Original Research Paper

Analysis of Big Data Football Club Market Value Using K-Means and Linear Regression Mining Methods

[1]Johanes Fernandes Andry, [2]Riama Sibarani and [1]Vlavio Nathanael Yefta

[1]*Department of Information Systems, Bunda Mulia University, Jakarta, Indonesia*
[2]*Department of Informatics, Satya Negara Indonesia University, Jakarta, Indonesia*

Abstract: Football clubs store a lot of data about their players, squad, transfer market, and market value, so big data is needed to process this data. This data can be analyzed to gain insight into the club's market value. Changes in player performance, age, and squad size can be analyzed using statistical methods. These digital data will provide information about the club's market value. Data mining is the process of analyzing data using various methods to produce useful information. The software application used in this data analysis is RapidMiner Studio, which is one of the best data mining tools. The purpose of this research is to analyze football clubs' market value according to squad size, players' value, and league. This study will use the clustering K-means and linear regression methods. The results of this study can be used by those who want to invest money in football clubs. An investor can use this data to predict and make decisions about whether to invest in specific clubs and leagues.

Keywords: Big Data, Football Clubs, RapidMiner, K-Means, Regression

Introduction

Data mining refers to extracting or mining knowledge from a very large amount of data. The term data mining can be referred to as knowledge mining from data and information or knowledge mining (Jain and Srivastava, 2013; Balusamy *et al*., 2021).

Big data is a very important concept, which is implemented on a large amount of data, which does not conform to the normal structure of the former database. Big data itself consists of several important technologies, namely Hadoop, HDFS, NoSQL, PIG, MapReduce, MongoDB, Cassandra, HIVE, and HBASE which of course can achieve goals to achieve final goals such as extracting values from data that were previously considered useless (Zakir *et al*., 2015; Ghavami, 2016). Data to be classified as large, the data must have three Vs, namely: Volume, Variety, and Speed. Many people assume that big data only has volume (Korcheva, 2021), the other two V's are no less important. Big data is not only big but varied. It comes in a variety of formats and can be structured or unstructured. The speed of the software, of course, refers to the data generated (Russom, 2013).

With millions of players and billions of fans from all over the world, soccer is a very popular sport world. A time when the club of a professional soccer team made a huge income; they are not just a club but a giant company with shareholders and managers and sales and profits (Amir and Livne, 2005; Al-Radaideh *et al*., 2013). This research is very important in analyzing soccer to determine which players have the maximum potential in achieving their professional careers in this sport. Of course, this helps the coach and the team able to identify which players are the most promising and create a solid team composition. Furthermore, players can of course also benefit from this. in principle, they can adapt to career schemes that have potential according to their abilities in football (Acs and Toka, 2022). Football clubs are not really after big profits, even football is not a big business and not a good business either (Muhammad, 2017).

This research will use the K-means and linear regression algorithms, K-means is used to group the football clubs and linear regression is used to predict their market values. K-means is a distance-based clustering algorithm that partitions data into a predefined number of clusters. Each cluster will have a centroid (center of gravity). Cases (individuals in the population) that are in a cluster near the centroid. For example, grouping demographic data into clusters and ranking the probability that an individual will belong to a particular cluster (Chitra and Subashini, 2013). The purpose of clustering is to find out the similarity and design of a large data set by dividing the data into several groups. Since it is assumed

that the data set is unlabeled, clustering is often considered the most valuable problem of unsupervised learning (Bryant and Cios, 2017). In the K-means algorithm, each piece of data must belong to a certain cluster and it is usually possible for any data that belongs to a certain cluster at one stage of the process, at the next stage it moves to another cluster (Metisen and Sari, 2015).

Regression predicts a numerical value (Gharehchopogha and Khalifehlou, 2012). Regression performs operations on the dataset in which the target value has been defined and the result can be extended by adding new information (Gharehchopogh *et al.*, 2013).

Soccer is the world's largest sport in terms of TV ratings, stadium visits, and total income (Collignon and Sultan, 2014). In 2011, football alone accounted for 43% of the global sports market in terms of ticket, media, and sponsorship revenue. It makes football more interesting than a sporting point of view. Over the last few decades, the industry has evolved into a billion-dollar global industry with global players and large investments (Hagen and Cunha, 2019).

