IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 5, May 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205103 |

Student Database Graduation Prediction Using Machine Learning

Pallavi R E[1], Pooja L K[2], Shreesha[3], Ruthvi R Chawla[4], Kiran B[5]

UG Student, Department of CSE, ATME College of Engineering, Mysuru, India[1-4]

Assistant Professor, Department of CSE, ATME College of Engineering, Mysuru, India[5]

ABSTRACT: this project is to improve prediction techniques regarding the future performance of students in select university courses through the utilization of multiple logistic regressions.

This is achieved with the aid of statistical computing software which applies forward step-wise variable selection methods that identify influential variables sufficient to accurately predict student success.

Once a logit model is constructed with the required parameters and predictors, the inverse logit function outputs a probability of student success.

In all cases, logistic prediction models matched or exceeded the performance of current prediction methods while using an equal or lesser number of explanatory variables.

These findings show that current prediction methods can improve by using a statistically justified procedure. It also suggests the inefficacy of some predictors used to currently estimate student performance.

I. INTRODUCTION

The specific problem is to identify data about students and access it and also their study achievements that is their regular grades and their behavioral characteristics stored in the university information system should be taken in order to predict if the student can graduate on time or not and what measures should be taken can be inferred from the risk score thus generated by the predictive model. The predictions are useful at the beginning of each semester to help students with planning their workload in the whole semester.

And also these predictions can help instructors to plan how and what amount of support each student need to graduate on time with better results than before.

The goal of the project is thus to predict students risk score with the major emphasis on the detection of students who can fail to meet the course requirements.

Therefore, we are dealing with the tasks of prediction ofstudent's success or failure and prediction of the students risk scores.

This project describes a predictive analytics framework to identify such students, describes features that are useful for this task, applies several classification algorithms, and evaluates those using metrics important to educational organizations.

This project also focuses on students who are at risk of not finishing graduation on time, but the framework lays a foundation for future work on other adverse academic outcomes.

II. EXISTING SYSTEM

- Identifying at-risk students often requires manual review and intervention by academic advisors, which can be time-consuming and may lack consistency.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205103 |

- The existing system may not fully leverage the wealth of data available within educational institutions. Factors beyond academic performance, such as extracurricular activities and demographic information, are often underutilized.
- Educational institutions typically utilize basic statistical measures and historical performance data to gauge student progress.

Disadvantages of Existing System

The model was able to predict only the final grades of all the students but could not identify or predict those students who are at highest probability of not graduating on time and show bad results. The models obtain grades with the error of only one degree in the grade scale for all the rest of courses.

III. METHODOLOGY AND DISCUSSION

This paper presents a student graduation prediction model using the ensemble method. An ensemble method is a type of learning method that integrates multiple models to solve a problem. Unlike traditional learning methods that use a single model to train data, ensemble methods use a collection of models to train data and combine the results. Ensembles generally provide more accurate predictions than individual models. The purpose of this approach is to enable an accurate assessment of the factors that influence a student's graduation time prediction. The steps of the proposed methodology are shown in Figure.

