International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 11 | Nov - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Predicting Student Dropout Rates in Higher Education: A Comparative Study of Machine Learning Algorithms

Prof.(Dr) Ganesh G.Dongre
Vishwakarma Institute of Technology,Pune
ganesh.dongre@vit.edu

Shreyas D. Khamkar
Vishwakarma Institute of Technology, Pune, India
shreyas.khamkar22@vit.edu

Om P. Jathar Vishwakarma Institute of Technology, Pune, India
om.jathar22@vit.edu

Adwait V. Kharmale
Vishwakarma Institute of Technology, Pune, India
adwait.kharmale22@vit.edu

Ayush A. Kude
Vishwakarma Institute of Technology, Pune, India
ayush.kude22@vit.edu

Aditya L. Jatade
Vishwakarma Institute of Technology, Pune, India
aditya.jatade22@vit.edu

Ishan Herekar
Vishwakarma Institute of Technology, Pune, India
ishan.herekar22@vit.edu

Abstract:

Recent years have seen a significant increase in study interest in the areas of predicting student performance, avoiding failure, and identifying the variables affecting student dropout. One important indicator in online and open distance learning courses is the student dropout rate. We purpose the naive bayes classification method to construct the student dropout prediction using naive bayes.

This work examines the critical topic of forecasting student dropout rates in higher education using machine learning approaches, with a particular emphasis on the random forest algorithm and the naive bayes algorithm. The study's goal is to properly anticipate dropout rates using data mining methods and machine learning algorithms after conducting a thorough evaluation of existing literature and approaches. The systematic method consists of data collection from a Kaggle dataset, data preparation to solve class imbalance via SMOTE oversampling, and algorithm selection. Random forest and naive Bayes approaches outperform other machine learning algorithms in terms of accuracy, sensitivity, specificity, and precision. The study underscores the importance of considering diverse factors such as demographic data, socioeconomic factors, and academic performance in dropout prediction models. The implications of this research extend beyond academia, with the potential to inform proactive interventions and support systems, ultimately leading to improved student outcomes and institutional effectiveness. According to this paper, the paper outputs that for the binary classification on the data set used in this project has best performed with Naive Bayes and Random Forest Algorithm with SMOTE oversampling.

Keywords- *SMOTE oversampling, machine learning, Random forest, naive bayes.*

I. INTRODUCTION

The quick development of online and open courses has shown to have a big influence on distance learning and open education, enabling universities and organizations to create various learner involvement and course distribution approaches. Additionally, academics in the disciplines of data analytics and education research have been interested in remote learning courses that are supported by virtual learning environments (VLEs), such as Massive Online Open Courses (MOOCs).

Virtual learning environment providers collaborate with more than 100 top-ranked academic institutions to offer open, frequently free, education [19]. The early prediction of student dropout and the identification of key factors influencing this phenomenon would enable improved decision-making. Concerning the development of new programs and strategies targeted at providing first-year university students with academic support and assistance throughout their transition. The purpose of these techniques is to decrease the impact of the factors that have been identified in order to reduce the dropout rate.

Dropout rates have far-reaching effects on educational institutions as well as society at large, in addition to having an impact on specific students. Predictive methods that can detect at-risk pupils in their early stages and preventive approaches are needed to address this situation. Predicting with high accuracy the possibility of high first-semester university dropout rates has been the aim of the most recent research in this topic. Some research only looked at the first year of study for students, while others looked at five or more years of college [4,17]. These studies have looked at the incidence over wide time periods. Similar to this, the criteria for designating someone as a dropout have evolved over time. Currently, it can mean missing more than two semesters of their first year of education, or it might include longer periods of time—up to three years—during which they do not attend. In order to prevent dropout scenarios, prediction will aid with early action. ch introduces a systematic methodology for building a student dropout prediction system using various algorithms. Utilizing a dataset sourced from Kaggle, encompassing demographic, socioeconomic, and academic performance data, the study aims to predict student dropout rates in higher education. After data preprocessing and balancing, the dataset is split into training and testing sets. During the preprocessing we had used the SMOTE oversampling methode. The performance of Decision Tree, Random Forest, Logistic Regression, KNN, and Naive Bayes algorithms is compared, with Naive Bayes exhibiting the highest accuracy, sensitivity, specificity, and precision.