In the past, investing in football club stock has meant high risk, low returns, and loss of money. Supporting your club was done from an emotional state or as a childhood dream, rather than purely for financial reasons. However, recent developments show that football stock returns outperform the returns of common benchmark indexes such as the FTSE 100 (Hagen and Cunha, 2019).

Looking at Deloitte's first football-money league report covering the 1997/1998 season, the total revenue of the 20 wealthiest clubs in Europe was € 1.4 billion. Looking at the numbers 20 years later, Deloitte's FML report for the 2017/2018 season shows that the 20 wealthiest clubs have total revenues of € 8.3 billion (Deloitte, 2018). That's 593% sales growth.

Looking at the financial figures for the European football industry, operating profit has increased by 77% in the last decade alone. From 2008 to 2017, the operating revenue of 4 of all of Europe's top clubs increased from € 11.4 to 20.1 billion (UEFA, 2019). Apart from match results, few studies have been performed on which factors affect the stock price of listed football clubs. Allouche and Soulez (2008) are among the few who have examined player transfers and the effect of a club's financial statements on the corresponding share price. The two studied 23 English clubs over three seasons from 1998/99 to 2000/01. How can we choose clubs to invest in with all these financial problems? (Hagen and Cunha, 2019).

The goal of this advanced analytical process is to extract information from the dataset and transform it into an understandable structure for further use (Zernik, 2010). Although this description gives a fairly high-level overview of data mining, this concept has been defined in various forms in the past. (Maksood and Achuthan, 2016).

Financial investing 101 states that shareholders generally want a company to invest in to maximize profits. Shareholders want the highest possible return on investment. For football shareholders, the club's desired behavior may look different. Cheffins (1999) found that many people who bought stock in football clubs bought it because of their love for the club. As a sentimental gesture, fans wanted to own part of their club or influence the club through voting rights Duque and Ferreira (2005).

Through this research, the investor can find which club has better market value so they can invest in that club (Adiwiyana and Harymawan, 2021).

Materials and Methods

Authors analyze and solve the problems encountered. In the research, the stages used are literature study, data collection, data processing, data analysis, and implementation.

Research Stages

The research stage starts from data collection to implementation. Data collecting, collection, and analysis of data using secondary data from https://www.kaggle.com/datasets/ (Provost and Fawcett, 2013).

Processing the data that has been obtained will then be processed by entering the dataset from www.kaggle.com into RapidMiner Studio and will proceed to the next stage, namely data analysis.

Data analysis, at this stage the data is analyzed using rapid mining software. The data is analyzed by grouping the data into several clusters, then the data will be predicted based on the reference of some selected attributes.

Implementation, the result of data mining techniques applied in the RapidMiner application can be seen in Fig. 1.

Research Subject

The subject of this study is the market value of football clubs. In this study, there are 102 units/amount of data in the dataset to be tested. Data collection is done by searching the appropriate data in the documentation on Kaggle.

Instrument Development

Instrument development is very important in research activities. Research instruments have a very important contribution in determining the quality of a study, where the validity or clarity of the data obtained will be largely determined by the quality of the instruments used by researchers, in addition to the data collection procedures used. In developing this instrument, will be used in this study is a documentation guide. A documentation guide is a tool used to collect data archives and documentation. Documentation using Kaggle data.

Table 1: Data attributes and description

Instance	Type data
Club name, competition name	Polynomial
Squad size	Integer
Average age of players, the market value of the club in millions (£)	Real
Average market value of players in millions (£), the market value of the top 18 players in millions (£)	

