigos que se tornaram minha família de Londrina, pelo poio e suporte incondicional durante toda a graduação.

Agradeço ao meu orientador Prof. Dr. Bruno Samways dos Santos, pela sabedoria e o apoio com que me guiou nesta trajetória.

À Universidade Tecnológica Federal do Paraná pelo apoio e suporte e pela concessão da cota voluntária de iniciação científica PVB2021130000093 que está vinculado ao edital PROPPG – 05 B/2021 – PIVICT.

Enfim, a todos os que por algum motivo contribuíram para a realização desta pesquisa.

RESUMO

O Inovação é frequentemente vista como o alicerce para o sucesso em muitas áreas, como empresas e até mesmo países, mas há dificuldades em mensurá-la. Apesar de estar presente em diversas áreas, inclusive no contexto de inovação, o *machine learning* (ML) ainda não foi utilizado para destacar características inovadoras em empresas. Este trabalho tem como objetivo empregar técnicas de *machine learning* (ML) não supervisionadas para a formação de grupos (*clusters*) de empresas de Londrina e região para discutir como as variáveis relacionadas à inovação se diferenciam entre os *clusters* formados. Para isso, elaborou-se um instrumento de coleta de dados com base na CIS 4 (Quarta Pesquisa de Inovação da Comunidade) criada pela Eurostat e na PINTEC (Pesquisa Industrial de Inovação Tecnológica). Após a criação do instrumento de coleta, enviou-se a uma amostra de empresas de Londrina e região. Para extração de conhecimento da base de dados obtida, foram utilizadas quatro técnicas de agrupamento: K-*means*, K-*means* com PCA, agrupamento hierárquico, e agrupamento hierárquico com PCA. Como resultado, foram semelhantes as distribuições das empresas nos dois *clusters* criados para todos os algoritmos. Identificou-se que em um dos *clusters* foram alocadas empresas com maior grau de inovação, enquanto no outro, as empresas com menor grau. Então foram feitas análises, com embasamento teórico, de três variáveis no contexto de inovação que obtiveram mais diferenças entre *clusters* formados: investimento regular em P&D, parcerias feitas com diferentes categorias para inovar e solicitações e/ou registros. Ao final, verificou-se que essas variáveis foram consideradas importantes para a caracterização do *cluster* mais inovador, porém os resultados indicaram que uma maior amostra deve ser coletada para validar os resultados obtidos inicialmente por esta pesquisa.

Palavras-chave: inovação; clusterização; k-means; agrupamento hierárquico.

ABSTRACT

Innovation is often seen as the foundation for success in many areas, such as companies and even countries, but there are difficulties in measuring it. Despite being present in several areas, including in the context of innovation, machine learning (ML) has not yet been used to highlight innovative features in companies. This work aims to employ unsupervised machine learning (ML) techniques for the formation of groups (clusters) of companies in Londrina and region to discuss how the variables related to innovation differ between the formed clusters. For this, a data collection instrument was developed based on CIS 4 (Fourth Community Innovation Survey) created by Eurostat and PINTEC (Industrial Research on Technological Innovation). After creating the collection instrument, a sample of companies in Londrina and region was sent. To extract knowledge from the obtained database, four clustering techniques were used: K-means, K-means with PCA, hierarchical clustering, and hierarchical clustering with PCA. As a result, the companies' distributions in the two clusters created for all algorithms were similar. It was identified that in one of the clusters were allocated companies with a higher degree of innovation, while in the other, companies with a lower degree. Then, analyzes were carried out, with theoretical basis, of three variables in the context of innovation that obtained more differences between clusters formed: regular investment in R&D, partnerships made with different categories to innovate and requests and/or registrations. In the end, it was verified that these variables were considered important for the characterization of the most innovative cluster, however the results indicated that a larger sample should be collected to validate the results obtained initially by this research.

Keywords: innovation; clustering; k-means; hierarchical clustering.

SUMÁRIO

1 INTRODUÇÃO

Inovação é frequentemente vista como o alicerce para o sucesso em muitas áreas, como empresas e até mesmo países (ROBINSON, STUBBERUD, 2012). Além disso, Claudino *et al.* (2017) dizem que a inovação é considerada um instrumento importante para as empresas aumentarem sua competitividade e se manterem fortes em cenários de constante mudanças e variações de mercado.

Segundo Drobyazko (2019), o ambiente de negócios competitivo da atualidade e com a rápida disseminação de conhecimento, aumentar a competitividade tem sido de grande importância. Foi o que ocorreu com grande parte das empresas que sobreviveram à pandemia do Covid-19, elas buscaram formas de se reinventar para se manterem competitivas no mercado.

Segundo uma pesquisa com executivos de 500 indústrias de médio e grande porte, feita pela CNI (Confederação Nacional da Indústria) e aplicada pelo Instituto FSB Pesquisa em outubro de 2021, 80% da amostra das empresas inovaram durante a pandemia e obtiverem aumento de lucro, produtividade e competitividade no mercado. Esta informação corrobora com a afirmação de Carvalho, Reis, Cavalcante (2011, p. 11), que diz “normalmente, quanto mais inovadora uma empresa for, maior será sua competitividade e melhor sua posição no mercado em que atua”.

Rogers (1998) afirma que a medição do conceito de inovação é difícil, uma vez que a natureza do escopo das atividades inovadoras é vasta, pois há várias formas de inovar. Segundo o Manual de Oslo (2005) criado pela OCDE (Organização para a Cooperação e Desenvolvimento Econômico) há quatro tipos de inovação: produto, processo, marketing e organizacional. A partir das definições descritas no manual, a Eurostat criou o *Fourth Community Innovation Survey* (CIS 4), um questionário para avaliar a capacidade inovadora das empresas. Ter esses dados não é o suficiente, é preciso saber como extrair desses conjuntos de dados, informações e conhecimentos úteis e/ou valiosas.

Para fazer a extração do conhecimento usa-se a mineração de dados (MD), que é uma das etapas com maior importância no processo de descoberta de conhecimento em banco de dados (*Knowledge Discovey in Databases* - KDD) (GALVÃO; MARIN, 2009).

O KDD busca padrões ocultos em um grande volume de informação utilizando alguma metodologia específica, que é útil para várias aplicações (ALAM *et al.*, 2014).

Dessa forma, é possível utilizá-las em diversas áreas, como agronegócio, melhoramento genético, mercados de ações entre outras. O único requisito para aplicação do KDD é que haja dados passíveis de manipulação, e a área da inovação pode gerar dados sobre o nível ou maturidade de empresas com um grande potencial de exploração.

Tendo em vista a importância da inovação e a dificuldade em mensurá-la, este trabalho utilizou o questionário CIS 4 e a PINTEC (Pesquisa Industrial de Inovação Tecnológica) para a criação e aplicação de um instrumento de coleta de dados em empresas de Londrina e região. Estes dados foram agrupados por técnicas de machine learning não supervisionadas e analisados de forma visual, destacando-se sempre as variáveis mais relevantes nos clusters encontrados.

1.1 Objetivo Geral

Empregar técnicas de *machine learning* (ML) não supervisionadas para a formação de grupos (*clusters*) de empresas de Londrina e região, para discutir como variáveis relacionadas à inovação se diferenciam entre os *clusters* formados.

1.2 Objetivos Específicos

- Construir um conjunto de dados, a partir de um instrumento de coleta baseado no Fourth Community Innovation Survey (CIS 4) de 2004 (adaptado) e na Pesquisa Industrial de Inovação Tecnológica (PINTEC);
- Aplicar o questionário em empresas de Londrina e região;
- Aplicar as técnicas de ML não supervisionadas, para verificar possíveis similaridades, dissimilaridades e padrões para os clusters formados;
- Discutir sobre os grupos formados.

1.3 Justificativa

Na última década houve muitos trabalhos utilizando mineração de dados no contexto da inovação como Kong *et al.* (2017) que utilizaram mineração de dados para avaliar as lacunas de inovação em países em desenvolvimento. Outra contribuição dos pesquisadores Yan *et al.* (2021) que utilizaram ML para pesquisar sobre avaliação

combinada da estabilidade operacional do ecossistema de inovação do setor de energia. Já Sun (2021) realizou uma simulação sobre os benefícios econômicos da inovação tecnológica utilizando ML. Uma outra aplicação interessante foi a de Saura *et al.* (2019) que utilizaram *Text Mining* para identificar fatores chaves para criação de *startups* de sucesso a partir da rede social *Twitter.* Porém não há relatos de pesquisa que utilizam técnicas de mineração de dados ou ML que buscam identificar o grau de inovação das organizações, tema que pode gerar vários *insights* em relação aos perfis das empresas localizadas em Londrina e região. Logo o presente trabalho justifica-se pelo fato de preencher esta lacuna na área de pesquisa.

1.4 Estrutura do trabalho

Após a exposição do contexto, objetivos e justificativas do trabalho presentes no primeiro capítulo, é apresentado no Capítulo 2 um referencial teórico, onde são abordados embasamentos sobre Mineração de Dados e KDD, tipos de ML, seguidos das técnicas de agrupamentos (K-*means* e hierárquico) e método PCA e por fim, a fundamentação teórica sobre inovação.

No Capítulo 3 é apresentado a metodologia do trabalho, iniciando com a explicação do instrumento de coleta, seguido do apontamento das ferramentas utilizadas e por último uma descrição detalhada das etapas da pesquisa.

Em seguida, são apresentados no Capítulo 4 os resultados e discussões obtidos a partir da aplicação dos algoritmos, comparando as técnicas. Inicialmente é abordado a análise demográfica dos grupos, em seguida é feita uma análise de variáveis relacionadas à inovação que se diferenciam entre os *clusters* formados.

Finalmente no Capítulo 5, é apresentada a conclusão e limitações do estudo.

2 REFERENCIAL TEÓRICO

Esta seção apresenta o embasamento teórico do trabalho, primeiramente abordando a metodologia do KDD e Mineração de Dados. Em seguida, discorre sobre os tipos de ML e as tarefas de mineração de dados, então descreve as técnicas de agrupamento K-*means* e agrupamento hierárquico, assim como a análise de componentes principais (PCA). Por fim, traz uma breve revisão de literatura sobre inovação tecnológica.

2.1 KDD e Mineração de dados

Segundo Fayyad *et al.* (1996), KDD é um processo não trivial, interativo e iterativo, que consiste em várias etapas e tem como objetivo identificar padrões, sejam eles compreensíveis, válidos, novos e potencialmente úteis, utilizando-se de grandes conjuntos de dados.

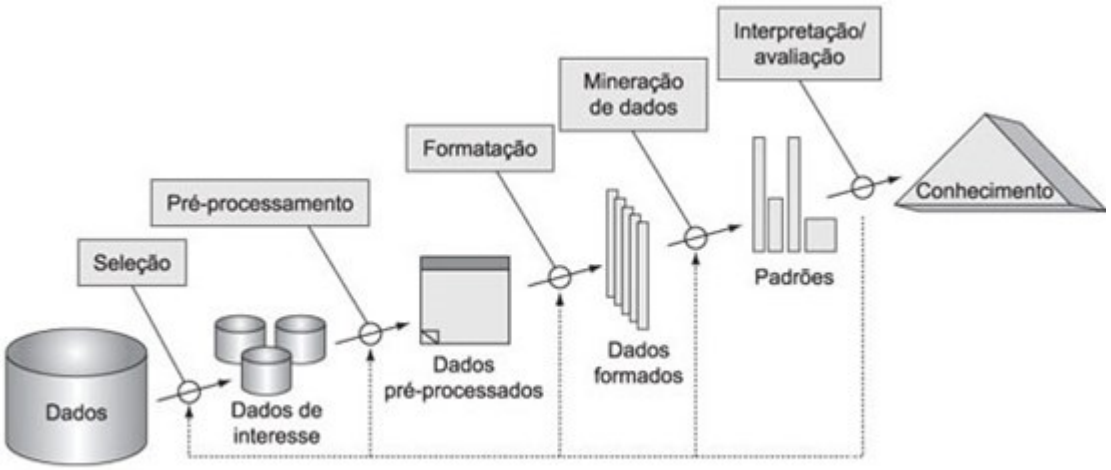
A expressão “não trivial” é utilizada referir-se à complexidade que os processos de KDD, os quais geralmente são divididos em várias etapas a fim de atingir o objetivo de identificar padrões potencialmente úteis e de fácil compreensão por meio da análise da base de dados (FAYYAD, 1996). Já o termo interativo faz referência à necessidade de um ser humano atuando como responsável pelo controle do processo. Por fim, o processo KDD é iterativo pois há possibilidade de repeti-lo, integral ou parcialmente, com a finalidade de encontrar melhores resultados, a partir do refinamento sucessivo (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Castro e Ferrari (2016) afirma que as etapas que compõem o KDD consistem em: definição da base de dados, pré-processamento, formatação dos dados, MD (processamento) e pós-processamento. Para dar o início no processo de KDD é preciso selecionar a base de dados que, podendo ser um conjunto ou subconjunto de dados que serão analisados (MOURA, 2018).

