THE ANALYSIS OF RELATIONSHIP BETWEEN HEIGHT OF BODY AND GAIN OF MEDALS IN OLYMPICS

[1]Willy Tri Wijaya [2] Rosita Herawati
[1,2]Program Studi Teknik Informatika Fakultas Ilmu Komputer, Universitas Katolik Soegijapranata
[2]rosita@unika.ac.id

ABSTRACT

The Olympics is an international sporting event held every four years and has lasted for 120 years, encompassing both summer and winter sports. The general opinion of people at the Olympics, in general, is that if short people enter a basketball competition, they won't win a medal. This study will prove whether public opinion about height affects an athlete's medal win. by using the BIRCH clustering algorithm to prove it is true or false. the data used is 120 years of Olympic history, the data taken is not completely intact, and there are missing data, the missing data will be replaced by using a deterministic regression method. The results of the research that has been done the results are that people's opinions are not entirely correct, besides the height factor other factors that affect the athlete's victory, namely the athlete's physical condition, the athlete's mental condition, and also how often athletes train to take part in the Olympics.

Keywords: BIRCH algorithm, Deterministic Regression, Olympics

INTRODUCTION

The Olympics is an international sporting event that is held once every four years and has been around for 120 years, which includes summer and winter sports. attended by more than 200 countries and tens of thousands of athletes. Various kinds of athletes take part in the Olympics to win medals, there are gold, silver, and bronze medals. From old to young, men and women, lightweight and heavyweight, tall people and short people. The general opinion of people at the Olympics, in general, is that if short people take part in basketball competitions, they will not win medals. From there the problem can be found, the public stigma of short athletes if they participate in the basketball Olympics will definitely lose or have trouble when competing, because the average height of basketball people is very high. This study will prove whether the stigma of society is true if the height of a basketball athlete can determine whether an athlete wins a medal or not.

RELATED WORK

There are many types of cluster algorithms, the cluster algorithms that are often used are K-Means, DBSCAN, Gaussian Mixture, Birch, etc. BIRCH is a text-mining clustering algorithm that is employed to perform Balanced Iterative Reducing and Clustering using Hierarchies [1]. It can effectively manage datasets that contain both numerical and categorical data. The algorithm introduces the concept of clustering feature (CF) and utilizes the CF tree to simplify the cluster representations. A type of tree used for clustering features

is known as a height-balanced tree. This tree stores the clustering features and has two parameters: the branching factor and the threshold. At first, BIRCH searches the database to form an initial CF tree. Following that, the CF tree's leaf nodes are clustered through the utilization of the suitable clustering algorithm. In 2015, Mamta Gupta presented a modified version of the BIRCH algorithm that calculates the average of all document points and utilizes the Jaccard measure to determine the distance between the mean and all points. After going through this process for every point, the BIRCH algorithm is then utilized. The enhanced BIRCH algorithm was discovered to be more effective compared to the current methods used for document clustering.

Besides being used as a regular clustering method, the BIRCH algorithm can also be used as a pre-clustering algorithm, state Kovacs and Bednarik [2]. They emphasize the importance of providing good intra-cluster similarity in the pre-clustering process for efficient data reduction in large data sets, rather than using the traditional method. The weakness of traditional methods do not consider this aspect and generate weak clusters. To illustrate, the HAC method is capable of producing clusters of any size, even if the distance between two elements within the same cluster is significantly greater than the distance between an element from another cluster. This implies that the elements in a cluster may not be inherently more alike to each other compared to elements in separate clusters. This contradicts the casual understanding of clustering.

BIRCH has the ability to reduce the time required for solving large-scale data problems, although its capability in generating high-quality clusters is severely lacking. The solution to these weaknesses includes adjusting the threshold value in a way that can adapt to the scale of the data point. The creation of the dynamic threshold involves increasing the scale of the leaf entry. The effectiveness of the proposed solution is assessed by using the silhouette coefficient (SC). The altered BIRCH algorithm has the potential to be a powerful method for managing the clustering of large datasets. The modified BIRCH algorithm shows a 60% decrease in the number of CF-Node results, total CF-Entries, and total CF-Leaf Entries, in comparison to the standard BIRCH algorithm [3].