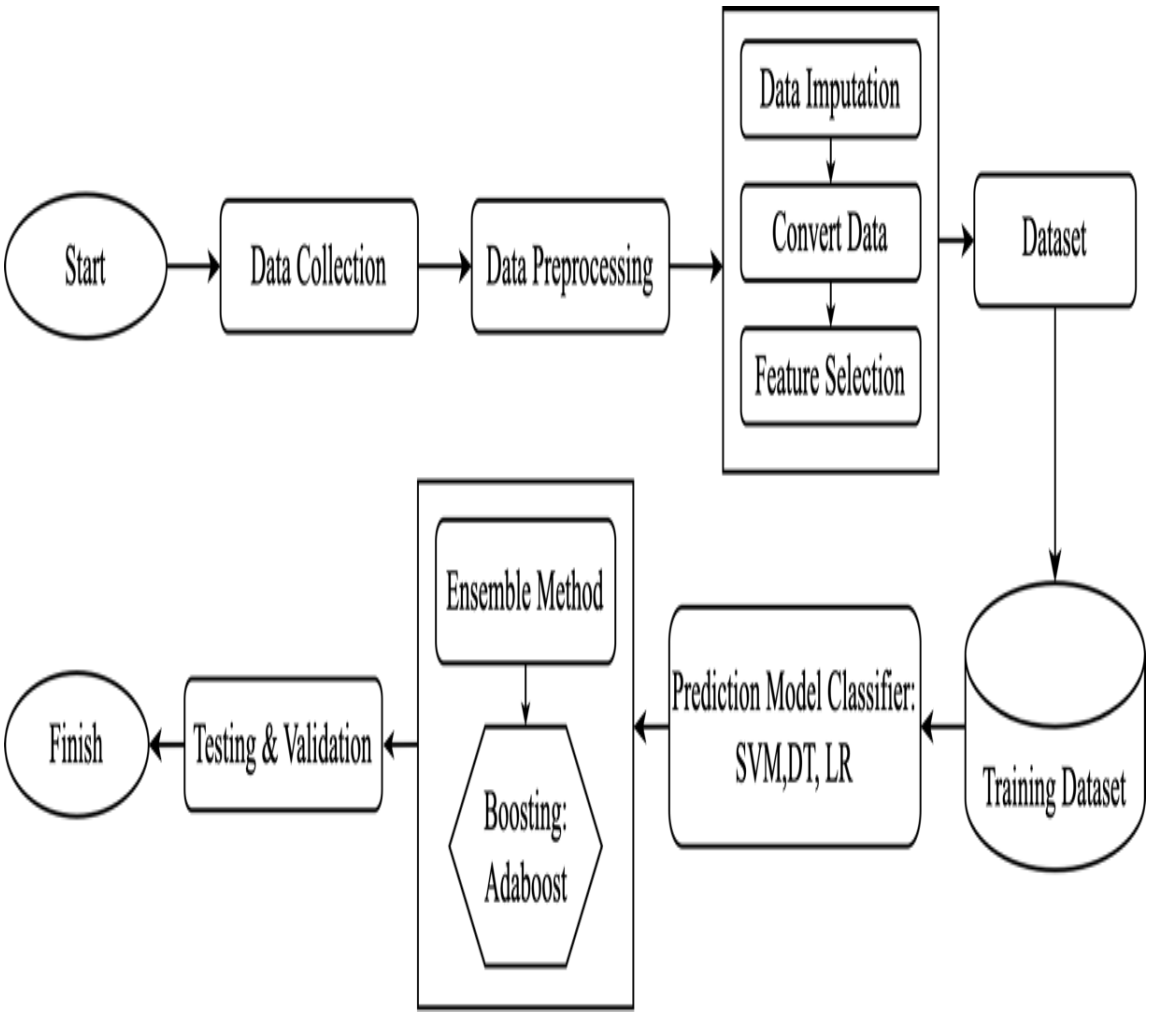


FIG: Student's Graduation Prediction Model Research Steps

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205103 |

Modules and their Description

There are three modules in this system

Logistic Regression

- Data Preprocessing
- Model Training and Evaluation
- Prediction and Deployment
- Manage projects
- Share secret key and messages(encode)
- Logout
- Mutual Information Statistic
- Mutual information measures the dependency between variables by quantifying the amount of information obtained about one variable through the other.
- In this project, mutual information statistic is utilized to determine the strength of association between each feature and the target variable (graduation status).
- Features with higher mutual information scores are considered more relevant for predicting graduation and are selected for model training.
- Chi-Squared
- The chi-squared test is a statistical method used to determine the association between categorical variables.
- In this project, the chi-squared test is applied to assess the independence between each feature and the target variable (graduation status).
- Features with higher chi-squared statistics and lower p-values indicate a significant association with graduation and are selected for model training.

Implementation

Data Preprocessing:

- **Data cleaning:** Handling missing values, outliers, and inconsistencies in the dataset.
- **Feature selection:** Identifying the most relevant features that significantly impact graduation prediction.
- **Feature encoding:** Converting categorical variables into numerical format using techniques like one-hot encoding.
- **Data normalization:** Scaling numerical features to ensure uniformity across different scales.

➢ **Logistic Regression Model:**

Logistic regression is a popular classification algorithm used for binary outcome prediction.

It models the probability of a binary outcome (in this case, graduation or not) based on one or more independent variables.

The logistic regression model estimates coefficients for each feature, which represent the impact of that feature on the probability of graduation.

The logistic function (sigmoid function) is used to map the output of the linear combination of features to a probability score between 0 and 1.

➢ **Model Training and Evaluation:**

The dataset is split into training and testing sets (e.g., 70% training, 30% testing) to train and evaluate the model.

The logistic regression model is trained on the training set using gradient descent or other optimization techniques to minimize the loss function.

Model performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC).

Cross-validation techniques like k-fold cross-validation may be employed to ensure robustness of the model.

➢ **Prediction and Deployment:**

Once the model is trained and evaluated, it can be used to predict the likelihood of graduation for new students.

