MINISTÉRIO DA EDUCAÇÃO
Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Colegiado do curso de Nome do Curso

CARACTERIZAÇÃO DE PROJETOS NA GESTÃO DE RECURSOS HUMANOS DE UMA EMPRESA FARMACÊUTICA UTILIZANDO A MINERAÇÃO DE DADOS

Milena Freitas Araujo
Pedro Henrique Abrahão Monteiro

João Monlevade, MG
2023

Milena Freitas Araujo
Pedro Henrique Abrahão Monteiro

CARACTERIZAÇÃO DE PROJETOS NA GESTÃO DE RECURSOS HUMANOS DE UMA EMPRESA FARMACÊUTICA UTILIZANDO A MINERAÇÃO DE DADOS

Trabalho de conclusão de curso apresentado ao curso de Engenharia de Produção do Instituto de Ciências Exatas e Aplicadas da Universidade Federal de Ouro Preto, como parte dos requisitos necessários para a obtenção do título de Bacharel em Engenharia de Produção.

Orientador: Profª. Helen de Cássia Sousa da Costa Lima

João Monlevade, MG

2023

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

M775c Monteiro, Pedro Henrique Abrahao.
Caracterização de projetos na gestão de recursos humanos de uma empresa farmacêutica utilizando a mineração de dados. [manuscrito] / Pedro Henrique Abrahao Monteiro. Milena Freitas Araujo. - 2023.
38 f.: il.: color., gráf., tab..

Orientadora: Profa. Dra. Helen de Cassia Sousa da Costa Lima.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Graduação em Engenharia de Produção .

1. Administração da produção - Processamento de dados. 2. Administração de pessoal. 3. Administração de projetos. 4. Mineração de dados (Computação). I. Araujo, Milena Freitas. II. Lima, Helen de Cassia Sousa da Costa. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 519.25:658.5

Bibliotecário(a) Responsável: Flavia Reis - CRB6-2431

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE COMPUTACAO E SISTEMAS

FOLHA DE APROVAÇÃO

Milena Freitas Araujo
Pedro Henrique Abrahão Monteiro

CARACTERIZAÇÃO DE PROJETOS NA GESTÃO DE RECURSOS HUMANOS DE UMA EMPRESA FARMACÊUTICA UTILIZANDO A MINERAÇÃO DE DADOS

Monografia apresentada ao Curso de Engenharia de Produção da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Engenharia de Produção.

Aprovada em 22 de junho de 2023, com nota 9,0 (nove).

Membros da banca

Doutora - Helen de Cassia Sousa da Costa Lima - Orientador(a) - Universidade Federal de Ouro Preto
Doutora - Janniele Aparecida Soares Araujo - Universidade Federal de Ouro Preto
Doutor - Thiago Augusto de Oliveira Silva - Universidade Federal de Ouro Preto

Helen de Cassia Sousa da Costa Lima, orientadora do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 13/07/2023.

Documento assinado eletronicamente por **Helen de Cassia Sousa da Costa Lima**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 13/07/2023, às 21:52, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0557797** e o código CRC **B1A7F401**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3808-0819 - www.ufop.br

Resumo

Este trabalho apresenta um estudo de análise de dados aplicado à gestão de projetos, utilizando o processo *Cross Industry Standard Process for Data Mining* (CRISP-DM) e técnicas de mineração de dados. A revisão da literatura aborda conceitos fundamentais de Gestão de Projetos, Gerenciamento de Recursos Humanos e o método CRISP-DM. Também são discutidas técnicas de mineração de dados, como clusterização, medidas de correlação e a análise de componentes principais. A metodologia CRISP-DM foi aplicada sequencialmente para coletar, compreender e preparar os dados do departamento de *Project Management Office* (PMO) de uma empresa farmacêutica. A análise de dados foi realizada utilizando a linguagem de programação *Python* e as bibliotecas *NumPy*, *Pandas* e *Scikit-learn*. Os resultados obtidos revelaram uma relação direta entre a complexidade dos projetos, sua duração e o tempo médio dedicado pelas pessoas envolvidas. Foram utilizadas técnicas estatísticas para identificar correlações entre os atributos dos projetos, como o tempo de planejamento e a duração do projeto. A análise de componentes principais permitiu reduzir a dimensionalidade dos dados e identificar os atributos mais relevantes para a caracterização dos projetos. Por fim, utilizou-se de bibliotecas específicas de visualização de dados do *Python* para facilitar a compreensão dos resultados e destacar as percepções obtidas. O estudo demonstrou a aplicabilidade da mineração de dados na categorização de projetos, fornecendo informações valiosas para a tomada de decisões estratégicas. Ao final dos resultados, foi possível separar os projetos selecionados em quatro *clusters*. O *Cluster* 0 é composto por projetos de complexidade Fácil e está relacionado a construções e reformas civis, além de áreas correlatas como redes de infraestrutura, elétrica, sistemas de alarme de incêndio e HVAC. Esses projetos são implementados em ambientes controlados e utilizam recursos de controle e gerenciamento, como FMS e BMS. O *Cluster* 1 está diretamente ligado à implementação e/ou alterações de sistemas ou equipamentos de automação que exigem o envolvimento da área de IT e QA é necessária para aprovar o funcionamento desses sistemas. O *Cluster* 2 abrange projetos de complexidade Difícil que envolvem adaptações ou construções civis, implementações de sistemas, máquinas e equipamentos. A área de QA está envolvida em todos os projetos desse *cluster*, pois estão relacionados à produção e é necessário comprovar sua conformidade. O *Cluster* 3 é composto por projetos de complexidade Médio a Fácil e envolve principalmente os recursos das áreas em que estão sendo implementados. Observamos uma relação direta entre a complexidade dos projetos e o envolvimento de múltiplas áreas. Projetos mais complexos demandam mais tempo e recursos humanos para sua execução. Além disso, projetos relacionados ao atributo *Capacity* estão associados a reformas estruturais e à aquisição de equipamentos que requerem peças de reposição. Este estudo contribui para a literatura ao aplicar técnicas de mineração de dados à gestão de projetos, ampliando o conhecimento sobre a utilização dessas ferramentas no contexto organizacional. Recomenda-se a realização de estudos futuros com dados mais recentes e a inclusão de outras variáveis, visando aprimorar as análises e previsões.

Palavras-chaves: Gestão de Projetos. Análise de Dados. Mineração de Dados. Clusterização. Alocação de Recursos Humanos.

Abstract

This work presents a data analysis study applied to project management using the Cross Industry Standard Process for Data Mining (CRISP-DM) and data mining techniques. The literature review covers fundamental concepts of Project Management, Human Resource Management, and the CRISP-DM method. Data mining techniques such as clustering, correlation measures, and principal component analysis are also discussed. The CRISP-DM methodology was sequentially applied to collect, understand, and prepare data from the Project Management Office (PMO) of a pharmaceutical company. Data analysis was performed using the Python programming language and libraries such as NumPy, Pandas, and Scikit-learn. The results revealed a direct relationship between project complexity, duration, and the average time dedicated by the people involved. Statistical techniques were used to identify correlations between project attributes, such as planning time and project duration. Principal component analysis helped reduce the data dimensionality and identify the most relevant attributes for project characterization. Finally, specific data visualization libraries in Python were used to facilitate result comprehension and highlight insights. The study demonstrated the applicability of data mining in project categorization, providing valuable information for strategic decision-making. At the end of the results, the selected projects were separated into four clusters. Cluster 0 consists of projects with low complexity and is related to civil construction and refurbishments, as well as related areas such as infrastructure networks, electrical systems, fire alarm systems, and HVAC. These projects are implemented in controlled environments and use control and management resources like FMS and BMS. Cluster 1 is directly linked to the implementation and/or modification of automation systems or equipment that require involvement from the IT area, and QA validation is necessary to approve the operation of these systems. Cluster 2 encompasses high-complexity projects that involve adaptations or civil construction, system implementations, machines, and equipment. QA area is involved in all projects in this cluster as they are related to production and compliance verification is necessary. Cluster 3 consists of projects with medium to low complexity and mainly involves resources from the areas in which they are being implemented. We observed a direct relationship between project complexity and the involvement of multiple areas. More complex projects require more time and human resources for their execution. Additionally, projects related to the Capacity attribute are associated with structural renovations and the acquisition of equipment that requires spare parts. This study contributes to the literature by applying data mining techniques to project management, expanding knowledge about the use of these tools in an organizational context. Further studies are recommended with more recent data and the inclusion of other variables to enhance analysis and predictions.

Keywords: Project Management. Data Analysis, Data Mining. Clustering. Allocation of Human Resources

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Em todos os lugares, é possível observar o volume de dados transmitidos diariamente (GOLDSCHMIDT; PASSOS; BEZERRA, 2015). Diante disso, tem crescido as ferramentas que ajudam as pessoas a modificar todos esses dados, a fim de torná-los úteis (CASTRO; FERRARI, 2016).

A partir disso, a mineração de dados, que é um tema multidisciplinar, foi desenvolvida com o objetivo de transformar esses dados, através da aplicação de algoritmos de maneira automática e inteligente, em conhecimentos que auxiliarão nas tomadas de decisões (CASTRO; FERRARI, 2016). Por meio da obtenção desses conhecimentos é possível utilizá-los para o controle de recursos, reconhecimento de padrões, entre outros (CASTRO; FERRARI, 2016).

Em conjunto com a mineração de dados, pode ser utilizado o PMI (do inglês, *Project Management Institute*; ou, Instituto de Gerenciamento de Projetos, em tradução livre), que é uma das principais associações mundiais em Gerenciamento de Projetos e é, atualmente, referência de padrão a ser seguido nessa área, tendo como guia prático o PMBOK (do inglês, *Project Management Body of Knowledge*; ou, Guia de Conhecimento sobre Gerenciamento de Projetos, em tradução livre) - cuja edição mais atual foi publicada em 2021. O referido guia é formado por processos distribuídos nas diversas fases de um projeto, desde a iniciação ao encerramento (PMI, 2021).

De acordo PMI (2021), um projeto conta com disciplinas relacionadas a custos, escopos, recursos humanos, comunicação, qualidade, tempo, risco, partes interessadas e aquisições. Neste trabalho, foca-se na área de gerenciamento de recursos humanos.

A análise de recursos, segundo Kerzner (2006), é um método que busca, primeiramente, mensurar a necessidade de se existirem pessoas trabalhando em um projeto ou processo. Esse é o objetivo básico desta ferramenta. Sendo assim, há uma busca constante em aumentar a confiabilidade dos recursos disponíveis para suprir a necessidade básica de cada área.

O estudo será desenvolvido em uma indústria farmacêutica multinacional possuindo como foco a produção de insulina. O departamento PMO (do inglês, *Project Management Office*; ou, Escritório de Gerenciamento de Projetos, em tradução livre) da empresa surgiu a fim de atender demandas de projetos de toda a fábrica da região atuante. Sua missão é ser um PMO estratégico que entregue soluções em gerenciamento de projetos para suprir as necessidades da empresa, tendo como visão ser reconhecido pela excelência e simplicidade na gestão de projetos.

Dentre todos os problemas enfrentados pelo departamento, pode-se citar a carência na indicação da quantidade de pessoas de outras áreas, com antecedência, que serão indispensáveis na atuação de seus projetos. Dessa forma, os gestores das áreas detentoras dos recursos são impactados por não conseguirem providenciar tudo que é necessário para a execução dos projetos.

Ao analisar a literatura sobre gestão de projetos, observou-se que são poucos os estudos que classificam projetos relacionando-os à alocação de recursos humanos e a clusterização.

Contudo, Opatha (2020), evidencia a importância da gestão de recursos humanos para o ambiente corporativo, pontuando que é uma das áreas mais críticas, e com isso, a organização necessita ter atenção. Ainda, Tang (2022), ciente da importância em gerenciar adequadamente os recursos humanos, desenvolveu um modelo de previsão para as empresas evitarem a evasão de seus colaboradores.

Com base nas perspectivas supracitadas, o trabalho justifica-se uma vez que evidencia a necessidade de tratar e analisar os dados históricos para elaborar adequadamente um projeto. Assim, este trabalho contribui para uma melhoria na alocação dos recursos humanos para a realização de projetos, visto que os supervisores de cada área terão insumos que explicam a necessidade de contratação de mão de obra fixa, temporária ou de terceiros para suprir as demandas.