II. LITERATURE REVIEW

Student dropout analysis with application of data mining methods by Mario Jadric Zeljko Garaca, Maja Cukusic (2010) paper proposes this paper uses data mining to tackle high dropout rates among junior students in higher education [1]. It analyzes student data using techniques like logistic regression, decision trees, and neural networks to identify areas for improvement. By comparing models, it suggests strategies to reduce dropout rates. The outcomes demonstrated in the above three models accuracy is 77.4%, 92%, 92% respectively. The decision tree and neural network provided the same accuracy. The comparative analysis of the models neural network and decision tree having a good accuracy in student dropout prediction[1].

Predicting student dropout: A machine learning approach by Lorenz Kemper, Gerrit Vorhoff, Berthold U. Wigger (2019) Using machine learning approaches, logistic regression, and decision trees, to forecast student dropout at the Karlsruhe Institute of Technology (KIT). Using examination data available across universities, our methodical approach can be easily replicated elsewhere. Decision trees slightly outperform logistic regressions, but both methods achieve impressive prediction accuracies, reaching up to 95% after three semesters. Even after just one semester, we achieve a classification accuracy of over 83%. The outcomes demonstrated above that model has a 90% accuracy rate in students dropout prediction[2].

Yafeng Zheng, Zhanghao Gao, Yihang Wang, Qian Fu3 "MOOC Dropout Prediction Using FWTS-CNN Model Based on Fused Feature Weighting and Time Series" (2020).This article addresses the challenge of high dropout rates in Massive Open Online Courses (MOOCs) by proposing a convolutional neural network model called FWTS-CNN. Unlike traditional CNNs, FWTS-CNN integrates feature weighting and behavioral time series to enhance dropout prediction accuracy. It extracts continuous behavioral features from learner logs, filters and ranks them based on importance using decision trees, weights these features accordingly, and constructs a CNN model incorporating both behavioral time series and weighted features. Experiments on the KDD Cup 2015 dataset demonstrate that FWTS-CNN achieves high accuracy, surpassing 87%, an improvement of approximately 2% compared to standalone CNN algorithms. By incorporating behavioral features and behavior time series, FWTS-CNN effectively enhances dropout prediction accuracy in MOOCs.[3]

"Dropout from Higher Education: A Theoretical Synthesis of Recent Research" (2014). The literature on dropout from higher education is extensive, yet much about the dropout process remains unknown. Two major shortcomings contribute to this lack of clarity: inadequate attention to definition and the absence of theoretical models that explain rather than merely describe dropout processes.[4]

Haarika Dasi, Srinivas Kanakal "Student Dropout Prediction Using Machine Learning Techniques"

(2022). Recent years have seen a rise in student dropout rates, posing challenges for educational institutions. This research focuses on predicting dropout using machine learning and educational data. Despite limited progress, the study demonstrates promising results, with prediction accuracy reaching 92-93% on four years' worth of data. By comparing machine learning classifier performance, the study highlights the potential to accurately predict dropout even with a small set of features.[5]

Title	DataSet	Data Preprocessing	Algorithms	Evaluations
Early prediction of dropout in online courses using Artificial Neural Networks	Collected data from the Open edX platform	NA	MLP	MLP was able to predict early dropout in online courses with an accuracy of 80%. This means that the model was able to correctly classify 80% of the students who would eventually drop out of the course.
Using Ensemble Decision Tree Model to Predict Student Dropout in Computing Science	Collected from a single institution in the South Pacific region. Past five years of historic data	1. Missing value imputation: Missing values were imputed using the mean value of the respective column.	The authors used a random forest (RF) model to predict student dropout.	The results of the study showed that the RF model was able to predict student dropout with an accuracy of 82%. The sensitivity and specificity of the model were 78% and 86%, respectively. The AUC of the model was 0.89.
A machine learning approach to Predict the Engineering Students at risk of dropout and factors behind: Bangladesh Perspective	Bangladeshi universities students dataset	We had some missing data which we imputed by mean[12]	SVM, Random Forest and Neural Network	Support Vector Machines (SVM), Random Forests, and Neural Networks for creating the prediction model. Each of these models has its strengths and weaknesses.

