INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN PROGRESSIVE EDUCATION & DEVELOPMENT

OPEN ACCESS JOURNAL

Vol 14, Issue 4, (2025) E-ISSN: 2226-6348

A Class Density-Weighted Gain Ratio Feature Selection for Multiclass Student Engagement Classification

Chong Ke Ting
Faculty of Computing, Universiti Teknologi Malaysia
Email: keting.phd@gmail.com

Noraini Ibrahim
Faculty of Computing, Universiti Teknologi Malaysia
Email: noraini_ib@utm.my

Sharin Hazlin Huspi
Faculty of Computing, Universiti Teknologi Malaysia
Email: sharin@utm.my

Wan Mohd Nasir Wan Kadir
Faculty of Computing, Universiti Teknologi Malaysia
Email: wnasir@utm.my

DOI Link: http://dx.doi.org/10.6007/IJARPED/v14-i4/26682

Published Online: 01 October 2025

Abstract

Educational Data Mining (EDM) uses vast educational datasets for discovering meaningful student participation patterns and academic achievements. Developing accurate multiclass classification models remains challenging due to it difficulties caused by class imbalance issues and irrelevant as well as redundant attributes. Filter-based feature selection methods demonstrate efficiency yet prove ineffective at resolving these problems so they create biased output performance which targets majority classes specifically. This study introduces Equitable Gain Ratio Feature Selection (EquiGR) which utilizes k-nearest neighbors to weight the class density levels for better minority group representation. The uses of Spearman Correlation Coefficient to detect and remove both strongly related redundant features along with low-ranking ones. The evaluation of proposed EquiGR method relied on four machine learning algorithms: Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM) and Logistic Regression (LR) as different learning paradigms for assessment. The experimental analysis of the imbalanced dataset with AE:PE:NE class distribution = 3624:4264:1183 showed EquiGR delivered better outcomes than baseline feature selection techniques for accuracy measures alongside precision and recall and F1-score metrics. The combination of RF with EquiGR reached 92.23% accuracy and 92.48% value for the NE-class F1-score. The proposed

method demonstrates effective enhancement of classification results while showing remarkable improvements for minority class predictions in educational predictive modeling scenarios.

Keywords: Student Engagement, Feature Selection, Class Imbalance, Machine Learning, Classification

Introduction

Due to the evolution of digitalize learning management system (LMS) in higher education has encourage the implementation of online learning. Therefore, student engagement has now become a significant interest for researcher as well as higher education itself. The accuracy of identifying the engagement level of students has become an important area in Educational Data Mining (Altaf et al., 2019; Zainol et al., 2021). Furthermore, due to the digital technologies getting more advance and the increase of LMS usage, large amount of educational data is generated which provide new opportunities and challenges for the classification. The exponential increase of educational data which include the learning management system, students' information database, and the transformation information into new insight is the largest challenges faced to optimized the benefit for students, lecturers, as well as administrators (Romero & Ventura, 2020).

While the researchers are studying on the improvement of student engagement classification, there are few challenges that faced in educational dataset. First, the class imbalance nature of educational dataset is one of the challenges that faced in educational data mining in optimizing the performance of student engagement level classification (Ayouni et al., 2021; Binali et al., 2021; Gledson et al., 2021; Orji et al., 2023). In the meanwhile, the overlapping of attribute values and their vulnerability due to human nature has made the classification process become more complex and presented obstacles for the development of extremely precise models for student engagement level classification (Al-Ashoor & Abdullah, 2022; Nand et al., 2020; Orji & Vassileva, 2020). The presence of redundant and irrelevant attributes in educational dataset causing the classification of student engagement level unable to be optimized.

A good feature selection is able to identify an appropriate subset and enhance the generalization abilities of the classification model (Al-Shabandar et al., 2019). By improving the quality of the training dataset, feature selection can optimize the performance of classification models and providing a faster, and less resource-intensive classification models (Chen et al., 2021). Even though the existing feature selection approaches is a significant step in data pre-processing to remove the irrelevant attributes from a dataset, they are facing some limitation (Al-Shabandar et al., 2019; Ramaswami et al., 2020). Feature selection can mainly divided into three categories which are filter-based, wrapper-based, and embedded-based (Cherrington et al., 2019).

Wrapper-based evaluates the subset features based on the classification outcome and embedded-based is involved in ML training, which limits its implementation in other learning algorithms, while filter-based is independent of the induction algorithm and is based on the general character of the data (Bolón-Canedo & Alonso-Betanzos, 2019). Filter-based feature selection is an approach that is good in dealing with irrelevant attributes and ensure the generalization of the attribute subset selected while using lower computational

cost. However, filter feature selections are not dealing well with the severely imbalanced samples, where the capabilities of minority classes are frequently being ignored during the process of feature selection (He et al., 2019). Moreover, they are facing the issues that it evaluates all the attributes in the dataset independently, removing those are not important without considering the attributes interrelated.

Therefore, in this research proposed an equitable gain ratio feature selection (EquiGR) to identify the most relevant attributes for classification of student engagement level while handling with the class imbalance and overlapping attribute values issues. This approach plays an important role in reducing the dimensions of attributes and increase the effectiveness of student engagement level classification model. The EquiGR is aimed to increase reliability and applicability of classification models and reduce the influence of class biases issues during the process of feature selection and remove the redundant attributes.

Literature Review

Educational Data Mining (EDM) is an interdisciplinary application of data mining that combines disciplines to provides insightful information about student engagement, academic achievement, and other aspects that necessitate in depth data analysis (Hasan et al., 2020; Luo & Wang, 2020). The identification of student engagement key elements is essential in EDM because it assist the strategies planning for enhancing experiences and institutional success (Karalar et al., 2021). Students can only undergo efficient online efficient online learning when they are actively engaging (Hu & Li, 2017). Therefore, higher education is increasing its effort to measure the students' engagement in online learning environment to indicate students' success (Nand et al., 2020). The lack of student engagement is the most critical reason that cause unfavourable online instruction perception, while the engagement of student in online learning is the critical elements for successful learning (Lasi, 2021; Tan et al., 2021).

