International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

# Student Data Clustering Using Machine Learning :Naïve Bayes Classifier & Decision Tree Algorithm

**Dr. S. Om Prakash[1], G. Harshvardhan[2]**

[1]Professor, Dept of CSE Raghu Institute of Technology, Visakhapatnam
[2]CSE Student, Raghu Institute of Technology, Visakhapatnam

**ABSTRACT**

Student dropout is a significant issue in educational institutions worldwide, posing challenges to both students and educators. Early identification of students at risk of dropping out can enable timely interventions and support, potentially mitigating dropout rates. In this study, we propose a machine learning-based approach for predicting student dropout, leveraging a variety of student-related data. The data include demographic information, academic performance metrics, attendance records, socio-economic factors, and extracurricular activities.

We preprocess the data by handling missing values, encoding categorical variables, and scaling numerical features. Feature selection techniques are employed to identify the most relevant predictors of dropout. Various classification algorithms, including logistic regression, decision trees, random forests, support vector machines, and neural networks, are evaluated for their predictive performance.

The models are trained on historical data, split into training and testing sets, and evaluated using metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC). Hyperparameter tuning is performed to optimize the models' performance further.

The results demonstrate the effectiveness of the proposed approach in accurately predicting student dropout. Additionally, we provide insights into the factors contributing most significantly to dropout risk, facilitating targeted interventions by educational institutions.

This study contributes to the field of educational data mining by offering a practical framework for dropout prediction, thereby empowering educators to proactively address the issue of student attrition. Future research directions include the integration of additional data sources and the development of personalized intervention strategies based on predictive analytics.

**Keywords:** Student dropout prediction, Machine learning, Educational data mining, Predictive modeling, Early intervention, Academic success.

**OBJECTIVES:**

- Help universities cope up with the phenomenon of adapting to rising rate of dropouts
- Prevention and intervention services to upgrade the students' retention rate.

**SCOPE OF THE PROJECT:**

- To study the how efficient the campus education methods are which will be extremely helpful for both students and Institution.
- To Build a predictive model which can be used to forecast the probability of a randomly chosen student, who will be tested if he/she will graduate or not.
- To Identity the factors which affect the probability of a student to dropout of a technical education.

**TECHNIQUES AND METHODS:**

- Data mining techniques
- Feature selections
- Classification algorithms in ML
- Prediction based on attributes and performance

**INTRODUCTION**

The term "dropout" indicates the termination by a student from school, college or any other educational institute without fulfilling the registered course. The word dropout suggests that if any student leaves from an educational institution for any reason they have even before finishing a their chosen program of studies. This act will be called drop out only if the said student do not join any other course in the same or an alternate college. Each year, more than a million students will leave the school or any other educational institution without full filling the course on various university or school that is around 8,000 students every day of the academic year. The action of dropping out is an effect of various causes, personal and professional. Our project is useful for University-based or School-based for examine the student behaviour, that leads to a dropout or not in the early stage and can prevent the dropout from taking necessary action towards the dropout reason. The existing method is very time consuming and not very accurate and focuses on only specific factors. The proposed method is a combined approach that takes into consideration factors such as demographics, academic performance, health issues, place of residence etc., which increases the accuracy and implements methods that reduce the time taken for prediction.

**DOMAIN:**

- Major Domain: MACHINE LEARNING
- Sub Domain: Data Mining

**Major Domain - Machine Learning:**

The essence of machine learning is the algorithms that give computers the ability to learn from data, and then make predictions and decisions. This decision process is called Classification. The algorithm that performs the decision process is called a classifier. Although, there are methods that can use the raw data for training things like photos and sounds, there are many algorithms to decrease the complexity of real objects of things what are called features. Features are values that helpfully characterize the objects that we desire to classify.

In order to train our machine learning classifier to make better predictions, we require training data which is

a sample case data from the original data set. The simple approach of dividing the decision space up into boxes can be represented by what's called a decision tree.

A machine learning algorithm that produces decision trees needs to choose what features to divide on and then for each of those features, what features to use for the decision. Decision Trees are just one basic example of a machine learning technique.

Techniques like Decision Trees and Support Vector Machines are strongly rooted in the field of statistics, which has dealt with making confident decisions, using data, long before computers ever existed.