O pré-processamento, a segunda etapa do KDD, engloba as funções responsáveis pela captação, organização e pelo tratamento de dados, estas etapas são executadas visando a preparação da base de dados para uma análise eficiente e eficaz. A preparação dos dados inclui a limpeza (remoção de dados inconsistentes e/ou dados considerados ruídos), seleção e/ou remoção dos dados (deixando apenas dados relevantes à análise), a integração (cruzamento de dados obtidos de diferentes

fontes) e por fim a transformação, a terceira etapa do processo, que seria a formatação dos dados, deixando-os mais apropriados para a etapa seguinte (CASTRO; FERRARI, 2017). É na etapa de transformação que são aplicadas técnicas como normalização de atributos, criação e redução de atributos e sintetização dos dados (MOURA, 2018).

Figura 1 - Etapas do KDD



Fonte: Teófilo (2015)

A quarta etapa, conhecida também pela expressão Mineração de Dados (do inglês *Data Mining* – DM), é a fase de processamento, a etapa considerada mais ampla do KDD. Segundo Goldschmidt *et al.* (2015), é responsável pela geração de modelo de conhecimento, o qual é visualizado, analisado e interpretado na etapa seguinte, o pós-processamento.

O pós-processamento é a etapa onde ocorre a avaliação dos resultados obtidos após a MD, com o objetivo de identificar conhecimentos genuinamente úteis e não ordinários. (CASTRO; FERRARI, 2017).

2.2 *Machine Learning* (ML)

ML é uma área da IA (Inteligência Artificial) baseada na ideia de que *softwares* (máquinas) podem aprender a alcançar um objetivo a partir da sua experiência anterior com um alto volume de dados, identificando padrões e até mesmo tomar decisões, minimizando ocorrência de interferência humana no processo (CASTRO; FERRARI, 2017).

Segundo McCue (2014), o ML pode ser dividido tradicionalmente em técnicas de aprendizado de máquina supervisionadas e não supervisionadas, porém existem mais dois tipos: o semissupervisionado e por reforço (BHAVSAR *et al*., 2017). Entretanto, apenas os aprendizados supervisionados e não supervisionados serão descritos na sequência.

O ML supervisionado necessita de uma maior interatividade humana em sua implantação. Isto porque esta abordagem é baseada em um conjunto de objetos (ou instâncias) para os quais as saídas desejadas são conhecidas (CASTRO e FERRARI, 2017). Desta forma, a partir desse conjunto de instâncias, que também pode ser chamado também de conjunto de treinamento, o aprenderá os padrões que resultaram na saída anteriormente conhecida.

Após o treinamento da máquina, o algoritmo deverá ser aplicado no conjunto de teste, ou seja, sem a saída desejada previamente definida, de modo que o próprio algoritmo deverá apresentar a saída que mais se enquadra nos padrões que foram definidos pelos atributos de entrada.

Já o aprendizado não supervisionado, ao contrário do supervisionado, não precisa do conjunto de teste, pois segundo Castro e Ferrari (2017) é fundamentado nos objetos da base, mas não possuem uma saída ou alvo definido, ou seja, os dados não são rotulados. Desta forma, o algoritmo deve aprender a rotular os objetos.

Os objetivos de cada tipo de aprendizado podem ser atingidos por meio da utilização das tarefas de MD comumente, as tarefas são divididas da seguinte forma: Classificação, Regressão, Associação e Agrupamento.

A tarefa de Classificação é, segundo Bramer (2016), uma das aplicações mais utilizadas quando o assunto é MD. Esta tarefa utiliza-se do método de aprendizagem supervisionada, ou seja, necessita de alvo definido e deve ser do tipo categórico.

Conforme Bezerra *et al*. (2010) a tarefa de classificação consiste na descoberta de uma função que seja capaz de mapear um conjunto de dados em um conjunto de rótulos categóricos pré-definidos, denominados classes. Em outras palavras, na classificação o intuito é descobrir ou descrever a classe de um objeto, sendo que objetos rotulados em uma mesma classe devem possuir padrões comuns entre si (AMARAL, 2016; FONTANA *et al.* 2009).

Segundo Fontana *et al.* (2009), esta tarefa é muito utilizada em sistemas de validação de compras de cartão de crédito, com a finalidade da loja decidir se deve aceitar ou não pagamento via cartão de crédito de determinado cliente.

Primeiramente, o algoritmo é treinado pelo conjunto de treino, o qual possui clientes classificados como bons pagadores e maus pagadores. Desta forma, o algoritmo cria um modelo dos padrões dos atributos dos clientes que são bons pagadores (adimplentes) e dos padrões dos maus pagadores (inadimplentes).

Assim, utilizando este modelo, quando o algoritmo for apresentado aos novos clientes, ainda sem classificação, ele deverá ser capaz de identificar a qual classe os novos clientes se enquadram melhor, a partir dos atributos de cada cliente (objeto).

A segunda tarefa a ser abordada é a tarefa de estimação ou regressão. Como a tarefa de classificação, a tarefa de regressão também utiliza o método de aprendizagem supervisionado, entretanto o atributo alvo neste caso é quantitativo (contínuo) e não mais categórico (discreto) (NING; KUMAR; STEINBACH, 2013). Fazendo uma analogia com o exemplo dado na tarefa de classificação, ao invés da regressão ter saída categórica (bons e maus pagadores), esta tarefa é capaz de estimar, por exemplo, o quanto de crédito aquele objeto (cliente) conseguiria na loja ou em bancos, etc.

Sobre as duas tarefas citadas acima pode-se afirmar que:

> A análise preditiva pode ser dividida em duas subtarefas: análise preditiva categórica, também chamada de tarefa de classificação; e análise preditiva numérica, também chamada de tarefa de regressão. A primeira subtarefa se manifesta quando os rótulos associados aos dados pertencem a um conjunto discreto e finito de categorias. Já a segunda se faz presente quando os rótulos associados aos dados são numéricos e pertencentes a um conjunto de valores contínuos. (SILVA; PERES; BOSCARIOLI, 2016, p. 9).

A próxima tarefa a ser apresentada é a tarefa de associação, a qual conforme Silva *et al.* (2016, p. 10) “é definida como a busca por ocorrências frequentes e simultâneas entre elementos de um contexto”, em outras palavras é a tarefa que procura por padrões de comportamento que se repetem frequentemente, a partir da aplicação de regras de associação determinadas. Esta tarefa é muito utilizada em dados sobre vendas, quando se associa regras entre os tipos de produtos vendidos, é possível ter indicações de oportunidades e *insights* sobre um determinado grupo de consumidores (SILVA, 2004).

Por fim, é importante falar sobre a última tarefa de MD neste trabalho, a tarefa de agrupamento ou clusterização. Zengin *et al.* (2011) diz que a clusterização é uma tarefa de aprendizado não-supervisionada, isto significa que utiliza uma base de dados não rotulados, sem um alvo previamente definido. Fontana *et al.* (2009) ainda afirmam que o objetivo desta tarefa é identificar padrões na base de dados, de forma

que objetos com características similares sejam alocados no mesmo *cluster*, e objetos com dissimilaridades são alocados em grupos diferentes.

Goldschmidt *et al.* (2015) acrescentam que esta é uma das tarefas básicas da MD apta a realizar partições naturais de registros (objetos) em uma base de dados. Já um dos maiores desafios desta tarefa é o desconhecimento do número ideal de *clusters*, seguido de sua análise (SILVA *et al.*, 2009; FONTANA *et al.*, 2009), isto significa que mesmo o algoritmo agrupando as instâncias de acordo com suas similaridades, é necessário analisar os grupos e verificar se o conhecimento extraído da segmentação é relevante para o objetivo do estudo a ser feito.

De acordo com Camilo e Silva (2009) a clusterização pode ser aplicada em auditorias, na separação de comportamentos suspeitos, assim como na pesquisa e segmentações de mercado, classificação de documentos na WEB e detecção de fraudes.

2.3 K-*means*

O K-*means*, ou K-médias, é o algoritmo mais antigo e mais popular do método de partição de acordo com Machado (2011), além de ser um algoritmo do tipo de agrupamento exclusivo (ou também chamado de *hard clustering*), isto significa que cada instância da base de dados é conferida a apenas um *cluster*. O algoritmo K-*means* segmenta o conjunto de dados em um número k de *clusters*, sendo necessário definir previamente e de forma randômica qual será este número (parâmetro de entrada para rodar o programa). Isto é visto como uma desvantagem do método, por isso, geralmente deve-se realizar testes alterando o valor de k de modo a encontrar o número de *clusters* que será melhor para a partição de base de dados em questão (SINAGA e YANG, 2020; RASCHKA, 2015; GOLDSCHMIDT, 2015).

É importante ressaltar que o K-*means* leva este nome porque é um algoritmo de agrupamento baseado em protótipos, desta forma, cada *cluster* se baseia em um protótipo, que neste caso é a média de pontos com similaridade, sendo denominado de “centróide” quando os atributos forem contínuos, e “medóide” quando forem categóricos (RASCHKA, 2015).

Silva *et al.* (2016) complementa que o algoritmo k-*means* opera com o intuito de minimizar a soma dos erros quadráticos intragrupos (a somatória da distância de

cada objeto até o centróide do *cluster* ao qual está inserido), de maneira que os grupos formados são compactos, com formato esférico e podem ser desbalanceados.

Segundo Goldschmidt *et al.* (2015) e Castro e Ferrari (2016), o funcionamento do K-*means* se dá por meio de uma técnica de refinamento iterativo, ou seja, a alocação dos objetos ao *cluster* cujo centróide está mais próximo, somada à atualização dos valores dos centróides, resultam em um processo iterativo de otimização de uma função custo que também calcula a soma dos erros quadráticos, e como já mencionado, o método busca minimizar esta função (Equação 1):

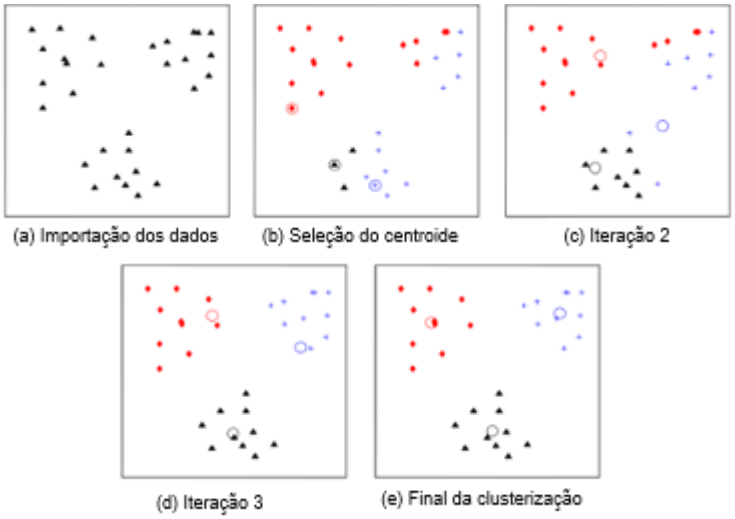
$$fc = \sum_{i=1}^{k} \sum_{x \in g_i} d \quad \textbf{(1)}$$

Onde fc é a função custo da base de dados, x é um exemplar da base, c_i é o centróide do *cluster* g_i e d é a distância entre o exemplar e o centróide do *cluster* (CASTRO; FERRARI, 2016).