Accuracy Enhancement of Prediction Method using SMOTE for Early Prediction Student's	Dataset of 8491 students from XYZ University	SMOTE (SMOTE is a data augmentation technique that creates synthetic data points for the minority class	1) Artificial Neural Networks(AN N) k-nearest neighbors(KN N), Support vector machine(SVM)	The results showed that SMOTE improved the accuracy of all three models. KNN model with SMOTE (69.3%) and the SVM model with SMOTE (69.8%). The authors concluded that SMOTE is an effective technique for improving the accuracy of early graduation prediction models.
A Time Series Classification	NA	NA	Time series forest (TSF)	Results show that the prediction accuracy on two selected datasets
Method for Behaviour-Bas ed Dropout Prediction			classification algorithm.	increases as the portion of data used in the model grows. However, a reasonable prediction accuracy of 0.84 is possible with only 5% of the dataset processed

III. METHODOLOGY

In this study, we describe the methodical process used to create a student dropout prediction system using multiple algorithms and compare the outcomes of these algorithms. The major purpose is to anticipate whether the student will drop out of college or not. We are using the data science process/pipeline to make the forecast. These processes have four steps:

1. Data Collection: We utilized the Kaggle dataset to train and test the model. The dataset's name is "predicts student dropout and academic success". The dataset has 34 columns, of which 17 are categories and 17 are numerical variables. This dataset offers a complete perspective of students enrolled in various undergraduate degrees provided by a higher education institution. It contains demographic data, socioeconomic characteristics, and academic performance information that may be utilized to identify potential predictors of student dropout and academic achievement. Finally, we know the region's unemployment rate, inflation rate, and GDP, which can help us better understand how economic issues influence student dropout rates and academic success results. This provides insight into what inspires students to continue or discontinue their studies in a variety of areas, including agronomy, design, education, nursing, journalism, management, social service, and technology.

2. Data Preprocessing: During data preparation, we first determined if the data was organized or unstructured. Our dataset is organized, so we don't need to do anything with it; nevertheless, it is in binary form. The class names are graduate and dropout. In these two classes, dropouts are the minority, while graduates are the majority. The minority class has 1420 rows, whereas the majority class contains 2210 variables; the overall dataset has 3630 rows and 34 columns. So, instead of unbalancing the classes, we will balance them.

Two methods are available for balancing: SMOTE and ROSE oversampling. We are using SMOTE oversampling in these methods since oversampling in ROSE increases the likelihood of overfitting the model. Through oversampling, a neural network is used to create a new data point that is situated between the original two data points. This states that it continues until there are an equal number of rows in each class or until both classes are in a balanced state. Following oversampling, there are 3614 rows in the graduation class and 4416 rows in the dropout class.

3. Data Splitting: After balancing the data, whole data is loaded into the single variable. The loading phase of our involves the critical step of loading the dataset, as exemplified in the provided code. This step is of paramount importance as it forms the foundation of our entire prediction analysis. Then we are going to split the data in training and testing. In the data splitting we split it into 70:30. For the training we take 70% of data and for testing 30% data. The data is taken for the training which is chosen randomly from the dataset, remaining 30% data is for the testing[16].

3.1.) Training Set: The training set is typically the larger of the two subsets, designed to serve as the basis for developing the predictive maintenance model. It contains a significant portion of the dataset, allowing the model to learn patterns, relationships, and correlations among the features and the outcome variable. By working with a substantial amount of data, the model can capture underlying trends and nuances within the dataset. In this case, training data is 70% of overall data and out of overall data, the first 5621 observations are allocated to the training set, which enables the model to understand the historical patterns associated with equipment performance and failure.

