Journal of Applied Artificial Intelligence
https://jaai.sabapub.com
ISSN: 2709-5908

2022 Volume 3, Issue 2: 1–12

DOI: https://doi.org/10.48185/jaai.v3i2.601

Development of a Predictive Model of Student Attrition Rate

Godwin Sani [1,*], Francisca O. Oladipo[2], Emeka Ogbuju[2], Friday J. Agbo[3,4]

[1,2]Department of Computer Science, Faculty of Science, Federal University Lokoja, Kogi State, 260101, Nigeria
[3]School of Computing, University of Eastern Finland, FI-80101 Joensuu, Finland.
[4]School of Computing and Data Science, Willamette University, Salem, OR 97301, USA.

Received: 10.10.2022 • Accepted: 01.11.2022 • Published: 30.12.2022 • Final Version: 31.12.2022

Abstract: Enrollment in courses is a key performance indicator in educational systems for maintaining academic and financial viability. Today, a lot of factors, comprising demographic and individual features like age, gender, academic background, financial capabilities, and academic degree of choice, contribute to the attrition rates of students at various higher education institutions. In this study, we developed prediction models for students' attrition rate in pursuing a computer science degree as well as those who have a high chance of dropping out before graduation using machine learning methodologies. This approach can assist higher education institutions in creating effective interventions to lower attrition rates and raise the likelihood that students will succeed academically. Student data from 2015 to 2022 were collected from the Federal University Lokoja (FUL), Nigeria. The data was preprocessed using existing WEKA machine learning libraries where our data was converted into attribute-related file form (ARFF). Further, the resampling techniques were used to partition the data into the training set and testing set, and correlation-based feature selection was extracted and used to develop the students' attrition model to identify the students' risk of attrition. Random Forest and decision tree machine learning algorithms were used to predict students' attrition. The results showed that Random Forest has 79.45% accuracy while the accuracy of Random tree stood at 78.09%. This is an improvement over previous results, where an accuracy of 66.14%. and 57.48% were recorded for random forest and Random tree respectively. This improvement was because of the techniques demonstrated in this study. It is recommended that applying techniques to the classification model will improve the performance of the model.

Keywords: Machine learning, Predictive model, Random Forest, Random Tree algorithm, Student Attrition, Feature selection method.

1. Introduction

Attrition among final-year students has significant effects on both the people and the affected institutions in today's educational system. Attrition results in costs for all parties, whether they are in terms of resources, time, or money [1], [2]. As a result, institutions of higher education face a significant challenge in preventing educational attrition [3].

The prediction of student attrition has generally been subjected to a machine learning model to make a better decision. The machine learning model is a computer program or similar software

*

designed to recognize patterns or behaviours based on previous experience or data. It is used to make decisions from a previously unseen dataset and to analyze data without an explicit program, This approach can help an institution to lower the rate of student attrition through the identification and maintenance of student relationships using the predictive data mining methods suggested by [4].

To improve retention strategies like learning aid or mentorship programs, the identification of risk cases is the first step. Evaluating attrition risks may be useful for effectively allocating pedagogical, psychological, or administrative resources. To determine whether it is feasible to forecast attrition using study progression data, this study will act as a pilot [5].

However, understanding and tackling the attrition problem at Nigeria's higher institutions is especially significant for two reasons. First off, attrition is a common domestic phenomenon in Nigerian higher education, with over 30% of students failing to complete their degrees. Second, there is a lack of qualified professionals in the Nigerian labour market because of attrition and economic expansion, and demographic change [6]. With the help of machine learning algorithms, this study aims to forecast student attrition issues, as well as to gain insight into and further examine the issue. It also aims to develop an application based on one of these algorithms that can accurately predict student attrition. The primary dataset of student assessment from 2012 to date was collected from the Federal University Lokoja (FUL), Kogi State Nigeria was used to perform this analysis.