De acordo com Goldschmidt *et al.* (2015), o procedimento iterativo do algoritmo k-*means* pode ser descrito nos seguintes passos:

1. Os centróides iniciais são escolhidos aleatoriamente;
2. Calcula-se a distância entre cada um dos objetos da amostra em relação a todos os centróides;
3. Cada objeto é agregado ao centróide que se encontra mais próximo;
4. Calcula-se os novos centróides utilizando-se a média dos objetos atribuídos a cada centróide. Neste momento, pode ocorrer reposicionamento dos centróides e uma nova alocação dos objetos aos *clusters*;
5. É necessário repetir o passo 4 até que o algoritmo não promova mais alterações nos centróides e nas alocações dos objetos.

Figura 2 - Ilustração do algoritmo K-*means*

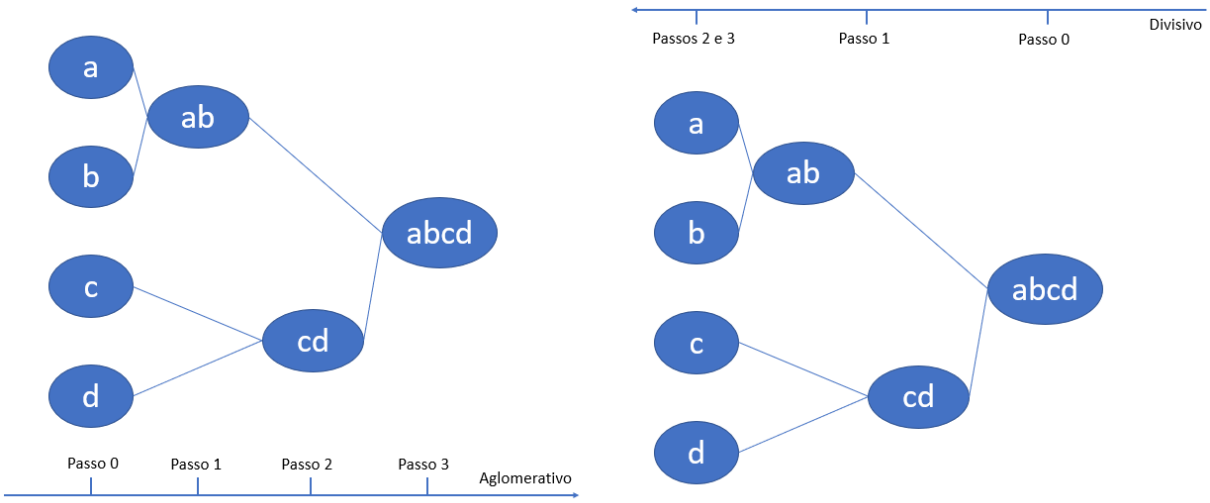


Fonte: Jain (2010) adaptado

2.4 Agrupamento Hierárquico

O método de agrupamento hierárquico, segundo Amaral (2016) consiste na inclusão progressiva dos objetos em grupos, de modo a formar uma estrutura hierárquica multinível. Pode ser implementado seguindo duas abordagens: aglomerativa e divisiva (SILVA *et al.*, 2016). A divisiva (*top-down*) é iniciada com todos os elementos alocados em um só grupo, a qual a cada iteração os objetos são divididos em múltiplos *clusters*. Já a aglomerativa (*bottom-up*) parte de n grupos, onde a amostra possui n elementos, que vão sendo fundidos a cada iteração até formar apenas um *cluster*. (AMARAL, 2016).

Figura 3 - Esquema gráfico para ilustrar o processo de agrupamento hierárquico aglomerativo e divisivo



Adaptado de Silva et al. (2016)

Geralmente os agrupamentos hierárquicos são representados graficamente pelo modelo denominado dendrograma e também utilizam como medida de distância a similaridade na formação dos clusters (SILVA *et al.*, 2016; AMARAL, 2016). O critério de similaridade (ou dissimilaridade) deve ser aplicado a cada par de objetos ou grupos. Este critério funciona da seguinte forma: quanto menor for o valor quantificado da distância, mais similares os objetos serão. Desta forma, os objetos que estiverem mais próximos serão fundidos, formando um cluster, que será identificado como um novo objeto, este processo será repetido até a formação de um único grupo (SILVA *et al.*, 2016; SHANNON, 2008).

De acordo com Silva et al. (2016), para aplicar a medida de distância em pares, utiliza-se a uma matriz de similaridade ou matriz de distância. Esta matriz é quadrada, de tamanho n x n, onde n é o número de elementos presentes na amostra, simétrica e sua diagonal principal é composta por zeros. Todos estes efeitos ocorrem, pois, cada célula presente na matriz representa a medida de distância entre dois objetos $dist\ x \to p,\ x \to q$, onde $x \to p$ e $x \to q$ são os objetos em questão, ou seja, a distância de um elemento a ele mesmo é zero (diagonal principal) e possuem similaridade máxima, ademais a distância dos elementos p e q é a mesma que a distância entre q e p $(dist\ x \to p, x \to q\ =\ dist\ x \to q, x \to p)$. Como mostrado na equação 2:

$$MS = \begin{bmatrix} 0 & & & & \\ dist_{x_{2,x_1}} & 0 & & & \\ dist_{x_{3,x_1}} & dist_{x_{3,x_2}} & 0 & & \\ \vdots & \vdots & \vdots & 0 & \\ dist_{x_{n,x_1}} & dist_{x_{n,x_2}} & \dots & dist_{x_{n,x_{n-1}}} & 0 \end{bmatrix} \quad \textbf{(2)}$$

Conforme Almeida (2013) e Silva *et al.* (2016), as técnicas para aplicar medida de distância em pares mais utilizadas no método hierárquico são: Menor distância ou Vizinhos mais próximos, Maior distâncias ou Vizinhos mais distantes, Distância média, Distância de centróides e por fim o método de Ward. Este último método, considerado um método mais complexo comparado aos outros, oferece uma maior precisão em relação aos resultados e ainda minimiza a variância entre os elementos, por isso, também é conhecido como o método da Mínima Variância (ESZERGÁR-KISS, 2016).

Eszergár-Kiss (2016) ainda afirma que o método de Ward pode ser resumido nos seguintes passos:

1. Normalização dos dados, se necessário;
2. Cálculo da distância entre *clusters*;
3. Aglutinação dos *clusters* que estão mais próximos;

É interessante ressaltar que no passo 3, se o novo *cluster* foi formado, sua distância em relação aos outros grupos deve ser recalculada (ESZERGÁR-KISS, 2016), ou seja, deve-se se repetir os passos 2 e 3 até restar apenas um *cluster*, sendo este z, o critério de parada do método hierárquico aglomerativo.

2.5 *Principal Component Analysis* (PCA)

A redução de dimensionalidade é um processo utilizado em situações em que os atributos da base de dados são redundantes ou correlacionados, ou seja, a dimensionalidade da base é maior do que o necessário (DIAS, 2013). Em casos como esse, fazem com que o tempo computacional seja maior do que preciso (SOUZA, 2019).

Schimitt (2005) acrescenta que reduzir a dimensionalidade da base de dados é uma solução a diminuição de tempo computacional necessário para a aplicação de um algoritmo, aumentando seu desempenho.

A técnica do PCA, em português “análise de componentes principais”, é uma técnica estatística multivariada utilizada para reduzir a dimensionalidade de uma base de dados que possui muitos atributos correlacionados (AIDOO, 2021).

Mishra *et al.* (2017) dizem que esta técnica usa princípios matemáticos subjacentes e sofisticados para reduzir o número dos atributos correlacionados, preservando as regiões de maior variância.

A PCA também é definida como uma técnica de extração de atributos de um determinado conjunto de dados, a qual executa uma combinação linear das variáveis originais e projeta novos dados que indicam as direções da variância máxima no espaço que explicam os principais padrões das variações nos dados originais e são adquiridos por meio do cálculo da variância-covariância da matriz da base de dados inicial. Este novo conjunto de atributos formado por variáveis não correlacionadas chamadas de componentes principais, do inglês *Principal Component* (PCs) (MAćKIEWICZ, 1993; AIDOO, 2021).

2.6 Inovação

De acordo com Feitosa (2011), na década de 1970 ocorreram muitas mudanças significativas nos processos produtivos das empresas em todo o mundo, isso devido à globalização e aos novos paradigmas tecnológicos. O aumento da concorrência internacional forçou as empresas a adotarem um processo de reestruturação industrial, procurando adaptar seu aparato produtivo às novas exigências do mercado, com produtos, serviços e processos em constante inovação, o que ainda é a realidade nos dias de hoje.

Antes mesmo das mudanças citadas acima, em 1961 foi fundada a OCDE, com o intuito de promover políticas que busquem:

- atingir o nível mais alto de desenvolvimento econômico sustentável e de emprego e um padrão de vida paulatinamente melhor em países pertencentes, ao mesmo tempo conservando a estabilidade financeira e consequentemente cooperando para o desenvolvimento da economia mundial;
- colaborar para a expansão econômica estável, para países membros e não membros no processo de desenvolvimento econômico; e

- colaborar para a expansão do comércio mundial embasada no multilateralismo e na não segregação, de conformidade com as obrigações internacionais.

No início década de 1990, a OCDE criou o Manual de Oslo, a principal fonte internacional de diretrizes para coleta e uso de dados sobre atividades inovadoras. Isso porque, segundo o Manual (2005), a habilidade de definir a escala de atividades inovadoras, os perfis das instituições inovadoras e os fatores internos e sistêmicos que são capazes de influenciar a inovação é um pré-requisito para o andamento e estudo de políticas que busquem incentivos à inovação tecnológica.

De acordo com Gault (2016), a primeira edição do Manual de Oslo foi criada em 1992 e era limitada principalmente às indústrias, apenas citando o setor de serviços, além disso abrangia somente produtos tecnológicos e inovação de processos. Já a segunda versão, elaborada em 1997, incorporou também o setor de serviços, mas ainda era limitado a produtos, processos e inserção de produtos no mercado. Por fim, em 2005, foi elaborada a terceira versão do Manual, a qual passou a definir inovação como sendo a implementação de bem ou serviço ou processo novo ou significativamente melhorado ou um novo método de marketing ou então o novo método organizacional nas práticas de negócio, dentro do local de trabalho ou em relações externas (MANUAL DE OSLO, 2005; GAULT, 2016).

Com base nas definições de inovação presentes no Manual de Oslo de 1997, o Eurostat, o Serviço de Estatística da União Europeia que é responsável por publicar estatísticas e indicadores de qualidade elevada e a nível europeu que possibilitam comparar países e regiões, elaborou, em 2004, o um instrumento de coleta chamado *The Fourth Community Innovation Survey*[1] (CIS 4), em português a Quarta Pesquisa de Inovação Comunitária, disponível em https://www.oecd.org/science/inno/37489901.pdf. Este questionário tem foco em inovação de produtos e processos, com interesse principalmente nos efeitos percebidos em relação à inovação, fontes de informação, sobre atividades inovadoras e gastos com inovação.

[1] *Community Innovation Survey* foi o primeiro *survey* baseado na metodologia do Manual de Oslo da OCDE. É realizado pela Eurostat de dois em dois anos. O CIS-4 é a quarta versão do questionário, criada em 2004.

Além disso, o instrumento também investiga os fatores que impedem ou dificultam a inovação e o uso dos direitos de propriedade intelectual. Por fim, há uma seção menor sobre inovação de marketing e organizacional.

O questionário CIS 4, foi lançado na maioria dos países em 2005, e utilizou como período de observação de 2002 a 2004, ou seja, um intervalo de três anos. Este questionário foi aplicado por volta de 30 países europeus e mais alguns não europeus.

