Chi Square Oversampling to Improve Dropout Prediction Performance in Massive Open Online Courses

Put authors name here[1i], Put author name here[1], Put author name here[1], Put author name here[2], Put authors name here[2]

[1]*Put author's affiliation here, complete with address, postal code, and country*
[2]*Put author's affiliation here, complete with address, postal code, and country*

Note: Due to our double-blind review policy, please include authors' information (including in the header and footer) only after acceptance at the peer review process.

Abstract. MOOCs have good potential for the Indonesian Government's efforts for educational equality. However, low retention rates are a global problem that must be addressed. One method used is to build a prediction model to prevent dropout. However, the prediction model has an obstacle in that comparing major and minor data is not proportional. In this research, the 141 datasets collected from the questionnaire results consisted of 95% participant data who completed the course and 5% dropout data. Therefore, in this study, oversampling was carried out to balance the data using the SMOTE-N chi-square method and the SMOTE-ENC chi-square method using chi-square. Next, the dataset formed is processed using the Support Vector Machine (SVM) machine learning method. In the testing process, the prediction model's performance with SMOTE-N chi-square and SMOTE-ENC chi-square oversampling data was compared with the prediction model's performance with regular SMOTE-N and SMOTE-ENC oversampling data. The data processing results show a significant increase in accuracy from each oversampling method with weighting. The SMOTE-N weighting modification using the chi-square value has the best value, where the F1-measure value reaches 95.33%, and there is a decrease in error in predicting dropout data using this method. This indicates that the model formed using the SMOTE-N chi-square method has good predictive ability.

Keywords: Chi-square; Dropout Prediction; Indonesia; MOOC; Oversampling;

1. Introduction

Massive Open Online Courses (MOOC) is an independent online learning (Cidral et al., 2020), where students can choose the desired material. The concept of openness in MOOCs allows anyone to have the same opportunity to learn, regardless of location, and not be tied to the flow of the curriculum. Currently, MOOCs have been widely used worldwide, and the certificates provided by several large MOOCs have good legality.

MOOCs in Indonesia have been used since 2014, and the MOOCs used also vary, depending on whether they come from abroad, are licensed from abroad, or are made domestically. With Indonesia's geographic condition as an archipelago and the uneven distribution of infrastructure and economic conditions, MOOCs can be a solution for equal distribution of education in Indonesia. The education system in Indonesia has changed severally due to various factors, such as the current Covid-19 pandemic (Izzati et al., 2024). Almost all education lines are moving online like the Merdeka Belajar Program launched by the Ministry of Education and Culture (Direktorat Jenderal Guru dan Tenaga Kependidikan, 2020). This program incorporates distance learning, which prioritizes the quality of human resources and is not limited by space and time. In the 6th National Dialogue Sukses Indonesiaku (BRIN, 2018), the Minister of Ristekdikti stated that online learning should be given more attention. This would enable students to learn without being bound by space

and time constraints, consistent with the Industrial Revolution 4.0 program. The number of internet users in Indonesia has increased significantly.

At the beginning of 2020, a survey showed that internet users reached 175.3 million or 64% of the Indonesian population aged 15 to 64. Furthermore, 98% of internet users surf the virtual world using mobile media (Jayani, 2018). Surveys show that 56.11% of people aged 20-25 have heard of MOOC. Additionally, 79.77% of the respondents had yet to try to access MOOC held for Indonesians. MOOC user data in Indonesia reaches only 695,000, or 0.4% of internet users (Lubis et al., 2020). Therefore, the government supported MOOC by issuing Permendikbud number 109 of 2013, stating that Indonesians could use online education services. The Ministry of Communication and Information of Indonesia has also promoted a program to use MOOC for equal distribution of digital literacy among government employees (Andreya, 2022).

In general, the problems faced by MOOCs are the same, namely low retention rates, reaching 5-10% (Bozkurt, 2019), which is inversely proportional to high dropout rates. Various studies have been conducted to increase this retention, including finding determining factors in MOOC use (Joo et al., 2018), building better designs (Haugsbakken, 2020), providing exciting delivery (Ahmad, 2021), or predicting dropout early (Abu Zohair, 2019).

Research related to predicting dropout in MOOCs still has a good trend. This happens because dropouts can be identified early with a suitable prediction method. By identifying dropouts early, the system can likely do something to retain course participants until course completion. Thus, retention rates in MOOC courses can be improved. A search on Google Scholar found 8290 publications with the keyword "MOOC dropout prediction", with a filter for 2020-2024. This number increased by 60% from searches with the exact keywords, with the 2015-2019 filter. However, there are not many similar publications that are filtered with the additional keyword Indonesia. Some of the publications that appear, when opened, do not show research locations specifically in Indonesia. Most of the research was conducted in developed countries, where various supporting infrastructures have used and supported MOOCs (Deng et al., 2019). Meanwhile, in developing countries, such as Indonesia, the use of MOOCs is still less popular (Lambert, 2020; Van De Oudeweetering & Agirdag, 2018) as a result of basic needs that have not been appropriately met, such as infrastructure conditions and resource ownership (Alhazzani, 2020; Dillah et al., 2023), financial conditions (Arhin & Wang'Eri, 2018), the use of foreign languages (Ruipérez-Valiente et al., 2020), and the ability to master technology (Hong et al., 2021).

Machine Learning is widely used to help predict (Sari et al., 2023). In a literature study conducted by Dalipi (Dalipi et al., 2018), the three most frequently used machine learning methods are Logistic Regression, Support Vector Machine (SVM), and Decision Tree. The SVM machine learning method was developed by Vladimir Vapnik (Schölkopf, 2002). SVMs are supervised machine-learning models that can be used for classifying, regression, or detecting outliers (Naghipour et al., 2024). SVM is known to have good accuracy for small datasets (Abu Zohair, 2019), has poor performance for unbalanced data (Rezvani & Wang, 2023; Wang et al., 2021), can be used in research that uses many parameters (Nurhayati et al., 2015) and can be used in various research fields (Cervantes et al., 2020). SVM is a supervised machine learning algorithm that classifies data into two groups by creating vectors (hyperplanes) (Pribadi & Shinoda, 2022). This study uses 141 data, where the predicted targets are classified that successfully passed and those that did not (classification). With SVM's capabilities and the available data conditions, SVM can be used in this research.