1.1 Objetivos

O objetivo geral deste trabalho é caracterizar projetos de modo a auxiliar a liderança de uma empresa na tomada de decisões quanto ao planejamento de alocação de recursos humanos. Nesse contexto, serão consideradas as características de cada projeto, como áreas abordadas, complexidade, custo, duração e fases de execução.

Para cumprimento do objetivo geral é necessário atender aos seguintes objetivos específicos:

- Obtenção de dados relativos a projetos executados e em execução na referida empresa;
- Realização de tratamentos e limpeza da base de dados para utilizar técnicas de mineração de dados;
- Detecção das principais características a serem analisados para agrupar os projetos;
- Aplicação de algoritmos de mineração de dados para geração de *clusters* contendo projetos com aspectos semelhantes e os relacionando às áreas existentes;

1.2 Organização do Trabalho

O Trabalho de Conclusão de Curso foi estruturado da seguinte forma. No Capítulo 1, foi realizada a introdução do trabalho, apontando os objetivos que orientam a organização do trabalho, esclarecendo os passos adotados no desenvolvimento da pesquisa. No Capítulo 2, explicam-se os conceitos de Gestão de Projetos, *Cross Industry Standard Process for Data Mining* (CRISP-DM), Clusterização e suas especificações, Medidas de Correlação, *Principal Component Analysis* (PCA) e as tecnologias utilizadas, buscando embasamento teórico para dar prosseguimento à pesquisa. No Capítulo 3, são descritos trabalhos correlacionados ao gerenciamento de recursos humanos e clusterização. O Capítulo 4 consiste na classificação da pesquisa, levantamento e pré-processamento da base de dados, realizando todos os ajustes necessários para aplicar o algoritmo e analise dos resultados alcançados. Por fim, no Capítulo 5, conclui-se o trabalho.

2 Revisão da Literatura

Neste capítulo haverá uma revisão dos principais conceitos teóricos que nortearão o trabalho. Na Seção 2.1, apresenta-se o tema Gestão de Projetos, especificamente, sobre a área que diz respeito ao Planejamento de Recursos Humanos na Seção; na Seção 2.2, é mostrado a metodologia CRISP-DM; na Seção 2.3, evidencia-se o método da Clusterização e suas particularidades; na Seção 2.4, é discutido a Medidas de Correlação; na Seção 2.5, apresenta o método *Principal Component Analysis* (PCA); e por fim, na Seção 2.6 apresenta as tecnologias utilizadas durante a elaboração do trabalho.

2.1 Gestão de Projetos

De acordo com PMI (2021), projeto é um conjunto de atividades temporárias, previamente definidas em cronograma e realizadas em grupo, com o intuito de apresentar um resultado único. Dessa forma, faz-se necessário ter escopo e recursos bem definidos.

2.1.1 Gerenciamento de Recursos Humanos

Uma das áreas mais importantes do PMBOK consiste no gerenciamento de recursos humanos (PMI, 2021). A equipe deve possuir as competências e habilidades particulares necessárias à realização do projeto, sendo assim, o gerente de projeto é responsável por coordenar e direcionar a equipe envolvida conduzindo as atividades, alinhando prazos e analisando os resultados. O PMO do projeto descreve os processos que organizam e gerenciam a equipe do projeto. De acordo com PMI (2021), há quatro principais áreas nessa gestão:

- Planejamento de recursos humanos: é o processo de determinar como os recursos humanos serão adquiridos e gerenciados para cumprir as necessidades do projeto. Este processo envolve identificar quais tarefas do projeto requerem recursos humanos, quantos recursos serão necessários e quando eles serão necessários;
- Contratação e alocação dos colaboradores: é um processo que envolve identificar, selecionar, contratar e alocar os membros da equipe necessários para realizar o trabalho do projeto. Esta área é responsável por garantir que a equipe de projeto seja composta por pessoas com as habilidades e conhecimentos necessários para realizar as tarefas do projeto;

- Desenvolvimento da equipe do projeto: é um processo que se concentra em aprimorar a eficiência e eficácia da equipe do projeto. Este processo envolve a criação de um ambiente de trabalho colaborativo e positivo, aprimorando as habilidades interpessoais e técnicas da equipe, e fornecendo treinamento e suporte para garantir o sucesso do projeto. O objetivo é ajudar a equipe a se desenvolver, transformando-a em uma equipe coesa e eficaz, capaz de alcançar os objetivos do projeto de maneira assertiva.

- Gerenciamento da equipe do projeto: é um processo que tem como objetivo garantir que a equipe do projeto esteja alinhada e capacitada para alcançar os objetivos do projeto. Também é responsável por garantir a comunicação adequada, resolução de conflitos, motivação e desenvolvimento da equipe ao longo do projeto. O gerenciamento efetivo da equipe é fundamental para o sucesso do projeto.

2.2 *Cross Industry Standard Process for Data Mining*

O CRISP-DM (Processo Padrão de Indústria Cruzada para Mineração de Dados, em tradução livre) é uma metodologia utilizada na mineração de dados para produzir uma visão geral de todo o projeto de mineração de dados (CHAPMAN *et al.*, 2000). Conforme apresentado na Figura 1, a metodologia se divide em seis fases e possibilita a elaboração e implementação de um projeto de mineração de dados que servirá de suporte nas decisões do negócio aplicado (MORO; LAUREANO; CORTEZ, 2011). As fases são descritas em concordância com suas particularidades a seguir:

- Entendimento do problema: a parte essencial do projeto. Nessa etapa, deve-se determinar e compreender o problema. Isso é de suma importância, uma vez que é a partir dessa etapa que decidirá qual o tipo de dado será analisado posteriormente (SHEARER, 2000).

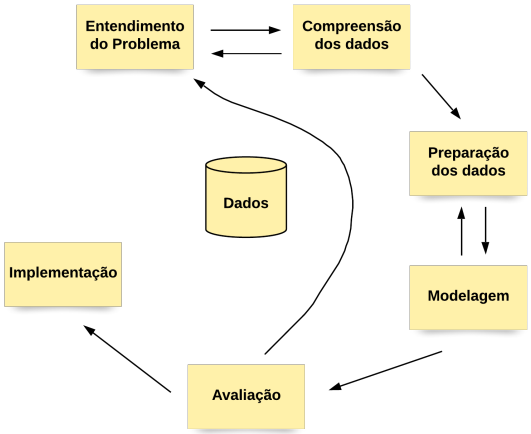
- Compreensão dos dados: nessa fase é realizada a coleta e a organização dos dados. A partir disso, iniciará o processo de mineração de dados. Isto é, a qualidade dos dados será testada, de forma que os dados mais importante serão mantidos e estudados afundo, e os menos importantes serão descartados (WIRTH, 2000);

- Preparação dos dados: é realizada as atividades necessárias para a construção da base de dados final, por meio das ferramentas de modelagem (CHAPMAN *et al.*, 2000);

- Modelagem: são escolhidas e testadas as ferramentas de modelagens de acordo com os objetivos definidos na primeira fase. Alguns métodos possuem certas particularidades, dessa forma, é necessário calibrar o modelo para encontrar valores ótimos e revisar a etapa anterior para adequar os dados conforme for preciso (SHEARER, 2000);

- Avaliação: é essencial para avaliar detalhadamente todo o projeto antes de entregá-lo, com isso, será possível alinhar o que foi produzido com os objetivos do negócio, identificando se algum ponto não foi suficientemente atendido ou se está de acordo com o planejado (WIRTH, 2000);
- Implementação: por mais que seja a última etapa, não necessariamente é o fim do projeto. Tudo que foi elaborado anteriormente deve ser organizado minuciosamente e entregue para o cliente, visto que será ele quem aplicará o modelo e, portanto, deve saber realizar todas as ações de implementação adequadamente. (CHAPMAN *et al.*, 2000).

Figura 1 – Fases do modelo de referência CRISP-DM



Fonte: Adaptado de Shearer, 2000

Dentro da fase de Modelagem, haverá a utilização da ferramenta de clusterização que é fundamental na caracterização de dados de negócios e reconhecimento de padrões (HAN; KAMBER; PEI, 2012).

2.3 Clusterização

De acordo com Ochi, Dias e Soares (2004), o processo de obtenção da solução de um problema de clusterização é realizado por meio do agrupamento de elementos de uma base de dados, resultando em grupos ou *clusters* que representam uma configuração em que cada elemento pertencente ao mesmo *cluster* possuirá maior similaridade. À vista disso, Doni (2004) fundamenta que o método tem como objetivo estabelecer uma homogeneidade entre os elementos de um mesmo *cluster* e uma heterogeneidade entre os grupos.

A clusterização é amplamente utilizada em diversas áreas de estudo com propósitos diferentes, com o intuito de organizar uma grande quantidade de dados para que possam ser entendidos de modo simples e eficiente (EVERITT *et al.*, 2011). Em consequência disso, a dificuldade de analisar todas as relações estabelecidas de todos os grupos possíveis torna-se necessário elaborar vários métodos capazes de auxiliar a formação dos *clusters* (DONI, 2004).

Ainda de acordo com Ochi, Dias e Soares (2004), os problemas de clusterização se distribuem em duas classes: *Problema de K - Clusterização* e *Problema de Clusterização Automática*. O primeiro diz respeito a problemas onde o número de *clusters* é preestabelecido, já o segundo, é o caso em que o número de *clusters* não é conhecido previamente, exigindo do algoritmo definir a quantidade necessária de *clusters* para resolver o problema (BERKHIN, 2002).

Zaïane *et al.* (2002), estabelecem critérios essenciais para que a aplicação de clusterização seja satisfatória:

- escalabilidade: o método necessita ser capaz de ser aplicado em grandes base de dados e o seu desempenho precisa diminuir conforme os dados aumentam;
- flexibilidade: os objetos a serem estudados possuem tipologias distintas, portanto, é essencial que o método de clusterização consiga ser versátil a ponto de tornar-se útil em diferentes tipos de dado;
- sintetizar dados de formas diferentes: ser capaz de condensar dados que possuem diferentes formatos e tamanhos;
- parâmetro de entrada: os parâmetros de entrada exigem uma quantidade mínima de conhecimento para que o método seja aplicado adequadamente;
- resistência a ruídos: no geral, é inevitável a presença de ruídos - pontos tangentes aos padrões da base de dados, nos problemas práticos. A partir disso, é requisito de um ótimo método de clusterização a capacidade de funcionar com sucesso na presença de ruídos;
- imparcialidade: é necessário que os resultados da aplicação do método sejam precisos. Para isso, não pode existir subjetividade no tratamento dos dados.

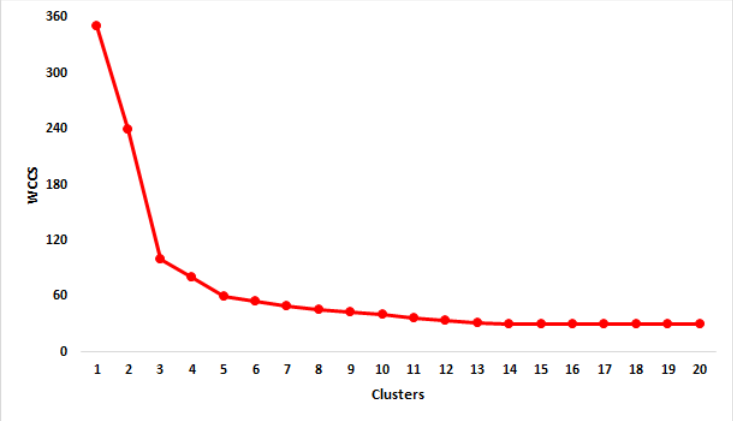
Nessa circunstância, é válido ressaltar que não existe um método de clusterização que seja apto para ser aplicado em todas as bases de dados. Dessa forma, é necessário conhecer as características da base e realizar os ajustes necessários no método para obter resultados consistentes (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001).

Ainda Halkidi, Batistakis e Vazirgiannis (2001), Ochi, Dias e Soares (2004) evidenciam que um dos métodos de clusterização amplamente utilizado é o algoritmo *k-means*. O *k-means* é um algoritmo iterativo que agrupa os objetos analisados em K *clusters*, onde K é um número pré-definido que melhor se encaixa à base de dados estudada (OCHI; DIAS; SOARES, 2004). Conforme Han, Kamber e Pei (2012), o algoritmo tem o objetivo de minimizar a soma dos quadrados das distâncias entre os pontos e os centróides dos *clusters*.