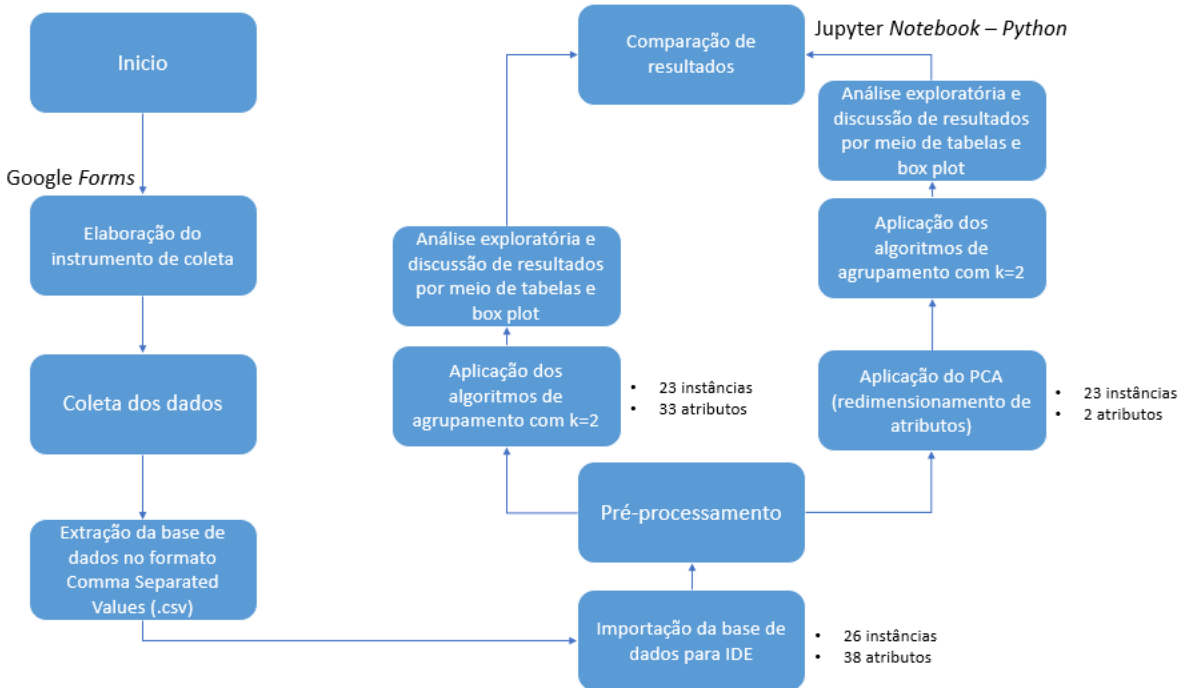
3 METODOLOGIA

Nesta seção é apresentado as etapas de pesquisa, seguido de como foi montado o questionário que deu origem ao conjunto de dados, assim como as ferramentas, bibliotecas, a linguagem de programação utilizada no trabalho e por fim uma explicação detalhada sobre o seu desenvolvimento.

3.1 Etapas da pesquisa

A Figura 4 apresenta um fluxograma das etapas da pesquisa de forma resumida e na sequência a explicação detalhada dos procedimentos realizada durante o trabalho.

Figura 4 - Etapas da pesquisa



Fonte: Autoria própria

Como representado na Figura 4, o trabalho iniciou-se pela elaboração do instrumento de coleta de dados, em seguida enviou-se o questionário por e-mail para uma amostra de empresas de Londrina e região. Depois os dados foram extraídos em formato *Comma Separeted Values* (.cvs) e importados pela IDE *Jupyter Notebook* utilizando a biblioteca *pandas* contendo 26 registros e 38 atributos.

3.2 Instrumento de coleta

O instrumento de coleta de dados utilizado neste trabalho foi construído a partir de adaptações dos questionários CIS 4 e PINTEC e enviado para algumas empresas de Londrina e Região. É importante salientar que em várias questões foram feitas em relação a um intervalo de tempo de três anos, este intervalo é o mesmo que foi utilizado em 2004 no CIS 4 e decidiu-se mantê-lo pois este questionário já estava validado. O instrumento de coleta é composto por 38 perguntas divididas em nove dimensões, as quais serão descritas nas subseções a seguir:

i. **Informações gerais da empresa:** esta dimensão tem como objetivo montar o perfil descritivo da empresa e conta com nove perguntas. Cinco destas perguntas possuem respostas categóricas nominais, como o nome da empresa (não obrigatória), cidade onde está situada, qual segmento em que está inserida, o cargo de quem está respondendo à pesquisa e o mercado geográfico que a organização atende. Em seguida há três questões com respostas categóricas ordinais, como o CEP da organização (não obrigatória), o número aproximado de funcionários e o ano em que foi fundada (não obrigatória). Por fim, tem-se uma pergunta que questiona o grau de concordância em escala Likert, de 4 pontos, onde 1 significa “Discordo Totalmente” e 4 equivale a “Concordo Totalmente”, em relação à seguinte afirmação: "A empresa considera importante ter fornecedores e/ou parceiros inovadores", esta questão tem como resposta um valor numérico escalar.

ii. **Inovação de bens e serviços:** esta dimensão conta com três perguntas com respostas categóricas nominais apresentadas no Quadro 1:

Quadro 1 – Perguntas e tipos de resposta das questões da dimensão Inovação de bens e serviços

Pergunta	Resposta
A empresa introduziu algum produto ou serviço novo ou significativamente melhorado nos	Questão dicotômica com opções de resposta: sim; não.

últimos três anos? (desconsidere mudanças de natureza exclusivamente estética);	
Quem desenvolveu essas inovações de bens ou serviços?	Questão de múltipla escolha com opções de resposta: Principalmente a própria organização ou grupo; em conjunto com outras empresas ou instituições; principalmente outras empresas ou instituições.
Essas inovações de bens ou serviços das questões anteriores podem ser consideradas:	Questão dicotômica com opções de resposta: Novas para o mercado; nova para a empresa.

Nesta dimensão, caso o entrevistado responder negativamente à primeira questão, as questões *b* e *c* serão puladas;

iii. **Inovação de processos:** esta dimensão aborda as mesmas questões da anterior, mas em relação à inovação em processos, onde a inovação de processo é considerada como a implementação de um processo de produção novo, ou melhorado de forma significativa, método de distribuição, ou atividade de suporte para os seus bens ou serviços;

iv. **Atividades inovadoras em implantação ou abandonadas:** é constituída por apenas uma questão com resposta categórica nominal e dicotômica. Nesta seção é questionado se a organização tinha alguma atividade de inovação para desenvolver produto ou processo inovador que foi abandonada nos últimos três anos ou então que ainda está em curso, onde atividades inovadoras são representadas como aquisição de maquinários, equipamentos, softwares e licenças, assim como treinamentos, marketing e Pesquisa e Desenvolvimento (P&D) quando forem especificamente realizadas para desenvolver e/ou implementar uma inovação de produto ou processo;

v. **Atividades inovadoras e despesas:** esta dimensão é composta por cinco questões com diferentes tipos de respostas, as quais serão descritas no Quadro 2:

Quadro 2 - Questões da seção Atividades inovadoras e despesa

Questões	Tipo de resposta
"A empresa investe regularmente em pesquisa e desenvolvimento (P&D)"	Grau de concordância utilizando escala Likert de quatro pontos
Nos últimos três anos, a empresa fez a aquisição de máquinas, equipamentos e hardware ou software de	Categórica nominal e dicotômica

computador avançados PARA PRODUZIR produtos e/ou processos novos ou significativamente melhorados?	
Foi feita a compra ou licenciamento de patentes e invenções não patenteadas, know-how, treinamentos ou outros tipos de conhecimento de outras empresas ou organizações (VOLTADOS para criação de um novo produto ou melhorias significativas)?	Categórica nominal e dicotômica
Houve investimento para a introdução no mercado de seus produtos e serviços novos ou significativamente melhorados, incluindo pesquisa de mercado e publicidade de lançamento?	Categórica nominal e dicotômica
Nos últimos três anos, a sua empresa recebeu algum apoio financeiro (inclua apoio financeiro por meio de créditos ou deduções fiscais, subsídios, empréstimos subsidiados e garantias de empréstimos) de órgãos públicos PARA atividades de inovação?	Categórica nominal e dicotômica

vi. **Fontes de informação e cooperação para atividades inovadoras:** esta dimensão é composta por duas questões que têm como resposta atributos categóricos nominais. A primeira é uma pergunta dicotômica e questiona se nos últimos três anos, a empresa participou de alguma atividade inovadora em cooperação com outras empresas ou instituições, neste caso cooperação em inovação é definida como uma parceria ativa com outras empresas ou instituições sem fins lucrativos a fim de realizar atividades inovadoras, onde os participantes não precisam se beneficiar comercialmente, excluindo contratação pura, ou seja, sem cooperação ativa. Em seguida é pedido que se indique a categoria da empresa ou instituição parceira, para esta pergunta é possível ter mais de uma resposta, podendo ser: fontes internas (informações advindas da própria empresa, cooperação entre setores) e/ou fontes de mercado (clientes, fornecedores, concorrentes, empresas de consultoria, institutos de desenvolvimento privado) e/ou fontes institucionais sem fins lucrativos (universidades, SENAI, SEBRAE, etc.) e/ou outras fontes como feiras, exibições, publicações, etc. A segunda questão desta seção, só deverá ser respondida se a resposta da anterior

for afirmativa, em caso negativo o respondente será direcionado à próxima seção;

vii. **Efeitos percebidos da inovação na empresa:** esta dimensão é constituída por nove questões e todas questionam o grau de concordância, ou seja, suas respostas são do tipo numérico escalar, em relação às afirmações listadas no Quadro 3:

Quadro 3 – Afirmações da seção de efeitos percebidos da inovação na empresa

"As inovações introduzidas aumentaram a gama de bens ou serviços";
"As inovações introduzidas fizeram com que a empresa entrasse em novos mercados ou aumentasse a participação no mercado";
"As inovações introduzidas aumentaram a qualidade de bens ou serviços";
"As inovações introduzidas proporcionaram uma maior flexibilidade de produção ou prestação de serviços";
"As inovações introduzidas aumentaram a capacidade de produção ou prestação de serviços";
"As inovações introduzidas fizeram com que os custos de mão de obra fossem reduzidos por unidade de produção";
"As inovações introduzidas fizeram com que o gasto de energia e matéria-prima fossem reduzidos por unidade de produção";
"As inovações introduzidas fizeram com que os impactos ambientais fossem reduzidos ou acarretaram melhoria da saúde e segurança";
"As inovações introduzidas fizeram com que a empresa atendesse a requisitos regulamentares";

viii. **Fatores que dificultam as atividades de inovação:** nesta dimensão são apresentadas cinco afirmações sobre os possíveis fatores que impossibilitam a execução de atividades ou projetos inovadores dentro da organização nos últimos três anos e o respondente deve assinalar o valor do grau de concordância em relação a cada afirmação, novamente o tipo de resposta será numérico escalar. Os possíveis fatores são apresentados no Quadro 4:

Quadro 4 – Fatores que dificultam as atividades de inovação

Fator custo (quando há falta de fundos dentro da organização ou falta financiamento de fontes externas ao empreendimento ou então consideram custos com inovação muito altos);

Fator conhecimento (escassez de mão de obra qualificada, ou falta de informação sobre tecnologias e/ou mercados, ou há dificuldade em encontrar parceiros para cooperações inovadoras);
Fator mercado (o mercado é dominado por empresas estabelecidas ou a demanda para bens ou serviços inovadores é incerta);
Fator resistência interna (a própria cultura da empresa é resiste à mudanças e a última afirmação diz que não houve razões para inovar (podendo ser pelas inovações anteriores ou então que não há demanda para inovações em seu segmento);

ix. **Direito de propriedade intelectual:** a última dimensão é composta por apenas uma pergunta, a qual pode ter mais de uma resposta, mas todas serão do tipo categórica nominal, ela questiona se nos últimos três anos a empresa solicitou uma patente, registrou um desenho industrial, registrou uma marca, reivindicou direitos autorais ou não realizou nenhuma das opções.

