Subject Stream Recommender System for Sri Lankan Advanced Level Students Using Data Mining Techniques

D.N.H. Munasingha, C.P. Wijesiriwardana, and W.G.C.W. Kumara

Abstract: Subject stream selection at the General Certificate of Education Advanced Level (G.C.E. (A/L)) is a critical decision for students in Sri Lanka, influenced by numerous internal and external factors. This study focuses on a Subject Stream Recommender System that leverages data mining techniques to enhance the decision-making process for students based on internal factors, specifically academic performance. The proposed system consists of three main stages: data preprocessing, training, and testing. The dataset, comprising academic records of 1,000 students from two national schools in the Southern Province, was pre-processed to remove irrelevant attributes, handle missing values, and encode categorical data. Three data balancing techniques namely, Resample, SpreadSubSample, and SMOTE, were applied to address class imbalance. The study was conducted in two experiments using different attribute sets. The experiments evaluated the performance of three classification algorithms—Random Forest (RF), Decision Tree (DT), and Naive Bayes (NB)—on both the original and balanced datasets. Results indicate that the Resample technique consistently produced the highest accuracy across both experiments. Random Forest algorithm, combined with Resample, achieved an accuracy of 88.61% in Experiment 1 and 93.32% in Experiment 2, making it the most suitable model for the recommender system. It is suggested that academic performance, particularly in Science and Mathematics, significantly impacts subject stream selection. The study concludes that the Random Forest algorithm with the Resample balancing technique provides the most accurate predictions, answering the research questions concerning algorithm selection, data balancing, and relevant predictors for the Subject Stream Recommender System.

***Index Terms*—recommendation system, data mining, data imbalance**

I. Introduction

UNITED Nations says “education must be the main priority of the global, political, and development agendas” and “education is not only a right but a path that leads to human development which creates opportunities and freedom”. Article 26 of the 1948 Universal Declaration of Human Rights asserts that each individual is entitled to education. This right encompasses free education, particularly at the primary and essential levels. It also stipulates that elementary education must be compulsory. Furthermore, the article emphasises the importance of providing widespread access to technical, professional, and higher education, ensuring equal opportunities for all based on merit.

D.N.H. Munasingha is with the Department. of Information Technology, Faculty of Information Technology, University of Moratuwa, Sri Lanka (e-mail: munasinghadnh.20@uom.lk).

C.P. Wijesiriwardana is with the Department. of Information Technology, Faculty of Information Technology, University of Moratuwa, Sri Lanka (e-mail: chaman@uom.lk).

W.G.C.W. Kumara is with the Department of Computer Science and Engineering, South Eastern University of Sri Lanka, Sri Lanka (e-mail: chinthakawk@seu.ac.lk).

The education system in the world consists of nine levels: childhood education, primary education, lower secondary education, upper secondary education, postsecondary education, tertiary education, bachelor’s, master’s, and doctoral level [1]. Many Countries in the world categorise their formal Education into many levels. Table I shows the Education systems and their categorizations in 3 countries.

TABLE I
EDUCATION SYSTEM IN SELECTED COUNTRIES

Country	Category	Age (Years)	School Level (Grades)
USA	Pre	4 – 5	Pre School
	Elementary	5 – 10	Kindergarten – 4
	Middle	10 – 14	5 – 8
	High	14 – 18	9 – 12
UK	Primary	5 – 11	1 – 6
	Secondary	11 – 16	7 – 11
	Further	16 – 18	12 – 13
China	Pre	3 – 5	Kindergarten
	Primary	6 – 11	1 – 6
	Junior (Lower) Secondary	12 – 14	7 – 9
	Senior (Upper) Secondary	15 – 17	10 - 12

Every student in Sri Lanka should follow the mandatory education in primary and secondary education. Education

ISSN: 2961-5992

provided in Sri Lankan schools can be divided into four levels: primary education, junior secondary education, senior secondary first stage, and senior secondary second stage [2]. As shown in Table II, at the end of each stage of education, the government conducts a general examination. Grade 5 Scholarship Examination is held after completing primary Education, the General Certificate of Education Ordinary Level Examination (OL) is held after the senior secondary first stage, and the General Certificate of Education Advanced Level Examination (AL) is held after the senior secondary second stage (referred as OL and AL hereafter).