The article by Lorbeer, et al. [4] mentioned that the BIRCH algorithm requires three important parameters: branching factor, threshold, and cluster count. The branching factor specifies the maximum number of CF sub-clusters in each node (internal node), a threshold the maximum number of data points a sub-cluster in the leaf node of the CF tree can hold, and cluster count the number of clusters to be returned after the entire BIRCH algorithm is complete. Using deep trees in tree-BIRCH is beneficial when dealing with data that contains a larger number of clusters. This text discusses the concept of supercluster splitting and presents MBD-BIRCH as an improved version of tree-BIRCH. MBD-BIRCH effectively reduces or eliminates supercluster splitting, all the while maintaining a quicker performance than tree-BIRCH with flat trees.

The traditional BIRCH algorithm uses distance to control the shape of clusters, but the clustering results for non-spherical datasets are not good. Guha, et al. [5] propose a novel clustering algorithm called ROCK for data with boolean and categorical attributes. The

improved algorithm draws on the link concept of the ROCK algorithm to overcome the traditional BIRCH algorithm limitation. Guo, et al. [6] also assess the effectiveness of a new algorithm called the LBIRCH algorithm and its performance compared to other clustering algorithms by utilizing three datasets: Spiral, Aggregation, and Double hemisphere. The Spiral dataset consists of 312 data objects in two dimensions. These objects can be separated into three clusters based on their shape. The Aggregation dataset comprises 788 sets of two-dimensional data points that can be categorized into seven clusters. The Double hemisphere dataset consists of 1500 objects with three dimensions, resembling two hemispherical surfaces with added random variation. Dongwei, et al. [6] state that the LBIRCH algorithm, which utilizes Link, is able to successfully cluster clusters of any shape. Additionally, it is found to be superior to other clustering algorithms in terms of both accuracy and efficiency. The findings from the experiments demonstrate that the LBIRCH algorithm is capable of effectively managing large datasets with complex dimensions and can produce superior clustering outcomes compared to the traditional BIRCH algorithm and other clustering algorithms. The suggested method has the potential to be utilized in a range of fields, including but not limited to image processing, data mining, and pattern recognition.

Madan and Dana [7] adopt a classic online clustering algorithm called BIRCH to incrementally cluster large datasets of features commonly used in visual clustering. The adapted version of BIRCH is called m-BIRCH, which enables data-driven parameter selection and effectively handles varying density regions in the feature space. In this research, the author utilizes extensive collections of visual clustering features, including 840,000 color SIFT descriptors, 1.09 million color patches, 60,000 grayscale patches affected by outliers, and 700,000 grayscale SIFT descriptors. The algorithm is employed to group sets of data that contain non-convex clusters, like the Hopkins 155 3D motion segmentation dataset. The m-BIRCH algorithm required only 10-20% of the dataset memory to perform the clustering.

RESEARCH METHODOLOGY

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

ANALYSIS AND DESIGN

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.

BIRCH Algorithm

BIRCH (Balanced Reducing and Clustering using Hierarchies) algorithm is a hierarchical algorithm. The BIRCH algorithm is a clustering algorithm that is well-suited for extremely large data sets [8]. The algorithm constructs a CF-tree by ensuring that each leaf node contains entries that meet a consistent threshold T. Additionally, the CF-tree is updated at each step with a new threshold value. However, the use of a single threshold in the birch algorithm leads to several drawbacks [9].

As 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[3]:

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.

In general, BIRCH serves as a beneficial clustering algorithm for text mining; however, one must acknowledge its constraints when implementing it in practical situations. The limitations and weaknesses of the BIRCH algorithm will be discussed in the following sections [1].

1. BIRCH is not appropriate for datasets with a high number of dimensions because it employs a distance measure that is greatly impacted by the curse of dimensionality.
2. BIRCH is not capable of effectively managing data that contains outliers or noise, as it lacks any means to address or filter out such anomalies within the data.