Commented [L1]: R1-1 However, there is an issue for the fundamental idea. Please provide comment and strong reference to against these issue

Commented [L2R1]: Added facts about the condition of MOOCs in Indonesia (continued in another paragraph)

Goopio (Goopio & Cheung, 2021) examined the factors that influence the dropout phenomenon in MOOCs and developed strategies to increase retention. Understanding the MOOC dropout phenomenon and increasing knowledge about factors influencing retention will enable MOOC providers to improve MOOCs' design features and course quality. In MOOC itself, Badali stated that by influencing the learning process of participants, such as providing distractions to avoid boredom and increasing participant engagement during the learning process, it can increase participant retention in the courses they take (Badali et al., 2022). In several studies in other fields, it was found that systems that provide recommendations or personalization using artificial intelligence or machine learning to users can increase user retention on the system, like e-commerce (Abdollahpouri et al., 2020; Acharya et al., 2023), digital marketing (Behera et al., 2020) and music player (Anderson et al., 2020). In this research, publications related to retention in MOOCs, namely research on predictions, sustainability of use, and level of motivation. The prediction process is grouped based on the data used, namely data on previous MOOC usage, user activity data on ongoing courses, and demographic data on course participants.

Commented [L3]: R1-1 However, there is an issue for the fundamental idea. Please provide comment and strong reference to against these issue

Commented [L4R3]: The finding that AI or machine learning can increase a person's retention in using the system has been added

The problem many people face in research related to predicting dropout in MOOCs is the imbalance in the data between those who complete the course and those who drop out. An imbalance in the amount of data can cause prediction accuracy on minor data to be less than optimal (Gyoten et al., 2020). In some conditions, minor data has a considerable influence on the prediction process, including in predicting course completion (Fahrudin et al., 2019). Several methods are used by researchers to overcome this problem, such as adjusting the training to get balanced data (Pazzani et al., 1994) or modifying the data (Japkowicz, 2000). Data modification can be done by oversampling minor data so that the amount of minor data equals significant data, or reducing the amount of major data to match the amount of minor data (Japkowicz, 2000).

A technique that is widely used to overcome data imbalance is synthetic data oversampling (Gyoten et al., 2020) (Limanto et al., 2024). Several oversampling methods have been developed in previous research. One oversampling method that is widely used is Synthetic Minority Over-sampling Technique (SMOTE) (Fahrudin et al., 2016), but SMOTE itself was developed for oversampling quantitative data. Several modifications of SMOTE that can be used for oversampling qualitative data are SMOTE for Nominal (SMOTE-N), SMOTE for Encoded Nominal and Continuous (SMOTE-ENC), and SMOTE for Nominal Continuous (SMOTE-NC) (Limanto et al., 2024). The difference between these three methods is that SMOTE-N can be used for qualitative data only, while SMOTE-ENC can be used for a mixture of qualitative and qualitative data, and SMOTE-NC can be used for a mixture of quantitative and qualitative data, but cannot be used for qualitative data.

Previous research suggested that oversampling by weighting underpopulated data effectively improved machine learning model performance (Fahrudin et al., 2019; Limanto et al., 2024). In his research, Tora developed the AWH-SMOTE method (Fahrudin et al., 2019), which can be used to oversampling quantitative data. The Information Gain used in the AWH-SMOTE method can improve the performance of the prediction model (minority recall, minority precision, and minority f-measure) when compared with other weighting methods. Meanwhile, Limanto et al built the GLoW SMOTE-D model (Limanto et al., 2024) for oversampling qualitative data. The trial results show that this method can improve the performance of predicting student failure in taking subjects compared to other techniques.

In this study, 5% of respondents stated that they had not completed the MOOC course. This happened because most of the respondents involved took MOOC courses due to work obligations, so motivation factors could not be measured in this research. To handle this

imbalance in the comparison of major and minor data, an oversampling process is necessary to balance the data.

This research aims to improve the ability of the MOOC course participant dropout prediction model by modifying the weighting of the SMOTE-N and SMOTE-ENC oversampling methods. In this research, weighting was carried out using the chi-square method, which measures the correlation between indicators and output in qualitative data. The primary dataset used in this research was obtained from distributing questionnaires to respondents who had used MOOCs as a learning tool. The dataset was processed using the SEM method, and from the blindfolding process, it was found that the predictive relevance value had a moderate prediction indication. Therefore, the data is then processed using machine learning to get predictions for dropout from the MOOC. Data processed using SVM is eliminated according to the factors accepted in the model formed. The trial was conducted by comparing the prediction results from chi-square SMOTE with the dataset prediction results from oversampling using SMOTE-N and SMOTE-ENC. Prediction results are measured using accuracy levels, including recall, precision, and F1-measure.

The presentation in this article is divided into three parts. The first part explains the research methodology carried out, followed by a presentation of the research results. The final section presents conclusions and outlines possibilities for future exploration.

2. Research Methods

This research was carried out in stages, as in Figure 1. First, an instrument was prepared based on previous research, and then primary data was collected by distributing questionnaires to various respondents who met the requirements. Because data processing using SEM shows that the data has predictive capabilities, planning is carried out to build a prediction model. The data composition was balanced at the preprocessing stage by developing the SMOTE-N chi-square and SMOTE-ENC chi-square oversampling methods. The dataset and oversampling results are processed using the SVM machine learning method, and model performance is measured. Details of the research methodology are described in the following section.