3.2.) Testing Set: The prediction set, on the other hand, is reserved for assessing the model's predictive capabilities. For the prediction data 30% of overall data are taken. It serves as a set of unseen data that the model has not encountered during training. The aim is to evaluate how well the model can generalize its findings to new, unseen data. In this code, remaining rows from the training data set are allocated to the testing set, enabling us to test the model's ability to accurately predict equipment failures on fresh, real-world data. By splitting the dataset into these two sets, we can gauge the model's performance by comparing its predictions on the test set with the actual outcomes. This evaluation process provides valuable insights into the model's accuracy, precision, and its capacity to make reliable predictions, a vital aspect of predictive maintenance. If the model can effectively generalize its findings from the training set to the prediction set, it is a strong indicator of its robustness and practicality in real-world scenarios. In conclusion, data splitting into

training and prediction sets is a standard and essential procedure in predictive maintenance modeling. It enables us to not only develop effective models but also rigorously evaluate their performance and reliability in real-world applications, which is critical for optimizing equipment reliability and minimizing downtime.

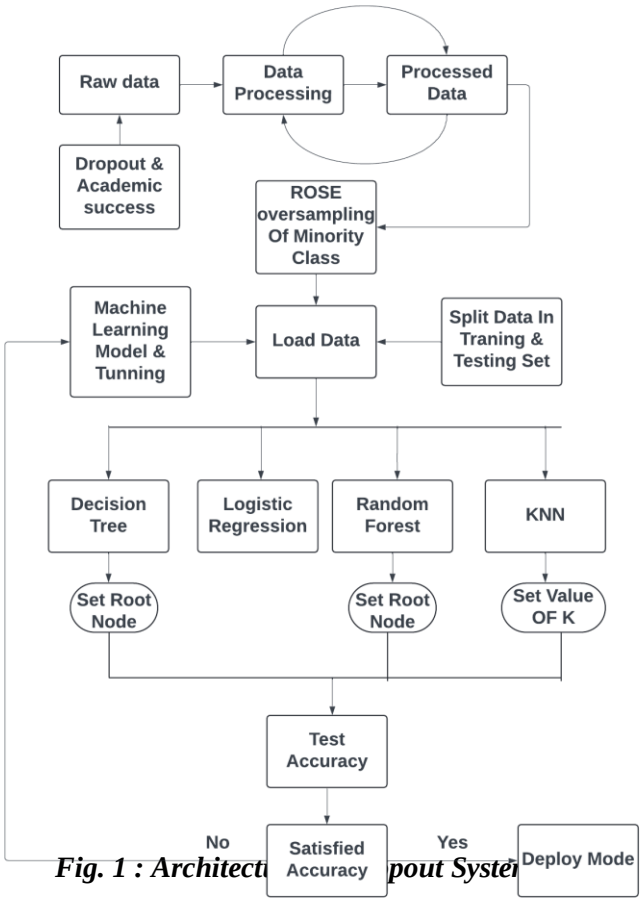

Fig. 1 : Architecti pout Syste

Fig.1. presents a schematic illustration of a student dropout prediction. It shows a structured process with some key parts. In the data processing it includes cleaning the data. Then the processed data is not balanced data, so in the next step is to oversample the dada and balance it. In the next step, load all the data in one variable and split it into training and testing data and then apply the algorithm on the training dataset. In the last step check the accuracy of the model and deploy the model, if the accuracy is not that much then again the tuning the machine learning model.

4.] Algorithms:

1. **Decision Tree:** A decision tree is a hierarchical model that is used in machine learning for classification and regression. It operates by recursively partitioning the input space into subsets, based on the values of input features, to make decisions. Each internal node represents a feature, each branch represents a decision based on that feature, and each leaf node represents the outcome or class label. Decision trees are popular due to their simplicity, interpretability, and ability to handle both numerical and categorical data effectively. They are widely used in various fields including data mining, pattern recognition, and bioinformatics for their intuitive nature and ability to handle complex datasets efficiently.

$$Entropy = \sum_{i=1}^{C} - p_i \times log_2(p_i)$$

The above question for calculating the entropy. We calculate the entropy to decide the root node. Root node depends on the Information Gain (IG), entropy is used for calculating the IG.

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

The above equation is for calculating the IG, it is totally based on the entropy of the total entropy of the dataset and the entropy of every attribute in the dataset. IG is the total entropy minus entropy of the attribute. The attribute having maximum entropy makes it a root node.