Segundo Dangeti (2017), para auxiliar na escolha do número ideal de *clusters* a ser utilizado no algoritmo *k-means*, é aplicado o *Elbow Method* (Método do Cotovelo, em tradução livre) que calcula o custo associado a cada valor de k, através da execução do algoritmo de agrupamento para diferentes quantidades de *clusters*. O custo é determinado pela métrica WCSS (do inglês, Within-Cluster Sum of Squares; ou, Soma dos Quadrados Intra-*clusters*, em tradução livre) que é obtida a partir da soma dos quadrados das distâncias dentro de cada *cluster* (DANGETI, 2017).

Em seguida, os valores de custos são plotados em um gráfico, conforme a Figura 5, com o número de *clusters* no eixo x e o WCSS para cada *cluster* no eixo y.

Figura 2 – Método do Cotovelo



Fonte: elaborado pelos autores.

Pode ser observado na Figura 5 que à medida que o número de *cluster* aumenta, os valores de WCSS diminuem rapidamente. De acordo com Dangeti (2017), o valor de k no qual a linha do gráfico se torna mais estável é chamado de cotovelo, momento em que o número de *clusters* ideal é encontrado. No caso exemplificado, o número de *clusters* seria 3.

Dangeti (2017) pontua ainda que outra métrica amplamente utilizada na clusterização é o *Silhouette Coefficient* (Coeficiente de Silhueta, em tradução livre), que é uma medida de avaliação da qualidade dos *clusters*, podendo assumir valores no intervalo de $[-1, 1]$. Portanto, valores mais altos indicam uma qualidade melhor de clusterização. O Coeficiente de Silhueta é descrito pela Equação (2.1)

$$s(i) = \frac{b(i) - a(i)}{max(b(i), a(i))} \tag{2.1}$$

Onde $a(i)$ é a distância média entre o item i e todos os outros pontos em seu *cluster* e $b(i)$ é a distância média para todos os itens no *cluster* mais próximo do item i fornecido (SILVA; BITTENCOURT; CALUMBY, 2019). Analisando a Equação (2.1), podemos observar que para $a(i) >> b(i)$ teremos $s(i)$ se aproximando de -1, indicando que o item i tem qualidade ruim. Sob outra perspectiva, para $b(i) >> a(i)$ obteremos um coeficiente $s(i)$ mais próximo de 1, indicando que o item i tem boa qualidade.

As duas métricas são usadas em conjunto para obter um número de *cluster* ainda mais efetivo.

2.4 Medida de Correlação

A correlação faz parte de uma das etapas no processo de análise de dados e é um conceito importante em estatística que descreve a relação entre duas variáveis. A correlação pode ser positiva, negativa ou neutra. Quando as variáveis aumentam juntas, a correlação é positiva, quando uma variável aumenta enquanto a outra diminui, a correlação é negativa. Quando não há relação entre as variáveis, a correlação é neutra (HAIR *et al.*, 2005).

Existem diferentes medidas de correlação que podem ser utilizadas para analisar a relação entre as variáveis. Uma medida de correlação amplamente utilizada é o coeficiente de correlação de *Spearman*, que é baseado na ordem dos dados, em vez dos valores reais das variáveis. O método de *Spearman* é descrito pela Equação (2.2).

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2.2}$$

Onde ρ representa o coeficiente de correlação de *Spearman*; d_i é a diferença entre as posições das observações nas duas variáveis; e n é o número total de observações. O coeficiente de correlação de *Spearman* assume valores entre -1 a 1, sendo que um coeficiente próximo a 1 indica uma correlação positiva perfeita, enquanto um coeficiente próximo a -1 indica uma correlação negativa perfeita (SCHOBER; BOER; SCHWARTE, 2018).

Na gestão de recursos humanos, a análise da correlação pode ser útil para identificar a relação entre diferentes variáveis, como o desempenho dos colaboradores e o tempo de trabalho, ou a satisfação dos colaboradores e a remuneração. Essa análise pode ser usada para desenvolver estratégias mais eficazes de gerenciamento de recursos humanos, com o objetivo de melhorar o desempenho dos colaboradores e aumentar a satisfação no trabalho (CHUANG; LIN, 2015).

2.5 *Principal Component Analysis*

A utilização de grandes conjuntos de dados está se tornando cada vez mais comum em diversas áreas e interpretá-los, muitas vezes, não é uma tarefa fácil (JOLLIFFE; CADIMA, 2016). Dessa forma, uma das ferramentas mais antigas e utilizadas, é o PCA (do inglês, Principal Component Analysis; ou, Análise de Componentes Principais, em tradução livre), por meio dela, é possível encontrar uma nova representação dos dados que minimize a perda de informação, ao passo que reduz a dimensionalidade deles (JOLLIFFE; CADIMA, 2016).

Conforme Abdi e Williams (2010), a ferramenta é aplicada através da elaboração da matriz de covariância dos dados e, em seguida, são encontrados os autovetores e autovalores dessa matriz. Por meio desses valores, é possível determinar as componentes principais dos dados e projetar os mesmos em um novo espaço de menor dimensão.

Bro e Smilde (2014), elencam quatro objetivos desta técnica:

- obter as informações mais relevantes do conjunto de dados;
- reduzir o tamanho do conjunto de dados mantendo somente as informações relevantes;
- tornar a descrição do conjunto de dados mais simples;
- analisar a estrutura das variáveis.

Esses objetivos destacam a capacidade do PCA de transformar um grande conjunto de dados em uma representação menor, porém mais significativa, permitindo uma análise mais eficiente e uma compreensão mais clara dos dados (BRO; SMILDE, 2014).

Ainda, Jolliffe e Cadima (2016), ressalta que a interpretação dos resultados gerados com o PCA deve ser feita com atenção, uma vez que as novas variáveis geradas não possuem uma interpretação explícita.

2.6 Tecnologias

Para a realização deste trabalho, foram utilizadas algumas tecnologias específicas, as quais serão descritas nesta seção.

2.6.1 *Python*

O *Python* [1] é uma linguagem de programação de alto nível, interpretada e interativa. Uma das principais vantagens do *Python* é a sua vasta bibliotecas disponíveis prontas para uso, bem como a possibilidade de adicionar *frameworks* de terceiros, o que amplia suas capacidades. Essa flexibilidade permite a integração do *Python* com outras linguagens de programação (BORGES, 2010).

As bibliotecas disponíveis na linguagem *Python* para pré-processamento de dados, análise estatística, algoritmos de aprendizado de máquina e técnicas de mineração de dados desempenham um papel fundamental na geração de ideias a partir de grandes volumes de dados. Essas bibliotecas capacitam a realização de análises detalhadas, permitindo uma abordagem analítica que auxilia nas tomadas de decisão estratégicas (MCKINNEY, 2012).

2.6.1.1 *NumPy*

A biblioteca *NumPy* [2] do *Python* oferece ao desenvolvedor uma ampla gama de estruturas de dados e algoritmos essenciais para realizar uma variedade de tarefas de análise de dados. Conforme McKinney (2012), a biblioteca tem os seguintes pontos:

- Funções matemáticas padrão que realizam operações rapidamente em conjuntos de dados completos;
- Possui funções que são de fácil utilização, não necessitando de uma linguagem de alto nível.

2.6.1.2 *Pandas*

O *Pandas* também é uma biblioteca do *Python* que possui ferramentas de manipulação de alto nível projetadas para tornar a análise de dados eficaz (MCKINNEY, 2012). O principal recurso desta biblioteca é o *DataFrame*, o qual desempenha um papel fundamental na análise e manipulação dos dados neste trabalho.

2.6.1.3 *Scikit-learn*

O *Scikit-learn* [3] é uma biblioteca de aprendizagem de máquina na linguagem de programação *Python* que é amplamente utilizado na área de ciência de dados. Essa biblioteca oferece implementações de diversos métodos de aprendizagem de máquina e tratamento de dados (HAO; HO, 2019). Por meio dessa biblioteca que o método da Seção 2.5 é aplicado.

[1] <https://www.python.org/>
[2] <https://numpy.org/>
[3] <https://scikit-learn.org/stable/>

2.6.1.4 Visualização de Dados

Através das bibliotecas *Matplotlib* [4], *Seaborn* [5] e *Ploty* [6] são gerados todas os gráficos presentes neste trabalho. A biblioteca *Matplotlib* é amplamente utilizada para a criação de gráficos estáticos e oferece uma ampla gama de opções de personalização. A biblioteca *Matplotlib* é amplamente utilizada na criação de gráficos estáticos, oferecendo diversas opções de personalização. O *Seaborn*, por sua vez, é uma extensão do *Matplotlib* que simplifica a criação de gráficos complexos, sobretudo para visualizações estatísticas, ao fornecer estilos estatísticos exclusivos. Já o *Plotly* é uma biblioteca que dá suporte à criação de gráficos e visualizações interativas, sendo particularmente útil na elaboração de visualizações dinâmicas para a web.

[4] <https://matplotlib.org/>
[5] <https://seaborn.pydata.org/>
[6] <https://plotly.com/>

3 Trabalhos Relacionados

Neste capítulo, serão apresentados e discutidos cinco trabalhos científicos que abordam as temáticas da alocação de recursos humanos em projetos e classificação de projetos sob diferentes perspectivas e contextos. Os trabalhos selecionados oferecem percepções valiosas, apresentando abordagens inovadoras, técnicas avançadas e boas práticas relacionadas à alocação de recursos.

Ao explorar esses trabalhos, será possível compreender as diferentes vertentes de estudo e análise que permeiam a alocação de recursos em projetos. Os estudos abrangem desde a busca por fatores de sucesso específicos (DVIR *et al.*, 1998) até a aplicação de técnicas de aprendizado de máquina em conjuntos de dados limitados (BANG *et al.*, 2022), passando pela importância da classificação de projetos para a recuperação de conhecimento em gerenciamento de projetos (BēRZIšA, 2015) e a análise da alocação de recursos humanos em projetos (KIELING *et al.*, 2021). Além disso, será apresentada uma abordagem sistemática para a alocação de recursos em projetos de *software* (OTERO *et al.*, 2009).

Com base nessas contribuições, este capítulo tem como objetivo fornecer um panorama abrangente sobre o tema da alocação de recursos em projetos, destacando os principais conceitos, desafios e soluções propostas pela literatura acadêmica. Esses estudos oferecem subsídios para aprimorar a tomada de decisões relacionadas à alocação de recursos, permitindo um planejamento mais eficiente, uma melhor utilização dos recursos disponíveis e, consequentemente, um aumento na probabilidade de sucesso dos projetos.

3.1 In search of project classification: a non-universal approach to project success factors

Dvir *et al.* (1998) oferecem uma abordagem não universal para identificar os fatores de sucesso em projetos por meio da classificação. A abordagem não universal reconhece que diferentes projetos podem ter fatores de sucesso específicos, dependendo do contexto e das características individuais. O estudo apresenta uma análise abrangente dos fatores que influenciam o sucesso dos projetos, incluindo aspectos organizacionais, técnicos e gerenciais. O objetivo é fornecer percepções sobre como identificar e gerenciar os fatores de sucesso de forma adaptável e personalizada para cada projeto, considerando suas características únicas e contexto específico.

3.2 Application of machine learning to limited datasets: prediction of project success

Bang *et al.* (2022) conduziram um estudo no qual aplicaram técnicas de aprendizado de máquina para analisar os fatores de sucesso em projetos de construção noruegueses. A pesquisa foi realizada com uma amostra de 160 projetos e utilizou uma abordagem quantitativa, obtendo dados por meio de questionários detalhados respondidos pelos membros relevantes das equipes de cada projeto. A análise quantitativa desses dados foi realizada utilizando o algoritmo *Random Forest Classifier* (Classificador de Floresta Aleatória, em tradução livre).

A Floresta Aleatória é um algoritmo de aprendizado de máquina que combina múltiplas árvores de decisão individuais para obter previsões mais precisas e assertivas (BANG *et al.*, 2022). No estudo, os pesquisadores demonstraram que é possível identificar os fatores de sucesso por meio da aplicação de técnicas de aprendizado de máquina. Os fatores de sucesso identificados corroboram a importância teoricamente reconhecida do planejamento e análise minuciosos e precoces, da complexidade ao longo do projeto, do envolvimento da liderança e dos processos que sustentam o sucesso do projeto.

