Predicting Undergraduates Dropouts Using Classification Techniques

***Abstract*—Undergraduate dropout is one of the biggest concerns in higher education institutes. Student retention has gained more attention from university administrators, especially those in private-sector higher education institutes as the competition is quite high in the private sector. This research's main objective is to predict undergraduate dropouts in the Information Technology Degree Program of a non-state higher education institution in Sri Lanka. Logistic Regression, Random Forest, Naïve Bayes, Artificial Neural Network (ANN), Decision Tree, and Support Vector Machine (SVM) classification techniques were used for the prediction. According to the results, SVM has the best F1 score which is 90%, ANN, Decision Tree, and Logistic regression got 88%, and Random Forest and Naïve Bayes have an 87% of F1 score. It has also been identified that dropouts are high in those who have done Advanced Level in Art Stream and under Other Category. Therefore before students get register from those categories if faculty can give them an aptitude test and select the relevant candidates, will be helpful to reduce the dropouts. Data mining techniques can improve the quality of education in non-state higher education institutes as this helps to identify the hidden patterns of educationally linked data.**

Keywords—classification, dropout, higher education institute, undergraduate

I. Introduction

The undergraduate dropouts affect not only the individual student but also the institute and, ultimately the society [1]. A higher level of education in students is one of the required conditions to improve human capital in society [2]. Therefore, preventing students from dropping out of their higher studies is important. Also, the smaller number of higher education means a lack of skilled human capital, leading to unemployment, and poor standards of living. On the other hand, non-state degree awarding institutes also can be considered as business organizations where that need to generate profit to facilitate their employees and provide necessary resources to the students. Hence, one student dropout means overall a huge financial loss to the institute. Indirectly it will affect the long-term retention of the institute. Not only the institute but also students and their families also suffered ultimately in terms of costs and students' time. The early identification of these students is not an easy task, and it is crucial for the success of the retention strategy. As today Sri Lanka is facing more economic instability than ever in recent history, the education sector is also affected. Thus, identification of these students in their early stages is an important factor as then the institute can take precautionary steps to prevent the students from dropout.

The primary aim of this study is to predict the number of student dropouts in a non-state higher education institute. To achieve the main aim, the following objectives have been formulated:

1) To identify the factors that have affected dropouts.

2) To determine the most suitable algorithm for dropout prediction in non-state higher education institutes.

This research used educational data mining which is a growing research area in educational institutions over the world. According to the researcher Shakeel et al., [3] it has identified that educational data mining techniques can be highly used for institutional benefits. There is very limited research done regarding dropouts' predictions, especially in non-state higher education institutes in Sri Lanka. According to the Ministry of Education in Sri Lanka, there are 25 degree-awarding institutes in Sri Lanka, other than state universities [4]. The primary aim of this study is to reduce the number of student dropouts in non-state higher education institutes in Sri Lanka. The prediction has done after students' first-semester examinations by considering major subjects in their first semester with other attributes like age, gender, and loan facilities.

II. Literature Review

Over the last 50 years, several theoretical and empirical studies have been done about student retention [5]. Among all the theories Tinto's student Integration Model is the most important which emphasizes the social and academic integration of the students [2] [5]. It has been identified through literature that most of the research work has used Tinto's Model.

Educational Data Mining (EDM) field is based on organizational theory [6]. EDM is an emerging discipline that used data mining tools in educational fields and EDM has different other disciplines such as learning theory, data mining, and machine learning. EDM is concerned to study student behavior by analyzing study-related data [1]. Data mining is also based on several theories including Data reduction, and probability theory based on statistics theory [7].

In the empirical study done by Lorenz et al., the single and most important factor that affects the students' dropouts in German universities is performance problems calculated by the combination of average passed and failed examinations. According to the authors, 70% of dropouts had performances issue [8]. The researcher Francesca [9] has identified 14.8% of students have dropout those in their first year and 21.6% have dropped out in the third year [9]. According to the research that Liga Paura has done it has identified nearly 64% of students dropped out after their first year of the degree program [10].

