Win Rate Prediction by NBA 2K Ratings Data Analysis with K-Mean and DBSCAN Algorithm

Rieva Putri Safa
Information System Department, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Tangerang, Indonesia
rieva.safa@student.umn.ac.id

Abstract — NBA 2K is a series of basketball sports simulation video games developed and published annually since 1999. The premise of each game in the series is to mimic the sport of basketball, more specifically the National Basketball Association. Therefore, researchers want to determine the relationship between NBA 2K's rating system and the actual NBA. Researchers want to use this data to explore new strategic opportunities that can be useful for the company's further innovations. This data can also be used for the game players to pick players according to its winning rate accuracy based on the NBA. Data will be analyzed using the K-Means algorithm and the DBScan algorithm which is included in the clustering algorithm group. This study aims to determine whether the winning rate data used on the NBA 2K is accurate enough as the real NBA statistics. This study shows that the K-Means clustering algorithm is more suitable to predict the accuracy of this data. The result shows that the accuracy of the K-Means result is 0.6471 (64.7%) which is higher than the result of DBSCAN with accuracy of 0.0191 (0.02%)

Index Terms — Algorithm, Basketball, Big Data, Clustering, DBSCAN, K-Means, NBA 2K

I. INTRODUCTION

A. Big Data Concept

In this modern era, every human activity never escapes something related to data. Data is raw facts that have not been processed. Every day humans continue to produce data so that along with the times, the amount of data that is available is increasing and is no longer able to be used using simple management tools and conventional methods. Therefore, the term Big Data emerged in society. Big Data refers to technologies and initiatives that involve data that is so diverse, rapidly changing, or so large that it is too difficult for conventional technology, expertise, or infrastructure to handle effectively. [1]

Big data can also be defined as a term that describes the large amounts of data, both structured and unstructured, that flood a business on a daily basis. But it is not the amount of data that matters. What matters is what companies do with the data. Big data can be analyzed to uncover insights that lead to better business strategic moves and decisions. As what Industry analyst, Doug Laney formulated, the definition of big data that is commonly used today as the three V's [2]:

Volume: Organizations collect data from a variety of sources, including business transactions, smart (IoT) devices, industrial equipment, videos, social media and more. In the past, storing it would have been a problem – but cheaper storage on platforms like data lakes and Hadoop have eased the burden.

Velocity: With the growth in the Internet of Things, data streams in to businesses at an unprecedented speed and must be handled in a timely manner. RFID tags, sensors and smart meters are driving the need to deal with these torrents of data in near-real time.

Variety: Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, emails, videos, audios, stock ticker data and financial transactions.

The phenomenon of Big Data is no stranger to human life today. Through technological developments, big data is becoming increasingly popular, especially in the realm of social media. Big data is a trend that covers a wide area in the business and technology world. There are three main things that trigger the development of Big Data technology [3] :

a. The rapid increase in data storage capabilities.

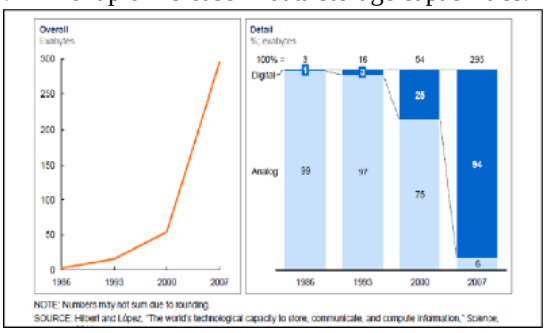


Figure 2. Graph of data storage growth [3]

c. Abundant data availability.

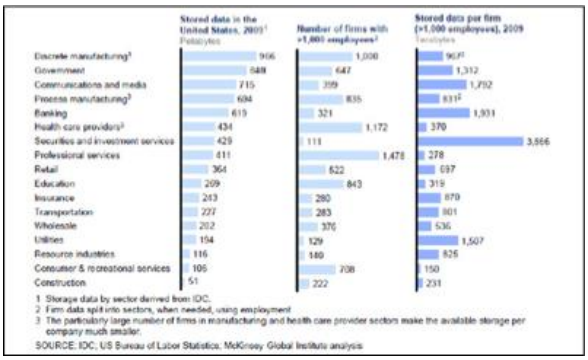


Figure 3. Graph of data availability in various sectors in US [3]

In the business world, the role of big data is very important for entrepreneurs to maintain and develop their business according to the market demands that are constantly changing due to trends. Data has a very strong influence on the continuity of a business, especially for large companies. The data generated by consumers from their transaction activities can no longer be processed with conventional methods, as very large amounts of data with a relatively high level of complexity are generated in a short time. Therefore, companies are now being asked to facilitate big data processing.

With big data, companies will be able to understand what product preferences consumers want, when, and how to adjust prices and sales models to meet consumer needs. If the stored data can be processed and used correctly, a business will certainly progress, such as increasing sales, consumer confidence, and others.

A. *NBA 2K*

NBA 2K is a series of basketball sports simulation video games in development since 1999 with annual release. The premise of each game in the series is to mimic the sport of basketball, more specifically the National Basketball Association. The NBA or National Basketball Association is a multinational company engaged in the media and entertainment industry as well as the sports industry. The NBA is a basketball league from the United States (US). The NBA was

b. The rapid increase in the capabilities of data processing engines.

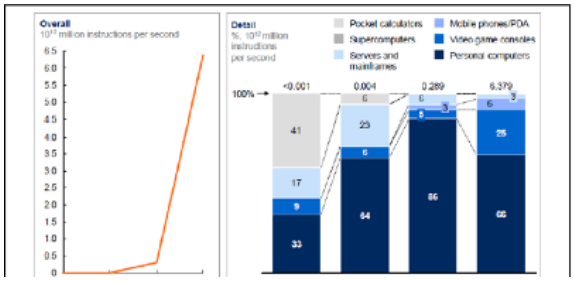


Figure 1. Graph of data processing engines capabilities increase [3]

founded in 1949 as a result of the merger of two basketball leagues, the American Basketball Association (NBA) and the National Basketball League (NBL) [4]. The latest edition is 2K21, released on September 4, 2020. The video games are now being published by 2K Sports. In each publication, all active NBA players and some Legends are individually rated on a 99-point scale. These ratings always lead to discussions, debates, reactions, and even from the players themselves. [5]