In this project we are using this algorithm, the algorithm gives us 85% accuracy and the remaining all the things like sensitivity, precision and specificity are also good.

2. **Random Forest:** Random Forest is the next version of the decision trees / number of decision trees that are together, it is an ensemble learning method used for classification and regression tasks. It operates by constructing multiple decision trees during training and outputs the mode (classification) or mean prediction (regression) of the individual trees. Each tree is trained on a random subset of the training data and a random subset of features, providing diversity and reducing overfitting. Random Forest combines the predictions of multiple trees to improve accuracy and robustness. It is widely used in various applications such as image classification,

bioinformatics, and financial modeling due to its high accuracy, scalability, and ability to handle large datasets with high dimensionality.

3. **Logistic Regression:** Logistic Regression is a statistical method used for binary classification tasks. It models the probability of a binary outcome based on one or more predictor variables by fitting a logistic curve to the observed data. Unlike linear regression, it employs the logistic function to map predictions to probabilities, ensuring outputs lie between 0 and 1. Logistic Regression is widely used in fields such as healthcare (for disease prediction), marketing (for customer segmentation), and finance (for credit scoring). It's valued for its simplicity, interpretability, and efficiency in handling large datasets.

$$Y = 1/1 + e^{-(a0+a1x)}$$

Above question represents the logistic regression equation, in the equation, Y is the dependent variable, a0 represents the value of y when x=0, and a1 represents when y changes per unit change in x.

4. **Knn:** K-Nearest Neighbors (KNN) is a non-parametric algorithm used for classification and regression tasks. It makes predictions based on the majority class (for classification) or the average value (for regression) of the k nearest data points in the feature space. KNN operates under the assumption that similar data points tend to belong to the same class or have similar values. It's used in various applications such as recommendation systems, pattern recognition, and anomaly detection due to its simplicity and effectiveness, especially with small to medium-sized datasets.

$$d = \sqrt{(x_2 - x_1) + (y_2 - y_1)^2}$$

This formula represents a weighted Euclidean distance calculation. The terms x1 and y1 denote the coordinates of the reference point, while x2 and y2 correspond to the coordinates of a second point in a two-dimensional space. The weight assigns asignificance to each dimension, allowing for the adjustment of the contribution of each coordinate to the overall distance calculation. This expression is instrumental in determining the distance between data points, offering a foundation for various proximity-based algorithms and analyses, including our approach in this research.

5. **Naive Bayes:** Naive Bayes is a probabilistic classifier based on Bayes' theorem with the "naive" assumption of independence between features. It calculates the probability of each class given a set of features and selects the class with the highest probability as the prediction. Naive Bayes is widely used in text classification tasks such as spam detection and sentiment analysis, as well as in medical diagnosis and recommendation systems. Despite its simplistic assumptions, Naive Bayes often performs well and is computationally efficient, especially with high-dimensional data.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

The above equation represents the probability equation in the naives theorem. In equestion P(A) & P(B) represents the probabilities of the observing A & B, P(A|B) & P(B|A) represent the Probability of hypothesis A on the observed event B & Probability of the evidence given that the probability of a hypothesis is true respectively.

According to these algorithms it gives us an accuracy of 98%. As compared to the previous algorithm it is a more accurate model.

IV. RESULTS AND DISCUSSIONS

The main aim of this project is to predict the students dropping out of college or not by considering different factors, as mentioned above. This section describes the results obtained for the proposed work in terms of precision, recall, and F-measure. From these parameters, we can know the prediction accuracy of the proposed system to analyse employee satisfaction without the intervention of human beings. The precision parameter is used to analyse accurate positive samples from the total positive samples. The recall is also used to examine the positive predicted values but from the total available samples in measure is used the data. F finds out the relation between the precision and recall values. Mathematically, Accuracy, Sensitivity, Specificity, Precision, recall, and F-measure are given by equation.