A total sample of 4407 student assessments was manually extracted from the FUL database of the previous year of student assessment, which was used for predicting student attrition. The dataset was used, and each line begins with the class label of each student evaluation before the data is annotated. And feature extraction from the data and pre-processing, machine learning techniques e.g., random forest and random tree models trained in other to make predictions. The need to prevent student dropouts from easily accessible sources is underpinned by the prediction of attrition among students using machine learning techniques. Thus, by choosing to rely on the conventional study as progression information or data that may appear on the average student's transcript of records, as opposed to survey student data, this can appear as a positive finding and be easily replicated at other schools. The primary goal of this student attrition is undoubtedly to make predictions using the dataset's presented characteristics. To forecast the attrition rate, the dataset will be subjected to the categorization techniques known as random forest and random tree. Overfitting for feature selection will be dealt with using the cross-validation method. One of this paper's main objectives is feature selection. Correlation-based feature selection is used to pick features and improve the outcomes.

Below is the working flow of student attrition prediction shown in Figure 1.

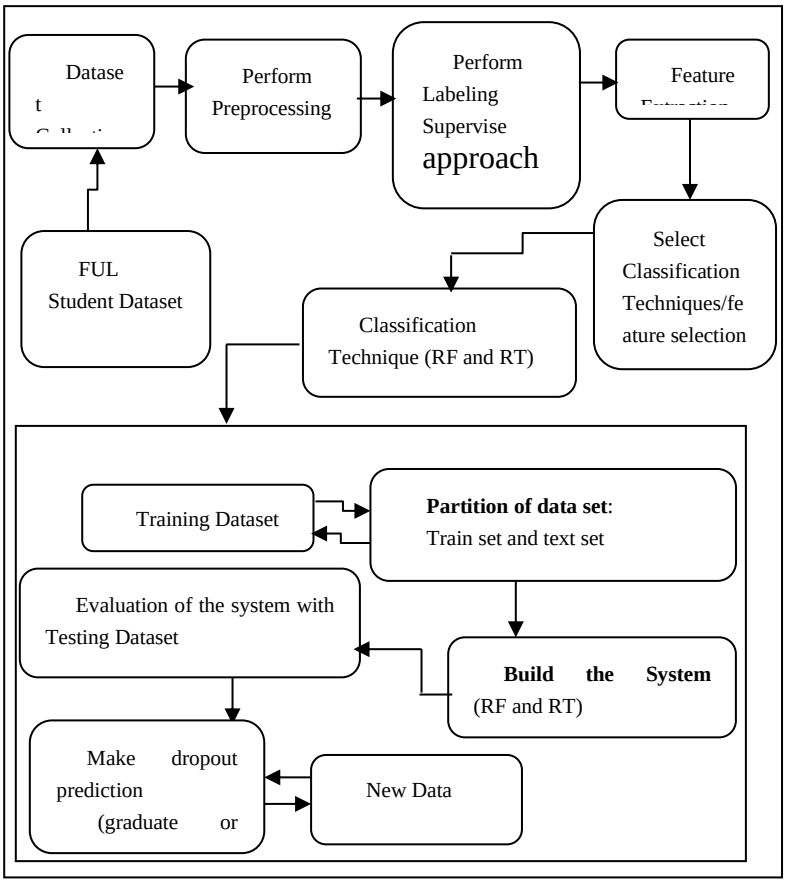

Figure 1. working/data flow of Student Attrition Prediction

Finally, the performance of the classifier was then tallied and evaluated. To build the model, the machine learning technique was applied using the NetBeans IDE (weka library plugin) and Weka for feature extraction and early data analysis.

The paper is structured into sections: Section II discusses related data mining works, but the study is primarily concerned with random trees and forests. In part III, the system problem was studied, together with the methodology, dataset, feature set, and experimental setup information. The system's implementation, evaluation, and discussion of the findings were presented in Section VI. Section V presents the report's conclusion, recommendations, and areas for upcoming work.

