# CHAPTER 3

## RESEARCH METHODOLOGY

1. Literature Study

   This study uses the python programming language and the birch algorithm to perform the clustering process, and linear regression to fill in the missing values. Before doing clustering requires pre-processing data to select the data to be used, after selecting the data to be used the next step is to fill in the blank data. The next step after pre-processing is to cluster. From the results of the cluster will be able to make conclusions about the data relationship between height and medals in the Olympics.

2. Dataset

   The dataset used in this study is the 120 years of the Olympics, the data that will be used are all the athlete's height data and medals that have been obtained. This data was obtained from Kaggle, which was made by Mysar Ahmad Bhat.

3. Programs

   The program used in this study uses Jupiter Notebook 6.3.0 which uses the Python programming language. The data type used is CSV, the library used is NumPy, pandas, sklearn, missingno, the use of the library will be demonstrated and explained later.

4. Implementation & Testing

After preparing the pre-processed data, the next step is to cluster using the BIRCH algorithm. With the aim of and from the results of the cluster, it can be concluded whether height is related to athlete medals at the Olympics. And calculate the cluster performance generated by the BIRCH algorithm using V-Measure.