Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Departamento de Computação e Sistemas

Caracterização de uma base de dados resultante da resolução de instâncias do problemas de escalonamento de projetos com restrições de recursos

Guilherme Augusto Rodrigues de Jesus

TRABALHO DE CONCLUSÃO DE CURSO

ORIENTAÇÃO:
Profa. Dra. Janniele Aparecida Soares Araújo

Outubro, 2024
João Monlevade–MG

Guilherme Augusto Rodrigues de Jesus

Caracterização de uma base de dados resultante da resolução de instâncias do problemas de escalonamento de projetos com restrições de recursos

Orientador: Profa. Dra. Janniele Aparecida Soares Araújo

Monografia apresentada ao curso de Sistemas de Informação do Instituto de Ciências Exatas e Aplicadas, da Universidade Federal de Ouro Preto, como requisito parcial para aprovação na Disciplina "Trabalho de Conclusão de Curso II".

Universidade Federal de Ouro Preto

João Monlevade

Outubro de 2024

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

J585c Jesus, Guilherme Augusto Rodrigues de.
Caracterização de uma base de dados resultante da resolução de instâncias do problemas de escalonamento de projetos com restrições de recursos. [manuscrito] / Guilherme Augusto Rodrigues de Jesus. - 2024.
47 f.: il.: color., gráf., tab..

Orientadora: Profa. Dra. Janniele Aparecida Soares Araujo.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Graduação em Sistemas de Informação .

1. Análise por agrupamento. 2. Banco de dados - Gerência. 3. Classificação. 4. Mineração de dados (Computação). I. Araujo, Janniele Aparecida Soares. II. Universidade Federal de Ouro Preto. III. Título.

CDU 004.6:025.4.036

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

Guilherme Augusto Rodrigues de Jesus

Caracterização de uma base de dados resultante da resolução de instâncias do problemas de escalonamento de projetos com restrições de recursos

Monografia apresentada ao Curso de Sistemas de Informação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de bacharel em Sistemas de Informação

Aprovada em 17 de Outubro de 2024

Membros da banca

Dra Janniele Aparecida Soares Araujo - Orientadora - Universidade Federal de Ouro Preto
Dra Helen de Cassia Sousa da Costa Lima - Universidade Federal de Ouro Preto
Dr Samuel Souza Brito - Universidade Federal de Ouro Preto

Janniele Aparecida Soares Araujo, orientadora do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 11/04/2025

Documento assinado eletronicamente por **Janniele Aparecida Soares Araujo**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 11/04/2025, às 13:13, 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 **0894818** e o código CRC **78BC1EB0**.

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

Dedico este trabalho de conclusão de curso aos meus queridos pais, cujo apoio incondicional foi minha maior fonte de inspiração. Agradeço pela paciência, pela compreensão e por sempre acreditarem em mim, mesmo nos momentos mais difíceis. Também dedico este trabalho aos meus professores e orientadores, que compartilharam seu conhecimento e sabedoria, guiando-me ao longo deste percurso acadêmico.

Agradecimentos

Primeiramente, eu agradeço a Deus por todas as bênçãos que tive em minha vida. Gostaríamos de expressar minha sincera gratidão a todos aqueles que contribuíram para o sucesso deste estudo.

Gostaria de agradecer à minha orientadora, Janniele, por sua orientação e apoio contínuo ao longo deste projeto. Suas valiosas sugestões e ideias foram fundamentais para o desenvolvimento desta pesquisa. Aos meus pais e irmãos, que me ajudaram e apoiaram durante minha vida.

"Science is more than a body of knowledge; it is a way of thinking."

— Carl Sagan (1934 – 1996),
in: The Demon-Haunted World: Science as a Candle in the Dark.

Resumo

O problema de escalonamento de múltiplos projetos com restrições de recursos e múltiplos modos de execução possui grande complexidade e é essencial em diversas áreas na atualidade. Esta monografia tem como objetivo principal realizar uma análise descritiva de uma base de dados que contém diversas instâncias resolvidas dos problemas *Multi-Mode Resource-Constrained Multi-Project Scheduling Problem* (MMRCMPSP), também conhecidos como problema de escalonamento de múltiplos projetos com restrição de recursos e múltiplos modos de execução. Foram utilizadas várias técnicas para extrair informações relevantes, como por exemplo, componente principal e *k-means*, o que facilitou a compreensão das características das instâncias e a identificação de padrões que podem contribuir para o desenvolvimento de soluções mais eficientes. Como resultado, novas descobertas sobre o comportamento e as particularidades dos dados foram obtidas, permitindo a categorização e a identificação de semelhanças entre eles. Esses achados possuem o potencial de facilitar a compreensão de problemas similares no futuro.

Palavras-chaves: Problema de Escalonamento. Mineração de Dados. Componente principal. Clusterização.

Abstract

The problem of scheduling multiple projects with resource constraints and multiple execution modes is highly complex and essential in various fields today. The main objective of this thesis is to conduct a descriptive analysis of a database containing various solved instances of *Multi-Mode Resource-Constrained Multi-Project Scheduling Problem* (MMRCMPSP) problems. Several techniques were used to extract relevant information, such as principal component and k-means, which facilitated the understanding of the characteristics of the instances and the identification of patterns that may contribute to the development of more efficient solutions. As a result, new discoveries regarding the behavior and particularities of the data were made, allowing for the categorization and identification of similarities among them. These findings have the potential to enhance the understanding of similar problems in the future.

Key-words: Scheduling Problem. Data Mining. Component Analysis. Clustering.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

MMRCMPSP *Multi-Mode Resource-Constrained Multi-Project Scheduling Problem*

MMRCPSP *Multi-Mode Resource-Constrained Project Scheduling Problem*

SMRCPSP *Single-Mode Resource-Constrained Project Scheduling Problem*

RCPSP *Resource Constrained Project Scheduling Problem*

PCA *Principal Component Analysis*

KDD *Knowledge Discovery in Databases*

AON *Activity-on-Node*

SHAP *SHapley Additive exPlanations*

Sumário

1 Introdução

A otimização combinatória é uma área essencial e amplamente aplicada em setores como engenharia e ciências, como ressaltado por Santos, Soares e Toffolo (2014). Nesses campos, há diversas aplicações práticas, incluindo a melhoria na gestão de projetos com um grande número de atividades. Entre os diferentes ramos da otimização combinatória, destacam-se os problemas de escalonamento de projetos, cadeias de suprimentos e outros problemas de otimização linear. Esta monografia focará no *Multi-Mode Resource-Constrained Multi-Project Scheduling Problem* (MMRCMPSP).

Segundo Ahmeti e Musliu (2021), o MMRCMPSP reflete um problema real, onde o objetivo é organizar de forma eficiente as atividades de diversos projetos, respeitando recursos disponíveis, as ordens de precedência e prazos de execução. Tseng (2008) observa que, embora exista uma vasta literatura sobre o *Resource Constrained Project Scheduling Problem* (RCPSP), os estudos focados em MMRCMPSP são significativamente menos frequentes, sendo este considerado uma generalização do RCPSP e um desafio tanto para acadêmicos quanto para matemáticos.

Uma solução para o MMRCMPSP envolve o agendamento de todas as atividades dos projetos, respeitando as restrições de precedência e alocação de recursos, conforme descrito por Santos, Soares e Toffolo (2014). O objetivo deste estudo é realizar uma análise descritiva dos dados resultantes da resolução de instâncias do MMRCMPSP, que utiliza um algoritmo metaheurístico que combina diversas estratégias de vizinhança para lidar com esta classe de problema.

1.1 O problema de pesquisa

Os problemas de escalonamento de projetos com restrição de recursos ou RCPSP são classificados como NP-Difíceis, conforme Blazewicz, Lenstra e Kan (1983). De acordo com Araujo (2019), a solução do RCPSP consiste em organizar o cronograma de atividades, levando em consideração recursos, tempo disponível e relações de precedência.

Diversas variações do RCPSP são abordadas na literatura, como o *Single-Mode Resource-Constrained Project Scheduling Problem* (SMRCPSP),*Multi-Mode Resource-Constrained Project Scheduling Problem* (MMRCPSP), e *Multi-Mode Resource-Constrained Multi-Project Scheduling Problem* (MMRCMPSP). Um grande desafio é identificar as classes de problemas e determinar as melhores configurações de algoritmos e parâmetros para cada classe.

Os experimentos feitos por Araujo et al. (2016) envolvem dados que caracterizam

cada instância, além de resultados experimentais com diferentes configurações de parâmetros. Mesmo algoritmos com desempenho médio podem ser úteis para subconjuntos específicos de problemas. O objetivo desta monografia é aplicar algoritmos de classificação e agrupamento para identificar classes de problemas e determinar suas melhores parametrizações.

A literatura possui diversos métodos heurísticos e exatos para a resolução desses problemas. No entanto, a escolha das melhores configurações de vizinhança para cada problema ainda é pouco explorada. Este trabalho foca em uma análise descritiva dos dados de instâncias resolvidas, utilizando técnicas de mineração de dados para identificar tendências e descobrir novos padrões.

1.2 Objetivos

O objetivo geral deste projeto é realizar uma análise descritiva dos dados de instâncias resolvidas do MMRCMPSP disponíveis na literatura. O objetivo específico é explorar técnicas de mineração de dados para extrair *insights* dessa base.

1.3 Organização do trabalho

Este trabalho está organizado da seguinte forma: a introdução contextualiza o tema e apresenta o problema de pesquisa. A Seção 2 traz uma revisão crítica da literatura relacionada ao MMRCMPSP, assim como elementos de contextualização que foram julgados importantes para entendimento do trabalho. A Seção 3, que constitui o desenvolvimento principal, detalha a metodologia utilizada e as análises efetuadas ao longo do estudo. Por fim, a Seção 4 sintetiza os principais resultados e descobertas.

2 Revisão bibliográfica

Neste capítulo, apresenta-se o referencial teórico e os trabalhos relacionados que fundamentam o desenvolvimento desta monografia.

2.1 Análise e Mineração de Dados

A análise e a mineração de dados são abordagens complementares que desempenham papéis essenciais na exploração e extração de informações a partir de grandes volumes de dados. Segundo PINTO (2020), existem diferentes tipos de análise de dados que podem ser utilizados, cada um com um papel específico:

- Análise Descritiva: Fornece uma visão geral dos dados, organizando, resumindo e apresentando-os de forma a evidenciar padrões e características importantes. Uma técnica frequentemente utilizada nesta etapa é a clusterização, como o *K-Means*, que agrupa os dados em subconjuntos homogêneos (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).
- Análise Preditiva: Envolve a construção de modelos, como os de aprendizado de máquina ou estatísticos, que utilizam dados históricos para prever eventos futuros (LEE; CHEANG; MOSLEHPOUR, 2022). Técnicas como *data mining* e análise de *big data* são amplamente aplicadas nesse contexto.
- Análise Prescritiva: Concentra-se em fornecer recomendações e diretrizes com base nas previsões geradas, ajudando na tomada de decisões e otimização de resultados (LEPENIOTI et al., 2020).
- Análise Diagnóstica: Investiga eventos passados para identificar as causas subjacentes e entender por que determinados resultados ocorreram. Essa análise é valiosa para melhorar processos e solucionar problemas.

A mineração de dados, por sua vez, envolve o uso de técnicas de aprendizado supervisionado e não supervisionado para explorar e identificar padrões ocultos nos dados. De acordo com Zhang, Zhang e Yang (2003), técnica como a de clusterização é amplamente utilizada para revelar esses padrões. Além disso, Jiawei (1996) destaca a crescente popularidade da mineração de dados devido às suas inúmeras aplicações.

No contexto do aprendizado supervisionado, os algoritmos são treinados com conjuntos de dados rotulados, onde as entradas são mapeadas para saídas conhecidas, como

descrito por Nasteski (2017). Em contraste, no aprendizado não supervisionado, os algoritmos descobrem padrões e estruturas nos dados sem a necessidade de rótulos predefinidos. Segundo Mahesh (2020), esse método é chamado de "não supervisionado"devido à ausência de um "professor"para guiar o treinamento.

Ambas as abordagens têm aplicações em várias áreas, e a escolha entre aprendizado supervisionado e não supervisionado depende da natureza dos dados e dos objetivos do problema a ser resolvido. A combinação dessas técnicas, tanto na análise quanto na mineração de dados, permite uma exploração mais profunda e abrangente, levando a *insights* valiosos para a tomada de decisões e a otimização de processos.