Este estudo contribui para o campo de conhecimento ao aplicar o aprendizado de máquina em um contexto específico, fornecendo informações valiosas sobre a previsão do sucesso de projetos de construção e a identificação de fatores críticos para esse sucesso.

3.3 Project Management Knowledge Retrieval: Project Classification

No artigo, Bērziša (2015) discute a importância da classificação de projetos como uma maneira de organizar e acessar eficientemente o conhecimento acumulado em projetos anteriores. A classificação permite agrupar projetos com características semelhantes, facilitando a identificação de padrões, boas práticas e lições aprendidas, facilitação assim a tomada de decisões e a promoção da reutilização de soluções e práticas bem-sucedidas. Além disso, a classificação permite o compartilhamento eficiente de informações entre os membros da equipe do projeto e outras partes interessadas.

Bērziša (2015) propõe um modelo de classificação de projetos baseado em três dimensões principais: contexto, conteúdo e processo. A dimensão do contexto refere-se ao ambiente em que o projeto está inserido, incluindo a indústria, o setor e a organização. A dimensão do conteúdo aborda os objetivos, produtos e resultados do projeto. A dimensão do processo engloba as etapas, atividades e técnicas utilizadas para executar o projeto.

3.4 Human resource allocation in projects: a systematic mapping study)

Kieling *et al.* (2021) apresenta um mapeamento sistemático sobre a alocação de recursos humanos em projetos. O estudo tem como objetivo identificar as principais técnicas utilizadas para alocar recursos humanos em projetos e fornecer uma visão geral das abordagens e tendências encontradas na literatura.

Durante o estudo, foram identificadas diversas técnicas de alocação de recursos humanos em projetos, incluindo métodos baseados em modelos matemáticos, abordagens heurísticas, algoritmos genéticos e técnicas de otimização. Além disso, foram investigados fatores que influenciam a alocação de recursos humanos, como habilidades e competências dos profissionais, disponibilidade de recursos, restrições de tempo e custo, entre outros.

Em suma, o artigo de (KIELING *et al.*, 2021) conclui que a alocação eficaz de recursos humanos em projetos é um desafio complexo e multidimensional. Não há uma abordagem única que seja adequada para todos os casos, sendo necessário considerar as características específicas de cada projeto e as necessidades da organização.

3.5 A systematic approach for resource allocation in software projects

Otero *et al.* (2009) apresentam uma abordagem sistemática para otimizar a alocação de recursos humanos em projetos de *software*. Os autores identificaram que a indústria de *software* entrega poucos projetos dentro dos prazos preestabelecidos. Uma das causas para tal atraso está diretamente ligada ao tempo gasto para treinar os profissionais para adquirem as habilidades necessárias para concluir as etapas dos projetos.

Portanto, é importante desenvolver processos sistemáticos de alocação de recursos humanos que considerem conjuntos completos de habilidades dos candidatos, com o objetivo de reduzir o tempo de treinamento. Para isso, é apresentado a metodologia *Best-Fitted Resource* (Recurso mais bem ajustado, em tradução livre), que visa encontrar o recurso mais adequado para uma determinada tarefa ou projeto com base em suas habilidades e competências. A ideia é selecionar o recurso que melhor se encaixe nas exigências e requisitos específicos da tarefa, mesmo que as habilidades mais desejáveis não estejam disponíveis na equipe atual.

Os resultados do estudo fornecem diretrizes práticas para auxiliar os gestores de projetos de *software* na alocação eficiente de recursos humanos, levando em consideração as restrições e objetivos específicos de cada projeto. Destaca-se a importância de uma abordagem sistemática e baseada em dados para a alocação de recursos humanos, contribuindo assim para a melhoria da eficiência e qualidade dos projetos.

3.6 Considerações Finais

Esse capítulo apresenta cinco trabalhos científicos com foco na classificação de projetos e alocação de recursos humanos. Esses trabalhos destacaram a importância de considerar as características específicas de cada projeto, adaptando as estratégias de alocação de recursos humanos de acordo com as necessidades e requisitos específicos. Além disso, ressaltaram a relevância da utilização de abordagens baseadas em conhecimento e dados, como a aplicação de técnicas de aprendizado de máquina e a recuperação de conhecimento em gerenciamento de projetos.

As pesquisas apresentadas corroboraram para a realização do estudo desenvolvido neste trabalho, apresentando formas de coleta de dados, utilizando metodologias de alocação de recursos humanos e algoritmos de aprendizado de máquina. Portanto, com a variedade de métodos e aplicações observados nos estudos mencionados anteriormente, essas pesquisas foram guias na seleção das aplicações e métodos mais adequados ao objetivo proposto neste trabalho.

4 Metodologia e Resultados

Este capítulo continuará com a aplicação das etapas do CRISP-DM, descritas na Seção 2.2. Inicialmente, na Seção 4.1, apresenta a classificação da pesquisa; na Seção 4.2 é evidenciada a coleta de dados; na Seção 4.3, é apresentada todas as variáveis que compõem a base de dados; na Seção 4.4 são explicitados os procedimentos realizados para compreender e preparar os dados; na Seção 4.5 é iniciada a etapa de aplicação de técnicas para transformação e mineração de dados, realizando a análise de correlação dos dados na Subseção 4.5.1, e na Subseção 4.5.2, são identificados os atributos que possuem maior variabilidade; na Seção 4.6 é apresentada os *clusters* formados; e por fim, na Seção 4.6.1 é realizada as considerações finais das análises.

4.1 Classificação da Pesquisa

Este trabalho apresentará a mensuração de variáveis de pesquisa que segundo Cauchick (2012), é considerada a característica mais marcante da abordagem quantitativa que pode ser segmentada em dois tipos de pesquisas: axiomática e empírica.

De acordo com Lakatos e Marconi (2022), o conhecimento empírico é transmitido de geração para geração, baseado na experiência pessoal ou pode ser determinado através da coleta de dados em campo. Segundo Silva e Menezes (2001), a pesquisa empírica possui três classificações: empírica quantitativa, normativa e descritiva. Essa última, segundo Silva e Menezes (2001), visa descrever características de um grupo envolvendo o uso de técnicas padronizadas de coleta de dados: questionário e observação sistemática, geralmente assumindo a forma de levantamento.

Neste sentido, essa pesquisa é classificada como quantitativa empírica descritiva. Além disso, a base de dados para o estudo em questão será formada por registros históricos de arquivos padronizados e aplicados a projetos.

4.2 Coleta de dados

A coleta de dados é uma etapa fundamental no processo de análise e modelagem do problema, sendo necessário definir a fonte de referência e os dados a serem utilizados. É essencial analisar a estrutura dos dados, identificar possíveis relacionamentos entre eles e compreender os diferentes tipos de variáveis presentes, a fim de aplicar técnicas de inteligência artificial de forma adequada.

Neste trabalho, foi realizado um trabalho em conjunto com os especialistas da área do PMO para analisar e coletar os dados disponíveis de maneira mais eficiente. Durante o início de cada projeto, o departamento categoriza informações importantes e realiza o levantamento das áreas da empresa que precisaram trabalhar em conjunto para a execução do projeto.

Esses dados foram registrados em uma planilha específica, permitindo o tratamento adequado das informações contidas. A colaboração com os SMEs (do inglês, Subject Matter Experts; ou, Especialistas no Assunto, em tradução livre) da área do PMO contribuiu para garantir a integridade e relevância dos dados coletados, facilitando o processo de análise posterior.

A qualidade dos dados é crucial para desenvolver modelos de aprendizado de máquina eficientes capazes de resolver o problema em questão. A colaboração com os especialistas da área do PMO ajudou a garantir que os dados coletados fossem confiáveis e representativos dos projetos em análise.

4.3 Base de dados

A base de dados utilizada neste estudo foi construída com base no portfólio de projetos do departamento de PMO, que consta com 104 projetos e utilizando informações dos documentos base relacionados. Conforme a Tabela 1, podem ser vistos os atributos escolhidos para caracterização dos projetos.

Atributos	Nomes	Descrição	Exemplo
attr0	Complexidade	Define o grau de complexidade dos projetos	1 (Fácil) 2 (Média) 3 (Difícil)
attr1	Civil	Responsável pela construção e reformas das estruturas físicas, como edifícios, fundações e infraestrutura	0 (Não utiliza) 1 (Utiliza)
attr2	PS - HVAC	Heating, Ventilation and Air Conditioning (Aquecimento, Ventilação e Ar Condicionado) Trata do projeto, instalação e manutenção de sistemas de aquecimento, ventilação e ar condicionado	0 (Não utiliza) 1 (Utiliza)
attr3	PS - BMS	Building Management System (Sistema de Gerenciamento Predial) Controla e monitora sistemas como segurança, climatização e iluminação	0 (Não utiliza) 1 (Utiliza)
attr4	PS - FMS	Facility Management System (Sistema de Gerenciamento de Instalações) Responsável pelo gerenciamento e manutenção de instalações, incluindo planejamento, monitoramento e manutenção preventiva	0 (Não utiliza) 1 (Utiliza)
attr5	PS - Fire System	Sistema de Incêndio Engloba sistemas de detecção, alarme e combate a incêndio, garantindo a segurança contra incêndios.	0 (Não utiliza) 1 (Utiliza)
attr6	PS - Elétrica	Departamento responsável pela instalação e manutenção de sistemas elétricos e de energia	0 (Não utiliza) 1 (Utiliza)
attr7	PS - Metrologia	Responsável pela calibração, medição e certificação de equipamentos de medição utilizados na produção	0 (Não utiliza) 1 (Utiliza)
attr8	PS - Utilidades Industriais	Engloba sistemas como água, vapor, ar comprimido e gás, necessários para operações industriais	0 (Não utiliza) 1 (Utiliza)
attr9	PS - Utilidades Clean	Trata dos sistemas de instalações para produção e fornecimento de ar e água dentro dos parâmetros definidos e sua purificação	0 (Não utiliza) 1 (Utiliza)
attr10	FP - Processo	Área dedicada ao processo final de produção do produto	0 (Não utiliza) 1 (Utiliza)
attr11	IT - Network/Infra	Responsável pela infraestrutura de rede, incluindo hardware, software e conectividade de TI	0 (Não utiliza) 1 (Utiliza)
attr12	IT - SM	Information Technology Service Management (Gestão de serviços de Tecnologia da Informação) Gestão, suporte, resolução de problemas e implementação de novas soluções	0 (Não utiliza) 1 (Utiliza)
attr13	IT - OT - Project	Information Technology Operational Technology (Tecnologia da Informação - Tecnologia Operacional) Engloba sistemas de tecnologia operacional usados na produção e automação industrial	0 (Não utiliza) 1 (Utiliza)
attr14	QA - QA	Quality Assurance (Garantia da Qualidade) Garante a qualidade dos produtos e processos, incluindo testes e documentação em conformidade com padrões	0 (Não utiliza) 1 (Utiliza)
attr15	QC - QC	Quality Control (Controle de Qualidade) Controla a qualidade dos produtos durante o processo de fabricação, incluindo inspeções e testes	0 (Não utiliza) 1 (Utiliza)
attr16	Planejamento | Execução	Duração entre a fase de Planejamento e a Execução	0 a 1 (dados escalonados)
attr17	Execução | Término	Duração entre a fase de Execução e Término	0 a 1 (dados escalonados)
attr18	Business Drive	Razões e objetivos para criação de um projeto	0 a 1 (dados escalonados)
attr19	Investimento	Identifica se o projeto é classificado como Investimento de Capital (CAPEX)	0 (Não utiliza) 1 (Utiliza)
attr20	Capacity	Identifica se o projeto é classificado como Despesas Operacionais (OPEX)	0 (Não utiliza) 1 (Utiliza)
attr21	Faixa de custo	Classifica os projetos de acordo com sua faixa de custo	0 a 1 (dados escalonados)
attr22	Duração	Tempo médio dedicado por outros departamentos	0 a 1 (dados escalonados)

Tabela 1 – Descrição da base de dados

Esses atributos foram escolhidos com base na relevância para a caracterização dos projetos em *clusters* e foram extraídos dos documentos base relacionados ao portfólio de projetos do departamento de PMO.

A utilização desses atributos permitirá uma análise abrangente e multifacetada dos projetos em *clusters*, explorando aspectos financeiros, objetivos, duração, colaboração interdepartamental e contribuições específicas de cada área da empresa.

