ComTech: Computer, Mathematics and Engineering Applications, 15(2), December 2024, 101−108 P-ISSN: 2087-1244
DOI: 10.21512/comtech.v15i2.11948 E-ISSN: 2476-907X

Association Analysis Using Apriori Algorithm of GANs-Expanded Student Performance Dataset

Rannie M. Sumacot*

Department of Public Administration, Faculty of Governance and Development Studies,
Southern Leyte State University
Southern Leyte, Philippines 6600
rsumacot@southernleytestateu.edu.ph

Received: 24th July 2024/ **Revised**: 5th November 2024/ **Accepted**: 5th November 2024

How to Cite: Sumacot, R. M. (2024). Association Analysis Using Apriori Algorithm of GANs-Expanded Student Performance Dataset. *ComTech: Computer, Mathematics and Engineering Applications, 15*(2), 101−108. https://doi.org/10.21512/comtech.v15i2.11948

Abstract **-** Traditional datasets are often limited, which can affect the accuracy of analyses. Additionally, the use of students' real data raises privacy concerns. Generative Adversarial Networks (GANs) offer a solution by generating synthetic data that closely mirrors real-world data without compromising sensitive information. The research explored the application of GANs to enhance student performance datasets by addressing challenges related to data scarcity and privacy in educational research. In the research, GANs were utilized to generate synthetic student performance data. The accuracy of the data was assessed using Mean Absolute Percentage Error (MAPE), with values ranging from 0.004% to 19.92% across various statistical measures and means. These results demonstrated the reliability of the synthetic data, making it suitable for further analysis. The synthetic datasets were then analyzed using the Apriori Algorithm, a well-known method in data mining for discovering significant patterns and relationships. A lower bound minimum support of 0.1 (10%) and a minimum confidence threshold of 0.6 (60%) were applied, ensuring the identification of meaningful associations. The analysis reveals important patterns and relationships among student attributes and behaviors. The research highlights the potential of GANs to advance data-driven educational research. By generating high-quality synthetic data, GANs allow researchers to conduct comprehensive analyses while addressing privacy concerns. The research contributes to the methodological approach to data augmentation in education, offering new opportunities for ethical and robust research.

Keywords**:** association analysis, Apriori algorithm, Generative Adversarial Networks (GANs), student performance dataset

I. INTRODUCTION

The field of data mining has experienced significant advancements with the advent of intelligent data analysis and the development of smart and automated applications primarily driven by machine learning algorithms (Sarker, 2021; Wu et al., 2021). The researcher aims to delve into the utilization of Generative Adversarial Networks (GANs) to analyze student performance through an association analysis approach.

GANs consist of two neural networks, the generator and the discriminator that are trained through an adversarial process. The generator produces synthetic data from random noise, while the discriminator evaluates the authenticity of the data, distinguishing between real samples from the training dataset and the generated samples. During training, the generator aims to create increasingly realistic data to fool the discriminator. In turn, it improves its ability to identify fake data. This process continues until the generator produces data which are indistinguishable from real data (Goodfellow et al., 2014).

GANs have gained recognition as a potent tool for creating synthetic data, particularly in situations where data scarcity or privacy concerns hinder data collection (Figueira & Vaz, 2022). Previous researchers have found successful applications in various domains, such as generating high-quality images (Goodfellow et al., 2020; Karras et al., 2020) and medical informatics (Choi et al., 2017). Additionally, numerous studies have underscored GANs' ability to generate authentic synthetic data (Pan et al., 2019). The potential applications of GANs in education are significant yet largely underexplored. Their ability to generate synthetic student performance data can facilitate various analyses while ensuring the

*Corresponding Author

protection of student privacy (Wang & Yeung, 2016).

Currently, student performance analysis predominantly relies on traditional statistical methods and machine learning techniques, which often require extensive real-world data. Such data may be difficult to obtain due to privacy concerns or logistical challenges. Therefore, the urgent need for comprehensive educational datasets can be addressed through GANs, which can produce synthetic datasets that closely mirror actual student performance data (Goyal & Mahmoud, 2024; Ramzan et al., 2024; Rather & Kumar, 2024).

