Comparison Between K-Means and K-Medoids Clustering Model in Clustering the Stage of Basketball Players

Clustering Algorithm

Maria Darlene Kusnadi[1]
[1] Information System Study Program, Faculty of Engineering and Informatics, Multimedia Nusantara University, South Tangerang, Indonesia

Accepted on 17 December 2021
Approved on 17 December 2021

Abstract—**This report contains the research of basketball players stage clustering based on their stats using clustering algorithm. Clustering algorithm is an unsupervised classification algorithm. Clustering is a set of techniques for partitioning a set of data points into groups that contains very similar data points to find subgroups or clusters in a dataset [1]. The purpose of this research is to form a model that will cluster the correct stage of a basketball player based on the supporting facts and stats data using K-Means and K-Medoids algorithms. Both algorithms would then be compared to decide which model is better suited for this dataset. Hopefully, this analysis result could become a reference in clustering the qualification of a basketball players to better group players according to which stage they should play on.**

Index Terms— K-Means; K-Medoids; Machine Learning; Clustering; Basketball Player; Stage **(;)**

I. Introduction

As the era progresses, technology develops in a tremendously quick rate, and becomes more and more integrated into the everyday life. One of the technology field that enjoys this rapid development is the information technology. Through the development of information technology, information can now flow freely, spread vastly, and instantaneously throughout the world. Data has become a precious resource [2]. Due to it's relevancy in the modern era, the knowledge, science, and technique to gather and process data into useful information is also experiencing a rapid development. One of such developing field is the Big Data.

Big Data is a term used for a massive and complex collection of data that is hard to parse or process using traditional or simple processing techniques. Recently, it isn't unfamiliar to hear this term used in conjunction to the usage and processing of the data itself in Information Technology fields. The defining characteristics of big data could be concluded in 4V, which are volume, velocity, variety, and veracity. Volume is the size of the data that needs to be processed. As the name suggest, Big Data processes a very large amount of data, it could even reach petabytes of data each day. Velocity refers to the speed of the produced data. The data processed by big data are often of high speed, therefore needing a specialized form of data processing. Variety refers to how Big Data is sourced from many sources, and contain different types of data, which are structured data, semi-structured data, and unstructured data. Finally, veracity refers to the accuracy or the consistency of the data. A highly accurate data would produce good quality analysis, in contrast, data with low accuracy would also produce low quality analysis, therefore data processing is highly important in Big Data usage[3].

Adequate data, information, and knowledge has become a necessity to survive or to get ahead in various fields, even in fields that don’t seem connected to information technology at first glance. One such field is the basketball sport. Basketball is a very popular sport that is played around the world. The format used in a basketball game is usually 5-on-5, however there are also alternative formats that are sometimes used when playing this game, such as 3-on-3 and 1-on-1. This game was founded by Dr.James Naismith, a physical education instructor in the YMCA International Training School (now Springfield College), in 1891[4].

In professional basketball, there are three stages where basketball player could compete with each other. The first one is Regular Season, which is a regularly scheduled competition between NBA Teams, beginning on the first day and ending on the last day of the season. The regular season is the most casual out of all the three stages. The second one is Playoff, which is the best-of-seven tournament, where the seven best teams from the regular season would compete against each other to gain a spot in the international tournament. The third stage is the International stage, which is the basketball tournament on the international scale. In the International stage, the teams who compete are representing each of their own country, therefore each country would pick the best of their players to represent them. The example of International stage is the FIFA Basketball World Cup where basketball teams from all over the world compete against each other [5].

The success of a basketball team relies on how well the players play for their team. Therefore, selecting good players is a vital process to ensure the team’s victory. In this case, one of the technique in Big Data processing, which is the clustering algorithm, could aide in creating clusters of players based on their stats. It could help in categorizing players according to their skills and stats, therefore increasing the chances to win in each stages.

II. Literature Review

A. *Clustering Algorithms*

Clustering could be interpreted as a term for a very broad set of techniques regarding the search of subgroups or clusters in a dataset. Clustering is an algorithm that would partition observations from a dataset into distinct groups of similar observations. The observations within a group tend to be similar to one another, however the observations in outside the group would tend to be quite different from one another. The clustering algorithms are unsupervised methods [6]. Clustering could be used for a lot of things, such as market research, pattern recognition, customer segmentation, develop animals and plants’ taxonomies, and many more. It is an important tool for data mining and analyzing big data. The reason for it could be found in the requirements of clustering in data mining. For data mining, clustering has to be scalable to deal with large databases, adaptive so it could be applied on any kind of data, able to detect arbitrary shape, have high dimensionality, able to deal with noisy data, interpretable, comprehensible, and usable [7].

B. *K-Means Algorithm*

K-Means Algorithm is one of the oldest and most popular clustering algorithm. In K-Means algorithm, for a chosen value of k, the algorithm will identify clusters of observations based on its’ proximity to the center of the k groups, or centroids. In this algorithm, the objective function of the clustering is quantified by the sum of squares of the Euclidean distance of data points to their closest representatives [8]. The distance could be summarized in this formula.

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \left\| x^{i} - \mu_{k} \right\|^{2}$$

Figure 1. 1. K-Means formula

The K-Means algorithm isn't fully unsupervised like other clustering algorithm because the user still needs to give the k, or the number of clusters for the algorithm to work [9]. The k for this algorithm is given by the user according to their need, so the k could be derived from the algorithm's application, however if there are no set clusters needed by the algorithm's application, there are alternate ways to decide on the amount of k.