3. BIRCH is inadequate for dealing with clusters that are not globular in shape due to its reliance on a spherical clustering criterion.
4. BIRCH is not appropriate for managing extensive datasets due to its high memory requirements for storing the CF tree and clustering features.

The BIRCH algorithm uses 3 important parameters: branching factor, number of clusters, and threshold. While the data points of a 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 superclusters. The branching factor Br is the maximum number of children a node can have. Then, when a leaf is reached, a 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 to a newly created cluster as its only member. As a result, the size of the clusters is obviously controlled by the threshold parameter T.

Regression Deterministic

Deterministic regression is a method of regressing missing values with exact predictions from the regression model. The process carried out in the first deterministic regression process is to do simple random imputation in the missing data first before this empty data is used in the regression model, after doing simple random imputation the next step is to enter data that has been processed by simple random imputation into the regression model. deterministic. What is done in a deterministic regression model is to combine the missing data and the predicted data with the right value without considering the random variance around the regression line.

Homogeneity, Completeness, and 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. The V-measure is the harmonic mean between homogeneity and completeness [10]. Homogeneity is homogeneous clustering 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, this study will use the sklearn library to calculate homogeneity, completeness, and V-Measure.

$$v = 2 \cdot \frac{h \cdot c}{h + c} \tag{1}$$

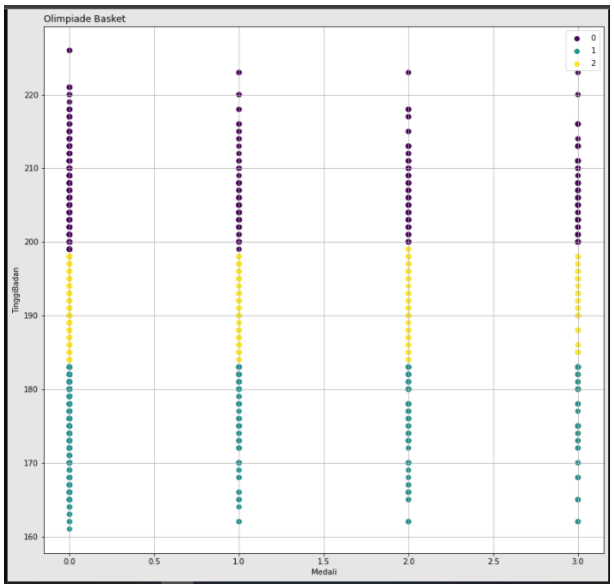
In the documentation, sklearn explains how to calculate the V-Measure using homogeneity and completeness and calculates the average using the formula as described in Figure 1. The formula above is a mathematical way to find the V-Measure result using numbers that have been found in inhomogeneity and completeness. In Figure 1 here *v* is V-Measure, 2 is values of beta, *h* is homogeneity and *c* is completeness.

IMPLEMENTATION AND TESTING

This study will analyze the relationship between athletes' height in the Olympics and medals. by using the BIRCH clustering algorithm, to calculate the performance of a given cluster from the BIRCH clustering algorithm, it will be seen from the results given by the v-measure, if the v-measure score is closer to 1 then the resulting cluster is better.