TABLE II
EDUCATION SYSTEM IN SRI LANKA

Age (Years)	Name	Grades
	G.C.E Advanced Level Examination	
17 - 18	Senior Secondary Second Stage	12 - 13
	G.C.E Ordinary Level Examination	
15 - 16	Senior Secondary First Stage	10 - 11
11 - 14	Junior Secondary Education	6 - 9
	Grade 5 Scholarship Examination	
6 - 10	Primary Education	1 - 5
4 - 6	Pre School	Pre School

The AL held at the end of the senior secondary education stage is very important to students as the results determine the next subject stream. In the senior secondary education stage, students must choose a subject stream and subjects according to their preferences. This subject stream will determine the future career path. Subject streams for senior secondary education are selected based on the results of the OL. Samaranayake and Caldera, 2012 found a direct relationship between OL and AL results [3].

The AL examination in five main subject streams: Physical Science, Biological Science, Commerce, Technology, and Arts [4]. Under each subject stream, students must study three subjects relevant to the stream. Apart from these subjects, all the students are required to take General English and Common General Knowledge subjects. The selection to the university is based on the grades scored at the AL for the three subjects relevant to the stream. Over time, a range of subjects has been introduced to the field of education through the expansion of scopes and the emergence of new subjects. Students must choose a subject stream and 3 subjects for their AL.

Table III shows the general and special qualifications for applying for AL subject streams. According to the education circulars issued by the Ministry of Education, the minimum criterion for applying to the AL is six core subjects including Language and Mathematics, and obtaining credits for three subjects. Students must complete special criteria for each subject stream [5], [6].

TABLE III
GENERAL AND SPECIAL QUALIFICATIONS FOR APPLYING FOR AL SUBJECT STREAMS IN SRI LANKA [7], [8]

A/L Stream	General Criteria	Special Criteria
Science	Should Pass the G.C.E O/L examination in Six core subjects including Language and Mathematics. And should obtain Credits for three subjects.	A Credit passes for science and should pass Mathematics.
Mathematics (Physical science)		A Credit pass (C) for Mathematics and should pass Science.
Technology		Should pass the Subjects Mathematics and Science
Commerce		A Credit pass (C) in Either Mathematics, Entrepreneur Education or Business studies and Accountancy
Arts		A Credit pass (C) in one of the subjects you wish to follow in A/L.

The choice of a suitable subject stream should be made with understanding and the intention of reaching goals for the future. Students may prefer to study particular subjects over others depending on their preferences and areas of interest. Sadly, they might not fully comprehend each, including the depth and scope of content covered and the subject's future value. As a result, they will be unable to make an informed choice that would increase their chances of succeeding academically and enjoyment of academics.

Picking unrelated subjects that do not go well together might result in a student failing or quitting school. Choosing the best subject stream and subjects for the AL is a problem that many students face. Although some students have a clear understanding and aim in subjects for AL, the majority of students do not have a clear understanding. A student's performance in Science and Mathematics subjects in OL directly impacts the performance in AL [7].

This paper proposes a recommender system, using different data mining algorithms, for predicting suitable subject streams for AL students. Predicting the suitable subject stream will be based on student performance in the OL and student academic performance in term tests in Grades 10 and 11.

The facts mentioned before have led to the following Research Questions.

- RQ 1: How to identify the most accurate data mining classification algorithm for the Recommender System?
- RQ 2: What is the best data-balancing technique for our dataset?
- RQ 3: Which features are selected to be the most relevant predictors for subject stream selection actions?
- RQ 4: Can one develop a reasonably accurate recommender system to predict students' subject stream according to their academic performance?

ISSN: 2961-5992

II. LITERATURE REVIEW

A. Educational Data Mining

Identification of patterns and relationships of a large dataset from different perspectives to get useful information is known as data mining. Educational Data mining (EDM) is applying data mining techniques to the specific types of data from the education system and implementing the data mining techniques to find a solution for the problems associated with the Education System [8].

EDM improves educational processes by improving performance by using data mining technologies to describe pedagogical approaches for future decision-making [9].

Various applications in EDM allow improvement and enhancement of the quality of the learning and teaching process [10]. The applications of EDM can target learners, educators, educational administrators, researchers, etc. There are most common methods which are used in EDM such as classification and prediction, clustering, outlier detection, relationship mining, social network analysis, process mining, text mining, the distillation of data for human judgment and discovery with models, etc. [11]. Classification, prediction, and clustering can be commonly used in many EDM practices [12]. According to Papamitsiou and Economides, there are six main applications of EDM, namely, student behavior modeling, prediction of academic performance, an increase of self-reflection and self-awareness, prediction of drop-out and retention, improvement feedback and assessment services, and recommendation of resources [13].