Table 2: Data attributes and description

No.	Attributes	Description
1.	Club name	Club name is an attribute that contains data on the names of clubs listed in the "top 100 football club market value 2021" in the dataset
2.	Competition name	Competition name is an attribute that contains data on the name or type of competition listed in the "top 100 football club market value 2021". The competition name data in this the dataset can be doubled because one type of competition can be followed by several clubs in the data set
3.	Squad size	Squad size is an attribute that displays overall information about the club's finances and how rich the league or type of competition is in the dataset
4.	Average age of players	Average age of players is an attribute that contains the average age of the players in the club and the name of the competition in the dataset
5.	Market value of the club in millions (£)	Market value of club in millions is an attribute that contains the market value (in millions) of the clubs listed in the 2021 football market value in the dataset
6.	Average market value of players in millions (£)	Average market value of players in millions is an attribute that contains the average market value of players (in millions) of clubs and competition names listed in the "top 100 football club market value 2021" in the data set
7.	Market value of top 18 players in millions (£)	Market value of top 18 players in millions is an attribute that contains the market value of the top 18 players (in millions) of the club and the name of the competition listed in the "top 100 football club market value 2021" in the dataset
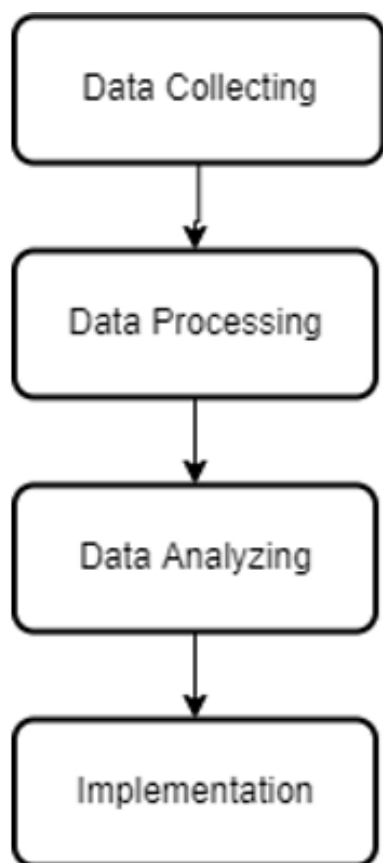
Fig. 1: Research stage (Anggraeni and Rizaldi, 2020; Madyatmadja *et al.*, 2021)

Data Analysis Technique

The technique used for data analysis that will be used in this research is descriptive. The two types of data in the research will be conducted in a way that is qualitative data and quantitative data. When this research is used quantitative data, which is in the form of numbers and is also the result of measurements and several calculations are carried out. The calculations carried out the will of course use the RapidMiner software and will use two methods in stages namely clustering and forecasting. The data will be grouped into several clusters and will be predicted. The proposed algorithm is K-means and Linear Regression, which then the results will be evaluated and validated the results. Later the level of accuracy of the data mining method will be measured based on the level of accuracy of the algorithm. K-means with Davies-Bouldin values and linear regression with Root Mean Squared Error (RMSE) and Squared Error (SE) values.

Results

This section will discuss the data set, namely the data attributes and descriptions along with their explanations, data grouping, the process of implementing the RapidMiner Studio application, forecasting with the RapidMiner, testing, and results.

The dataset in this study uses data from Kaggle entitled the top 100 football club market value. This data set has 102 agencies with data in the form of numbers and club names. The dataset has attributes or

features that describe each instance. The following is a list of attributes in this study and their explanations, Table 2.

Based on Table 1, the data types of the attributes used are described. most of the attributes are numeric, indicating that the data can be grouped and predicted based on a certain value. To run several algorithms in the RapidMiner studio application, it is necessary to import data.

When it is done to import data, it should also be noted that each column in the existing table will be assigned the correct attribute mark, so that if the wrong attribute is given, the result will be incorrect as well.

Process Implementation Using RapidMiner

Based on Fig. 2, there is a design model that is carried out systematically and according to the standards of the related algorithm, the researcher uses the k-means

algorithm in clustering. The entire process in the design model will run the K-means algorithm to get the required results.

When the Authors "click open" RapidMiner then in the repository table create a new repository and enter the market value data for the top 100 football clubs. Select the fetch operator then enter the dataset into the fetch operator in the fast miner design process field, after that use the select attribute to filter the attributes that will be used in the clustering process with this K-means method. The attributes used for this process include average player age, player average market value in millions, club market value in millions, top 18 player market value in millions, and team size. The selection of attributes can affect the result of the data mining process of this method.