**Sub-Domain – Data Mining:**

Data mining is the process of finding similarities, patterns and relations in huge data sets to predict outcomes in future. Using a wide range of methods, this information can be used to elevate the revenues, costs to be cut, improve the relationships with clients, reduce risk factors and much more.

There are wide and diverse applications of Data Mining. Initially, started as the technique and method to investigate interesting patterns, data mining now works as the basis to a number of machine learning and Artificial Intelligence applications worldwide. Few of the Data mining applications are:

- Market Segmentation
- Fraud detection
- Market based analysis
- Trend Analysis
- Prediction Systems
- Customer churn
- Insurance and health car

## PROBLEM DEFINITION INCLUDING THE SIGNIFICANCE AND OBJECTIVE

There is a distinct rise in the demand for engineers in the IT industry too in our country. Though there is a dire need, most of the students are not aware of this. Therefore, though the industry is rising in demand, the students are not able to meet their requirements and the eligibility criteria. Due to various factors, many may undergo severe issues that may force them to discontinue their studies. It might be financial, personal, or any unforeseen demanding situations. Thus, the administration and management of the colleges waste huge chunks of money on discontinued students' resources. A drop out prediction system can help the institutions to balance and use their finances and resources.

## LITERATURE SURVEY

### INTRODUCTION TO THE PROBLEM DOMAIN TERMINOLOGY

Today Artificial Intelligence (AI) is beyond the technologies of blockchain and quantum computing. It has also found its place in smaller projects and is made easier as even the common man also to work with. Machine Learning models are created to re-train the existing models for better performance and resulted in an efficient way to solve a problem. High-Performance Computing (HPC), which are now easily in reach to everyone which resulted in an unexpected surge and for IT professionals having Machine Learning skills.

Predictive analytics is a section of data mining sighted at obtaining predictions about future results based on factual data and techniques such as statistical modelling algorithms and machine learning. The science of

predictive analytics can produce insights with a vital degree of precision and accuracy. With the assistance of sophisticated data mining tools and algorithm models, any industry can now use past and current data to reliably predict trends and performances which are seconds, days, or years into the future.

Few examples of the various industries which use the data mining techniques to predict the future trends and analyze them are:

Aerospace: Used to forecast the result of particular maintenance operations on aircraft reliability, use of fuel and availability.

- **Automotive:** To consolidate records of segment sturdiness and predict the failure rate in the upcoming plans of vehicle manufacturing. The driver's behaviour is studied to improve the driver assistance technologies and, ultimately, autonomous vehicles.
- **Energy:** Predict long-term price and supply-demand ratios and to determine the impact of weather disasters, failure of equipment, rules and additional variables on service costs.
- **Financial services:** To develop credit-risk models. Predict the trends of the financial market. Predict the influence of new strategies, laws and regulations on business markets.
- **Manufacturing:** Foretell the location and machine failure rates. To improve the efficiency of raw material deliveries based on predicted future demands.
- **Law enforcement:** Use the trends of crime data to determine neighbourhood that may require further security at definite times of the year.
- **Retail:** Develop an online client in real-time to decide whether providing extra product knowledge or incentives will enhance the probability of a finished transaction.

## EXISTING SYSTEM:

An educational institute contains student records that are affluence of information but are too huge for one person to understand in their entirety. Finding essential features from this data is an important task in educational research. Finding the academic and financial status of each student in an institute is a wearisome task. Hence, the modification of the system includes time consumption, less efficiency and less user satisfaction. Furthermore, this system is a manual process that supplements the limitation.

## DESIGN OF THE PROPOSED SYSTEM:

The proposed system uses two data mining techniques for predictive analysis.

- Decision Tree Classifier
- Gaussian Naïve Bayes Algorithm

## RELATED WORKS:

1. Data processing Approach for Predicting Student and Institution's Placement Percentage by Professor Ashok M, professor Apoorva A ,2016 International Conference on Computational Systems and knowledge Systems for Sustainable Solutions during this paper author has used the info mining technique for the prediction of the student's placement. For the prediction of student's placement author has divided the information into the 2 segments, first segment is that the training segment which is historic data of passed out students. Another segment consists of current data of scholars, supported the

historic data author has designed the algorithm for calculating the position chances. Author has used the varied data processing algorithms like decision tree, Naive Bayes, neural network and therefore the prosed algorithm were applied, and decision are made with the assistance of confusion matrix.