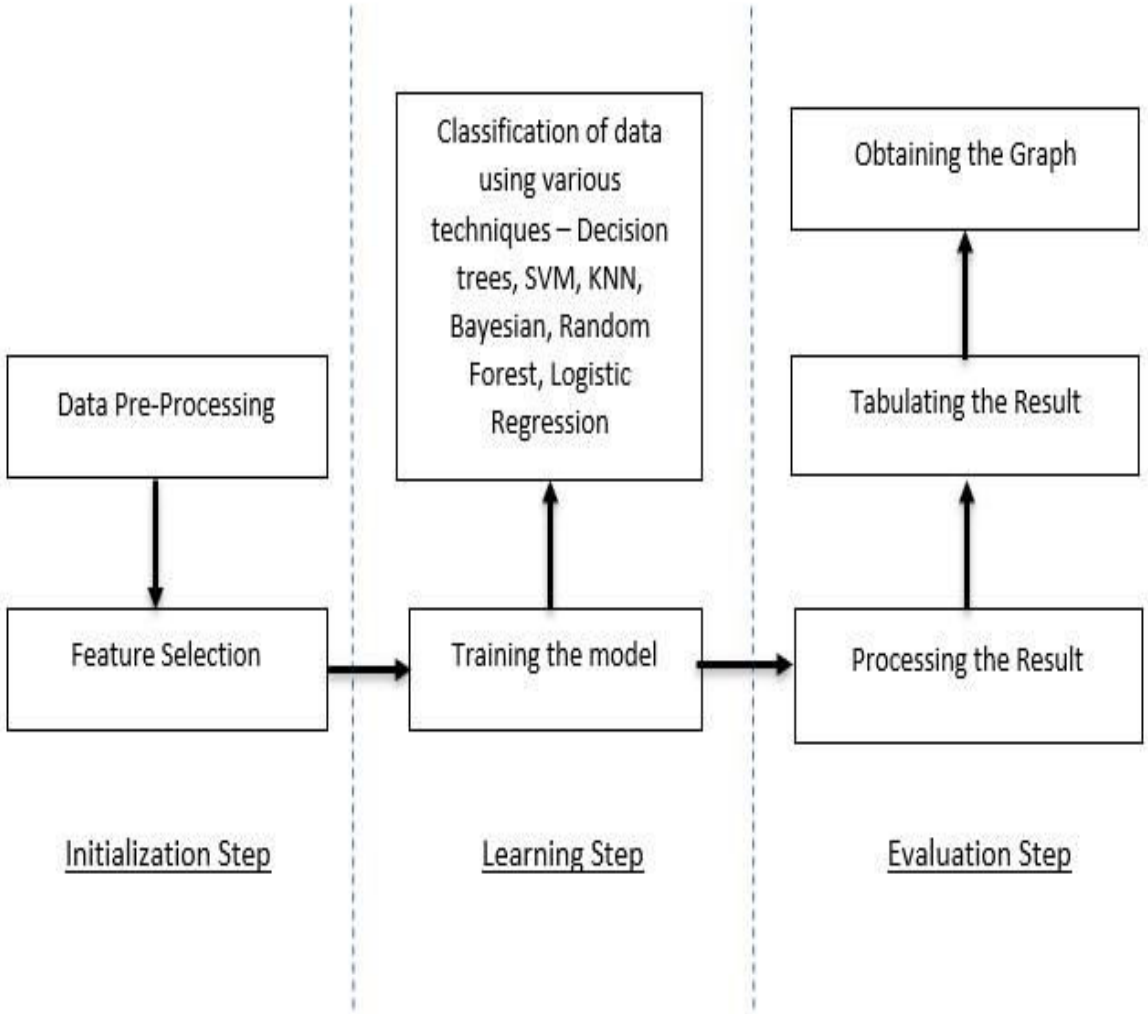
International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205103 |

System Architecture



IV. RESULTS

Results of Features Correlation

TABLE 2. Features Correlation

Features	Graduation Time
Parent's Income	679485.684630
CGPA 3rd Semester	10.338820
Educational Track	8.293161
CGPA_5th Semester	5.728623
CGPA 4th Semester	5.666952
CGPA_2nd Semester	4.094044

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205103 |

Based on the correlation of the 6 features used in Table 2 above, The variable with the highest correlation to student graduation is parent's income. Students with low parent's income can motivate themselves to graduate on time because if they don't graduate on time, they will have to extend their time in college, incurring additional expenses. Other variables include CGPA 2nd Semester, CGPA 3rd Semester, CGPA 4th Semester, and CGPA 5th Semester. This is based on the common practice of new students who typically aim for a high GPA in the early semesters of their studies.

V. CONCLUSION

In recent years, the field of data analytics is growing quickly, driven by intense market demand for systems that tolerate the intense requirements of data, as well as people who have the skills needed for manipulating data queries and translating results.

Thus, it has changed the way we look at data and how we draw conclusions from analysis to obtain future predictions for current events.

More and more advancements occurring in the near future changes the whole perception of data and its analysis to prediction dramatically.

Our lives will get much easier. Performing such a project in this field of computer science will help us gain exposure and get ready for the future of behavioral data analytics.

REFERENCES

1. CHAO-YING JOANNE PENG KUK LIDA LEE GARY M. INGERSOLL – "An Introduction to Logistic Regression Analysis and Reporting "
2. Vijayalakshmi Sampath, Andrew Flagel, Carolina Figueroa – "A LOGISTIC REGRESSION MODEL TO PREDICT
3. Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. CoRR, abs/1106.181
4. Vladimir Cherkassky and Yunqian Ma. Practical selection of svm parameters and noise estimation for svm regression. Neural networks, 17(1):113–126, 2004.
5. S. Colby and J. Ortman. Projections of the size and composition of the u.s. population: 2014 to 2060, Mar 2015.
6. Tom Fawcett. An introduction to roc analysis. Pattern Recognition Letters, 27(8):861 – 874, 2006. ROC Analysis in Pattern Recognition.
7. Isabelle Guyon and Andr´e Elisseeff. An introduction to variable and feature selection. J. Mach. Learn. Res., 3:1157–1182, March 2003.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details