NBA 2K games have come a long way from being a distinguished basketball game to being the best basketball game of all time. From its graphics, gameplay, community, and many more reasons why the game became so successful today, NBA 2K has become the best basketball game by improving its overall structure throughout the year. The NBA 2K series marked a milestone in the history of the sports video game world. When the first NBA 2K was released on November 10, 1999, developed by Visual Concepts Entertainment and published by Sega Sports, it ushered in the best basketball video game series to come to life today. Every year since NBA 2K was released, the game has been updated to provide a more realistic way to play basketball in a video game. The first console to hit the market was the console called Dreamcast, which ranged from NBA 2K to NBA 2K2. Thereafter, Playstation, Xbox, and Gamecube released NBA 2K games, starting with NBA 2K2, and Gamecube left the game in NBA 2K3 for many years. [6]

B. *Purpose of the Research*

eSports is the competitive game of video games. It takes many forms, from local casual tournaments to global international events. In terms of structure and supporting industries, it is now very similar to (conventional) sport. Esports have been around since the beginning of games. In Cabinet days, players competed with each other in gambling halls and in local tournaments. [7]

It wasn't until the late 2000s, when Internet connections improved, that the e-sports boom really began, that it reached people's homes and the bigger picture. While esports are generally viewed via online streaming, most high-level events take place on

physical venues. As what Galov stated on HostingTribunal about some eSports statistics [7] :

1. eSports viewership was growing at a rate of **11.7% in 2020**.
2. There were a total of **496 million** eSports viewers in 2020.
3. **4.4 million** monthly eSports players streamed on Twitch in 2020.
4. The average pro eSports player earned **$5,000** in 2020.
5. The eSports industry reached **$950 million in revenue** in 2020.
6. **$4.5 billion** was invested in eSports in 2018.
7. The industry is projected to reach **$1.5 billion** by 2023.
8. **42%** of viewers enjoy the adverts they see on stream.

Mentioning about eSports, NBA 2K is also an eSport game. NBA 2K is not just a 'game', but it has been a serious business for all. Including players, team owners, and fans of the 2K themselves has big expectations for this game league. The NBA 2K League holds a big tournament every season with hundred thousand dollars as the prize pool. Brendan Donahue, the NBA's Managing Director of NBA 2K League mentioned that this game has a high chance in being globally recognized compared to any other game in eSports, remembering all of the demographics on the game. Donohue also stated that [8]:

> "… there are another 200 million people who are eSports enthusiasts and play regularly. They may be fans of other games and other titles, like the mega-hits 'League of Legends', 'Dota 2', the 'Call of Duty' series and the like. But the NBA thinks it league can engage them as well …"

By that, they expected that the NBA 2K will keep taking the lead on eSports because there are 1.4 billion people worldwide who are NBA fans, with whom the league engages daily through its various platforms. The NBA 2K League is indeed hoped to provide a new way to engage with them. Also, there are an estimated 1.6 million people, who play NBA 2K every day, at an average of 90 minutes per day. [8]

Table 1. NBA2K - NBA2K21 sales [9]

Title	**Year**	**Sales** (mill)
NBA 2K	1999	
NBA 2K1	2000	
NBA 2K2	2001/2002	
NBA 2K3	2002	
ESPN NBA Basketball	2003	
ESPN NBA 2K5	2004	1.6
NBA 2K6	2005	
NBA 2K7	2006	
NBA 2K8	2007	
NBA 2K9	2008	2
NBA 2K10	2009	2
NBA 2K11	2010	5.5
NBA 2K12	2011	4
NBA 2K13	2012	4
NBA 2K14	2013	7
NBA 2K15	2014	7
NBA 2K16	2015	4
NBA 2K17	2016	8.5
NBA 2K18	2017	10
NBA 2K19	2018	12
NBA 2K20	2019	14
NBA 2K21	2020	8

According to the NBA2K - NBA2K21 sales from Video Game Sales Wiki, it can be told that there is a significant drop on the sales on the latest season. By implementing big data and the right data analysis, NBA will then know what the players want, expect, and hope from the game in order to prevent and overcame a decline in sales.

This study aims to find the best analytical model with a high level of accuracy using K-Means and DBSCAN for the clustering of win rate accuracy predictions in NBA 2K games against NBA real statistics as one of the factors supporting the success of this game's sales.

II. Literature Review

A. Data Mining

Data mining is a step in Knowledge Discovery in Databases (KDD). Knowledge discovery as a process consists of data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation. Data mining refers to the process of mining knowledge from a very large set of data. The framework of the data mining process is composed of three stages, namely data collection, data transformation, and data analysis. The process begins with pre-processing which consists of collecting data to produce raw data required by data mining, which is then followed by data transformation to convert raw data into a format that can be processed by data mining, for example through filtration or aggregation. The results of the data transformation will be used by data analysis to generate knowledge using techniques such as statistical analysis, machine learning, and information visualization [10]. Data mining can also be interpreted as a series of processes carried out to explore added value in the form of information that has not been known manually from a database by extracting patterns from data that aims to manipulate data into more valuable information obtained by extracting data. and recognize important or interesting patterns from existing data in a database [11].

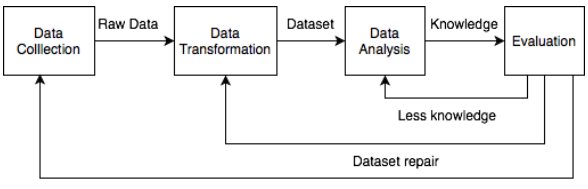


Figure 4. Data Mining Information Flow [10]

B. Clustering

Clustering is an activity to consider an important approach to find similarities in data and place similar data into groups. Clustering is considered as the most important unsupervised learning where every problem of this type relates to finding structures in unlabeled data sets. A cluster grouping divides the data set into a number of groups where the similarity in one group is greater than in other groups. The use of clustering algorithms depends on the type of data present for a particular purpose and application. If cluster analysis is used as a descriptive or exploratory tool, it is possible to try several algorithms on the same data to obtain what the data reveals. In general, clustering methods can be classified into several categories, one of which is the partitioning method category. This partitioning method is based on initial determination of the number of groups, then iteratively reallocating objects to rediscover groups that are in one point [12].