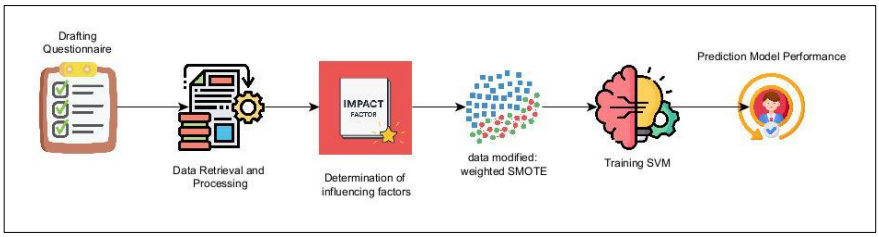

Figure 1 Research Methodology

2.1. Dataset Collection

The research instrument was prepared based on previous research (Liliana et al., 2022), where in this research it was found that the factors that influence retention in MOOCs in developing countries are resources (Sánchez-Prieto et al., 2016), social influence (Dewberry & Jackson, 2018), self-efficacy (Briz-Ponce et al., 2017), perceived ease of use (Taghizadeh et al., 2021) and perceived usefulness (Reparaz et al., 2020). The first five factors are based on findings from a literature study of 89 Scopus papers sorted by research location (Liliana et al., 2022). A review of 89 papers found 18 factors used in the studies, with 26 papers conducted in developing countries. Of the 89 publications used, 86% came from the Scopus-indexed Q1 quartiles, such as the Journal of Economic Perspectives,

Computers and Education, Computers in Human Behavior, Information and Management, American Educational Research Journal, Future Generation Computer Systems, and International Journal of Information. Management. In addition, 10% came from Q2 quartiles such as Information Economics and Policy, Information Systems and e-business Management, Electronics (Switzerland), and Asia Pacific Education Review.

The factors used in each study were sorted and ranked, and the five most used were selected. The categorization found a slight difference between factor rankings worldwide and in developing countries. Meanwhile, the power distance and Uncertainty Avoidance factors are used in this research because Indonesia, the location of this research, is a country with high scores on the power distance and uncertainty avoidance indicators (Insights, 2022). This shows that Indonesian people still have a tendency to be controlled by people who are more powerful, such as parents, teachers, or seniors in the office, and prefer to avoid conflict by obeying what other people want.

This instrument was prepared based on instruments used in previous research and validated by education, language, and statistics experts. Survey questions were prepared by educational research psychologists in the form of a combination of closed-ended questions (closed-ended questionnaires) using the Likert scale method (Awang et al., 2016). The study adopted measuring tools used in previous studies because they were considered fit for reuse. Most of the instruments were adopted from similar studies. Instruments on cultural factors in power distance and uncertainty avoidance were prepared independently by the research team. All survey questions were prepared with educational psychologist researchers in the form of a combination of closed-ended and open-ended questions through interview techniques (McLafferty, 2016). The experts involved in preparing this research instrument were educational psychologists to determine factors, English and Indonesian language experts to validate the results of instrument translation, and statisticians to build research models. Instruments are presented in Table 1, and the details can be seen at https://tinyurl.com/bdh53mbr.

Data collection was carried out through questionnaires, and 141 MOOC users were collected in Indonesia. The study respondents comprised college students with a working age of up to 45 years. They constitute a digital generation mature enough to make decisions. The characteristics of the respondents are the age of digital learners, with a distribution of 37% aged 17-25 years, 53.5% aged 26-40 years, and 9.4% over 40 years, of which 79.5% are already working. 88.2% of respondents took 1-5 courses at MOOC within one year, while 7.9% were outside Java. The most commonly used MOOCs by respondents are Coursera and Udemy. Another finding is that the types of MOOCs used mainly by respondents in this study are Coursera and Udemy. This is in line with the findings on Google Scholar, with the keywords "Coursera/udemy/futurelearn in Indonesia", which shows that MOOC research in Indonesia is dominated by Coursera (820 papers), Udemy (393 papers) and Futurelearn (125 papers).

The distribution of questionnaires was carried out via social media such as Facebook and Instagram, conversation media such as WhatsApp and WhatsApp groups, and email media. The problem faced during data collection was that MOOCs were less popular, as indicated by many information recipients who questioned what a MOOC was and had never heard of Coursera and the like. Apart from that, many groups on Facebook and WhatsApp are inactive, so the response received is less than expected.

Most of the respondents were from the island of Java. Respondents from outside Java only made up 8% of the total respondents. This is due to the gap in the quality of infrastructure on the island of Java and outside Java. The infrastructure referred to here is the smooth running of the internet network and the existence of supporting hardware.

Several studies found that the quality of this infrastructure, apart from influencing productivity in the economic sector (Sukwika, 2018), is also closely related to the quality of public education (Sinta & Wahyuni, 2022). The more difficult it is to reach, the lower the quality of the internet available, so the internet literacy of people in that area is also low. This is in line with findings in the field. Many colleagues outside Java need to learn what a MOOC is, so they cannot be respondents in this research. Most of the respondents involved in this research stated that they could complete the MOOC course they took because providing the course was a work obligation. Meanwhile, several people not burdened with the obligation to learn from their work and were contacted to be respondents in this research stated that they had never been involved in MOOC courses because they could learn from YouTube or blogs. This causes an imbalance in the amount of data on MOOC course participants who passed and those who still need to complete the course.

The data collected in this research was 141 data, of which 5% were respondents who had used MOOCs but still needed to complete the course. There needs to be more data between participants who can complete the course and those who drop out in the MOOC learning process. Data collected from the questionnaire was processed using the SEM model and the SMART-PLS application. Based on the test results, several things can be concluded. Social influence, self-efficacy, and perceived ease of use factors directly influence behavioral intention. Meanwhile, perceived usefulness was found to have no positive influence on behavioral intention. This fact is in line with previous findings (Issa & Isaias, 2016), which stated that in developing countries, the perceived ease of use of the system is more important than the perceived benefits captured by a person.