C. *K-Medoids Algorithm*

K-Medoids is a clustering algorithm similar to k-means that is proposed by Kaufman and Rousseeuw in 1987. K-medoids group a set of objects into k number of clusters, where the clusters are represented by objects in a collection of objects. These representative objects are called the medoids. This algorithm is similar to k-means because both of them breaks the dataset into groups of k, and appoints objects or observations as the center of a cluster. However, different from k-means, k-medoids appoint actual data points as centers, therefore it could give more interpretability of the cluster. One of the most used algorithm in K-Medoids algorithm is the Partitioning Around Medoids (PAM) algorithm. The PAM algorithm is a greedy search algorithm where the solution it finds might not be the most optimal, however it works faster than an exhaustive search [10].

$$j = \sum_{i=1}^{k} \sum_{p\Omega c_j}^{n} \left\| P - O_j \right\|$$

Figure 1. 2 K-Medoids formula

D. *Deciding K*

There are several ways to decide on the amount of K for k-means and k-medoid algorithms if there are no set clusters needed by the algorithm's application. One of the common approach to decide on a k value is the elbow method. Elbow method identifies when the set of clusters explains the most of the variance in the data, where adding new clusters beyond the ones that elbow method suggests would only give little incremental contribution in the variance explained. The second approach is the silhouette method. This method measure how much a point is similar to its' cluster compared to other clusters by computing the silhouette coefficients of each point. The third approach is the gap statistics method. This method would compare the expected values under null reference distribution of the data to the total intracluster variation for different values of k [11].

E. *Benefit and Weakness of both algorithms*

K-Means is a well-known partitioning method for clustering. This algorithm works well in small to medium-sized data points and finding spherical-shaped clusters. Its' advantage lies in its' faster execution time compared to a lot of other algorithms. However, the main drawback of this algorithm is it's not a fully unsupervised algorithm, therefore it need an input of k from its' user [12].

K-Medoids is also a partitioning algorithm method for clustering. Compared to k-means, k-medoids is known to be more complex, more accurate, and have a shorter execution. This algorithm is also known to be less sensitive to outliers than k-means, therefore more robust in handling data with outliers. However, k-medoids also has drawbacks such as it is more costly than k-means and doesn't scale well for large datasets [13].

F. *Stratified Sampling*

Sampling is a technique to systematically select a relatively smaller number of representatives or a subset from a pre-defined population to serve as data source for observations or experimentations. Stratified sampling is a method of sampling that divides a population into smaller groups called strata. Stratified sampling is categorized into probability sampling, which is a sampling scheme where the probability to choose each individual is the same, or known so it could be adjusted mathematically. The strata made during the sampling are formed based on the members' shared attributes or characteristics. After the strata are formed, a random sample of each stratum would then be taken in proportion to each stratum's size compared to the population. The sample taken would then be pooled together to form a random sample. This type of sampling is often used if the population size of each category is too unbalanced to do simple random sampling [14].

G. *External validation*

Clustering validation is needed to evaluate the goodness of clustering algorithm results. The validation technique is needed to prevent users in finding patterns in a random dataset. This validation is also needed when comparing two clustering algorithms. The external validation method is one of clustering validation method. The external validation would compare results of a clustering algorithm to an externally known results, such as predetermined or externally provided class labels. In other words, this validation technique will attempt to find how much the result from the clustering algorithm matches the "true" labels. Therefore, it is mainly used to decide which clustering algorithm is better for a specific dataset [15].

III. METHODOLOGY

A. *Research Object*

The research object for this research is the ability of Basketball Players from 49 leagues around the world. This data used in the research are data related to the players' details and performance in each Leagues. It contains data about players' height, weight, scoring stats, minutes played, rebounds, blocks, the stage they play at, et cetera. The target of this research is to decide which clustering is better in grouping the players according to their abilities, therefore helping in determining which stage is the best for each player to play at.

B. *Data Gathering*

The dataset used in this research is the Basketball Player Stats dataset. The dataset was taken from a data science learning community website called Kaggle. In total, this dataset contains 53.949 observations with 34 variables that includes both categorical and numerical variables. The data used in this research are quantitative data from the dataset, of which there are 20 variables, including the target variable, Stage. The other 19 variables are Games Played (GP), Minutes Played (MIN), Field Goals Made (FGM), Field Goals Attempt (FGA), 3 Points Made (X3PM), 3 Points Attempt (X3PA), Free Throws Made (FTM), Free Throws Attempt (FTA), Turnovers (TOV), Personal Fouls (PF), Offensive Rebounds (ORB), Defensive Rebounds (DRB), Rebounds (REB), Assist (AST), Steals (STL), Block (BLK), Points (PTS), players' height in cm (height_cm), and players' weight in kg (weight_kg).

C. *Framework*

These are the steps done to conduct this research analysis.

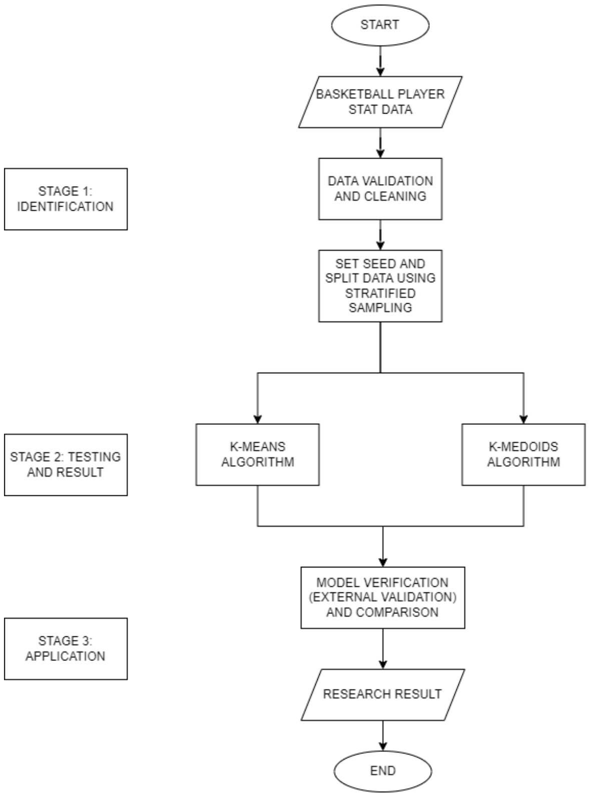

START
BASKETBALL PLAYER STAT DATA
STAGE 1: IDENTIFICATION
DATA VALIDATION AND CLEANING
SET SEED AND SPLIT DATA USING STRATIFIED SAMPLING
STAGE 2: TESTING AND RESULT
K-MEANS ALGORITHM
K-MEDOIDS ALGORITHM
MODEL VERIFICATION (EXTERNAL VALIDATION) AND COMPARISON
STAGE 3: APPLICATION
RESEARCH RESULT
END

Figure 3. 1 Framework work-flow

D. *Identification*

Identification is the first stage. In this stage the data, which is in the form of comma-separated values (.csv) would be inputted into R before it could be analyzed using the same program. In preparation of processing the data, it is also necessary to install or call all the packages needed. After the data and the packages has been imported, then it is time to decide on the target variable. In this research case, the target variable is the Stage variable. The Stage variable is a categorical variable, that has 3 levels, “Regular Season”, “Play-offs”, and “International”.