4.4 Compreensão e preparação dos dados

A limpeza de dados é uma etapa fundamental na preparação da base de dados para análise, pois visa eliminar dados inválidos, irrelevantes ou que não agregam informações relevantes. Essa etapa inclui a inspeção dos dados ausentes, que podem representar anomalias na coleta de dados, falta de conhecimento do valor em questão, situações nas quais o dado está descrito por outra variável, ou simplesmente ser a consequência da aplicação da regra de negócio. A avaliação e tratamento dessas situações são importantes para identificar valores discrepantes e deixar a base preparada para análise sem um possível viés advindo dos dados ausentes.

De acordo com a literatura, a limpeza de base de dados é uma prática amplamente utilizada em projetos de análise de dados, e tem como objetivo garantir a qualidade dos dados e eliminar possíveis erros e inconsistências. A falta de limpeza pode levar a resultados imprecisos e interpretações errôneas, comprometendo a confiabilidade das análises e decisões tomadas com base nessas análises (KANDEL *et al.*, 2012).

Diversas técnicas são utilizadas na limpeza de dados, incluindo a remoção de dados duplicados, correção de valores discrepantes tratamento de dados ausentes, entre outras. É importante ressaltar que a limpeza de dados é um processo iterativo, que pode ser realizado em várias etapas, conforme a necessidade da análise (BATISTA *et al.*, 2013).

No contexto do presente projeto, a base de dados original contemplava inicialmente 48 atributos referentes a 104 projetos. Com o intuito de simplificar e focar naqueles atributos mais relevantes para a análise proposta, foi adotado o critério de remoção dos atributos que apareciam em menos de 10% dos projetos. Esse procedimento resultou na seleção de 23 atributos que foram considerados mais significativos para o estudo em questão, possibilitando uma análise mais objetiva e direcionada.

Os atributos Planejamento | Execução, Execução | Término, Duração, *Business Drive* e Faixa de Custo tiveram seus dados escalonados utilizando a função de escalonamento da biblioteca *sklearn*[1]. Essa etapa foi realizada com o objetivo de estabelecer um limite para os dados, uma vez que eles estavam fora dos padrões dos outros atributos.

[1] <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.minmax_scale>

No entanto, é importante mencionar que os atributos *Business Drive* e Faixa de Custo são atributos que possuem sigilo de dados da empresa estudada, dessa forma, não é possível fornecer informações específicas sobre esses atributos.

4.5 Modelagem

Após a conclusão da etapa anterior, foi iniciada a etapa de aplicação de técnicas para transformação e mineração de dados, por meio da utilização de algoritmos de agrupamento.

4.5.1 Análise de Correlação

Com o intuito de obter um conhecimento útil inicial, é necessário, primeiramente, entender a relação entre as variáveis presentes na base de dados, para isso, foi aplicado o método de correlação de *Spearman*, explicado na Seção 2.4. A Tabela 2 mostra as correlações entre os pares de atributos da base de dados, em que as maiores correlações observadas estão destacadas em vermelho.

Os valores destacados representam os maiores pares de atributos correlatos:

a) Correlações positivas:

- Complexidade e Duração: possui correlação de 0,75; os projetos mais complexos geralmente exigem uma maior alocação de tempo por parte das pessoas envolvidas, dessa forma, a medida que o projeto se torna mais complexo, a quantidade de horas exigidas também aumenta;
- Planejamento | Execução e Duração: possui correlação de 0,72; o tempo de duração do planejamento para execução de um projeto e o tempo médio de horas das pessoas que trabalharam nele significa que quanto maior for o tempo de planejamento do projeto, maior tende a ser o tempo médio de horas que as pessoas trabalham nele. Um projeto que requer um tempo maior de planejamento normalmente é mais complexo e envolve diversas etapas e atividades;

– PS - BMS e PS - FMS: possui correlação de 0,88; essas áreas de BMS (do inglês, *Building Management System*; ou, Sistema de Gerenciamento de Edifícios, em tradução livre) e FMS (do inglês, *Facility Management System*); ou, Sistema de Gerenciamento de Instalações, em tradução livre) é fundamental para a eficiência e o desempenho das instalações em um contexto empresarial. A integração dessas áreas permite uma gestão mais precisa e eficiente dos sistemas prediais, otimizando o controle e o monitoramento em tempo real de parâmetros como segurança, climatização, iluminação e controle de acesso. O BMS fornece informações valiosas ao FMS, permitindo uma tomada de decisão informada e estratégica para a manutenção preventiva, a gestão de ativos e a otimização do consumo de energia. Essa correlação positiva contribui para a melhoria da qualidade e do conforto dos ocupantes, além de resultar em redução de custos operacionais e sustentabilidade no gerenciamento das instalações.

	attr0	attr1	attr2	attr3	attr4	attr5	attr6	attr7	attr8	attr9	attr10	attr11	attr12	attr13	attr14	attr15	attr16	attr17	attr18	attr19	attr20	attr21	attr22
attr0	1	0,098079	0,250940	0,273491	0,302419	0,154064	0,216767	0,230751	0,342956	0,342580	0,088449	0,186231	0,332940	0,318188	0,446547	0,035607	0,396174	0,274637	0,001126	-0,045710	0,114416	0,445358	0,753527
attr1	0,098079	1	0,417855	0,337701	0,378836	0,530580	0,308116	0,099567	0,283999	0,153406	-0,133957	0,007280	-0,109659	-0,044740	-0,148354	0,041447	0,027049	0,145034	-0,130191	0,058347	0,567790	0,313116	0,100825
attr2	0,250940	0,417855	1	0,788405	0,732467	0,700877	0,466667	0,265908	0,407795	0,285239	-0,069145	0,214187	0,088823	0,111111	0,118345	-0,016786	0,150998	0,041797	-0,023173	0,056888	0,405853	0,301896	0,153128
attr3	0,273491	0,337701	0,788405	1	0,882304	0,534881	0,491167	0,391403	0,423503	0,344060	-0,084591	0,195836	0,188353	0,168555	0,160225	0,017798	0,228769	0,038826	-0,061221	0,051037	0,303584	0,175689	0,172161
attr4	0,302419	0,378836	0,732467	0,882304	1	0,528883	0,458261	0,434484	0,464129	0,378101	-0,063113	0,161966	0,146385	0,127712	0,132026	0,110657	0,264305	0,068681	-0,121455	0,048079	0,350849	0,240249	0,244235
attr5	0,154064	0,530580	0,700877	0,534881	0,528883	1	0,457239	0,135022	0,481801	0,376536	-0,143309	0,289515	0,000000	0,063413	-0,039103	-0,098321	0,009032	0,077996	-0,001854	0,059807	0,454545	0,281052	0,077271
attr6	0,216767	0,308116	0,466667	0,491167	0,458261	0,457239	1	0,439803	0,444044	0,355298	-0,004610	0,260782	0,153960	0,232593	0,178307	0,276409	0,296051	-0,050009	-0,180746	0,094813	0,209498	0,130494	0,213491
attr7	0,230751	0,099567	0,265908	0,391403	0,434484	0,135022	0,439803	1	0,391403	0,568570	-0,079323	0,185422	0,285112	0,350323	0,272431	0,439339	0,402445	0,071238	-0,241997	0,058347	0,018060	0,053805	0,296266
attr8	0,342956	0,283999	0,407795	0,423503	0,464129	0,481801	0,444044	0,391403	1	0,662407	-0,143241	0,148026	0,188353	0,168555	0,110034	0,088988	0,207579	0,256094	-0,016514	0,051037	0,303584	0,284792	0,278830
attr9	0,342580	0,153406	0,285239	0,344060	0,378101	0,376536	0,355298	0,568570	0,662407	1	-0,157575	0,212731	0,130013	0,113095	0,197210	0,226045	0,133912	0,116513	-0,085852	0,043556	0,160330	0,173213	0,187539
attr10	0,088449	-0,133957	-0,069145	-0,084591	-0,063113	-0,143309	-0,004610	-0,079323	-0,143241	-0,157575	1	0,033206	0,071858	0,052550	0,044188	-0,117692	0,171260	0,025136	0,272490	0,049562	-0,173668	0,208118	0,164774
attr11	0,186231	0,007280	0,214187	0,195836	0,161966	0,289515	0,260782	0,185422	0,148026	0,212731	0,033206	1	0,761492	0,755292	0,413840	0,022140	0,202087	-0,085864	-0,015699	0,117358	0,093726	0,177636	0,270915
attr12	0,332940	-0,109659	0,088823	0,188353	0,146385	0,000000	0,153960	0,285112	0,188353	0,130013	0,071858	0,761492	1	0,885270	0,614940	0,029074	0,381408	0,065668	0,008225	0,098533	-0,100423	0,089771	0,440441
attr13	0,318188	-0,044740	0,111111	0,168555	0,127712	0,063413	0,232593	0,350323	0,168555	0,113095	0,052550	0,755292	0,885270	1	0,560168	0,014548	0,403181	0,049688	-0,018437	0,102398	-0,129100	0,165248	0,450704
attr14	0,446547	-0,148354	0,118345	0,160225	0,132026	-0,039103	0,178307	0,272431	0,110034	0,197210	0,044188	0,413840	0,614940	0,560168	1	0,201435	0,330358	0,126334	-0,031209	-0,068671	-0,081515	0,104766	0,483804
attr15	0,035607	0,041447	-0,016786	0,017798	0,110657	-0,098321	0,276409	0,439339	0,088988	0,226045	-0,117692	0,022140	0,029074	0,014548	0,201435	1	0,113389	-0,169504	-0,302405	0,037242	0,022774	-0,113431	0,038258
attr16	0,396174	0,027049	0,150998	0,228769	0,264305	0,009032	0,296051	0,402445	0,207579	0,133912	0,171260	0,202087	0,381408	0,403181	0,330358	0,113389	1	0,013198	-0,044199	-0,031202	-0,053893	0,293695	0,727025
attr17	0,274637	0,145034	0,041797	0,038826	0,068681	0,077996	-0,050009	0,071238	0,256094	0,116513	0,025136	-0,085864	0,065668	0,049688	0,126334	-0,169504	0,013198	1	0,166134	-0,018054	0,119436	0,360037	0,332581
attr18	0,001126	-0,130191	-0,023173	-0,061221	-0,121455	-0,001854	-0,180746	-0,241997	-0,016514	-0,085852	0,272490	-0,015699	0,008225	-0,018437	-0,031209	-0,302405	-0,044199	0,166134	1	-0,134849	-0,129877	0,109106	-0,034344
attr19	-0,045710	0,058347	0,056888	0,051037	0,048079	0,059807	0,094813	0,058347	0,051037	0,043556	0,049562	0,117358	0,098533	0,102398	-0,068671	0,037242	-0,031202	-0,018054	-0,134849	1	-0,132592	0,141010	-0,011489
attr20	0,114416	0,567790	0,405853	0,303584	0,350849	0,454545	0,209498	0,018060	0,303584	0,160330	-0,173668	0,093726	-0,100423	-0,129100	-0,081515	0,022774	-0,053893	0,119436	-0,129877	-0,132592	1	0,320619	0,094675
attr21	0,445358	0,313116	0,301896	0,175689	0,240249	0,281052	0,130494	0,053805	0,284792	0,173213	0,208118	0,177636	0,089771	0,165248	0,104766	-0,113431	0,293695	0,360037	0,109106	0,141010	0,320619	1	0,533685
attr22	0,753527	0,100825	0,153128	0,172161	0,244235	0,077271	0,213491	0,296266	0,278830	0,187539	0,164774	0,270915	0,440441	0,450704	0,483804	0,038258	0,727025	0,332581	-0,034344	-0,011489	0,094675	0,533685	1