2. Related Works

This section reviews the concept of relevant work on student attrition prediction using a machine learning model. Here we discuss the various stage of attrition prediction in higher institutions and the well-known survival model has been the foundation for models and analyses of student attrition in higher education for forty years [7]. While most of the data mining perspective on predicting individual student attrition adds to our knowledge not yet been taken on in Nigeria use cases. Meanwhile, several international studies have been conducted. In [8], explores student retention probability based on freshmen' background data, including demographic, academic, and course participation data. Methodologically, they find that, transferred hours, residency, and background as critical retention determinants utilizing multivariate adaptive regression splines, neural networks, and classification trees.

A description of some of the most significant data mining research is also provided in this part for forecasting study success merely based on enrollment data [9]. Socio-demographic and course data make up his dataset, which he uses to apply classification/regression trees and feature selection. The most accurate indicators are found to be ethnicity, course plans, and course blocks. However, the greatest prediction performance only achieves a 60% accuracy. As a result, Kovacic concludes that data from the enrolling process alone is insufficient to accurately predict research success.

[10] conducts yet another study to comprehend university student retention from a data mining standpoint. In addition to analyzing how to identify students who are in danger of dropping out, the study offers helpful guidance on how to utilize the information acquired to increase the retention of those students going forward. They base their study on a dataset that incorporates data from some university-affiliated sources, including the student record system, online resources, and library use. They provide a diverse view of the student's academic activity as a result. They analyze the algorithms and discover, intriguingly, that a naïve Bayes method performs better at forecasting student achievement than support vector machines and decision trees.

Additionally, [11] found that balanced datasets provided better predictions than unbalanced datasets when comparing individual classification algorithms with those of ensemble learning algorithms for identifying students who are more likely to drop out. [12] has shown the superiority of the unique clustering with affinity measures (UCAM) clustering algorithm over conventional clustering algorithms by using affinity measures rather than individual classification techniques.

2.1. Conceptual Frameworks

[13] shown that a multi-layered neural network-based technique performed more accurately when classifying students into high, medium, and low categories than conventional machine learning algorithms without variable reduction.

Using three different student representations, [14] in his finding using classification algorithms to predict dropouts at higher levels of educational institutions, discovered that gradient boost ensemble and random forest algorithms outperformed Naive Bayes, which underperformed due to its strong interdependence assumption. [15] used a c4.5 algorithm (j48) to predict student performance on the English exit exam and discovered that English placement test results were a significant indicator of performance on the exam. The classification accuracy of the different algorithms of the decision tree, such as reptree, c4.5 decision tree and random tree, in forecasting, if a student will graduate at the right time was compared in research published in [16]. The ICRM2 algorithm was suggested by [17] as a method for predicting student attrition and it was shown that it performed more than other classifiers in terms of "True Negative rate and Geometric Mean of True Negative and True Positive rate", in addition to overall classification accuracy. These studies are notable examples of those that have evaluated classifiers besides general classification precision. At Indonesia Open University, researchers compared Naive Bayes, Bagging, and C4.5 for identifying inactive students using cross-validation, confusion matrices, ROC curves, recall, precision, and F measures in [18].

The Random Tree classifier demonstrated a 57.48 per cent accuracy, indicating that it fares poorly when applying machine learning methods to predict student failure in university examinations. In conclusion, classifiers of tree shape are more effective in predicting student attrition from pursuing a computer science degree. This is because the accuracy of student attrition prediction was greatly improved by applying an "algorithm with feature selection" throughout the entire procedure.

This study addresses a research gap while adhering to the principles of earlier notable research efforts. As suggested by [19], new metrics can be used to evaluate classification algorithms, including accuracy, precision, recall, and f score.

The results of the study demonstrate that numerous enrolments and attrition rate predictions have used "machine learning techniques (MLT)" to pinpoint the pattern of undergraduate enrolment and attrition. Along with various feature sets, classification methods and algorithms like SVM, NB, J48, and CCM are used. The results of a study on student performance as a predictor of freshman-year dropout rates were published in [20]. In this study, the ideal model that predicts the attrition rate with an accuracy of 80% is created using naive Bayes, decision trees, and rule-based induction machine learning techniques.