C. K-Means

K-Mean clustering is a cluster analysis method that aims to break up objects into k clusters and then observe where each cluster object is obtained through the nearest average. This algorithm is one of the famous learning simple and easy to learn as a problem solving grouping of a dataset. The K-Means algorithm is an evolutionary algorithm whose method of operation has the same meaning as the name of the algorithm. This algorithm groups observations into k groups, where k is the input parameter. Each data is then assigned to each cluster observation based on the proximity of the observations to the cluster mean value. The average value in the cluster is then calculated repeatedly in the initial process [12]. The stages of doing K-Means Clustering are as follows [13]:

1) Determine the desired number of clusters
2) Allocate data according to the predetermined number of clusters.
3) Determine the centroid value in each cluster.
4) Calculate the closest distance using the Euclidean formula.
5) Show the results based on the lowest distance from the calculation results of step 4
6) If the appropriate results have not been obtained, the iteration is continued again using step 3. The iteration will be stopped if the clustering results are the same as the previous iteration.

The weaknesss of K-Means, includes [14]:

1) If the amount of data is not too much, it is easy to determine the initial cluster.
2) The number of clusters, as many as K, must be determined before calculating.
3) Never know a real cluster using the same data, but if it is entered in a different way, it may produce different clusters if the amount of data is small.
4) Do not know the contribution of the attributes in the grouping process because it is assumed that each attribute has the same weight.

The solution to overcome this weakness is to use K-means clustering but only if there is a lot of data available.

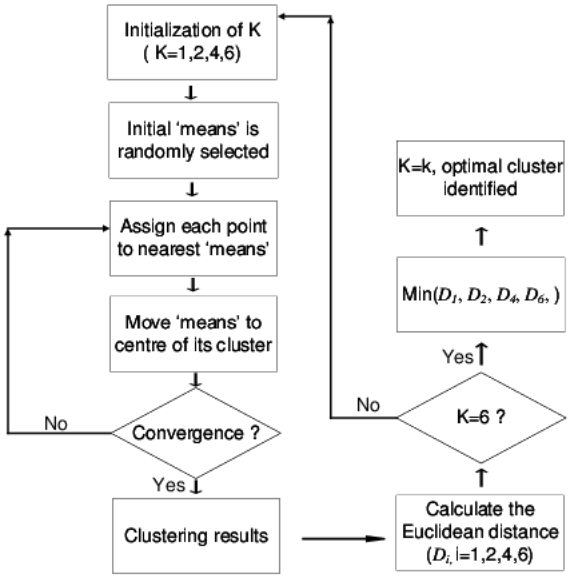


Figure 5. Flowchart of K-Means

D. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Density – Based Spatial Clustering Algorithm With Noise (DBSCAN) is a clustering algorithm based on data density. The concept of density referred to in DBSCAN is the number of data (minPts) that are within the Eps radius (ε) of each data. The concept of density like this produces three kinds of status for each data, namely core, border, and noise. Core data is data whose amount of data within the eps radius is more than minPts, noise data is data whose number of data within the eps radius is less than minPts, and boundary data is data where the amount of data within the eps radius is less than minPts but makes the neighboring data become core data. The DBSCAN grouping process is to calculate the distance from the center point (p) to another point using the Euclidean distance and is expressed in the following equation [15]:

$$D(x_1, p_1) = \sqrt{\sum_{j}^{q} = 1\,(x_{ij} - p_{ij})^2}$$

E. Confusion Matrix

In data mining to measure or there are several ways to measure the performance of the resulting model, one of them uses a confusion matrix (accuracy). Confusion matrix is a method used to perform accuracy calculations on the concept of data mining. Precision or confidence is the proportion of positive predicted cases that are also true positives in the actual data. Recall or sensitivity is the proportion of true positive cases that were correctly predicted positive [16].

Table 2. Confusion Matrix Model

Aktual	Classified as	
	+	-
+	True positives (A)	False negatives (B)
-	False positives (C)	True negatives (D)

The calculation of accuracy with the confusion matrix table is as follows:

Accuracy = (A+D)/(A+B+C+D)

Precision is defined as the ratio of selected relevant items to all selected items. Precision can be interpreted as a match between the request for information and the answer to the request. The precision formula is:

Precision = A/(C+A)

Recall is defined as the ratio of selected relevant items to the total number of items relevant available. Recall is calculated by the formula:

Recall = A/(A+D)

Precision and Recall can be assigned a numerical value by using a percentage calculation (1-100%) or by using a number between 0-1. The recommendation system will be considered good if the value of precision and recall is high.

The ROC curve shows accuracy and compares classifications visually. ROC expresses the confusion matrix. ROC is a two-dimensional graph with false positives as horizontal lines and true positives as vertical lines. AUC (the area under curve) is calculated to measure the difference in the performance of the method used. ROC has a diagnostic value level, namely [16]:

a. Accuracy is 0.90 – 1.00 = excellent classification
b. Accuracy is 0.80 – 0.90 = good classification
c. Accuracy is 0.70 – 0.80 = fair classification
d. Accuracy is 0.60 – 0.70 = poor classification
e. Accuracy is 0.50 – 0.60 = failure

III. METHODOLOGY

A. Object of the Research

This study focuses on processing NBA league game data with original statistics used for rating player skills in NBA 2K. The dataset that will be used for the purposes of this research includes data that starts from NBA 2K16 (2014-15 season), up to NBA 2K21 (2019-20 season). The dataset contains 2,412 observations and 23 variables. The 2,412 observations represent the number of players in the NBA league. On the other hand, the 23 variables in the dataset are divided into two types based on the data type, namely numeric variables and categorical variables. If detailed, the 8 variables consist of 2 categorical variables and 21 numeric variables. The categorical variables are Player and Season, while the numerical variables are Age, Games Played, Wins, Loss, Mins, Points, Goal Made, Goal Percentage, Goal Attempt, Three Point Made, Three Point Attempt, Three Point Percentage, Free Throw Made, Free Throw Attempt, Free Throw Percentage, Rebound, Assist, Steal, Block, and Rankings. Each of these numerical variables describes the calculation of the average NBA play of each player.

B. Method of Data Collection

The data collected and used for the research is not primary, as it does not come from survey results or other data collection methods that require researchers to participate in the actual data collection process. The data used in this study is a secondary data, because data was previously collected by other parties. Primary data collection cannot be done on the basis of this research because of the limited scope and ability to collect data independently in a relatively short period of time.