Tabela 2 – Análise de Correlação Spearman

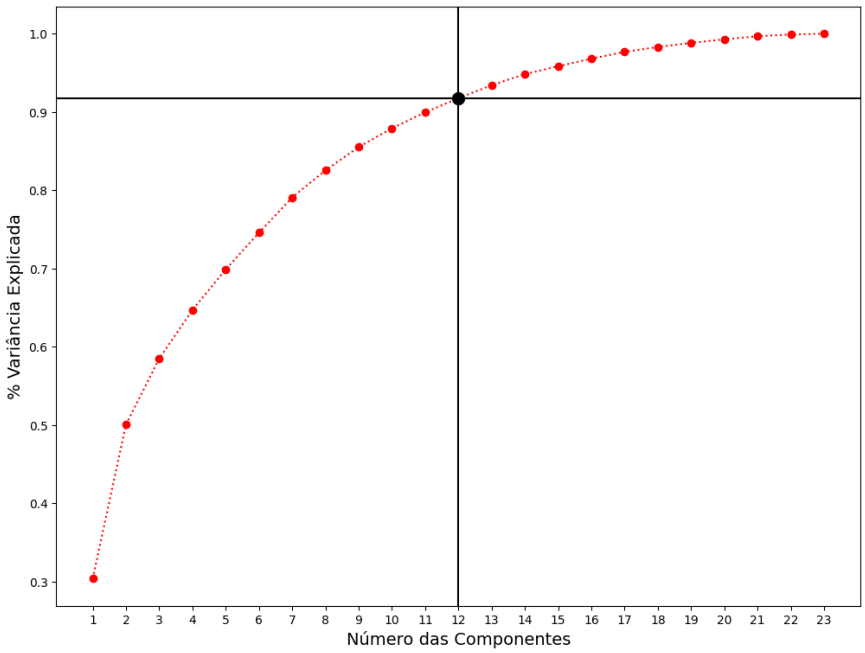
b) Correlações negativas:

- QC - QC e *Business Drive*: possui correlação de -0,3; o atributo *Business Drive* e a demanda de recursos de QC - QC sugere uma relação inversa entre esses dois atributos nos projetos analisados. Os projetos com um *Business Drive* mais alto tendem a ter uma menor necessidade de recursos de QC - QC, enquanto os projetos com um *Business Drive* mais baixo apresentam uma maior demanda por esses recursos.
- PS - Metrologia e *Business Drive*: possui correlação de -0,24; a correlação negativa pode ser explicada pelo fato de que a área de Metrologia desempenha um papel fundamental na garantia do controle de medidas e calibrações. Os projetos que estão mais focados na conformidade e nas regulamentações, como os representados pelo *Business Drive* 1, naturalmente exigirão uma maior aplicação de recursos de Metrologia. Essa relação negativa sugere que, à medida que os projetos se afastam do foco em conformidade e regulamentações, a demanda por recursos de Metrologia diminui.

4.5.2 Identificando os Atributos de Maior Variabilidade

Por causa da alta dimensionalidade da base de dados final, foi aplicada a técnica de análise de componentes principais descrita na Seção 2.5, com o objetivo de reduzir a quantidade de variáveis e facilitar a identificação das variáveis que possuem variabilidade maior.

Figura 3 – Variância Explicada Cumulativa



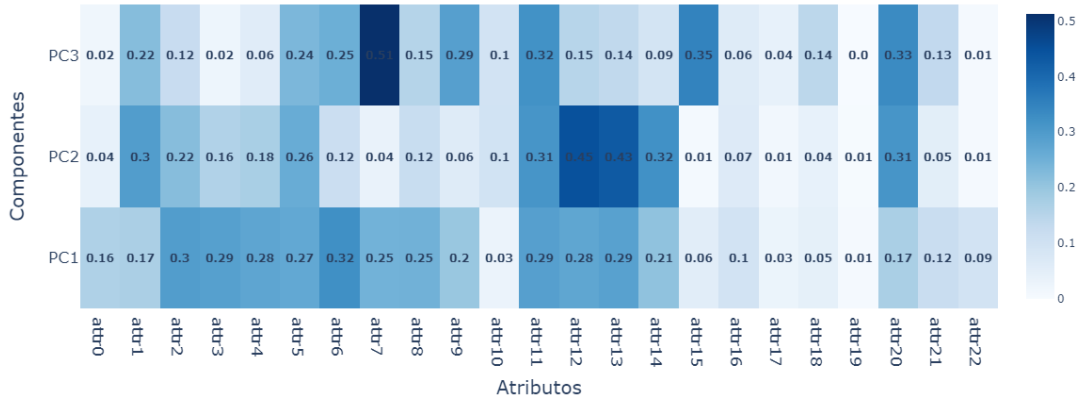
Fonte: elaborado pelos autores.

Após aplicar o PCA, foi obtido o gráfico ilustrado pela Figura 3. Esse gráfico mostra as porcentagens das variabilidades cumulativas no eixo y conforme o número de componentes aumenta no eixo x. A medida que adicionamos mais componentes, o gráfico nos mostra como a porcentagem da variabilidade total acumulada vai aumentando. Em outras palavras, ele representa a contribuição de cada componente para a explicação total da variabilidade dos dados.

Com o objetivo de preservar pelo menos 90% da variabilidade dos dados, foram escolhidas 12 componentes principais. Essas componentes foram selecionadas para serem utilizadas como entrada no algoritmo de agrupamento, garantindo a preservação de aproximadamente 91,75% das informações originais enquanto a dimensionalidade da base é reduzida.

É relevante ressaltar que foram escolhidas as três primeiras componentes principais para analisar a porcentagem dos atributos originais e que juntas representam uma variabilidade acumulada de aproximadamente 60% das informações originais da base. Essa escolha foi feita com o propósito específico de investigar a importância dos atributos originais e não com o intuito de redimensionar o tamanho da base de dados. Dessa forma, essa seleção nos permite avaliar a contribuição desses atributos na caracterização dos *clusters*, fornecendo uma compreensão mais aprofundada dos padrões presentes nos projetos após a redução da dimensionalidade.

Figura 4 – Contribuição dos 23 atributos em cada uma das 3 primeiras componentes



Fonte: elaborado pelos autores.

O gráfico da Figura 4 evidencia a contribuição de cada um dos atributos nas componentes principais. Observa-se que o eixo x representa os atributos da base de dados e o eixo y representa as componentes escolhidas. Cada célula mostra o valor correspondente a importância daquele atributo naquela componente. A medida que a célula fica escura, maior é o grau de informação que esse atributo carregará. Por outro lado, coloração mais clara indica que o atributo possui pouca importância naquela componente.

Isso significa que atributos com valores mais expressivos na componente principal têm maior contribuição na definição dos padrões e características dos *clusters*. Portanto, essa análise nos permite selecionar quais atributos são mais significativos para cada *cluster*, auxiliando na compreensão dos fatores que influenciam na sua formação e diferenciação.

Com o auxilio da Figura 4, foi possível determinar os atributos que serão utilizados para explicar os *clusters*. Os atributos estão organizados na Tabela 3.

Atributos	PC1	PC2	PC3
attr0	0,16	0,04	0,02
attr1	0,17	0,3	0,22
attr2	0,3	0,22	0,12
attr3	0,29	0,16	0,02
attr4	0,28	0,18	0,06
attr5	0,27	0,26	0,24
attr6	0,32	0,12	0,25
attr7	0,25	0,04	0,51
attr9	0,2	0,06	0,29
attr11	0,29	0,31	0,32
attr12	0,28	0,45	0,15
attr13	0,29	0,43	0,14
attr14	0,21	0,32	0,09
attr20	0,17	0,31	0,33

Tabela 3 – Atributos selecionados para caracterização dos clusters

Na Tabela 3 é possível observar que os valores em vermelho indicam a componente principal na qual cada atributo contribuiu de forma mais significativa. Esses valores indicam a magnitude da contribuição de cada atributo para as respectivas componentes principais.

Importante ressaltar que, embora o atributo Complexidade tenha valor menor em comparação aos demais atributos, ele foi escolhido para ser analisado devido à sua importância no contexto da empresa.

4.5.3 Clusterização

Em seguida à redução de dimensionalidade, utilizou-se o conjunto de componentes resultante como entrada para aplicar o algoritmo de agrupamento *k-means* [2], com o objetivo de identificar padrões nas características dos projetos. O *k-means* faz parte da classe de abordagens de clusterização, conforme explicado na Seção 2.3.

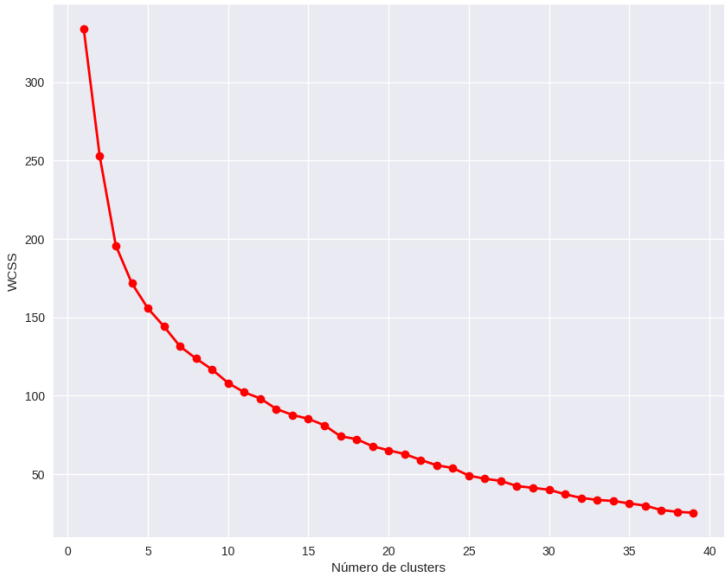
Uma etapa crucial desse processo é a escolha do número ideal de *clusters*. Para determinar o número adequado, foram empregadas duas técnicas: o método do cotovelo e o coeficiente de silhueta, ambas descritas na Seção 2.3. Essas abordagens avaliam a variabilidade interna dos dados e a estrutura dos grupos formados, permitindo encontrar o número de *clusters* que maximiza a separação entre eles e minimiza a dispersão dentro de cada grupo.

[2] <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

Portanto, é essencial realizar a escolha prévia do número ideal de *clusters*, garantindo uma análise precisa e uma interpretação significativa dos padrões identificados nas características dos projetos após a redução de dimensionalidade.

Adicionalmente, o método do cotovelo foi a primeira métrica a ser utilizada para avaliar a qualidade da clusterização, bem como auxiliar a definição do número ótimo de *clusters* para a base de dados.

Figura 5 – Método do Cotovelo



Fonte: elaborado pelos autores.

Após aplicar o método, foi obtido o gráfico da Figura 5. Observa-se que o eixo y representa o WCSS e o eixo x indica o número de *clusters*. Conforme o número de *clusters* aumenta, os valores de WCSS diminuem. Isto é, a redução dos valores de WCSS indica um ganho de informação ao aumentar a quantidade de *clusters*.

O objetivo é encontrar o número ideal de *clusters*, que é determinado quando a curva de redução do WCSS apresenta uma queda mais acentuada e, em seguida, se estabiliza. Esse ponto de inflexão no gráfico indica que, a partir desse número de *clusters*, a melhoria no ganho de informação é menor.

Posteriormente, utilizou-se a métrica coeficiente de silhueta para definir de uma forma mais efetiva a quantidade de *clusters* e comparar com o método anterior. Conforme explicado na Seção 2.3, o coeficiente ilustra o quão bem cada item foi classificado.

Quantidade de *clusters*	**Média *Silhouette***
2	0.277559
3	0.262254
4	0.278683
5	0.260173
6	0.269301
7	0.253720
8	0.269928
9	0.234400
10	0.264029
11	0.268661
12	0.266740
13	0.275499
14	0.275432
15	0.285228

Tabela 4 – Valores da Média *Silhouette*

À visto disso, calculou-se a Média *Silhouette*[3] de todos os coeficientes entre 2 a 15 *clusters*, como pode ser visto na Tabela 4. Apesar da média ter sido maior para 15 *clusters*, foi constatado que não poderia utilizar essa quantidade de *clusters*, pois não houve uma distribuição semelhante entre eles, além de possuir alguns *clusters* abaixo da média.

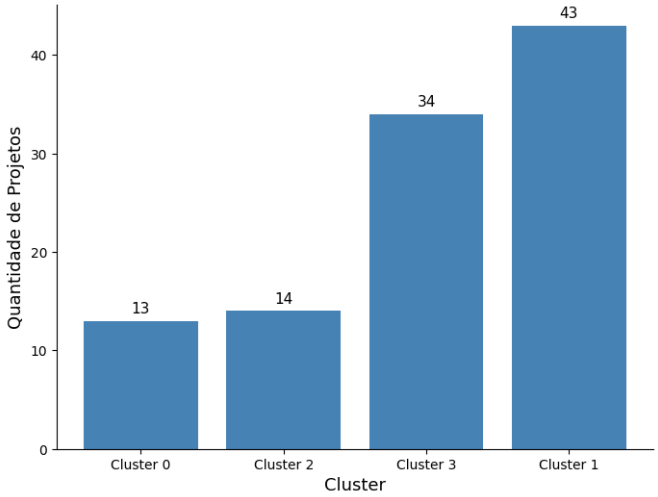
Portanto, foi determinado que utilizar 4 *clusters* é a quantidade ideal para a base de dados, uma vez que apresentou resultados mais satisfatórios e uma distribuição mais semelhante dos projetos dentro desses *clusters*, mesmo não possuindo a maior média.