The research aims to bridge this gap by employing GANs-enhanced datasets to support association analysis of student performance and using the Apriori algorithm to identify patterns and associations within the dataset. The Apriori algorithm is a popular data mining technique that also analyzes patterns in linguistic data, highlighting its versatility in different fields. The Apriori algorithm stands as the pioneering association rule mining algorithm (Gan et al., 2024; Ye, 2020). It employs support-based pruning techniques to manage the exponential proliferation of candidate sets. This algorithm holds a paramount position as the most influential method for mining frequent Boolean itemsets. According to Li et al. (2021), association rules, serving as indicators of interdependence and correlations among various elements, constitute a pivotal research methodology within the domain of graphic pattern data mining. Similarly, educational data mining is integrated with deep neural networks and the Apriori algorithm to predict student performance and major selection (Ouassif & Ziani, 2024). An association rule represents a practical and straightforward knowledge model derived from quantified data that uncovers relationships among valuable data items within extensive datasets. These previous studies highlight the versatility and potential of the Apriori algorithm in various fields. The goal is to unveil concealed patterns and connections that can furnish valuable insights into the factors influencing student performance. This approach is innovative as it harnesses GANs' capabilities to enhance traditional data mining techniques, potentially yielding more robust and dependable outcomes (Liu et al., 2015).

In conclusion, the research aspires to augment the existing knowledge base by investigating an innovative approach to student performance analysis using GANs-enhanced datasets. It is anticipated that the findings will offer valuable insights for educators and policymakers. It can aid them in making well-informed decisions to enhance students' outcomes.

II. METHODS

The research is conducted by studying literature. At this stage, the literature study is carried out by looking for reference materials in the form of books, journals, final assignments, theses, and the Internet in accordance with the discussed issues.

The research applies a quantitative approach to investigate associations between various factors and student performance. The dataset, sourced from Arvidsson (2023), includes 28 variables related to students' demographics, academic behaviors, family background, and extracurricular activities, with data types ranging from categorical to numerical. To augment the dataset comprehensively, the researcher uses GANs, an advanced artificial intelligence technique within the deep learning domain, to generate synthetic data. This innovative process is executed through Google Colab, a cloud-based platform known for supporting the creation and execution of Python code (Google, n.d.).

Following data collection, preprocessing is conducted to ensure data quality. This step involves handling missing values, binning, and normalizing the data. For the GANs-generated synthetic data, additional validation is performed to ensure consistency with the real-world data. The quality of the synthetic data is assessed using Mean Absolute Percentage Error (MAPE) calculations. It is a metric that measures the accuracy of data replication as a percentage. MAPE measures prediction accuracy by calculating the ratio of the sum of absolute prediction errors to the sum of actual values. A lower MAPE value signifies a more accurate model (Tan et al., 2024). This metric is employed to evaluate the accuracy of the synthetic data in replicating the statistical properties of the original dataset, ensuring its reliability for further analysis. The formula for MAPE is in Equation (1). It has *MAPE* as the Mean Absolute Percentage Error, n as the number of data points or observations, A_t as the actual value at time t, and P_t as the predicted or forecasted value at time t.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - P_t}{A_t}\right| \text{x } 100 \quad (1)$$

Equation (1) computes the absolute percentage error for each data point. It aggregates these errors by summation, subsequently computes the average across all data points, and presents the result as a percentage. The use of absolute values and multiplication by 100 ensures that all errors are positive and expressed as percentages.

Next, for association analysis, the researcher applies data mining techniques, specifically association rule mining through the Apriori algorithm, to uncover patterns and associations within the dataset. This analysis focuses on identifying frequent itemsets and association rules that reveal relationships between various variables, such as students' attributes, online forum participation, and their impact on student performance. The algorithm is implemented using the Weka tool, a popular open-source data mining software. According to Utkarsh (2023), Weka features a user-friendly Graphical User Interface (GUI) and a comprehensive suite of machine learning algorithms designed for classification, regression, clustering,

association rule mining, and feature selection. Weka also supports scripting, integrates seamlessly with various programming languages, and is compatible with multiple data formats.