2.2 Técnicas de correlação entre variáveis, redução de dimensionlidade e agrupamento

A análise de dados de grandes dimensões exige o uso de técnicas robustas para redução de dimensionalidade, agrupamento eficiente e validação de resultados. Dentre essas técnicas, a Análise de Componentes Principais (*Principal Component Analysis* (PCA)) e o algoritmo de agrupamento *K-Means* se destacam, além de métodos estatísticos, como a correlação de *Pearson* e a análise da variância explicada, que contribuem para uma interpretação eficaz dos dados.

A PCA é uma técnica estatística essencial para redução de dimensionalidade em conjuntos de dados com muitas variáveis. Conforme discutido por Ding e He (2004), o PCA transforma variáveis originais em novas coordenadas chamadas componentes principais, de modo que elas capturam a maior parte da variabilidade nos dados. O principal objetivo do PCA é reduzir a quantidade de variáveis enquanto mantém a maior parte da variação original, facilitando a interpretação e o processamento dos dados.

Ao calcular as componentes principais, a análise da variância explicada desempenha um papel fundamental. A variância explicada indica quanta variabilidade total dos dados é capturada por cada componente principal. Assim, quanto maior a variância explicada pela primeira componente, maior sua importância no conjunto de dados. Isso ajuda a determinar quantas componentes principais são necessárias para representar os dados de forma eficaz. Segundo Jolliffe (2002), a escolha adequada de componentes pode ser feita selecionando aquelas que, em conjunto, explicam uma proporção significativa da variabilidade total, geralmente em torno de 70-90%.

Além disso, a PCA pode identificar correlações entre variáveis. A correlação de Pearson é um método amplamente utilizado para medir a relação linear entre duas variáveis. Ela varia entre -1 e 1, onde valores próximos a 1 indicam correlação positiva forte, e valores próximos a -1 indicam correlação negativa forte (BENESTY et al., 2009). Isso permite

a eliminação de variáveis redundantes antes de aplicar técnicas de agrupamento, como o *K-Means*, otimizando a performance do algoritmo.

O *K-Means* é um dos algoritmos mais utilizados para o agrupamento de dados, e sua popularidade se deve à sua simplicidade e eficiência. Ele organiza os dados em k *clusters*, com base em critérios de minimização da soma das distâncias quadráticas entre os pontos e seus respectivos centroides (SHUKLA; NAGANNA, 2014). O valor k é definido pelo usuário e indica o número de *clusters* que o algoritmo deve formar. O processo iterativo de recalcular os centroides com base nas atribuições dos pontos continua até que uma condição de convergência seja satisfeita, como mostrado no Algoritmo 1.

Algorithm 1 Algoritmo de Agrupamento *K-Means*

Entrada: Base de Dados D, Número de clusters k
Saída: k centróides C
Inicializar k centróides com valores aleatórios pertencentes a D
while Condição de Convergência Não Satisfeita **do**
    for cada ponto $d \in D$ **do**
        Atribuir d ao cluster com o centróide mais próximo
    end for
    for cada cluster c **do**
        Recalcular o centróide c como a média dos pontos atribuídos
    end for
end while

Uma das questões mais críticas no uso do *K-Means* é a escolha do valor de k, ou seja, o número de *clusters*. O Método do Cotovelo (*Elbow Method*) é uma técnica simples e eficaz para determinar o valor ideal de k. Ao plotar a soma das distâncias quadráticas internas (inércia) em relação ao número de *clusters*, o ponto em que a redução da inércia começa a ser menos significativa (formando um "cotovelo"no gráfico) indica o número ótimo de *clusters* (CUI et al., 2020).

Para verificar a qualidade dos agrupamentos gerados pelo *K-Means*, é fundamental aplicar métricas de avaliação que indiquem quão bem os *clusters* refletem a estrutura dos dados. Uma das principais métricas é o *Silhouette Score*, que avalia o quão próximo um ponto está de seu próprio *cluster* em comparação com outros *clusters*. O valor do *Silhouette Score* varia entre -1 e 1, onde valores próximos a 1 indicam que os pontos estão corretamente agrupados, e valores próximos a -1 indicam que os pontos podem estar no *cluster* errado (ROUSSEEUW, 1987).

Por fim, o coeficiente de correlação de *Pearson* pode ser aplicado entre variáveis dentro de cada *cluster* para validar se os agrupamentos respeitam as correlações lineares inerentes aos dados. Isso permite verificar se as variáveis dentro de um *cluster* compartilham uma relação significativa, validando a coesão dos grupos formados pelo algoritmo *K-Means*.

Dessa forma, a combinação de técnicas como PCA, *K-Means*, e métodos estatísticos

de avaliação oferece uma abordagem robusta para a análise de grandes volumes de dados. A PCA facilita a compreensão e a simplificação de dados complexos, enquanto o *K-Means* permite uma eficiente identificação de grupos. A avaliação cuidadosa por meio de métricas como o *Silhouette Score* e a correlação de *Pearson* garante que os agrupamentos resultantes são de alta qualidade e refletem a estrutura intrínseca dos dados.

2.3 Linguagens e Tecnologias Utilizadas

A escolha das linguagens e tecnologias em um projeto é um fator determinante para alcançar seus objetivos e atender aos requisitos estabelecidos. Uma seleção criteriosa dessas ferramentas é essencial para garantir o sucesso do desenvolvimento, promovendo eficiência e qualidade no produto final. Nesta seção, são apresentadas as principais linguagens e tecnologias empregadas ao longo do desenvolvimento deste trabalho.

O Google Colab foi a plataforma escolhida para o desenvolvimento de todo o código deste trabalho de conclusão de curso. Esta plataforma *online*, oferecida gratuitamente pelo Google, foi projetada para facilitar o desenvolvimento colaborativo em Python, com foco em ambientes de aprendizado de máquina e análise de dados. Uma das principais vantagens do Google Colab é sua integração com o Google Drive, que permite o armazenamento e a manipulação de dados em nuvem (KANANI; PADOLE, 2019).

Python foi a linguagem de programação escolhida para o desenvolvimento do projeto. De acordo com Menezes (2010), Python é uma linguagem extremamente versátil, amplamente utilizada em diversas áreas devido à sua simplicidade, legibilidade e vasta coleção de bibliotecas e *frameworks*. Sua sintaxe clara, com comandos baseados no inglês, facilita o aprendizado e a aplicação, conforme destacado por Sahoo et al. (2019). A disponibilidade de bibliotecas *open-source* torna Python uma escolha ideal para projetos de ciência de dados e aprendizado de máquina.

No desenvolvimento deste projeto, diversas bibliotecas de Python foram empregadas para otimizar o processamento e a visualização dos dados. A biblioteca *pandas*, conforme mencionado por Sahoo et al. (2019), oferece poderosas ferramentas para manipulação, limpeza e análise de grandes conjuntos de dados, além de funcionalidades avançadas para armazenamento e visualização de informações.

Para melhorar a visualização dos dados, foi utilizada a biblioteca *matplotlib*. Segundo Sial, Rashdi e Khan (2021), *matplotlib* é amplamente compatível com outras bibliotecas de Python e proporciona uma ampla gama de opções para criação de gráficos, desde visualizações simples até mais complexas, o que facilita a extração de *insights* a partir dos dados.

Outra biblioteca fundamental no desenvolvimento deste trabalho foi a *scikit-learn*,

principalmente pela disponibilidade do algoritmo de clusterização *k-means*. De acordo com Hao e Ho (2019), *scikit-learn* é uma poderosa biblioteca de aprendizado de máquina, conhecida por sua simplicidade de uso e pelo fato de ser *open-source*. Ela oferece uma vasta gama de ferramentas para modelagem de dados, como algoritmos de classificação, regressão e clusterização, além de métodos de avaliação de performance de modelos.

Com essa seleção de ferramentas e bibliotecas, foi possível realizar as diversas análises, visualizações e comparações ao longo da monografia, garantindo uma abordagem eficiente para o processamento dos dados e a implementação dos modelos de aprendizado de máquina.

2.4 Trabalhos Relacionados

A seção de trabalhos relacionados nesta monografia são referências acadêmicas, ou seja, estudos anteriores relacionados a área desse trabalho de conclusão de curso.

Os autores Araujo et al. (2016) que realizaram a pesquisa que gerou o artigo científico *Neighborhood Composition Strategies in Stochastic Local Search*. Esse estudo investiga estratégias de composição de vizinhança em algoritmos de *Stochastic Local Search*, focando na otimização da seleção de vizinhanças para melhorar a busca em problemas do tipo MMRCMPSP. Foi graças a este estudo que foi possível acesso a base de dados utilizada na Seção 3.

O artigo *Software Project Schedule Management Using Machine Learning e Data Mining* feito por Wei e Rana (2019) detalha como o uso de *Machine Learning* e *Data Mining* pode transformar o gerenciamento de projetos, aumentando a precisão e eficiência. É citado que, atualmente, muitas empresas ainda utilizam métodos mais antigos que não conseguem lidar com a complexidade dos projetos modernos. Diversos fatores como mudanças de escopo, restrições de recursos e a subestimação de desafios técnicos frequentemente causam atrasos. O estudo conclui que, no futuro, o gerenciamento de cronogramas poderá ser amplamente automatizado, reduzindo custos e melhorando as taxas de sucesso dos projetos.

Já em relação ao artigo *A data-driven meta-learning recommendation model for multi-moderesource constrained project scheduling problem* desenvolvido por Chu et al. (2023), apresenta como os problemas de agendamento com restrições de múltiplos modos são um desafio complexo, onde a eficácia das meta-heurísticas varia significativamente conforme as características de cada instância utilizada. A abordagem tradicional de selecionar algoritmos por tentativa e erro é ineficiente e pode levar a soluções não ideais. Para resolver essa questão, os autores propõem o Modelo de Recomendação de Meta-Heurísticas, que utiliza técnicas de meta-aprendizado para extrair informações dos problemas e assim, recomenda a melhor meta-heurística utilizando classificadores baseados em Máquinas de Vetores de Suporte Multiclasse.

2.5 O problema MMRCMPSP

O *Multi-Mode Resource-Constrained Multi-Project Scheduling Problem* (MMRCMPSP) é uma versão mais completa e complexa do problema de agendamento de projetos. Segundo Hartmann e Briskorn (2022), esse conjunto de problemas do qual ele faz parte é essencial em gestão de projetos e normalmente, as atividades possuem dependências, sendo elas a competição por recursos limitados e as restrições de precedência entre cada par de atividades.

As principais características do MMRCMPSP podem ser simplificadas em:

- Múltiplos projetos: Projetos consistem em grupos de atividades com dependências de recursos e precedências.
- Múltiplos Modos: Nesse caso cada atividade apresenta diversas maneiras de serem executadas, cada uma dessas maneiras podem ter diferentes custos e duração.
- Recursos: Também podem ser de diferentes tipos e devem ser respeitados em relação à quantidade disponível, seja ela, em período ou total para todo o conjunto do problema.
- Objetivo: Exemplo de objetivos inclui a diminuição de tempo de execução do projeto, diminuição de custos ou otimização de recursos.

De acordo com Toffolo et al. (2016) para apresentar uma formulação completa para problemas do tipo MMRCMPSP, os seguintes dados devem ser considerados:

- Variáveis descritivas: Como número de projetos, número de atividades, número de atividades em cada projeto e número de modos de cada uma das atividades.
- Variáveis temporais: Como duração de cada atividade para cada configuração e tempo em cada atividade pode ser realizada.
- Variáveis restritivas: Como recursos renováveis disponíveis, recursos não renováveis disponíveis, quantidade de recursos disponíveis para o projeto, quantidade de cada recursos necessário para realizar cada configuração de atividade e lista de precedentes de cada atividade.

2.5.1 Representação e exemplo

Como pôde ser visto os problemas dessa categoria são complexos pois envolvem muitas informações, muitas configurações e diversas restrições. Para lidar com essa complexidade, diversas formas de representação podem ser adotadas. Dentre elas diagramas de rede, como o *Activity-on-Node* (AON) ajudam a visualizar as precedências e a sequência

de todas as atividades do projeto. Além disso, tabelas detalhadas organizam informações essenciais, como tempos, modos de execução e alocações de recursos. E em relação a representação de possíveis soluções para instâncias desse problema o gráfico de *Gantt* é muito utilizado principalmente pela fácil apresentação visual de grandes grupos de atividades e seu escalonamento em relação ao tempo.