The secondary data collected was obtained from a site named Kaggle (with the link https://www.kaggle.com/willyiamyu/nba-2k-ratings-with-real-nba-stats). Kaggle is a site or a platform that hosts competitions in the field of data science. Furthermore, this site is also one of the common learning resources for data science. In order to support researchers, Kaggle provides several data sets with different data variations, with which research on interesting topics can be conducted without difficulty in data collection.

C. Research Methods

Research methods are procedures that researchers use to solve problems that arise in research activities. In this way, it can be said that the research method is the main method used by researchers to achieve the objectives and obtain answers to the problems that are being carried out. In this study, the researchers used quantitative research method designs, supported by data analysis capabilities, to allow the processed data to have more value in the decision-making process of certain parties.

Data processing is done with R. R is a programming language and a computer program used to support statistical and graphical analysis activities. R is accessed through RStudio. RStudio is an integrated development environment (IDE) for R. Additionally, data must first be validated prior to data processing to ensure the accuracy and security of the

data to be used in research. R also facilitates this validation function.

The dataset of the research the NBA 2K Ratings for each player, as well as their corresponding real-life NBA statistics for that season. In which the data starts from NBA 2K16 (2014-15 season), up to NBA 2K21 (2019-20 season). This data will be analyzed and researched by using two kinds of algorithms, K-Means and DBSCAN algorithm. Both of these algorithms are in the scope of data mining in clustering. As stated above, data mining is the process of looking for interesting patterns or information in selected data using certain techniques or methods [17].

IV. Results and Discussion

A. *Data Validation*

The data from the dataset needs to be validated in order to is intended so that the truth and completeness of the data used can be guaranteed. Data validation is done by aligning the data type that a variable should have. In the research data that will be used, the data type for the variable 'wins' is still not correct when imported into R because it is still an integer type. The 'wins' variable must be changed to a factor data type because it is included in a numeric variable.

After changing the data type to factor, the data must also be checked for completeness in order to further validate the data. The check is done with the *na omit* function available in R to eliminate incomplete cases. In this research, we re-validated the *na omit* checking by using the *missmap* function of the Amelia package in R. The results from both checks showed that the datasets used were all complete and no data was lost at all. Thus, it can be concluded that the data is valid for use in research because it has passed the stage of checking the correctness and completeness of the data.

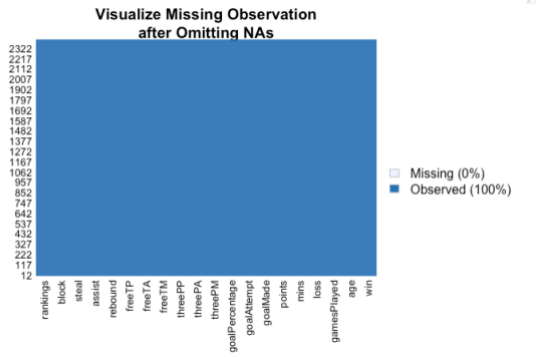


Figure 6. Missmap on Dataset

In addition on checking the validity of the data using *na omit* and *missmap*, the researcher will also provide additional detailed information to support the validity of the data. The information is obtained from the Kaggle website, where this dataset was obtained. It is proven that among all of the variables, this dataset is perfect and complete, without any data being lost.