However, the approach to improve student engagement is still an open research questions that needs further exploration (Orji & Vassileva, 2020). Hussain et al. (2018) emphasize that performance of classification algorithms for student engagement level can be affected due to the concerned depend on the selection of attributes. However, different course has different norm to constitute engaged student activities (Motz et al., 2019). Due to the lack of directly available predictors in LMS, different attributes are being utilized in different studies of student engagement level. Therefore, the selection of attributes is important in classifying student engagement level. Other than that, due to the educational data streaming, class imbalance is one of the issues that encountered in educational research that negatively impact the effectiveness of classification algorithms (Hassan et al., 2021; Palli et al., 2024). However, it is crucial to solve the problem of data imbalance to avoid unnecessary losses, while maintaining the characteristics of the data (Jamaluddin & Mahat, 2021). Even though class imbalance is a challenge in educational research, the issues of class imbalance is not focuses on the existing student engagement level research.

However, the existing filter feature selection approaches are fail to eliminate the redundant attributes since it of the attributes are weighted independently(Cherrington et al., 2019). Furthermore, severely imbalanced samples do not response well to the feature

selection approaches. This is because, it's possible to undervalue the capabilities of minor classes and neglect the distributional traits of minority class samples. Due to this, there will be a greater bias against those who belong to the majority classes (He et al., 2019). Moreover, Zou et al. (2021) mentioned that by considering the class imbalance difficulties, it is possible to provide a more accurate feature selection for multiclass classification (Zou et al., 2021).

Pooja (2024) mentioned that handling the class imbalance issues by implementing feature selection is getting more attention by recent research. This is because proper feature selection is one of the significant approach to resolve the bias-to-majority issues other than resampling technique (Liu et al., 2018). However, the standard guideline on dealing with the multidimensional issues with data imbalance does not exist (Bach & Werner, 2018). Therefore, there are numerous studies that analyze feature selection approaches that include the weight-adjustment mechanism to handle imbalanced classes for enhancing classification performance.

Li et al. (2018) created Weighted Gini Index as an imbalance dataset solution which modifies feature weights according to imbalance ratio. Using this approach creates enhanced sensitivity towards minority instances but it potentially makes the specific attributes appear more relevant when they primarily occur with minority cases that reducing the overall class discrimination. Furthermore, Ghosh et al. (2022) developed a binary differential evolution-based feature selection approaches which utilize mutual information together with Manhattan distance-based mutation. This research implements a weighting approach which balances class impact through inverse relationship between class weights and sample number distribution. However, the disproportionally increase weighting function might cause the selected attributes highly specific towards minority class but less relevant to overall separation.

In the research of Tiwari (2014) which applied ReliefF with weight-based implementation which enhances minority instances through the application of weighting factors. Weighting minority class instances for feature selection improves their representation but may cause overfitting when selecting attributes that highly specialize for the minority class at the expense of attributes necessary for separation between all classes. Sagoolmuang and Sinapiromsaran (2020) solved the challenges of overlapping instances in imbalanced dataset through the development of a class overlapping-balancing entropy gain ratio for decision tree. The approach minimizes the significance of minority class objects that overlap each other by giving them reduced scoring values in uncertain areas. This approach provides boundary noise reduction but its effectiveness against imbalanced multiclass dataset exists only when a single class occupies separate regions because overlapping minority classes decrease their boundary impact.

Objectives

The main research goal of this research is focusing on address the challenges posed by imbalanced data and redundant attributes in feature selection, which can significantly affect the performance of classification model. This paper aims to develop and evaluate a equitable gain ratio feature selection approach that effectively incorporates class imbalance information, ensuring a more balanced representation of both majority and minority classes while eliminating the redundant attributes. Therefore, the three research objectives for this

paper are mainly as follows: (1) to design and develop a feature selection approach that addresses the underestimation of attributes relevant to the minority class in imbalanced datasets, (2) To enhance the proposed feature selection method by incorporating redundancy attributes elimination through Spearman correlation analysis and (3) to evaluate and validate the performance of proposed feature selection approach using Random Forest, Support Vector Machine, Naïve Bayes, and Logistic Regression classifiers.

Methodology

In this research, EquiGR is proposed to overcome the limitation of existing gain ratio feature selection that do not consider class imbalance during the process of feature selection and fail to eliminate redundant attributes issues. Therefore, this proposed approach is mainly divided into two main phases which are weighted gain ratio which increase the weight of minority samples that enhance its importance during the process of feature selection and the Spearman Correlation Coefficient that implemented after phase 2 to eliminate the redundant attributes. The overall workflow for this research is as shown in Figure 1.

Data Collection

In this research, the student interactions and student demographic that collected from Moodle Learning Management System (LMS) and Student Information System at Universiti Teknologi Malaysia. The dataset is mainly made up of total 36 attributes, where 19 attributes are measuring the student interaction, while 17 of the attributes are measuring the student interaction in LMS. The total of 9071 students records are extracted, and the labelling approach proposed by Chong et al. (2023) is implemented to categorize and label the student engagement into three different levels. The extracted attributes are facing limitation including the missing of values particularly in the demographic dataset that collected from the SIS where the students' details of some samples are not recorded. Since the data is collected from eleven different educational course which include Technology & Information System, Discrete Structure, Programming Techniques I, Digital Logic, Database, System Analysis & Design, Data Structure & Algorithm, Network Communication, Human Computer Interaction, Artificial Intelligence, and Application Development that implemented diverse education contact might inherent biases due to the course structure or instructor-specific teaching styles that will influence the student behaviour patterns and engagement measurement. The distribution of student engagement level for AE:PE:NE is 3624:4264:1183. Furthermore, the class imbalance issue is also established in the dataset where the dataset is biased towards PE level, while the NE level is underrepresented. The details of student demographic and student interaction attributes are shown in respectively.