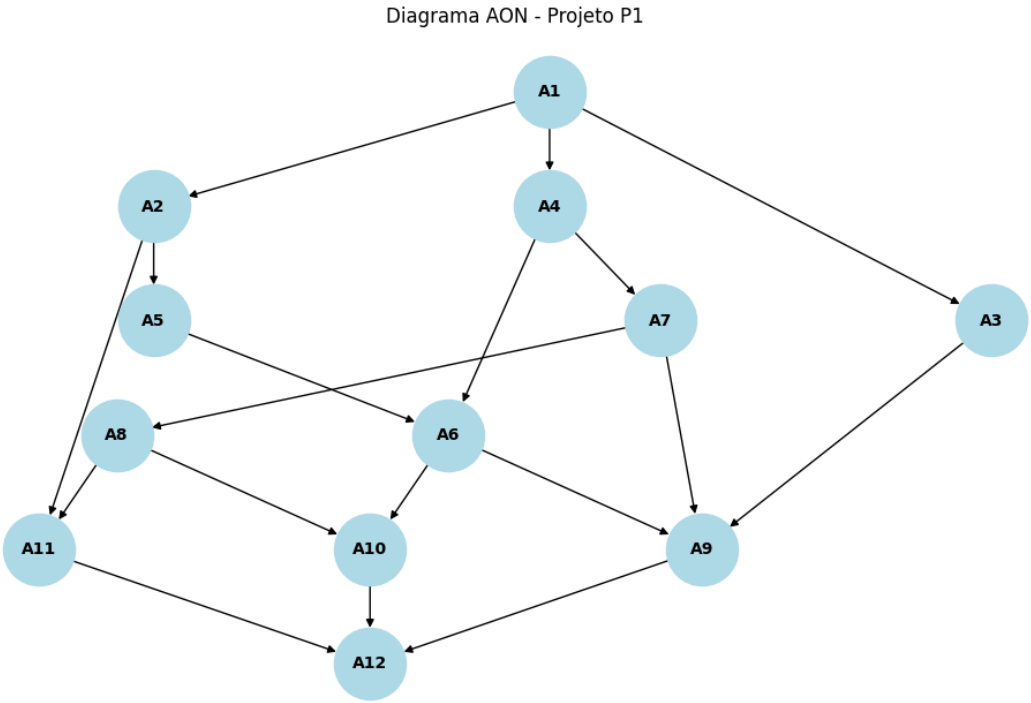
A Tabela 1 apresenta o exemplo de instância A-1 do problema MMRCMPSP, disponível no site oficial do desafio MISTA 2013. Esse exemplo ilustra dois projetos com atividades que podem ser executadas em múltiplos modos, cada um com diferentes durações e exigências de recursos. Ela inclui informações organizadas em colunas, como projeto, atividade, precedências, modo, duração e recursos. Apesar de ser o exemplo mais simples do desafio, ele já permite perceber a complexidade envolvida nesse tipo de problema, evidenciando as dificuldades associadas ao seu manuseio e à obtenção de soluções viáveis.

Tabela 1 – Instância A-1 MMRCMPSP Mista 2013

Projeto	Atividade	Precedências	Modo	Duração	Recursos
Projeto 1	A-1	–	1	0	R1: 0, R2: 0, N1: 0, N2: 0
Projeto 1	A-2	A-1	1	1	R1: 8, R2: 0, N1: 0, N2: 7
Projeto 1	A-3	A-1	1	4	R1: 9, R2: 0, N1: 9, N2: 0
Projeto 1	A-4	A-1	1	1	R1: 0, R2: 5, N1: 7, N2: 0
Projeto 1	A-5	A-2	1	3	R1: 0, R2: 8, N1: 9, N2: 0
Projeto 1	A-6	A-4, A-5	1	2	R1: 8, R2: 0, N1: 4, N2: 0
Projeto 1	A-7	A-4	1	5	R1: 7, R2: 0, N1: 0, N2: 6
Projeto 1	A-8	A-7	1	1	R1: 3, R2: 0, N1: 0, N2: 6
Projeto 1	A-9	A-3, A-6, A-7	1	3	R1: 6, R2: 0, N1: 0, N2: 6
Projeto 1	A-10	A-6, A-8	1	2	R1: 0, R2: 4, N1: 8, N2: 0
Projeto 1	A-11	A-2, A-8	1	3	R1: 1, R2: 0, N1: 0, N2: 6
Projeto 1	A-12	A-9, A-10, A-11	1	0	R1: 0, R2: 0, N1: 0, N2: 0
Projeto 2	A-1	–	1	0	R1: 0, R2: 0, N1: 0, N2: 0
Projeto 2	A-2	A-1	1	1	R1: 6, R2: 9, N1: 7, N2: 0
Projeto 2	A-3	A-1	1	1	R1: 6, R2: 4, N1: 0, N2: 7
Projeto 2	A-4	A-1	1	1	R1: 6, R2: 2, N1: 0, N2: 7
Projeto 2	A-5	A-4	1	7	R1: 4, R2: 9, N1: 6, N2: 0
Projeto 2	A-6	A-5	1	1	R1: 2, R2: 6, N1: 0, N2: 5
Projeto 2	A-7	A-5	1	5	R1: 2, R2: 5, N1: 8, N2: 0
Projeto 2	A-8	A-5	1	5	R1: 8, R2: 9, N1: 0, N2: 10
Projeto 2	A-9	A-7, A-8	1	1	R1: 5, R2: 8, N1: 0, N2: 6
Projeto 2	A-10	A-2, A-3, A-8	1	4	R1: 2, R2: 7, N1: 8, N2: 0
Projeto 2	A-11	A-6, A-7, A-8	1	6	R1: 7, R2: 9, N1: 0, N2: 7
Projeto 2	A-12	A-9, A-10, A-11	1	0	R1: 0, R2: 0, N1: 0, N2: 0

Com as restrições de precedências da instância A-1 é possível criar os diagramas de rede que são mais fáceis de serem analisados. A Figura 1 representa um desses diagramas com todas as dependências precedentes da tabela referentes ao projeto P1 e a Figura 2 representa a organização do projeto P2.

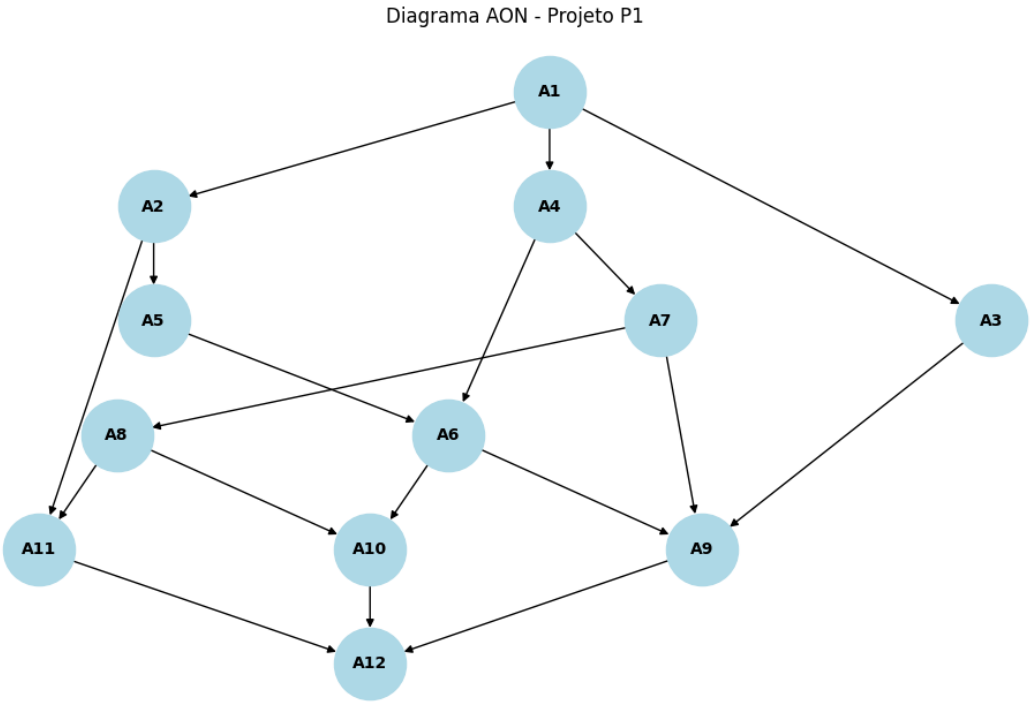
Figura 1 – Gráfico de rede Projeto P1



Fonte: Os Autores

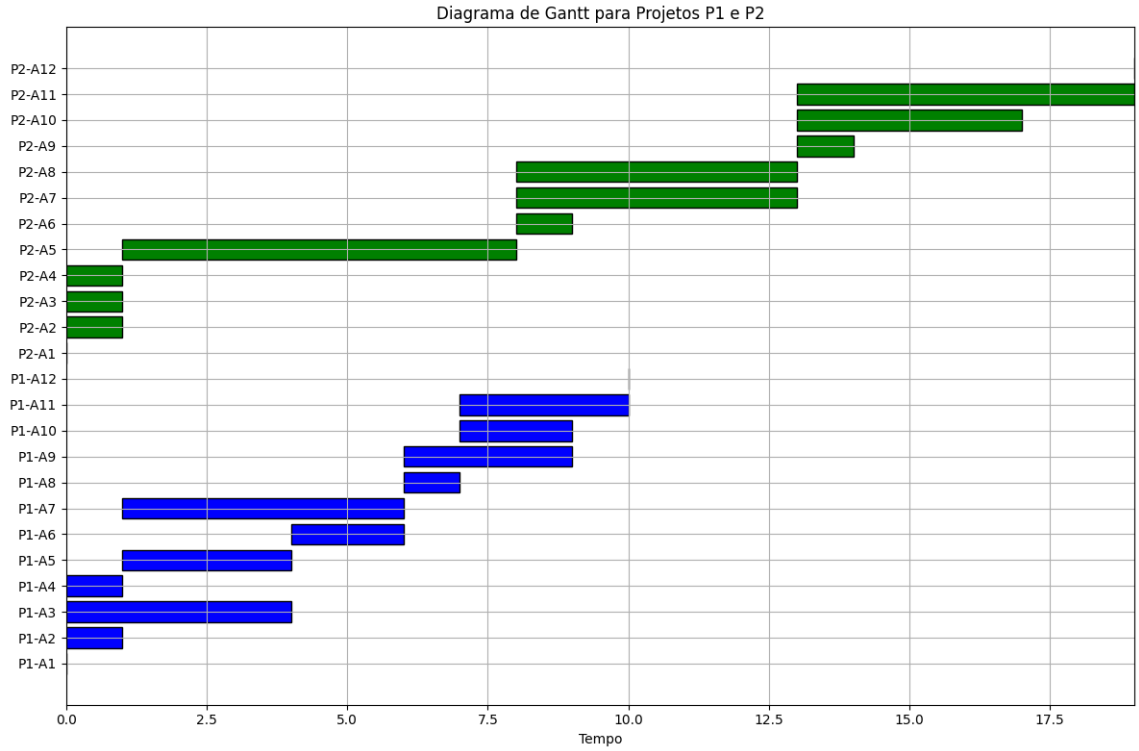
O diagrama de *Gantt* apresentado na Figura 3 ilustra uma das possíveis soluções para o exemplo do problema analisado nessa seção, oferecendo uma representação visual clara do escalonamento das atividades dos dois projetos. Nele, cada atividade é exibida como uma barra horizontal, com o comprimento indicando sua duração e a posição no eixo temporal mostrando os tempos de início e término. A solução respeita às relações de precedência definidas no problema, garantindo que nenhuma atividade comece antes da conclusão de suas predecessoras, mas utiliza o modo de execução 1 para cada atividade, simplificando a visualização. No entanto, embora o diagrama represente uma sequência válida, ele não leva em consideração as restrições de recursos renováveis e não renováveis, nem o uso de diferentes modo de execução para tentar achar um melhor resultado.