The data used in this research are qualitative data. Therefore, to ensure the algorithm could function well, the dataset deeds to be filtered and cleansed. It is necessary to check whether the data has missing values or not. This process could be done using the missmap function in R Studio. The missing data was then omitted, making the formerly 53.949 rows of observations into 36.292 rows.

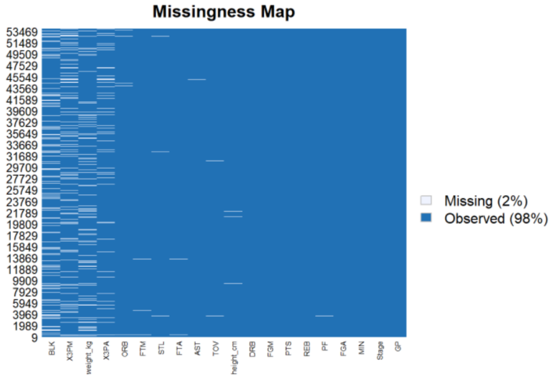


Figure 3. 2 Missingness Map before data cleansing

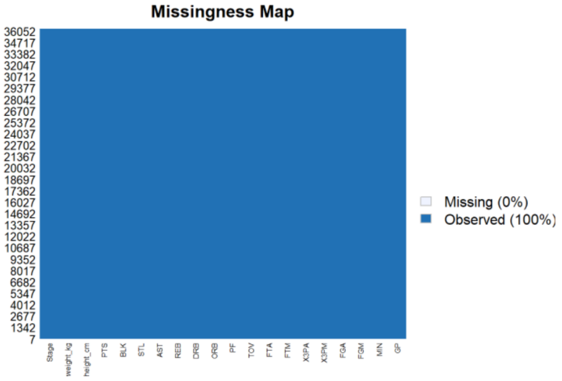
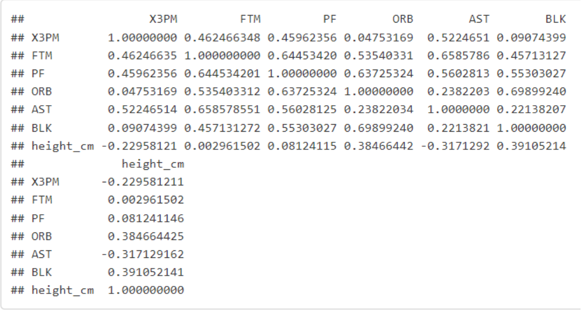


Figure 3. 3 Missingness Map after data cleansing

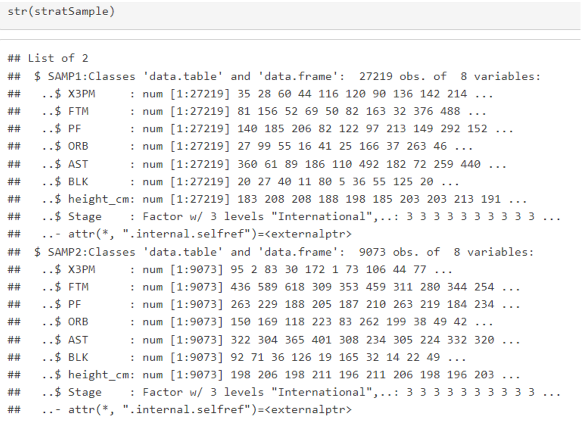
The correlation of the data was then checked to ensure that it is suitable for clustering algorithm. Independent variables with high correlation are omitted to prevent multicollinearity, making the formerly 20 variables into 8 variables.



```
##                 X3PM         FTM         PF        ORB        AST        BLK
## X3PM       1.00000000 0.462466348 0.45962356 0.04753169  0.5224651 0.09074399
## FTM        0.46246635 1.000000000 0.64453420 0.53540331  0.6585786 0.45713127
## PF         0.45962356 0.644534201 1.00000000 0.63725324  0.5602813 0.55303027
## ORB        0.04753169 0.535403312 0.63725324 1.00000000  0.2382203 0.69899240
## AST        0.52246514 0.658578551 0.56028125 0.23822034  1.0000000 0.22138207
## BLK        0.09074399 0.457131272 0.55303027 0.69899240  0.2213821 1.00000000
## height_cm -0.22958121 0.002961502 0.08124115 0.38466442 -0.3171292 0.39105214
##              height_cm
## X3PM      -0.229581211
## FTM        0.002961502
## PF         0.081241146
## ORB        0.384664425
## AST       -0.317129162
## BLK        0.391052141
## height_cm  1.000000000
```