The researcher configures the generic object editor within the Weka tool with a lower bound minimum support of 0.1 (10%) and a minimum confidence threshold of 0.6 (60%). These thresholds are selected to ensure that only strong association rules are identified. The strength of the association rule is measured by support, which is calculated as the proportion of transactions in the dataset that contain the itemset. The formula for support is in Equation (2).

$$Support(X) = \frac{Number\ of\ transactions\ containing\ X}{Total\ number\ of\ transactions} \quad (2)$$

Confidence refers to the likelihood that Y itemset appears in transactions that also contain X itemset. It measures the strength of the association rule $X \rightarrow Y$ by calculating the proportion of transactions where X is present, and Y is also observed. In other words, confidence quantifies the probability of Y occurring given that X has already occurred. The formula for calculating confidence is in Equation (3).

$$Confidence(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)} \quad (3)$$

A rule is considered strong if it satisfies both minimum support and minimum confidence thresholds set by the user. For example, if Support (X→Y) ≥ Minimum Support and Confidence (X→Y) ≥ Minimum Confidence, the rule is considered strong (Dino, 2022).

In summary, the research employs a quantitative approach that combines data collection, preprocessing, and association analysis to investigate student performance using GANs-enhanced datasets. Through the generation of synthetic data using GANs, the research meaningfully expands the dataset, thereby enhancing the analytical process with a diversified set of data points that closely replicate students' real-world profiles. The data preprocessing phase is meticulously conducted to uphold data integrity and consistency by addressing missing values, standardizing variables, and ensuring that the synthetic data accurately reflect students' authentic characteristics. This rigorous preprocessing establishes a reliable foundation for association analysis, in which the Apriori algorithm is employed to identify significant patterns and associations among variables, including students' demographics, academic behaviors, and extracurricular engagement. The analysis reveals key insights into how specific factors may contribute to or impact student performance, ultimately supporting the objective of using these associations to guide educational strategies and interventions. The research offers a comprehensive and data-driven understanding of the factors influencing student outcomes by integrating advanced machine learning techniques and data mining methodologies.

III. RESULTS AND DISCUSSIONS

The researcher unveils the results of the MAPE analysis for the student performance dataset, which has been augmented, enhanced, and expanded using GANs. The GANs play a pivotal role in this process by employing a two-part architecture consisting of a generator and a discriminator. The generator's task is to create synthetic data that closely resemble the original dataset, while the discriminator assesses the authenticity of the generated data in comparison to the real data.

The process begins with the collection of the original student performance dataset, ensuring it is comprehensive and representative of the target population. Following this, the data undergo preprocessing to normalize values and address any missing data, making it suitable for input into the GAN. As the training commences, the generator learns to produce synthetic data by capturing the underlying patterns present in the original dataset. Concurrently, the discriminator is trained to differentiate between real and synthetic data. This interplay between the two networks is characterized by a competitive dynamic, where the generator continually refines its output based on feedback from the discriminator. The training iterates until the generator successfully creates synthetic data that the discriminator cannot reliably distinguish from the real data, achieving a delicate balance between the two.

This section also highlights the best rules identified by applying the Apriori algorithm, which analyzes the associations and patterns within the dataset. The MAPE serves as a critical metric in this analysis, offering a standardized measure of the accuracy of the synthetic data when compared to the original dataset. By quantifying the deviation of the synthetic dataset from the actual values, MAPE expresses this error as a percentage, allowing for an easy interpretation of performance. The results, summarized in Table 1, present the MAPE for each statistic and the mean of the student data.

The MAPE for each statistic reflects the accuracy of the synthetic dataset compared to the original data for individual variables or statistics, capturing different aspects such as grades, attendance, or other performance indicators. This MAPE shows how well the GANs-generated data aligns with the original data for these specific metrics. In contrast, the MAPE for the mean provides an overall measure of accuracy across all data points, aggregating the deviations of the synthetic data from the original data into a single metric. It represents the average performance across the entire dataset and gives a comprehensive sense of the quality of the data generation process, summarizing the performance of the GANs across all observed statistics. A lower MAPE value indicates a closer match between the synthetic and original data, thereby demonstrating the quality of the data generation process (Liu et al., 2024; Shahul Hameed et al., 2024). From Table 1, it is evident that the MAPE

for each statistic varies across different categories, providing insight into the performance of the GAN in generating data that reflects the original dataset's characteristics.

The analysis of the MAPE for various statistics related to the dataset provides insights into the accuracy of the synthetic data compared to the original dataset. Each statistic reflects a different characteristic of the data, and the MAPE values indicate how closely the synthetic data replicates these characteristics. Lower MAPE values suggest a closer match between the synthetic and original data.