4.6 Análise de Resultados

Finalizando a etapa de transformação e mineração de dados com a rotulação de cada observação em seu respectivo *cluster*, inciou-se a análise dos resultados gerados. No primeiro momento, foi quantificado os projetos pertencentes em cada um dos *clusters*.

[3] <https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html>

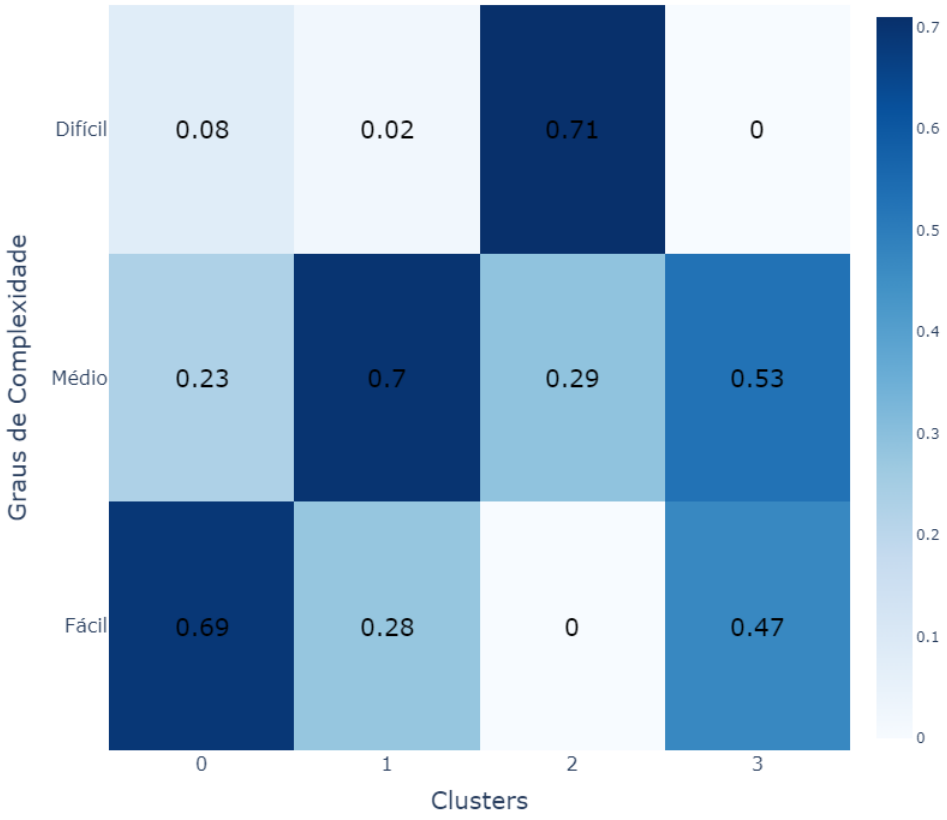
Figura 6 – Quantidade dos projetos em cada cluster



Fonte: elaborado pelos autores.

Analisando a Figura 6, observamos que 12,5% dos projetos foram agrupados no *Cluster* 0, correspondendo a 13 projetos, 41,3% dos projetos foram agrupados no *Cluster* 1, correspondendo a 43 projetos, 13,5% dos projetos foram agrupados no *Cluster* 2, correspondendo a 14 projetos, e 32,7% projetos foram agrupados no *Cluster* 3, correspondendo a 34 projetos.

Figura 7 – Relação do atributo Complexidade em cada *cluster*



Fonte: elaborado pelos autores.

O gráfico da Figura 7 relaciona a complexidade dos projetos de cada *cluster*. No eixo x representa a numeração dos *clusters* e no eixo y representa os graus de complexidade. Cada célula mostra o valor correspondente ao grau de complexidade de um *cluster* específico. A medida que a célula fica escura, uma porcentagem maior de projetos nesse *cluster* possui a complexidade indicada. Por outro lado, coloração mais clara indica uma porcentagem menor de projetos nesse *cluster* com a complexidade indicada.

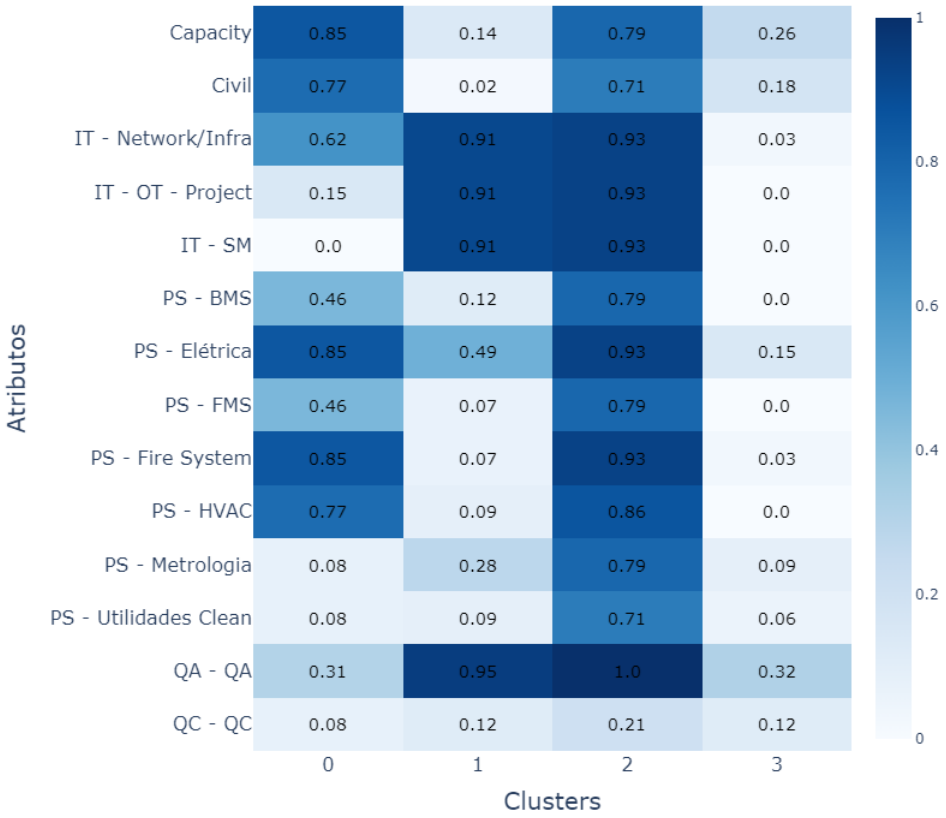
Os valores presentes na matriz são calculados dividindo-se cada grau de complexidade pelo total de projetos no respectivo *cluster*. Isso significa que cada valor representa a porcentagem da complexidade em relação ao total de projetos dentro daquele *cluster*.

A partir da Figura 7, notamos que o *Cluster* 0, que contém 13 projetos, concentra projetos com complexidade Fácil. O *Cluster* 1, que contém a maior quantidade de projetos, possui uma maior concentração de projetos com complexidade Média. O *Cluster* 2, que contém 14 projetos, está mais concentrado em projetos com complexidade Difícil. Já o *Cluster* 3, que contém 34 projetos, possui uma distribuição de complexidade Fácil e Média.

Considerando a Figura 7 é possível observar que 69% dos projetos do *Cluster* 0 foram classificados como complexidade Fácil. No entanto, uma análise minuciosa da base de dados, em conjunto com o especialista da área, revelou que os projetos com complexidade Média e Difícil não se restringem apenas a reformas ou construções, mas também envolviam a implementação de equipamentos. Essa ampliação do escopo resultou em um aumento no custo e na necessidade de controle de qualidade para a conclusão desses projetos, tornando-os logicamente mais complexos do que os demais.

Ainda, o *Cluster* 2 é composto por 71% de projetos classificados com complexidade Difícil, não incluindo projetos de complexidade Fácil em sua composição. Para uma análise mais aprofundada, foi realizada uma investigação detalhada na base de dados, com o auxílio de um especialista da área, a fim de compreender os projetos de complexidade Média nesse *Cluster*. Identificou-se que os projetos inicialmente classificados como complexidade Média deveriam ser reclassificados como complexidade Difícil. Essa constatação levou a uma ação junto ao departamento responsável, com o objetivo de revisar a matriz de complexidade utilizada e avaliar a necessidade de reformulação ou melhoria dela.

Figura 8 – Utilização dos atributos em cada *cluster*



Fonte: elaborado pelos autores.

A Figura 8 apresenta a utilização dos atributos em cada *cluster*. No eixo x é representada a numeração dos *clusters*, e no eixo y, os atributos presentes em cada projeto. As células representam a porcentagem de projetos que utilizaram a área no *cluster* específico. Ainda, a coloração escura, indicam uma maior utilização da área nesse *cluster*, enquanto células mais claras, indicam uma baixa utilização da área nesse *cluster*.

Os valores presentes na matriz são calculados dividindo-se a quantidade total de utilização da área pelo total de projetos no respectivo *cluster*. Dessa forma, cada valor representa a porcentagem de utilização em relação ao total de projetos dentro daquele *cluster*. Essa abordagem permite ter uma visão proporcional da utilização de cada área em relação ao tamanho do cluster.

Ao analisar os *Clusters* 2 e 0, observou-se uma maior alocação nas áreas de Elétrica, Civil, HVAC, BMS, FMS, *Fire System* e IT - Network/Infra . Essas áreas citadas estão intrinsecamente relacionadas, abrangendo aspectos de automação, controle de sistemas, infraestrutura e segurança. A presença conjunta dessas áreas nos projetos desses *Clusters* indica uma correlação lógica, em que o escopo dos projetos envolve reformas ou construções estruturais, reforçado pela área de Civil presente nesses *Clusters* e, consequentemente, alteração de sistemas que são utilizados para que os prédios atinjam os padrões estabelecidos de ar, ventilação, temperatura, redes e alarmes de incêndio.

Comparativamente, o *Cluster* 0 apresenta uma maior proporção de projetos concentrados nessas áreas específicas em comparação ao *Cluster* 2, sugerindo que o primeiro tenha um enfoque mais direcionado para reformas e construções estruturais.

Além disso, o atributo *Capacity* também é mais utilizado nos *Clusters* 2 e 0, uma vez que, em sua maioria, esses projetos envolvem a aquisição de mobiliários ou itens multimídia para a adequação do espaço. Para o *Clusters* 2, diferentemente do *Cluster* 0, a predominância desse atributo em seus projetos ocorre também pelo fato de ser necessária a compra de peças de reposições.

Analisando os três atributos relacionados a IT, ou seja, IT - Network/Infra, IT - SM e IT - OT - Project, nos *Clusters* 1 e 2, verifica-se que essas áreas são mais amplamente utilizadas, representando 91% e 93% dos projetos, respectivamente, em cada uma delas. Embora pertençam à mesma área, a utilização simultânea desses três atributos ocorre quando os projetos envolvem tanto alterações prediais quanto a necessidade de modificações nos sistemas, o que aumenta a complexidade dos projetos.

O *Cluster* 1 é caracterizado pela representatividade dos atributos de IT e QA (95%) em seu conjunto. Com o auxílio da especialista, verificou-se que essas áreas caminham juntas, uma vez que as alterações nos sistemas de automação da empresa geralmente exigem validação por parte QA para retomar a operação normal. Portanto, são projetos majoritariamente ligados a implementação de sistemas.

Foi possível constatar que o *Cluster* 2 apresenta uma maior porcentagem de utilização em quase todas as áreas, com exceção da área Civil, que é mais representativa no *Cluster* 0. Diferentemente dos outros *clusters*, o *Cluster* 2 é o único que apresenta uma maior proporção de uso da área de Metrologia (79%) em sua execução. Os projetos que indicam a necessidade de alocação dessa área são aqueles ligados a aquisição de máquinas, equipamentos ou sistemas que precisam ser calibrados e quando estes estão ligadas a produção, a calibração precisa ser validada pelo QA, que é departamento que garante a qualidade dos processos e aparece em 100% dos projetos deste agrupamento. Analisando a base de dados, juntamente com o especialista da área, foi possível notar a alocação de QA está presente em 90% dos projetos que contemplam Metrologia. Essa integração entre as áreas é essencial para garantir a qualidade global dos projetos, considerando não apenas a precisão das medições, mas também outros aspectos relevantes, como a conformidade com normas e regulamentações específicas.