Figura 2 – Gráfico de rede Projeto P2



Fonte: Os Autores

Figura 3 – Diagrama *Gantt* A-1



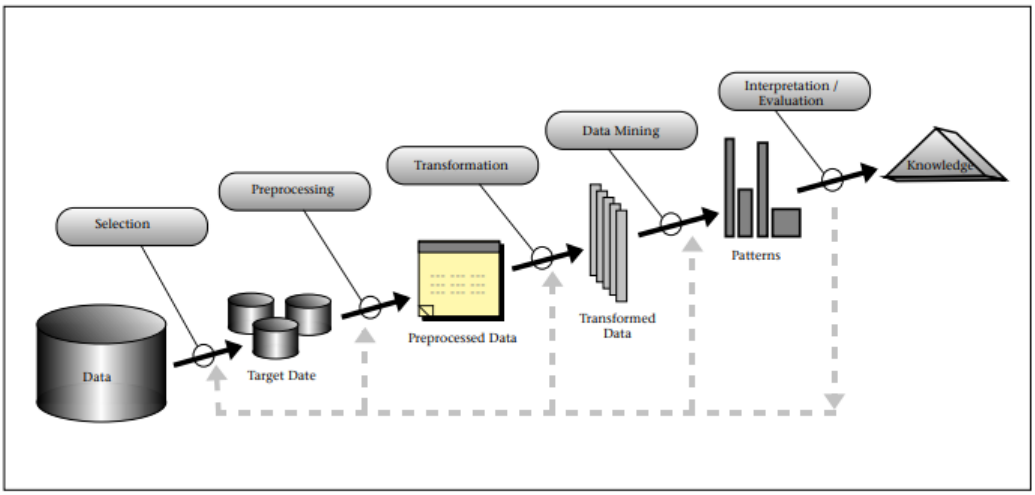
Fonte: Os Autores

3 Metodologia

Neste capítulo, são detalhadas as etapas da metodologia *Knowledge Discovery in Databases* (KDD), um processo estruturado para a descoberta de conhecimento em bases de dados, conforme proposto por Fayyad, Piatetsky-Shapiro e Smyth (1996). Inicialmente, foi realizada uma pesquisa para entender os problemas da área e identificar a metodologia mais adequada para este trabalho. Foram consideradas diversas metodologias, cada uma com suas particularidades e aplicações. No entanto, a metodologia KDD foi escolhida devido a sua característica estruturada para a descoberta de conhecimento. Seu foco na extração de padrões relevantes a partir de grandes volumes de dados permite uma análise aprofundada e eficiente, garantindo *insights* valiosos para a tomada de decisão.

O KDD oferece um processo sistemático e abrangente, que vai desde a seleção dos dados até a interpretação dos resultados, facilitando a descoberta de padrões, tendências e informações relevantes. A aplicação desta metodologia permitiu uma análise robusta e bem estruturada ao longo do desenvolvimento da monografia. Conforme ilustrado na Figura 4, a metodologia é composta por etapas interconectadas, como seleção, pré-processamento, transformação, mineração de dados e interpretação dos resultados.

Figura 4 – Etapas do KDD



Fonte: Fayyad, Piatetsky-Shapiro e Smyth (1996)

As principais etapas da metodologia KDD incluem:

- **Seleção**: Identificação e escolha dos dados relevantes para a análise, baseando-se em critérios específicos para garantir que apenas informações significativas sejam

consideradas.

- **Pré-processamento**: Preparação dos dados, envolvendo a remoção de valores ausentes, tratamento de dados discrepantes e normalização, assegurando que os dados estejam prontos para a mineração.
- **Transformação**: Transformação ou redução de dimensionalidade dos dados, criando novas características quando necessário, para simplificar a análise sem perda de informações essenciais.
- **Mineração de Dados**: Aplicação de algoritmos para identificar padrões, associações ou outras informações relevantes. Neste trabalho, foi utilizado o método *K-Means* para clusterização dos dados.
- **Interpretação e análise dos resultados**: Etapa final, em que os resultados obtidos são interpretados e validados, transformando os *insights* obtidos em conhecimento prático e aplicável.

A análise descritiva da base de dados foi conduzida utilizando técnicas de mineração de dados, com ênfase na caracterização e categorização dos dados. Diversas bibliotecas, conforme descrito na Seção 2.3, foram empregadas, proporcionando métodos eficientes para a manipulação, organização e visualização dos dados. Essas ferramentas facilitaram a execução das etapas de leitura, processamento e geração de gráficos, contribuindo significativamente para a compreensão das variáveis envolvidas. A aplicação da metodologia KDD permitiu uma análise eficiente, possibilitando a caracterização detalhada dos dados e a extração de padrões relevantes, fundamentais para o desenvolvimento e conclusão deste trabalho. A descrição de cada uma dessas etapas é abordada nas seções subsequentes: a seleção e o pré-processamento são discutidos na Seção 3.1, a transformação na Seção 3.2, a mineração de dados na Seção 3.3 e por fim a análise dos resultados na Seção 3.4.

3.1 Seleção e Pré-processamento de Dados

A seleção de uma base de dados de qualidade é essencial em qualquer projeto de análise de dados. A base de dados utilizada neste trabalho consiste em 1.170 exemplos de resolução do problema MMRCMPSP, contendo 52 atributos que descrevem cada instância, como número de projetos, atividades, configurações relacionadas às intensidades de aplicação das 14 vizinhanças e os custos de resolução associados. Esses atributos permitem uma análise aprofundada do problema e a identificação de padrões e similaridades entre as instâncias. A Tabela 2 apresenta um resumo desses atributos.

Inicialmente, foi realizada uma análise geral da base de dados para verificar a existência de valores ausentes ou nulos. Nenhum desses problemas foi identificado.

Tabela 2 – Base de dados

ID	Atributos	Tipo	Exemplo	Descrição
0	nProjects	int	2	Número de projetos da instância
1	minJobsProj	int	12	Número mínimo de atividades por projeto
2	maxJobsProj	int	12	Número máximo de atividades por projeto
3	starts	int	6	Possíveis inícios
4	avgStartProj	int	3	Média dos inícios
5	ends	int	6	Possíveis finais
6	avgEndProj	int	3	Média dos finais
7	nJobs	int	24	Número de atividades
8	minDuration	int	0	Duração mínima
9	maxDuration	int	10	Duração máxima
10	avgDuration	float	5.19	Média de duração
11	minModes	int	1	Mínimo número de modos
12	maxModes	int	3	Máximo número de modos
13	avgModes	float	2.67	Média do número de modos
14	minNumRRModes	int	0	Mínimo de recursos renováveis por modos
15	maxNumRRModes	int	2	Máximo de recursos renováveis por modos
16	avgNumRRModes	float	1.41	Média de recursos renováveis por modos
17	minNumNRModes	int	0	Mínimo de recursos não renováveis por modos
18	maxNumNRModes	int	1	Máximo de recursos não renováveis por modos
19	avgNumNRModes	float	0.94	Média de recursos não renováveis por modos
20	minPrec	int	0	Mínimo número de precedentes
21	maxPrec	int	3	Máximo número de precedentes
22	avgPrec	float	1.50	Média de precedentes
23	nRR	int	3	Número de recursos renováveis
24	nNR	int	4	Número de recursos não renováveis
25	minConsumptionRR	int	1	Mínimo consumo de recursos renováveis
26	maxConsumptionRR	int	9	Máximo consumo de recursos renováveis
27	avgConsumptionRR	float	5.39	Média do consumo de recursos renováveis
28	minConsumptionNR	int	2	Mínimo consumo de recursos não renováveis
29	maxConsumptionNR	int	10	Máximo consumo de recursos não renováveis
30	avgConsumptionNR	float	5.88	Média do consumo de recursos não renováveis
31	maxCapRR	int	24	Capacidade máxima de recursos renováveis
32	minCapRR	int	13	Capacidade mínima de recursos renováveis
33	avgCapRR	float	17.67	Média da capacidade de recursos renováveis
34	maxCapNR	int	66	Capacidade máxima de recursos não renováveis
35	minCapNR	int	39	Capacidade mínima de recursos não renováveis
36	avgCapNR	float	51.25	Média da capacidade de recursos não renováveis
37	Intensity1	float	0.2	Intensidade da Vizinhança *invert sequence of jobs*
38	Intensity2	float	0.3	Intensidade da Vizinhança *offset job*
39	Intensity3	float	0.4	Intensidade da Vizinhança *swap two jobs*
40	Intensity4	float	0.5	Intensidade da Vizinhança offset project
41	Intensity5	float	0.6	Intensidade da *swap and compact two projects*
42	Intensity6	float	0.7	Intensidade da *compact project on percentage*
43	Intensity7	float	0.8	Intensidade da Vizinhança *change one mode*
44	Intensity8	float	0.9	Intensidade da Vizinhança *change two modes*
45	Intensity9	float	1.0	Intensidade da Vizinhança *change three modes*
46	Intensity10	float	1.1	Intensidade da Vizinhança *change four modes*
47	Intensity11	float	1.2	Intensidade *successive swap of a job in a window*
48	Intensity12	float	1.3	Intensidade *successive insertions of a job in a window*
49	Intensity13	float	1.4	Intensidade da Vizinhança *squeeze project on extreme*
50	Intensity14	float	1.5	Intensidade da Vizinhança *compact subsequent projects*
51	cost	float	100	Custo da solução

Foram realizadas alterações na estrutura da base de dados. A variável que representava um conjunto de intensidades foi desmembrada em 14 novas variáveis do tipo *float*, antes era um objeto que continha a informação de todas as intensidades concatenadas. Depois dessa transformação a estrutura ficou como representado na Tabela 2. Após a criação dessas novas colunas, a coluna original foi removida. Em seguida, foi feita uma análise mais detalhada utilizando a biblioteca *pandas*, para leitura e apresentação dos resumos descritivos da base de dados. Essa análise permitiu uma compreensão mais profunda das características das instâncias, como quantidade, tipos de variáveis, nomes e número de colunas, entre outros aspectos relevantes. Foram eliminadas as variáveis com valores constantes, ou seja, aquelas que apresentavam o mesmo valor em todas as instâncias. As variáveis removidas foram:

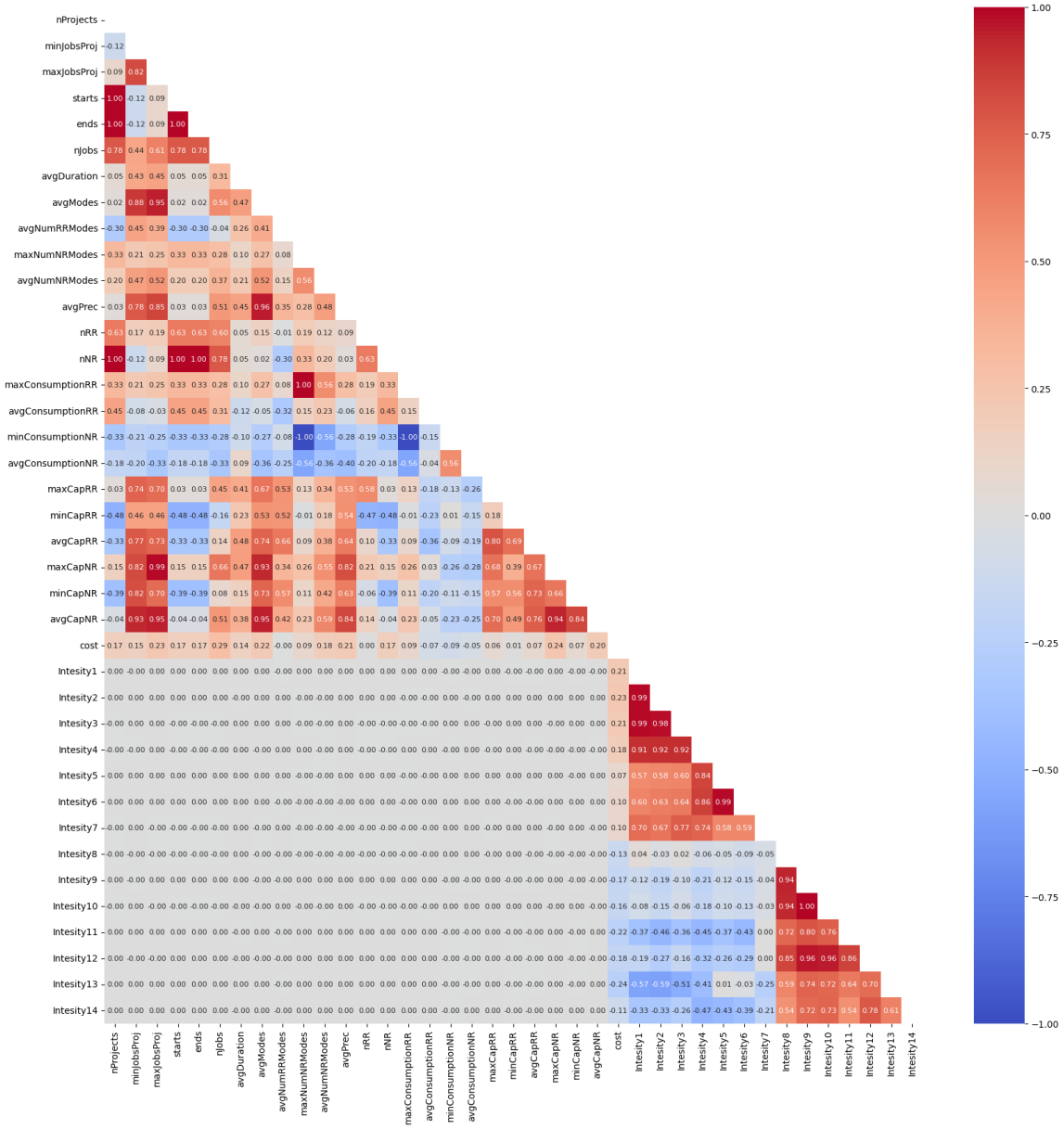
- `avgStartProj`;
- `avgEndProj`;
- `minDuration`;
- `maxDuration`;
- `minModes`;
- `maxModes`;
- `minPrec`;
- `maxPrec`;
- `minNumRRModes`;
- `maxNumRRModes`;
- `minNumNRModes`;
- `minConsumptionRR`;
- `maxConsumptionNR`.