For "STUDENT AGE", the MAPE is approximately 0.1527, indicating a reasonably accurate replication with an error of about 15.27%. The MAPE for "SEX" is approximately 0.1705, showing a relatively close match with an error of about 17.05%. For "GRADUATED HIGH SCHOOL TYPE", the MAPE is approximately 0.1486, suggesting a replication error of about 14.86%. Then, the MAPE for "SCHOLARSHIP TYPE" is lower at approximately 0.0562, reflecting a more accurate replication with an error of about 5.62%. Meanwhile, "ADDITIONAL WORK" has a MAPE of approximately 0.1647, indicating an error of about 16.47%.

The MAPE for "REGULAR ARTISTIC OR SPORTS ACTIVITY" is approximately 0.1610, with an error of about 16.10%. For "DO YOU HAVE A PARTNER", the MAPE is approximately 0.1596. It suggests an error of about 15.96%. The MAPE for "TOTAL SALARY (IF AVAILABLE)" is relatively low at approximately 0.0646. It reflects a replication error of about 6.46%. Then, "TRANSPORTATION TO THE SCHOOL" has a MAPE of approximately 0.1850, with an error of about 18.50%. For "ACCOMMODATION TYPE", the MAPE is approximately 0.1744, suggesting an error of about 17.44%.

Table 1 Comparison of Mean Absolute Percentage Error (MAPE) for Each Statistic and for Mean in Students' Data

	Label	MAPE for Each Statistic	MAPE for Mean
1.	STUDENT AGE	0.152696	0.045225
2.	SEX	0.170516	0.050446
3.	GRADUATED HIGH SCHOOL TYPE	0.148621	0.037311
4.	SCHOLARSHIP TYPE	0.056246	0.003975
5.	ADDITIONAL WORK	0.164680	0.046735
6.	REGULAR ARTISTIC OR SPORTS ACTIVITY	0.160949	0.009224
7.	DO YOU HAVE A PARTNER	0.159612	0.000783
8.	TOTAL SALARY (IF AVAILABLE)	0.064598	0.119663
9.	TRANSPORTATION TO THE SCHOOL	0.184952	0.066562
10.	ACCOMMODATION TYPE	0.174414	0.078661
11.	MOTHER EDUCATION	0.089836	0.072066
12.	FATHER EDUCATION	0.093310	0.035271
13.	NUMBER OF SIBLINGS	0.061558	0.024757
14.	PARENTAL STATUS	0.173763	0.114389
15.	MOTHER OCCUPATION	0.052750	0.030474
16.	FATHER OCCUPATION	0.041704	0.018061
17.	WEEKLY STUDY HOURS	0.042841	0.064680
18.	READING FREQUENCY OF NON-SCIENTIFIC BOOKS OR JOURNALS	0.148109	0.014096
19.	READING FREQUENCY OF SCIENTIFIC BOOKS OR JOURNAL	0.108571	0.003655
20.	ATTENDANCE TO SEMINARS OR CONFERENCES RELEVANT TO THE COURSE	0.192056	0.026601
21.	IMPACT OF PROJECT OR ACTIVITIES ON STUDENT SUCCESS	0.171442	0.090288
22.	ATTENDANCE TO CLASSES	0.199202	0.055199
23.	PREPARATION TO EXAMINATIONS BASED ON COMPANIONSHIP	0.166703	0.078450
24.	PREPARATION TO EXAMINATIONS BASED ON TIME	0.164114	0.127410
25.	TAKING NOTES IN CLASSES	0.111188	0.059009
26.	LISTENING IN CLASSES	0.115669	0.007285
27.	DISCUSSION IMPROVES THE STUDENTS' INTEREST AND SUCCESS IN THE COURSE	0.120686	0.042916
28.	OPINION IN FLIPPED CLASSROOMS	0.142415	0.015856

The MAPE for "MOTHER EDUCATION" is approximately 0.0898, indicating a replication error of about 8.98%, while "FATHER EDUCATION" has a MAPE of approximately 0.0933, reflecting an error of about 9.33%. For "NUMBER OF SIBLINGS", the MAPE is approximately 0.0616, with an error of about 6.16%. Next, "PARENTAL STATUS" shows a MAPE of approximately 0.1738, with an error of about 17.38%. The MAPE for "MOTHER OCCUPATION" is relatively low at approximately 0.0528. The value indicates an error of about 5.28%. Then, "FATHER OCCUPATION" has a MAPE of approximately 0.0417, with an error of about 4.17%.

