A Real-time Web Application using Machine Learning for Predicting University Dropouts

[1]Sushmitha K, [2]Dr. Shivamurthy R C

[1]Student, [2]Professor and Head of the Department.
Computer Science & Engineering,
Maharaja Institute of Technology
Mysore, India.

Abstract- **In order to prevent dropouts from occurring, it is necessary to identify pupils who are likely to stop attending school or a college. This field of study is crucial because it enables educational institutions to help difficult students and potentially increase the likelihood that they will graduate from college. Dropout rates from engineering colleges place a significant strain on a nation's citizens' ability to pursue higher education and build successful careers. The prosperity of a nation rests,on its capacity to produce graduates from higher education who can advance the nation. Various institutions are looking to artificial intelligence's (AI) potential to anticipate dropout as early as feasible to help solve the dropout problem. Because it can be challenging to obtain personal information and does so with privacy concerns, schools must be dependent on the theoretical statistics of their students to build precise and dependable extrapolative algorithms. The goal of this effort is to construct the finest prediction model based just on the academic information; therefore it must have the greatest ability to infer knowledge.**

Comprehensions and forecasting skills to boost academic results is very important for any student to achieve greater heights. On relevant datasets, machine learning algorithms can be used to find patterns, trends, and important variables that influence the achievement of students. Many students61.8% are High level. 38.2% are in the low level. In total, there are 36.5% female students and 63.5% male pupils. Just 5% of female students are in low level classes. So, the academic growth of female students is higher than that of male students. In comparison, female students engage significantly more in educational events than their male counterparts. According to our observation, this means that female students perform better than male pupils. In this project, machine learning is being used to forecast the student's performance. Decision Tree Algorithm (DT), Random Forest Classifier (RFC), and Logistic Regression Algorithm (LR) constitute the model. Here, we developed a web application that will analyze user input to figure out if a student will drop out of university or not.

Index Terms - **Dropout, Machine Learning, Flask, Web Application.**

I. INTRODUCTION

The performance of the student is mostly what decides whether a student drops out of college. Student performance is the collective academic success and advancement of students in their studies. It includes a variety of factors, including grades, test results, involvement in class projects, and overall subject mastery. Student performance offers information about their skills, areas of strength, and areas that need growth, making it an important gauge of their development as students.

Teachers, caregivers, and educational organizations can assess student performance in order to assess whether or not a certain teaching tactics, course of study, or student support system is functioning. It provides targeted measures which enhance students' learning experiences by allowing the recognition of areas where students may be underperforming. In addition, the achievement of a pupil is a significant consideration when assessing future opportunities, such as doorway to universities, financial aid, and job prospects.

The components influencing the achievement of a pupil might be many and varied. Personal characteristics including inspiration, routines for studying, preferences for learning, and previous experience may be among them. Academic achievement is additionally affected by outside factors such as socioeconomic status, availability of resources, educational materials efficiency, and the educational environment.

The combination of individualized instruction, differentiated learning tactics, frequent assessments, mechanisms for feedback, and supportive learning settings is frequently used to increase student performance. Educators and institutions work to maximize student potential and support their academic success by attending to individual requirements and offering appropriate advice and facilities. Realizing the significance of student proficiency, universities around the world are always working to improve learning outcomes and guarantee that all students have equitable access to excellent educational opportunities. By concentrating on pupil achievement, both educators and stakeholders hope to deliver children with the talents and acquaintance needed to excel in the classroom and be successful in their goals for the future.