Figure 3. 4 The multicollinearity test of 8 variables

After cleaning the data, descriptive statistics were applied to see the distributions of the data. Unfortunately, it was found that the distributions are unbalanced, with one of the Stage is significantly larger in population compared to the other Stage. The data was then split into training and testing data. The split was done using stratified sampling to provide a more accurate sampling due to the imbalance in data. The training data contains 75% of the whole data, which contains 27.219 rows of data, and the testing data contains 25% of the whole data, which contains 9.073 rows of data. The data is split as a preparation to make clustering algorithm models. This step concludes the identification stage of the analysis process.



```
str(stratSample)
```

```
## List of 2
##  $ SAMP1:Classes 'data.table' and 'data.frame':  27219 obs. of  8 variables:
##   ..$ X3PM     : num [1:27219] 35 28 60 44 116 120 90 136 142 214 ...
##   ..$ FTM      : num [1:27219] 81 156 52 69 50 82 163 32 376 488 ...
##   ..$ PF       : num [1:27219] 140 185 206 82 122 97 213 149 292 152 ...
##   ..$ ORB      : num [1:27219] 27 99 55 16 41 25 166 37 263 46 ...
##   ..$ AST      : num [1:27219] 360 61 89 186 110 492 182 72 259 440 ...
##   ..$ BLK      : num [1:27219] 20 27 40 11 80 5 36 55 125 20 ...
##   ..$ height_cm: num [1:27219] 183 208 208 188 198 185 203 203 213 191 ...
##   ..$ Stage    : Factor w/ 3 levels "International",..: 3 3 3 3 3 3 3 3 3 3 ...
##   ..- attr(*, ".internal.selfref")=<externalptr>
##  $ SAMP2:Classes 'data.table' and 'data.frame':  9073 obs. of  8 variables:
##   ..$ X3PM     : num [1:9073] 95 2 83 30 172 1 73 106 44 77 ...
##   ..$ FTM      : num [1:9073] 436 589 618 309 353 459 311 280 344 254 ...
##   ..$ PF       : num [1:9073] 263 229 188 205 187 210 263 219 184 234 ...
##   ..$ ORB      : num [1:9073] 150 169 118 223 83 262 199 38 49 42 ...
##   ..$ AST      : num [1:9073] 322 304 365 401 308 234 305 224 332 320 ...
##   ..$ BLK      : num [1:9073] 92 71 36 126 19 165 32 14 22 49 ...
##   ..$ height_cm: num [1:9073] 198 206 198 211 196 211 206 198 196 203 ...
##   ..$ Stage    : Factor w/ 3 levels "International",..: 3 3 3 3 3 3 3 3 3 3 ...
##   ..- attr(*, ".internal.selfref")=<externalptr>
```

Figure 3. 5 Stratified sampling

E. *Testing and result*

The second stage is the testing and result stage. In this stage, the clustering models using K-Means and K-Medoids are made. The models are made using testing data. The models are made by clustering the cleansed and prepared data, using 7 numeric variables. K-Means algorithm will group observations based on its' proximity to the centroids by calculating the sum of sqares of the Euclidean distance of data points to their closest representatives, creating clusters accoriding to the calculation. Whereas K-Medoids would cluster data in a similar way, but it will appoint medoids, an actual data points from the dataset. The clustering algorithm would be applied to the original dataset and the scaled dataset, as scalability is an important factor in deciding the quality of clustering algorithm. The cluster result would then be visualized in graphs.

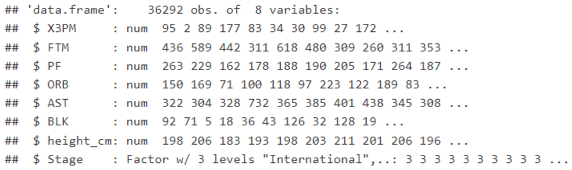
F. *Application*

The final stage is the application stage. In the application stage, the two models that were made in prior stage would be validated using the external validation technique. If the accuracy exceed 50%, then the model is usable because it means that using the model is better than wild guessing. However, if the model's accuracy is below 50%, than it is not usable. After validating the two models, both models would be compared to each other. The comparison could be seen from each model's external validation.

IV. Results and Discussion

A. *Identification*

The data used in this research is a secondary data retrieved from the data science learning community website, Kaggle. It is a Basketball Player Stat Dataset from 49 Leagues around the world, with 36.292 rows of data and 8 variables. Below is the data used in this research after filtering and cleansing.



```
## 'data.frame':    36292 obs. of  8 variables:
##  $ X3PM     : num  95 2 89 177 83 34 30 99 27 172 ...
##  $ FTM      : num  436 589 442 311 618 480 309 260 311 353 ...
##  $ PF       : num  263 229 162 178 188 190 205 171 264 187 ...
##  $ ORB      : num  150 169 71 100 118 97 223 122 189 83 ...
##  $ AST      : num  322 304 328 732 365 385 401 438 345 308 ...
##  $ BLK      : num  92 71 5 18 36 43 126 32 128 19 ...
##  $ height_cm: num  198 206 183 193 198 203 211 201 206 196 ...
##  $ Stage    : Factor w/ 3 levels "International",..: 3 3 3 3 3 3 3 3 3 3 ...
```

Figure 4. 1 data structure

After the data was filtered and cleansed, descriptive statistics was applied to see the distribution of the data. Below are the barplot and boxplots of the filtered and cleansed Basketball Player Stat Dataset.