3.2 Transformação

O coeficiente de correlação de *Pearson* foi utilizado para medir a força e direção das relações lineares entre variáveis contínuas, com o *heatmap* variando do azul ao vermelho. Embora a correlação de *Spearman* possa ser uma alternativa em casos de relações não lineares, observou-se uma distribuição de cores similar à do *heatmap* de Pearson, o que indicou que as relações entre as variáveis seguiam majoritariamente uma tendência

linear. Esse processo foi crucial para identificar padrões e destacar variáveis fortemente correlacionadas. A Figura 5 apresenta o *heatmap* da correlação de Pearson para todas as variáveis da base de dados.

Figura 5 – *Heatmap* da correlação de Pearson

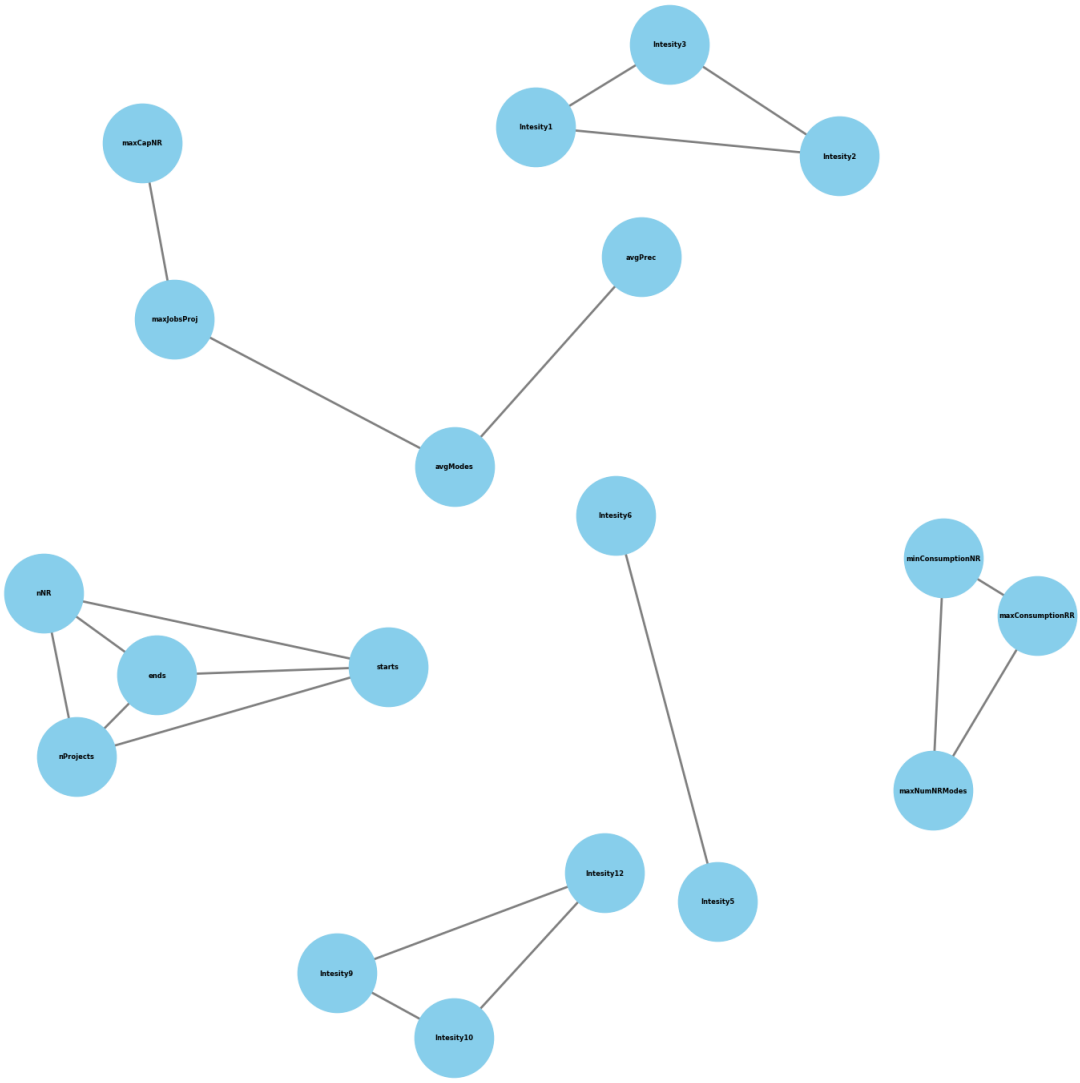


Fonte: Os autores

Para evitar multicolinearidade, variáveis com correlação superior a 0,9 foram removidas, conforme sugerido por Akoglu (2018), que indica que coeficientes próximos de 0,9 são considerados fortes. Essa estratégia simplifica o conjunto de dados e elimina redundâncias. A remoção de uma variável de cada par altamente correlacionado permite manter apenas as informações mais relevantes. A Figura 6 apresenta o diagrama de rede

de correlações, evidenciando os pares de variáveis correlacionadas.

Figura 6 – Diagrama de rede de correlações entre pares altamente correlacionados



Fonte: Os autores

A Tabela 3 apresenta os pares de variáveis com alta correlação e os respectivos valores.

A seguir são apresentadas as variáveis escolhidas para remoção. As variáveis `starts, ends e nNR` foram removidas por apresentarem correlação 1.0 com a variável `nProjects` e entre elas mesmas nos diagramas de rede. A variável `maxJobsProj` foi a que permaneceu em relação às demais no diagrama. Foi observada entre `maxConsumptionRR` e `minConsumptionNR` uma correlação negativa forte, assim como entre `minConsumptionNR` e `maxNumNRModes`. O coeficiente de -1 indica uma relação inversa, sugerindo que, à medida que uma dessas variáveis aumenta, a outra tende a diminuir, isto representa apenas a configuração de como cada grupo de instâncias da literatura foi criada considerando os

Variável 1	**Variável 2**	**Correlação**
nProjects	starts	1.0
nProjects	ends	1.0
nProjects	nNR	1.0
maxJobsProj	avgModes	0.9508
maxJobsProj	maxCapNR	0.9904
starts	ends	1.0
starts	nNR	1.0
ends	nNR	1.0
avgModes	avgPrec	0.9601
maxNumNRModes	maxConsumptionRR	1.0
maxNumNRModes	minConsumptionNR	-1.0
maxConsumptionRR	minConsumptionNR	-1.0
Intesity1	Intesity2	0.9898
Intesity1	Intesity3	0.9876
Intesity2	Intesity3	0.9831
Intesity5	Intesity6	0.9915
Intesity9	Intesity10	0.9971
Intesity9	Intesity12	0.9639
Intesity10	Intesity12	0.9603

Tabela 3 – Pares de variáveis com alta correlação

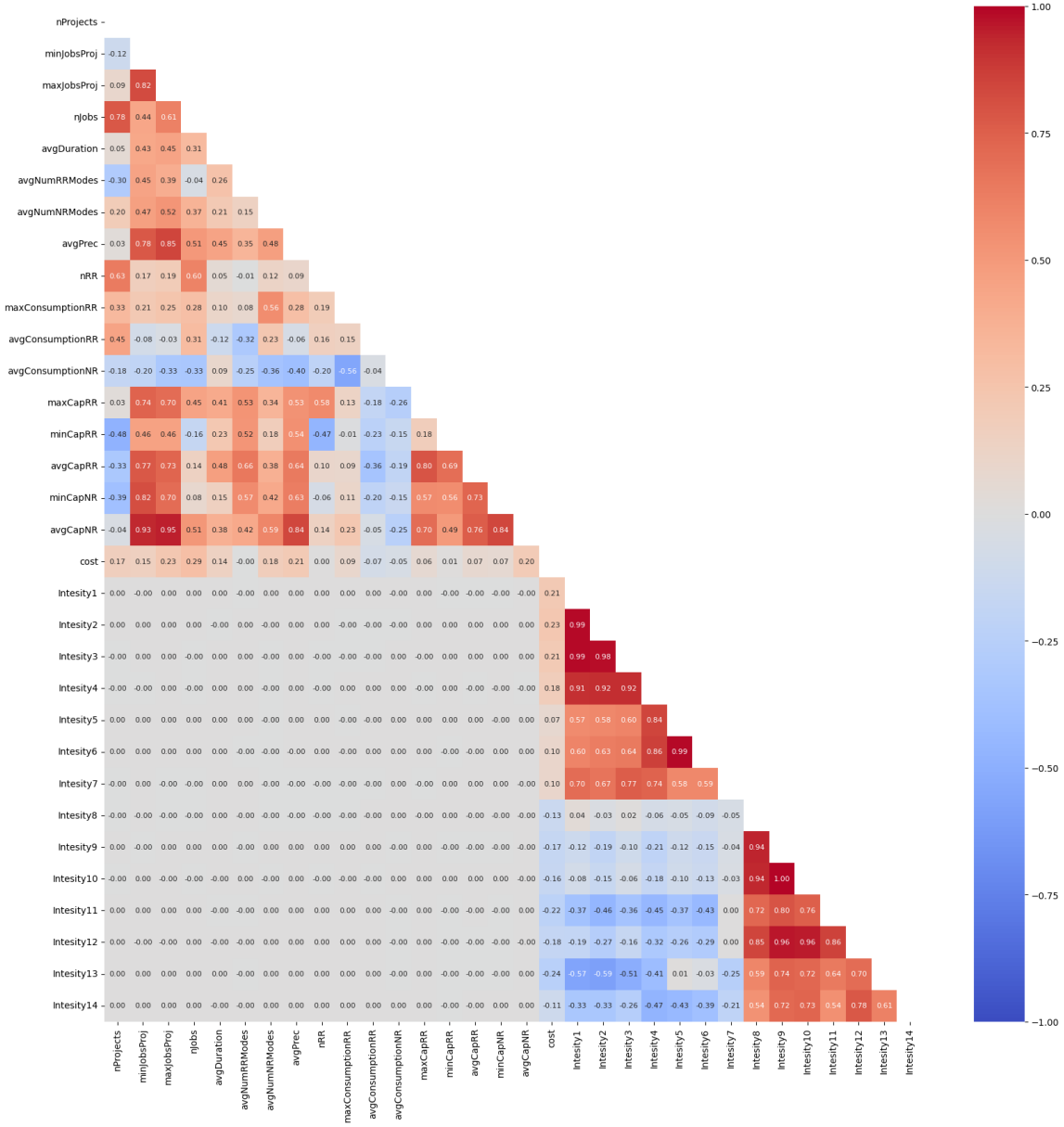
recursos renováveis e não renováveis. Como a correlação de `maxConsumptionRR` estava positivamente forte `maxNumNRModes` e negativamente forte com as outras, ela foi escolhida para permanecer no conjunto de dados. Nesta abordagem, não serão removidas as variáveis de intensidade. As intensidades mostram-se não relacionadas com as características dos projetos, como era de se esperar, além disso foram parametrizações preestabelecidas por Araujo et al. (2016) e são informações relevantes para identificações de padrões, estas mostram-se relacionadas com os valores de custo da função.

É importante lembrar que as variáveis de intensidades não foram removidas da base de dados, mesmo apresentando uma alta correlação entre si, devido à sua importância para a análise final dos resultados. Embora a correlação possa indicar redundância, essas variáveis desempenham um papel fundamental na interpretação dos padrões e na extração de conhecimento relevante. Sua manutenção permite uma avaliação mais completa e precisa, garantindo que informações essenciais não sejam perdidas no processo de análise.

Por fim, a Figura 7 apresenta o novo *heatmap* da correlação após a remoção das variáveis altamente correlacionadas, sem considerar as de intensidade, proporcionando uma visão mais clara das relações remanescentes, permanecendo um total de 32 variáveis.