2.2. Oversampling

In this study, data oversampling was carried out by developing the SMOTE N and SMOTE ENC chi-square methods. Oversampling was also performed using the SMOTE-N and SMOTE-ENC methods for performance comparison. In general, SMOTE looks for the closest minor data points and creates new data based on similarities to the closest data. In SMOTE-N, the determination of data duplication is calculated using the Value Difference Metric (VDM) distance formula, as can be seen in Equation (1) (Chawla et al., 2002). The VDM equation calculates the value difference matrix for each nominal feature in a particular set of feature vectors, as seen in algorithm 1. Meanwhile, in SMOTE-ENC, data duplication is determined using the Euclidean Distance formula, as seen in Equation (2) (Gyoten et al., 2020).

$$\delta\ (x, y) = \sum_{i=1}^{N} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right|^k \quad (1)$$

Where

x	= value of field a of record r1
Y	= value of field a of record r2
δ (x, y)	= the distance between x and y
K	= a constant that usually has a value of one or two
C_x	= the number of occurrences of the value x in field/column a (c1)
Cy	= the number of occurrences of the y value in field/column a (c2)
i	= 0, 1 --> 0 passes, 1 dropout
Cx,i	= the number of occurrences of the value x in field/column a, which has an output column/field = i

Algorithm 1 How SMOTE works
1. Identify minor data
2. Randomly select 1 minor data point
3. Select K-nearest neighbors (k=3) from the minor data group using the VDM formula
4. Select 1 nearest neighbor data
5. Repeat steps 1–4 until the amount of data is balanced

$$E(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \quad (2)$$

Where
M = the number of fields/features/columns other than the output
X = first record
Y = second record
E(x, y) = the distance between x and y

Development is carried out by weighting when calculating distance. The weights given are obtained based on the results of correlation calculations with chi-square, as can be seen in Equation (3). Chi-square is helpful for testing the relationship between indicators and output from research data (Nihan, 2020). In other research, weighting using the chi-square method improved the quality of the flawed data detection process (Gol & Abur, 2015).

$$\chi^2(X,Y) \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(O_{ij}-E_{ij})^2}{Eij} \quad (3)$$

Where
X,Y = variables whose correlation will be calculated
O_{ij} = frequency of observations
E_{ij} = expected frequency
n = the number of possible values of variable X
m = the number of possible values of variable Y

Four types of oversampling are carried out, namely with values N=1000, 1300, 1500 and 2000, aiming to make predictions with proportionally different data. The variable N determines the number of duplicate data, where the N value must be a multiple of 100. If N=1000, then the amount of minor data will be 11x the original amount. For example, the initial number of minor data is 7; with N=1000 duplication, the number of minor data will be 77.

2.3. Prediction

The prediction model will be built using the SVM method, which is implemented in the Python programming language. To ensure the quality of the model, predictions are carried out using 10-fold cross-validation (Malakouti et al., 2023), where the dataset is grouped into 10 parts. One part is used for testing, and the other 9 parts are used for training. The training and testing process is repeated ten times so that every part is used as testing data.

2.4. Prediction Model Performance

The performance of the prediction model used is accuracy, as seen in Equation (4), recall, as seen in Equation (5); precision, as seen in Equation (6) and F1-measure, as seen in Equation (7) (Radha & Nelson Kennedy Babu, 2020), according to the Confusion Matrix (Chawla et al., 2002). The confusion matrix is a table that shows the comparison between the actual value and the model's predicted value, as seen in Table 1. The True Negative (TN) condition indicates that the model can predict the value 0 correctly, where 0 indicates the

condition of the participant who passed the course. False Negative (FN) indicates the model's prediction error regarding the actual value of 0. Meanwhile, the True Positive (TP) condition indicates that the model is able to predict the value 1 correctly, where 1 indicates the condition of the participant who dropped out of the course. False Positive (FP) indicates the model's prediction error regarding actual value 1.

Tabel 1 Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

$$Accuration = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$F1 = \frac{2}{\frac{1}{Recall}+\frac{1}{Precision}} = 2\ x\ \frac{Precision\ x\ Recall}{(Precision\ x\ Recall)\ x\ \frac{Precision+recall}{Precision\ x\ Recall}} = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall} \quad (7)$$

The accuracy value shows the model's ability to make predictions on major and minor data, compared to the overall prediction results, whether the prediction results are correct or incorrect. Recall measures how well the model can find existing positive conditions (dropouts), compared to actual dropout data. Precision measures how accurate the model is when making dropout predictions compared to dropout prediction data. F1-measure is the harmonic value of recall and precision, providing a balanced picture between the two matrices to ensure that the model can predict both major and minor data.

3. Results and Discussion

The questionnaire has 30 parameters involved in the machine learning process, where passing is marked with an output of 0 and dropout is marked with a value of 1. The composition of major (pass) and minor (dropout) data is very different, namely 95% and 5%. Therefore, data oversampling was carried out using the SMOTE-N and ENC chi-square methods. Meanwhile, for comparison, oversampling was carried out on the same dataset using the usual SMOTE-N and SMOTE-ENC methods.

Processing of the oversampling dataset using SVM can be seen in Table 2. The measurement results show that the initial dataset cannot predict minor data (dropout). This is indicated by the average recall, precision, and F1-measure values of 0.0%, which indicates a TP value = 0 (the model cannot predict dropout at all). Meanwhile, the oversampling dataset in the F1-Measure column shows an increase in the prediction accuracy of minor data, and the level of accuracy increases as the comparison of major and minor data becomes more balanced.

Figure 2 compares F1-measure accuracy for each method. Based on the calculation results, the SMOTE-N oversampling method shows stable performance when N=1500 (minor data duplication reaches 16x the initial data amount). Meanwhile, other methods show increased performance in all accuracy calculations (Accuracy, Recall, Precision, F1-measure columns), even in minor data duplication, reaching 21x the initial data amount.

