CHAPTER 4

ANALYSIS AND DESIGN

4.1 Analysis

This chapter discusses what the problem will be solved, the data to be used, the data used, the explanation of the BIRCH algorithm. This explanation is in the form of a narrative and a table is given for further explanation.

4.1.1 Problem and Data

The problem that will be discussed in this research is whether the height of athletes in the Olympics is related to the medals they have won, the stigma of society is that height is very influential in winning medals for an athlete, especially in basketball. This study will prove whether the height of the basketball athlete has an effect on the athlete's medal achievement or maybe height has nothing to do with the medal won by the basketball athlete. In addition to basketball which was made for this research, other sports can also be investigated whether height is related to the athlete's victory. The sports that will be studied besides basketball are football, volleyball, beach volleyball, and all sports in the Olympics.

4.1.2 Data Collection and Data Usage

The data that will be used is the 120-year history of the Olympics which is obtained at www.kaggle.com in which there are four CSV datasets provided. CSV that needs to be used is "athlete_events.csv" this CSV contains data on athletes from name, gender, age, height, country, etc. Full data explanation is in the table

by pre-processing the data and filling in the missing data, the data used is height and medals won.

Table 4.1: athlete_events CSV data

Colums Names	Description
ID	Unique number for each athlete
Name	Athlete's name
Sex	Male (M) or Female (F)
Age	Integer
Height	In centimeters
Weight	In kilograms
Team	Team name
NOC	National Olympic Committee 3-letter code
Games	Year and Season
Year	Integer
Season	Summer or Winter
City	Host City
Sport	Sport
Event	Event
Medal	Gold, Silver, Bronze, or NA

4.1.3 Algorithm BIRCH

BIRCH (Balanced Reducing and Clustering using Hierarchies) algorithm is a hierarchical algorithm. The BIRCH algorithm uses a tree structure to create clusters, which are commonly called Clustering Feature trees (Cf-trees). BIRCH builds on the idea that points that are close enough to one another should always be considered as a group. The CFs provide this level of abstraction. In other words, the core of the BIRCH clustering algorithm is the CF[3]. BIRCH (Balanced Reducing and Clustering using Hierarchies) algorithm is a hierarchical algorithm. The BIRCH algorithm uses a tree structure to create clusters, which are commonly called Clustering Feature trees (Cf-trees). BIRCH builds on the idea that points that are close enough to one another should always be considered as a group. The CFs provide this level of abstraction. In other words, the core of the

BIRCH clustering algorithm is the CF. The BIRCH algorithm consists of four stages:

1. Scanning a database to formulate an in-memory CF tree.
2. Building smaller CF trees.
3. Performing a global clustering.
4. Refining clusters, which is not mandatory and requires more scans of the dataset.

The weakness of the BIRCH algorithm is that it can only manage data in the form of numbers. BIRCH algorithm uses 3 important parameters: branching factor, number of clusters, threshold. While the data points of given dataset are entered into BIRCH, a height-balanced CF tree of hierarchical clusters is built. Each node represents a cluster in the cluster hierarchy where leaf nodes are the actual clusters and intermediate nodes are super clusters. The branching factor Br is the maximum number of children a node can have. Then, when a leaf is reached, the new point is added to this leaf cluster, which will not increase the radius of the cluster beyond the threshold (T). Otherwise, the new point is assigned into a new created cluster as its only member. As a result, the size of the clusters is obviously controlled by the threshold parameter T.