2. Student Placement Analyzer: A Recommendation System Using Machine Learning, by Senthil Kumar Thangavel, Divya Bharathi P, Abijith Sankar, International Conference on Advanced Computing and Communication Systems (ICACCS -2017) ), Jan. 06 - 07, 2017, Coimbatore, INDIA during this paper author is concern about the challenges face by any institute regarding the position. the location prediction is extremely complex when the quantity of the entities increases in any institute. With the assistance of machine learning this complex problem of prediction are often easily solved. during this paper all the tutorial record of student is taken into consideration. Various classification and data making algorithms are used like Naïve Bayes, Decision Tree, SVM and Regressions. After the prediction of the scholars will be placed in of the given category that's core company, dream company or support services.
3. A Placement Prediction System Using K-Nearest Neighbors Classifier, by Animesh Giri, M Vignesh V Bhagavath, Bysani Pruthvi, Naini Dubey, Second International Conference on Cognitive Computing and data Processing (CCIP), 2016 the position prediction system predicts the probability of scholars getting placed in various companies by applying K-Nearest Neighbors classification. The result obtained is additionally compared with the results obtained from other machine learning models like Logistic Regression and SVM. the tutorial history of student together with their skill sets like programming skills, communication skills, analytical skills and team work is considered which is tested by companies during recruitment process. Data of past two batches are taken for this system.
4. Class Result Prediction using Machine Learning, by Pushpa S K, professor, Manjunath T N, Professor and Head, Mrunal T V, Amartya Singh, C Suhas, International Conference on Smart Technology for Smart Nation, 20172017 during this paper, the results of a category is predicted using machine learning. Performance of scholars in past semester together with ample internal examinations of this semester is taken into account to predict whether the scholar passes or fails within the current semester before attempting the ultimate examination. The author uses SVM, Naive Bayes, Random Forest Classifier and Gradient Boosting to compute the result. Boosting is an ensemble learning algorithm which mixes various learning algorithm to get better predictiveperformance.
5. Student Placement Analyzer: A Recommendation System Using Machine Learning, Apoorva Rao R, Deeksha K C, Vishal Prajwal R, Vrushak K, Nandini, JARIIE-ISSN(O)-2395-4396 Now-a-days institutions face many challenges regarding student placements. For educational institutions it's much difficult task to stay record of each single student and predict the location of student manually. to beat these challenges, concept of machine learning and various algorithms are explored to predict the results of class students. For this purpose, training data set is historical data of past students and this is often wont to train the model. This software package predicts placement status in 5 categories viz dream company, core company, mass recruiter, not eligible and not inquisitive about placements. this method is additionally helpful to weaker students. Institutions can provide extra care towards weaker students in order that they'll improve their performance. By use Naïve Bayes algorithm all the information are going to be monitor and appropriate decision are going to be provided.

## TECHNOLOGIES AND METHODS

### PYTHON

**History of Python:**

Guido van Rossum developed Python in the early nineties at the National Research Institute for Mathematics and Computer Science, Netherlands. Python is a derivation of many other languages, namely ABC, Modula-3, C, C++, Algol68, Smalltalk, and Unix shell and few other scripting languages. Python has been copyrighted immediately after its establishment. Python source code is open to all with the GNU General Public License (GPL). A dedicated technical team works for the development of Python while the major decisions regarding the technology are still taken by Mr Guido van Rossum.

**Input as CSV File:**

Data Science often requires the action of reading and taking inputs from CSV(Comma Separated Values) files as a prerequisite to several tasks. Usually, the inputs from which come in from varied sources are converted into CSV files. To make it easier, Python offers a library called the Pandas library which renders different features to read the CSV files according to the developer's need. The CSV file is a text file, similar to a Microsoft Excel sheet, the only difference is that the values in the columns are separated by a comma in CSV files.

For example, suppose the data is present in the file named inputfile.csv. This file can be created using windows notepad by just copying and pasting the data and saving it with an extension ".csv"

```
import pandas as pd
dt_data= pd.read_csv('path/inputfile.csv')
print(dt_data)
```

**Operations using NumPy:**

NumPy (Numerical Python) is a python library that provides high level functions to operate mathematically on arrays. It consists of multidimensional array objects and a compilation of methods for processing array.

Using NumPy, the following operations can be performed −

Mathematical and logical operations.

Fourier transforms and routines, which are used for shape manipulation.

Operations related to linear algebra. NumPy also has preprocessed functions for linear algebra and random number generation.