[21] analyzed the dropout rate for "189 students enrolled in the online information technology certificate program" as part of their study. The study employs the following machine learning algorithms— Decision Tree (DT), K Nearest Neighbor (KNN), Neural Network (NN) and Naive Bays (NB)—and finds that student demographic data is extremely important in determining dropout rates, with NN and DT having accuracy rates of 87 per cent and 80 per cent, respectively. In [22], a study was conducted on "220 undergraduate students enrolled in information technology courses" to determine the elements that contribute to student dropout rates using machine learning algorithms. According to this study, personal variables account for 28 per cent of dropout rates and are the most significant factor affecting student attrition. According to the same study, institutional variables including the university setting and course costs account for 17% of dropout rates. It is also noted that a small number of students are most likely to leave school due to concerns about homesickness and transition issues [23].

After three semesters of student enrollment, [24] used logistic regression and decision trees to estimate the student dropout rate and success factor on a resample review of the data from the Karlsruhe Institute of Technology (KIT) with 95% accuracy.

On the other hand, several researchers and authors have used data science and machine learning predictive methodologies to forecast student behaviour and performance in educational contexts. The researchers [25] built three prediction models that forecast the student's enrolment at the department level for higher education institutions in both private and public universities using three distinct machine learning techniques, primarily J48, naive Bayes, and neural networks.

3. Materials and Method

This section focuses on the notion of predicting students' attrition from pursuing computer science degree utilizing methods of machine learning. The approaches were based on the sample of students' data from 2015 to date concerning their class labels. This fact led to the system's construction using the obtained data set from Federal University Lokoja (FUL), Nigerian with another related literature review such as journals or articles for the smooth running of this paper.

The analysis of the proposed system methodology is based on the concept of collecting a sample of student data dataset; this sample was used to form the basis of our approach toward solving the problem definition as follows:

a. **Machine learning Approach**

1. Sample data collection (student's data in this case)

ii. Pre-processing- the data were provided with two labels, graduate or dropout, since it is a supervised learning approach, then it is a binary classification.

iii. Feature extraction – Utilize Weka library's feature et=xtraction techniques (to convert the dataset into binary classification analysis)

iv. Datasets Resampling - done by applying the training set and testing set during system development analysis using Weka tools.

To create the system with all the above-mentioned requirements, develop the model with the Weka library and utilize the java programming language with the Netbeans IDE environment. The provided algorithm was then used to perform the classification model and structured data analytics.

b. Random Forest

Classification accuracy for RF has been greatly enhanced by growing an ensemble (group) of trees and letting them vote on the most popular class. These ensembles are frequently grown using random vectors, which control the development of each tree in the ensemble. An early example is bagging, where each tree is grown using a random pick from the training set of instances (without replacement).

Another illustration is random split selection, in which each node chooses a split at random from the training sets it generates by "randomizing the outputs from the initial training set". Another strategy is to choose the training set at random from a set of weights applied to the training sets. This study work used the strategy that randomly chooses a subset of features to employ to grow each tree.

✓ **Random forest:** employing random forest to forecast student attrition, this is taken to a new level by fusing the idea of an ensemble with trees. As a result, in terms of ensemble learning, the random forest is a strong learner while the trees are weak learners.

For a certain number of trees T, the following is how such a system is trained:

1. To build a subset of the data, randomly choose N examples using replacement. There should be at least 76% of the whole set in the subgroup.
2. For every node:

   1. From among all the predictor variables, m predictor variables must be chosen at random for some integer m.

   2. A binary split is performed on that node using the predictor variable that, according to some objective function, offers the optimal split.