Students with lower grades in Computer related courses also dropped out compared to the Non-dropout students [11]. Student dropouts in Computer Science degree programs have considered four different binary classification models, Support Vector Machine (SVM), Logistic Regression, Naïve Bayes, and Artificial Neural Network (ANN) for prediction [12]. According to the results, it has been identified that the Naïve Bayes classifier has achieved the highest accuracy level of 96%.

The researchers [3] implemented a predictive model for Business Studies students who enrolled in the University of Gujrat (UOG), Pakistan which has a considerable number of dropouts after two semesters. The authors have used different classification techniques and the main objectives are

- extract the most appropriate attributes from the data source
- identify the other attributes that determine learning behavior
- identify the most suitable classification algorithm and
- provide information to the administration about risk students.

It has been identified that other than academic performance students' level of preference to do the course also affects the completion of the degree program [13]. The study was done for undergraduates at Guru Gobind Singh Indraprastha University [14] has identified that attributes like secondary level education, gender, and medium have less relevant to the study. According to the author's results, it has been identified that True Positive rate, Recall, and Precision are higher in the Random Tree algorithm than in J48.

The authors Kasthuriarachchi et al., [15] analyzed student performance in tertiary education and identify the factors that affect the students' performances. These authors also have used several data mining techniques like Decision Tree (DT), Naïve Bayes (NB), and SVM to identify the factors. The selected sample is based on students who have enrolled in an Information Technology degree program for three years. The performance study was done after six semesters by getting each semester's Grade Point Average (GPA) values (numeric value 0-4). Other than the GPA authors also considered Sex, Age, and Number of Failure subjects.

This author, Kabakchieva [16] has predicted student performances using four classification algorithms. The highest accuracy gets from the Neural Network algorithm is 73.59% and other classifiers are Rule Leaner (One R), Decision Tree (J48), and K-Nearest Neighbour. Much research found through the literature has been done to identify the influence factors for students' dropouts, but there is limited research done related to undergraduates' dropouts from Information Technology degree programs in non-state higher education institutes in Sri Lanka. Thus, there is a need to do research to identify the factors that influenced dropouts in Information Technology (IT) degree programs.

III Methodology

Dataset Description and Preprocessing

Once students get registered in the institute the following data can be collected Age, Gender, Advanced Level (A/L) Stream, and whether the student has obtained Loan or not. The above-mentioned data were gathered from the faculty and altogether there were 159 instances. The data set was cleaned to ensure that there is no missing records. According to the selected dataset the gender distribution of the Information Technology Degree program is mentioned below. As the above Fig. 1, shows it has been identified that the Male student count (55%) is higher than the Female students(45%). There were several students who were absent

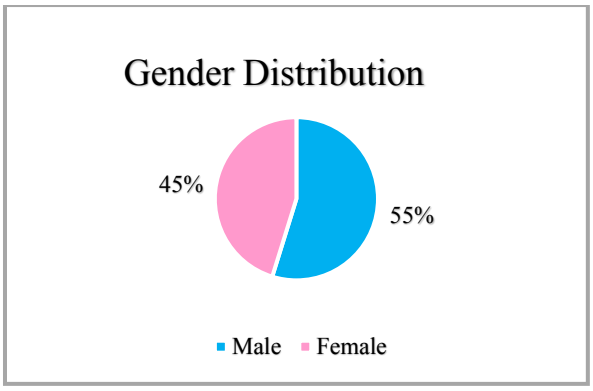

Fig. 1. Gender Distribution

from the final examination may be due to medical reasons or maybe they had already dropped out before the first end-semester examination; those instances have been renamed to zero.