Table 2 Level of Accuracy of Prediction Results

		Accuracy	Recall	Precision	F1-Measure	Process Time (second)
preliminary data		**94.42%**	0.00%	0.00%	0.00%	3.42
SMOTE-N	N=1000	**91.89%**	90.98%	86.97%	88.00%	**0.42**
SMOTE-ENC		91.84%	92.70%	**87.70%**	**89.51%**	**0.42**
chi-square SMOTE-N		**91.89%**	**93.17%**	87.62%	89.49%	0.84
chi-square SMOTE-ENC		**91.89%**	91.63%	87.44%	89.26%	0.84
SMOTE-N	N=1300	93.12%	**93.90%**	89.73%	91.37%	**0.3**
SMOTE-ENC		92.60%	93.88%	90.79%	91.90%	0.96
chi-square SMOTE-N		93.10%	93.12%	**91.40%**	**92.05%**	1.2
chi-square SMOTE-ENC		**93.14%**	93.73%	91.27%	91.98%	0.78
SMOTE-N	N=1500	**93.51%**	**95.17%**	91.72%	92.97%	1.26
SMOTE-ENC		93.06%	94.80%	91.79%	92.99%	**0.36**
chi-square SMOTE-N		93.50%	93.41%	**92.83%**	92.79%	0.6
chi-square SMOTE-ENC		**93.51%**	94.54%	92.47%	**93.29%**	0.6
SMOTE-N	N=2000	94.37%	**96.34%**	93.80%	94.85%	1.08
SMOTE-ENC		94.34%	95.59%	93.88%	94.54%	0.9
chi-square SMOTE-N		**94.77%**	95.72%	**95.30%**	**95.36%**	2.04
chi-square SMOTE-ENC		94.76%	96.01%	94.79%	95.33%	**0.42**

The F1-measure value in predictions using a weighted dataset is better than the value produced using the SMOTE-N and SMOTE-ENC methods without weighting. Meanwhile, the SMOTE-N chi-square and SMOTE-ENC chi-square did not show a significant difference. This indicates that the SMOTE-N chi-square and SMOTE-ENC chi-square methods improve prediction model performance.

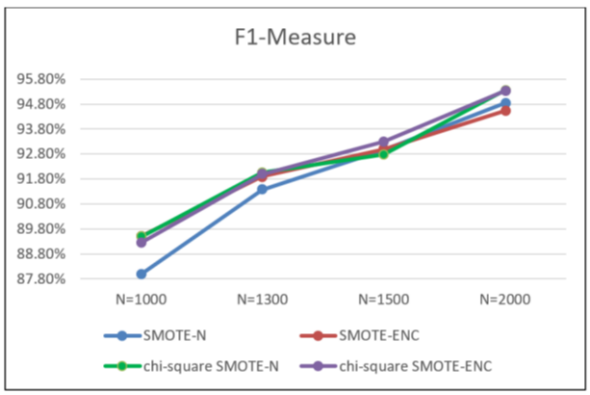

Figure 2 Comparison of F1-Measure from various oversampling datasets

Commented [L5]: Editor - Please revise your graphical abstract to the SmartArt Graphics which improves the reader's interpretation of the paper with jpg or png format, not same with another figure on paper

Commented [L6R5]: graphical abstract has been changed to PNG

The level of accuracy shows the model's performance in predicting major and minor data. A comparison of the accuracy values of the initial dataset with the modified results can be seen in Figure 3. From the results of this research, the initial dataset has an accuracy level of 94.42%, but 0% in the recall value (TP=0), which indicates that the model can predict major data (graduation) only. The level of accuracy in the dataset resulting from oversampling using the SMOTE and SMOTE-ENC methods is lower than the level of accuracy from the initial dataset. Still, on the contrary, there is an increase in the recall value and precision value. This indicates that the model that uses the oversampled dataset can predict minor data (dropout).

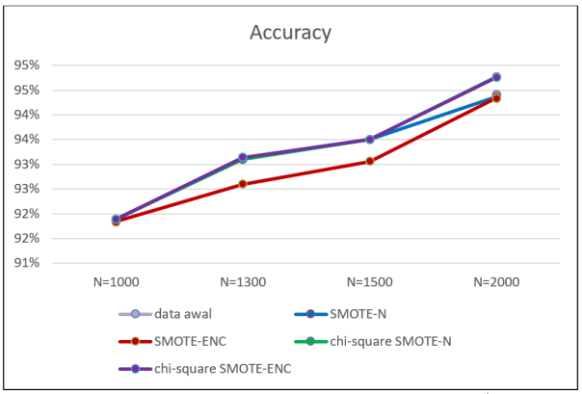

Figure 3 Accuracy comparison of various oversampling datasets

Judging from the precision value (Fig 4) in the oversampling dataset, an increase in the N value indicates an increase in the precision value. Apart from that, the precision value in the dataset resulting from oversampling using the SMOTE-N chi-square method consistently shows a higher average for all N values than the other three methods. This indicates that the higher the precision value (Equation 6), the error in the dropout prediction (FP) decreases. Therefore, the SMOTE-N chi-square oversampling method produces a lower dropout prediction error.

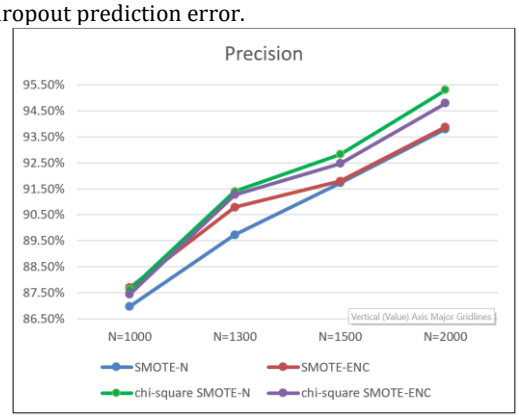

Figure 4 Precision comparison of various oversampling datasets

Commented [L7]: Editor - Please revise your graphical abstract to the SmartArt Graphics which improves the reader's interpretation of the paper with jpg or png format, not same with another figure on paper

Commented [L8R7]: graphical abstract has been changed to PNG

Commented [L9]: Editor - Please revise your graphical abstract to the SmartArt Graphics which improves the reader's interpretation of the paper with jpg or png format, not same with another figure on paper

Commented [L10R9]: graphical abstract has been changed to PNG

Overall, with N=2000, the best method in this trial was chi-square SMOTE-N (Table 3). Calculations are carried out by processing accuracy data at N=2000 with the Rank.Avg formula in Excel, which can be seen in Equation (8). From these calculations, it can be concluded that the modification of the SMOTE-N chi-square method significantly impacts the dataset processed in SVM.