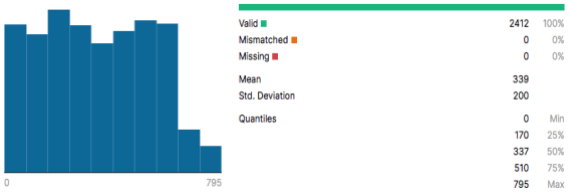


Figure 7. Validity of the Dataset

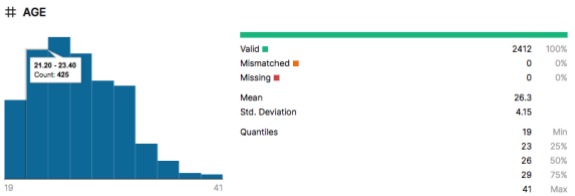


Figure 8. Validity of Age Variable

Figure 9. Validity of Games Played Variable

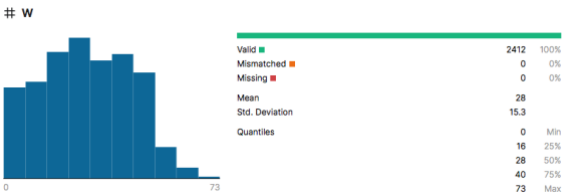


Figure 10. Validity of Wins Variable

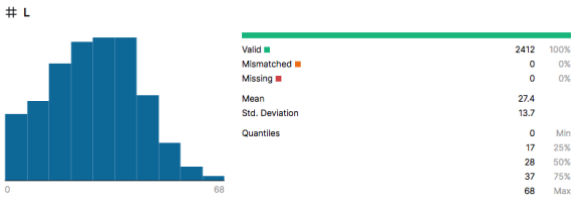


Figure 11. Validity of Loss Variable

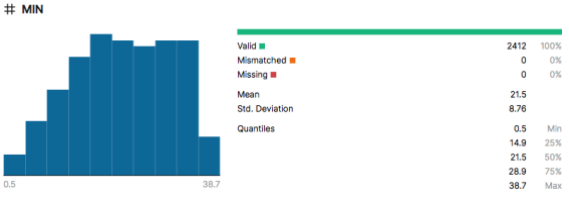


Figure 12. Validity of Mins Variable

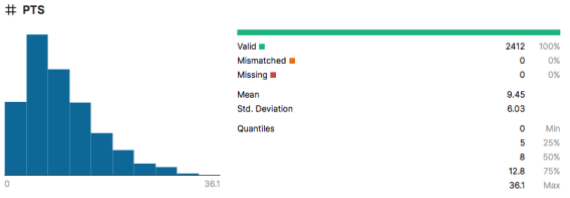


Figure 13. Validity of Points Variable

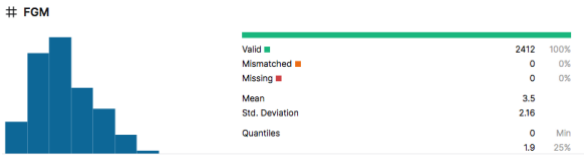


Figure 14. Validity of Goal Made Variable

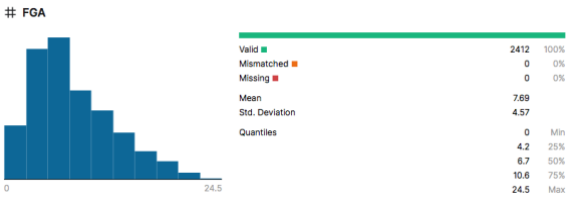


Figure 15. Validity of Goal Attempt Variable

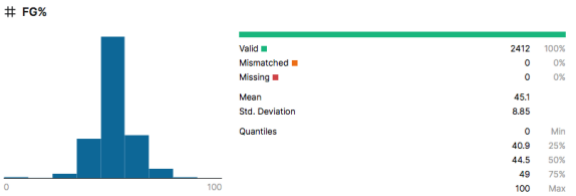


Figure 16. Validity of Goal Percentage Variable

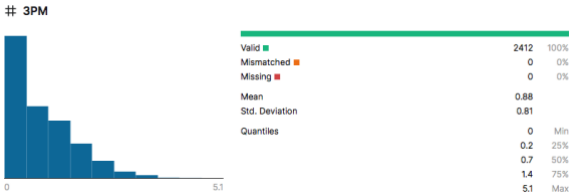


Figure 17. Validity of Three Point Made Variable

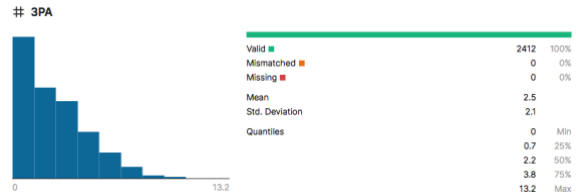


Figure 18. Validity of Three Point Attempt Variable

Figure 19. Validity of Three Point Percentage Variable

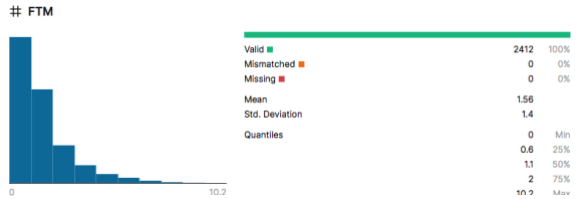


Figure 21. Validity of Free Throw Made Variable

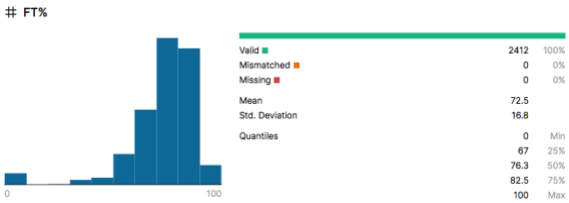


Figure 22. Validity of Free Throw Percentage Variable

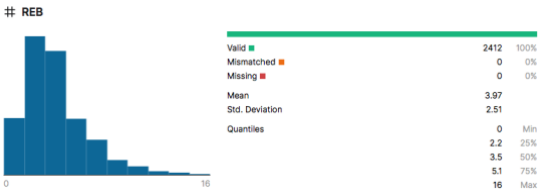


Figure 23. Validity of Rebound Variable

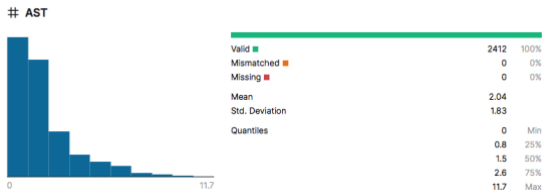


Figure 24. Validity of Assist Variable

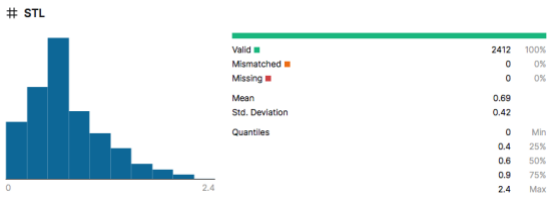


Figure 25. Validity of Steal Variable