Accuracy = TP+TN/TP+TN+FP+FN

Precision = TP/TP+FP Recall/Sensitivity =

TP/TP+FN Specificity= TN/TN+FP

TP=True Positive

TN=True Negative FP

= False Positive FN =

False Negative

Table 1: Confusion matrix for All Algorithm

Model	TP	TN	FP	FN
Naive Bayes	535	533	2	18
Random Forest	510	517	27	34
Logistic Regression	453	483	84	68
Decision Tree	453	476	84	75
KNN	362	506	175	45

After applying the algorithms on the training dataset and testing, it gives us the accuracy, sensitivity, specificity and precision. We apply decision tree, random forest, logistic regression Knn and naive bayes algorithm. Comparative Analysis of various algorithms on testing data.

Model	Accuracy	Sensitivity	Specificity	Precision
Decision Tree	85.38	85	85.79	85
Random Forest	94.39	93.42	94.97	95.03
Logistic Regression	86.02	87.65	84.35	85.18
KNN	79.77	91.83	67.41	74.3
Naive Bayes	98.16	96.73	99.63	99.62

Table 2

As we compare the accuracy column it is easier to find the algorithm which gives us the maximum accuracy i.e. naive bayes. On the other hand sensitivity ,specificity and precision are also maximum as compared to other models. After the naive bayes, random forest also has the best accuracy. Table 3 shows the variation in accuracy in various modes.

Table 3

After going through the accuracy, sensitivity is also important for the model. As we observe in the table 2 sensitivity of the naive bayes is maximum, after the naive bayes random forest has maximum sensitivity. In table 4 we can observe the various models sensitivity through the histogram. In the graph it shows us the minimum sensitivity is of decision tree and logistic regression and decision tree has slightly same sensitivity.

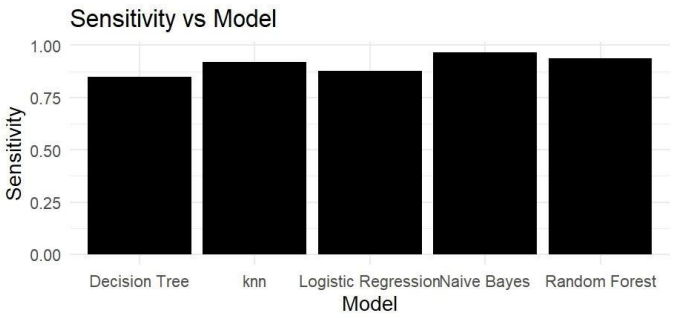

Table 4

Now we move on to the specificity, it is not that much important but we will go through it. As we can see the naive bayes algorithm has very high specificity, on the other hand Knn has very less specificity in this model. Knn is based on the regression but the naive bayes is based on the probability.

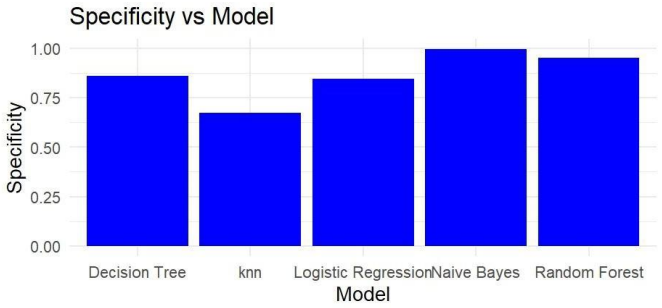

Table 5

Precision, it last but not least. It is a very important point in the testing. In result, naive bayes have a better performance as compared to other algorithms.

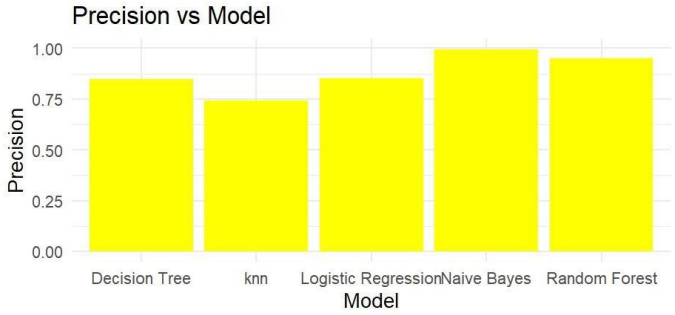

Table 6

V. CONCLUSIONS

Our research underscores the potential of machine learning techniques, particularly random forest algorithms, in predicting student dropout rates. Moving forward, further research and implementation of predictive models in educational settings can lead to improved student outcomes and institutional effectiveness.