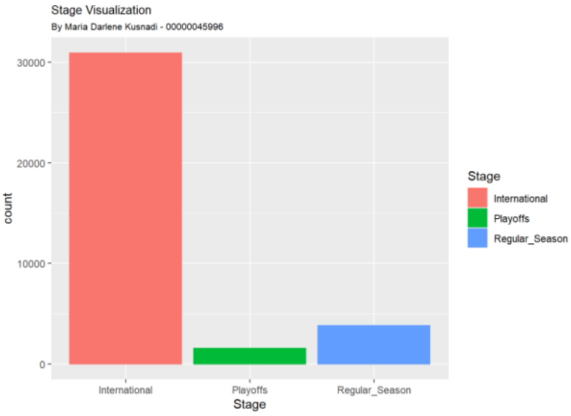


Figure 4. 2 International, Playoffs, and Regular_Season Stage comparison

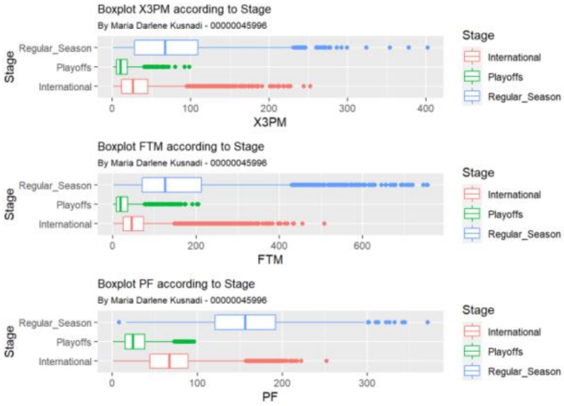


Figure 4. 3 X3PM, FTM, and PF boxplots according to Stage

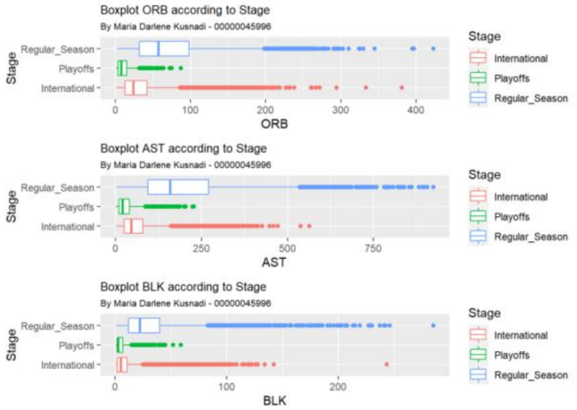


Figure 4. 4 ORB, AST, and BLK boxplots according to Stage

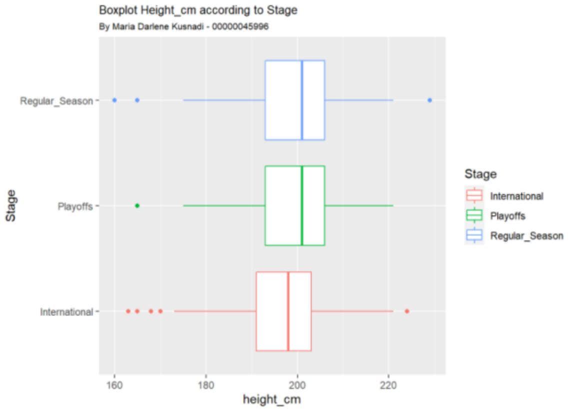
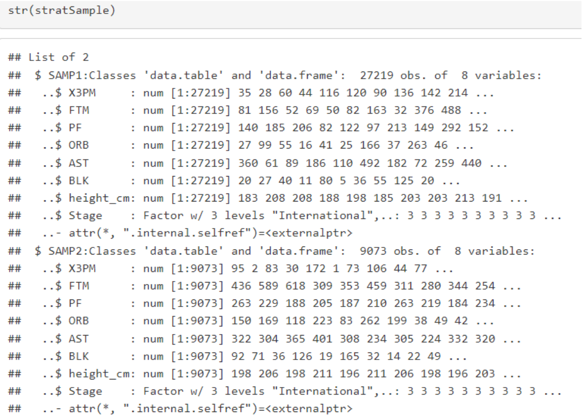


Figure 4. 5 Height_cm boxplot according to Stage

Because of the unbalanced data, the data was split using stratified sampling, to ensure a more accurate sample of data from the three stages.



```
str(stratSample)
```

```
## List of 2
##  $ SAMP1:Classes 'data.table' and 'data.frame':  27219 obs. of  8 variables:
##   ..$ X3PM     : num [1:27219] 35 28 60 44 116 120 90 136 142 214 ...
##   ..$ FTM      : num [1:27219] 81 156 52 69 50 82 163 32 376 488 ...
##   ..$ PF       : num [1:27219] 140 185 206 82 122 97 213 149 292 152 ...
##   ..$ ORB      : num [1:27219] 27 99 55 16 41 25 166 37 263 46 ...
##   ..$ AST      : num [1:27219] 360 61 89 186 110 492 182 72 259 440 ...
##   ..$ BLK      : num [1:27219] 20 27 40 11 80 5 36 55 125 20 ...
##   ..$ height_cm: num [1:27219] 183 208 208 188 198 185 203 203 213 191 ...
##   ..$ Stage    : Factor w/ 3 levels "International",..: 3 3 3 3 3 3 3 3 3 3 ...
##   ..- attr(*, ".internal.selfref")=<externalptr>
##  $ SAMP2:Classes 'data.table' and 'data.frame':  9073 obs. of  8 variables:
##   ..$ X3PM     : num [1:9073] 95 2 83 30 172 1 73 106 44 77 ...
##   ..$ FTM      : num [1:9073] 436 589 618 309 353 459 311 280 344 254 ...
##   ..$ PF       : num [1:9073] 263 229 188 205 187 210 263 219 184 234 ...
##   ..$ ORB      : num [1:9073] 150 169 118 223 83 262 199 38 49 42 ...
##   ..$ AST      : num [1:9073] 322 304 365 401 308 234 305 224 332 320 ...
##   ..$ BLK      : num [1:9073] 92 71 36 126 19 165 32 14 22 49 ...
##   ..$ height_cm: num [1:9073] 198 206 198 211 196 211 206 198 196 203 ...
##   ..$ Stage    : Factor w/ 3 levels "International",..: 3 3 3 3 3 3 3 3 3 3 ...
##   ..- attr(*, ".internal.selfref")=<externalptr>
```

Figure 4. 6 Stratified sampling data structure

B. *Testing and Result*

After the identification stage, comes the testing and result stage. The clustering algorithm models made in this stage are K-Means algorithm and K-Medoids algorithm. The algorithm models are made using the testing data, which has 9.073 rows of data and 7 numerical variables.

K-Means algorithm will identify clusters of observations based on its' proximity to the center of the k groups, or centroids. In this algorithm, the objective function of the clustering is quantified by the sum of squares of the Euclidean distance of data points to their closest representatives. As stated before, to make a k-means algorithm model, the user needs to input the amount of k. The amount of k is primarily decided on the user's need and the application of the algorithm, or by using methods such as elbow method, silhouette method, and gap statistics method. In this case, the algorithm is used to cluster the stages that the basketball players should play at, therefore the k that is used in creating this model has the value of 3. Below is the visualization of the clusters made by the K-Means Clustering.