**Key Features of Pandas:**

- Fast and efficient DataFrame object with efficient indexing methods
- Different formats of the inputs can be stored into the memory using the library tools
- Alignment of data and integrated handling of data that is missing.
- Data sets can be reshaped and pivoted.
- Label-based slicing
- large data sets can be efficiently indexed and subsetting can be used.
- Insertion and deletion of Columns from a data structure.
- Group by data used for aggregation and transformations.
- High-performance merging and joining of data.

- Time Series functionality.

**NAÏVE BAYES ALGORITHM**

Naïve Bayes classifier is an algorithm that is based on the Bayes theorem. It has a substantial independence assumption. It is also called "independent feature model". It presupposes the presence or absence of a particular feature of a class is irrelevant to the presence or absence of any other feature in a supposed class. Naïve Bayes classifier can be trained in a supervised learning model too. It applies a method of highest similarity. It has worked in a complicated real-world problem. It needs only a small quantity of training data. It determines parameters for classification. The variance of the variable needs to be calculated for each class and not the entire matrix. Naïve bayes is mainly used when the inputs are high. It gives output in a better sophisticated form. Every attribute's probability is displayed in the Prediction table. Machine learning and data mining methods are based on naïve bayes classification.

**Bayes theorem**:

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)}$$

- Where P(H|X ) is the posterior probability of H conditioned on X
- P(X|H) is the posterior probability of X conditioned on H
- P(H)is the prior probability of H
- P(X) is the prior probability of X

**Types of Naive Bayes Model**

There are three types of Naive Bayes Model:

***Gaussian Naive Bayes classifier***

In Gaussian Naive Bayes, continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution. A Gaussian distribution is also called Normal distribution When plotted, it gives a bell shaped curve which is symmetric about the mean of the feature values as shown below:

Now, we look at an implementation of Gaussian Naive Bayes classifier using scikit-learn.

| | Yes | No | P(Yes) | P(No) |
|---|---|---|---|---|
| **Sunny** | 3 | 2 | 3/9 | 2/5 |
| **Rainy** | 4 | 0 | 4/9 | 0/5 |
| **Overcast** | 2 | 3 | 2/9 | 3/5 |
| **Total** | 9 | 5 | 100% | 100% |

Python

```
# load the iris dataset
from sklearn.datasets import load_iris
iris = load_iris()

# store the feature matrix (X) and response vector (y)
X = iris.data
y = iris.target

# splitting X and y into training and testing sets
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, random_state=1)

# training the model on training set
from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()
gnb.fit(X_train, y_train)

# making predictions on the testing set
y_pred = gnb.predict(X_test)

# comparing actual response values (y_test) with predicted response values (y_pred)
from sklearn import metrics
print("Gaussian Naive Bayes model accuracy(in %):", metrics.accuracy_score(y_test, y_pred)*100)
```

**Output:**

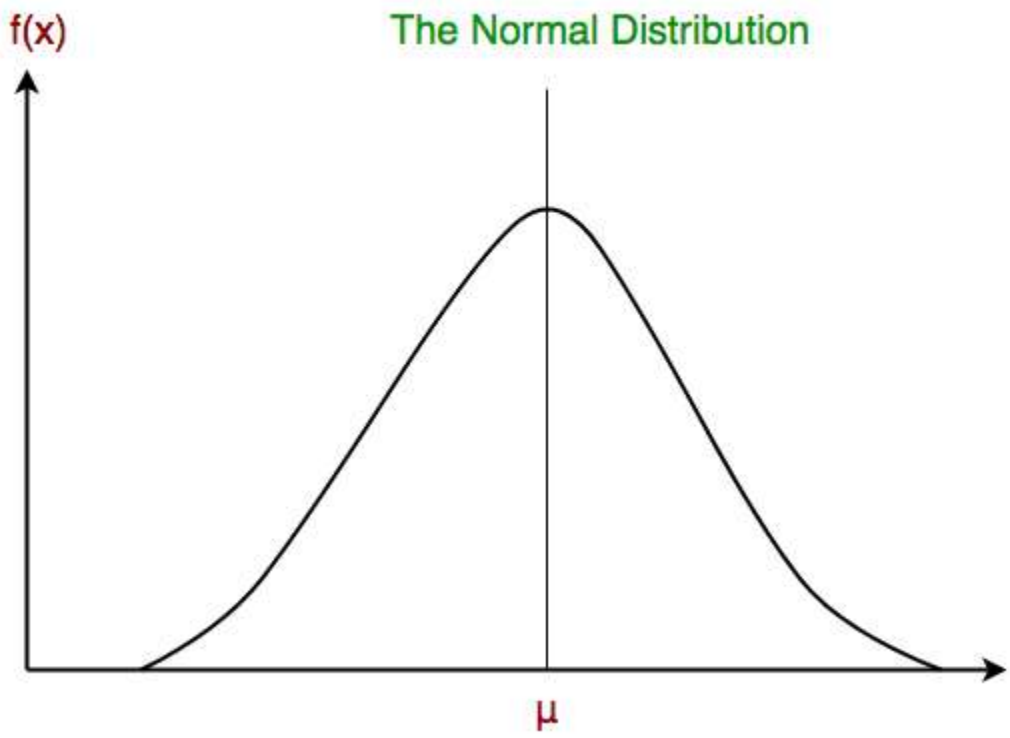
```
Gaussian Naive Bayes model accuracy(in %): 95.0
```