The MAPE for "WEEKLY STUDY HOURS" is approximately 0.0428, suggesting an error of about 4.28%. For "READING FREQUENCY OF NON-SCIENTIFIC BOOKS OR JOURNALS", the MAPE is approximately 0.1481, reflecting an error of about 14.81%. Then, "READING FREQUENCY OF SCIENTIFIC BOOKS OR JOURNALS" has a MAPE of approximately 0.1086, indicating an error of about 10.86%. The MAPE for "ATTENDANCE TO SEMINARS OR CONFERENCES RELEVANT TO THE COURSE" is approximately 0.1921, with an error of about 19.21%. Meanwhile, "IMPACT OF PROJECT OR ACTIVITIES ON STUDENT SUCCESS" shows a MAPE of approximately 0.1714, indicating an error of about 17.14%. The MAPE for "ATTENDANCE TO CLASSES" is approximately 0.1992, reflecting an error of about 19.92%.

For "PREPARATION TO EXAMINATIONS BASED ON COMPANIONSHIP", the MAPE is approximately 0.1667, with an error of about 16.67%. The MAPE for "PREPARATION TO EXAMINATIONS BASED ON TIME" is approximately 0.1641, indicating an error of about 16.41%. Then, "TAKING NOTES IN CLASSES" has a MAPE of approximately 0.1112, suggesting an error of about 11.12%. For "LISTENING IN CLASSES", the MAPE is approximately 0.1157, with an error of about 11.57%. The MAPE for "DISCUSSION IMPROVES THE STUDENT INTEREST AND SUCCESS IN THE COURSE" is approximately 0.1207, reflecting an error of about 12.07%. Finally, the MAPE for "OPINION IN FLIPPED CLASSROOMS" is approximately 0.1424, indicating an error of about 14.24%.

On the other hand, the analysis of the MAPE for the mean values of various columns in the dataset provides insights into how closely the synthetic data matches the original data in terms of average values. Lower MAPE values indicate better accuracy in replicating the mean values. Thus, it shows a higher quality of the synthetic data generation process.

For "STUDENT AGE", the MAPE for the mean is approximately 0.0452, signifying a very close match between the synthetic and original data, with an error of about 4.52%. The MAPE for "SEX" is approximately 0.0504, suggesting an extremely accurate replication with an error of about 5.04%. For "GRADUATED HIGH SCHOOL TYPE", the MAPE for the mean is approximately 0.0373. It indicates an extremely close match with an error of about 3.73%. The MAPE for "SCHOLARSHIP TYPE" is very low, at approximately 0.0040, reflecting an exceptionally accurate replication with an error of about 0.40%.

Next, the MAPE for the "ADDITIONAL WORK" mean is approximately 0.0467. It suggests an extremely close match with an error of about 4.67%. For "REGULAR ARTISTIC OR SPORTS ACTIVITY", the MAPE is approximately 0.0092, indicating an exceptionally accurate replication with an error of about 0.92%. The MAPE for "DO YOU HAVE A PARTNER" is extremely low, at approximately 0.0008. It signifies an almost perfect match with an error of about 0.08%. Next, "TOTAL SALARY (IF AVAILABLE)" has a MAPE of approximately 0.1197, suggesting a reasonably close match with an error of about 11.97%. The MAPE for "TRANSPORTATION TO THE SCHOOL" is approximately 0.0666, reflecting an extremely accurate replication with an error of about 6.66%.

For "ACCOMMODATION TYPE", the MAPE for the mean is approximately 0.0787, signifying a very close match with an error of about 7.87%. The MAPE for "MOTHER EDUCATION" is approximately 0.0721, suggesting a very accurate replication with an error of about 7.21%. For "FATHER EDUCATION", the MAPE is approximately 0.0353, indicating an extremely close match with an error of about 3.53%. The MAPE for "NUMBER OF SIBLINGS" is approximately 0.0248, signifying a very accurate replication with an error of about 2.48%. Next, the MAPE for "PARENTAL STATUS" is approximately 0.1144, suggesting a reasonably close match with an error of about 11.44%.

