Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol 7, No.1, March 2024, pp. 104 – 109
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Improving K-Means Clustering Accuracy for Academic Success Investigation With Extreme Gradient Boosting Algorithm

[1]Irma Darmayanti, [2]Laily Farkhah Adhimah, [3]Rizki Sadewo, [4]Nurul Hidayati, [*5]Pungkas Subarkah
[1,4]Departement of Information Technology, Universitas Amikom Purwokerto
[2,5]Departement of Informatics, Universitas Amikom Purwokerto
[3]Departement of Information System, Universitas Amikom Purwokerto
Email: [1]irmada@amikompurwokerto.ac.id, [2] lailyfarkhaha@gmail.com, [3]rizkisadewo00@gmail.com, [4]nurulhid371@gmail.com, [5]subarkah@amikompurwokerto.ac.id

Article Info
Article history:
Received Nov 15th, 2023
Revised Jan 28th, 2024
Accepted Mar 4th, 2024

Keyword:
Academic
Algorithm
Education
Extreme Gradient Adaboost
K-Means

ABSTRACT
Human Resources (HR) has a very important role in the development of the nation, so to improve the quality of human resources, education is needed. Education has a role in developing science, disseminating, socializing, and applying it. So that education is one of the important factors in advancing a nation. However, there are still many challenges in achieving quality education, especially in developing countries such as Indonesia, such as parental education level, socioeconomic status, and environmental conditions can also affect the quality of education and students' opportunities for academic success. The research methods used in this research are problem identification, data collection, data analysis, and evaluation. The results in this study are an increase in accuracy of 38.55% from the difference in the K-Means accuracy value of 14% resulting from the David Bounded Index and the use of the extreme gradient adaboost algorithm.

Corresponding Author:
Pungkas Subarkah,
Departement of Informatics,
Amikom Purwokerto University,
Jl. Letjend Pol. Soemarto No.127, Watumas, Purwanegara, Kec. Purwokerto Utara,
Banyumas Regency, Central Java.
Email: subarkah@amikompurwokerto.ac.id
DOI: http://dx.doi.org/10.24014/ijaidm.v7i1.26657

1. INTRODUCTION

Humans learn not to pursue grades but to prepare for a better life and in line with the thoughts of Theodore Meyer Greene "Education is an effort to prepare human resources to achieve a meaningful life"[1]. Human Resources (HR) has a very important role in the development of the nation, so to improve the quality of human resources, education is needed. Education has a role in developing science, disseminating, socializing, and applying it. So that education is one of the important factors in advancing a nation.

A quality education can improve academic success. And academic success is seen as a reflection of a person's qualities in intellectual ability, perseverance, and adaptability [2]. Academic success plays an important role in the lives of individuals and society as a whole. For individuals, academic success can provide access to opportunities such as scholarships, better jobs, and the possibility of further study. Academic success can also improve self-confidence, analytical skills, and problem-solving skills. On the community side, academic success can contribute to building quality human resources. Qualified human resources will be able to increase the productivity and competitiveness of a country in the global market. Academic success can also help in overcoming social and economic problems such as poverty and unemployment [3].

However, there are still many challenges in achieving quality education, especially in developing countries like Indonesia. Some of the challenges in Indonesia are limited access to education, disparities between regions, inequality in the quality of education, and problems with the availability of skilled human resources in the field of education.

In the growing digital era, analytical and decision-making capabilities are relied upon in classifying or predicting. So that it can help make it easier for an agency to make policies. Many studies have been conducted to perform classification and prediction [4][5]. One data analysis technique that is often used is data mining, which is the process of extracting useful information from large and complex data.

Some of the research that was attempted included the initial research attempted by Gustientiedina et al with the title "Application of the K-Means Algorithm for Clustering Drug Information at Pekanbaru Regional Hospital". The research attempted to implement clustering in information mining can be used to analyze drug usage, planning and drug control in hospitals. The procedure to be used for clustering drug information is the K-Means algorithm which aims to group drug information in the Pekanbaru Regional Universal Hospital which can be used as a reference in decision making in planning and controlling medical supplies in hospitals. The clustering results that the group of drugs listed with little consumption has an average annual demand for drugs of less than 18000 pieces, and the drugs listed with more consumption have an average annual demand for drugs between 18000-70000 pieces, on the other hand, the drugs listed in the group of drugs with large consumption have an average annual demand for drugs above 70000 pieces. From the results of the cluster analysis, it is necessary to try to increase the accuracy value to make it more valid by setting the best centroid value [6].