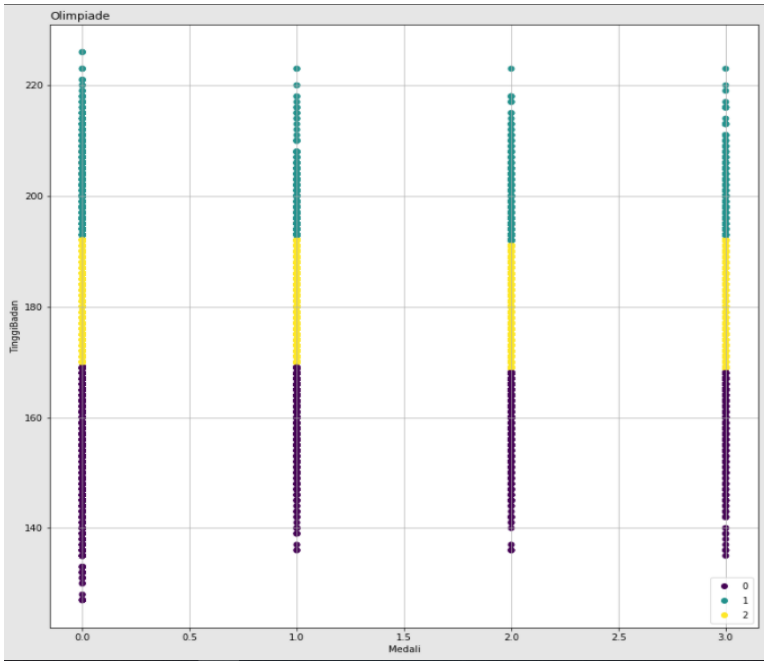
Testing

For the results of the basketball olympiad cluster, it can be seen in the picture above, that the cluster is divided into 3, and from the results it can be seen that basketball athletes with a height above 200cm can also not win medals at the Olympics, and other results show that basketball athletes with a height of 180cm and below also can win a gold medal. and other athletes with a height of 200cm and over or athletes with a height of 180cm and below also won bronze and silver medals.



Cluster Basketball

From the image of the cluster results above, it can be concluded that the results of the soccer, volleyball, and beach volleyball clusters can be drawn. It is divided into 3 clusters and shows that those with a height of 193cm and above and 172cm and below are more likely to not win a medal at all, and in terms of winning medals, they win more silver medals compared to gold and bronze when viewed from the results of the Olympic sports cluster. soccer, volleyball and beach volleyball.



Cluster All Sport Olympic

From the picture of the results of the cluster above, it can be concluded that the results of the cluster of all sports at the Olympics can be concluded. By looking at the results of the clusters, many clusters did not win medals for athletes with a height of 193 and above and also 169cm and below. but the results are different for athletes with height 193 and above and also 169cm and under who won fewer medals than those who did not win any medals at all. for the comparison of athletes who won gold medals only, more in height 169cm and below.

Evaluation

In this evaluation discussion, we will discuss the results of the homogeneity, completeness, and V-Measure calculations from the basketball cluster, Beach Volleyball Football cluster, and all sports clusters in the Olympics, to see whether the cluster performance produced by the BIRCH algorithm is good or not. By looking at the results of the V-Measure. The results of homogeneity and completeness are used to calculate the V-Measure. The closer the V-Measure value is to the number 1, the better the resulting cluster.

The table below describes the results of the homogeneity, completeness, and measure of the basketball cluster. It shows that the completeness result is 1, and the homogeneity is

0.28, the completeness is 0.99, and the result of V-Measure is 0.44. The 0.44 V-Measure result means that the basketball cluster's performance results are not good.

Tabel 1. Result Evaluation Basketball

Homogeneity	0.28341135525242406
Completeness	0.997462750812483
V-Measure	0.4414052383181702

After seeing the results of the basketball cluster, then looking at the results of the Football, Volleyball, and Beach Volleyball clusters The table below describes the results of the homogeneity, completeness, and measure of the Football, Volleyball, and Beach Volleyball cluster. It shows that the result is, homogeneity is 0.24, completeness 0.99, and the result of V-Measure is 0.39. The 0.39 V-Measure result means that the Football, Volleyball, and Beach Volleyball cluster's performance results are not good.

Tabel 2. Result Evaluation Football,Volleyball, and Beach Volleyball

Homogeneity	0.24744240114897698
Completeness	0.9999999999999997
V-Measure	0.39671956143396464

The table below describes the results of the homogeneity, completeness and cluster size of the overall sports in the Olympics. This shows that the homogeneity result is 0.20, the completeness is 0.99, and the V-Measure result is 0.34. The measurement result of 0.34 means that the performance results of the Football, Volleyball, and Beach Volleyball clusters are not good. From the results of all clusters that have been calculated using homogeneity, completeness and V-Measure, the V-Measure is still very far from approaching number 1 so the performance results of the BIRCH algorithm are still not good enough.