A técnica de PCA foi aplicada para a redução de dimensionalidade, permitindo transformar o conjunto original de variáveis em um número menor de componentes principais. Conforme indicado na literatura (ver Seção 2), pode-se considerar as compo-

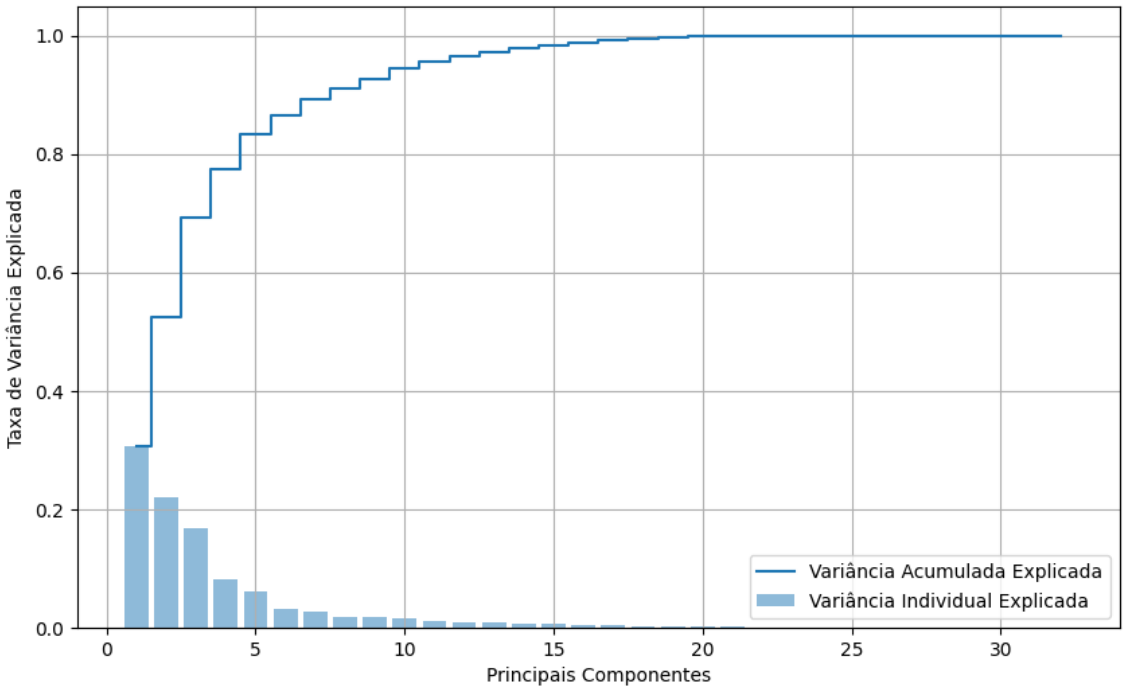
Figura 7 – *Heatmap* de correlação após redução



Fonte: Os autores

nentes principais que representam certa de 70% da variância total dos dados. Para esta primeira análise, foram realizados experimentos com 7, 4 e 3 componentes principais, preservando aproximadamente 90%, 80% e 70% da representatividade da base, respectivamente. Observou-se que, independentemente da quantidade de componentes utilizadas, o número ótimo de grupos após a aplicação do algoritmo *k-means* permaneceu em 6 em todas as situações.. Logo, para visualizarmos melhor as informações em apenas 3 dimensões serão apresentadas análises descritivas para as 3 componentes principais da Figura 8. Isso permite uma representação eficiente da base de dados com menor número de variáveis, mantendo as principais fontes de variação.

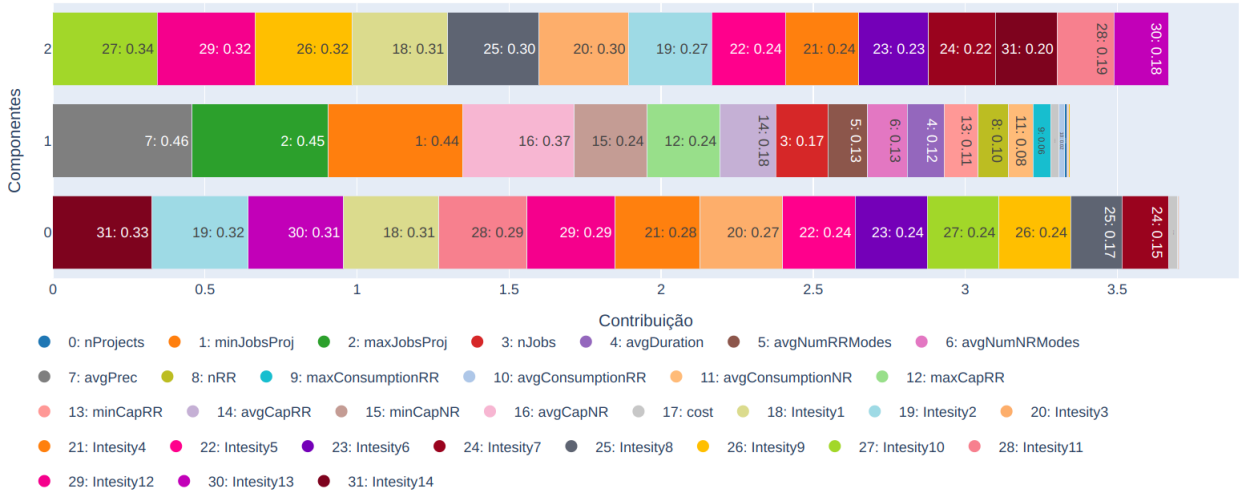
Figura 8 – Variância explicada



Fonte: Os autores

A Figura 9 mostra a contribuição acumulada de cada variável na criação das componentes principais.

Figura 9 – Contribuição acumulada das variáveis em cada uma das 3 componentes principais



Fonte: Os autores

A Tabela 4 apresenta as variáveis que mais contribuem para cada componente principal criada, descrevendo elas e nos mostrando alguns *insights* encontrados durante essa etapa.

PC	**Variáveis**	***Insight***
PC 0	`Intensity14`, `Intensity2`, `Intensity13`, `Intensity1`, `Intensity11`, `Intensity12`, `Intensity4`, `Intensity3`, `Intensity5`, `Intensity6`, `Intensity10`, `Intensity9`, `Intensity8`, `Intensity7`, `cost`	Esta componente reflete a influência das intensidades de vários fatores na configuração das vizinhanças, sugerindo que as intensidades são críticas na determinação do desempenho das soluções.
PC 1	`avgPrec`, `maxJobsProj`, `minJobsProj`, `avgCapNR`, `minCapNR`, `maxCapRR`, `avgCapRR`, `nJobs`, `avgNumRRModes`, `avgNumNRModes`, `avgDuration`, `minCapRR`, `nRR`, `avgConsumptionNR`, `maxConsumptionRR`, `cost`, `avgConsumptionRR`, `nProjects`	Esta componente reflete a complexidade das instâncias em relação à média de predecessores das atividade, ao número de atividades dos múltiplos projetos e as capacidades de recursos, além de sugerir também que a capacidade dos recursos, tanto renováveis quanto não-renováveis é um fator crítico para uma alocação eficiente em relação ao custo.
PC 2	`Intensity10`, `Intensity12`, `Intensity9`, `Intensity1`, `Intensity8`, `Intensity3`, `Intensity2`, `Intensity5`, `Intensity4`, `Intensity6`, `Intensity7`, `Intensity14`, `Intensity11`, `Intensity13`	Esta componente também reflete a importância das diferentes configurações de intensidades das vizinhanças, indicando que ajustar essas intensidades pode levar a melhores resultados em diferentes parametrizações.

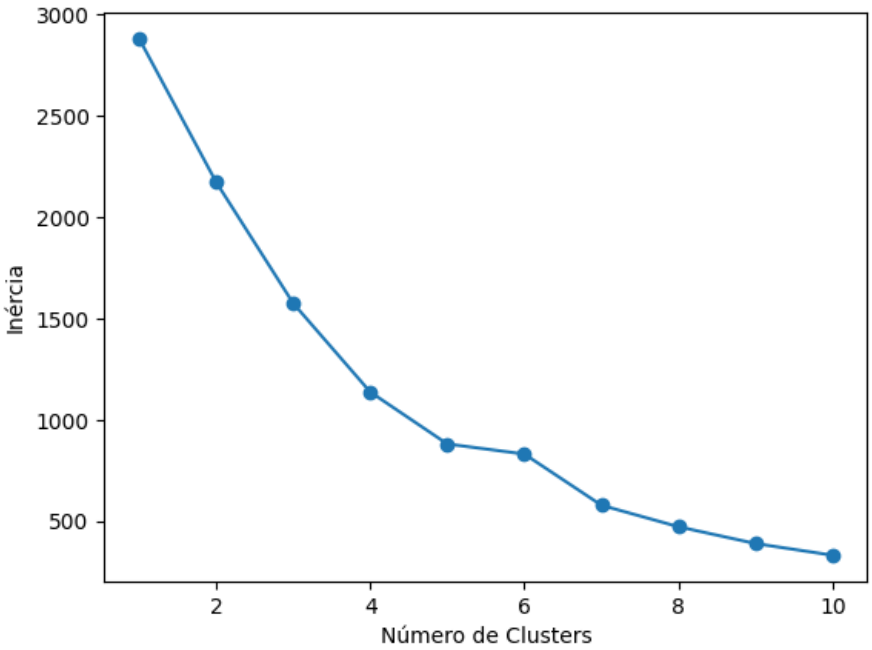
Tabela 4 – Análise das contribuições e *insights* por componente

3.3 Mineração de dados

A etapa de mineração de dados no processo KDD desempenha um papel essencial na extração de padrões, conhecimentos e *insights* significativos. Essa fase aplica técnicas computacionais e algoritmos para explorar conjuntos de dados em busca de padrões previamente desconhecidos e informações ocultas.

Para determinar o número ideal de *clusters*, foi utilizado o método do cotovelo (*elbow method*), uma técnica amplamente empregada em algoritmos de agrupamento como o *k-means*. Na análise de *K-Means* com diferentes números de *clusters*, foi calculada a variação intra-cluster para cada configuração, conforme ilustrado na Figura 10.

Figura 10 – Elbow



Fonte: Os autores

O ponto em que a adição de *clusters* deixa de trazer melhorias significativas na variação intra-cluster, identificado como "cotovelo", foi calculado considerando o ponto de equilíbrio (ponto da curva mais distante de uma reta traçada entre os pontos, considerando a fórmula da Equação 3.1.

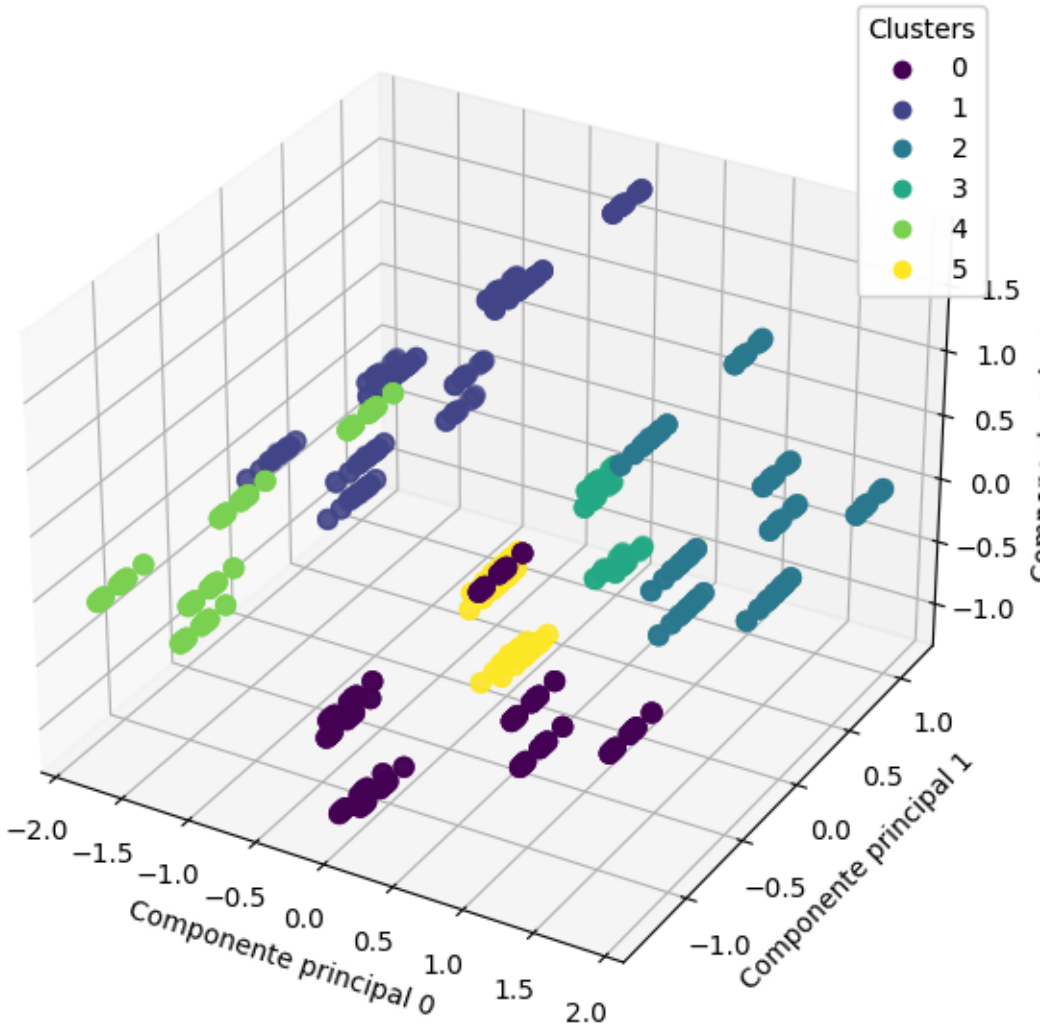
$$distância(P_0, P_1, (x, y)) = \frac{|(y_1 - y_0)x - (x_1 - x_0)y + x_1y_0 - y_1x_0|}{\sqrt{(y_1 - y_0)^2 + (x_1 - x_0)^2}} \tag{3.1}$$

Para ambos os experimentos com representatividade 70–90% das componentes o melhor valor de k foi 6. Esse ponto sinaliza a quantidade ideal de *clusters*, capturando adequadamente a estrutura dos dados e proporcionando uma base sólida para o agrupamento com o algoritmo *k-means*.