**Table 3:** Result by using different algorithm model

| No.	Algorithm model	Result
1.	K-means	The results of the data have been grouped into 10 clusters; the Davies-Bouldin value obtained from the use of the K-means algorithm is -0.904. In addition, it is shown that the most clusters are cluster 3 with 23 items/club and the smallest clusters are cluster 2 and cluster 8 with 1 item/club
2.	Linear regression	The prediction results shown are based on a split ratio of 30 data samples, resulting in a description of the root mean squared error value of 25,041+/-0.00 and squared error of 627,060+/-1680,827. From the RMSE and SE results, it can be concluded that the prediction results from the algorithm used are strong and valid, this conclusion is in line with the statement that the smaller the RMSE value or the closer to zero, the more accurate the prediction results will be
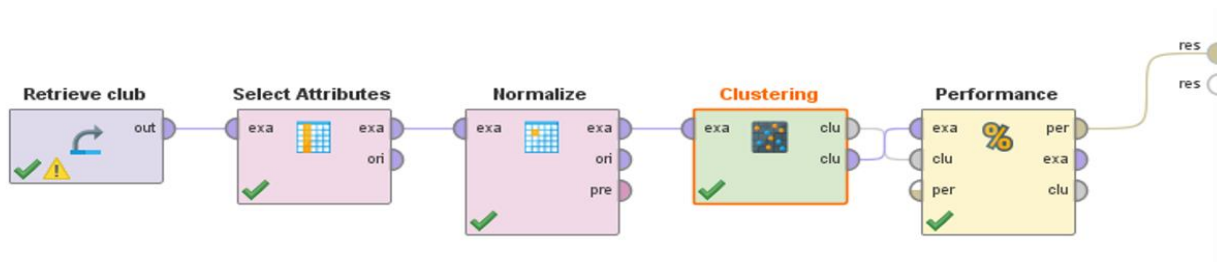
Fig. 2: Process to determine K-means algorithm using RapidMiner
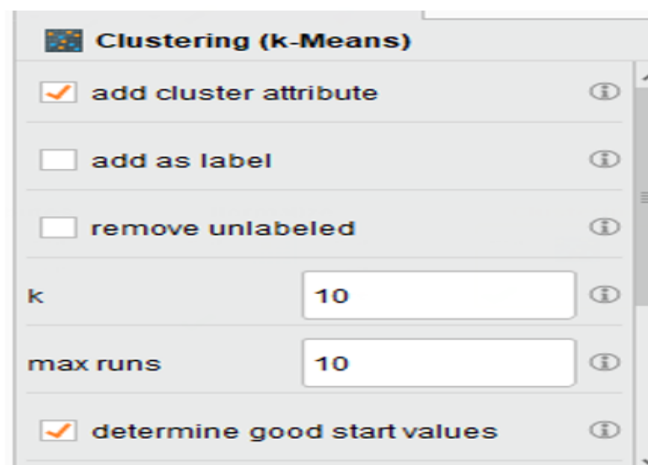
Fig. 3: K-means settings
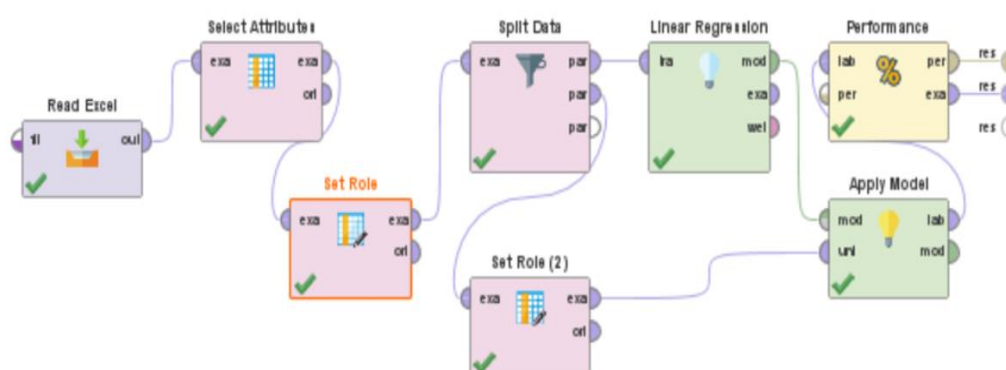
Fig. 4: Process to determine linear regression algorithm using RapidMiner
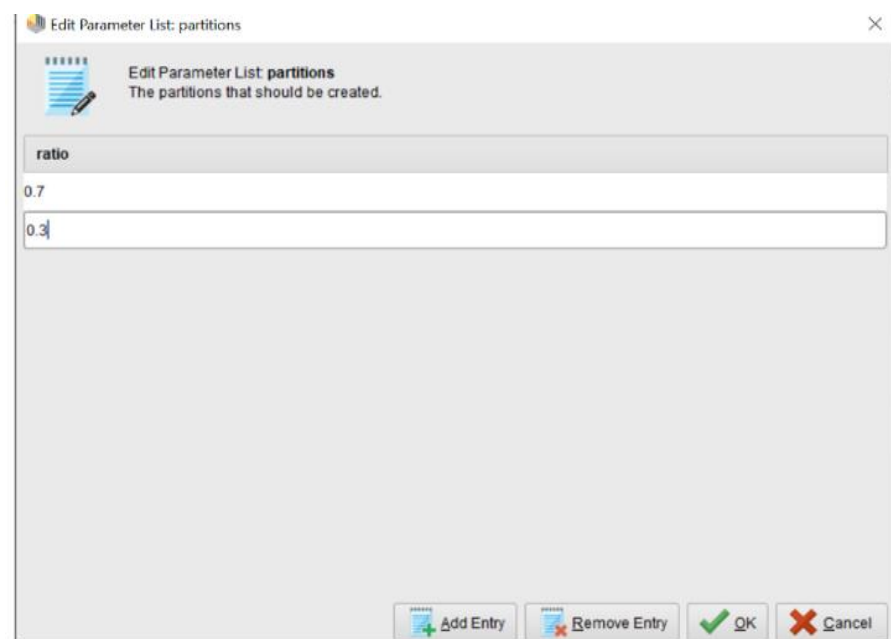
Fig. 5: Split data ratio