Data Pre-Processing

Student engagement classification received an enhancement through combining student interaction data with SIS student demographic data. When both sources are integrated that provide students's learning experience in its entirely. A structured database join using matric_no, year, course, and section attributes merged the datasets each time matric_no served as the unique key field for accurate record association.

However, the joined dataset consist of some missing values and inconsistencies. Therefore, data preprocessing is needed to improve the data quality and model performance.

1. Missing value imputation: The k-Nearest Neighbor (kNN) imputer is implemented to estimate and impute the missing values based on similar data points using Euclidean distance
2. Data encoding: Categorical attributes such as course, religion, and scholarship are encoded into numerical format by utilizing Python's LabelEncoder so that making the attributes are more compatible with machine learning algorithms.
3. Data Normalization: The numerical attributes are undergone Z-score normalization which standardize them to have 0 mean and 1 standard deviation so that the attributes with highly magnitude ranges could not become dominant during the learning process. The pre-processing steps prepared the dataset for the machine learning algorithm by improving its accuracy, consistency, and suitability for model training.

Figure 1 Overall research workflow

Table 1
Details of student demographic

Name of Attributes	Datatype	Description
Course	String	The subject the student enrolled in such as 'AAA', 'BBB', 'CCC' and so on.
Section	Integer	The specific class student is assigned to within a course, labelled as 1, 2, 3, and so on.
Pointer	Float	The student pointer achieve for the particular course.
Religion	String	The student's religious such as 'Buddha', 'Islam', 'Kristian', 'Hindu' and so on
Descent	String	The student's ancestral such as 'Melayu', 'Cina', 'India', Dusun (Sabah), 'Bangladeshi' and so on

Stu_year	Integer	The current academic year of the student, it represented as "1" for first-year, "2" for second-year, and so on.
Stu_course	String	The major course that student enrolled in such as SCSB, SCSJ, SCSR, SCSV and SCSP
Race	String	The student's racial or ethnic background such as 'Melayu Semenanjung', 'Cina', 'India', 'Bumiputera Sabah', 'Kaum Bukan Warganegara' and so on.
Scholarship	String	The type of scholarship that student receives such as 'PTPTN-JB', 'SELF SPONSOR', 'YAYASAN PELAJARAN', and so on.
Sem_No	Integer	The current semester number for student, it is a sequential number.
Father_Income	Float	The monthly income of the student's father in MYR.
Mother_Income	Float	The monthly income of the student's mother in MYR.
State	String	The geographical state that their current house located such as 'Johor', 'Perak', 'Sarawak' and so on.
Place_of_Birth	String	The state where the student was born such as 'Johor', 'Perak', 'Sarawak' and so on.
Accumulated_Credit	Float	The total number of academic credits the student has earned up to the current point in their studies.
Current_SEM_Credit	Float	The number of academic credits the student is registered for during the current semester.
Previous_Education	String	The details of the student's prior educational background, such as 'STPM', 'Matriculation' and so on.
Previous_Education_Result	Float	An attribute represents the pointer from the student's previous education.
Age	Integer	An attribute denotes the student's current age which is measured my year and month.

Table 2
Details of Student Interaction

Dimension Measure	**Name of attributes**	**Datatype**
Behavioural	No. of Login Clicks	Float
	No. of the individual assignment submitted	Float
	No. of the group assignment submitted	Float
	No. of quizzes completed	Float
	Total no. of assignment submitted	Float
Cognitive	No. of access to course material	Float
	Total time spend on individual assignment	Float
	Total time spend on group assignment	Float
	Average time spend on individual assignment	Float
	Average time spend on group assignment	Float
Social	No. of forum views	Float
	No. of forum participation	Float
Emotional	No. of individual assignment ontime	Float
	No. of individual assignment late	Float
	No. of group assignment ontime	Float
	No. of group assignment late	Float

The Figure 1 shows the overall workflow of this research which include the design and implementation of the proposed EquiGR and the evaluation and validation phase by using Random Forest, Support Vector Machine, Naïve Bayes, and Logistic Regression classifiers. In the meanwhile, the Table 1 and Table 2 are showing the student demographic and student interaction attributes that collected from the LMS and SIS respectively that will be used in this research for the experiment. The student demographic data is made up of string, integer, and float datatypes, while student interaction is mainly made up of float datatype.

Proposed Equitable Gain Ratio Feature Selection

The main objective of the research is to identify the best attributes subset to improve the multiclass imbalance issues in student engagement classification model. There are two tier processes executed, where first tier is to rank the attributes by implementing weighted gain ratio filter feature selection. Then, the correlation coefficient between the attributes is analyzed to identify the highly correlated attributes. This phase is using Python 3.7.7 and Linux as the software and platform for development.

First Tier Process

In the first tier, there are three stages involved to analyze the top rank attributes which are identification of nearest neighbor of minority class, compute the weight for each minority samples, and compute gain ratio for each attribute based on class weight. The flowchart for first tier of proposed equitable gain ratio feature selection is illustrated in Figure 2.