4.1.4 Homogeneity, Completeness, V-Measure

V-Measure is a method to calculate the performance of a cluster that is created, if the V-Measure number is closer to the number 1 then the performance of the cluster created is getting better. By calculating the average homogeneity and completeness. homogeneity is homogeneous clustering is one where each cluster has data points belonging to the same class label. Homogeneity describes the closeness of the clustering algorithm to this perfection. completeness is complete clustering is one where all data points belonging to the same class are clustered into the same cluster. Completeness describes the closeness of the clustering algorithm to this perfection, in this study will use the sklearn library to calculate homogeneity, completeness, and V-Measure. In the documentation, sklearn explains how to calculate the V-Measure using homogeneity and completeness and calculates the average using the formula as described in the image below.

$$v = 2 \cdot \frac{h \cdot c}{h + c}$$

Figure 1 : Formula V-Measure

The formula above is a mathematical way to find the V-Measure result using numbers that have been found in homogeneity and completeness. By using sklearn, you only need to use the provided library, an example of how to use sklearn to calculate homogeneity and completeness is as shown below.

```
>>> labels_true = [0, 0, 0, 1, 1, 1]
>>> labels_pred = [0, 0, 1, 1, 2, 2]
>>> metrics.homogeneity_completeness_v_measure(labels_true, labels_pred)
(1.0, 0.68..., 0.81...)
```

Figure 2 : Example Use Library Sklearn for Homogeneity, Completeness, and V-Measure

by using the code as above using the sklearn library, homogeneity, completeness can be calculated and even numbers have been found in the V-Measure. In the example above, the table containing the sample data for the numbers 0,0,0,1,1,1 is compared with the data for the numbers 0,0,0,1,2,2 and the results of homogeneity, completeness and V-Measure are found. with the result 1.0 in homogeneity, 0.68 in completeness, and 0.81 in V-Measure.

4.2 Desain

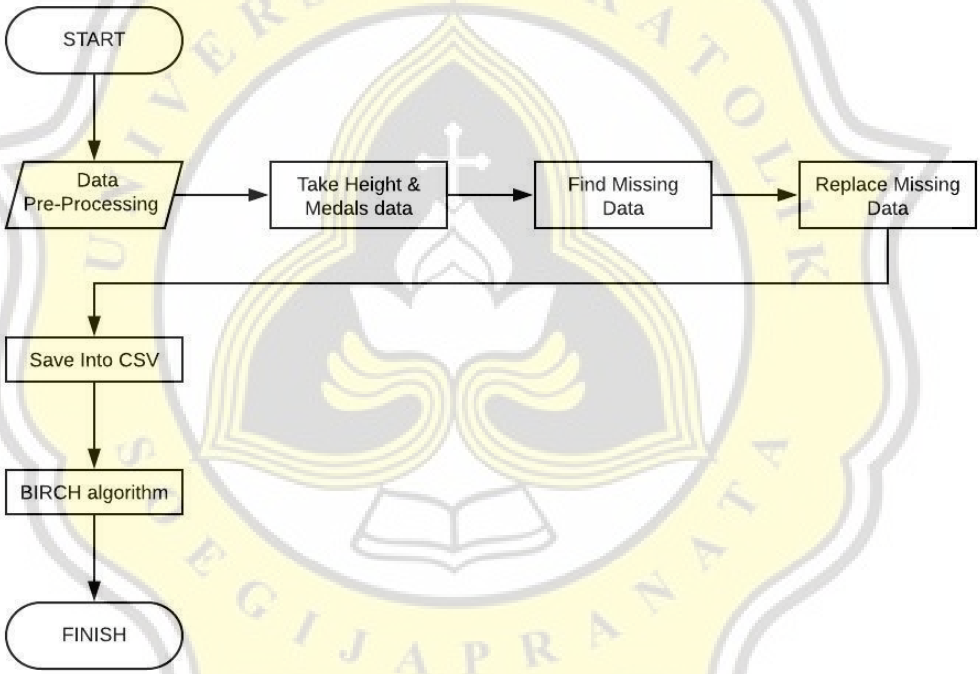

Figure 3: Flowchart Program

The flowchart above explains how the whole program runs. The first thing to do after getting the data is pre-processing the data. In the pre-processing data, height and medal data is collected, after that, looking for missing data to be replaced in the next step by using deterministic regression. After the data has finished pre-processing, it is then stored in the form of a CSV file, the data that has been saved earlier will be used for the clustering process.