In the normalize operator, the number in the data will be smaller (simple) than a large number value which will complicate data mining. In addition, the normalization operator is used to scale the values to fit a specific range. use normalize will be applied to all previously selected data in the select attribute. Then the normalization method that the researcher uses for the data set is Z-transformation.

The attributes that the researcher uses and all the data will be divided into 10 clusters, each of which has a maximum run of 10 according to the settings installed on the K-means operator in Fig. 3. This operator performs clustering using the K-means algorithm by grouping examples together that are like each other. After grouping, the distance between the clusters can be measured.

Based on Fig. 4 this is the process of applying the data mining forecasting method using a linear regression algorithm.

The first step is to input the dataset through the read excel operator, this operator can directly set several attributes that have missing values by replacing them automatically based on the minimum or maximum value or with the average value for these attributes. Furthermore, the attributes used for this process include the average age of players, average market value of players in millions, market value of clubs in millions, market value of top 18 players in millions, and team size where the researcher chose the attribute market value of clubs in millions. The prediction will be made. Then the set role operator is connected which is used to determine the role of an attribute to be predicted. The researcher sets the attribute name, namely market value of club in millions with a role label because it will make predictions on these attributes.

The next step is to add a data split operator. This operator is used to separate data which aims to separate the data into two samples. The researcher sets a data set Fig. 5 with a comparison of 0.7 and 0.3, which means 0.3 is the sample data to be taken from the sampling technique while 0.7 is the remaining amount of data that is not sampled, so from 102 data are contained in the dataset, forecasting will be done based on the 30 selected data.

Next, connect the linear regression operator to connect the data with the algorithm used and continue to relate to the applied model operator. The apply model operator serves to test datasets from other operators and usually, this operator can also produce data predictions that are not involved or change data by applying a preprocessing model. The last stage is entering operator performance to get RMSE and SE levels to get a measure of forecasting accuracy.

The results of using different algorithms are shown in Table 3, namely data grouping with the K-means algorithm and data forecasting with the linear regression algorithm. The results obtained are in accordance with the standards of each algorithm, where the K-means algorithm produces the Davies-Bouldin value and the linear regression algorithm produces the root means squared error and squared error values so that the accuracy of the use of the algorithm can be proven to be strong and valid.