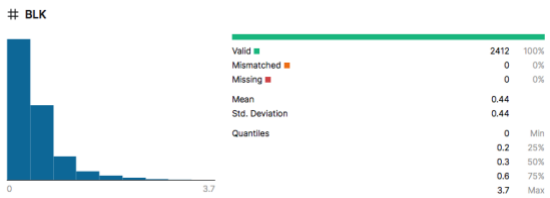


Figure 26. Validity of Block Variable

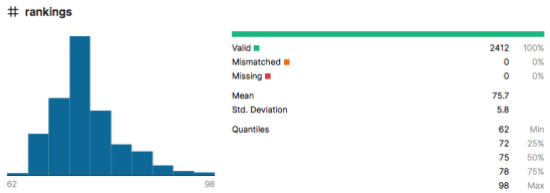


Figure 27. Validity of Rankings Variable

B. *Data Visualization*

1) *Data Visualization using Barplot*

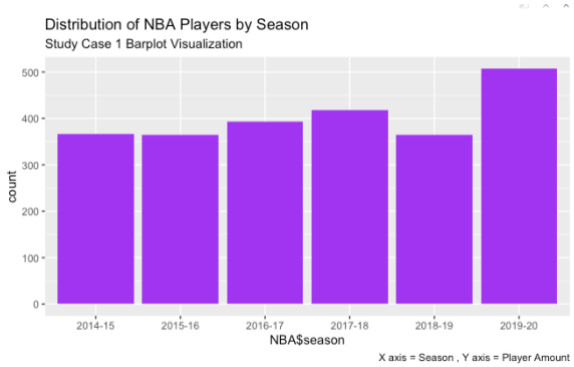


Figure 28. Distribution of NBA Players by Season Barplot

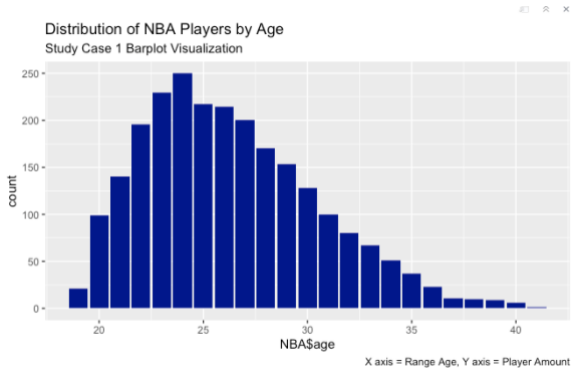


Figure 29. Distribution of NBA Players by Rank Boxplot

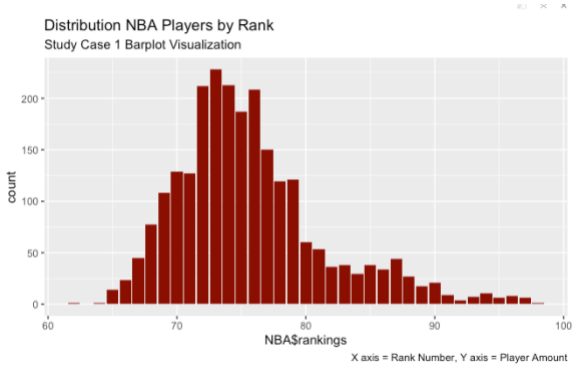


Figure 30. Distribution of NBA Players by Age Boxplot

2) *Data Visualization using Boxplot*

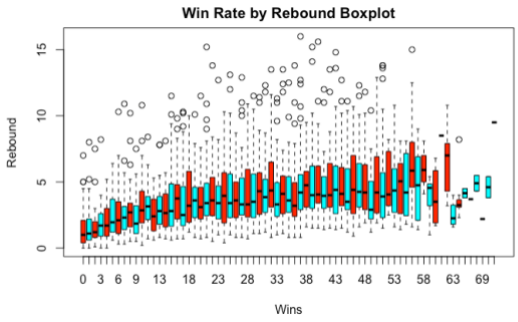


Figure 31. Win Rate by Rebound Boxplot

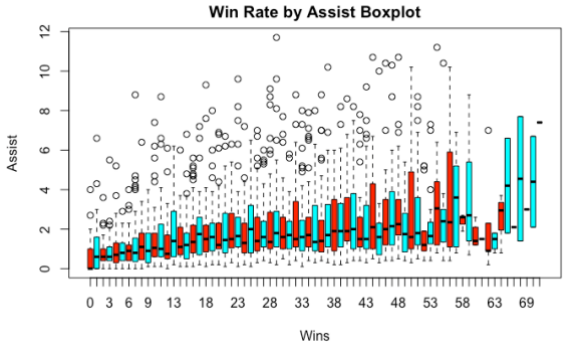


Figure 32. Win Rate by Assist Boxplot

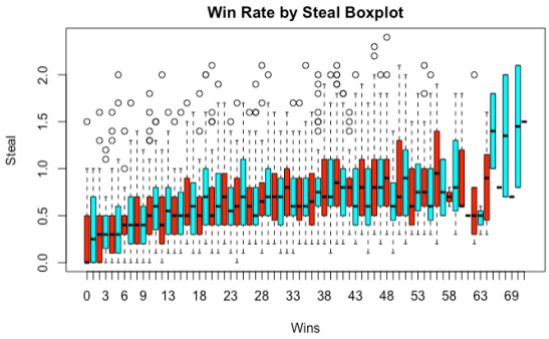


Figure 33. Win Rate by Steal Boxplot

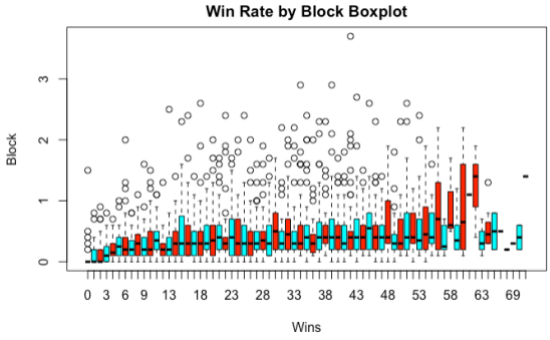


Figure 34. Win Rate by Block Boxplot

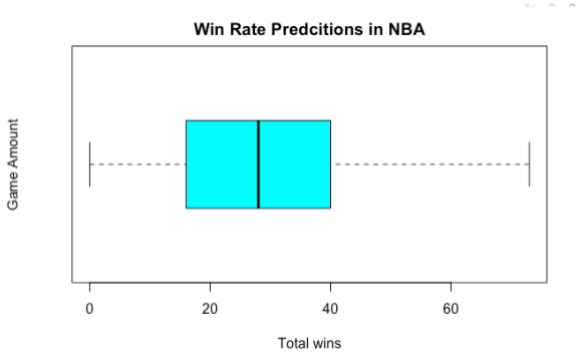


Figure 35. Win Rate in NBA Boxplot

3) *Data Visualization using Scatterplot*

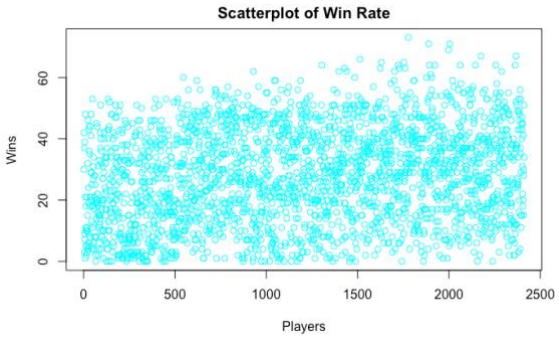


Figure 36. Win Rate Scatterplot

C. *Application of K-Means and DBSCAN*

After changing the data types as numeric and as factor, checking data validity, and making data visualizations, researchers then splits the data into training and testing, with the sample ratio of 80:20. There are 1,929 samples used in training and 483 samples used in testing. In the assessment and testing, both algorithms uses selected and target independent variables.