B. Educational Recommender Systems

TABLE IV
RELATED WORK OF EDUCATIONAL RECOMMENDER SYSTEMS

Research Work	Recommending Element	Scope	User	Instances	Data mining Algorithm	Metrics
[14]	Academic Guidance	HE	Student	330	DT, NB, and NN.	Accuracy Kappa
[15]	Academic advice	HE	Student	200	One R, Zero R, JRIP, PART, J48, Random tree, REP tree, and Decision stump.	Accuracy
[16]	Course Recommendation	HE	Student	8700	LR, NB, SVM, DT, *k*NN	Accuracy
[17]	School Recommendation	HE	Student	Not mentioned	*k*NN	Accuracy
[18]	Subject Recomm=endation	HE	Student	6948	LR, SVM, *k*NN, RF	Accuracy F1 Score
Proposed Approach	Subject Stream Recommendation	HE	Student	767	DT, RF, NB	Accuracy

HE: Higher Education

A summary of the previous research on educational recommender systems is provided in Table IV. The potential of using EDM for academic advice was evaluated by examining socioeconomic data, academic records, and the motivation for a course [14]. The suggested model is tested using Decision Trees (DT), Neural Networks (NN), and Naive Bayes (NB) algorithms using data from 330 Moroccan students. Accuracy and kappa values were used to assess the performance of each data mining algorithm. According to their research, the DT performs best, with an accuracy of 54.71% and kappa values of 0.354. They go on to say that the rules produced by DT offer useful information for wise academic judgments.

Mobasher, Shawish, and Ibrahim in 2017 proposed a recommender system to assess the student's demographic data, and study-related and psychological characteristics to predict academic performance prediction using data mining techniques [15]. This system highlights the student's weak points and provides appropriate recommendations by analyzing 200 student details in Egypt. The researchers trained the dataset under four rule induction algorithms: One R, Zero R, JRIP, and PART; and four decision algorithms: J48, Random tree, REP tree, and Decision stump. In their research, they stated that REP Tree outperforms other algorithms with a prediction accuracy of 73.6% and the lowest prediction accuracy was for Zero R with 45.5%. They conclude that the student's academic details, educational-related attributes, and psychological characteristics can be considered when creating the recommender system.

By proposing appropriate courses to students, [16] aimed to address the issue of university students being misdirected in courses and reduce the waste of course vacancies. 8,700 student records from two Nigerian institutions were used in training and testing. The dataset was trained using the DT, LR, NB, SVM, and k-NN algorithms. They discovered throughout their investigation that every algorithm used was effective, with all of them having an efficiency above 90%. RMSE of LR is nearly zero ($2.614 \times 10-14$). The accuracy of NB and SVM is 99.94%, DT is 98.10%, and k-NN is 99.87%. The Researchers state that all the algorithms can be applied to the system as all of them have an accuracy greater than 90%. [17] conducted the study to develop a recommender system for graduate students in Bangladesh to help to choose universities according to their academic details. In their research, they created a relational database for students and developed an algorithm to find similarities between trained and test data. The data was trained using k-NN. The researcher states the experimental results show that applicants get admission into universities from the recommended universities in 65-70% of cases of training and testing datasets.

A recommender system was suggested by [18] to assist students with subject selection. At the Public Spanish University in Spain, 323 students' 6,948 academic

ISSN: 2961-5992

observations were used to perform this study. The recommender system was created using data mining, encoding, feature engineering, scaling, and resampling approaches. To train the dataset, they used data mining techniques including LR, SVM, kNN, and RF. Their suggested technique achieves 83.5% accuracy. In their study, they used balancing techniques to balance the dataset to address the problem of data imbalance.

C. The Need for A Subject Stream Recommender System

Difficulty in choosing a suitable subject stream has led to the need for a support system to help students choose subjects. A subject stream system is a tool that aims to help users find items online by providing suggestions that closely match their interests [19]. The main purpose of a recommender system is to make useful determinations on possible links or patterns between the user of the system and some entities based on some form of data processing and thus to output these particular results. Depending on whether the user is seeking a list of suggestions or predictions, these outputs may take the shape of an ordered list or a unique value [20]. Patterns between subject stream selection and previous academic results can be identified through data mining techniques [21].