It has identified that students have major 6 streams for their A/Ls namely,

1. Physical Science
2. Biological Science
3. Commerce
4. Arts
5. Engineering Technology
6. Bio System Technology

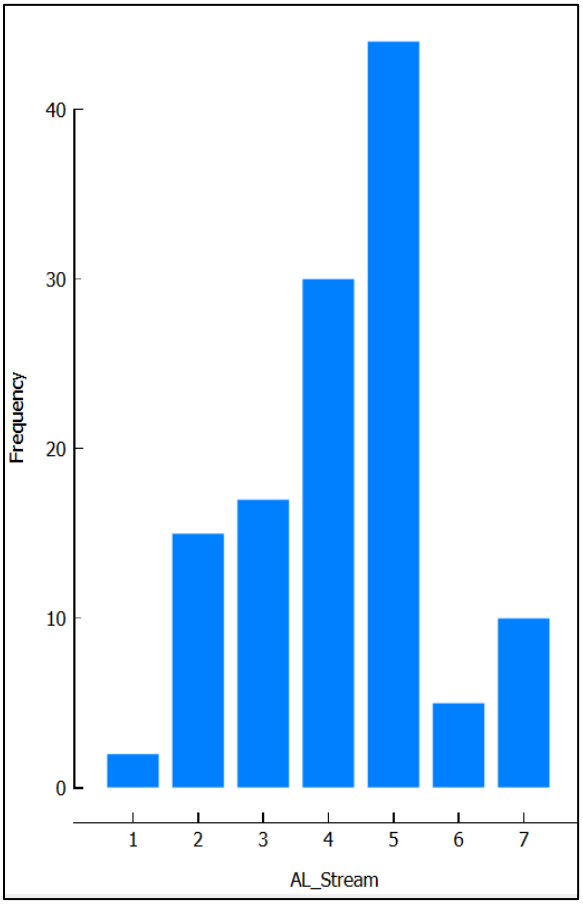

Fig. 2. Undergrduates by A/L Stream

Among those undergraduates, some of them have selected subjects from more than one stream. Those students' A/L stream has been considered as an "Other Category" which has given the categorical value of 7. Fig. 2. Shows the distribution of undergraduates according to the A/L Stream.

When the dataset was analyzed by the A/L stream it has identified that the majority of students who have registered for the Information Technology Degree Proram, have done their A/Ls in Engineering Technology (36%) and Arts streams (24%). After analysing these data it has been identified that among all the categories, the dropout rate is high, in those who have done A/Ls in Art Stream (33%) and Other Stream (27%) shown in the following Fig.3. Also, no dropouts were found from Physical Science and Bio System Technology streams.

After students did their end-semester examination next data were gathered, like Student performance (calculated by Garde Point Average value), and the number of failed exams. All the data were gathered from the examination department and the selected undergraduates were registered for the 2021 academic year. The Grade Point Average (GPA) was calculated for six subjects namely

- IT and Computing Fundamentals
- Structured Programming
- Discrete Mathematics
- Psychology
- Leadership Skills
- Business Management