1) *K-Means*

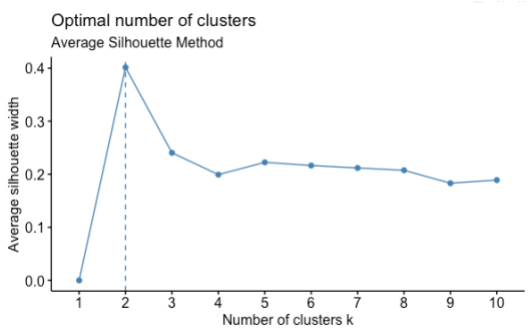


Figure 37. K-Means Optimal Cluster Plot

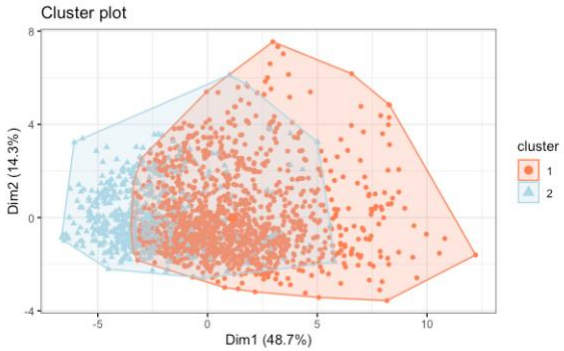


Figure 38. K-Means Cluster Plot

```
----------------------------------------
purity                          : 0.7714
entropy                         : 0.5765
normalized mutual information   : 0.3137
variation of information        : 1.2831
normalized var. of information  : 0.814
----------------------------------------
specificity                     : 0.5598
sensitivity                     : 0.7345
precision                       : 0.6251
recall                          : 0.7345
F-measure                       : 0.6754
----------------------------------------
accuracy OR rand-index          : 0.6471
adjusted-rand-index             : 0.2943
jaccard-index                   : 0.5099
fowlkes-mallows-index           : 0.6776
mirkin-metric                   : 1312416
----------------------------------------
[1] 0.2942676
```

Figure 39. K-Means External Validation Using Rand Index

From this assessment, by using factoextra function, it shows that this data has two optimal clusters. From the testing assessment using K-Means (unscaled), it shows that the accuracy or rand-index rate for the prediction is 64.7% (0.6471). That shows that using by using the K-Means clustering for this dataset is good enough based on the medium accuracy. From the validation result, we can also see that the sensitivity of this prediction is at 73% (0.7345), specificity is at 56% (0.5598), variation of information at 128% (1.2831), and corrected rand-index at 30% (0.2942).

2) *DBSCAN*

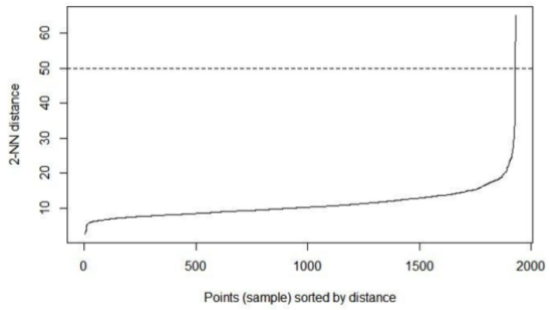


Figure 40. DBSCAN Points

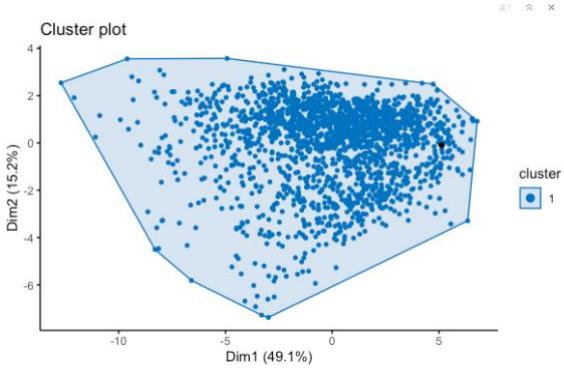


Figure 41. DBSCAN Cluster Plot

```
----------------------------------------
purity                          : 0.0264
entropy                         : 5e-04
normalized mutual information   : 0.0011
variation of information        : 5.8467
normalized var. of information : 0.9994
----------------------------------------
specificity                     : 0.001
sensitivity                     : 0.9993
precision                       : 0.018
recall                          : 0.9993
F-measure                       : 0.0354
----------------------------------------
accuracy OR rand-index          : 0.019
adjusted-rand-index             : 0
jaccard-index                   : 0.018
fowlkes-mallows-index           : 0.1342
mirkin-metric                   : 3648382
----------------------------------------
[1] 1.284432e-05
```

Figure 42. DBSCAN External Validation Using Rand Index

From this assessment, by using factoextra function, it shows that this data has two optimal clusters. From the testing assessment using DBSCAN, it shows that the accuracy or rand-index rate for the prediction is 2% (0.019). That shows that using by using the DBSCAN clustering for this dataset is not good and not suitable based on the very low accuracy. From the validation result, we can also see that the sensitivity of this prediction is at 99% (0.9993), specificity is at 0.1% (0.001), variation of information at 584% (5.8467), and corrected rand-index at 1.284432e-05.

V. CONCLUSION

Table 3. Validation Result Comparation

Parameter	**K-Means**	**DBSCAN**
Variation of Information	1.2831	5.8467
Specificity	0.5598	0.001
Sensitivity	0.7345	0.9993
Accuracy or Rand-Index	0.6471	0.019
Corrected Rand-Index	0.2942	1.284432e-05

From the results obtained from this research, as shown on table 3, researcher can conclude that K-Means model is better and a lot more accurate to use compared to DBSCAN. It is proven by the big gap of the accuracy rate between the two models.The accuracy on the K-Means is greater on 64% compared to DBSCAN with only 0.1%. As for the variation of information, sensitivity, and specificity rate between the two models, it is also shown that the K-Means model is winning by comparing the huge gap the result shows. By this, it is proven that the DBSCAN model is not suitable for predicting the win rate on NBA 2K based on the real statistics of NBA.

ACKNOWLEDGMENT

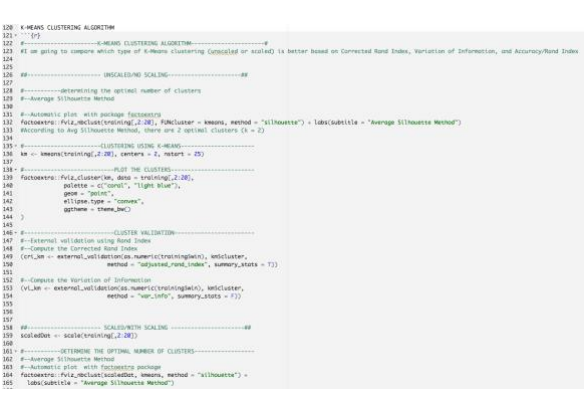
One of us would like to thank Mr. Ir. Raymond Sunardi Oetama, M.CIS as the lecturer for the Data Analysis course, Information Systems Department, Multimedia Nusantara University, for willing to share all the knowledge and insights that are very useful for the continuity of this research procession and for providing input, recommendations, support, and guidance since the very beginning of our study.