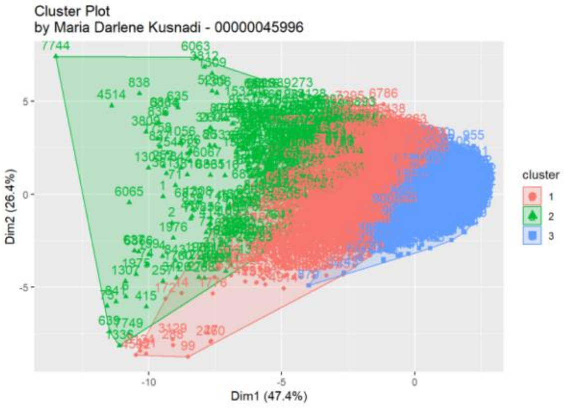


Figure 4. 7 K-Means Clustering without scaled data

Scalability is an important aspect in clustering algorithm, because it means that the clustering could be used in a large database. Therefore, the user should also perform scaling on the data and do another clustering algorithm model on the scaled data. The scaling of the data is done by using a function called scale() in RStudio. The scale() function will scale the columns of a numeric matrix. Below is the visualization of the clusters made by the scaled data.

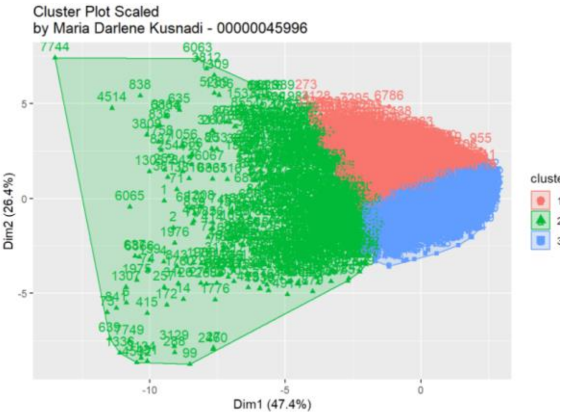
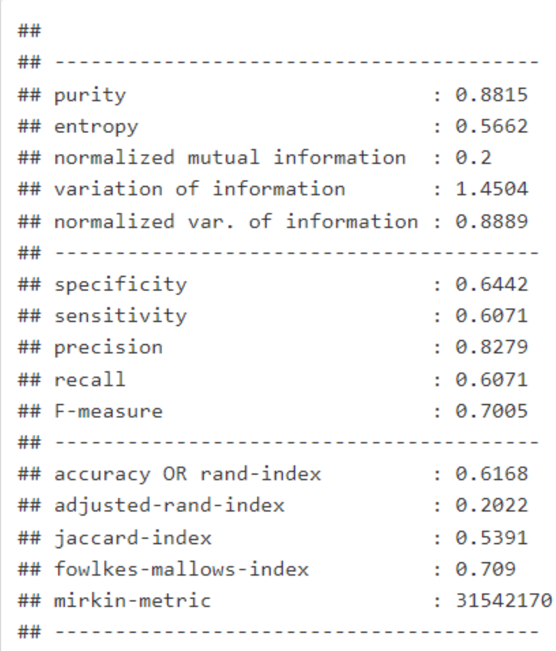


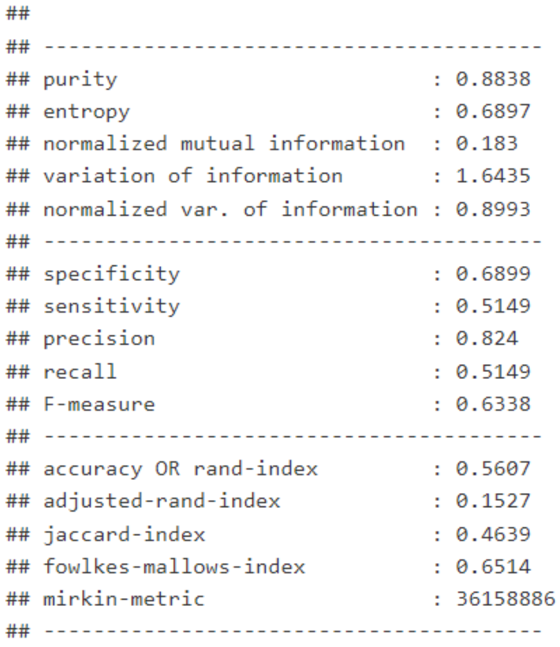
Figure 4. 8 K-Means Clustering with scaled data

After making the clustering algorithm model, both the unscaled k-means and the scaled k-means models should be validated to prove their accuracy and usability. In this case, the validation method was the external validation technique. To do it, both of the clustering results are matched with an external, original data, which is the Stage variable in the dataset. Below are the results of the external validation. The external



```
##
## ---------------------------------------
## purity                              : 0.8815
## entropy                             : 0.5662
## normalized mutual information       : 0.2
## variation of information            : 1.4504
## normalized var. of information      : 0.8889
## ---------------------------------------
## specificity                         : 0.6442
## sensitivity                         : 0.6071
## precision                           : 0.8279
## recall                              : 0.6071
## F-measure                           : 0.7005
## ---------------------------------------
## accuracy OR rand-index              : 0.6168
## adjusted-rand-index                 : 0.2022
## jaccard-index                       : 0.5391
## fowlkes-mallows-index               : 0.709
## mirkin-metric                       : 31542170
## ---------------------------------------
```

Figure 4. 9 External validation K-Means Clustering without scaled data



```
##
## ---------------------------------------
## purity                              : 0.8838
## entropy                             : 0.6897
## normalized mutual information       : 0.183
## variation of information            : 1.6435
## normalized var. of information      : 0.8993
## ---------------------------------------
## specificity                         : 0.6899
## sensitivity                         : 0.5149
## precision                           : 0.824
## recall                              : 0.5149
## F-measure                           : 0.6338
## ---------------------------------------
## accuracy OR rand-index              : 0.5607
## adjusted-rand-index                 : 0.1527
## jaccard-index                       : 0.4639
## fowlkes-mallows-index               : 0.6514
## mirkin-metric                       : 36158886
## ---------------------------------------
```

Figure 4. 10 External validation K-Means clustering with scaled data

The external validation of the unscaled k-means algorithm result shows the accuracy of 61,68%, specificity of 64,42%, sensitivity of 60,71% , and precision of 82,79%. The external validation of the scaled k-means algorithm result shows the accuracy of 56,07%, specificity of 68,99%, sensitivity of 51,49%, and precision of 82,4%. The scaled version shows a lower accuracy than the unscaled version. This means that this model is not very scalable, however the model is usable because both of the unscaled and scaled data accuracy is more than 50%, making it a better solution than pure wild guesses.