Table 3 Ranking Method with N=2000

	Acuration	Recall	Precision	F1-Measure	Rating
SMOTE-N	3	**1**	4	3	2.75
SMOTE-ENC	4	4	3	4	3.75
chi-square SMOTE-N	**1**	3	**1**	**1**	**1.5**
chi-square SMOTE-ENC	2	2	2	2	2

$$R = Rank.Avg(value, range, order) \quad (8)$$

Where

R = ranking obtained for each accuracy

Value = performance in each cell

Range = performance on the measured column

Order = sorting method, 0 for descending

4. Conclusions

The condition of the initial dataset used in this research cannot predict dropout data (minor data), because the proportion of major and minor data is unbalanced. In this study, oversampling of the dataset was carried out using chi-square SMOTE-N and chi-square SMOTE-ENC. The two modified methods can increase prediction accuracy on minor data compared to processing the SMOTE-N and SMOTE-ENC oversampling dataset without weighting. This is in line with the results of previous research, where weighting carried out using the SMOTE method gave better results for quantitative data (Fahrudin et al., 2019). The weakness of the SMOTE-N chi-square and SMOTE-ENC chi-square methods is that the processing time is relatively longer than the SMOTE N and SMOTE-ENC methods, but is still faster than processing the initial dataset. Therefore, this method still has potential to be developed so that datasets can be processed more quickly.

Conflict of Interest

The authors declare no conflicts of interest.

References

Abdollahpouri, H., Adomavicius, G., Burke, R., Guy, I., Jannach, D., Kamishima, T., Krasnodebski, J., & Pizzato, L. (2020). Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction*, *30*(1), 127–158. https://doi.org/10.1007/s11257-019-09256-1

Abu Zohair, L. M. (2019). Prediction of Student's performance by modelling small dataset size. *International Journal of Educational Technology in Higher Education*, *16*(1). https://doi.org/10.1186/s41239-019-0160-3

Acharya, N., Sassenberg, A. M., & Soar, J. (2023). Effects of cognitive absorption on continuous use intention of AI-driven recommender systems in e-commerce. *Foresight*, *25*(2), 194–208. https://doi.org/10.1108/FS-10-2021-0200

Ahmad, E. A. (2021). Content presentation techniques for learning experience enhancement in Massive Open Online Course (MOOC). *International Journal of E-Learning and Higher Education*, *14*(1), 19–32. https://doi.org/10.24191/ijelhe.v14n1.1412

Alhazzani, N. (2020). MOOC's impact on higher education. *Social Sciences & Humanities Open*, *2*(1), 100030. https://doi.org/10.1016/j.ssaho.2020.100030

Anderson, A., Maystre, L., Anderson, I., Mehrotra, R., & Lalmas, M. (2020). Algorithmic Effects on the Diversity of Consumption on Spotify. *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020*, *2*, 2155–2165. https://doi.org/10.1145/3366423.3380281

Andreya, E. (2022). *Kominfo dan Kemendagri Luncurkan MOOC Literasi Digital Sektor Pemerintahan*. Kementrian Komunikasi Dan Informatika RI. https://aptika.kominfo.go.id/2022/11/kominfo-dan-kemendagri-luncurkan-mooc-literasi-digital-sektor-pemerintahan/

Arhin, V., & Wang'Eri, T. (2018). Orientation programs and student retention in distance learning: The case of university of cape coast. *Journal of Educators Online*, *15*(1). https://doi.org/10.9743/JEO2018.15.1.6

Awang, Z., Afthanorhan, A., Mamat, M., Awang, Z., & Afthanorhan, A. (2016). The Likert scale analysis using parametric based Structural Equation Modeling (SEM). *Computational Methods in Social Sciences*, *4*(1), 13–21. https://doi.org/10.5281/zenodo.1299429

Badali, M., Hatami, J., Banihashem, S. K., Rahimi, E., Noroozi, O., & Eslami, Z. (2022). The role of motivation in MOOCs' retention rates: a systematic literature review. *Research and Practice in Technology Enhanced Learning*, *17*(1). https://doi.org/10.1186/s41039-022-00181-3

Behera, R. K., Gunasekaran, A., Gupta, S., Kamboj, S., & Bala, P. K. (2020). Personalized digital marketing recommender engine. *Journal of Retailing and Consumer Services*, *53*(September 2018), 101799. https://doi.org/10.1016/j.jretconser.2019.03.026

Bozkurt, A. (2019). Dropout Patterns and Cultural Context in Online Networked Learning Spaces. *Open Praxis: International Council for Open and Distance Education*, *11*(1), 41–54.