The link between teachers' and students' perceptions of subject difficulty and subject choice in secondary education was examined [22]. They found that students were prepared to disregard a subject's difficulty when they enjoyed it or needed it to achieve their academic or professional goals and that both professors' recommendations and the school's standards had an effect. This study's most important conclusion was that the viewpoints of enjoyment, utility, and challenge seemed to be the main factors influencing topic choices.

Whiteley *et al.* identified the impact of school policies and practices on students as well as other influences that affect individual subject choices and career decisions [19]. The results of the research show that student perceptions and satisfaction with issues relating to subject selection may alter substantially over time and it was clear that student's Knowledge and Understanding of subject selection would improve over time.

Since the proposed system is to predict the subject stream, predictive data mining techniques should apply to the proposed System. There are Several Classification algorithms researchers use for their studies such as DT, NB, RF, LR, SVM, etc.

According to the literature, we found that other classification algorithms named NB, DT, and RF could be used to make predictions in EDM tasks. The results obtained were interesting in terms of improving performance. So, classification algorithms are the most suitable method to analyze the subject stream prediction.

III. METHODOLOGY

As can be seen in Fig. 1 the set of processes followed to create the Subject Stream Recommender System, the proposed approach consists of three main steps i.e., pre-processing, training, and testing. In the pre-processing stage, basic feature engineering and encoding have been undertaken. Once this pre-processing has been done, training the dataset is initiated. In training the dataset, each scenario is evaluated using each one of the selected data mining algorithms, and the highest accuracy generated algorithm is selected to create a model to recommend the subject stream.

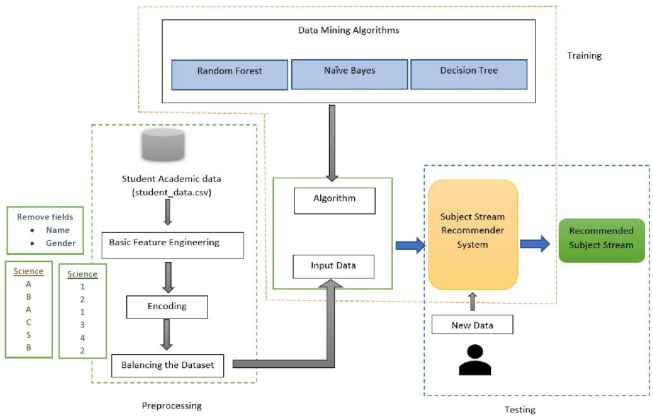

Fig. 1: Set of Processes followed to create the Subject Stream Recommender System

A. Dataset

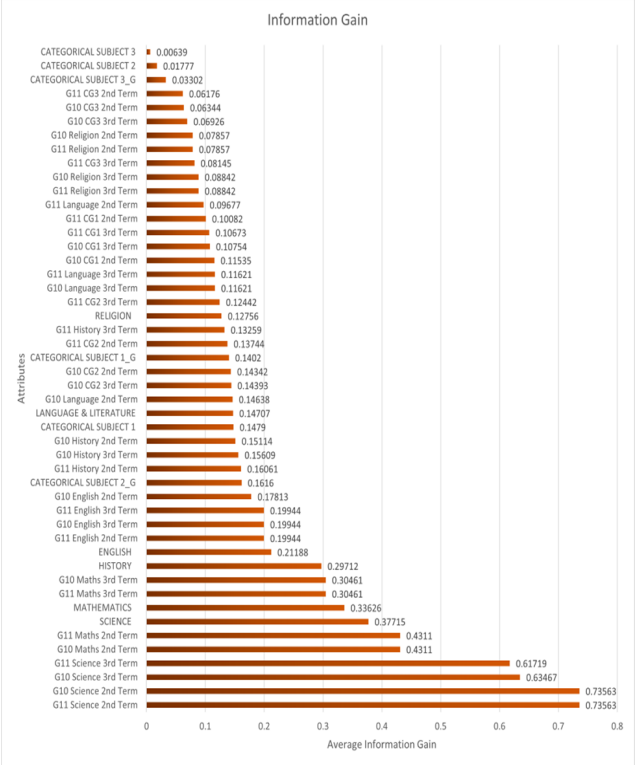

Fig. 2. The average information gained from each feature.