Além disso, o *Cluster* 3 é o que apresenta a menor proporção das outras áreas em sua execução, sendo composto majoritariamente por recursos da própria área para a realização dos projetos e apenas uma ligeira participação em alguns projetos da área de QA para realizar as validações necessárias, mas no geral, os projetos não dependem de outras áreas. Os projetos desse *cluster* estão especificados nas faixas iniciais de custo e não possuem projetos de complexidade Difícil em seu portfólio.

4.6.1 Considerações Finais

Com base nas análises realizadas, obtivemos informações valiosas sobre os projetos do departamento de PMO.

O *Cluster* 0 é caracterizado por projetos de complexidade Fácil e está diretamente relacionado a construções e reformas civis, bem como às áreas correlatas, como adequações de redes (*IT - Network/Infra*), elétrica, sistemas de alarmes de incêndio (*Fire System*) e HVAC. Os projetos inseridos nesse *cluster*, que são implementados em ambientes controlados, utilizam os recursos relacionados a FMS e BMS para controle e gerenciamento desses locais.

Por sua vez, o *Cluster* 1 está diretamente ligado à implementação e/ou alterações de sistemas ou equipamentos de automação que exigem o envolvimento das áreas de TI. Além disso, a validação por parte da área de QA é necessária para aprovar o funcionamento desses sistemas.

O *Cluster* 2 é caracterizado por uma considerável utilização dos atributos na maioria dos casos. Ele abrange projetos que envolvem não apenas adaptações ou construções civis, mas também implementações de sistemas, máquinas e equipamentos. Nesses casos, a área de QA precisa estar envolvida em 100% dos projetos, uma vez que todos estão relacionados à produção e é necessário comprovar sua conformidade. Esse *cluster* é predominantemente de complexidade Difícil devido às particularidades envolvidas em sua execução.

O *Cluster* 3 é caracterizado por projetos que não possuem atributos com contribuições em destaque em comparação com os demais. São projetos que circulam entre a complexidade Média e Fácil, há uma pequena participação de QA e envolvem principalmente os recursos das próprias áreas nas quais estão sendo implementados.

A caracterização dos *clusters* e a compreensão das relações entre os atributos contribuem para uma melhor compreensão dos projetos, permitindo a tomada de decisões informadas e estratégicas. Essas informações podem ser utilizadas para otimizar o planejamento e execução dos projetos, identificar áreas de maior complexidade e colaboração interdepartamental, e melhorar a eficiência e o desempenho das instalações da empresa.

5 Conclusão

Este estudo utilizou a metodologia CRISP-DM para realizar a análise de um conjunto de dados relacionados aos projetos do departamento de PMO de uma empresa farmacêutica. As etapas do CRISP-DM foram aplicadas de forma sequencial, começando pela classificação da pesquisa, seguida pela coleta de dados, compreensão e preparação dos dados, modelagem e, finalmente, a avaliação dos resultados.

Nesse contexto, verificou-se que projetos complexos demandam mais tempo e recursos humanos, especialmente quando envolvem múltiplas áreas. Projetos relacionados à *Capacity* estão associados a reformas estruturais e aquisição de equipamentos com necessidade de peças de reposição.

A integração entre as áreas de BMS e FMS é essencial para garantir eficiência e desempenho das instalações. As três áreas de IT estão interligadas, com IT - OT - Project e IT - SM necessitando da presença de IT - Network/Infra. A demanda de recursos humanos em Metrologia está relacionada a projetos em conformidade com regulamentações. Projetos com alto Business Drive tendem a demandar menos recursos de controle de qualidade. O tempo de planejamento está relacionado à duração do projeto e ao número de horas de trabalho necessárias.

As conclusões anteriores destacam a importância de considerar a complexidade dos projetos, a integração de áreas chave, as demandas regulatórias e os processos de controle de qualidade ao planejar e gerenciar projetos. Tais informações podem contribuir para a tomada de decisões estratégicas e o aprimoramento da gestão de projetos em diferentes contextos organizacionais.

O trabalho tem algumas limitações a serem consideradas. Primeiramente, os resultados são específicos para a empresa e o departamento de PMO analisados, não sendo necessariamente generalizáveis para outras organizações. Além disso, a análise se baseou em dados históricos, e as conclusões podem ser influenciadas por mudanças no contexto ou nas práticas de gestão de projetos ao longo do tempo.

Para trabalhos futuros, é possível elaborar uma base similar a utilizada neste trabalho, mas incluindo questões relacionadas as demandas por recursos humanos em cenários reais e previstos, com o objetivo de identificar comportamentos e obter estimativas das demandas em projetos.

Com base no que foi apresentado, obtivemos informações valiosas sobre os projetos do departamento de PMO da empresa estudada. A identificação de *clusters* e a compreensão das relações entre os atributos contribuem para uma melhor compreensão dos projetos, permitindo a tomada de decisões assertivas. As conclusões e conhecimentos obtidos podem ser utilizados para aprimorar a eficiência, qualidade e desempenho dos projetos, contribuindo para o sucesso organizacional.

Referências

ABDI, H.; WILLIAMS, L. J. Principal component analysis. **Wiley Interdisciplinary Reviews: Computational Statistics**, John Wiley Sons, Inc., v. 2, p. 433–459, 2010. ISSN 1939-0068.

BANG, S.; AARVOLD, M. O.; HARTVIG, W. J.; OLSSON, N. O. E.; RAUZY, A. Application of machine learning to limited datasets: prediction of project success. **Expert Systems with Applications**, v. 190, p. 115286, 2022.

BATISTA, G.; KEOGH, E.; TATAW, O.; SOUZA, V. Alves de. Cid: An efficient complexity-invariant distance for time series. **Data Mining and Knowledge Discovery**, v. 28, 04 2013.

BERKHIN, P. Survey of clustering data mining techniques. **A Survey of Clustering Data Mining Techniques. Grouping Multidimensional Data: Recent Advances in Clustering.**, v. 10, 08 2002.

BORGES, L. E. **Python para desenvolvedores**. [S.l.]: Novatec Editora, 2010. ISBN 9788590945116.

BRO, R.; SMILDE, A. K. Principal component analysis. **Anal. Methods**, The Royal Society of Chemistry, v. 6, p. 2812–2831, 2014. Disponível em: <http://dx.doi.org/10.1039/C3AY41907J>.

BēRZIšA, S. Project management knowledge retrieval: Project classification. **ENVIRONMENT. TECHNOLOGIES. RESOURCES. Proceedings of the International Scientific and Practical Conference**, v. 2, n. 0, p. 33–39, 2015. ISSN 2256-070X. Disponível em: <http://journals.ru.lv/index.php/ETR/article/view/968>.

CASTRO, L. D.; FERRARI, D. **Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações**. [S.l.]: Saraiva Uni, 2016. ISBN 9788547200985.

CAUCHICK, P. A. **Metodologia de Pesquisa em Engenharia de Produção e Gestão de Operações**. 2. ed. [S.l.]: Elsevier Editora Ltda., 2012.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. **CRISP-DM 1.0: Step-by-Step Data Mining Guide**. [S.l.], 2000.

CHUANG, T. H.; LIN, C. C. Data mining to improve human resources management in taiwanese enterprises. **Total Quality Management Business Excellence**, v. 26, p. 1301–1312, 2015.

DANGETI, P. **Advanced Statistics for Machine Learning**. [S.l.]: Packt Publishing, 2017.

DONI, M. V. Análise de *Cluster*: Métodos hierárquicos e de particionamento. 01 2004.

DVIR, D.; LIPOVETSKY, S.; SHENHAR, A.; TISHLER, A. In search of project classification: a non-universal approach to project success factors. **Research Policy**, v. 27, n. 9, p. 915–935, 1998. ISSN 0048-7333.

EVERITT, B.; LANDAU, S.; LEESE, M.; STAHL, D. **Cluster Analysis**. [S.l.: s.n.], 2011. ISBN 9780470749913.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. **Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações**. 2. ed. [S.l.]: GEN LTC, 2015. ISBN 978-85-352-7822-4.

HAIR, J.; ANDERSON, R.; TATHAM, R.; BLACK, W. **Analise Multivariada de Dados**. [S.l.]: Bookman, 2005. ISBN 9788536304823.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. **Journal of Intelligent Information Systems**, v. 17, 10 2001.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3. ed. [S.l.]: Morgan Kaufmann, 2012.

HAO, J.; HO, T. K. Machine learning made easy: A review of scikit-learn package in python programming language. **Journal of Educational and Behavioral Statistics**, v. 44, n. 3, p. 348–361, 2019.

JOLLIFFE, I.; CADIMA, J. Principal component analysis: A review and recent developments. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 374, 2016.

KANDEL, S.; PAEPCKE, A.; HELLERSTEIN, J. M.; HEER, J. Wrangler: Interactive visual specification of data transformation scripts. **ACM Transactions on Interactive Intelligent Systems (TiiS)**, v. 2, n. 3, p. 1–30, 2012.

KERZNER, H. **Gestão de projetos: as melhores práticas**. 2. ed. [S.l.]: Bookman, 2006.

KIELING, E. J.; RODRIGUES, F. C.; FILIPPETTO, A.; BARBOSA, J. L. V. Human resource allocation in projects: a systematic mapping study. **International Journal of Business Information Systems**, v. 37, n. 4, p. 505–521, 2021.

LAKATOS, E. M.; MARCONI, M. A. **Metodologia Científica**. 8. ed. [S.l.]: Editora Atlas, 2022.

MCKINNEY, W. **Python for Data Analysis**. [S.l.]: O'Reilly Media, Inc., 2012. ISBN 9781449319793.

MORO, S.; LAUREANO, R.; CORTEZ, P. Using data mining for bank direct marketing: an application of the crisp-dm methodology. EUROSIS-ETI, 2011. Disponível em: <https://hdl.handle.net/1822/14838>.

OCHI, L.; DIAS, C.; SOARES, S. Clusterização em mineração de dados. 01 2004.

OPATHA, H. P. J. Hr analytics: A literature review and new conceptual model. **International Journal of Scientific and Research Publications (IJSRP)**, v. 10, 06 2020.

OTERO, L. D.; CENTENO, G.; RUIZ-TORRES, A. J.; OTERO, C. E. A systematic approach for resource allocation in software projects. **Computers Industrial Engineering**, v. 56, n. 4, p. 1333–1339, 2009. ISSN 0360-8352. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0360835208001708>.

PMI. **A Guide to the Project Management Body of Knowledge and the Standard for Project Management (PMBOK Guide)**. 7. ed. [S.l.]: Project Management Institute, 2021.

SCHOBER, P.; BOER, C.; SCHWARTE, L. Correlation coefficients: Appropriate use and interpretation. **Anesthesia & Analgesia**, v. 126, p. 1763–1768, 2018.

SHEARER, C. The crisp-dm model: the new blueprint for data mining. **Journal of data warehousing**, THE DATA WAREHOUSE INSTITUTE, v. 5, n. 4, p. 13–22, 2000.

SILVA, D.; BITTENCOURT, R.; CALUMBY, R. T. Clustering similarity measures for architecture recovery of evolving software. In: **Anais do VII Workshop de Visualização, Evolução e Manutenção de Software**. [S.l.]: SBC, 2019.

SILVA, E. L.; MENEZES, E. M. **Metodologia da Pesquisa e Elaboração de Dissertação**. Dissertação (Programa de Pós-Graduação em Engenharia de Produção Laboratório de Ensino a Distância) — Universidade Federal de Santa Catarina, Brasil, 2001.

TANG, D. Optimization of human resource management system based on data mining technology and random forest algorithm. **Wireless Communications and Mobile Computing**, v. 2022, 08 2022.

WIRTH, R. Crisp-dm: Towards a standard process model for data mining. In: **Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2000. p. 29–39.

ZAïANE, O.; FOSS, A.; LEE, C.-H.; WANG, W. On data clustering analysis: Scalability, constraints, and validation. In: . [S.l.: s.n.], 2002. p. 28–39. ISBN 978-3-540-43704-8.