The MAPE for "MOTHER OCCUPATION" is extremely low, at approximately 0.0305, indicating an almost perfect match with an error of about 0.30%. For "FATHER OCCUPATION", the MAPE is approximately 0.0181, reflecting an extremely accurate replication with an error of about 1.81%. The MAPE for "WEEKLY STUDY HOURS" is approximately 0.0647, signifying a very close match with an error of about 6.47%. For "READING FREQUENCY OF NON-SCIENTIFIC BOOKS OR JOURNALS", the MAPE is very low, at approximately 0.0141, indicating an exceptionally accurate replication with an error of about 1.41%. Then, the MAPE for "READING FREQUENCY OF SCIENTIFIC BOOKS OR JOURNALS" is extremely low, at approximately 0.0037, reflecting an almost perfect match with an error of about 0.37%.

The MAPE for "ATTENDANCE TO SEMINARS OR CONFERENCES RELEVANT TO THE COURSE" is approximately 0.0266, suggesting an extremely accurate replication with an error of about 2.66%. For "IMPACT OF PROJECT OR ACTIVITIES ON STUDENT SUCCESS", the MAPE is relatively low, at approximately 0.0903, indicating a relatively accurate replication with an error of about 9.03%. The MAPE for "ATTENDANCE TO CLASSES"

is approximately 0.0552, reflecting an extremely close match with an error of about 5.52%. The MAPE for "PREPARATION TO EXAMINATIONS BASED ON COMPANIONSHIP" is relatively low, at approximately 0.0785. It suggests a reasonably accurate replication with an error of about 7.85%. For "PREPARATION TO EXAMINATIONS BASED ON TIME", the MAPE is approximately 0.1274, signifying a reasonably close match with an error of about 12.74%.

The MAPE for "TAKING NOTES IN CLASSES" is approximately 0.0590, indicating a very accurate replication with an error of about 5.90%. The MAPE is extremely low for "LISTENING IN CLASSES", at approximately 0.0073. It suggests an almost perfect match with an error of about 0.73%. The MAPE for "DISCUSSION IMPROVES THE STUDENT INTEREST AND SUCCESS IN THE COURSE" is approximately 0.0429, signifying a very close match with an error of about 4.29%. Finally, the MAPE for "OPINION IN FLIPPED CLASSROOMS" is extremely low, at approximately 0.0159. It indicates an almost perfect match with an error of about 0.16%.

These MAPE values are utilized to assess the quality of the synthetic data generation process, with lower values indicating better accuracy in replicating the corresponding statistics or mean values in the synthetic dataset compared to the original dataset. The association analysis of the GANs-expanded student performance dataset, which includes 28 attributes, reveals the ten best rules identified using the Weka tool. The 10 identified best rules are as follows:

1. TRANSPORTATION TO THE SCHOOL=Bus ==> ATTENDANCE TO SEMINARS OR CONFERENCES RELEVANT TO THE COURSE=Yes 130 <conf:(0.7)> lift:(1.22) lev:(0.04) [23] conv:(1.4)
2. IMPACT OF PROJECT OR ACTIVITIES ON STUDENT SUCCESS=positive 245 ==> ATTENDANCE TO SEMINARS OR CONFERENCES RELEVANT TO THE COURSE=Yes 170 <conf:(0.69)> lift:(1.2) lev:(0.04) [28] conv:(1.36)
3. MOTHER OCCUPATION=housewife 246 ==> ATTENDANCE TO SEMINARS OR CONFERENCES RELEVANT TO THE COURSE=Yes 161 <conf:(0.65)> lift:(1.13) lev:(0.03) [19] conv:(1.21)
4. PARENTAL STATUS=married 243 ==> ATTENDANCE TO SEMINARS OR CONFERENCES RELEVANT TO THE COURSE=Yes 159 <conf:(0.65)> lift:(1.13) lev:(0.03) [18] conv:(1.21)
5. READING FREQUENCY OF SCIENTIFIC BOOKS OR JOURNAL=Sometimes ATTENDANCE TO SEMINARS OR CONFERENCES RELEVANT TO THE COURSE=Yes 209 ==> READING FREQUENCE OF NON-SCIENTIFIC BOOKS OR JOURNALS=Sometimes 136 <conf:(0.65)> lift:(1.14) lev:(0.03) [16] conv:(1.21)
6. GRADUATED HIGH SCHOOL TYPE=state READING FREQUENCE OF NON-SCIENTIFIC BOOKS OR JOURNALS=Sometimes 202 ==> ATTENDANCE TO SEMINARS OR CONFERENCES RELEVANT TO THE COURSE=Yes 131 <conf:(0.65)> lift:(1.12) lev:(0.02) [14] conv:(1.19)
7. PREPARATION TO EXAMINATIONS BASED ON TIME=closest date to the exam 249 ==> ATTENDANCE TO SEMINARS OR CONFERENCES RELEVANT TO THE COURSE=Yes 161 <conf:(0.65)> lift:(1.12) lev:(0.03) [17] conv:(1.18)
8. PARENTAL STATUS=married 243 ==> READING FREQUENCY OF SCIENTIFIC BOOKS OR JOURNAL=Sometimes 156 <conf:(0.64)> lift:(1.17) lev:(0.03) [22] conv:(1.24)
9. PREPARATION TO EXAMINATIONS BASED ON COMPANIONSHIP=alone 227 ==> GRADUATED HIGH SCHOOL TYPE=state 145 <conf:(0.64)> lift:(1.17) lev:(0.03) [21] conv:(1.24)
10. WEEKLY STUDY HOURS=<5 hours 206 ==> ATTENDANCE TO SEMINARS OR CONFERENCES RELEVANT TO THE COURSE=Yes 131 <conf:(0.64)> lift:(1.1) lev:(0.02) [12] conv:(1.15)