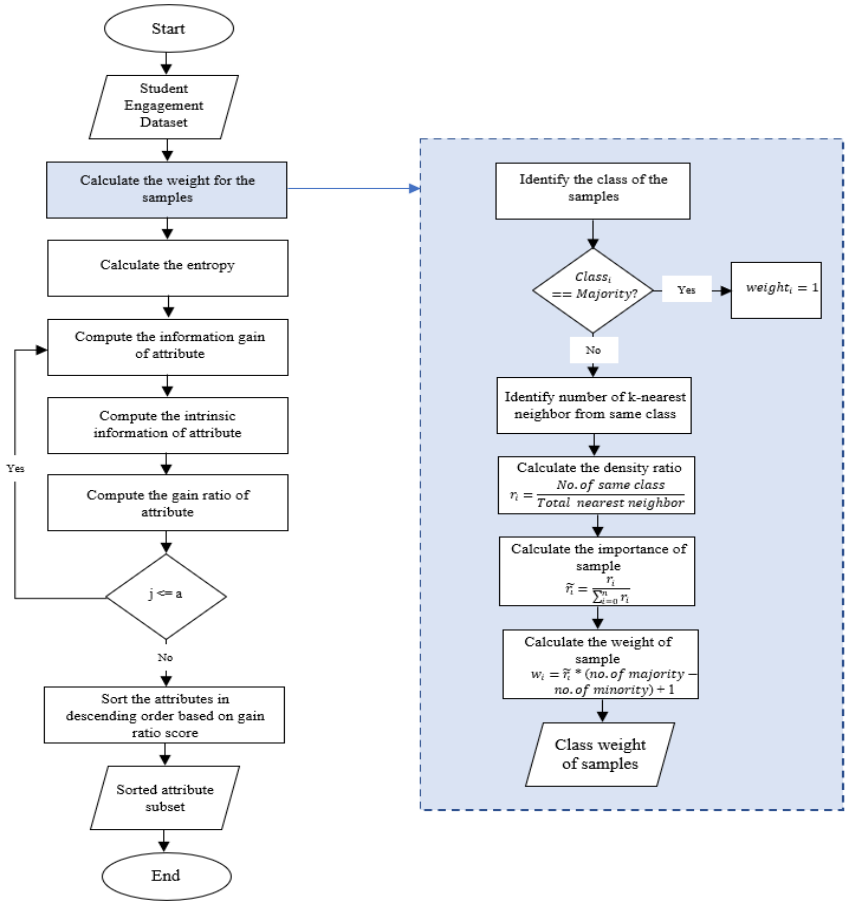


Figure 2 First Tier of Proposed Equitable Gain Ratio Flowchart

Figure 2 shows the first tier of EquiGR approach enhance the traditional gain ratio approach by addressing the class imbalance, with a focus on increasing the influence of minority class samples. Default weight values equal to one are assigned to the majority class. The weighting system provides higher significance values to minority class samples to overcome the negative impacts of unbalanced class distributions. The weight for each minority class gets determined through the k-Nearest Neighbour (kNN) method applying Minkowski distance to detect the local density for the minority class samples. The importance level resulting from minority class scoring gets adjusted by major class size differences and receives a weight increase of one to finalize the weight calculation. This weighting approach grants minority samples with higher significance during the feature selection phase. Then, the calculated weight serves as input for the gain ratio entropy computation process.

In imbalance dataset, the majority class tends to be dominated during the feature selection process, leading to the underrepresentation of minority class patterns. This step is aimed to increase the weight of minority samples based on their local density which making the minority samples become more influential during the feature selection process, ensure that the attributes that related to the minority classes are not ignored. This can result in a more balanced and fair evaluation of attribute importance, that consequently improve the classifiers performance, particularly for the minority class.

Second Tier Process

In the second tier, it involves the elimination of redundant attributes that can cause overfitting and reduced model interpretability. By removing the highly correlated attributes while retaining the most relevant one with higher weighted gain ratio score, this phase ensures the final attributes subset is both compact and informative. The flowchart for first tier of EquiGR is illustrated in

Figure 3.

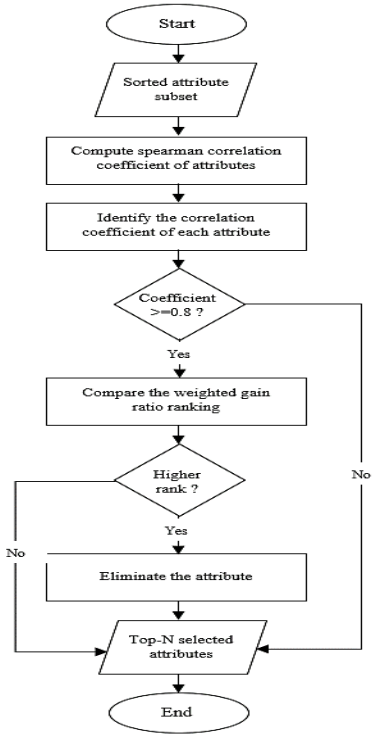


Figure 3 Second Tier of Proposed Equitable Gain Ratio Flowchart

Figure 3 shows the work flow of second tier of EquiGR where elimination of redundant attributes by assessing their correlation is being focused. The spearman correlation coefficient is implemented in this research to access the correlation between attributes. The Spearman correlation coefficients is less sensitive with uncertain data situation which involves non-linear patterns, outliers, or ordinal data. Firstly, the Spearman correlation coefficient is computed to measure the correlation between pairs of attributes. Then, identify the pairs of attributes with correlation coefficients above the threshold (>=0.8). This is because high correlation suggests that the attributes consist of similar information that will lead to redundancy in the dataset. Therefore, in the next step the highly correlated attributes with lower weighted gain ratio score that calculated previously are eliminated. This is to ensure that only the most informative attributes from each correlated pair kept in the final subset. After eliminating the redundant attributes, the final Top-N attributes are selected as the final subset for the prediction of student engagement level. The final subset will be both highly relevant based on the result of weighted gain ratio and non-redundant based on correlation analysis.

The elimination of redundant attributes with lower ranking is important to reduce attribute overlap which can cause vagueness on class boundaries and reduce the performance of classifiers while keeping the important information. It can help to improve the classifiers accuracy by focusing on the most informative attributes, reduces overfitting, and enhances computational efficiency. This consequently gives a clearer, and more discriminative attribute subset for better classification.