   3. Pick another m variable at random from all the predictor variables and repeat the process at the next node.

   Additionally, there are three somewhat different systems depending on the value of m:

   1. Using a random splitter, choose m = 1.

   Breiman's bagger, version 2, using m as the total number of predictor variables

   3. M predictor variables in a random forest. The three value forms that Brieman recommends are 12m, m, and 2m

Random tree

A random forest is a group of tree predictors, and a random tree [26] is one of them. It can handle issues with classification and regression. According to how it is classified: The random trees classifier takes a feature vector as input, classifies it using each tree in the forest, and outputs the class label that earned the most votes. Regression cases include classifier answers that are averaged

over all of the forest's trees. Despite using distinct training sets, all of the trees are trained using the same parameters.

c. System Design

The method to achieve this works as follows:

FUL Student data collection

FUL Student data pre-processing

FUL Student Feature extraction

Training-set and Test-set

Model Building

Based on the supervised learning described above, the algorithm will be trained and labelled according to the class to which it belongs. The algorithm classifies the unlabeled data based on the learnt relationship between the feature sets and the output after learning the relationship from the labelled data. Hence conceptual framework of the model as shown in Figure 2.

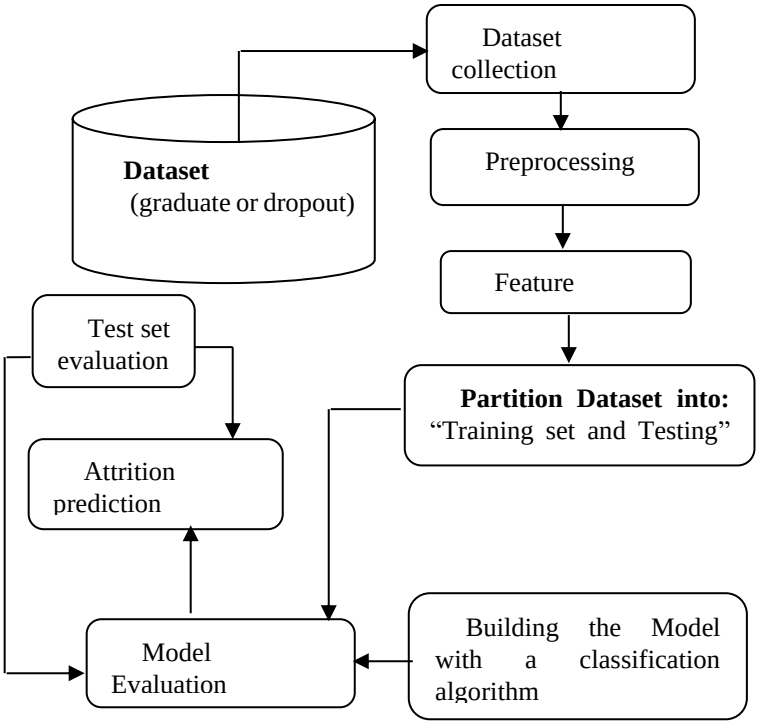

Figure 2. Steps for Student Attrition Prediction

1) Pre-Processing

In this step, ample geometric correction and sifting are done. The preprocessing uses the output of the classifier to take the required action to improve the performance.

2) Supervised Classification

Prior information must be obtained by the analyst for supervised categorization. To create representative parameters for each class and to determine decision limits, the analyst needs access to a large enough known dataset. The training stage is the name of this procedure. After training, the classifier makes

classifications based on the learned parameters. Techniques of supervised classification that are frequently utilized include maximum likelihood, minimal distance, and classification methods.

The main benefit of supervised categorization is that a mistake can be quickly found and attempted to be fixed by the operator. The drawback is that choosing training data can be expensive and time-consuming, and occasionally, the training data may not accurately reflect the conditions over the entire image. The analyst's choice of training sets is subject to inaccuracy.

3) **Unsupervised Classification**

In unsupervised classification, prior knowledge of the classes is not necessary, as the interference of humans is not required since the process is completely automated. The image and data are classified by clustering algorithms. Values within a given data type should be close enough in the measurement space. The spectral classes that are produced by the unsupervised classification are based on the organic grouping of image values. As a result of this system's usage of a clustering technique that is quick and requires few operating settings, it is becoming more and more popular in the field of GIS database maintenance. Migrating means clustering classifier (MMC). is the most widely used unsupervised classifier.