BRIN, K. R. (2018). *Mohamad Nasir : Indonesia Sukses Berkualitas di Tangan Kita*. https://ristekdikti.go.id/kabar/mohamad-nasir-indonesia-sukses-berkualitas-di-tangan-kita-2/

Briz-Ponce, L., Pereira, A., Carvalho, L., Juanes-Méndez, J. A., & García-Peñalvo, F. J. (2017). Learning with mobile technologies – Students' behavior. *Computers in Human Behavior*, *72*, 612–620. https://doi.org/10.1016/j.chb.2016.05.027

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*(xxxx), 189–215. https://doi.org/10.1016/j.neucom.2019.10.118

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *30*(2), 321–357. https://www.jair.org/index.php/jair/article/view/10302/24590

Cidral, W., Aparicio, M., & Oliveira, T. (2020). Students ' long-term orientation role in e-learning success : A Brazilian study. *Heliyon*, *6*(June), e05735. https://doi.org/10.1016/j.heliyon.2020.e05735

Dalipi, F., Imran, A. S., & Kastrati, Z. (2018). MOOC dropout prediction using machine

learning techniques: Review and research challenges. *IEEE Global Engineering Education Conference, EDUCON*, *2018-April*, 1007–1014. https://doi.org/10.1109/EDUCON.2018.8363340

Deng, R., Benckendorff, P., & Gannaway, D. (2019). Progress and new directions for teaching and learning in MOOCs. *Computers and Education*, *129*(July 2018), 48–60. https://doi.org/10.1016/j.compedu.2018.10.019

Dewberry, C., & Jackson, D. J. R. (2018). An application of the theory of planned behavior to student retention. *Journal of Vocational Behavior*, *107*, 100–110. https://doi.org/10.1016/j.jvb.2018.03.005

Dillah, A. U., Asbari, M., & Faris, M. (2023). Educations Guidelines: Merajut Sistem Pendidikan di Negara Berkembang. *Journal of Information Systems and Management (JISMA)*, *2*(5), 93–95.

Direktorat Jenderal Guru dan Tenaga Kependidikan. (2020). *Merdeka Belajar*. 2020. https://gtk.kemdikbud.go.id/read-news/merdeka-belajar

Fahrudin, T., Buliali, J. L., & Fatichah, C. (2016). RANDSHUFF: An algorithm to handle imbalance class for qualitative data. *International Review on Computers and Software*, *11*(12), 1093–1104. https://doi.org/10.15866/irecos.v11i12.10956

Fahrudin, T., Buliali, J. L., & Fatichah, C. (2019). Enhancing the performance of smote algorithm by using attribute weighting scheme and new selective sampling method for imbalanced data set. *International Journal of Innovative Computing, Information and Control*, *15*(2), 423–444. https://doi.org/10.24507/ijicic.15.02.423

Gol, M., & Abur, A. (2015). A modified Chi-Squares test for improved bad data detection. *2015 IEEE Eindhoven PowerTech, PowerTech 2015*, *1*, 1–5. https://doi.org/10.1109/PTC.2015.7232283

Goopio, J., & Cheung, C. (2021). The MOOC dropout phenomenon and retention strategies. *Journal of Teaching in Travel and Tourism*, *21*(2), 177–197. https://doi.org/10.1080/15313220.2020.1809050

Gyoten, D., Ohkubo, M., & Nagata, Y. (2020). **Imbalanced data classification procedure based on SMOTE**. *Total Quality Science*, *5*(2), 64–71. https://doi.org/10.17929/tqs.5.64

Haugsbakken, H. (2020). Five Learning Design Principles to Create Active Learning for Engaging with Research in a MOOC. *European Journal of Open, Distance and E-Learning*, *23*(1), 32–45. https://doi.org/10.2478/eurodl-2020-0003

Hong, J. C., Hsiao, H. S., Chen, P. H., Lu, C. C., Tai, K. H., & Tsai, C. R. (2021). Critical attitude and ability associated with students' self-confidence and attitude toward "predict-observe-explain" online science inquiry learning. *Computers and Education*, *166*(February 2020). https://doi.org/10.1016/j.compedu.2021.104172

Insights, H. (2022). *WHAT ABOUT INDONESIA?* https://www.hofstede-insights.com/country/indonesia/

Issa, T., & Isaias, P. (2016). Internet factors influencing generations Y and Z in Australia and Portugal: A practical study. *Information Processing and Management*, *52*(4), 592–617. https://doi.org/10.1016/j.ipm.2015.12.006

Izzati, B. M., Adzra, S. S., & Saputra, M. (2024). Online Learning Acceptance in Higher Education during Covid-19 Pandemic: An Indonesian Case Study. *International Journal of Technology*, *15*(1), 207–218. https://doi.org/10.14716/ijtech.v15i1.5078

Japkowicz, N. (2000). Learning from imbalanced data sets: a comparison of various strategies. *AAAI Workshop on Learning from Imbalanced Data Sets*, 0–5.

Jayani, D. H. (2018). *Orang Indonesia Habiskan Hampir 8 Jam untuk Berinternet*. https://databoks.katadata.co.id/datapublish/2020/02/26/indonesia-habiskan-