Furthermore, research with the title "Comparison of SVM, Random Forest and XGBoost Algorithms for Determining Credit Application Approval". With the aim of achieving the right and comfortable credit granting Credit analysis is an observation to view the feasibility of a credit problem. From this analysis will be known the feasibility of credit recipients. This research uses the CRISP-DM methodology which consists of 6 stages, namely Business Understanding, Information Understanding, Information preparation, Modeling Evaluation, and Deployment by practicing classification procedures by equating SVM, Random Forest, and XGBoost algorithms. This research uses open source datasets obtained from Kaggle. The results of research using SVM, random forest, and XGBoost algorithms obtained the highest accuracy, recall, precision values in the XGBoost model with an accuracy value of 82%, recall 70%, and precision 92% [7].

Finally, research conducted by Yuliansyah Ibrahim in 2022, with the title "K-Means Clustering and Extreme Gradient Boosting (Xg boost) for Point Prediction in Fantasy Premier League Games". Combining the K-Means unsupervised learning model with the Extreme Gradient Boosting (XGBoost) supervised learning model. The data used is four seasons of Fantasy Premier League data starting from the 2017 season to the 2020 season. The collected data is processed and aggregated, then combined into one whole dataset. Based on the experiments conducted, it was found that combining K-Means Clustering with XGBoost did not improve the performance of the model [8]. The advantage of the k-means algorithm is in rechecking the quality of the cluster at each iteration [9]–[11], allowing changes in the number of clusters to meet the validity of the cluster quality so as to produce optimal clusters [12]–[14]

Based on the description that has been presented above, the research that will be carried out aims to cluster academic success using student techniques with Extreme Gradient Boosting (XGBoost). One of the methods used to analyze academic data is by using classification. Classification aims to group data into certain categories or classes based on the attributes or features possessed by the data [15]–[17] Based on the description that has been presented above, the research that will be carried out aims to cluster academic success using student techniques with Extreme Gradient Boosting (XGBoost). One of the methods used to analyze academic data is by using classification. Classification aims to group data into certain categories or classes based on the attributes or features possessed by the data [18].

2. RESEARCH METHOD

The following is the research flow that will be applied to research on the success rate of students using XGBoost, can be seen in Figure 1.

Figure 1 shows the research flow, The following is the result of the explanation:

2.1. Problem Identification

The stage where the researcher searches and reviews it. So that it can sharpen existing problems, look for support for facts, information, theories in determining the theoretical basis or framework, know for sure whether the problem has been studied before or not and facilitate the completion of research and explain the limitations of the research to be carried out. The limitation of this research is that researchers will only perform data calculations to produce a prediction model from the Indonesian State data set.

2.2. Data Collection

This stage involves a meticulous data selection process aimed at obtaining clean and research-ready data. The data pre-processing stage encompasses attribute identification and selection, addressing missing attribute values or incomplete attributes (handling missing values), and discretizing the value process.

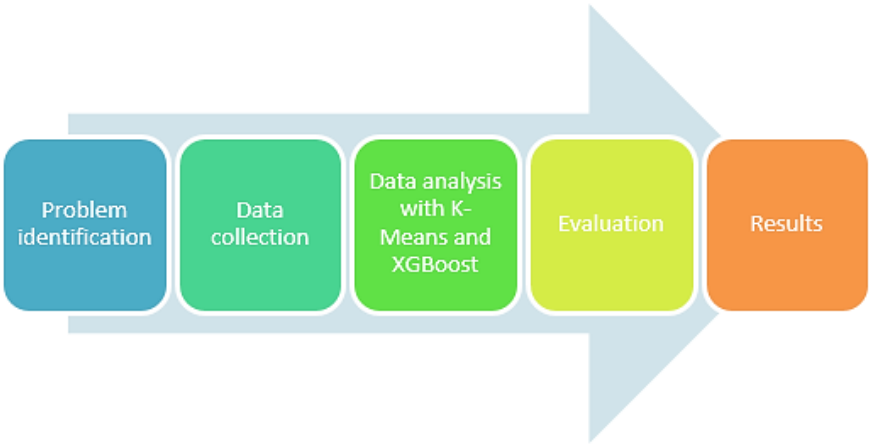

Figure 1. Research Flow

2.3. Data analysis with K-Means and XGBoost

Perform the calculation process using K-Means combined with XGBoost Algorithm. The following is an explanation of each algorithm used:

1. Kmeans Algorithm

The Kmeans algorithm is a umpteenth method that divides illustrations into K non-overlapping, predefined partitions with each illustration listed in just one partition[19], [20]. This algorithm attempts to protect intra-cluster illustrations as similar as possible while protecting that clusters are always distinct. It distributes point information to the clusters so that the sum of the squared distances in between them as well as the cluster centers is as small as possible Within the clusters, the variance can be minimal thus creating more similar clusters[21]–[23]. The process of a K-Means algorithm can be seen as follows:

a. Determine how many clusters you want to assign cluster center k to.
b. Using euclidean distance and then calculating each data to the cluster center. Here is the formula euclidean distance:

$$D\ (p,c)_n=\sqrt{\sum_{i=0}^{n}(pi-ci)^2} \quad (1)$$

c. Categorize the data into clusters with the shortest distance by using the equation

$$\text{Min}\ \sum_k^k = d\ ^{ik}\sqrt{\sum_j^m (Cij\ -\ Cik)} \quad (2)$$

d. Calculate the cluster center using the equation

$$C_{kj}\frac{\sum_{i=j}^{p} Xij}{p} \quad (3)$$

e. Please repeat steps two through four so that there is no more data moving to other clusters.

2. XGBoost Algorithm

XGBoost, a scalable tree boosting system, was proposed been widely used in Kaggle's Higgs sub-signal recognition contest. More recently, it has attracted wide attention due to its outstanding efficiency and high prediction accuracy. In fact, XGBoost is an improved GBDT algorithm[24]. The following is the equation for the XGBoost Algorithm:

$$W_j= -\frac{\sum_{i\epsilon Ij}\ Gi}{\sum_{i\epsilon Ij} Hi+\lambda} \quad (4)$$

2.4. Evaluation

This The testing process with Root Mean Squared Error (RMSE) to determine the level of accuracy of the calculation results of K-Means clustering and XGBoost Algorithm to increase accuracy in clustering student academic success.

3. RESULTS AND ANALYSIS

3.1. Data Collection

The Student Academic Performance dataset is obtained from the Kaggle site with details there are 480 data, has 16 attributes[25]. Below is the arrangement of attributes on the dataset used, can be seen in the table 2.

Table 2. Data Academic Dataset

No.	Atribute name	Description
1	Gender	Student's gender (nominal: 'Male' or 'Female')
2	Nationality	Student's nationality (nominal: 'Kuwait','Lebanon','Egypt','SaudiArabia','USA','Jordan', Venezuela','Iran','Tunis','Marocco','Syria','Palestina','Iraq','Lybia')
3	Place of birth	Student's Place of birth (nominal: 'Kuwait','Lebanon','Egypt','SaudiArabia','USA','Jordan', Venezuela','Iran','Tunis','Marocco','Syria','Palestina','Iraq','Lybia')
4	Educational stages	Educational level student belongs (nominal: 'lowerlevel','MiddleSchool','HighSchool')
5	Grade Levels	Grade student belongs (nominal: 'G-01','G-02','G-03','G-04','G-05','G-06','G-07','G-08','G-09','G-10','G-11','G-12')
6	Section ID	Classroom student belongs (nominal:'A','B','C')
7	Topic	Course topic (nominal:'English','Spanish','French','Arabic','IT','Math','Chemistry','Biology', 'Science','History','Quran','Geology')
8	Semester	School year semester (nominal:'First','Second')
9	Parent responsible for student	Nominal:'mom','father
10	Raised hand	How many times the student raises his/her hand on classroom (numeric:0-100)
11	Visited resources	How many times the student visited a course content (numeric:0-100)
12	Viewing announcements	How many times the student check the new announcements (numeric:0-100)
13	Discussion groups	How many times the student participate on discussion groups (numeric:0-100)
14	Parent Answering Survey	Parent Answered the survey which are provided from school or not (nominal:'Yes','No')
15	Parent School Satisfaction	The Degree of parent satisfaction from school (nominal:'Yes','No')
16	Student Absence Day	The number of absence days for each student (nominal:above-7. under-7)

3.2. Data analysis with K-Means and XGBoost

The next process is the analysis with K-Means and Adaboost, data selection is carried out on the Student Absence attribute. Thus, the data needed in the clustering stage is obtained. In the process of determining. This research analyzes the cluster by forming 3 trials, namely by applying K = 0, into K=9. The three trials were applied to each algorithm, namely K-Means and AdaBoost. In the results and evaluation stage, the best K value for the clustering process is determined by using the Davies Bouldin Index (DBI) method. After knowing the DBI value of each cluster, then the smallest DBI value shows the best cluster results and shows the optimal number of clusters optimal number of clusters. The following are the results of K=0 into K=9 DBI on the K-Means algorithm.