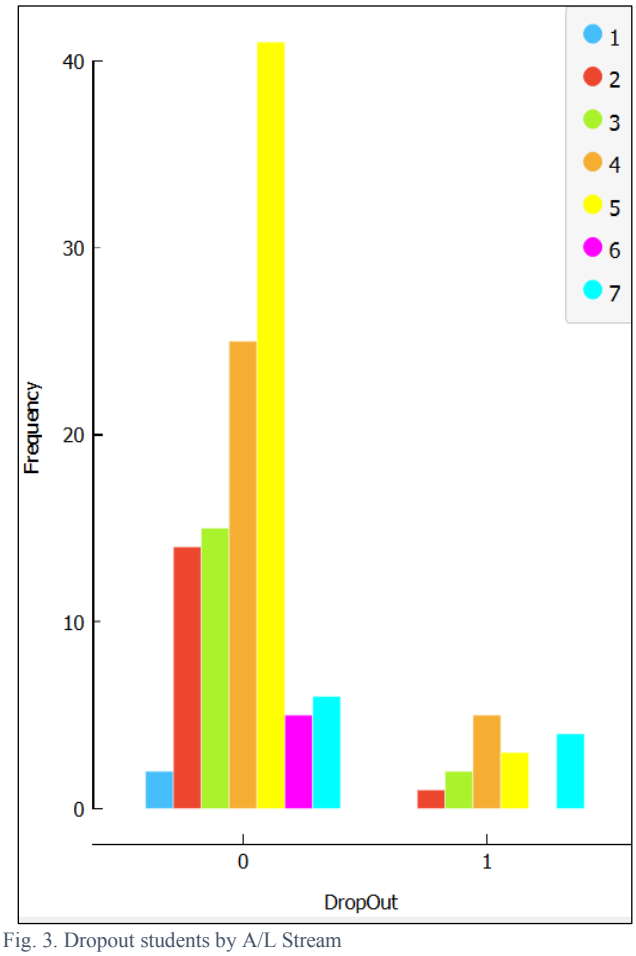

Fig. 3. Dropout students by A/L Stream

TABLE 1. DATASET DESCRIPTION

Variable		Value
Age		20-24 years, > 24
Gender		1 = Male 2 = Female
Loan		1= Yes 0 = No
AL_Stream		1. Physical Science 2. Biological Science 3. Commerce Stream 4. Arts Stream 5. Engineering Technology 6. Bio System Technology 7. Other Category
Student Performance (GPA)	IT & Computing Fundamentals	**1** - 1 to 1.99 **2** – 2.00 to 2.99 **3** - 3.00 to 4.00
	Structured Programming	
	Discrete Mathematics	
	Psychology	
	Leadership Skills	
	Business Management	
English Language Marks		**1** - 0 to 1.99 **2** – 2.00 to 2.99 **3** - 3.00 to 4.00
Count_npass		>=0

The variable selection has been done according to the previous research, Yaacob et al., and Boris et al., have used admission information (demographic data) and transcript records [11] [2] related to the degree program. Other than the demographic data "count of failed exams" variable has been selected by Lorenz et al [8]. Obtaining an Education Loan has been considered by the authors Mishra et al [14].

In this research, the Dropout variable is considered as the target variable and coded as Dropout = 1, Non-Dropout = 0. According to the analysis, it has been identified that the mean GPA value of dropout students is 1.50 as shown in Fig. 4.

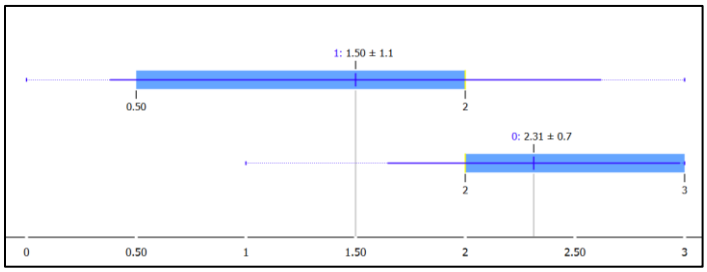

Fig. 4. Box plot of GPA by Dropout

As there are many machine learning algorithms it is important to select the best algorithm for dropout prediction. Even in machine learning, there are different types of algorithms like Regression, Classification, and Clustering, out of these, this research is going to use Classification techniques. The selected classification algorithms are Logistic Regression, Random Forest, Naïve Bayes, ANN, Decision Tree, and SVM.

Logistic Regression: This is suitable for binary or dichotomous classifications, which means true or false (0 and 1) type situations *[17]*.

Random Forest: The Random Forest algorithm is one of the most important algorithms in machine learning. It has been implemented based on ensemble learning. This algorithm will create many decision trees for the given subsets of datasets and then get the average to build the predictive models. Without taking one decision tree, this technique will consider each decision tree in the subset, and based on the most votes the output results will consider. This algorithm is comparatively slower than the Decision Tree *[18]*.

