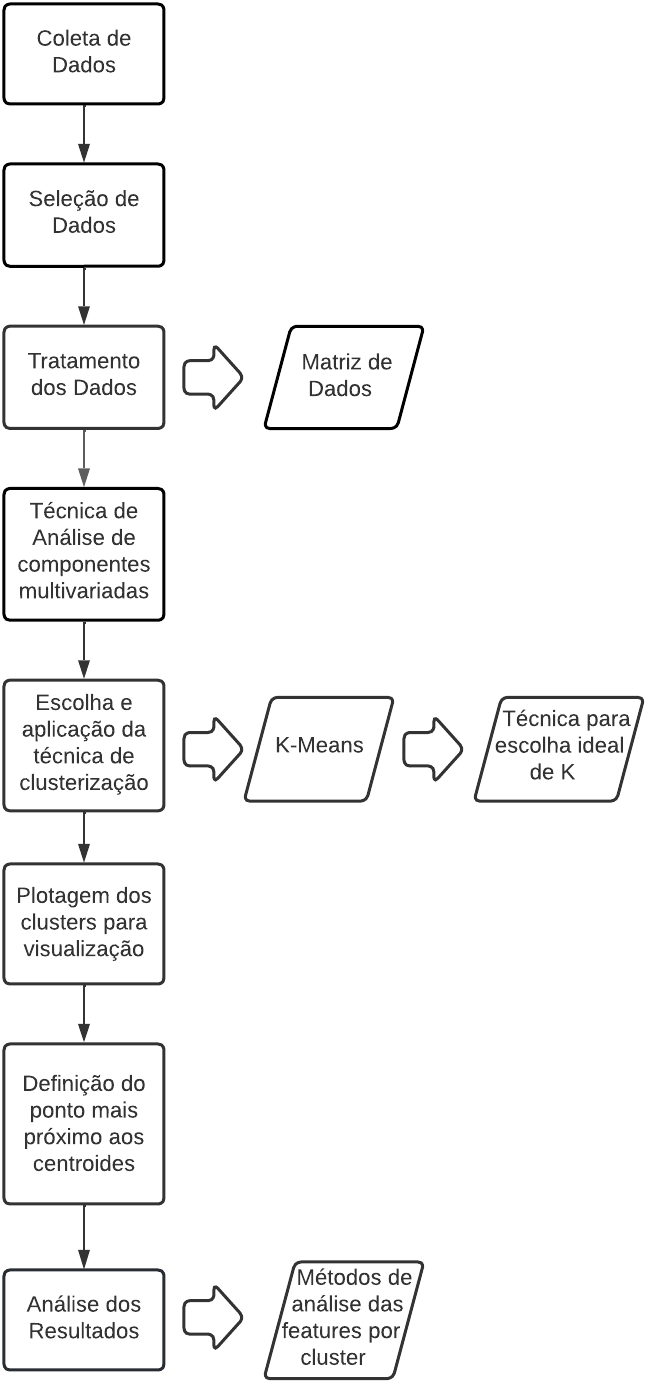
A metodologia adotada nesse trabalho pode ser vista na Figura a seguir. Nela, ilustra-se as principais etapas adotadas, assim como os seus produtos gerados.



(Figura 1. Visão geral da metodologia do trabalho)

Na primeira etapa, é realizado a coleta de dados. A fim de serem utilizados dados de diversas fontes, é realizado uma busca por dados de caracteristicas socioeconômicas e ambientais, buscando assim a pluralidade da informação levantada. Para o exemplo proposto, utilizou-se dados do Instituto Brasileiro de Geografia e Estatística (IBGE) e do Cadastro Ambiental Rural (CAR).

Já na segunda etapa, as principais *features* de interesse são escolhidas pelo usuário. Essa etapa pode contar ainda com a ajuda de um usuário especialista, o qual saberá definir quais são as informações mais relevantes a serem consideradas dado o problema. Caso haja dados privados, esses podem ser também utilizados pelo usuário. Para a terceira etapa, ocorre o tratamento dos dados. Os dados categóricos são convertidos para dados numéricos e os dados numéricos são adequados conforme o algoritmo de clusterização a ser utilizado. Dessa etapa, é obtido como resultado a matriz de dados a ser utilizado com a fonte de dados para o algoritmo. Na quarta etapa, é utilizado uma técnica de análise de componentes multivariadas. Essa etapa é importante para a metodologia adotada pois possiblita que a matriz multidimensional de dados seja convertida para uma matriz bidimensional, possibilitando assim a plotagem dos dados em um gráfico de dispersão bidimensional, por exemplo. Nesse caso, escolheu-se o *Principal Component Analysis* (PCA) como algoritmo para redução da dimensionalidade.

Na quinta etapa é realizada a escolha e aplicação de um algoritmo de clusterização, além da determinação do centroide de cada agrupamento formado. Para o exemplo, foi escolhido o algoritmo *Kmeans* por ser um algoritmo clássico e bem visto na literatura. Por escolher essa técnica, foi necessário a adoção também de um método para a escolha ideal do número de agrupamentos K. Para isso, usou-se o *Silhouette Score*.

Na sexta etapa, após a geração de todos os agrupamentos e a determinação de seus centroides, ocorre a plotagem dos dados em um gráfico. Dessa forma, os *clusters* formados podem ser vistos, assim como a distância *inter-clusters* e *intra-clusters*. Na etapa seguinte, é definido o ponto *intra-cluster* mais próximo do centroide determinado. Dessa forma, esse ponto é o elemento mais indicado para ser um representante do agrupamento, ou seja, um elemento que possuirá as propriedades mais caracteristicas do conjunto a que pertence. Por fim é realizado a análise dos resultados obtidos. Dessa forma, para cada agrupamento formado, ocorre a análise das *features* por meio do cálculo da média e mediana, e a plotagem desses valores em um *heatmap* para visualização geral dos dados. Assim, é possível identificar os principais atributos que levaram os dados àquele agrupamento e a consequente sugestão de priorização da alocação dos recursos disponíveis nesses mesmos atributos, buscando assim potencializar a recuperação da área atingida pelos desastres.