Discussion

Figure 6 shows the results of the detailed performance calculations below, namely:

- Avg. Within centroid distance serves to determine the distance from one cluster to another cluster
- Davies Bouldin serves to measure cluster validation on the clustering method. Davies Bouldin method produces a value of -0.904. The value is closer to zero, the better

The nominal values result shows the nominal values of the clustered data. Can be seen absolute count of each cluster. Here, there are 10 results from the clustering carried out which shows the results in the form of an index, nominal value, absolute count, and fraction. In addition, there is a fraction in each cluster. Fraction itself is the ratio of the data divided by the total is 1.

Figure 7 shows a data chart from the results of the k-means clustering. Each point shows a club with 100 items divided into several clusters. Each point has a color that indicates how many clusters. The club with the highest market value of club is in cluster 5. While in clusters 0, 9, 3, and 4, no club has a market value of club exceeding 0.25.

The results (example set) of the applied model. The predictions made for 30 examples (based on the data split ratio) on the market value of club in the millions attribute are marked with a green background.

The result of the performance of the implementation of the algorithm used. The description has a root mean squared error of 25,041+/- 0.00 and a squared error of 627,060+/-1680,827. From the results of the RMSE and SE, it can be concluded that the prediction results from the algorithm used are strong and valid, this conclusion is in line with the statement that the smaller the RMSE value or the closer to zero, the more accurate the prediction results will be.


```
PerformanceVector

PerformanceVector:
Avg. within centroid distance: -0.841
Avg. within centroid distance_cluster_0: -0.824
Avg. within centroid distance_cluster_1: -0.872
Avg. within centroid distance_cluster_2: -0.000
Avg. within centroid distance_cluster_3: -0.856
Avg. within centroid distance_cluster_4: -0.593
Avg. within centroid distance_cluster_5: -1.714
Avg. within centroid distance_cluster_6: -1.147
Avg. within centroid distance_cluster_7: -0.908
Avg. within centroid distance_cluster_8: -0.000
Avg. within centroid distance_cluster_9: -0.877
Davies Bouldin: -0.904
```

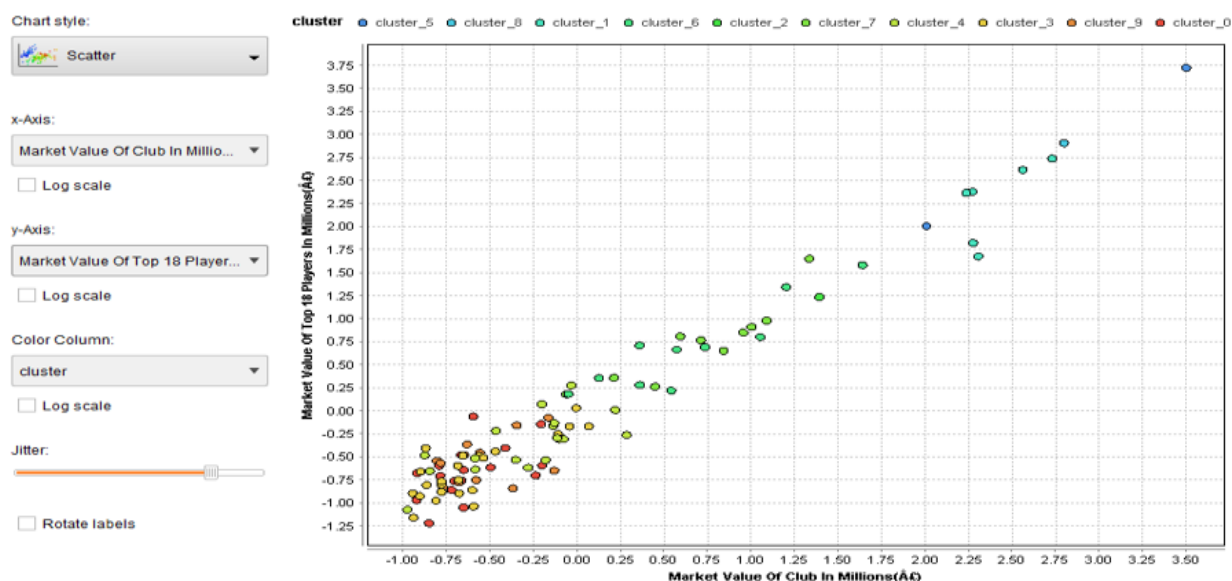**Fig. 6:** Performance vector cluster K-means

Fig. 7: Cluster K-means chart

Conclusion

Data have 7 attributes in this dataset. The attributes: Club name, competition name, squad size, average age of players, market value of club in millions (£), and market value of top 18 players in millions (£). In the clustering process, use the k-means algorithm in clustering. Next, selecting the attributes, the attributes include average player age, player average market value in millions, club market value in millions, top 18 player market value in millions, and team size.