3.3 Ferramentas utilizadas

A elaboração do instrumento de coleta foi feita no Google *Forms*. Desenvolvimento do foi feito na linguagem Python 3.10.2, que é uma linguagem de programação muito poderosa, por isso é considerada a mais popular na área de ciência de dados, mais informações sobre a linguagem disponível em: https://docs.python.org/3.10/. Para a importação e manipulação da base de dados (pré-processamento) foi utilizada a biblioteca *pandas,* disponível em: https://pandas.pydata.org/docs/. Em seguida, para aplicação dos algoritmos foi utilizada a biblioteca *Scikit-Learn* (https://scikit-learn.org/). Por fim, foram utilizadas as bibliotecas *Matplotlib* (https://matplotlib.org/stable/index.html) e *Seaborn* (https://seaborn.pydata.org/) para a visualização dos resultados estatísticos na forma de gráficos.

3.4 Pré-processamento

Para facilitar sua identificação e a manipulação da base de dados, foi feito uma troca em relação aos nomes das variáveis, pois estavam exatamente como as

questões do instrumento de coleta. O Quadro 5 apresenta os nomes das variáveis após a troca de nome e o tipo de dado que cada uma representa:

Quadro 5 - Variáveis após troca de nome e seus tipos

Variáveis	Tipo
nome	Categórica nominal
cidade	Categórica nominal
cep	Numérica
segmento	Categórica nominal
cargo	Categórica nominal
n_funcionarios	Numérica
mercado_geografico	Categórica nominal
ano_fundacao	Numérica
s1q10_lkt	Numérica e escalar
s2q11_bin	Binária ou dicotômica
s3q12_mtpe	Categórica nominal
s3q13_bin	Binária ou dicotômica
s4q14_bin	Binária ou dicotômica
s5q15_mtpe	Categórica nominal
s6q16_bin	Binária ou dicotômica
s7q17_lkt	Numérica e escalar
s7q18_bin	Binária ou dicotômica
s7q19_bin	Binária ou dicotômica
s7q20_bin	Binária ou dicotômica
s7q21_bin	Binária ou dicotômica
s8q22_bin	Binária ou dicotômica
s9q23_cxs	Categórica nominal
s10q24_bin	Binária ou dicotômica
s11q25_lkt	Numérica e escalar
s11q26_lkt	Numérica e escalar
s11q27_lkt	Numérica e escalar
s11q28_lkt	Numérica e escalar
s11q29_lkt	Numérica e escalar
s11q30_lkt	Numérica e escalar
s11q31_lkt	Numérica e escalar
s11q32_lkt	Numérica e escalar
s11q33_lkt	Numérica e escalar
s12q34_lkt	Numérica e escalar

s12q35_lkt	Numérica e escalar
s12q36_lkt	Numérica e escalar
s12q37_lkt	Numérica e escalar
s12q38_lkt	Numérica e escalar
s13q39_cxs	Categórica nominal
Observação 1: variáveis com final "_lkt" representam questões de escala Likert com quatro pontos; variáveis com final "_bin" representam questões dicotômicas; variáveis com final "_mtpe" representam questões de múltipla escolha; variáveis com final "_cxs" representam questões de caixa de seleção.	

O Quadro 5 mostra que, com exceção das questões descritivas da empresa presentes na seção 1 do instrumento de coleta, foi determinado um código para identificação do número da seção e da questão de cada parte do questionário. Por exemplo, tem-se a questão "s1q10_lkt", onde o número da seção é na sequência de do "s" e o número da questão é sequência de "q", identificando assim a seção 1, questão 10, escala Likert.

Após a análise inicial, preferiu-se por excluir algumas variáveis como "cargo", que indica o cargo do entrevistado na empresa, e por fim, o atributo "cep", pois devido ao baixo número de instâncias não foi possível observar a concentração geográfica das empresas. Em seguida, foi feita uma padronização da variável "cidade" uma vez que, por ser uma pergunta com resposta aberta havia formas diferentes de escrevê-la, como por exemplo, escrever em caixa alta ou sem acento.

Então foi feita uma análise exploratória estatística descritiva contendo medidas de tendência central e dispersão para atributos numéricos e uma análise de frequência para dados categóricos.

A partir disto, iniciou-se a discretização utilizando o método *Label Encoder*, importado da bliblioteca *scikit-learn*, para questões que apresentavam duas opções de resposta, pois este algoritmo realiza uma binarização das variáveis, respeitando ordem alfabética, ou seja, na variável "cidade" Cambé foi substituído por 0 e Londrina por 1, assim como nas questões com resposta de dicotômica onde "Não" se tornou 0 e "Sim" 1. Além da variável "cidade" este algoritmo foi aplicado para todas as questões com "_bin".

Seguindo com a discretização dos atributos categóricos, foi utilizado o método *get_dummies* da biblioteca *pandas* para realizar a discretização de variáveis que representam questões de múltipla escolha com mais de duas opções de resposta.

Este algoritmo transforma cada opção em uma nova coluna e binariza a resposta de modo que a coluna que apresentar o número “1”, significa a resposta selecionada. Os atributos que passaram por esse processo foram: “segmento”, “s3q12_mtpe” e “s5q15_mtpe”.

Em relação da variável numérica “ano_fundacao”, optou-se por transformá-la em uma variável categórica indicando intervalos de 20 em 20 anos de idade das empresas e ficou da seguinte forma: “menos que 20 anos”, “de 21 a 40 anos”, “de 41 a 60 anos” e mais de 61 anos”. Esta transformação foi feita utilizando a biblioteca *pandas*.

Após esta transformação, foi necessário discretizar as variáveis “ano_fundacao” e “mercado_geografico”. Este processo foi realizado utilizando o método *map* da biblioteca *pandas*, transformando as variáveis categóricas ordinais em numéricas.

No caso do atributo “mercado_geografico”, primeiramente foi definido que de cinco opções seria reduzido para apenas duas: “Nacional” e “Internacional”, ou seja, se a empresa atender o mercado municipal, estadual, regional ou nacional a resposta ficaria como 0, enquanto a que atende o mercado internacional ficaria com 1 de resposta.

Depois, para as questões presentes na seção 12 “Fatores que dificultam as atividades de inovação”, por serem os únicos valores escalares com ideia de negatividade foi feita a harmonização desses atributos utilizando novamente o método *map* da biblioteca *pandas*, uma vez que para todas as outras questões com escala Likert, o maior número indicava um maior nível de inovação, enquanto nesta seria o oposto. A substituição foi feita nas questões “s12q34_lkt”, “s12q35_lkt”, “s12q36_lkt”, “s12q37_lkt” e “s12q38_lkt”, invertendo de 4 para 1, 3 para 2 e assim sucessivamente.

Por fim, foi feita a discretização das variáveis que possuíam como tipo de resposta caixa de seleção, ou seja, questões em que poderia escolher mais de uma resposta ou então deixar em branco se fosse o caso. São elas: “s9q23_cxs” e “s13q39_cxs”.

O atributo “s9q23_cxs” refere-se a questão 23 da seção 9 “Fontes de informação e cooperação para atividades inovadoras”, a questão pedia para que o entrevistado indicar se havia cooperação com alguma das categorias descritas nas opções de resposta. Primeiramente foi criado um novo conjunto de dados no qual cada resposta foi transformada em coluna ao dividi-la utilizando como limitador da

função *str.split*. Depois, cada resposta foi substituída pelo número 1 enquanto a ausência de mesma foi substituída por 0, como demonstrado no Quadro 6.

Quadro 6 - Discretização inicial do atributo "s9q23_cxs"

Fontes Internas (Informações advindas da própria empresa, cooperação entre setores)	1
Fontes de mercado (clientes, fornecedores, concorrentes, empresas de consultoria, institutos de desenvolvimento privado)	1
Fontes institucionais sem fins lucrativos (universidades, SENAI, SEBRAE, etc.)	1
Outras fontes como feiras, exibições, publicações, etc.	1
NaN	0

Após estas transformações, foram somados todos os valores presentes em cada instância do conjunto de dados criado, de maneira que a instância com maior valor é a instância que tem cooperação com maior número de fontes de informação. Logo em seguida, o atributo “s9q23_cxs” recebeu os valores desta soma.

Por último foi feito a discretização da variável “s13q39_cxs”, que corresponde à última questão do instrumento de coleta relacionada ao direito de propriedade intelectual da empresa nos últimos três anos. Esta variável foi discretizada da mesma forma que a anterior mudando apenas a forma inicial da sua discretização, como é demonstrado no Quadro 7:

Quadro 7 - Discretização inicial do atributo "s13q39_cxs"

Solicitou uma patente	1
Registrou um desenho industrial	1
Registrou uma marca	1
Reivindicou direitos autorais	1
Nenhuma das anteriores	0

Após a discretização dos atributos categóricos, foi realizada a normalização pelo método *MinMax* da biblioteca *scikit-learn* para todas as variáveis escalares e variáveis do tipo caixa de seleção.

Na última etapa do pré-processamento, foi gerado um plotado um *heatmap*, mapa de calor em português, com a biblioteca *seaborn* para identificar instâncias com dados faltantes. Foram identificadas e excluídas três instâncias que possuíam respostas em branco na dimensão sobre os efeitos percebidos na empresa em relação à introdução de inovação para aplicação dos algoritmos de agrupamento, restando apenas 23 respostas válidas.

3.5 Clusterização

Neste trabalho foram aplicados quatro tipos de clusterização, sendo eles: K-*means*, agrupamento hierárquico, PCA e K-*means*, e PCA e agrupamento hierárquico. Além disso, optou-se por fazer uma seleção dos atributos relacionados com inovação para então, aplicar os algoritmos. Para isso, foi criado um novo conjunto de dados chamado df_agrupamento com os seguintes atributos: “s1q10_lkt”, “s2q11_bin”, “s4q14_bin”, “s6q16_bin”, s7q17_lkt”, “s7q18_bin”, “s7q19_bin”, “s7q20_bin”, “s7q21_bin”, “s8q22_bin”, “s9q23_cxs”, “s10q24_bin”, “s11q25_lkt”, “s11q26_lkt”, “s11q27_lkt”, “s11q28_lkt”, “s11q29_lkt”, “s11q30_lkt”, “s11q31_lkt”, “s11q32_lkt”, “s11q33_lkt”, “s12q34_lkt”, “s12q35_lkt”, “s12q36_lkt”, “s12q37_lkt”, “s12q38_lkt”, “s13q39_cxs”, “s3q12_mtpe”, “s3q13_bin” e “s5q15_mtpe”.

Para a aplicação dos algoritmos, primeiramente utilizou-se o coeficiente da silhueta para encontrar o número ótimo de *clusters*. O teste foi realizado para k = 2, k = 3 e k = 4, com os seguintes resultados apresentados no Quadro 9:

Quadro 9 – Resultados do teste de coeficiente da silhueta

Número de *clusters* (k)	Coeficiente da silhueta
2	0,131
3	0,111
4	0,107

Como o coeficiente da silhueta para k = 2 foi maior do que os outros números, optou-se por utilizar dois *clusters* na aplicação dos algoritmos de agrupamento. Também optou-se por utilizar dois componentes principais que para a aplicação do PCA, ou seja, os 38 atributos iniciais foram reduzidos a dois, buscando aqueles que melhor explicaram a variância dos dados.

4 RESULTADOS E DISCUSSÕES

Esta seção está dividida em duas etapas: na primeira serão apresentados os resultados gerais algoritmos de agrupamento, seguido de uma análise em relação ao segmento das empresas. Por fim, é feita uma análise a partir da comparação dos resultados das técnicas de agrupamento em relação aos atributos referentes à inovação, individualmente. É importante ressaltar que os algoritmos K-*means* e K-*means* com aplicação do PCA obtiverem exatamente o mesmo resultado, portanto optou-se por utilizar apenas os gráficos do primeiro.