O algoritmo *k-means*, importado da biblioteca `sklearn`, foi utilizado para realizar o agrupamento da base de dados. Como um algoritmo de aprendizado não supervisionado, ele analisa exclusivamente as características dos dados de entrada para formar grupos, ou *clusters*. A visualização em 3 dimensões das componentes principais em relação à distribuição dos *clusters* é apresentada na Figura 11.

A análise nos gráficos das componentes principais permite uma compreensão dos fatores, a serem apresentados na Seção 3.4 que influenciam a distribuição e agrupamento das instâncias, fornecendo pistas sobre como melhorar a parametrização das vizinhanças e otimizar o uso de recursos.

Figura 11 – Visualização dos *clusters* 0, 1, 2, 3, 4 e 5 para as componentes /0, 1 e 2/



Fonte: Os autores

3.4 Análise dos resultados

O conjunto de técnicas e métodos escolhidos oferece uma abordagem abrangente para a análise de dados. Começamos com a análise de correlação para entender as relações entre variáveis, seguida pela redução de dimensionalidade usando PCA. Em seguida, determinamos o número ideal de *clusters* utilizando o *Elbow Method* e finalmente aplicamos o algoritmo *k-means* para agrupar os dados. Essas etapas forneceram *insights* valiosos para a compreensão do conjunto de dados e podem orientar decisões futuras com base nos padrões identificados.

A Tabela 5 apresenta de forma simplificada os *insights* extraídos ao analisar os valores das correlações com a variável custo da Figura 5. É importante lembrar que a variável custo da base de dados utilizada representa a qualidade da solução, nesse caso,

quanto menor o valor do custo melhor é sua qualidade entre as possiveis soluções.

Variáveis	**Correlação com Custo**	***Insight***
`nProjects`, `minJobsProj`, `maxJobsProj`, `starts`, `ends`, `nJobs`	0.15 a 0.29	O número de projetos e atividades tem correlação positiva com o custo, indicando que mais projetos/atividades elevam a complexidade e o custo.
`avgDuration`, `avgModes`	0.14 a 0.22	Atividades mais longas e modos de execução variados podem aumentar o custo devido ao uso adicional de recursos.
`avgNumRRModes`, `maxNumNRModes`, `avgNumNRModes`	0.00 a 0.18	Modos com mais recursos não renováveis podem aumentar o custo, enquanto modos com recursos renováveis têm pouca influência.
`avgPrec`	0.21	Precedências podem aumentar a complexidade e, consequentemente, o custo.
`nRR`, `nNR`	0.00 a 0.17	O número de recursos renováveis sozinhos parecem não impactar o custo, mas mais recursos não renováveis podem elevar o custo.
`maxConsumptionRR`, `avgConsumptionRR`	-0.07 a 0.09	O alto consumo de recursos renováveis pode aumentar o custo, mas a média de consumo pode ser gerida eficientemente.
`minConsumptionNR`, `avgConsumptionNR`	-0.09 a -0.05	Otimizar o uso de recursos não renováveis pode reduzir o custo.
`maxCapRR`, `avgCapRR`, `maxCapNR`, `avgCapNR`	0.06 a 0.24	Capacidades de recursos não renováveis têm impacto mais forte no custo, enquanto recursos renováveis têm correlação fraca.
`Intensity1`, `Intensity2`, `Intensity3`, `Intensity4`	0.18 a 0.22	Estratégias simples de vizinhança tendem a aumentar o custo devido à complexidade introduzida.
`Intensity8`, `Intensity9`, `Intensity10`, `Intensity11`, `Intensity12`, `Intensity13`, `Intensity14`	-0.24 a -0.13	Estratégias avançadas de vizinhança podem ajudar a otimizar o uso de recursos, reduzindo o custo.

Tabela 5 – Correlação de variáveis com o custo e seus impactos

Ao analisar a tabela e os valores das correlações, percebe-se que o número de atividades e o número de recursos não renováveis têm as maiores influências positivas no custo, o que é esperado, pois projetos mais longos e com mais recursos aumentam a complexidade e o custo. Além disso, a otimização dos recursos não renováveis e a redução

de precedências são fatores críticos para minimizar o custo em soluções de MMRCMPSP. E em relação as intensidades de vizinhança, as estratégias avançadas tendem a diminuir o custo quando bem aplicadas, pois otimizam o uso de modos e recursos, enquanto estratégias mais simples podem aumentar o custo. As intensidades das primeiras vizinhanças (como inverter sequência de tarefas, deslocar ou trocar duas tarefas) mostram uma correlação positiva com o custo. Isso indica que a aplicação dessas estratégias de vizinhança pode aumentar o custo, talvez porque essas modificações introduzam mais complexidade de tempo ou alterações nos recursos. As vizinhanças mais avançadas, como mudanças em múltiplos modos ou compactação de projetos subsequentes, mostram correlação negativa, sugerindo que essas estratégias podem ser mais eficazes para reduzir o custo ao otimizar a utilização de recursos e tempo.

Após aplicar o algoritmo *k-means* sobre as base de dados reduzida pelas 3 componentes principais, pode-se compreender melhor os 6 *clusters* que agrupam a base de dados nas 3 dimensões do PCA como pode ser visto na Figura 11. Já a Tabela 6 apresenta as informações sobre os *clusters* encontrados.

Tabela 6 – Sumário dos Grupos Identificados pelos *Clusters*

Cluster	**Num.**	**Descrição e Características**
0	216	O grupo apresenta dispersão significativa, possivelmente contendo execuções das instâncias com características intermediárias em termos de intensidade e capacidades de recursos, sem valores extremos.
1	315	O maior *cluster*, representando a maioria das execuções das instâncias. Agrupa dados com características médias ou balanceadas, provavelmente relacionadas à intensidade e capacidade equilibradas.
2	252	Grupo intermediário com certa variabilidade entre as instâncias. Possivelmente, os dados aqui têm relação com o custo ou número de atividades dos projetos, além das capacidades de recursos.
3	120	Um dos menores grupos, possivelmente representando instâncias mais especializadas com características fora da média, como alta capacidade ou intensidades extremas.
4	135	*Cluster* pequeno, refletindo padrões diferenciados, possivelmente relacionados a durações e modos de execução de projetos com maior restrição de recursos.
5	132	Grupo menor com características específicas, como projetos com custos extremos ou intensidades controladas, indicando menor flexibilidade em termos de recursos.

A Tabela 7 detalha a distribuição nos 6 *clusters*, cada linha da tabela representa uma instância, enquanto cada coluna indica a quantidade de ocorrências, exemplos das resoluções da instância e seu agrupamento em um dos 6 *clusters*. Em Araujo et al. (2016) foram realizadas 39 execuções com as diferentes configurações de intensidades de

vizinhanças para cada instância. Essa distribuição é essencial para compreender como as classes das instâncias se agrupam de acordo com as características analisadas, permitindo identificar padrões e tendências entre os *clusters*.

Tabela 7 – Divisão das instâncias A, B e X

Instância	***Cluster 0***	***Cluster 1***	***Cluster 2***	***Cluster 3***	***Cluster 4***	***Cluster 5***
Instâncias A						
A-1	24	0	0	0	15	0
A-2	0	15	12	0	0	12
A-3	0	15	12	12	0	0
A-4	24	0	0	0	15	0
A-5	0	15	12	0	0	12
A-6	0	15	12	12	0	0
A-7	24	0	0	0	15	0
A-8	0	15	12	0	0	12
A-9	0	15	12	12	0	0
A-10	0	15	12	12	0	0
Instâncias B						
B-1	24	0	0	0	15	0
B-2	0	15	12	0	0	12
B-3	0	15	12	12	0	0
B-4	24	0	0	0	15	0
B-5	0	15	12	0	0	12
B-6	0	15	12	12	0	0
B-7	24	0	0	0	15	0
B-8	0	15	12	0	0	12
B-9	0	15	12	12	0	0
B-10	0	15	12	0	0	12
Instâncias X						
X-1	24	0	0	0	15	0
X-2	0	15	12	0	0	12
X-3	0	15	12	12	0	0
X-4	24	0	0	0	15	0
X-5	0	15	12	0	0	12
X-6	0	15	12	12	0	0
X-7	24	0	0	0	15	0
X-8	0	15	12	0	0	12
X-9	0	15	12	12	0	0
X-10	0	15	12	0	0	12

Através da tabela foi possível perceber que as instâncias são organizadas em grupos bem distribuídos. Esses grupos são definidos pelos números das instâncias (1 a 10), independentemente das letras A, B ou X.

É possível ver as principais características de cada *cluster* levando em consideração a complexidade de cada uma das instâncias que representam eles. Na Tabela 8 a seguir é possível ver as principais diferenças entre as instâncias de cada *cluster*.

Dessa forma é possível identificar grupos dentro dos *clusters* e às diferenças nas suas características:

- Grupo 1 (*Clusters* 0 e 4): Instâncias pequenas, com baixa complexidade e custo.

Tabela 8 – Características das instâncias de cada */clusters*

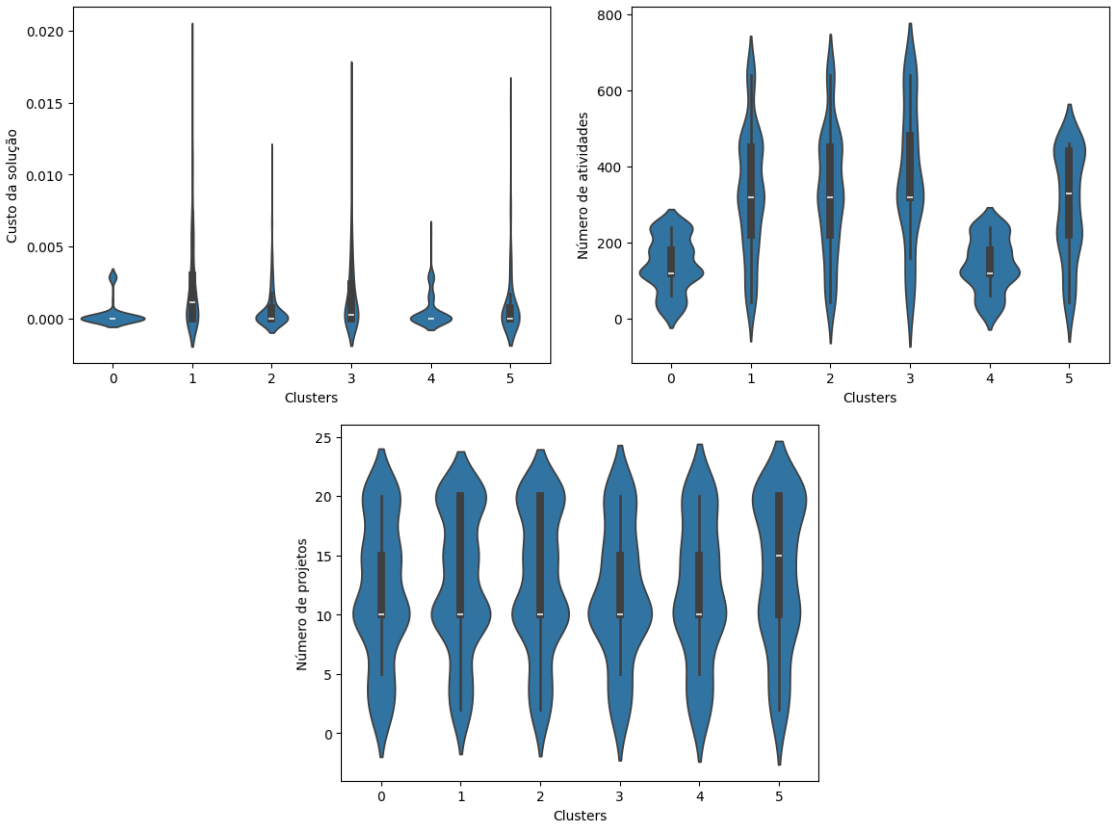
Cluster	Nº de Instâncias	nProjects	nJobs
Cluster 0	9	Predominantemente 2	24
Cluster 1	21	Varia de 2 a 20	Varia de 24 a 640
Cluster 2	15	Varia de 10 a 20	Varia de 320 a 640
Cluster 3	10	Predominantemente 20	640
Cluster 4	9	Predominantemente 2	24
Cluster 5	12	Varia de 10 a 20	Varia de 320 a 440

- Grupo 2 (*Cluster* 1): Instâncias heterogêneas, abrangendo desde pequenas até grandes, com alta variabilidade.
- Grupo 3 (*Clusters* 2 e 5): Instâncias intermediárias a grandes , com custos elevados.
- Grupo 4 (*Cluster* 3): Instâncias muito grandes, alta complexidade e custos altos.