Machine Learning Classification

Machine learning classification can mainly divide into four category which are tree-based, structure-based, and probabilistic-based, statistical-based (Al-Shabandar et al., 2019; Tomasevic et al., 2020). Therefore, in order to evaluate the performance EquiGR in different based of machine learning approach, one of the popular machine learning approaches from each of the based are being selected to evaluate the proposed model. The Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), and Logistic Regression (LR) which representing the tree-based, structure-based, probabilistic-based and statistical-based respectively are implemented in this research to predict the student engagement level. The 5-fold cross validation is implemented to divide the data into training and testing dataset for the training of the machine learning approach. Besides, the educational data mining commonly use performance measure which include accuracy, precision, recall, and F1-measure are used in this research to evaluate and validate the performance of the proposed approach. Accuracy is implemented is used in this research since is a commonly used performance measure used in educational data mining, however is might cause misleading when handling with imbalanced data. Therefore, the precision, recall, and f-score were computed for each class. Precision is used to indicate the reliable of the classification result for each class, while recall is used to understand the sensitivity of the model in classifying each of the student engagement level (Ayouni et al., 2021; Kabathova & Drlik, 2021). More importantly, F-score is the harmonic mean of precision and recall, it is implemented to assess the effectiveness of a classifier by considering both precision and recall rates in a complete manner (Zheng et al., 2020).

Result and Discussion

The performance measure which includes accuracy, precision, recall, and f-score in Actively Engaged, Passively Engaged, and Not Engaged by implementing different feature selection approaches which include gain ratio, first tier equiGR, and equiGR with second tier enhancement. The performance measures are shown in Table 3.

Table 3

Performance Measure of classifiers by implementing different feature selection approach

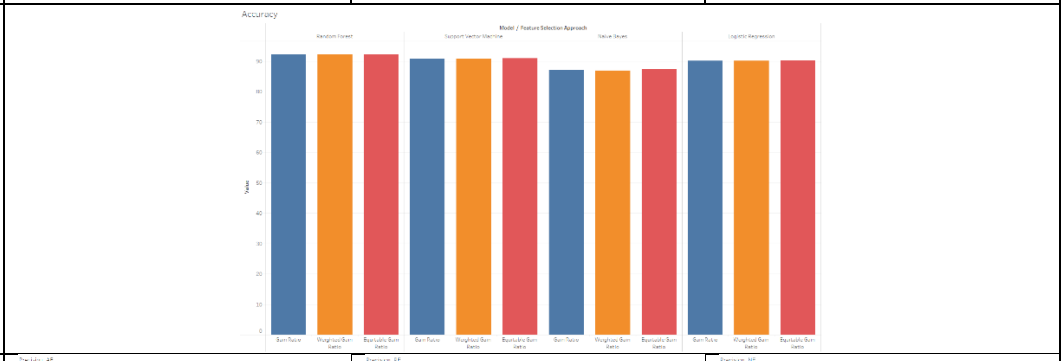
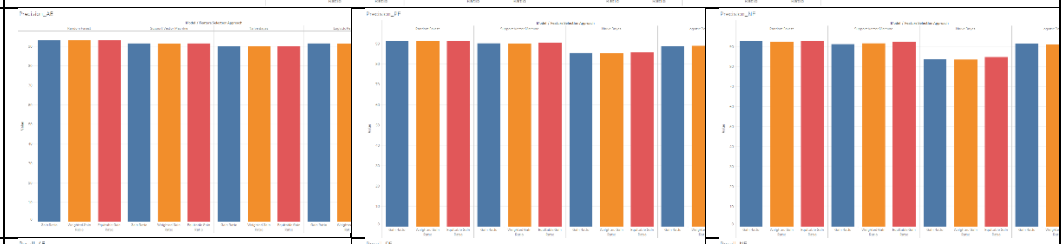
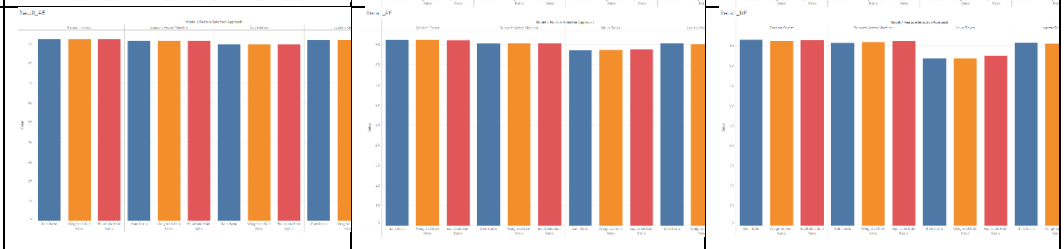
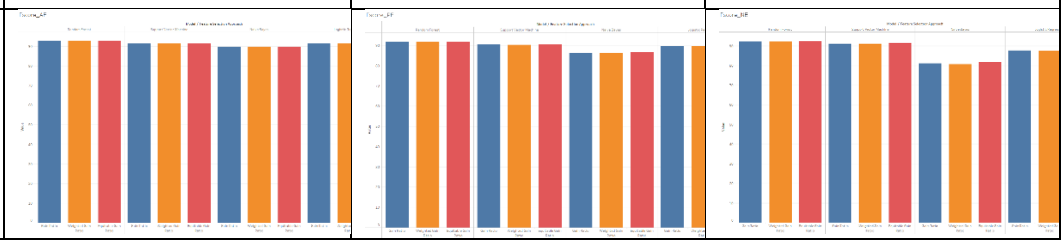
Performance Measure	Actively Engaged	Passively Engaged	Not Engaged
Accuracy			
Precision			
Recall			
F-score			

Table 3 shows the comparative analysis of Gain Ratio, Weighted Gain Ratio, and Equitable Gain Ratio Feature Selection with is represent by blue, orange, and red respectively by implementing Random Forest, Support Vector Machine, Naïve Bayes, and Logistic Regression respectively. The data evaluation occurred on unbalanced dataset featuring a distribution pattern of AE:PE:NE = 3624:4264:1183. The performance metrics included accuracy, precision, recall, and f-score to assess the performance of classifiers with different feature selection approach.