By simplifying the rules into simplified interpretation, the students who use the bus for transportation to school are likely to attend seminars or conferences relevant to their course with a confidence of 70%. Moreover, those who have experienced a positive impact from projects or activities on their success are also likely to attend such events with a confidence of 69%. Students whose mothers are housewives are predicted to attend seminars or conferences with a confidence of 65%, as are students whose parents are married. Additionally, students who sometimes read scientific books or journals and attend relevant seminars or conferences are likely to sometimes read non-scientific books or journals with a confidence of 65%. Moreover, graduates from state high schools who sometimes read non-scientific books or journals are also likely to attend relevant seminars with a confidence of 65%. Students who prepare for exams based on the nearest exam date are similarly predicted to attend relevant conferences or seminars with a confidence of 65%. Furthermore, students

with married parents are expected to sometimes read scientific books or journals with a confidence of 64%. Those who prepare for exams alone are likely to have graduated from state high schools with a confidence of 64%, and students who study less than 5 hours weekly are also likely to attend relevant seminars or conferences with the same confidence level. These rules can provide valuable insights for educational institutions in understanding student behaviors and improving educational outcomes.

IV. CONCLUSIONS

In the research, the researcher employs GANs to augment a student's performance dataset, aiming to enhance data availability for analysis. The quality assessment of the generated synthetic data reveals reasonably accurate replication of various statistics and mean values from the original dataset, with MAPE typically ranging from 4% to 19%. Subsequently, association analysis is conducted using the Apriori algorithm to identify meaningful patterns and relationships within the GANs-expanded dataset. Notable associations are found, such as students using buses for transportation being more likely to attend seminars or conferences relevant to their courses with a confidence of 70%. These findings suggest the potential of GANs-enhanced datasets in uncovering valuable insights into student performance analysis, offering opportunities for educators and policymakers to make informed decisions to improve student outcomes.

While this study demonstrates the potential of GANs-enhanced datasets in revealing valuable insights into student performance, it also acknowledges certain limitations. The synthetic data, although statistically robust, may not fully capture the intricate patterns present in the original dataset. It can impact the depth of the associations identified. The reliance on the Apriori algorithm, which primarily uncovers frequent patterns, suggests that alternative methods can further enrich the analysis.

The implications of these findings are noteworthy for educators and policymakers since they highlight the importance of utilizing advanced data generation techniques to inform decision-making. By embracing GANs-augmented datasets and exploring diverse analytical methods, future research can enhance understanding of student behaviors and outcomes, leading to more effective strategies for educational improvement.

For future research, it will be useful to try different data augmentation methods or more advanced GAN models to generate more detailed features. Using other association analysis methods or machine learning algorithms can also provide further insights. Lastly, future research should use larger or more diverse datasets to ensure the findings apply to a broader range of educational contexts.