4.1 Resultados gerais e análise de segmento em cada *cluster*

As distribuições de objetos dos quatro algoritmos em análise ficaram razoavelmente equilibradas, com poucas diferenças, como é mostrada no Quadro 8:

Quadro 8 - Distribuição de atributos em clusters de acordo com o seu algoritmo

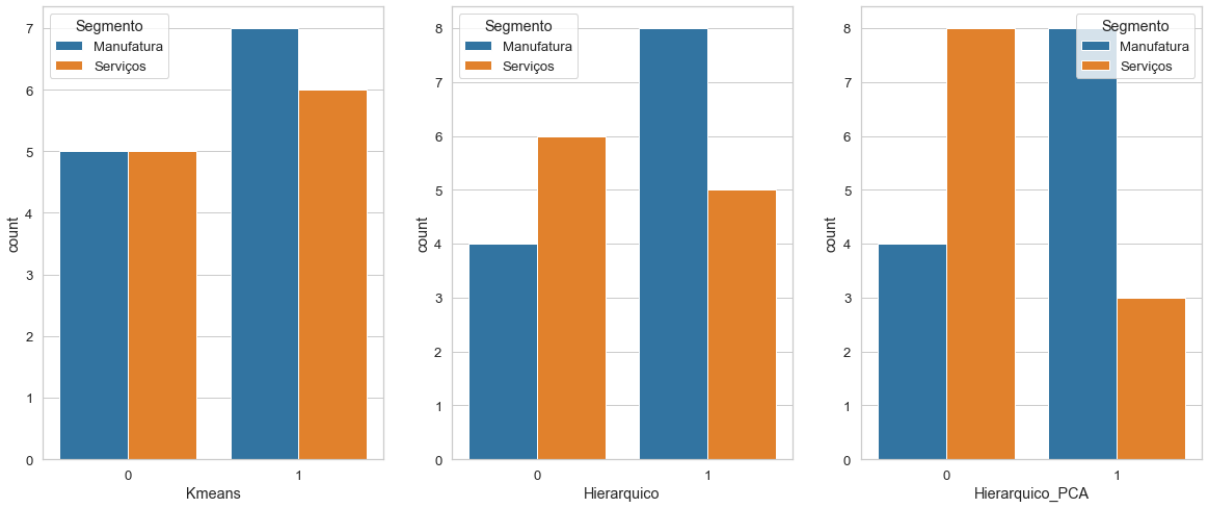
Algoritmo	***Cluster* 0**	***Cluster* 1**
K-*means*	10	13
K-*means com PCA*	10	13
Agrupamento Hierárquico	10	13
Agrupamento Hierárquico com PCA	12	11

As duas primeiras técnicas obtiveram a mesma distribuição, além de possuírem as mesmas instâncias em cada grupo.

Apesar das distribuições dos algoritmos K-*means*, K-*means* com PCA e agrupamento hierárquico estarem iguais em número, houve uma troca nas alocações de duas empresas, sendo uma do segmento da indústria e a outra prestadora de serviços, o que resultou em um mesmo número de empresas em cada *cluster*.

Para realizar a comparação entre os algoritmos no contexto das características demográficas, optou-se por selecionar ‘segmento’ por ser o único atributo que diferenciou os dois *clusters*.

Figura 5 – Comparação de resultados dos algoritmos de agrupamento em relação ao segmento das empresas



Fonte: Autoria própria

A Figura 5 mostra que como resultado, a técnica K-*means* alocou cinco empresas de serviços e cinco empresas manufatureiras no *cluster* 0. E no *cluster* 1 foram alocadas seis empresas prestadoras de serviços e sete de manufatura.

Em relação à técnica K-*means*, o agrupamento hierárquico apresenta uma divergência, o que já era esperado, uma vez que houve uma inversão na alocação de duas empresas de segmentos distintos como citado anteriormente. Neste caso a distribuição das organizações prestadoras de serviço foi de seis no *cluster* 0 e 5 no *cluster* 1 (uma a menos que a técnica anterior, em cada grupo). Já a distribuição de empresas manufatureiras foi de 4 no *cluster* 0 e 8 no 1, sendo uma a mais em cada grupo comparando com o algoritmo K-*means*.

Por fim, os resultados obtidos pelo algoritmo hierárquico com aplicação de PCA, os quais divergem ainda mais dos resultados das primeiras técnicas apresentadas. Neste algoritmo é possível verificar uma melhor separação entre empresas prestadoras de serviço e manufatureiras, o *cluster* 0 tem uma maior concentração de empresas de serviços sendo um total de 8 organizações (66,67%) para 4 empresas de manufatura (33,33%) e no *cluster* 1 foram atribuídas 8 empresas do setor industrial (62,5%) e apenas 3 do setor de serviços (37,5%).

4.2 Análise de resultados a partir de atributos relacionados à inovação

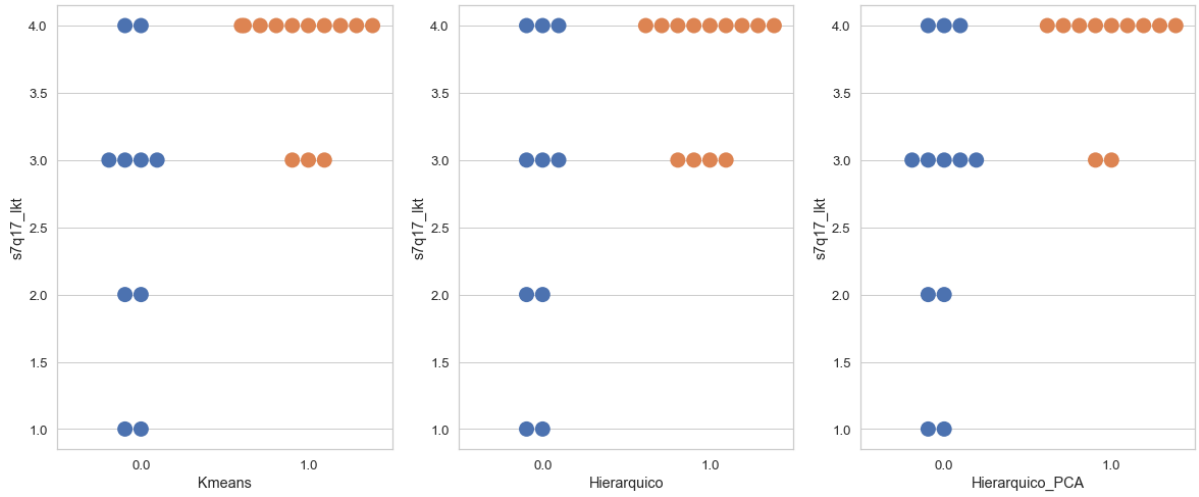
Esta seção apresenta a análise de resultados dos algoritmos agrupamentos a partir dos atributos “s7q17_lkt” (Investimento regular em P&D), “s9q23_cxs” (fontes de

cooperação para inovar) e “s13q39_cxs” (solicitações e/ou registros), individualmente. Estes atributos foram escolhidos por apresentarem diferenças entre *clusters* mais visíveis.

4.2.1 Análise de resultados dos algoritmos de agrupamento em relação ao atributo “Investimento regular em P&D”.

Nesta subseção é apresentado o gráfico *swarm plot* da biblioteca *seaborn* do atributo “s7q17_lkt”, o qual está relacionado com o grau de concordância em relação a investimentos regulares em P&D, variando entre 1 (mínimo) e 4 (máximo).

Figura 6 – Resultados dos algoritmos de agrupamento em relação a investimentos regulares em P&D



Fonte: Autoria própria

A Figura 6 mostra que os resultados obtidos por todos os algoritmos foram semelhantes. Nota-se grandes divergências entre *clusters*, sendo que os *clusters* 0 apresentam empresas que têm diferentes graus de concordância e discordância em relação à afirmação, isto é, nos *clusters* 0, há tanto empresas que investem em P&D regularmente, quanto empresas que não.

Enquanto que para os *clusters* 1, todas as empresas concordam com a afirmação, variando apenas na intensidade, com uma maior concentração no maior grau de concordância, 4. Ou seja, as empresas alocadas nos *clusters* 1 mostram um maior investimento regular em pesquisa e desenvolvimento que o outro.

Pesquisas anteriores como a de Furtado, Quadros e Domingues (2007) disseram que a intensidade de P&D é considerado um indicador internacional para comparar setores e países. Geralmente é a partir desse indicador que é mensurado o envolvimento em atividades inovadoras internas às empresas, isso ocorre por ser um dos principais fatores que impactam a introdução de inovações tecnológicas de sucesso (ANZOLA-ROMAN *et al.*, 2018; AARSTADA, KVITASTEIN, 2019).

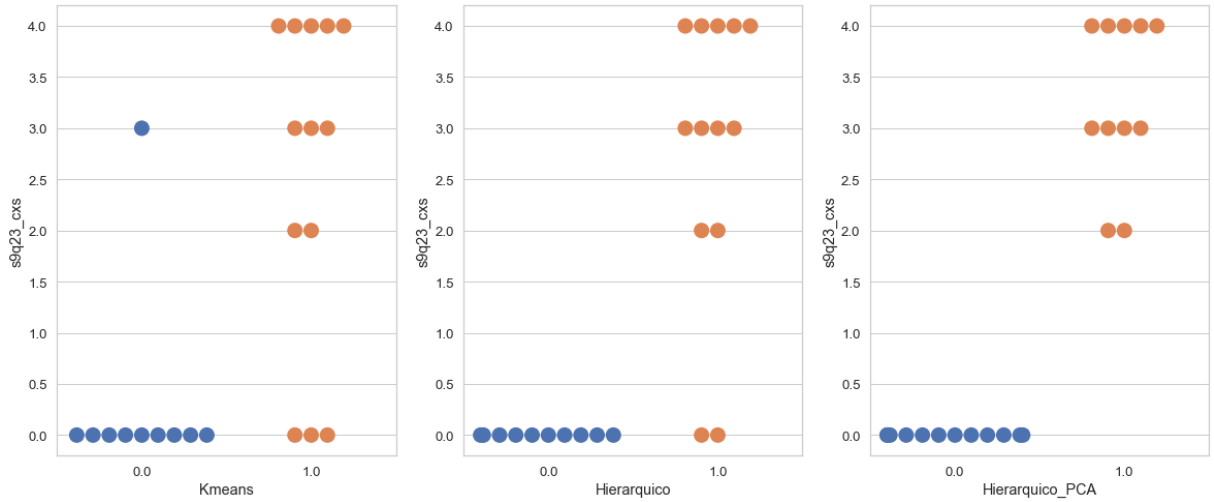
Outras pesquisas apresentaram a inovação empresarial como função de investimentos em P&D, uma vez que investimentos e esforços no desenvolvimento de competências internas do setor visam alavancar resultados inovadores (ANZOLA-ROMAN *et al.* 2018; MAIRESSE, MOHNEN, 2010).

Acs e Audretsch (2003) ainda afirmam que países mais inovadores são aqueles com maiores investimentos em P&D e as indústrias consideradas mais inovadoras tendem a ser caracterizadas por investimentos significativos em P&D e em novos conhecimentos econômicos.

4.2.2 Análise de resultados dos algoritmos de agrupamento em relação ao atributo "Fontes de cooperação para inovar"

Esta subseção apresenta uma segunda análise feita em relação ao atributo "s9q23_cxs", o qual aponta quais e quantas (variando de 0 a 4) cooperações foram feitas entre as empresas entrevistadas e outras instituições (empresas concorrentes, fornecedores, clientes, universidades, academias de ensino e governo) com o objetivo de inovar como é mostrado na Figura 7:

Figura 7 - Resultados dos algoritmos de agrupamento em relação ao número de fontes de cooperação para inovar



Fonte: Autoria própria

Os resultados em relação a este atributo também foram semelhantes, mas com algumas particularidades. Para o algoritmo K-*means,* 90% das empresas do *cluster* 0 não fizeram nenhuma parceria, apenas uma empresa desse *cluster* respondeu ter três tipos de parceria para inovar, o que pode ser considerado um *outlier*. Já para os outros dois algoritmos, hierárquico e hierárquico com PCA, todas as empresas alocadas em seus *clusters* 0 não fazem nenhum tipo de parceria com o objetivo de inovar.

Para os algoritmos K-*means* e hierárquico sem PCA, os resultados dos *clusters* 1 ficaram parecidos, no primeiro há três empresas (23%) que não faz nenhum tipo de parceria, e no segundo há duas (15%) na mesma situação.

Já o restante das organizações alocadas nesses *clusters* fazem pelo menos dois tipos diferentes de parceria, sendo 77% do K-*means* e 85% para o agrupamento hierárquico.