Gain Ratio feature selection act as the baseline feature selection approach, the overall classification performance was moderate across all models. The RF model achieved 92.17% accuracy and 92.22% f- score when GR was used for feature selection of the NE class. While the overall metrics appeared reasonable, a closer inspection revealed that the model exhibited reduced the recall towards NE class, with the most serious drop to 78.71% in NB classifiers. The performance metrics of LR and NB deteriorated to the extent that their F1-score reached below 81.05% after implementing the GR approach that does not consider minority weight and redundant attributes in the dataset induced generalization difficulties.

From the first tier of EquiGR GR through the incorporating weights that consider for class imbalance, which help to increase the importance of minority class during the process of feature selection with utilizing the k-NN local density estimation. SVM performance with WGR achieved better results as compared to GR because WGR enhanced the NE class increasing the precision from 91.23% to 92.21% which proved WGR capability on reducing the unimportant attributes with considering the importance of minority class. Besides, it also achieves the accuracy of LR to 90.18% while achieving NE class recall of 81.41% whereas GR achieved only 81.43%. The enhancements WGR introduced proved to be incremental while in some cases, insufficient to overcome the redundant attributes.

With the enhancement of EquiGR in second tier, where high redundant attributes with lower ranking attribute elimination demonstrated superior performance than GR and first tier of EquiGR when evaluated through all classification models. Its class-aware and redundancy-reducing approach led to balance feature subsets that delivered improvement model performance specifically for the NE class categories. The RF model using EquiGR delivered 92.23% accuracy, while NE class F-score rise to 92.48% and its precision reached 92.65% and recall received a 92.34%. In SVM the NE class F1-score of EquiGR reached 91.65% while WGR produced a score of 90.96%. The results demonstrated the advantages of EquiGR.

GR provides simple interpretable feature selection at first but the method fails against imbalanced datasets that remove important features from minority groups. This issue becomes somewhat manageable through EquiGR demonstrates exceptional capability balancing the importance of majority and minority class which boost classification results for every class group especially for sparse NE category. The results achieved by EquiGR with machine learning classifiers prove its ability to identify attribute which are relevant and non-redundant as well as well distributed thus demonstrating its excellence for imbalanced data classification.

The proposed EquiGR approach presents a novel contribution to gain ratio feature selection on handling with the class imbalance and redundant datasets by integrating the minority class weighting through k-NN local density estimation and redundancy elimination via Spearman correlation. The proposed approach has successfully improved the quality of the attribute subset and consequently improve the performance of the classifiers. Unlike traditional gain ratio approaches that treat all the samples equally, EquiGR incorporates the kNN with Minkowski distance to estimate the local density of the minority samples, therefore assigning greater significance to the minority samples that are in the denser regions. The proposed weighting approach ensure that informative attributes for minority class are not

undervalues, which directly overcoming the key limitation of the traditional gain ratio approaches.

Furthermore, EquiGR enhances the attribute subset quality by eliminating the redundant attributes with lower ranking using Spearman correlation coefficient, where only the most informative attribute from each highly correlated pair is kept. The two enhance approach in gain ratio feature selection not only improve the discriminatory power of the selected attributes but also contributes to a more compact and generalizable model. The integration of class-aware weighting and redundancy elimination into EquiGR from existing gain ratio feature selection approach, shows its potential in improving the classifiers performance.

As mentioned by Feng et al. (2020), the increase of the minority class weight in smaller sample sizes through various feature selection approaches has been proven with its effectiveness in enhancing the classification performance on imbalance datasets. However, this weighted approach is facing some limitations. Firstly, the effectiveness of the weighting approach is highly depended on the class distribution. In the case where the dataset is extremely imbalance, the weighted feature selection might not perform well unless combined with other techniques such as data resampling approach (Tsai et al., 2024). Furthermore, the weighted approach can be sensitive to noise, which may skew the results and affect the performance of the classification when the data is not well cleaned.

In the meanwhile, the implementation of k-Nearest Neighbors (k-NN) for local density computation to increase the weight of minority samples presents challenges due to its sensitivity to the choice of *k*. A small *k* may lead to overfitting, while large *k* can complicate the important local structures, making it difficult to identify the density peaks accurately. Moreover, k-NN is sensitive towards noise and irregular density distributions, which can distort the local density estimation and result in incorrect cluster assignments (Hou et al., 2024). In high-dimensional spaces, the distance between points becomes less meaningful due to data sparsity, further complicating the accurate local density estimation (Adarsh M.D, 2024).

In future, the hybrid approaches that combine EquiGR with resampling such as SMOTE and ADASYN to handle the extreme class imbalances issues Meanwhile, the approaches which include denoising encoder or robust distance metrics can be studied to improve the k-NN based density estimation in noisy environments. Furthermore, the alternative weighting approach such as cost-sensitive learning and adaptive samples reweighting approach need to be studied in future to overcome the influence of minority class samples during feature selection more effectively.