Naïve Bayes: This is known as a faster algorithm for classification problems. Naïve Bayes implements based on the Bayes theorem and can be used for Real-time prediction, recommendation systems, multi-class prediction, and text classification *[19]*.

Artificial Nural Network (ANN): This algorithm consists of three layers namely the input layer, hidden layer, and output layer. During the classification time input layer propagates the activation value to the hidden layer. The hidden layer itself will calculate the activation values and then send those to output units *[20]*

Decision Tree: The decision Tree is a supervised learning algorithm that starts with a root node and outgoing branchers which leads to internal nodes. After the evaluation outcome will be denoted by leaf nodes. The leaf node will provide the possible outcomes within the dataset. This is one of the popular algorithms *[21]*.

Support Vector Machine (SVM): This algorithm is trained to learn by example and assign a label to objectives. This model can be further divided into linear and nonlinear models. When it is linear the data set will divide into two separate subdomains and be distinguished by label *[22]*.

According to the research design after the data set was preprocessed data profiling has done to get descriptive statistical analysis. Then different machine-learning algorithms were used to find out the best model. Then the results were analyzed by the F1 Score value as mentioned the Fig.5.

To evaluate each model the Confusion Matrix (Table II) was used.

Research Design

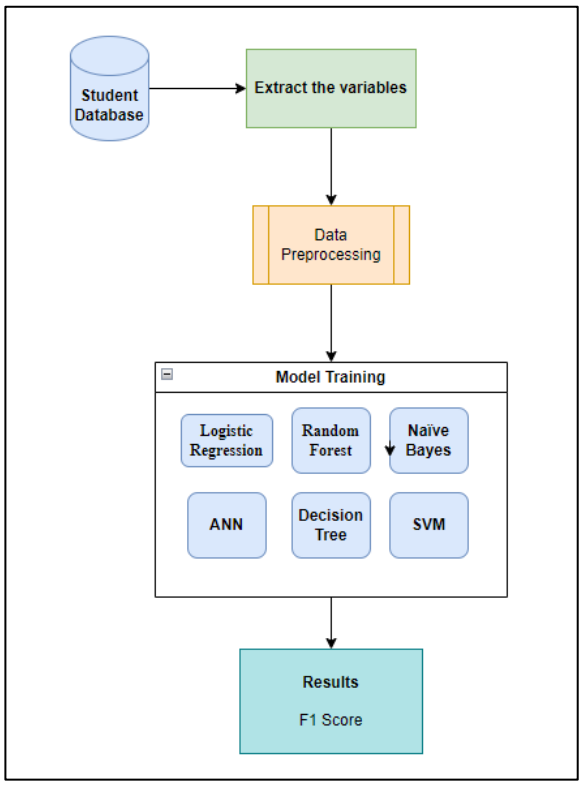

Fig. 5. Research Design

- The Accuracy (Acc) of the Model is calculated as follows,

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}$$

- True Positive (TP) Rate is known as recall or out of actual positives how many predicted as positive by the model, calculated as follows,

$$Recall = \frac{TP}{TP+FN}$$

- True Negative (TN) Rate measures the actual negatives that the model has predicted as negative