Many internal and external factors might affect the

ISSN: 2961-5992

decision of the subject stream selection of a student. Therefore, it is hard to find the most suitable attributes for the Subject Stream Recommender System. This study does not consider the external factors that might affect the subject stream selection and only focuses on internal factors such as academic performance. The Education Ministry of Sri Lanka has issued some circulars (i.e. Circular No 2008/17 and 25/2013) regarding the educational qualifications that students possess when applying for the AL.

To identify the impact of these attributes on the subject stream selection we used a feature selection algorithm to measure the impact. There are many feature selection algorithms in data mining i.e. Cfs Subset Eval, Chi-Squared Attribute Eval, Classifier Subset Eval, Consistency Subset Eval, Gain Ratio Attribute Eval, Info Gain Attribute Eval, OneR Attribute Eval, Principal Components, ReliefF Attribute Eval, SVM Attribute Eval, Symmetrical Uncert Attribute Eval, and Wrapper Subset Eval [23]. In this study information gain, evaluator was used to the measure impact of each feature on the AL Subject Stream. The values of the information gained for all the attributes are shown in Fig. 2.

Most of the factors considered in this study such as OL grades, term test marks, and selected subject stream for AL were based on the previous academic performance of the students.

As the research data for this study, using the convenience sampling method, information (results of students who sat for the OL in 2017, corresponding students who sat for the AL in 2020, and the marks obtained by the students in Grades 10 and 11 examinations) of 1,000 students who study in two main national schools in the Southern Province of Sri Lanka was used.

The initial dataset was extracted from the result sheets of the OL and AL and the term test mark result sheets. The result sheets contained many types of data such as index numbers. Summary of results etc. Since the goal is to identify the factors affecting the selection of the AL subject stream it was decided not to include all attributes in the initial dataset. All the potential fields were extracted into a single table where a tuple represents the data of a single student.

B. Data Pre-Processing

Data pre-processing involves the activities conducted to prepare the final dataset to be used in data analysis. In the phase of the data pre-processing selection of potential fields and cleaning of the final dataset. It is very crucial to identify and select the most suitable and relevant attributes to develop a model with the highest accuracy. The initial dataset contains 54 attributes, and the student name, gender, index number, and summary of results fields are not needed for the analysis, so these fields were removed from the dataset. The missing values of this dataset can be considered as the students who are absent for a certain subject or subjects in the OL or the school term tests. So, in this dataset, missing values are denoted by "AB". Since this study is to develop a model with the highest accuracy to recommend the subject stream the best method to deal with missing values is to ignore the entire tuple. This dataset contains 28 tuples of missing values, and the tuples were removed.

1) Basic Feature Engineering

In this stage, basic features, engineering, and encoding are initiated. The dataset contains fields such as student name, gender, index number, and summary of results which are unnecessary for the data mining process and, hence all the unnecessary fields are removed.

2) Encoding

Many data mining algorithms did not work properly with categorical data so non-numerical features have to be transformed into numerical features. Since the dataset contains nominal values, the encoding feature makes them properly interpretable for data mining algorithms. In this process, ordinal encoding has been performed to transform nominal data into numerical data. Ordinal encoding converts each data label into integer values and the encoded data represents the sequence of labels.

C. Balancing the Dataset

One of the common issues that affect the dataset is the class imbalance problem which refers to the imbalanced distribution of values of the dependent variable [24]. When analyzing the collected dataset, the imbalance distribution of values in the dependent variable has been identified. In this dataset, the AL subject stream is the dependent variable. Table V shows how values of the subject stream feature are distributed in the dataset.

TABLE V
VALUE DISTRIBUTION OF THE SUBJECT STREAM FEATURE

Subject Stream	Count	Weight (%)
Physical Science	270	35.2
Arts	116	15.1
Biological Science	237	30.9
Commerce	144	18.8
Total	767	100.0

IV. RESULTS AND DISCUSSION

The pre-processed dataset was trained using different classification techniques namely DT, RF, and NB with the WEKA tool. To evaluate a classification technique the confusion matrix can be taken into consideration. These three classifiers were applied to all two experiments and compared the Accuracy. As in Table VI, the dependent variable which is the subject stream feature does have a data imbalance problem. To deal with this situation balancing the dataset should be done.