Table 3. DBI value in each cluster

Cluster	Items	Value DBI (%)
0	55	11
1	45	9
2	52	11
3	42	9
4	38	8
5	54	11
6	39	8
7	42	9
8	47	10
9	66	14

From the table 3 regarding the number of members of each cluster referring to the DBI value, K = 9 was chosen, with a DBI value of 14%. Furthermore, the results in using the extreme gradient adaboost algorithm get an accuracy of 52.55% with details of Precision, Recall and F-Measure as figure 2.

Precision	Recall	F-Measure
0,478	0,578	0,524
0,607	0,913	0,730
0,412	0,099	0,159
0,493	0,525	0,470

Figure 2. Confusion matrix

3.3. Evaluation

This evaluation stage on the use of K-Means algorithm using DBI which gets the highest accuracy value of 14%, while on the use of extreme gradient boosting algorithm gets an accuracy value of 52.55%. So that there is an increase in accuracy of 38.55% on the dataset.

4. CONCLUSION

After going through the results and discussion stages of both the K-Means algorithm and the extreme boosting algorithm. That the use of K-Means only gets 14% accuracy from DBI and the use of the extreme gradient adaboost algorithm gets an accuracy value of 52.55%. Thus there is an increase in accuracy of 38.55%.

REFERENCES

[1] J. Hadi, "Meneropong Pendidika Indonesia," *Kompasiana*, 2023. [Online]. Available: https://www.kompasiana.com/hadiramadhan3129/63d0791bc3bdbf2fdc174932/meneropong-pendidikan-indonesia.

[2] L. A. Fitriyah, A. W. Wijayadi, and N. Hayati, "Efikasi Diri, Kestabilan Emosi dan Keberhasilan Akademik Mahasiswa Dalam Perkuliahan," *DWIJA CENDEKIA J. Ris. Pedagog.*, vol. 4, no. 1, p. 44, 2020.

[3] R. D. Aprilia, "Pengaruh Pertumbuhan Ekonomi, Upah Minimum, Pendidikan Dan Tingkat Pengangguran Terhadap Tingkat Kemiskinan," *J. Ilm. Mhs. Fak. Ekon. dan Bisnis*, pp. 1–19, 2016.

[4] I. Darmayanti, P. Subarkah, L. R. Anunggilarso, and J. Suhaman, "Prediksi Potensi Siswa Putus Sekolah Akibat Pandemi Covid-19 Menggunakan Algoritme K-Nearest Neighbor," *J. Sains dan Teknol.*, vol. 10, no. 2, pp. 230–238, 2021.

[5] J. Silva, N. Varela, L. A. B. López, and R. H. R. Millán, "Association rules extraction for customer segmentation in the SMES sector using the apriori algorithm," *Procedia Comput. Sci.*, vol. 151, no. 2018, pp. 1207–1212, 2019.

[6] G. Gustientiedina, M. H. Adiya, and Y. Desnelita, "Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan Pada RSUD Pekanbaru," *J. Nas. Teknol. dan Sist. Inf.*, vol. 5, no. 1, pp. 17–24, 2019.

[7] M. R. Givari, M. R. Sulaeman, and Y. Umaidah, "Perbandingan Algoritma SVM, Random Forest Dan XGBoost Untuk Penentuan Persetujuan Pengajuan Kredit," *Nuansa Inform.*, vol. 16, no. 1, pp. 141–149, 2022.

[8] B. A. B. Iii and M. Penelitian, "Yuliansyah Ibrahim, 2022 K-MEANS CLUSTERING DAN EXTREME GRADIENT BOOSTING (XGBOOST) UNTUK PREDIKSI POIN DI GAME FANTASY PREMIER LEAGUE Universitas Pendidikan Indonesia | repository.upi.edu | perpustakaan.upi.edu," pp. 36–44, 2022.

[9] P. Zeng, F. Sun, Y. Liu, Y. Wang, G. Li, and Y. Che, "Mapping future droughts under global warming across China: A combined multi-timescale meteorological drought index and SOM-Kmeans approach," *Weather Clim. Extrem.*, vol. 31, p. 100304, 2021.

[10] Y. Chen, F. M. Zahedi, A. Abbasi, and D. Dobolyi, "Trust calibration of automated security IT artifacts: A multi-domain study of phishing-website detection tools," *Inf. Manag.*, vol. 58, no. 1, p. 103394, 2021.

[11] H. G. Costa, M. H. T. da Silva, G. N. Santos, A. Bonamigo, and R. D. Callado, "Clustering Brazilian Public Emergency Healthcare Units," *IFAC-PapersOnLine*, vol. 55, no. 10, pp. 566–571, 2022.