Conclusion

This investigation developed an Equitable Gain Ratio Feature Selection (EquiGR) system to solve problems from both class imbalance and attributes redundancy within classification tasks that deal with minority classes. The research evaluated the EquiGR approach by testing it against two standard feature selection approaches namely GR and through RF, SVM, NB, and LR models. Experimental data modeling represented a realistic case of class imbalance because statistics showed AE:PE:NE = 3624:4264:1183. The

application of Gain Ratio proved rapid with straightforward interpretation but its framework failed to attract features linked to the scarce NE category thereby producing deficient outcome predictions. WGR achieved a slight improvement with its class-distribution-based weighing mechanisms which produced improved results for the detection of minority classes. The proposed EquiGR approach reached superior results than GR and WGR methods in every evaluation metric and delivered remarkable improvements to NE class recall and F1-score measurements. When RF was used with EquiGR able to give 92.23% accuracy and 92.48% F1-score for NE class detection which implies major improvements both in recognition precision and class sensitivity. The experiment produced comparable positive effects on all tested classification systems that demonstrated EquiGR has general applicability. The assessment highlights the need for feature selection approaches which feature both class weight and reduction of feature redundancy while going past information gain measurements. EquiGR serves as an effective procedure for improving machine learning model performance when working with unbalanced datasets. In future other domains that often facing with the class imbalance and redundant attributes issues such as medical diagnosis, fraud detection, and cybersecurity detection should be investigated for applying EquiGR specifically multi-class problems. Besides, other ensemble-based, neural network, and instance-based classifiers such as XGBoost, LightGBM, Multilayer Perceptron, Artificial Neural Networks, and kNN classifiers can be implemented in future to evaluate and validate the EquiGR approach.

References

Adarsh M.D, P. K. K. (2024). Addressing K-Nn Limitations Through Boosted Multi-Algorithm Nearest Neighbour Ensembles. 2024 7th International Conference on Circuit Power and Computing Technologies (ICCPCT),

Al-Ashoor, A., & Abdullah, S. (2022). Examining Techniques to Solving Imbalanced Datasets in Educational Data Mining Systems. *Int. J. Comput*, *21*(2), 205-213. https://doi.org/10.47839/ijc.21.2.2589

Al-Shabandar, R., Hussain, A. J., Liatsis, P., & Keight, R. (2019). Detecting at-risk students with early interventions using machine learning techniques [Article]. *IEEE access*, *7*, 149464-149478, Article 8847304. https://doi.org/10.1109/ACCESS.2019.2943351

Altaf, S., Soomro, W., & Rawi, M. I. M. (2019). *Student Performance Prediction using Multi-Layers Artificial Neural Networks: A Case Study on Educational Data Mining* Proceedings of the 2019 3rd International Conference on Information System and Data Mining, Houston, TX, USA. https://doi-org.ezproxy.utm.my/10.1145/3325917.3325919

Ayouni, S., Hajjej, F., Maddeh, M., & Al-Otaibi, S. (2021). A new ML-based approach to enhance student engagement in online environment. *PLOS ONE*, *16*(11), e0258788. https://doi.org/10.1371/journal.pone.0258788

Bach, M., & Werner, A. (2018). Cost-Sensitive Feature Selection for Class Imbalance Problem. Information Systems Architecture and Technology: Proceedings of 38th International Conference on Information Systems Architecture and Technology – ISAT 2017, Cham.

Binali, T., Tsai, C.-C., & Chang, H.-Y. (2021). University students' profiles of online learning and their relation to online metacognitive regulation and internet-specific epistemic justification. *Computers & Education*, *175*, 104315. https://doi.org/10.1016/j.compedu.2021.104315

Bolón-Canedo, V., & Alonso-Betanzos, A. (2019). Ensembles for feature selection: A review and future trends. *Information Fusion*, *52*, 1-12. https://doi.org/10.1016/j.inffus.2018.11.008

Chen, L.-q., Wu, M.-t., Pan, L.-f., & Zheng, R.-b. (2021). Grade prediction in blended learning using multisource data. *Scientific Programming*, *2021*(1), 4513610. https://doi.org/10.1155/2021/4513610

Cherrington, M., Airehrour, D., Lu, J., Thabtah, F., Xu, Q., & Madanian, S. (2019). Particle swarm optimization for feature selection: A review of filter-based classification to identify challenges and opportunities. 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON),

Chong, K. T., Ibrahim, N. B., & Huspi, S. H. B. (2023). Multiclass Student Engagement Level Prediction using Belief-Rule Based Labelling. 2023 Sixth International Conference of Women in Data Science at Prince Sultan University (WiDS PSU),

Feng, F., Li, K. C., Shen, J., Zhou, Q., & Yang, X. (2020). Using Cost-Sensitive Learning and Feature Selection Algorithms to Improve the Performance of Imbalanced Classification. *IEEE access*, *8*, 69979-69996. https://doi.org/10.1109/ACCESS.2020.2987364

Ghosh, S. K., Janan, F., & Ahmad, I. (2022). Application of the Classification Algorithms on the Prediction of Student's Academic Performance. *Trends in Sciences*, *19*(14), 5070-5070. https://doi.org/10.48048/tis.2022.5070

Gledson, A., Apaolaza, A., Barthold, S., Günther, F., Yu, H., & Vigo, M. (2021). Characterising Student Engagement Modes through Low-Level Activity Patterns. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 88–97). Association for Computing Machinery. https://doi.org/10.1145/3450613.3456818

Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., Sarker, K. U., & Sattar, M. U. (2020). Predicting student performance in higher educational institutions using video learning analytics and data mining techniques [Article]. *Applied Sciences (Switzerland)*, *10*(11), Article 3894. https://doi.org/10.3390/app10113894

Hassan, H., Ahmad, N. B., & Sallehuddin, R. (2021). An Empirical Study to Improve Multiclass Classification Using Hybrid Ensemble Approach for Students' Performance Prediction. In (Vol. 724, pp. 551-561): Alfred, R., Iida, H., Haviluddin, H., Anthony, P. (eds) Computational Science and Technology.