Por fim, no *cluster* 1 do agrupamento hierárquico com PCA, foram alocadas apenas empresas que fizeram dois ou mais tipos de parceria para inovar.

Edquist (1997) afirmou que “as empresas raramente inovam sozinhas”, para ele a inovação é o produto de uma *network* e não de uma única pessoa ou empresa. Para alguns autores, o modelo de inovação aberta e colaborativa representa uma maneira de apoiar os esforços em relação à inovação utilizando recursos externos e ausentes na empresa (LUZZINI *et al*., 2015).

Trabalhos anteriores apontam que a colaboração com diferentes atores externos (fornecedores, concorrentes, clientes e organizações de pesquisa como as universidades) melhora não só o compartilhamento de conhecimento como também a aquisição de conhecimento da empresa, o que faz com que a base de conhecimento existente se amplie e por consequência promova a capacidade de inovação de uma empresa (LUZZINI *et al*., 2015; NAJAFI-TAVANI *et al.,* 2018).

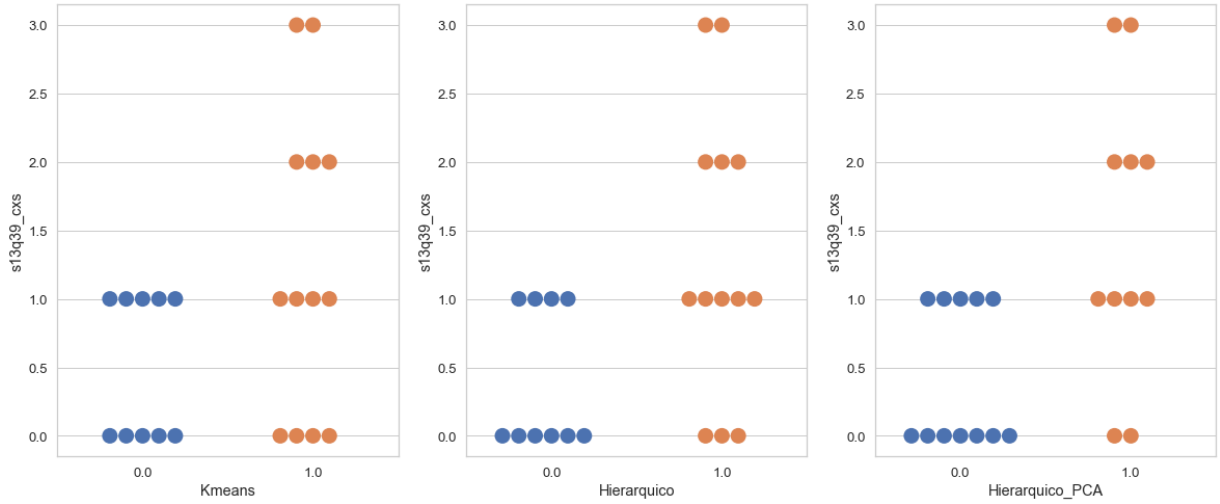
Outros autores abordam a relação entre empresas, universidades e governo como hélice tríplice e consideram essa interação como a chave para inovação, assim como crescimento econômico e social baseado em conhecimento (ETZKOWITZ, ZHOU, 2017; IATA, 2018). A hélice tríplice é um processo em contínuo desenvolvimento e tem objetivo criar um ambiente propício à inovação e empreendedorismo (ETZKOWITZ, ZHOU, 2017).

Um exemplo de sucesso da interação empresa-universidade-governo é o Vale do Silício, que inicialmente tinha interações dupla hélice entre empresa-universidade e empresa-governo, que posteriormente com auxílio dos governantes dos estados da região e com o grande intercâmbio entre as duplas hélices formaram a chama hélice tríplice (ETZKOWITZ, ZHOU, 2017).

4.2.3 Análise de resultados dos algoritmos de agrupamento em relação ao atributo "Solicitações e/ou registros"

Por fim, foi selecionado para discussão. o gráfico do atributo "s13q39_cxs", o qual questiona se a organização solicitou alguma patente e/ou registros nos últimos três anos, podendo variar de 0 a 4.

Figura 8 - Resultados dos algoritmos de agrupamento em relação ao número de solicitações e/ou registros



Fonte: Autoria própria

Os resultados dos três algoritmos foram semelhantes, onde as empresas dos *clusters* 0 estão divididas em empresas que não fizeram nenhuma solicitação ou registros nos últimos três anos e empresas que fizeram apenas uma solicitação ou registros nesse período.

Enquanto que nos *clusters* 1 formados, há tanto empresas que não fizeram nenhum tipo solicitação ou registros e empresas que fizeram até três tipos diferentes no período de três anos. Para o K-*means* 30,75% das empresas do *cluster* 1 não fizeram nenhum tipo de registro, contra 23% do agrupamento hierárquico e 18,2% do agrupamento hierárquico com PCA.

De acordo com Basberg (1987), ados e estatísticas relacionadas a este atributo são utilizados pare mensurar mudanças tecnológicas, assim como analisar a difusão de tecnologias entre países e empresas. O autor ainda afirma que o uso de estáticas sobre patentes apoia-se na suposição de que elas refletem o incentivo de atividades de inovação.

Bolívar-Ramos (2017) diz gastos com P&D e patentes são considerados dois recursos estratégicos críticos para o sucesso de uma empresa e estão amplamente relacionados, uma vez que com mais investimentos no setor de P&D mais conhecimentos as organizações irão produzir, aumentando as chances de desenvolver invenções patenteáveis. A autora também mostra o patenteamento como forma legal mais robusta de proteção dos resultados de P&D, uma vez que limita a capacidade dos concorrentes de copiar e realizar invenções duplicadas, garantindo

os retornos advindos dos investimentos em P&D à empresa e ajudando-a a manter sua vantagem competitiva derivadas da invenção (BOLÍVAR-RAMOS, 2017; CECCAGNOLI, 2009).

5 CONSIDERAÇÕES FINAIS

Este trabalho empregou técnicas de *machine learning* (ML) não supervisionadas para a formação de grupos (*clusters*) de empresas de Londrina e região, discutindo como as variáveis relacionadas à inovação se diferenciam entre os *clusters* formados.

A base de dados da pesquisa foi obtida a partir da aplicação de um instrumento de coleta criado pela autora com adaptações do CS1 4 e da PINTEC. Em seguida foi feito o pré-processamento dos dados para quantificar atributos categóricos (discretização) e em seguida a normalização de todos os atributos presentes na base de dados.

Na sequência, foi aplicado quatro algoritmos de clusterização sendo eles: K-*means* com e sem aplicação de PCA e agrupamento hierárquico com e sem aplicação de PCA. Os resultados obtidos nos primeiros algoritmos foram idênticos, tanto na quantidade quanto na alocação das instâncias em cada grupo. O algoritmo de agrupamento hierárquico obteve as mesmas quantidades de instâncias alocadas nos grupos que as técnicas de K-*means* mas com uma inversão de alocação de duas organizações, o que gerou resultados semelhantes com as primeiras técnicas.

Por fim, o último algoritmo (hierárquico com aplicação de PCA) foi o que mais teve diferença em relação a distribuição e alocação das empresas em cada *cluster*, este algoritmo conseguiu separar melhor as instâncias em relação ao atributo “segmento” (Figura 5), alocando mais prestadoras de serviços no *cluster* 0 e mais empresas manufatureiras no *cluster* 1. Apesar dessas diferenças visualizadas na seção 4.1., o algoritmo obteve resultados muito semelhantes aos do agrupamento hierárquico em relação às variáveis relacionadas à inovação.

Com os resultados discutidos na seção 4, percebe-se que para todos os algoritmos os *clusters* 1 obteve resultados melhores (*scores* maiores) no contexto de inovação, quando comparados com os *clusters* 0. Mostrando uma maior preocupação com fatores utilizados para mensuração de grau de inovação das empresas como investimento regular em P&D, maior número de parcerias com outros tipos de instituições com objetivo de inovar e solicitação e/ou registros nos últimos três anos.

Portanto, conclui-se que, para esta base de dados, os *clusters* 0 apresentam empresas com mais preocupações e esforços no contexto de inovação se comparado com os *clusters* 1.

Por fim, ao realizar uma análise entre os resultados demográficos da seção 4.1., em especial o resultado do algoritmo de agrupamento hierárquico com PCA em relação ao atributo “segmento”, onde o *cluster* 0 possui 67% de empresas no setor de serviços e o *cluster* 1 o qual apresenta 73% de indústrias, pressupõe que as indústrias dessa amostra possuem mais características inovadoras que as de serviços.

Segundo De Castro *et al*. (2020), isso ocorre devido ao conceito de inovação no setor de serviços ter surgido após a inovação em indústrias. Inicialmente o setor de serviços adotavam inovações tecnológicas produzidas nas indústrias, produziam poucas inovações em seu setor (KON, 2016).

Para Kinoshita *et al.* (2013), o aumento do número de empresas prestadores de serviço juntamente com a sua importância econômica devido à sua participação no PIB (70% do PIB brasileiro) e por ser o setor que mais emprega no Brasil, segundo IBGE (2021). Com a expansão deste setor e o aumento da concorrência, empresas prestadoras de serviços apostam na inovação para sobreviver ao mercado e alcançar vantagens competitivas (KINOSHITA *et al*., 2013).

Kon (2016) mostra que inovações elaboradas pelo setor de serviços muitas vezes passam despercebidas, por, geralmente, produzir um produto intangível se torna mais difícil definir mudanças de produção ou consumo daquele bem resultantes de inovação.

Apesar das afirmações anteriores, não há nenhum indício concreto de que isso ocorra em Londrina e região devido à pequena amostra do conjunto de dados analisado.

Para resultados melhores e mais assertivos, seria necessário obter mais respostas das empresas de Londrina e região, de forma a se obter *clusters* mais robustos buscando uma possível validação dos resultados encontrados nesta pesquisa.

REFERÊNCIAS

AARSTAD, Jarle; KVITASTEIN, Olav Andreas. Enterprise R&D investments, product innovation and the regional industry structure. **Regional Studies**, 2019.

ABREU, Diego; GARCIA, Neytfla. **Agência de notícias da indústria**, 2021. 80% das indústrias inovaram na pandemia e tiveram aumento de lucro e produtividade. Disponível em: https://noticias.portaldaindustria.com.br/noticias/inovacao-e-tecnologia/80-das-industrias-inovaram-na-pandemia-e-tiveram-aumento-de-lucro-e-produtividade/. Acesso em: 15 set. 2022.

ACS, Zoltan J.; AUDRETSCH, David B. **Innovation and technological change.** In: Handbook of entrepreneurship research. Springer, Boston, MA, 2003. p. 55-79.

AIDOO, E. N., Appiah, S. K., Awashie, G. E., Boateng, A., & Darko, G. (2021). **Geographically weighted principal component analysis for characterising the spatial heterogeneity and connectivity of soil heavy metals in Kumasi, Ghana. Heliyon,** 7(9), e08039.

ALAM, S., DOBBIE, G., KOH, Y. S., RIDDLE, P., & REHMAN, S. U. **Research on particle swarm optimization based clustering:** a systematic review of literature and techniques. Swarm and Evolutionary Computation, Elsevier, v. 17, 2014.

ALMEIDA, Juliana Pacheco de. **Estudo sobre imigrantes na Área Metropolitana de Brasília utilizando a técnica de análise de agrupamento**. 2013. https://bdm.unb.br/bitstream/10483/8040/1/2013_JulianaPachecoDeAlmeida.pdf

ANZOLA-ROMÁN, Paula; BAYONA-SÁEZ, Cristina; GARCÍA-MARCO, Teresa. **Organizational innovation, internal R&D and externally sourced innovation practices:** Effects on technological innovation outcomes. Journal of Business Research, v. 91, p. 233-247, 2018.

AMARAL, Fernando. **Aprenda Mineração de Dados: teoria e prática**. Rio de Janeiro: Alta Books, 2016.

BASBERG, Bjørn L. Patents and the measurement of technological change: a survey of the literature. **Research policy**, v. 16, n. 2-4, p. 131-141, 1987.

BEZERRA, Eduardo; GOLDSCHMIDT, Ronaldo. A Tarefa de Classificação em Text Mining. **Revista de Sistemas de Informação da FSMA, n**, v. 5, p. 42-62, 2010.