A. Data balancing with Resample, SpreadSubSample, and SMOTE

ISSN: 2961-5992

Table VI compares the performance of each balancing technique. SpreadSubSample (SSS) and Resample (RS) techniques got equally distributed datasets and every class has equal amounts of instances. In the original dataset, the Arts subject stream has the lowest count (116) after rebalancing with SpreadSubSample, all the classes have the count of the Arts subject stream count. The total number of instances was reduced from 767 to 464 with this balancing technique. In the original imbalanced dataset, the Physical Science class had the highest count of instances, and the Arts class had the lowest count. After balancing the dataset with the Resampling filter all the classes have the same count (191 total number of instances was reduced by 3 counts when balancing and the balanced dataset contains 764 instances.

TABLE VI
SUMMARIZATION OF THE DATA BALANCING TECHNIQUES

Subject Stream	Imbalanced	Balanced		
		SpreadSample	SMOTE	Resample
Physical Science	270	116	270	191
Arts	116	116	232	191
Biological Science	237	116	237	191
Commerce	144	116	144	191
Total	767	464	883	764

In the SMOTE filter, only the minority class will change their instances by adding data tuples. SMOTE algorithm calculates the number of instances that should be added to the minority class and changes only the minority class. The Arts subject stream changed from 116 to 232. The total number of instances was increased from 767 to 883 with this balancing technique by adding data tuples to the dataset.

In the Resample technique, the total number of instances was approximately equal to the original dataset. In SpreadSub-Sample the total instances were lesser than the original dataset while in SMOTE the total instances were higher than the original dataset.

TABLE VII
EXPERIMENT SPECIFICATIONS

Experiment	No of Attributes	Specification
01	13	G.C.E O/L All Subject Grades and Selected Subject Stream
02	49	G.C.E O/L All Subject Grades, G10, G11 Term Test marks for All subjects and Selected Subject Stream

B. Experimental Results

To find the suitable attributes, the dataset was divided into two sections, and the research in two experiments as shown in Table VII.

Experiment 1 has grades of all subjects in OL as attributes. The dataset was tested under RF, DT, and NB classification algorithms. Table VIII compares the accuracy of each classification algorithm with the dataset. With the original imbalanced dataset, the RF has the highest accuracy of 50.71% and the NB has the lowest accuracy of 49.67%. According to the performance of the dataset, which is balanced using the SpreadSubSample technique, RF has the highest accuracy of 50.86% and NB has the lowest accuracy of 49.67%. The balanced dataset accuracy was lower than the original imbalanced dataset. According to the performance of the dataset, which was balanced using the SMOTE technique, the RF has the highest accuracy of 85.73% which is higher than the imbalanced dataset accuracy of RF and the NB has the lowest accuracy of 56.85%. When the dataset is balanced using resample techniques, RF has the highest accuracy of 88.60% and NB has the lowest accuracy of 59.82%. According to Table X, RF has the highest accuracy in every dataset and the dataset which was balanced using Resample technique has the highest Accuracy. Therefore, RF data mining algorithm and the Resample data balancing method got the highest accuracy in Experiment 1.

TABLE VIII
ACCURACY OF EXPERIMENT 1

Classification Algorithm	Imbalanced (%)	Balanced (%)		
		SpreadSubSample	SMOTE	Resample
RF	50.71	50.86	85.73	88.60
DT	48.24	48.24	69.19	73.17
NB	49.67	49.67	56.85	59.82

Experiment 2 contains grades of the all subject in OL, Grades 10 and 11 term test marks for all subjects as attributes. The dataset was tested under RF, DT, and NB Classification Algorithms, and compares the accuracy of each classification algorithm with the dataset as in Table IX. With the original imbalanced dataset, DT has the highest accuracy of 75.88% and the NB has the lowest accuracy of 62.06%. According to the performance of the dataset, which is balanced using the SpreadSubSample technique, DT has the highest accuracy of 75.43% and NB has the lowest accuracy of 60.99%. According to the performance of the dataset, which was balanced using the SMOTE technique, RF has the highest accuracy of 77.12% which is higher than the imbalanced dataset accuracy of RF and the NB has the lowest accuracy of 63.65%. When the dataset is balanced using resample techniques, RF has the highest accuracy of 93.32% and NB has the lowest accuracy of 61.78%. According to Table X RF has the highest accuracy in SMOTE used dataset, and resample used dataset and the DT has the highest in the original dataset and SpreadSubSample used dataset. The dataset which was balanced using the Resample technique has the highest Accuracy 93.32%. Therefore, The RF Data mining algorithm and The Resample data balancing method got the highest accuracy in Experiment 2.