II. RELATED WORKS

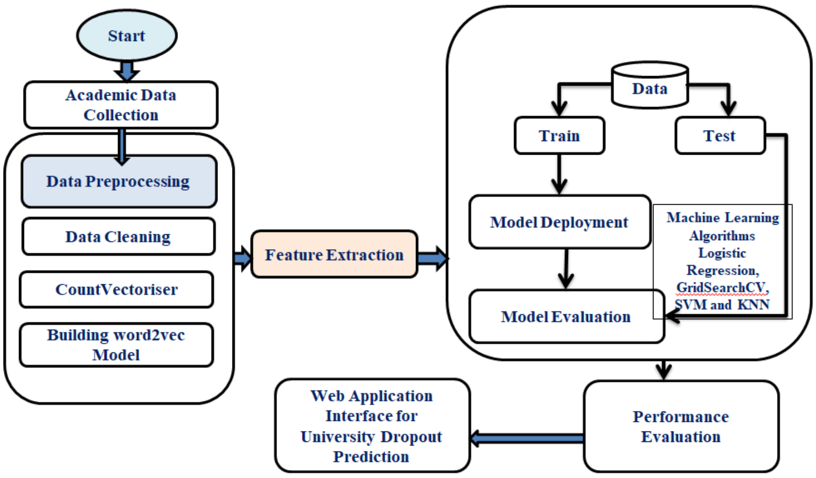
One of the most crucial aspects of virtual learning is the recognition of at-risk pupils. *Yan Chen et al.* have explained that Students exhibit diverse online learning behaviors at various points in a semester. As a consequence, we propose a progressive forecasting approach for recognizing vulnerable pupils during the course of a semester. They examine the distinctive characteristics of every pupil and their distance learning behaviors, identify characteristics that have a strong connection with their educational results, and then indicate combined sets of features based on time frame as well as time cutoff constraint approaches. The trials' findings show that the suggested model's precision varies between 90.4% to 93.6% contingent on the phase. The issue of university dropout rates has numerous unfavorable effects. It has a negative economic impact and is a problem in academia. The primary recognition of students who are going to be dropout from the university has received a lot of attention in recent years. In order to forecast such as dropout of student from 1st year or in 3rd year of studies is the central criteria of research examines data in Europe. We took the five main program areas into account while making this prediction. To find the factors most closely connected to dropout, a Feature Selection Process was employed first. *M Segura et al.* The issue of university dropout rates has numerous unfavorable effects. It has a negative economic impact and is a problem in academia. The primary recognition of students who are going to be dropout from the university has received a lot of attention in recent years. In order to forecast such as dropout of student from 1st year or in 3rd year of studies is the central criteria of research examines data in Europe. We took the five main program areas into account while making this prediction. To find the factors most closely connected to dropout, a Feature Selection Process was employed first. *H Dasi et al* used the academic record of four years is taken into account. The important parameters are selected in identifying course finishers and dropouts. The prediction accurateness varied in the middle of 92% and 93% using unobserved information from the upcoming educational year. The machine learning strategies are used alongside the performance measurements. The outcomes exhibited that a variety of machine learning techniques might be employed effectively to examine little educational records. Search engines may reject the page as a result. Make sure your abstract flows smoothly and adheres to proper grammar. F Del Bonifro et al. defines the school absences are a significant problem, particularly in underdeveloped nations. Understanding the root causes and spotting the symptoms are essential for dealing with this. The objective of this research is to correctly estimate the likelihood that a student will leave school. In order to identify the most significant causes of high dropout rates, I will quantify prediction accuracy and examine several features of the student data. In order to determine the best predictor, they have trained a neural network and deployed a number of classification methods.

III. PROPOSED METHODOLOGY

The aim of this project is to design system to predict student performance based on student credentials like there academics enactment as shown in figure 1. In the discipline of teaching, methods of learning can be supervised as well as unsupervised when it involves predicting dropping out of school. The fundamental concept of supervised learning involves gaining knowledge from a set of labeled examples in the training set to identify un-labeled examples in the test set as correctly as possible. In order to better retain students, educational institutions may find that utilizing machine learning to predict student dropout is a useful tool for spotting at-risk students early on and offering the right kind of support. Presented here is a machine learning-based dropout prediction system. In this perception, used Logistic Regression, GridSearchCV, Support Vector Machine and K Nearest Neighbor.

Gather relevant data about students, their academic performance, engagement, and other socioeconomic factors. The dataset should include features such as attendance records, course grades, demographic information, extracurricular activities, participation in online forums, library usage, and any other information that may impact student performance and engagement. We have applied data cleaning process for the collected data to handle missing values, outliers, and safeguard the data reliability. Feature engineering might also be essential to extract useful information from raw data and create meaningful predictors.

Figure 1: Proposed System of Student Dropout Prediction



We have divided the dataset into training dataset and testing dataset distinctly. Use the training data to train the selected machine learning models with hyperparameter tuning techniques, like Grid Search Cross Validation (GSCV) technique, to optimize model performance. Evaluate the trained model(s) using the testing dataset to measure its predictive dropout performance. Cross-validation techniques can also be employed to get a better estimate of the model's generalization ability. In educational settings, it's crucial to provide explanations for model predictions. Integrate the trained model into the institution's existing system for student monitoring and intervention. This can be in the form of a web application or API that takes in student information and returns the likelihood of dropout. The system should not only predict dropout but also provide actionable insights for intervention. Based on the prediction results, educational institutions can implement targeted support systems, counseling services, or mentoring programs to assist at-risk students and improve retention rates.