K Medoids will group a set of objects into k number of clusters, where the clusters are represented by objects in a collection of objects. Similar to K-Means, the k in K-Medoids is also mainly decided based on user's need and the application of the algorithm. Therefore, in this case, the k that is used in creating this model has the value of 3.

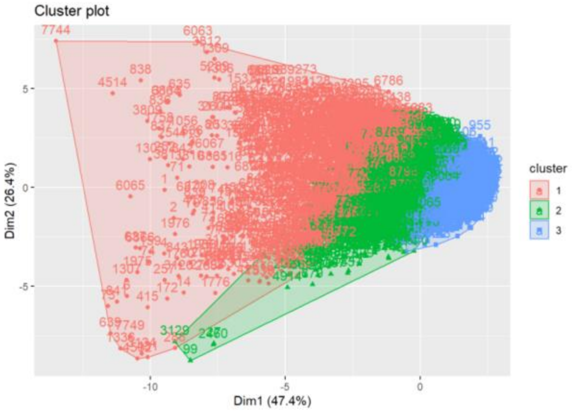


Figure 4. 11 K-Medoids Clustering without scaled data

K-Medoids also requires scalability. Therefore, the user should also perform scaling on the data and do another clustering algorithm model on the scaled data. The data for K-Medoids algorithm was also scaled by using the function scale(). Below is the visualization of the clusters made by the scaled data.

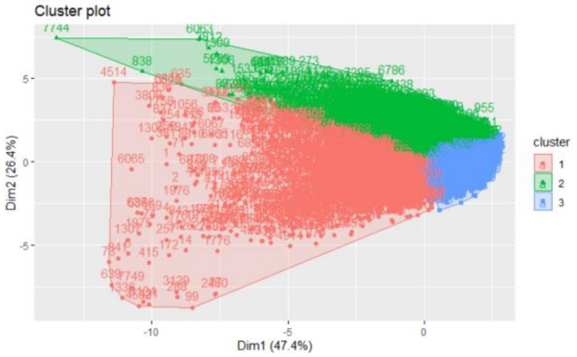
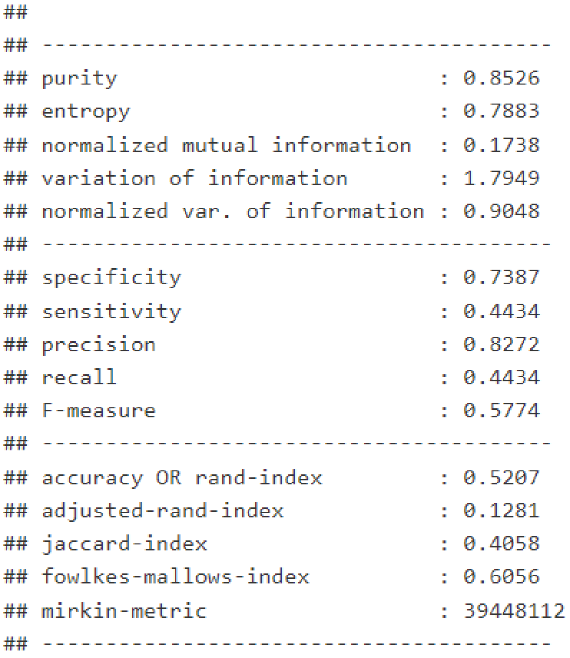


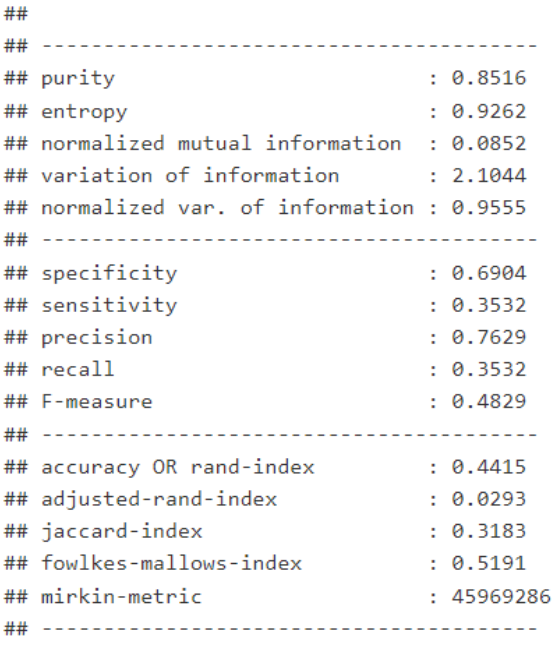
Figure 4. 12 K-Medoids Clustering with scaled data

After making the clustering algorithm model, both the unscaled K-Medoids and the scaled K-Medoids models were validated with the external validation technique. Both of the clustering results are matched with an external, original data, which is the Stage variable in the dataset. Below are the results of the external validation.



```
##
## ----------------------------------------
## purity                             : 0.8526
## entropy                            : 0.7883
## normalized mutual information      : 0.1738
## variation of information           : 1.7949
## normalized var. of information     : 0.9048
## ----------------------------------------
## specificity                        : 0.7387
## sensitivity                        : 0.4434
## precision                          : 0.8272
## recall                             : 0.4434
## F-measure                          : 0.5774
## ----------------------------------------
## accuracy OR rand-index             : 0.5207
## adjusted-rand-index                : 0.1281
## jaccard-index                      : 0.4058
## fowlkes-mallows-index              : 0.6056
## mirkin-metric                      : 39448112
## ----------------------------------------
```