The time taken in unsupervised classification is less and human errors are minimal and there is no need for erstwhile knowledge. The disadvantage is that sometimes the clusters in the spectral region may not match our perception of classes.

4) **Dataset Description**

For this investigation, the Federal University in Lokoja, Kogi State, Nigeria provided the student data from 2015 to the present. There are "attributes and occurrences" in this dataset. Graduate or drop out, which are just class names, are the two possible values for the first attribute, the class attribute. 4407 instances make up the dataset. The name of the label is represented by attribute one. The second characteristic is the student's course enrolment, whose values are limited to coursework or evaluations.

5) **Experimental Set Up**

All Weka, an open-source application that uses the Java programming language and the Netbeans IDE, was used to carry out and calculate all of the tests. The University of Waikato in New Zealand developed the data mining software application WEKA [27], which uses JAVA to execute data mining techniques. Weka is noteworthy because it is widely used in the fields of machine learning and data mining research.

Weka is the name of a bird in New Zealand. WEKA is a modern tool for developing machine learning (ML) strategies and using them to solve practical data mining problems. The content of the "dataset, pre-processing of the dataset, and executed classification and detection" concerning feature selection, random forest, and random tree are covered in greater detail in the next subsection.

6) **Selection of Training Data**

The traits that best represent the pattern for predicting whether attribution is graduate or dropout are picked in this stage.

7) **Classification of Outputs**

The output of the anticipated outcome is categorized into various groups under this, including graduate or dropout.

4. Implementation and Results

The general "implementation and outcomes" of this study are the focus of this section. Based on this data set, which was gathered from Federal University Lokoja in Kogi State, Nigeria, the system was constructed with two classes of students: dropouts and graduates. The suggested RF and RT algorithms were successful in predicting student attrition. These methods were used to train the acquired data, which was then used to develop a model. Before applying a classification strategy to this dataset, preprocessing and feature extraction were performed. The categorization used for this study was based on RF and RT, both of which were capable of capturing the whole training sample of data that was necessary to apply the test to predict student attrition.

When those components were organized into the Weka library using the Java programming language, this model was implemented with a set of features or qualities to identify their performance.

4.1. Model Evaluation

On two folds, which represent a sample of the dataset used to generate the prediction, a classification model experiment was run. After the model was built using a training set, it was used in the test set to predict the result with an unidentified class label as well as to forecast a new class label utilizing the classes involved. See the model assessment of RF below and table 1 with analysis.

Correctly Classified Instances	*3166*	*79.7883 %*
Incorrectly Classified Instances	*802*	*20.2117 %*
Kappa statistic	*0.5582*	
K&B Relative Info Score	*151764.6823 %*	
K&B Information Score	*1471.6516 bits*	*0.3709 bits/instance*
Class complexity | order 0	*3847.2872 bits*	*0.9696 bits/instance*
Class complexity | scheme	*2470.8673 bits*	*0.6227 bits/instance*
Complexity improvement (Sf)	*1376.4199 bits*	*0.3469 bits/instance*
Mean absolute error	*0.3104*	
Root mean squared error	*0.372*	
Relative absolute error	*64.791 %*	
Root relative squared error	*75.9992 %*	
Total Number of Instances	*3968*	

Table 1. Results and Analysis with Random Forest

Class	Precision	Recall	F-Measure	ROC Area
Graduate	0.783	0.920	0.846	0.876
Dropout	0.834	0.613	0.707	0.876

Below are the classification results in a random tree and table 2 with analysis:

Correctly Classified Instances	*3112*	*78.4274 %*

Incorrectly Classified Instances	*856*	*21.5726 %*
Kappa statistic	*0.5418*	
K&B Relative Info Score	*194843.3643 %*	
K&B Information Score	*1889.3826 bits*	*0.4762 bits/instance*
Class complexity | order 0	*3847.2872 bits*	*0.9696 bits/instance*
Class complexity | scheme	*436219.3642 bits*	*109.9343 bits/instance*
Complexity improvement (Sf)	*-432372.077 bits*	*-108.9647 bits/instance*
Mean absolute error	*0.2443*	
Root mean squared error	*0.4337*	
Relative absolute error	*50.9968 %*	
Root relative squared error	*88.6255 %*	
Total Number of Instances	*3968*	

Table 2. Results and Analysis with Random Tree

Class	Precision	Recall	F-Measure	ROC Area
Graduate	0.801	0.854	0.827	0.797
Dropout	0.755	0.678	0.714	0.800

Algorithm Comparison

The detailed Performance Evaluation on the test set by the class is shown in the Table 3 below

Table 3. Results and Analysis from the comaprison

Algorithm	Accuracy (%)	Train Set	Test Set	Binary Class
Random Forest	97.00%	3968	439	Graduate
Random Tree	54.98%	3968	439	Dropout

Incredibly, the random forest achieved an excellent performance in these findings compared to a random tree in terms of accuracy by algorithms, 54.98% in the random tree and 97.00% in the random forest respectively under the same circumstances. The results of this study utilized a set performance assessment by class as shown in table 3 above.

5. Conclusion and Future Work

A topical study showed that researchers have suggested numerous methods to predict student attrition, enrollment, failure etc. Major works have been done on enrollment and retention prediction which are biased towards limited feature space. Consequently, it's necessary to look for new features for predicting student attrition. The research project began by presenting an overview of the student attrition prediction in their degree course career analyzing the phenomenon's effects on organizations, institutions, and people. We also explored the behaviour of student attrition with a high risk of dropout. Related works are done on enrollment, retention, and student attrition prediction were also summarized and discussed.

The study of the current system, analysis of reports of student attrition prediction from the literature, and analysis of institution data were the first steps taken in this paper to develop a student attrition prediction model. The attrition process that could be converted into a feature set was then identified. We detected significant traits with the assistance of machine learning algorithms.

As part of our contribution, we also introduced a correlation-based feature i.e. the system didn't directly utilize the two sets of the algorithm proposed as a default, it was customized by updating the parameters with the choice of java programming language, those two sets of the algorithm were undergone finetune, and this show that the results obtained from the set of the feature was better than the default algorithm, therefore with the help of feature sets as well as to predict and classify the unknown student attrition with machine learning model.

References

[1] Gansemer-Topf, A. M. and Schuh, J. H. (2006). Institutional selectivity and institutional expenditures: Examining organizational factors that contribute to retention and graduation. *Research in Higher Education, 47(6):613–642.*

[2] Yu, C. H., DiGangi, S., Jannasch-Pennell, A., and Kaprolet, C. (2010). A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science, 8(2):307–325.*

[3] Zhang, Y., Oussena, S., Clark, T., and Kim, H. (2010). Using data mining to improve student retention in higher education: a case study. *International Conference on Enterprise Information Systems*.

[4] Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49(4):498–506.*

[5] Lorenz, k., (2018). Predicting Student Dropout: A Machine Learning https://www.researchgate.net/publication/322919234

[6] Office, F. L. (2017). Fachkrfteengpassanalyse. Technical report, Arbeitsmarktberichterstattung.

[7] Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research, 45(1):89–125*

[8] Yu, C. H., DiGangi, S., Jannasch-Pennell, A., and Kaprolet, C. (2010). A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science, 8(2):307–325.*

[9] Kovacic, Z. (2010). Early prediction of student success: Mining students' enrollment data. *Proceedings of Informing Science and IT Education Conference pages 647–665*

[10] Zhang, Y., Oussena, S., Clark, T., and Kim, H. (2010). Using data mining to improve student retention in higher education: *a case study. International Conference on Enterprise Information Systems*.