ISSN: 2961-5992

TABLE IX
ACCURACY OF EXPERIMENT 2

Classification Algorithm	Imbalanced (%)	Balanced (%)		
		SpreadSubSample	SMOTE	Resample
RF	74.96	70.25	77.12	93.32
DT	75.88	75.43	75.87	81.02
NB	62.06	60.99	63.65	61.78

C. Summary of Accuracy in Each Experiment

Table X represents the highest accuracy generated algorithm and resampling technique in every experiment. Accuracy in all balanced datasets is higher than in the original imbalanced dataset. In Experiment 1, the best data mining algorithm is RF, and the best balancing technique is Resample which generates an accuracy of 88.61%. In Experiment 2, the best data mining algorithm is a DT, and the best balancing technique is Resample which generates an accuracy of 93.32%. When comparing the performance of each experiment, Experiment 2 has the highest accuracy in the balanced dataset the best data mining algorithm is RF, best balancing technique is Resample.

TABLE X
HIGHEST ACCURACY IN EXPERIMENTS

Experiment	Imbalanced		Balancing Technique	Balanced	
	Accuracy (%)	Best Algorithm		Accuracy (%)	Best Algorithm
1	50.71	RF	Resample	88.61	RF
2	75.88	DT		93.32	RF

V. CONCLUSION

This research involved conducting two experiments using three data mining algorithms to determine the optimal attributes and achieve the highest accuracy for a Subject Stream Recommender System. The dataset, originally imbalanced, was balanced using various techniques, with the Resample method consistently providing the best accuracy across all experiments. In Experiment 1, the Random Forest (RF) algorithm combined with the Resample technique achieved 88.61% accuracy, while in Experiment 2, the same combination resulted in 93.32% accuracy. Across all experiments, RF and Resample were the most effective, making Experiment 2 the preferred model for the proposed recommender system due to its superior accuracy. Regarding the research questions, the first question about identifying the most accurate data mining algorithm was answered, with RF being the top performer. The second question on data balancing techniques showed that Resample was the most accurate. For the third question, concerning relevant predictors, the study considered 49 attributes, with the Information Gain Evaluator highlighting that performance in Science and Mathematics has a significant impact on subject stream selection in AL. Finally, the fourth research question, about improving the system's accuracy, indicated that increasing the dataset size could enhance model performance. The study faced limitations, particularly the exclusion of demographic and psychological factors that might influence AL results. Additionally, the research was limited by the small dataset size (767 instances) and the data being collected from a single year. These constraints led to high variance and error, potentially affecting the system's reliability. Future work could address these limitations by incorporating external factors such as parental influence, economic status, learning skills, and peer influence. Further, the system could be enhanced by selecting the most relevant attributes for each student, applying multiple classifiers, and using additional data balancing techniques. A larger dataset collected over an extended period would likely improve the model's ability to identify patterns and achieve higher accuracy.

REFERENCES

[1] UNESCO Institute for Statistics, "International Standard Classification of Education ISCED 2011," 2012. [Online]. Available: https://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf

[2] Ministry of Education, Sri Lanka, "National Curriculum Framework for Secondary Education in Sri Lanka," Nov. 2020. [Online]. Available: https://educationforum.lk/wp-content/uploads/2021/08/MoE_Framework_SecondaryEducation_2020Nov.pdf

[3] C. P. Samaranayake and H. A. Caldera, "A data mining solution on high failure rate in Physical Science stream at the university entrance examination," in 2012 Tenth International Conference on ICT and Knowledge Engineering, IEEE, 2012, pp. 163–170.

[4] Sri Lanka Information Zone, "Sri Lankan GCE AL Subjects: A Complete Guide," Sri Lanka Information Zone: Your Ultimate Information Zone. Accessed: Jan. 19, 2024. [Online]. Available: https://srilanka-information-zone.blogspot.com/2021/05/sri-lankan-gce-al-subjects.html

[5] Secretary, Ministry of Education Sri Lanka, "Admission of students to A/L (advanced level) classes," 2008. [Online]. Available: https://moe.gov.lk/wp-content/uploads/2020/07/2008-17is.pdf

[6] Secretary, Ministry of Education Sri Lanka, "Admission of students for G.C.E (advanced level) classes," Jun. 11, 2013. [Online]. Available: https://moe.gov.lk/wp-content/uploads/2020/07/2013-25e.pdf

[7] H. Aturupane, V. Dissanayake, R. Jayewardene, M. Shojo, and U. Sonnadara, "Strengthening mathematics education in Sri Lanka," 2011, doi: 10.11588/xarep.00003518.