IV. SYSTEM IMPLEMENTATION

The purposed System Implementation developed for student dropout prediction. The implementation can classified into different modules of project and are listed as Data Collection, Data Preprocessing, Feature Selection / Extraction, Model Training and Testing, Machine Learning Model Deployment, Performance Evaluation, Comparison Study of The Model and some Exploratory Data Analysis (EDA) on dataset.

1) *Dataset:* In this project we have used the dataset based on the *student records* as shown in the figure 2. Many researchers had also used this dataset. This dataset is being provided by the UC Irvine Machine Learning Repository and it is available on the UCI website. This dataset contains *362 student information* and *39* attributes with 1 target attribute. The targ*et attribute has labelled in three*-class to represent *Low, High or Medium*.

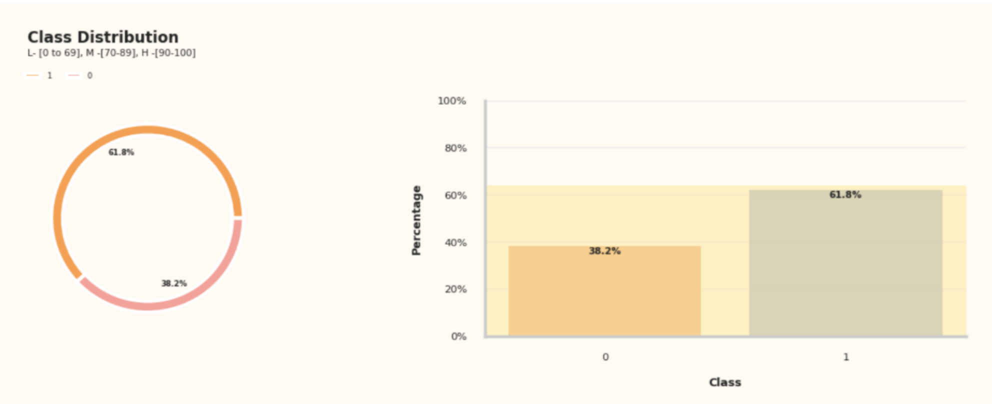
Figure 2: Student Dataset

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
	Name	Gender	Department	Year	SEM 1 SGP	SEM 2 SGP	SEM 3 SGP	SEM 4 SGP	SEM 5 SGP	SEM 6 SGP	SEM 7 SGP	SEM 8 SGP	How many	How many	How much	How much	Internet av	If yes, wha	What is yo	Can you si	D
0	Mayuresh Pitale	Male	Computer Eng	Second Year	7.93	7.93	7.38	7.3	7.4	7.6	7.9	7.2	0	1	1	3	1	0	1	0	
1	Vijeta Nayak	Female	Computer Eng	Fourth Year	7.75	7.78	7.68	8.93	8.33	8.5	8	9	0	0	2	2	1	0	1	0	
2	Ojas Nandkumar Thale	Male	Information T	Second Year	6.15	6.09	6	7.1	6.1	6.3	6.7	6.7	1	1	3	2	1	0	3	1	
3	Anush Madathumkara	Male	Computer Eng	First Year	5.75	6.54	5.9	5.98	6	6.2	6.3	6.8	0	0	3	1	1	0	1	0	
4	Parag Iyyani	Male	Computer Eng	Second Year	5.67	6.81	6.6	6.5	6.7	6.2	5.8	5.9	0	0	2	3	1	0	1	0	
5	Akshay Chakkungal	Male	Computer Eng	Second Year	6.1	6.1	6.5	5.9	5.4	5.2	6.1	6.2	1	0	3	3	1	0	1	1	
6	Shradha	Female	Information T	Second Year	6.2	6.3	5.9	5.8	6.5	6.2	5.7	5.9	1	0	1	4	1	0	1	0	
7	TEJAS AVINASH PATIL	Male	Information T	Second Year	6.51	7.42	6.73	7	7.1	7.2	7.6	7.2	0	0	3	3	1	0	1	1	
8	Prem Thamarakshan	Male	Information T	Second Year	6	6.15	6.2	5.9	6.8	6.4	6.2	6.7	0	0	2	4	1	0	1	0	
9	Kaustubh Nitin Deshmukh	Male	Mechanical E	Second Year	7.2	7.1	6.43	7.3	7.4	7.9	7.4	6.9	0	1	2	5	1	0	2	0	
10	Neha Mahesh Patil	Female	Information T	Second Year	8	7.85	8.54	8.1	8.3	7.9	7	7.5	0	0	3	2	1	0	3	0	
11	Sonali patil	Female	Information T	Second Year	6.7	6.67	6.69	6.9	6.8	6.4	6.2	6.1	0	0	3	2	1	0	1	1	
12	Shraddha pandey	Female	Information T	Second Year	6.5	6.1	5	6.2	6.7	6.4	6.8	6.2	0	0	1	5	1	0	1	0	
13	Gitanjali Singh	Female	Information T	Second Year	7.41	8	8.54	8.1	8.2	8.4	7.9	7.8	0	0	2	1	1	0	1	1	
14	CHIRAG BHOIR	Male	Mechanical E	Second Year	8.57	8.22	8.08	8.8	8.5	8.4	8.6	8.1	0	0	3	1	1	0	3	0	
15	Avdhoot Mane	Male	Information T	Second Year	5.14	5.2	5.6	5.9	5.7	5.8	5.9	5.7	0	0	2	3	1	0	1	0	