Use normalize to make the data number simple. The last step is to use clustering to make all data divided to into 10 clusters. The results of the data have been grouped into 10 clusters; the Davies-Bouldin value obtained from the use of the K-means algorithm is -0.904. In addition, it is shown that the most clusters are cluster 3 with 23 items/club and the smallest clusters are cluster 2 and cluster 8 with 1 item/club. In the Forecasting Process, the first step use read excel to set several attributes that have missing values by replacing them automatically.

Next, selecting the attributes, the attributes include is average age of players, average market value of players in millions, market value of clubs in millions, market value of top 18 players in millions, and team size. Set the role of attributes, and set the attribute name, namely market value of club in millions with a role label because it will make predictions on these attributes. Split data into two samples using split data, set data with 0.3 is sample data to be taken from the sampling technique while 0.7 is the remaining amount or data that is not sampled. Using linear regression to predict the value of a variable on the value of another variable. Then using apply model to serve

the test datasets from other operators. The last step is entering operator performance to get RMSE and SE levels to get a measure of forecasting accuracy. The prediction results shown are based on a split ratio of 30 data samples, resulting in a description of the Root Mean Squared Error value of 25,041+/-0.00 and Squared Error of 627,060+/-1680,827. From the RMSE and SE results, it can be concluded that the prediction results from the algorithm used are strong and valid, this conclusion is in line with the statement that the smaller the RMSE value or the closer to zero, the more accurate the prediction results.

Author's Contributions

Johanes Fernandes Andry: Lead research project, coordinated developer, experimented, was an instructor, did data analysis and wrote the manuscript.

Riama Sibarani: Advise research project, design the experiment, data analysis, and wrote the manuscript.

Vlavio Nathanael Yefta: Advise research project, design the application, data analysis, wrote the manuscript, proofread.

Ethics

This article is original and unpublished. Correspondence authors confirm that all other authors have read and agree that the manuscript does not involve ethical issues.

References

Acs, B., & Toka, L. (2022). A career in football: what is behind an outstanding market value? In *International Workshop on Machine Learning and Data Mining for Sports Analytics* (pp: 15-25). Springer, Cham. https://doi.org/10.1007/978-3-031-02044-5_2

Adiwiyana, H. I., & Harymawan, I. (2021). Factors that Determine the Market Value of Professional Football Players in Indonesia. *Jurnal Dinamika Akuntansi*, *13*(1), 51-61. https://journal.unnes.ac.id/nju/index.php/jda/article/view/26079

Al-Radaideh, Q. A., Assaf, A. A., & Alnagi, E. (2013, December). Predicting stock prices using data mining techniques. In *the International Arab Conference on Information Technology (ACIT'2013)* (pp: 1-8). http://www.acit2k.org/ACIT/2013Proceedings/163.pdf

Allouche, J., & Soulez, S., (2008). Determinants of share price variations of listed football clubs: Empirical evidence from English football leagues, In book: Myths and facts about football: The economics and psychology of the world's greatest sport chapter: 20 Publisher: *Cambridge Scholars Press*.

Amir, E., & Livne, G. (2005). Accounting, valuation and duration of football player contracts. *Journal of Business Finance & Accounting*, *32*(3-4), 549-586. https://doi.org/10.1111/j.0306-686X.2005.00604.x

Anggraeni, D., & Rizaldi, R. (2020, July). Rapidminer's Implementation in analyzing student mainstream data. In *International Conference on Social, Sciences and Information Technology*, *1*(1), 1-6. https://doi.org/10.33330/icossit.v1i1.819

Balusamy, B., Kadry, S., & Gandomi, A. H. (2021). *Big Data: Concepts, Technology and Architecture*. John Wiley & Sons.