He, Y., Zhou, J., Lin, Y., & Zhu, T. (2019). A class imbalance-aware Relief algorithm for the classification of tumors using microarray gene expression data. *Computational biology and chemistry*, *80*, 121-127. https://doi.org/10.1016/j.compbiolchem.2019.03.017

Hou, P., Zhou, L., & Yang, Y. (2024). Density clustering method based on k-nearest neighbor propagation. *Journal of Physics: Conference Series*, *2858*(1), 012041. https://doi.org/10.1088/1742-6596/2858/1/012041

Hu, M., & Li, H. (2017). Student engagement in online learning: A review. 2017 International Symposium on Educational Technology (ISET),

Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. R. (2018). Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores. *Computational intelligence and neuroscience*, *2018*(1), 6347186. https://doi.org/10.1155/2018/6347186

Jamaluddin, A. H., & Mahat, N. I. (2021). Validation assessments on resampling method in imbalanced binary classification for linear discriminant analysis. *Journal of Information and Communication Technology*, *20*(1), 83-102. https://doi.org/10.32890/jict.20.1.2021.6358

Kabathova, J., & Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques [Article]. *Applied Sciences (Switzerland)*, *11*(7), Article 3130. https://doi.org/10.3390/app11073130

Karalar, H., Kapucu, C., & Guruler, H. (2021). Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system. *International Journal of Educational Technology in Higher Education*, *18*(1), Article 63. https://doi.org/10.1186/s41239-021-00300-y

Lasi, M. b. A. (2021). Online Distance Learning Perception and Readiness During Covid-19 Outbreak: A Research Review. *Development*, *10*(1), 63-73. https://doi.org/10.6007/IJARPED/v10-i1/8593

Li, K., Yu, M., Liu, L., Li, T., & Zhai, J. (2018). Feature Selection Method Based on Weighted Mutual Information for Imbalanced Data. *International Journal of Software Engineering and Knowledge Engineering*, *28*(08), 1177-1194. https://doi.org/10.1142/s0218194018500341

Liu, H., Zhou, M., Lu, X. S., & Yao, C. (2018). Weighted Gini index feature selection method for imbalanced data. 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC),

Luo, J., & Wang, T. (2020). *Analyzing Students' Behavior in Blended Learning Environment for Programming Education* Proceedings of the 2020 The 2nd World Symposium on Software Engineering, Chengdu, China. https://doi-org.ezproxy.utm.my/10.1145/3425329.3425346

Motz, B., Quick, J., Schroeder, N., Zook, J., & Gunkel, M. (2019). The validity and utility of activity logs as a measure of student engagement. Proceedings of the 9th international conference on learning analytics & knowledge,

Nand, R., Chand, A., & Naseem, M. (2020). Analyzing students' online presence in undergraduate courses using Clustering. 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Gold Coast, Australia.

Orji, F., & Vassileva, J. (2020). Using Machine Learning to Explore the Relation Between Student Engagement and Student Performance. 2020 24th International Conference Information Visualisation (IV),

Orji, F. A., Fatahi, S., & Vassileva, J. (2023). Data-Driven Approach for Student Engagement Modelling Based on Learning Behaviour. International Conference on Human-Computer Interaction,

Palli, A. S., Jaafar, J., Gilal, A. R., Alsughayyir, A., Gomes, H. M., Alshanqiti, A., & Omar, M. (2024). Online Machine Learning from Non-stationary Data Streams in the Presence of Concept Drift and Class Imbalance: A Systematic Review. *Journal of Information and Communication Technology*, *23*(1), 105-139. https://doi.org/10.32890/jict2024.23.1.5

Pooja, K. S. (2024). Machine Learning Advancements In Education: An In-Depth Analysis And Prospective Directions. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(3), 3229–3237. https://www.ijisae.org/index.php/IJISAE/article/view/5928

Ramaswami, G. S., Susnjak, T., Mathrani, A., & Umer, R. (2020). Predicting Students Final Academic Performance using Feature Selection Approaches. 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE 2020,

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, *10*(3), e1355. https://doi.org/10.1002/widm.1355

Sagoolmuang, A., & Sinapiromsaran, K. (2020). Decision tree algorithm with class overlapping-balancing entropy for class imbalanced problem. *International Journal of Machine Learning and Computing*, *10*(3), 444-451. https://doi.org/10.18178/ijmlc.2020.10.3.955

Tan, K. H., Chan, P. P., & Mohd Said, N.-E. (2021). Higher education students' online instruction perceptions: A quality virtual learning environment. *Sustainability*, *13*(19), 10840. https://doi.org/10.3390/su131910840

Tiwari, D. (2014). Handling class imbalance problem using feature selection. *International Journal of Advanced Research in Computer Science & Technology*, *2*(2), 516-520. https://doi.org/10.1007/978-981-99-2602-2_30

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction [Article]. *Computers and Education*, *143*, Article 103676. https://doi.org/10.1016/j.compedu.2019.103676

Tsai, C.-F., Chen, K.-C., & Lin, W.-C. (2024). Feature selection and its combination with data over-sampling for multi-class imbalanced datasets. *Applied Soft Computing*, *153*, 111267. https://doi.org/10.1016/j.asoc.2024.111267

Zainol, S. S., Hussin, S. M., Othman, M. S., & Zahari, N. H. M. (2021). Challenges of online learning faced by the B40 income parents in Malaysia. *International Journal of Education and Pedagogy*, *3*(2), 45-52.

Zheng, Y., Gao, Z., Wang, Y., & Fu, Q. (2020). MOOC Dropout Prediction Using FWTS-CNN Model Based on Fused Feature Weighting and Time Series [Article]. *IEEE access*, *8*, 225324-225335, Article 9296213. https://doi.org/10.1109/ACCESS.2020.3045157

Zou, Y., Hu, X., Li, P., & Li, J. (2021). Multi-label streaming feature selection via class-imbalance aware rough set. 2021 International Joint Conference on Neural Networks (IJCNN),