16	Prachi Satbhaya	Female	Information T	Second Year	5.6	5.7	5.2	5.7	5.1	5.2	5.6	5.9	0	0	1	2	1	0	3	0	
17	Saad shaikh	Male	Information T	Second Year	5.5	5.2	5.85	5.9	5.7	5.6	5.3	5.7	0	0	2	3	1	0	1	0	
18	Prathamesh ganesh patil	Male	Automobile E	First Year	5.1	5.2	5.6	5.4	5.3	5.4	5.2	5.4	0	0	3	3	1	0	1	1	
19	Mansi Nimje	Female	Information T	Second Year	7.74	7.5	7.73	7.8	7.9	7.4	7.8	8	0	0	3	2	1	0	1	0	
20	Nikita patil	Female	Electronics Er	First Year	5.8	5.2	5.3	5.9	5.8	5.9	5.7	6	0	1	2	1	1	0	3	1	
21	Krutika Parvatikar	Female	Information T	Second Year	7.3	7.2	6.77	6.9	6.8	7.2	7.1	7.3	0	0	1	2	1	0	1	0	
22	Nayan Shelke	Male	Mechanical E	Second Year	9.48	9.69	8.89	8.9	8.7	9.2	9.1	9.4	1	0	3	1	1	0	2	0	
23	Suraj vijay Patil	Male	Mechanical E	Second Year	6.12	7.82	6.58	7.1	7	6.9	6.8	7.2	0	0	3	3	0	0	1	0	
24	Kajal Rajendra Borhade	Female	Information T	Second Year	5.6	5.8	5.8	6.2	6.1	6.3	6.4	6.2	0	0	3	5	1	0	1	1	
25	Karan nipurte	Male	Computer Eng	First Year	5.9	5.7	5.6	5.8	5.9	6.1	6.7	6.4	0	1	3	5	0	0	2	1	
26	Akshay Puranjay Shetty	Male	Information T	Second Year	6.89	7.6	7.9	7.8	7.5	7.6	7.4	7.9	0	1	2	1	1	0	1	0	
27	Pooja zagade	Female	Information T	Second Year	7.89	7.37	7.65	7.6	7.9	7.7	7.4	8	0	0	3	1	1	0	1	1	
28	Mustafa Shaikh	Male	Information T	Second Year	7.11	7.44	6.08	7.1	7.2	7.4	7.1	6.9	0	0	2	4	1	0	1	0	

2) Feature Selection: Feature selection is needed for trained each machine learning classifier because without removing unnecessary attributes from the dataset result may be affected. The classifier algorithm with feature selection gives better performance and reduces the execution time of the model. For this process, three different feature selection methods were used in this research.

3) Machine Learning Algorithms: Machine Learning Classification algorithms are deployed and performed student dropout prediction and provided insight evaluation model values on the dataset. The data analysis also performs to view all the insight of the dataset. The Figure 3 shows the class distribution such high, low & medium with the ratio of how many student dropout or not.

Figure 3: Class distribution



V. EXPERIMENTAL RESULT

The experiment is carried on student datasets; here we have used machine leaning algorithms for our study as prediction algorithms. The predictions are shown by using the web applications by giving the input to the machine learning implemented models. Here we input the values for both apps and get the results.

The implementation of student dropout prediction is as shown in Figure 5. Here we need to provide the inputs such as all semester CGPA, Assignment, travel, studies, internet, speed, mode trans, lectures_2, submissions, lectures_5, practicals_5, coaching classes and social skill.