Bryant, A., & Cios, K. (2017). RNN-DBSCAN: A density-based clustering algorithm using reverse nearest neighbor density estimates. *IEEE Transactions on Knowledge and Data Engineering*, *30*(6), 1109-1121. https://doi.org/10.1109/TKDE.2017.2787640

Chitra, K., & Subashini, B. (2013). Data mining techniques and its applications in banking sector. *International Journal of Emerging Technology and Advanced Engineering*, *3*(8), 219-226.

Cheffins, B. R. (1999). Playing the Stock Market: "Going Public" and Professional Club Sports. *The Journal of Corporation Law*. *24*(3), 641-684.

Collignon, H. & Sultan, N. (2014). Winning in Business Sports. AT Keaney Report. https://www.kearney.com/telecommunications/article/-/insights/winning-in-the-business-of-sports

Deloitte. (2018). Annual review of football finance 2018. *Retrieved from*. https://www2.deloitte.com/bd/en/pages/about-deloitte/articles/annual-review-of-football-finance.html

Duque, J., & Ferreira, N. A. (2005). Explaining share price performance of football clubs listed on the Euronext Lisbon. *ISEG-Universidade Tecnica de Lisboa Business Administration Working Paper*, (05-01). http://doi.org/10.2139/ssrn.675633

Gharehchopogh, F. S., Bonab, T. H., & Khaze, S. R. (2013). A linear regression approach to prediction of stock market trading volume: A case study. *International Journal of Managing Value and Supply Chains*, *4*(3), 25. https://doi.org/10.5121/ijmvsc.2013.4303

Gharehchopogha, F. S., & Khalifehlou, Z. A. (2012). A new approach in software cost estimation using a regression-based classifier. *Global Journal on Technology*, *2*.

Ghavami, P. (2016). Big Data Governance: Modern Data Management Principles for Hadoop, NoSQL & Big Data Analytics. *Washington, DC*.

Hagen, J., & Cunha, M. N. (2019). The history of investing in football and factors affecting the stock price of listed football clubs. *International Journal of Financial Management*, *9*(4), 31-41. http://publishingindia.com/ijfm/

Jain, N., & Srivastava, V. (2013). Data mining techniques: a survey paper. *IJRET: International Journal of Research in Engineering and Technology*, *2*(11), 2319-1163.

Zernik, J. (2010). Data Mining as a Civic Duty-Online Public Prisoners' Registration Systems. *International Journal on Social Media: Monitoring, Measurement, Mining*, *1*(1), 84-96.

Korcheva, A. (2021). Big Data and Data Mining. *Journal: Encyclopedia of Sustainable Management*, 1-3. https://doi.org/10.1007/978-3-030-02006-4_1014-1

Maksood, F. Z., & Achuthan, G. (2016). Analysis of data mining techniques and its applications. *International Journal of Computer Applications*, *140*(3), 6-14.

Metisen, B. M., & Sari, H. L. (2015). Analisis clustering menggunakan metode K-means dalam pengelompokkan penjualan produk pada Swalayan Fadhila. *Jurnal Media Infotama*, *11*(2). https://doi.org/10.37676/jmi.v11i2.258

Madyatmadja, E. D., Sembiring, D. J. M., Angin, S. M. B. P., Ferdy, D., & Andry, J. F. (2021). Big data in educational institutions using RapidMiner to predict learning effectiveness. *Journal of Computer Science*, *17*(4), 403-413.

Muhammad, S. (2017). Pengaruh performance, umar, dan cost of input terhadap market value pemain sepakbola profesional (studi empiris pada klub sepakbola yang berkompetisi dalam Liga Spanyol pada periode 2014/2015). https://dspace.uii.ac.id/123456789/27762

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*." O'Reilly Media, Inc. "https://www.kaggle.com/kimjihoo/coronavirusdataset

Russom, P. (2013). Managing big data. *TDWI Best Practices Report, TDWI Research*, *1*, 40.

UEFA. (2019). UEFA report details European football's journey to profitability, in web: https://www.uefa.com/insideuefa/news/024d-0f8e6732b0e4-eb06c43d2edb-1000--uefa-report-details-european-football-s-journey-to-profitabilit/

Zakir, J., Seymour, T., & Berg, K. (2015). Big data analytics. *Issues in Information Systems*, *16*(2). https://doi.org/10.48009/2_iis_2015_81-90