REFERENCES

Arvidsson, J. (2023). *Students performance*. Kaggle. https://www.kaggle.com/datasets/joebeachcapital/students-performance/data

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2017). Generating multi-label discrete patient records using generative adversarial networks. In *Machine Learning for Healthcare Conference* (pp. 286–305). PMLR. https://doi.org/10.48550/arXiv.1703.06490

Dino, L. (2022, May 1). *Association mining — Support, association rules, and confidence*. Medium. https://medium.com/@24littledino/association-mining-support-association-rules-and-confidence-60132a37e355

Figueira, A., & Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and GANs. *Mathematics, 10*(15), 1–41. https://doi.org/10.3390/math10152733

Gan, D., Numtong, K., Li, H., & Jiang, S. (2024). Exploring the application of the Apriori algorithm in knowledge mining for linguistic data within Chinese studies. *Eurasian Journal of Applied Linguistics, 10*(1), 279–298.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems 27 (NIPS 2014).*

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139–144. https://doi.org/10.1145/3422622

Google. (n.d.). *Frequently asked questions*. https://research.google.com/colaboratory/faq.html

Goyal, M., & Mahmoud, Q. H. (2024). A systematic review of synthetic data generation techniques using generative AI. *Electronics, 13*(17), 1–38. https://doi.org/10.3390/electronics13173509

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020). Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)* (pp. 12104–12114).

Li, Z., Li, X., Tang, R., & Zhang, L. (2021). Apriori algorithm for the data mining of global cyberspace security issues for human participatory based on association rules. *Frontiers in Psychology, 11*, 1–12. https://doi.org/10.3389/fpsyg.2020.582480

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3730–3738).

Liu, R., Wei, J., Liu, F., Si, C., Zhang, Y., Rao, J., ... & Dai, A. M. (2024). Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling.*

Ouassif, K., & Ziani, B. (2024). Predicting university major selection and academic performance through the combination of Apriori algorithm and deep neural network. *Education and Information Technologies*. https://doi.org/10.1007/s10639-024-13022-1

Pan, Z., Yu, W., Yi, X., Khan, A., Yuan, F., & Zheng, Y. (2019). Recent progress on Generative Adversarial Networks (GANs): A survey. *IEEE Access, 7,* 36322–36333. https://doi.org/10.1109/ACCESS.2019.2905015

Ramzan, F., Sartori, C., Consoli, S., & Reforgiato Recupero, D. (2024). Generative adversarial networks for synthetic data generation in finance: evaluating statistical similarities and quality assessment. *AI, 5*(2), 667–685. https://doi.org/10.3390/ai5020035

Rather, I. H., Kumar, S. (2024). Generative adversarial network based synthetic data training model for lightweight convolutional neural networks. *Multimedia Tools and Applications, 83*, 6249–6271. https://doi.org/10.1007/s11042-023-15747-6

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2,* 1–21. https://doi.org/10.1007/s42979-021-00592-x

Shahul Hameed, M. A., Qureshi, A. M., & Kaushik, A. (2024). Bias mitigation via synthetic data generation: A review. *Electronics, 13*(19), 1–14. https://doi.org/10.3390/electronics13193909

Tan, H. M., Minh, L. G., Minh, T. C., Quyen, T. T. B., & Cao-Van, K. (2024). Comparing LSTM models for stock market prediction: A case study with Apple's historical prices. In Nature of Computation and Communication (ICTCC 2023). Springer. https://doi.org/10.1007/978-3-031-59462-5_12

Utkarsh. (2023, May 16). *Weka in data mining*. Scaler. https://www.scaler.com/topics/data-mining-tutorial/weka-tool-in-data-mining/

Wang, H., & Yeung, D. Y. (2016). Towards Bayesian deep learning: A framework and some existing methods. *IEEE Transactions on Knowledge and Data Engineering, 28*(12), 3395–3408.

Wu, W. T., Li, Y. J., Feng, A. Z., Li, L., Huang, T., Xu, A. D., & Lyu, J. (2021). Data mining in clinical big data: the frequently used databases, steps, and methodological models. *Military Medical Research, 8*, 1–12. https://doi.org/10.1186/s40779-021-00338-z

Ye, F. (2020). Research and application of improved Apriori algorithm based on hash technology. In *2020 Asia-Pacific Conference on Image Processing (IPEC)* (pp. 64–67). IEEE. https://doi.org/10.1109/IPEC49694.2020.9115141