hampir-8-jam-untuk-berinternet
Joo, Y. J., So, H. J., & Kim, N. H. (2018). Examination of relationships among students' self-determination, technology acceptance, satisfaction, and continuance intention to use K-MOOCs. *Computers and Education*, *122*(April 2017), 260–272. https://doi.org/10.1016/j.compedu.2018.01.003
Lambert, S. R. (2020). Do MOOCs contribute to student equity and social inclusion? A systematic review 2014–18. *Computers and Education*, *145*(November 2018), 103693. https://doi.org/10.1016/j.compedu.2019.103693
Liliana, L., Santosa, P. I., & Kusumawardani, S. S. (2022). Completion factor in massive open online course in developing countries: A literature review in 2015-2021. *World Journal on Educational Technology: Current Issues*, *14*(2), 456–472. https://doi.org/10.18844/wjet.v14i2.6919
Limanto, S., Buliali, J. L., & Saikhu, A. (2024). GLoW SMOTE-D: Oversampling Technique to Improve Prediction Model Performance of Students Failure in Courses. *IEEE Access*, *12*(November 2023), 8889–8901. https://doi.org/10.1109/ACCESS.2024.3351569
Lubis, A. H., Idrus, S. Z. S., & Rashid, S. A. (2020). The exposure of MOOC usage in Indonesia. *International Journal of Scientific and Technology Research*, *9*(2), 2716–2720. https://www.researchgate.net/profile/Andre-Lubis/publication/339400732_The_Exposure_Of_MOOC_Usage_In_Indonesia/links/5e4f9a37a6fdccd965b458e0/The-Exposure-Of-MOOC-Usage-In-Indonesia.pdf
Malakouti, S. M., Menhaj, M. B., & Suratgar, A. A. (2023). The usage of 10-fold cross-validation and grid search to enhance ML methods performance in solar farm power generation prediction. *Cleaner Engineering and Technology*, *15*(February), 100664. https://doi.org/10.1016/j.clet.2023.100664
McLafferty, S. L. (2016). Conduction Questionnare Survey. In N. Clifford, M. Cope, T. Gillespie, & S. French (Eds.), *Key Methods in Geography* (3rd ed., p. 129). SAGE. https://books.google.co.id/books?hl=en&lr=&id=7hcFDAAAQBAJ&oi=fnd&pg=PA129&dq=type+of+questionnaire&ots=TDKNuk5Rcv&sig=OV_yNcnV5h2YafM_ycfiFQhBibc&redir_esc=y#v=onepage&q=type of questionnaire&f=false
Naghipour, M., Ling, L. S., & Connie, T. (2024). A Review of AI Techniques in Fruit Detection and Classification: Analyzing Data, Features and AI Models Used in Agricultural Industry. *International Journal of Technology*, *15*(3), 585–596. https://doi.org/10.14716/ijtech.v15i3.6404
Nihan, S. T. (2020). Karl Pearsons chi-square tests. *Educational Research and Reviews*, *15*(9), 575–580. https://doi.org/10.5897/err2019.3817
Nurhayati, S., Luthfi, E. T., & Papua, U. Y. (2015). Prediksi Mahasiswa Drop Out Menggunakan Metode Support Vector. *Prediksi Menggunakan SVM*, *3*(6), 82–93.
Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., & Brunk, C. (1994). Reducing Misclassification Costs. *Proceedings of the 11th International Conference on Machine Learning, ICML 1994*, 217–225. https://doi.org/10.1016/B978-1-55860-335-6.50034-9
Pribadi, T. W., & Shinoda, T. (2022). Hand Motion Analysis for Recognition of Qualified and Unqualified Welders using 9-DOF IMU Sensors and Support Vector Machine (SVM) Approach. *International Journal of Technology*, *13*(1), 38–47. https://doi.org/10.14716/ijtech.v13i1.4813
Radha, S., & Nelson Kennedy Babu, C. (2020). Enterprise big data analysis using SVM classifier and lexicon dictionary. *International Journal of Enterprise Network Management*, *11*(1), 65–75. https://doi.org/10.1504/IJENM.2020.103913
Reparaz, C., Aznárez-Sanado, M., & Mendoza, G. (2020). Self-regulation of learning and

MOOC retention. *Computers in Human Behavior*, *111*(May). https://doi.org/10.1016/j.chb.2020.106423
Rezvani, S., & Wang, X. (2023). A broad review on class imbalance learning techniques. *Applied Soft Computing*, *143*(August). https://doi.org/10.1016/j.asoc.2023.110415
Ruipérez-Valiente, J. A., Halawa, S., Slama, R., & Reich, J. (2020). Using multi-platform learning analytics to compare regional and global MOOC learning in the Arab world. *Computers and Education*, *146*(November 2019). https://doi.org/10.1016/j.compedu.2019.103776
Sánchez-Prieto, J. C., Olmos-Migueláñez, S., & García-Peñalvo, F. J. (2016). Informal tools in formal contexts: Development of a model to assess the acceptance of mobile technologies among teachers. *Computers in Human Behavior*, *55*, 519–528. https://doi.org/10.1016/j.chb.2015.07.002
Sari, M., Berawi, M. A., Larasati, S. P., Susilowati, S. I., Susantono, B., & Woodhead, R. (2023). Developing Machine Learning Model to Predict HVAC System of Healthy Building: A Case Study in Indonesia. *International Journal of Technology*, *14*(7), 1438–1448. https://doi.org/10.14716/ijtech.v14i7.6682
Schölkopf, B. (2002). An Introduction to Support Vector Machines. In *Computing and Information Sciences: Recent Trends* (pp. 3–17). Narosa Publishing House. https://doi.org/10.1016/B978-044451378-6/50001-6
Sinta, T. Della, & Wahyuni, B. D. (2022). Kesenjangan Sosial dalam Mengakses Pendidikan di Indonesia. *Edukasia Multikultura*, *4*, 11–28.
Sukwika, T. (2018). Peran Pembangunan Infrastruktur terhadap Ketimpangan Ekonomi Antarwilayah di Indonesia. *Jurnal Wilayah Dan Lingkungan*, *6*(2), 115. https://doi.org/http://dx.doi.org/10.14710/jwl.6.2.115-130
Taghizadeh, S. K., Rahman, S. A., Nikbin, D., Alam, M. M. D., Alexa, L., Ling Suan, C., & Taghizadeh, S. (2021). Factors influencing students' continuance usage intention with online learning during the pandemic: a cross-country analysis. *Behaviour and Information Technology*, *0*(0), 1–20. https://doi.org/10.1080/0144929X.2021.1912181
Van De Oudeweetering, K., & Agirdag, O. (2018). MOOCS as Accelerators of Social Mobility? A Systematic Review. *Journal of Educational Technology & Society*, *21*(1), 1–11. https://doi.org/10.2307/26273863
Wang, L., Han, M., Li, X., Zhang, N., & Cheng, H. (2021). Review of Classification Methods on Unbalanced Data Sets. *IEEE Access*, *9*, 64606–64628. https://doi.org/10.1109/ACCESS.2021.3074243

Additional Note:
Similarity max 20% with Turnitin or iThenticate

[i]Corresponding author's email: name@ai.ue.oa, Tel.: +00-00-000000; Fax: +00-00-000000
doi: 10.14716/ijtech.v0i0.0000