O gráfico de violino é uma ferramenta importante na análise de dados, pois combina aspectos de um *boxplot* com a distribuição de densidade dos dados. Dessa forma, ele permite uma visualização mais detalhada da distribuição, destacando tanto a concentração dos valores quanto a forma da distribuição, seja ela simétrica ou assimétrica. Além disso, o gráfico de violino é útil para comparar múltiplos grupos ou categorias em uma única visualização, facilitando a identificação de padrões, diferenças e *outliers* entre os grupos. Na Figura 12 é possível comparar os diferentes *clusters* e suas características individuais a cerca do custo das soluções, número total de atividades e de projetos.

A avaliação dos gráficos de violino, que mostram a distribuição das variáveis "Número de Atividades", "Número de Projetos"e "Custo", indicou que os clusters 0 e 4 agrupam instâncias menos complexas, com menor variabilidade no número de atividades e mantêm custos consistentemente baixos. Por outro lado, os clusters 1, 2, 3 e 5, associados a instâncias mais complexas, apresentam maior variabilidade nos custos.

Figura 12 – Gráfico de violino



Fonte: Os autores

4 Conclusão

O foco principal deste trabalho foi fazer uma análise descritiva de uma base de dados que continha casos resolvidos, com o propósito de identificar padrões e tendências que não só facilitassem a compreensão da base, mas também ajudassem na classificação de outros exemplos encontrados na literatura.

Por meio da metodologia utilizada, foi observado conexões relevantes entre o custo e as diferentes variáveis e intensidades. Variáveis como número de projetos e número de atividades de cada projeto possuem uma correlação positiva com o custo, com valores entre 0.15 e 0.29, o que indica que um aumento na quantidade de projetos/atividades eleva a complexidade e, consequentemente, os custos. Outras variáveis importantes que também aumentam o custo/complexidade são uma maior duração média de atividades e mais opções de modos de execução que possuem correlação entre os valores 0.14 e 0.22.

Foi percebido também que em as intensidades utilizadas para resolução das instâncias impactaram positivamente e negativamente o custo. Nos casos negativos onde houve um aumento do custo, foi notado que em grande maioria eram estratégias simples como inverter a sequência de trabalhos e trocar a posição de dois trabalhos, com seus valores de correlação variando entre 0.18 e 0.22. Já em relação aos casos positivos onde a correlação foi negativa eram estratégias mais complexas como alteração do modo de execução duas ou três vezes, por exemplo, seus valores por sua vez variam entre os valores de -0.13 a -0.24.

Outra observação feita, foi em relação às componentes principais, onde foi possível ver quais variáveis contribuíram para a formação de cada componente. Além disso, foi possível perceber para a componente principal PC2, a importância das diferentes configurações de intensidades.

A técnica de agrupamento foi essencial para reconhecer e estruturar *clusters* com instâncias semelhantes. Foram encontrados seis grupos distintos ordenados de 0 a 5, onde cada um possui suas características e tendências. Por exemplo o *cluster* 1 que agrupa 315 instâncias e possuem dados com características médias e em contrapartida o *cluster* 3 que agrupa somente 120 instâncias e possuem características mais extremas.

Foi possível concluir que o processo de *clustering* revelou uma estrutura clara baseada nos números das instâncias (1 a 10), organizando-as em três grupos distintos: Grupo 1 com instâncias 1, 4, 7, associado aos *clusters* 0 e 4; Grupo 2 com instâncias 2, 5, 8, 10, ligado aos *clusters* 1, 2 e 5 e por fim o Grupo 3 com as instâncias 3, 6, 9, conectado aos *clusters* 1, 2 e 3. As letras A, B e X, que acompanham as instâncias, não influenciaram o agrupamento, sugerindo que representam apenas variações dentro do mesmo tipo de instância.

Nova Analise

Em resumo, os dados mostram que a utilização de técnicas de mineração de dados auxiliou na compreensão também permitiu novas oportunidades para abordagens mais efetivas na resolução dos problemas. Os achados têm a capacidade de melhorar modelos futuros.

Finalmente, este estudo faz uma contribuição importante para estudos futuros, enfatizando a relevância das análises descritivas como uma etapa essencial na exploração e entendimento de grandes conjuntos de dados.

4.1 Trabalhos Futuros

Quanto aos próximos passos desta monografia, algumas opções de aprimoramento e expansão podem ser consideradas para aprimorar tanto a profundidade quanto a eficácia da análise de dados.

Uma sugestão é expandir e variar a base de dados. Com um aumento da quantidade de dados ou inclusão de outras fontes de informação, será viável realizar uma análise mais sólida, podendo ainda descobrir padrões inéditos na base de dados existente.

Outra opção é usar métodos de análise mais avançados. Utilizar outras estratégias de *machine learning* ou adotar modelos estatísticos mais elaborados pode auxiliar na detecção de padrões e tendências de forma mais precisa. Esta metodologia pode proporcionar *insights* mais profundos e relevantes, sendo de grande ajuda para pesquisas futuras.

Como por exemplo a aplicação de métodos supervisionados de aprendizado de máquina, uma vez que este trabalho concentrou-se exclusivamente em técnicas não supervisionadas para a análise da base de dados. A utilização de abordagens supervisionadas poderá permitir uma avaliação mais precisa do desempenho dos modelos. Em conjunto com o método *SHapley Additive exPlanations* (SHAP) para a interpretação dos resultados, para uma análise mais aprofundada da importância das variáveis.

Essas trajetórias de progresso oferecem vastas oportunidades de crescimento e criação, tanto em termos da quantidade de informações a serem investigadas quanto da complexidade das análises e da eficácia do procedimento.

Referências

AHMETI, A.; MUSLIU, N. Hybridizing constraint programming and meta-heuristics for multi-mode resource-constrained multiple projects scheduling problem. In: *Proceedings of the 13th international conference on the practice and theory of automated timetabling-patat*. [S.l.: s.n.], 2021. v. 1, p. 14. Citado na página 13.

AKOGLU, H. User's guide to correlation coefficients. *Turkish journal of emergency medicine*, Elsevier, v. 18, n. 3, p. 91–93, 2018. Citado na página 29.

ARAUJO, J. et al. Neighborhood composition strategies in stochastic local search. *test*, 2016. Citado 4 vezes nas páginas 13, 19, 31 e 38.

ARAUJO, J. A. S. Mixed-integer linear programming based approaches for the resource constrained project scheduling problem. 2019. Citado na página 13.

BENESTY, J. et al. Pearson correlation coefficient. In: _____. *Noise Reduction in Speech Processing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 1–4. ISBN 978-3-642-00296-0. Disponível em: <https://doi.org/10.1007/978-3-642-00296-0_5>. Citado na página 16.

BLAZEWICZ, J.; LENSTRA, J. K.; KAN, A. R. Scheduling subject to resource constraints: classification and complexity. *Discrete applied mathematics*, Elsevier, v. 5, n. 1, p. 11–24, 1983. Citado na página 13.

CHU, X. et al. A data-driven meta-learning recommendation model for multi-mode resource constrained project scheduling problem. *Computers and Operations Research 157 (2023)*, 2023. Citado na página 19.

CUI, M. et al. Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance*, Clausius Scientific Press, v. 1, n. 1, p. 5–8, 2020. Citado na página 17.

DING, C.; HE, X. K-means clustering via principal component analysis. In: *Proceedings of the twenty-first international conference on Machine learning*. [S.l.: s.n.], 2004. p. 29. Citado na página 16.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI Magazine Volume 17 Number 3 (1996) (© AAAI)*, 1996. Citado 2 vezes nas páginas 15 e 25.

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*, SAGE Publications Sage CA: Los Angeles, CA, v. 44, n. 3, p. 348–361, 2019. Citado na página 19.

HARTMANN, S.; BRISKORN, D. An updated survey of variants and extensions of the resource-constrained project scheduling problem. *European Journal of operational research*, Elsevier, v. 297, n. 1, p. 1–14, 2022. Citado na página 20.

JIAWEI, H. Data mining techniques. In: *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*. [S.l.: s.n.], 1996. p. 545. Citado na página 15.

JOLLIFFE, I. T. *Principal Component Analysis*. [S.l.]: Springer New York, NY, 2002. Citado na página 16.

KANANI, P.; PADOLE, M. Deep learning to detect skin cancer using google colab. *International Journal of Engineering and Advanced Technology Regular Issue*, v. 8, n. 6, p. 2176–2183, 2019. Citado na página 18.

LEE, C. S.; CHEANG, P. Y. S.; MOSLEHPOUR, M. Predictive analytics in business analytics: decision tree. *Advances in Decision Sciences*, Asia University, Taiwan, v. 26, n. 1, p. 1–29, 2022. Citado na página 15.

LEPENIOTI, K. et al. Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, Elsevier, v. 50, p. 57–70, 2020. Citado na página 15.

MAHESH, B. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, v. 9, n. 1, p. 381–386, 2020. Citado na página 16.

MENEZES, N. N. C. Introdução a programação com python. *São Paulo: Novatec*, 2010. Citado na página 18.

NASTESKI, V. An overview of the supervised machine learning methods. *Horizons. b*, v. 4, p. 51–62, 2017. Citado na página 16.

PINTO, D. F. *Uma análise exploratória sobre o uso das competências em analytics na gestão de projetos*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2020. Citado na página 15.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, p. 53–65, 1987. ISSN 0377-0427. Disponível em: <https://www.sciencedirect.com/science/article/pii/0377042787901257>. Citado na página 17.

SAHOO, K. et al. Exploratory data analysis using python. *International Journal of Innovative Technology and Exploring Engineering*, v. 8, n. 12, p. 4727–4735, 2019. Citado na página 18.

SANTOS, H. G.; SOARES, J.; TOFFOLO, T. Hybrid local search for the multi-mode resource-constrained multi-project scheduling problem. In: *Proceedings of the 10th International Conference on the Practice and Theory of Automated Timetabling (PATAT'14)*. [S.l.: s.n.], 2014. p. 397–407. Citado na página 13.

SHUKLA, S.; NAGANNA, S. A review on k-means data clustering approach. *International Journal of Information & Computation Technology*, v. 4, n. 17, p. 1847–1860, 2014. Citado na página 17.

SIAL, A. H.; RASHDI, S. Y. S.; KHAN, A. H. Comparative analysis of data visualization libraries matplotlib and seaborn in python. *International Journal*, v. 10, n. 1, 2021. Citado na página 18.

TOFFOLO, T. A. et al. An integer programming approach to the multimode resource-constrained multiproject scheduling problem. *Journal of Scheduling*, Springer, v. 19, n. 3, p. 295–307, 2016. Citado na página 20.

TSENG, C.-C. Two heuristic algorithms for a multi-mode resource-constrained multi-project scheduling problem. *Journal of Science and Engineering Technology*, v. 4, n. 2, p. 63–74, 2008. Citado na página 13.

WEI, W.; RANA, M. E. Software project schedule management using machine learning data mining. *INTERNATIONAL JOURNAL OF SCIENTIFIC E TECHNOLOGY RESEARCH VOLUME 8, ISSUE 09, SEPTEMBER 2019*, 2019. Citado na página 19.

ZHANG, S.; ZHANG, C.; YANG, Q. Data preparation for data mining. *Applied artificial intelligence*, Taylor & Francis, v. 17, n. 5-6, p. 375–381, 2003. Citado na página 15.