[8] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., vol. 40, no. 6, pp. 601–618, 2010.

[9] C. Silva and J. Fonseca, "Educational Data Mining: A Literature Review," in Europe and MENA Cooperation Advances in Information and Communication Technologies, vol. 520, Á. Rocha, M. Serrhini, and C. Felgueiras, Eds., Cham: Springer International Publishing, 2017, pp. 87–94.

[10] B. Bakhshinategh, O. R. Zaiane, S. ElAtia, and D. Ipperciel, "Educational data mining applications and tasks: A survey of the last 10 years," Educ. Inf. Technol., vol. 23, no. 1, pp. 537–553, Jan. 2018.

[11] S. Ray and M. Saeed, "Applications of Educational Data Mining and Learning Analytics Tools in Handling Big Data in Higher Education," in Applications of Big Data Analytics, M. M. Alani, H. Tawfik, M. Saeed, and O. Anya, Eds., Cham: Springer International Publishing, 2018, pp. 135–160.

[12] S. Hussain, R. Atallah, A. Kamsin, and J. Hazarika, "Classification, Clustering and Association Rule Mining in Educational Datasets Using Data Mining Tools: A Case Study," in Cybernetics and Algorithms in Intelligent Systems, vol. 765, R. Silhavy, Ed., in

ISSN: 2961-5992

Advances in Intelligent Systems and Computing, vol. 765. , Cham: Springer International Publishing, 2019, pp. 196–211.
[13] Z. Papamitsiou and A. A. Economides, "Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence," J. Educ. Technol. Soc., vol. 17, no. 4, pp. 49–64, 2014.
[14] M. Mimis, M. El Hajji, Y. Es-saady, A. Oueld Guejdi, H. Douzi, and D. Mammass, "A framework for smart academic guidance using educational data mining," Educ. Inf. Technol., vol. 24, no. 2, pp. 1379–1393, Mar. 2019.
[15] G. Mobasher, A. Shawish, and O. Ibrahim, "Educational Data Mining Rule based Recommender Systems.," in CSEDU (1), 2017, pp. 292–299. Accessed: Jan. 19, 2024. [Online]. Available: https://pdfs.semanticscholar.org/ebb9/b826de4aae1eb9087b4409a6a6c07004f1a2.pdf
[16] M. Isma'il, U. Haruna, G. Aliyu, I. Abdulmumin, and S. Adamu, "An autonomous courses recommender system for undergraduate using machine learning techniques," in 2020 international conference in mathematics, computer engineering and computer science (ICMCECS), IEEE, 2020, pp. 1–6. Accessed: Jan. 19, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9077626/
[17] M. Hasan, S. Ahmed, D. M. Abdullah, and M. S. Rahman, "Graduate school recommender system: Assisting admission seekers to apply for graduate studies in appropriate graduate schools," IEEE, 2016, pp. 502–507.
[18] A. J. Fernández-García, R. Rodríguez-Echeverría, J. C. Preciado, J. M. C. Manzano, and F. Sánchez-Figueroa, "Creating a recommender system to support higher education students in the subject enrollment decision," IEEE Access, vol. 8, pp. 189069–189088, 2020.
[19] S. Whiteley, J. Porter, and T. E. P. Authority, "Student perceptions of subject selection: Longitudinal perspectives from Queensland schools," in Australian Association for Research in Education 1998 Conference Proceedings, 1998.
[20] S. B. Aher and L. Lobo, "Prediction of course selection by student using combination of data mining algorithms in E-learning," Int. J. Comput. Appl., vol. 40, no. 15, pp. 1–7, 2012.
[21] P. Dixit, H. Nagar, and S. Dixit, "Decision Support System Model for Student Performance Detection using Machine Learning," vol, vol. 10, pp. 25–31, May 2021.
[22] F. Ünal, "Data mining for student performance prediction in education," Data Min.-Methods Appl. Syst., vol. 28, pp. 423–432, 2020.
[23] M. Arif, A. Jahan, M. I. Mau, and R. Tummarzia, "An Improved Prediction System of Students' Performance Using Classification model and Feature Selection Algorithm," Int. J. Adv. Soft Comput. Its Appl., vol. 13, no. 1, 2021.
[24] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," Inf. Sci., vol. 513, pp. 429–441, 2020.

ISSN: 2961-5992