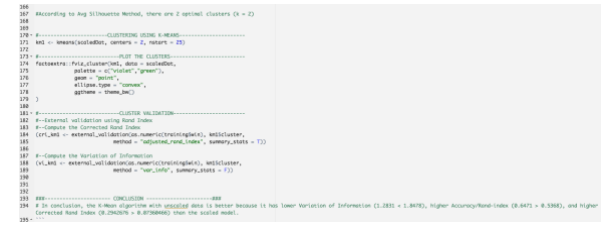
ATTACHMENTS

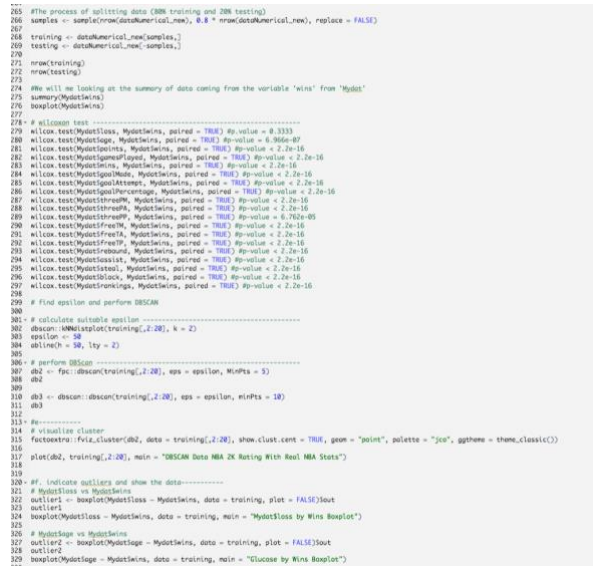
The attachments containing the R code in analyzing the win rate predictions in NBA 2K based on the real NBA statistics provided in the following:

LOAD LIBRARY

```{r, warning = FALSE, message = FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(dplyr)
library(readxl)
library(Amelia)
library(ggplot2)
library(factoextra)
library(ClusterR)
```







```
# MydatSassist vs MydatSwins
outlier7 <- boxplot(MydatSassist ~ MydatSwins, data = training, plot = FALSE)$out
outlier7
boxplot(MydatSassist ~ MydatSwins, data = training, main = "Diabetes Pedigree Function by Wins Boxplot")

# MydatSsteal vs MydatSwins
outlier8 <- boxplot(MydatSsteal ~ MydatSwins, data = training, plot = FALSE)$out
outlier8
boxplot(MydatSsteal ~ MydatSwins, data = training, main = "Age by Wins Boxplot")

# MydatSblock vs MydatSwins
outlier7 <- boxplot(MydatSblock ~ MydatSwins, data = training, plot = FALSE)$out
outlier7
boxplot(MydatSblock ~ MydatSwins, data = training, main = "Diabetes Pedigree Function by Wins Boxplot")

# MydatSrankings vs MydatSwins
outlier8 <- boxplot(MydatSrankings ~ MydatSwins, data = training, plot = FALSE)$out
outlier8
boxplot(MydatSrankings ~ MydatSwins, data = training, main = "Age by Wins Boxplot")

#g. validation---------
# validation using confusion matrix --------------------------------
# pred vs truth

(cri_db <- external_validation(as.numeric(trainingSwins), db2$cluster, method = "adjusted_rand_index", summary_stats = T))

# Compute the Variation of Information
(vi_db <- external_validation(as.numeric(trainingSwins), db2$cluster, method = "var_info", summary_stats = F))
```

REFERENCES

[1] Pujianto, A., Mulyati, A., & Novaria, R. (2018). Pemanfaatan Big Data dan Perlindungan Privasi Konsumen di Era Ekonomi Digital. *Majalah Ilmiah BIJAK*, *15*(2), 127-137.

[2] *Big data: What it is and why it matters*. (n.d.). https://www.sas.com/en_id/insights/big-data/what-is-big-data.html

[3] Maryanto, B. (2017). Big data dan pemanfaatannya dalam berbagai sektor. *Media Informatika*. 16 (2).

[4] Hasudungan, A. A. (2018). *Upaya-upaya NBA dalam pemasaran produk olahraganya di Tiongkok*. Bandung: Universitas Katolik Parahyangan.

[5] Rallet, B. (2021, January 10). Sport Analytics — NBA 2K ratings prediction. Medium. https://towardsdatascience.com/sport-analytics-nba-2k-ratings-prediction-b7b72e2e72eb

[6] Joseph Saludo. (2018). NBA 2K. *ART 108: Introduction to Games Studies*.

[7] Nick Galov. (2021, April 2). *30 eSports stats to frag the filthy casuals with in 2021*. HostingTribunal. https://hostingtribunal.com/blog/esports-stats/#gref

[8] David Aldrige, D. A. (2018, April 9). *Not just a game: NBA 2K league quickly becoming a serious business for all*. NBA.com. https://www.nba.com/news/morning-tip-nba-2k-league-draft-serious-business-players-owners-fans

[9] *Nba 2k*. (n.d.). Video Game Sales Wiki. Retrieved October 6, 2021, from https://vgsales.fandom.com/wiki/NBA_2K

[10] Firdaus, D. (2017). Penggunaan data mining dalam kegiatan sistem pembelajaran berbantuan komputer. *Jurnal Format*. 6(2).

[11] Binjori, A. S. (2020). Implementasi data Mining untuk pengembangan sistem rekomendasi pemilihan SMK dengan menggunakan algoritma Cart. *KLIK: Kajian Ilmiah Informatika & Komputer*, *1*(2), 42-48.

[12] Kamila, I., Khairunnisa, U., & Mustakim. (2019). Perbandingan algoritma K-Means dan K-Medoids untuk pengelompokan data transaksi bongkar muat di Provinsi Riau. *Jurnal Ilmiah Rekayasa dan Manajemen Sistem Informasi*, *5*(1).

[13] Sani, A. (2018). *Penerapan metode K-Means clustering pada perusahaan*. STMIK Widuri.

[14] Sibuea, F. L, & Sapta, A. (2017). Pemetaan siswa berprestasi menggunakan metode K-Means clustering. *Jurnal Teknologi dan Sistem INformasi*, *4*(1), 85-92.

[15] Ashari, B. S., Otniel, S. C., & Rianto. (2019). Perbandingan kinerja K-Means dengan DBSCAN untuk metode clustering data penjualan *online retail*. *Jurnal Siliwangi*, *5*(2).