BHAVSAR, Parth. SAFRO, Ilya; BOUAYNAYA, Nidhal; POLIKAR, Robi; DERA, Dimah**. Machine learning in transportation data analytics**. In: Data analytics for intelligent transportation systems. Elsevier, 2017. p. 283-307.

BOLÍVAR-RAMOS, María Teresa. The relation between R&D spending and patents: The moderating effect of collaboration networks. **Journal of engineering and technology management**, v. 46, p. 26-38, 2017.

BRAMER, M. **Principles of Data Mining**. 3. ed. Londres: Springer London, 2016.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, v. 1, n. 1, p. 1-29, 2009.

CARVALHO, Hélio Gomes de; REIS, Dálcio Roberto dos; CAVALCANTE, Márcia Beatriz. **Gestão da inovação**. Curitiba, PR: Aymará Educação, 2011.

CASTRO, Leandro Nunes de; FERRARI, Daniel Gomes. **Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações**. São Paulo: Saraiva, v. 5, 2016.

CECCAGNOLI, Marco. **Appropriability, preemption, and firm performance**. Strategic Management Journal, v. 30, n. 1, p. 81-98, 2009.

CLAUDINO, Tiago Bomfim; SANTOS, Sandra Maria; CABRAL, Augusto Cézar Aquino; PESSOA, Maria. Naiula. **Fostering and limiting factors of innovation in Micro and Small Enterprises**. RAI Revista de Administração e Inovação, v. 14, n. 2, p. 130-139, 2017.

DE CASTRO, Rachel Gonçalves; DA SILVA, Jorge Ferreira; DE OLIVEIRA PAULA, Fábio. **Inovação de serviço e seu impacto no desempenho financeiro**. Revista Pretexto, p. 86-102, 2020.

DIAS, M. S. Regressão Construtiva Em Variedades Implícitas. Tese (Doutorado) — Pontifícia Universidade Católica Do Rio De Janeiro- PUC-RIO, 2013

DROBYAZKO, Svetlana et al. **Innovative entrepreneurship models in the management system of enterprise competitiveness**. Journal of Entrepreneurship Education, v. 22, n. 4, p. 1-6, 2019.

ESZERGÁR-KISS, Domokos; CAESAR, Bálint. Definition of user groups applying Ward's method. **Transportation Research Procedia**, v. 22, p. 25-34, 2017.

ETZKOWITZ, Henry; ZHOU, Chunyan. **Hélice Tríplice: inovação e empreendedorismo universidade-indústria-governo**. Estudos avançados, v. 31, p. 23-48, 2017.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. AI Magazine, vol. 17, no. 3, p. 37–53, 1996. Disponível em: https://ojs.aaai.org//index.php/aimagazine/article/view/1230. Acesso em: 12 dez. 2021.

FERREIRA, Anderson Vinícius Alves et al. Projeto 500 Cities: Detecção de Comunidades Utilizando Algoritmos de Clusterização. **Revista de Engenharia e Pesquisa Aplicada**, v. 3, n. 3, 2018.

FONTANA, A., NALDI, M. C. **Estudo de Comparação de Métodos para Estimação de Números de Grupos em Problemas de Agrupamento de Dados**. 2009. Universidade de São Paulo. ISSN - 0103-2569. Disponível em:http://repositorio.icmc.usp.br/bitstream/handle/RIICMC/6697/Relat%c3%b3rio%20T%c3%a9cnico_340_2009.pdf?sequence=1. Acesso em: 03 jun. 2022.

FURTADO, André; QUADROS, Ruy; DOMINGUES, Silvia **A. Intensidade de P&D das empresas brasileiras**. Inovação Uniemp, v. 3, n. 6, p. 26-27, 2007.

GALVÃO, Noemi Dreyer; MARIN, Heimar de Fátima**. Técnica de mineração de dados:** uma revisão da literatura. Acta Paulista de Enfermagem, v. 22, p. 686-690, 2009.

GAULT, Fred**. Defining and measuring innovation in all sectors of the economy. Research policy relevance**. OECD Blue Sky Forum III. Disponível em: http://www.oecd.org/sti/blue-sky-2016-agenda.htm#ps4_d2. Acesso em 20 out. 2022.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel; BEZERRA, Eduardo**. Data mining: conceitos, técnicas, algoritmos, orientações e aplicações**. 2. ed. Rio de Janeiro: Elsevier, 2015.

GUEDES, Luís. **Era da Informação: o que é e quais são os efeitos nas empresas**. FIA – Fundação Instituto de Administração. São Paulo, 25 jan. 2019. Disponível em: <https://fia.com.br/blog/era-da-informacao/#:~:text=Era%20da%20Informa%C3%A7%C3%A3o%20%C3%A9%20um,mundo%20conectado%20o%20tempo%20todo.>.

IATA, Cristiane; CUNHA, Cristiano. **A atuação da tríplice hélice em Santa Catarina pela visão dos núcleos de inovação tecnológica (NITs) do Estado**. Navus: Revista de Gestão e Tecnologia, v. 8, n. 4, p. 180-188, 2018.

KINOSHITA, KAROLINE FERREIRA; CIRANI, CLAUDIA BRITO SILVA; SILVA, W. N**. A Inovação em Serviços no Brasil: uma Comparação Internacional**. XVI Seminários em Administração, SemeAd. Anais... São Paulo, 2013.

KON, Anita**. Ecossistemas de inovação: a natureza da inovação em serviços**. Revista de Administração, Contabilidade e Economia da Fundace, v. 7, n. 1, 2016.

LUZZINI, Davide et al. The path of innovation: purchasing and supplier involvement into new product development. Industrial Marketing Management, v. 47, p. 109-120, 2015.

N, A. K.; DUBES, R. C. **Algorithms for clustering data**. Upper Saddle River, NJ, EUA: Prentice-Hall, Inc., 1988. Disponível em: http://portal.acm.org/citation.cfm?id=46712. Acesso em 14 fev. 2022.

KONG, Dejing; ZHOU, Yuan; LIU, Yufei; XUE, Lan. **Using the data mining method to assess the innovation gap:** A case of industrial robotics in a catching-up country. Technological Forecasting and Social Change, v. 119, p. 80-97, 2017.

MAćKIEWICZ, W. R. A**. Principal components analysis (pca). Computers Geosciences, v. 19, p. 118–173, 1993.**

MAIRESSE, Jacques; MOHNEN, Pierre. Using innovation surveys for econometric analysis. In: **Handbook of the Economics of Innovation**. North-Holland, 2010. p. 1129-1155.

Manual de Oslo: Diretrizes para a coleta e interpretação de dados sobre inovação. 2 edição. Paris: OCDE, 1997.

MCCUE, Colleen. **Data mining and predictive analysis: Intelligence gathering and crime analysis**. Butterworth-Heinemann, 2014.

MOURA, Karina Vargas de**. Data Science**: um estudo dos métodos no mercado e na academia. 2018.

NAJAFI-TAVANI, Saeed et al. How collaborative innovation networks affect new product performance: Product innovation capability, process innovation capability, and absorptive capacity. **Industrial marketing management**, v. 73, p. 193-205, 2018.

NATÁRIO, Maria Manuela Santos. **A Importância da Inovação no Desenvolvimento de Regiões Transfronteiriças**. Revista Chão Urbano, p. 6-54, 2011.

NING, T. P. KUMAR, V.; STEINBACH, M. **Introdução ao Data Mining – mineração de dados**. [S.l.]: Ciência Moderna, 2013.

RASCHKA, SEBASTIAN. **Python Machine Learning**. Birmingham – Mumbai, 2015. ISBN 978-1-78355-513-0.

REIS, Luiz Claudio Rezende; SÁ, Maria Irene da Fonseca e. **Big Data: Um novo campo de atuação para bibliotecários**. Prisma. com, n. 41, p. 231-250, 2020.

ŘEZANKOVÁ, H. A. N. A. Different approaches to the silhouette coefficient calculation in cluster evaluation. In: **21st International Scientific Conference AMSE Applications of Mathematics and Statistics in Economics**. 2018. p. 1-10

ROBINSON, Sherry; STUBBERUD, Hans Anton**. Issues in innovation for Norwegian SMEs**. Journal of International Business Research, v. 11, n. 1, p. 53, 2012.

SANTOS, M. O., GOLDSCHMIDT, R., CAVALCANTI, M. C., **Uma Estratégia baseada em Técnicas de KDD para apoiar o Projeto Físico em SGBD's XML Nativos**. XXII Simpósio Brasileiro de Banco de Dados. 2007. Disponível em: <https://d1wqtxts1xzle7.cloudfront.net/38678085/SBBD20_1.pdf?1441500646=&response-content-disposition=inline%3B+filename%3DUma_Estrategia_baseada_em_Tecnicas_de_KD.pdf&Expires=1639532293&Signature=UJyP22hwiZ~9Px9ezHov~awJgagrVXNSvq4z5T6kLBbTnbNbgFXd0~zcz~smOHRMKMRpcJFEYfdEZfZnC3Ed6JgnZVcUcP0yG69XsWDUL3dgn1SrORxmzM~tqd1y81zSyOEI5huFavkMEoIk-r1-jtpqo~PqrgnpLUA-p30PaLVOoIO~EG9G2NZDmybV9TL3wHZMlstooq9RNiN-pNm4~pfCq6nRHB6b4HeW1npSdamUekSz~sVBUA5XxUrLFOR3-5zqa25CPPhfqhxGn4V8AzZx~Lf84R8s6GJGvJyp-fbJ9WDq>. Acesso em: 12 dez. 2021.

SAURA, Jose Ramon; PALOS-SANCHEZ, Pedro; GRILO, Antonio. **Detecting indicators for startup business success:** Sentiment analysis using text data mining. Sustainability, v. 11, n. 3, p. 917, 2019.

SHANNON, William D. 11 Cluster Analysis. **Handbook of statistics**, v. 27, p. 342-366, Elsevier B.V. 2008.

SILVA, Douglas Eder Uno et al. **Avaliação da recuperação arquitetural de visões modulares de software a partir de técnicas de agrupamento**. 2019.

SILVA, Glauco Carlos. **Mineração de regras de associação aplicada a dados da Secretaria Municipal de Saúde de Londrina PR**. 2004. Disponível em <https://www.lume.ufrgs.br/bitstream/handle/10183/8696/000586835.pdf?sequence=1>. Acesso em: 05 jun. 2022.

SILVA, Leandro Augusto; PERES, Sarajane M.; BOSCARIOLI, Clodis. **Introdução à Mineração de Dados – Com Aplicações em R**. Rio de Janeiro: Elsevier Editora Ltda., 2016.

SCHMITT, Jeovani et al. Pré-processamento para a mineração de dados: uso da análise de componentes principais com escalonamento ótimo. 2005.

SINAGA, Kristina P.; YANG, Miin-Shen. **Unsupervised K-means clustering algorithm**. IEEE access, v. 8, p. 80716-80727, 2020.

SOUSA, Raul Pedro de Vasconcelos. **Análise dos componentes principais supervisionada: uma abordagem não-paramétrica**. 2019. Trabalho de Conclusão de Curso. Brasil.

SUN, Yexia. **Simulation of economic benefits of technological innovation based on FPGA and machine learning**. Microprocessors and Microsystems, v. 80, p. 103549, 2021.

J.H. (1963) Hierarchical groupings to optimize an objective function, Journal of the American Statistical Association, Vol. 58, pp. 236–244

YAN, Yongcai; XIA, Jing; SUN, Dong; HU, Qiqi. Research on combination evaluation of operational stability of energy industry innovation ecosystem based on machine learning and data mining algorithms. **Energy Reports**, v. 8, p. 4641-4648, 2022.

ZENGIN, Kenan; ESGI, Necmi; ERGINER, Ergin; AKSOY, Mehmet E.. **A sample study on applying data mining research techniques in educational science**: Developing a more meaning of data. Procedia Social and Behavioral Sciences, 2011. v. 15, p. 4028–4032, 2011. ISSN 1877-0428. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877042811009542> . Acesso em: 01 jun. 2022.