TABLE II. CONFUSION MATRIX

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

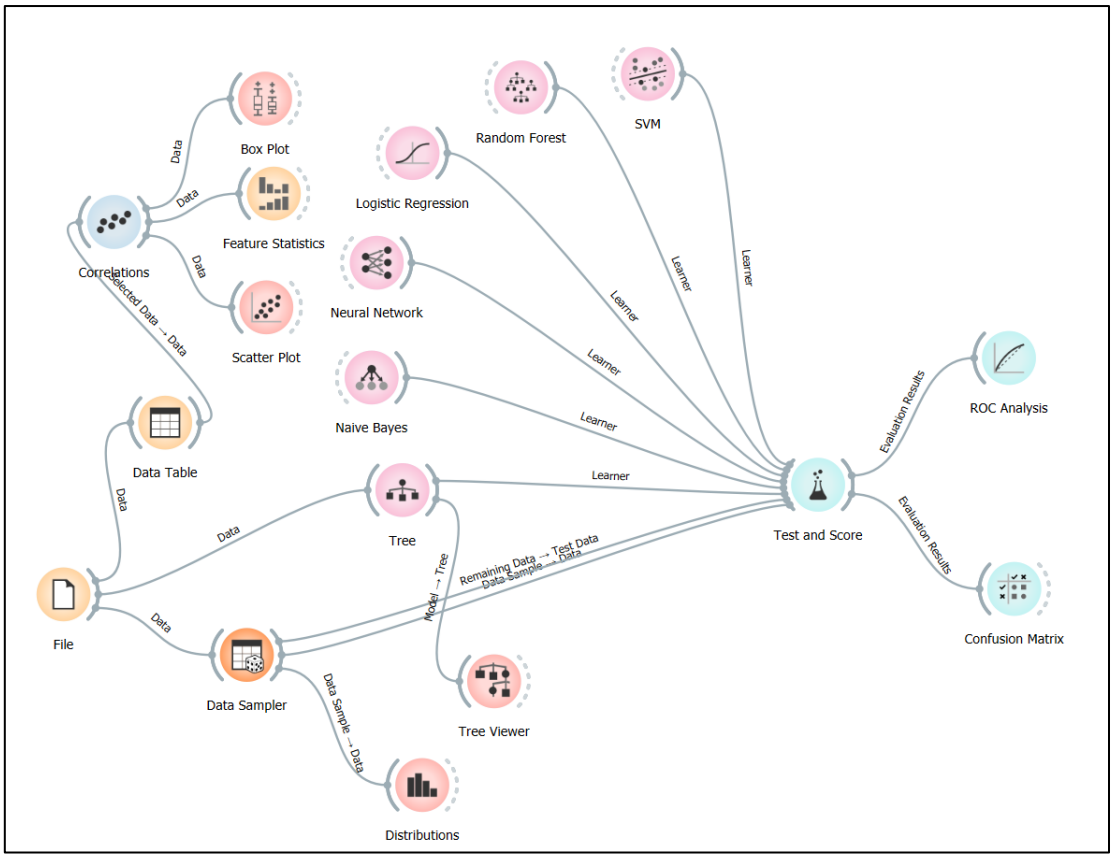

Fig. 6. Orange Tool Flow Diagram

The Orange data mining software [23] tool has been used to do the experiment. All classification algorithms used a cross-validation procedure with 10-fold repeated 10 times with different sets of the dataset. The flow diagram of the Orange Tool is mentioned above in Fig 6.

IV Results of the Experiment

This section is elaborate on the results of the above-mentioned algorithms. The Decision Tree results shows 27 nodes, and 14 leaves. It has identided from the decision tree if the undergraduates has zero or less GPA definitely it is a dropout. Then if a student has not obtained a loan, and the A/L stream is either 5, 6, or 7 the dropout rate is 8.7%. if the A/L stream is either 1, 2, 3, or 4 the dropout rate is 26.7%. if the undergraduate's GPA is 2 and has a 1 or 2 for the English Language and if the Gender is 1, the Dropout rate is 60%. According to the results of the F1 score it has been identified that the SVM model has the best score of 90% Artificial Neural Networks, Decision Tree, and Logistic Regression models have a score of 88%, Naïve Bayes and Random Forest model have 87% scores, According to the given TABLE III, the accuracy range models are between 87% to 90%.