**Multinomial Naive Bayes**

Feature vectors represent the frequencies with which certain events have been generated by a multinomial distribution. This is the event model typically used for document classification.

**Bernoulli Naive Bayes**

In the multivariate Bernoulli event model, features are independent booleans (binary variables) describing inputs. Like the multinomial model, this model is popular for document classification tasks, where binary term occurrence(i.e. a word occurs in a document or not) features are used rather than term frequencies(i.e. frequency of a word in the document).

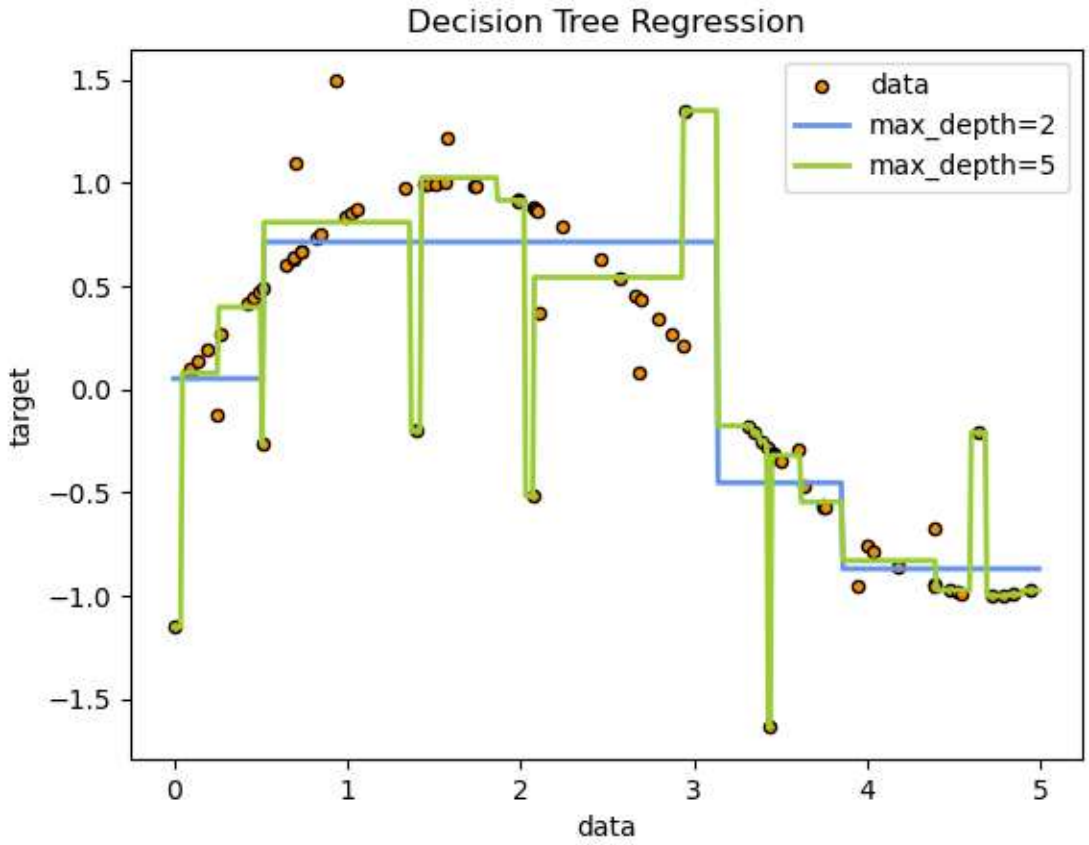



**DECISION TREE CLASSIFIER**

The Decision Tree Classifier algorithm comes under Supervised Learning. The decision tree algorithm can also be used for solving regression and classification problems which none of the other supervised learning algorithms can do. The purpose of using a Decision Tree as a classifier is to build a training model to predict the target variable's class/value by acquiring simple decision tree rules deduced from earlier data(training data).