Tabel 3. Result All Sport Olimpic

Homogeneity	0.2080736632245153
Completeness	0.9900922803821292
V-Measure	0.34387912427104816

CONCLUSION

The research that has been carried out, resulting in conclusions that can be drawn from the results of this study are: the results of the cluster show that the athlete's height in the Olympics has nothing to do with medals, the results given by the BIRCH algorithm show that athletes with a height above 200cm in ball sports basketball did not get a medal but with a height below 180 cm there were those who got a gold medal in the sport of basketball. Not

only basketball but also soccer, volleyball and beach volleyball and the overall results from the given clusters also show that height has nothing to do with medals, and height is not a major factor in winning an athlete. There are many other factors that affect an athlete's victory in the Olympics, namely the athlete's health, age, and how often the athlete trains to practice his skills. From the results of the V-Measure evaluation to calculate cluster performance, the cluster results for the entire basketball, soccer, volleyball, and beach volleyball clusters are still far from approaching number 1 to improve cluster performance, we must rearrange three important parameters for the BIRCH algorithm, namely breaching factor, number of clusters, and threshold. For further research, we can find the best option for the three important parameters of the BIRCH algorithm.

REFERENCES

[1] N. Garg dan R. K. Gupta, "Exploration of Various Clustering Algorithms for Text Mining," *Int. J. Educ. Manag. Eng.*, vol. 8, no. 4, hlm. 10–18, Jul 2018, doi: 10.5815/ijeme.2018.04.02.

[2] L. Kovacs dan L. Bednarik, "Parameter optimization for BIRCH pre-clustering algorithm," dalam *2011 IEEE 12th International Symposium on Computational Intelligence and Informatics (CINTI)*, Budapest, Hungary: IEEE, Nov 2011, hlm. 475–480. doi: 10.1109/CINTI.2011.6108553.

[3] F. Ramadhani, M. Zarlis, dan S. Suwilo, "Improve BIRCH algorithm for big data clustering," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 725, no. 1, hlm. 012090, Jan 2020, doi: 10.1088/1757-899X/725/1/012090.

[4] B. Lorbeer, A. Kosareva, B. Deva, D. Softić, P. Ruppel, dan A. Küpper, "Variations on the Clustering Algorithm BIRCH," *Big Data Res.*, vol. 11, hlm. 44–53, Mar 2018, doi: 10.1016/j.bdr.2017.09.002.

[5] S. Guha, R. Rastogi, dan K. Shim, "Rock: A robust clustering algorithm for categorical attributes," *Inf. Syst.*, vol. 25, no. 5, hlm. 345–366, Jul 2000, doi: 10.1016/S0306-4379(00)00022-3.

[6] D. Guo, J. Chen, Y. Chen, dan Z. Li, "LBIRCH: An Improved BIRCH Algorithm Based on Link," dalam *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, Macau China: ACM, Feb 2018, hlm. 74–78. doi: 10.1145/3195106.3195158.

[7] S. Madan dan K. J. Dana, "Modified balanced iterative reducing and clustering using hierarchies (m-BIRCH) for visual clustering," *Pattern Anal. Appl.*, vol. 19, no. 4, hlm. 1023–1040, Nov 2016, doi: 10.1007/s10044-015-0472-4.

[8] W. H. E. Day dan H. Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods," *J. Classif.*, vol. 1, no. 1, hlm. 7–24, 2014, doi: 10.1007/BF01890115.

[9] N. Ismael, M. Alzaalan, dan W. Ashour, "Improved Multi Threshold Birch Clustering Algorithm," *Int. J. Artif. Intell. Appl. Smart Devices*, vol. 2, hlm. 1–10, Mei 2014, doi: 10.14257/ijaiasd.2014.2.1.01.

[10] K. Aggarwal *dkk.*, "Robust Assessment of Clustering Methods for Fast Radio Transient Candidates," *Astrophys. J.*, vol. 914, no. 1, hlm. 53, Jun 2021, doi: 10.3847/1538-4357/abf92b.