[11] Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. Decision Support Systems, 49(4), 498–506. https://doi.org/10.1016/j.dss.2010.06.003.

[12] Banumathi, A., & Pethalakshmi, A. (2012). A novel approach for upgrading Indian education by using data mining techniques. 2012 IEEE International Conference on Technology Enhanced Education (ICTEE), 1–5. https://doi.org/10.1109/ICTEE.2012.6208603

[13] Alam, M. M., Mohiuddin, K., Das, A. K., Islam, Md. K., Kaonain, Md. S., & Ali, Md. H. (2018). A Reduced feature-based neural network approach to classify the category of students. *Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence - ICIAI '18, 28–32. https://doi.org/10.1145/3194206.3194218*

[14] Manrique, R., Nunes, B. P., Marino, O., Casanova, M. A., & Nurmikko-Fuller, T. (2019). An Analysis of Student Representation, Representative Features and Classification Algorithms to Predict Degree Dropout. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge - LAK19, 401–410. https://doi.org/10.1145/3303772.3303800*.

[15] Puarungroj, W., Boonsirisumpun, N., Pongpatrakant, P., & Phromkhot, S. (2018). Application of Data Mining Techniques for Predicting Student Success in English Exit Exam. *Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication - IMCOM '18, 1–6. https://doi.org/10.1145/3164541.3164638*.

[16] Supianto, A. A., Julisar Dwitama, A., & Hafis, M. (2018). Decision Tree Usage for Student Graduation Classification: A Comparative Case Study in Faculty of Computer Science Brawijaya University. *2018 International Conference on Sustainable Information Engineering and Technology (SIET), 308–311. https://doi.org/10.1109/SIET.2018.8693158*

[17] Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems, 33(1), 107–124. https://doi.org/10.1111/exsy.12135*

[18] Ratnaningsih, D. J., & Sitanggang, I. S. (2016). Comparative analysis of classification methods in determining non-active student characteristics in Indonesia Open University. *Journal of Applied Statistics, 43(1), 87–97*

[19] Hossin, M., Sulaiman, M.N. (2015). A Review of Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process, 5(2), 01–11*.

[20] Petkovski A., Stojkoska B., Trivodaliev K., and Kalajdziski S., (2016) "Analysis of Churn Prediction: A Case Study on Telecommunication services in Macedonia," *in Proceedings of 24th Telecommunications Forum, Belgrade, pp. 1-4, 2016*.

[21] Yukselturk E., Ozekes S., and Türel Y.,(2014) "Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program," *European Journal of Open, Distance and e-learning, vol. 17, no. 1, pp.118-133, 2014*

[22] Rai S. and Jain A., (2013) "Students' Dropout Risk Assessment in Undergraduate Courses of ICT at Residential University-A Case Study," *International Journal of Computer Applications, vol. 84, no. 14, pp. 31-36, 2013*.

[23] Rai S. and Jain A., (2013) "Students' Dropout Risk Assessment in Undergraduate Courses of ICT at Residential University-A Case Study," *International Journal of Computer Applications, vol. 84, no. 14, pp. 31-36, 2013*

[24] Kemper L., Vorhoff G., and Wigger B., (2020), "Predicting Student Dropout: A Machine Learning Approach," *European Journal of Higher Education, vol. 10, no. 1, pp. 28-47, 2020.*

[25] Mulugeta M. and Borena B.,(2013) "Higher Education Students' Enrolment Forecasting System Using Data Mining Application in Ethiopia" *HiLCoE Journal of Computer Science and Technology, vol. 2, no. 2, pp. 37-43, 2013*

[26] Wikipedia contributors, "Random_tree", Wikipedia, The Free Encyclopedia. *Wikimedia Foundation, 13 - Jul-2014.*

[27] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. itten, and L. Trigg,(2005) Weka, in Data Mining and Knowledge Discovery Handbook, *Springer, 2005, pp. 1305 – 1314*