**Decision Trees (DTs)** are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

For instance, in the example below, decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model.




To predict a class label of a record, the values of the root attribute are compared with the attribute of the record. Based on this comparison, we understand the branch corresponding to that value and move to the next node.

## CONCLUSION

Looking at the statistics of the last few years, the number of student dropout from the educational institute is rapidly increasing. It affects the educational institutions besides the future of the student hence posing as a major threat to all educational institutions. The reasons for dropping out may vary. It depends on various factors that lead to predicting the student who may drop out. In this project, we concentrate on the reason behind the student's drop out. So, data collection plays a vital role in this paper. The collected data is evaluated by several techniques under data preprocessing methods. The data which is collected by various resources contains many determinants like Academics, Demographical factors, Psychological factors, Health issues etc. which play important roles in a student's decision to drop out. This prediction system will help to predict the student who will be choosing to drop out of the registered course. Identifying them at an early stage will prevent the student to drop out and the institution can monitor and give valuable counselling to them in persuading to change the mind of the student from dropping out. It will also help the student to know the right path to achieve their dreams.

**REFERENCES**

1. K. Bonneau and D. Management, "Brief 3: What is a Dropout?" pp. 14–17, 2007.
2. A. Pradeep, S. Das, and J. J. Kizhekkethottam,"Students dropout factor prediction using EDM techniques," Proc. IEEE Int. Conf. Soft- Computing Netw. Secur. ICSNS 2015, 2015. A. Cheah et al., "Analyzing Students Records to Identify Patterns of Students' Performance," pp. 544–547.
3. K. Chai, H. T. Hn, and H. L. Cheiu, 'Naive-Bayes Classification Algorithm', Bayesian Online Classif. Text Classif. Filter., pp. 97–104, 2002.
4. S. S. Panwar, 'of Computer © I a E M E Data Reduction Techniques To Analyze Nsl-Kdd Dataset', pp. 21–31, 2014.
5. V. Hegde and S. G. Kini, 'Multivariate and Multi-Behavioral Student Dropout Prediction Using Naïve Bayesian Algorithm'.
6. A. Cheah et al., "Analyzing Students Records to Identify Patterns of Students' Performance," pp. 544–547.
7. V. Hegde, "Dimensionality Reduction Technique for Developing Undergraduate Student Dropout Model using Principal Component Analysis through R Package," pp. 1–6, 2016.
8. M. Nasiri, B. Minaei, and F. Vafaei, "Predicting GPA and academic dismissal in LMS using educational data mining: A case mining," 3rd Int. Conf. eLearning eTeaching, ICeLeT 2012, no. Dm, pp. 53–58, 2012.
9. B. Guo, R. Zhang, G. Xu, C. Shi, and L. Yang, "Predicting students performance in educational data mining," Int. Symp. Educ. Technol. ISET 2015, pp. 125–128, 2016.
10. M. G. M. Mohan, S. K. Augustin, and V. S. K. Roshni, 'A BigData approach for classification and prediction of student result using MapReduce', 2015 IEEE Recent Adv. Intell. Comput. Syst. RAICS 2015, no. December, pp. 145–150, 2016.
11. The article on Higher education student Dropout Rate https://indianexpress.com/article/explained/in-higher-education-dropout-rates-decline-in-last-five-years-6261594/#:~:text=The%20dropout%20rate%20in%20the,from%207.49%25%20to%202.82%25.
12. "Cause Analysis of students' dropout rate in higher education study program", by Liga Paura and Irina Arhipova, Latvia University of Agriculture, Faculty of Information Technologies, Liela street 2, Jelgava, LV 3001, Latvia from 2nd World Conference on Business, Economics and management in the journal supported by Science Direct, Procedia – Social and Behavioural Sciences 109 (2014) 1282 – 1286.