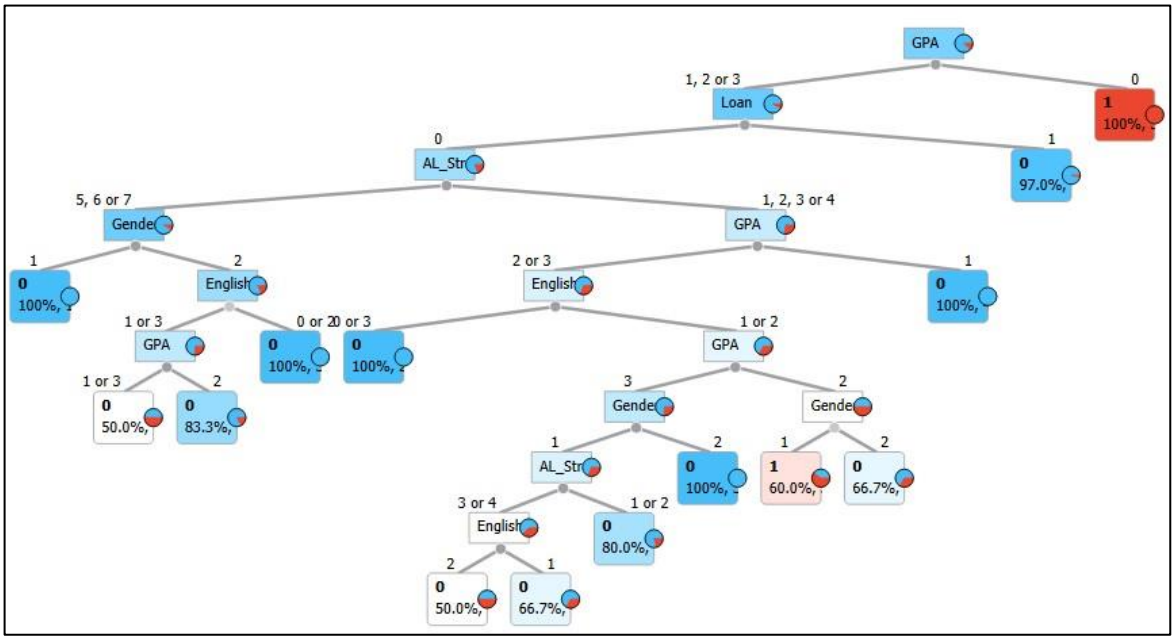

Fig. 7. Results of Decision Tree

TABLE III. EXPERIMENTAL VALUES

Model	AUC	CA	F1	Precision	Recall
SVM	64%	92%	90%	93%	92%
Neural Network	78%	89%	88%	88%	89%
Tree	46%	90%	88%	88%	90%
Logistic Regression	82%	91%	88%	92%	91%
Naive Bayes	78%	89%	87%	86%	89%
Random Forest	78%	89%	87%	86%	89%

V CONCLUSION AND DISCUSSION

According to the previous discussion, it has been identified that selected models are suitable for the prediction of undergraduate dropouts. It has been identified that other than the Information Technology degree program we can consider the entire faculty of the Information Technology degree programs to do the prediction as the results will be helpful for decision-makers to get the necessary precautionary actions. The dataset that was used for this experiment is limited hence in the future we can increase the amount of data for better results.

As the dropout rate is high in Art Stream and Other Category streams, it is better to consider the pre-requisite subjects or aptitude tests for those students, before they registered for the degree program.

According to the time being it has considered only first-semester end examination results, but undergraduate dropouts can happen at any stage of their education ladder. As this degree program is an Honours degree, there are eight semesters hence, experiments can be continued until students get graduated.

REFERENCES

[1] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble," *Computers and Education: Artificial Intelligence,* vol. 3, pp. 12, 26 March 2022.

[2] B. Perez, C. Castellanos and D. Correal, "Predicting Student Drop-Out Rates Using," *Springer Nature Switzerland,* pp. 111-125, 2018.

[3] K. Shakeel, "Educational Data Mining to Reduce Student Dropout Rate by Using Classification," in *OMICS International Conference on Big Data Analysis & Data Mining*, USA, 2015.