The results of our literature research indicated that the most often used approaches were the random forest and decision tree algorithms. Their accuracy is less than 89%, But in our paper, we go through the probability based algorithm for binary classification. According to the paper we used naive bayes algorithm for prediction, Naive bayes algorithm gives more than 95%+ accuracy in the prediction. It also gives more sensitivity, specificity and precision. After the prediction we compare the results of some previous mostly used algorithms with naive bayes, Naive bayes is best for the binary prediction dataset.

VI. REFERENCES

1] Mario Jadric Zeljko Garaca, Maja Cukusic. "Student dropout analysis with application of data mining methods." (2010).

[2] Lorenz Kemper, Gerrit Vorhoff, Berthold U. Wigger . "Predicting student dropout: A machine learning approach." (2019)

[3]Yafeng Zheng, Zhanghao Gao, Yihang Wang, Qian Fu3"MOOC Dropout Prediction Using FWTS-CNN Model Based on Fused Feature Weighting and Time Series." (2020)

[4] "Dropout from Higher Education: A Theoretical Synthesis of Recent Research" (2014)"- A review.

[5] Haarika Dasi, Srinivas Kanakal "Student Dropout Prediction Using Machine Learning Techniques" (2022).

[6]M.Revathy Research Scholar, S. Kamalakkannan Associate Professor. "Collaborative learning for improving intellectual skills of dropout students using data mining techniques" (2021)

[7] Diogo E. Moreira da Silva , Eduardo J. Solteiro Pires , Arsénio Reis, Paulo B. de Moura Oliveira, João Barroso . "Forecasting Students Dropout: A UTAD University Study." (2022)

[8] Marcell Nagy, Roland Molontay. "Predicting Dropout in Higher Education based on Secondary School Performance." (2018)

[9] Ainul Yaqin, Majid Rahardi, Ferian Fauzi Abdulloh "Accuracy Enhancement of Prediction Method using SMOTE for Early Prediction Student's Graduation in XYZ University" (2022)

[10]Mohammed Naseem, Kaylash Chaudhary, Bibhya Sharma, Aman Goel Lal. "Using Ensemble Decision Tree Model to Predict Student Dropout in Computing Science" (2019)

[11] Hermel Santiago Aguirre Montaño, Ma. Carmen Cabrera-Loayza. "Early prediction of dropout in online courses using Artificial Neural Networks" (2021)

[12] Kelly J. de O. Santos, Angelo G. Menezes, Andre B. de Carvalho, Carlos A. E. Montesco. "Supervised Learning in the Context of Educational Data Mining to Avoid University Students Dropout." (2019)

[13] Sheikh Arif Ahmed, Shahidul Islam Khan."A machine learning approach to Predict the Engineering Students at risk of dropout and factors behind: Bangladesh Perspective" (2019)

[14] Haiyang Liu, Zhihai Wang, Phillip Benachour, Philip Tubman, "A Time Series Classification Method for Behaviour-Based Dropout Prediction," (2018)

[15] Ace C. Lagman, Lourwel P. Alfonso, Marie Luvett I. Goh, Jay-ar P. Lalata, Juan Paulo H. Magcuyao, Heintjie N. Vicente,"Classification Algorithm Accuracy Improvement for Student Graduation Prediction Using Ensemble Model".(2020)

[16] Martín Solís, Tania Moreira, Roberto González, Tatiana Fernández, María Hernández "Perspectives to Predict Dropout in University Students with Machine Learning ." (2018)

[17] Aguiar, E., Ambrose, G.A., Chawla, N.V., Goodrich, V & Brockman, J, "Engagement vs Performance: Using Electronic Portfolios to Predict First Semester Engineering Student Persistence".(2014)

[18] Plagge, Mark, "Using artificial neural networks to predict first-year traditional students second year retention rates."(2013)

[19] M. H. Baturay, "An overview of the world of MOOCs," ProcediaSocial and Behavioral Sciences, vol. 174, 2015, pp. 427-433.