Figure 4. 13 External validation K-Medoids Clustering without scaled data



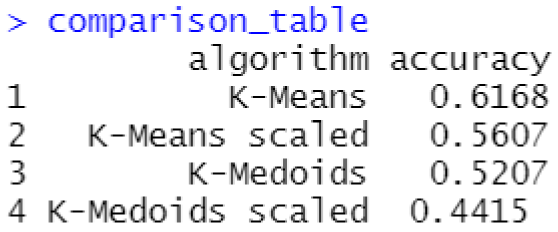
```
##
## ----------------------------------------
## purity                            : 0.8516
## entropy                           : 0.9262
## normalized mutual information     : 0.0852
## variation of information          : 2.1044
## normalized var. of information    : 0.9555
## ----------------------------------------
## specificity                       : 0.6904
## sensitivity                       : 0.3532
## precision                         : 0.7629
## recall                            : 0.3532
## F-measure                         : 0.4829
## ----------------------------------------
## accuracy OR rand-index            : 0.4415
## adjusted-rand-index               : 0.0293
## jaccard-index                     : 0.3183
## fowlkes-mallows-index             : 0.5191
## mirkin-metric                     : 45969286
## ----------------------------------------
```

Figure 4. 14 External validation K-Medoids Clustering with scaled data

The external validation of the unscaled K-Medoids algorithm result shows the accuracy of 52,07%, specificity of 73,87%, sensitivity of 44,34% , and precision of 82,72%. The external validation of the scaled k-means algorithm result shows the accuracy of 44,15%, specificity of 69,04%, sensitivity of 35,32%, and precision of 76,29%. The scaled version shows a lower accuracy than the unscaled version. This means that this model is not very scalable. The model is also unsuitable because the scaled version has an accuracy of below 50%, making it worse than wild guesses.

C. *Application*

In the final stage, the application stage, the two algorithm models, K-Means and K-Medoids, are compared. The two models are created using the same dataset, and tested with the same data. However, they are also created using different methods and approach, therefore producing different results.



```
> comparison_table
         algorithm accuracy
1          K-Means   0.6168
2   K-Means scaled   0.5607
3        K-Medoids   0.5207
4 K-Medoids scaled  0.4415
```

Figure 4. 15 Comparison table

The results show that K-Means algorithm has a better accuracy with 61,68% accuracy for the unscaled data and 56,07% accuracy for the scaled data, compared to the K-Medoids accuracy with 52,07% accuracy for the unscaled data and 44,15% for the scaled data. Therefore, K-Means algorithm is a better choice for clustering basketball player stat data based on their Stages.

V. Conclusion

The two classification algorithm models produce different results based on their approach in each methods. K-Means is an algorithm that will identify clusters of observations based on its' proximity to the centroids by calculating the sum of squares of the Euclidean distance of data points to their representatives. K-Medoids, similar to K-Means, is an algorithm that partitions dataset into clusters by calculating observations' proximity to the medoids. To group the basketball player stat dataset into correct groups that represent the players' ability, the correct clustering model is needed. In this research, it is found that in determining the Stages of the players in the Basketball player stat dataset, K-Means algorithm is the more suitable choice than the K-Medoids algorithm in doing the clustering. Therefore, it would be better to use the K-Means algorithm when dealing with similar dataset and target variable in the future.

References

[1] Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. D. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PloS one*, *14*(1). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0210236

[2] Kasemin, H. K. (2016). *Agresi Perkembangan Teknologi Informasi*. Prenada Media. https://books.google.co.id/books?hl=en&lr=&id=R_ouDwAAQBAJ&oi=fnd&pg=

[3] PA1&dq=perkembangan+teknologi&ots=xD53wRnKNh&sig=r3Odq_BPHHdjgFDJKTx915SeoZ4&redir_esc=y#v=onepage&q=perkembangan%20teknologi&f=false

[4] Supriyanto, E., Bakti, I., & Furqon, M. (2021). The Role of Big Data in The Implementation of Distance Learning. *Paedagoria : Jurnal Kajian, Penelitian dan Pengembangan Kependidikan*, *12*(1), 61-69. http://journal.ummat.ac.id/index.php/paedagoria/article/view/3902/pdf

[5] Oliver, Jon A. (2004). Basketball fundamentals. Human Kinetics. https://books.google.co.id/books?hl=en&lr=&id=5R9dGBuuG0MC&oi=fnd&pg=PR6&dq=basketball+is&ots=q8ulHJrrJy&sig=XEsp4y_opb7CB_MVJONyiWxHWrA&redir_esc=y#v=onepage&q=basketball%20is&f=false

[6] Dehesa, R., Vaquera, A., Gomez-Ruano, M. A., Gonçalves, B., Mateus, N., & Sampaio, J. (2019). KEY PERFORMANCE INDICATORS IN NBA PLAYERS'PERFORMANCE PROFILES. *Kinesiology*, *51*(1), 92-101.

https://hrcak.srce.hr/ojs/index.php/kinesiology/article/view/5456

[7] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* Springer.

[8] Anju, & Gulia, P. (2016). Clustering in Big Data: A Review. *International Journal of Computer Applications (0975 – 8887), 153*(3), 44-47. Retrieved from https://www.researchgate.net/publication/310754534_Clustering_in_Big_Data_A_Review

[9] Aggarwal, C. C. (2015). *Data Mining The Textbook.* Springer

[10] Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, *8*, 80716-80727. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9072123

[11] Park, H.-S., & Jun, C.-H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications, 36,* 3336-3341. https://isiarticles.com/bundles/Article/pre/pdf/79087.pdf

[12] Yuan, C., & Yang, H. (2019). Research on K-value selection method of K-means clustering algorithm. *J*, *2*(2), 226-235. https://www.mdpi.com/2571-8800/2/2/16/htm

[13] Arora, P., & Varshney, S. (2016). Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, *78*, 507-512. https://www.sciencedirect.com/science/article/pii/S1877050916000971

[14] Soni, K. G., & Patel, A. (2017). Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data. *International Journal of Computational Intelligence Research*, *13*(5), 899-906. http://www.ripublication.com/ijcir17/ijcirv13n5_21.pdf

[15] Etikan, I., & Bala, K. (2017). Sampling and sampling methods. *Biometrics & Biostatistics International Journal*, *5*(6), 00149. https://bit.ly/3ITaZ90

[16] Xiong, H., & Li, Z. (2018). Clustering validation measures. In *Data Clustering*. Chapman and Hall/CRC.