[4] M. o. E. (. Education, "Recognized Institutes & Degree," 2022. [Online]. Available: https://www.mohe.gov.lk/index.php?option=com_content&view=article&id=352&Itemid=336&lang=en#recognized-degrees. [Accessed December 2022].

[5] A. Siri, "Predicting Students' Dropout at University Using Artificial Neural Networks," *Italian Journal of Sociology of Education, 2015,* pp. 225-247.

[6] R. A. Huebner, "A survey of educational data-mining research," *Research in Higher Education Journal,* vol. 19, p. 13, April 2013.

[7] S. Sivanandam and S. Sumathi, "Theoretical Frameworks for Data Mining," *g, Studies in Computational Intelligence, 2006,* pp. 265-270.

[8] L. Kemper, G. Vorhoff and B. U. Wigger, "Predicting student dropout: A machine learning approach," *European Journal of Higher Education,* vol. 10, 2020, pp. 28-47.

[9] F. D. Bonifro, M. Gabbrielli, G. Lisanti and S. P. Zingaro, "Student Dropout Prediction," *Springer Nature Switzerland, 2020* pp. 129-140.

[10] I. A. Liga Paura, "Cause Analysis of Students' dropout rate in higher education study programme," 2013.

[11] W. F. W. Yaacob, N. M. Sobri, S. A. M. Nasir and W. F. W. Yaacob, "Predicting Student Drop-Out in Higher Institution Using Data," *Journal of Physics: Conference Series,* 2020.

[12] N. Shynarbek, A. Orynbassar, Y. Sapazhanov and S. Kadyrov, "Prediction of Student's Dropout from a University Program," *IEEE,* 2021.

[13] M. Segura, J. Mello and A. Hernández, "Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role?," *Mathematics,* p. 20, 2022.

[14] T. Mishra, D. D. Kumar and D. S. Gupta, "Mining Students' Data for Performance Prediction," 2014.

[15] K. Kasthuriarachchi and S. Liyanage, "Knowledge Discovery with Data Mining for Predicting Students' Success Factors in Predicting Students' Success Factors in," Moratuwa, 2017.

[16] D. Kabakchieva, "Student Performance Prediction by Using Data Mining Classification Algorithms," *International Journal of Computer Science and Management Research ,* vol. 1, no. 4, 2012.

[17] D. Hosmer, Jr.Lemeshow, S and R. Sturdivant, "Introduction to the Logistic Regression Model," in *Applied Logistic Regression*, John Wiley & Sons, Ltd, 2013, pp. 1-33.

[18] M. Chasudary, "Random Forest Algorithm - How It Works & Why It's So Effective," 2022. [Online]. Available: https://www.turing.com/kb/random-forest-algorithm. [Accessed 2022].

[19] P. Pedamkar, "Naive Bayes Algorithm," 2022. [Online]. Available: https://www.educba.com/naive-bayes-algorithm/. [Accessed 2022].

[20] S. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," in *Emerging Artificial Intelligence Application in Computer Engineering*, IOS Press, 2007.

[21] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. M. Fardoun and S. Ventura, "Early dropout prediction using data mining: a case study with high school students," *John Wiley & Sons, Ltd,* vol. 33, pp. 107-124, 2016.

[22] S. Suthaharan, "Support Vector Machine," in *Machine Learning Models and Algorithms for Big Data Classification* , Bostan MA, Springer, 2016, pp. 207-235.

[23] J. Demsar, C. T, A. Erjavec, H. T, M. Milutinovic, M. Mozina, M. Polajnar, M. Toplak, A. Staric, M. Stajdohar, L. Umek, L. Zagar, J. Zbontar, M. Zitnik and B. Zupan, "Orange: Data Mining Toolbox in Python," *Journal of Machine Learning Research,* vol. 14, p. 2349−2353, 2013.