[12] K. Ariasa, I. G. A. Gunadi, and I. M. Candiasa, "Optimasi Algoritma Klaster Dinamis pada K-Means dalam Pengelompokkan Kinerja Akademik Mahasiswa (Studi Kasus: Universitas Pendidikan Ganesha)," *J. Nas. Pendidik. Tek. Inform. JANAPATI*, vol. 9, no. 2, pp. 181–193, 2020.

[13] H. Asri, "Big Data and IoT for real-time miscarriage prediction A clustering comparative study," *Procedia Comput. Sci.*, vol. 191, pp. 200–206, 2021.

[14] A. Nowak-Brzezinska and C. Horyn, "Outliers in rules - The comparision of LOF, COF and KMEANS algorithms.," *Procedia Comput. Sci.*, vol. 176, pp. 1420–1429, 2020.

[15] A. A. Abdulnassar and L. R. Nair, "Performance analysis of Kmeans with modified initial centroid selection algorithms and developed Kmeans9+ model," *Meas. Sensors*, vol. 25, no. December 2022, p. 100666, 2023.

[16] J. Li *et al.*, "Hierarchical and partitioned planning strategy for closed-loop devices in low-voltage distribution network based on improved KMeans partition method," *Energy Reports*, vol. 9, pp. 477–485, 2023.

[17] J. Dogra, S. Jain, and M. Sood, "Segmentation of MR Images using Hybrid kMean-Graph Cut Technique," *Procedia Comput. Sci.*, vol. 132, pp. 775–784, 2018.

[18] A. Solichin and K. Khairunnisa, "Klasterisasi Persebaran Virus Corona (Covid-19) Di DKI Jakarta Menggunakan Metode K-Means," *Fountain Informatics J.*, vol. 5, no. 2, p. 52, 2020.

[19] M. Cui, "Introduction to the K-Means Clustering Algorithm Based on the Elbow Method," pp. 5–8, 2020.

[20] M. Ahmed, R. Seraj, S. Mohammed, and S. Islam, "The k-means Algorithm : A Comprehensive Survey and Performance Evaluation," pp. 1–12, 2020.

[21] K. P. Sinaga and M. Yang, "Unsupervised K-Means Clustering Algorithm," vol. 8, 2020.

[22] A. F. Jahwar, "META-HEURISTIC ALGORITHMS FOR K-MEANS CLUSTERING : A REVIEW," vol. 17, no. 7, pp. 1–20, 2021.

[23] M. Jahangoshai, M. Eshkevari, M. Saberi, and O. Hussain, "Knowledge-Based Systems GBK-means clustering algorithm : An improvement to the K-means algorithm based on the bargaining game," *Knowledge-Based Syst.*, vol. 213, p. 106672, 2021.

[24] K. Song, F. Yan, T. Ding, L. Gao, and S. Lu, "A steel property optimization model based on the XGBoost algorithm and improved PSO," *Comput. Mater. Sci.*, vol. 174, no. December 2019, p. 109472, 2020.

[25] I. Aljarah, "Students' Academic Performance Dataset," *Kaggle*, 2015. [Online]. Available:

https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data/data. [Accessed: 30-Jul-2023].

BIBLIOGRAPHY OF AUTHORS

Irma Darmayanti was born in Purwokerto, May 6, 1991. she holds a Master's degree in Informatics Engineering and currently works as a lecturer at Amikom Purwokerto University since 2020 in the Information Technology study program. Scientific fields of Programming and Data Scientist

Laily Farkhah Adhimah is a 7th semester student majoring in Informatics, Faculty of Computer Science at Amikom Purwokerto University. Home address in Terban Village, Warungasem district, Batang Regency. Hobbies Reading.

Rizki Sadewo is a 7th semester student, information systems study program, Faculty of Computer Science, Amikom University, Purwokerto. Lives in Watumas Hamlet, Purwanegara, North Purwokerto, Banyumas district. Her hobbies are dancing and singing

Nurul Hidayati is a 5th semester student manjoring in Information Technology, Faculty of Computer Science at Amikom Purwokerto University. Home address in Kedungpring Village, Kemranjen District, Banyumas Regency.

Pungkas Subarkah is a lecturer at the Informatics Study Program, Amikom Purwokerto University. The author's education continued his S-1 Study at Amikom Purwokerto University and S-2 Master of Informatics Engineering at Amikom University Yogyakarta. Research interests in Data Mining, Machine Learning, and Information System