Figure 4: Home Page of the Application

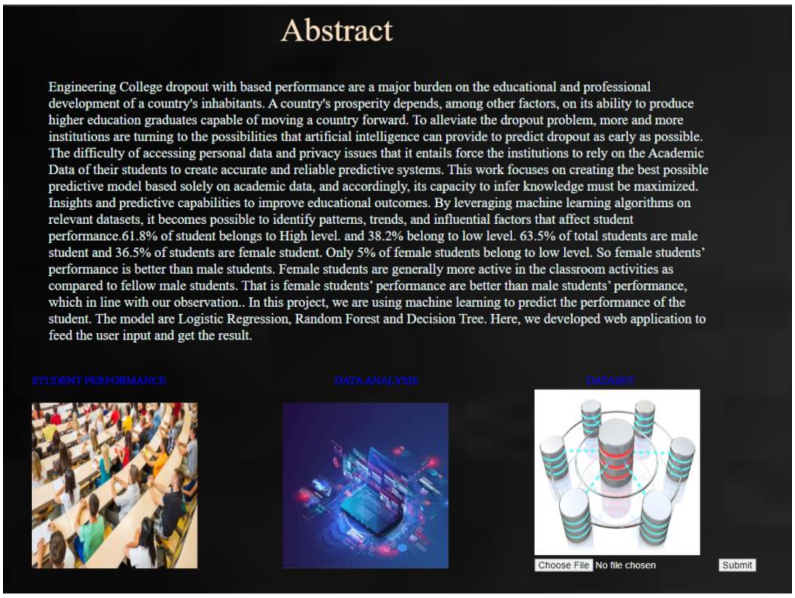


Figure 5: User Input to the Student Dropout Application

The figure 5 shows user input to Prediction Model. Figure 6 shows the result of the prediction model. The algorithm accuracy can be found by using various parameters such as Precision, Recall, F1 score and Accuracy. The experimental values as show as in Table 1.

Figure 6: Prediction Result

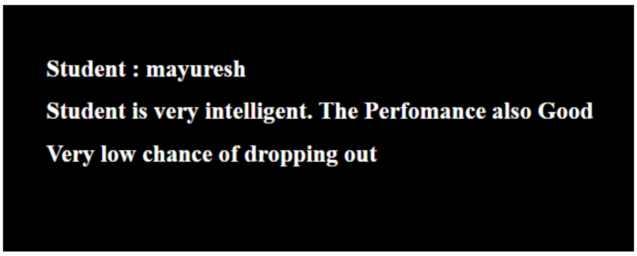


Table 1: Model Accuracy

Machine Algorithm	Accuracy
Random Forest	100%
Logistic Regression	100%
Decision Tree	98.5%

VI. CONCLUSION

By applying machine learning approaches to analyze student performance, educational results might be improved by way of the useful insights and prediction powers these techniques provide. Finding patterns, trends, and significant variables that influence student performance can be done by utilizing machine learning algorithms on pertinent datasets.38.2% of students are in low level, compared to 61.8% of students who are in high level. Men make up 63.5% of all pupils, while women make up 36.5 %. The low level is only represented by 5% of female students. As a result, female students perform better than male students. Compared to their male classmates, female students are typically more engaged in classroom activities. The performance of female students is therefore superior to that of male students, which is consistent with our data. To anticipate the student's performance in this project, we have used the core algorithms on ML. Logistic Regression, Random Forest Algorithm, and Decision Tree make up the model. We created this web application to receive user input and output the results and show the different level of severity among the students.

REFERENCES:

1. Y. Chen, Q. Zheng, S. Ji, F. Tian, H. Zhu, and M. Liu, ``Identifying at-risk students based on the phased prediction model,'' *Knowl. Inf. Syst.*, vol. 62, no. 3, pp. 987_1003, Mar. 2020.
2. H. A. Mengash, ``Using data mining techniques to predict Student performance to support decision making in university admission systems,'' *IEEE Access*, vol. 8, pp. 55462_55470, 2020.
3. J. Fernandez-Garcia, R. Rodriguez-Echeverria, J. C. Preciado, J. M. C. Manzano, and F. Sanchez-Figueroa, ``Creating a recommender system to support higher education students in the subject enrolment decision,'' *IEEE Access*, vol. 8, pp. 189069_189088, 2020.
4. S. Theodoridis, ``Learning in parametric modeling: Basic concepts and directions,'' in *Machine Learning*, 2nd ed. S. Theodoridis, Ed. New York, NY, USA: Academic, 2020, pp. 67_120.
5. *Pandas-Dev/Pandas: Pandas*, The Pandas Development Team, Road Dallas, TX, USA, Feb. 2020.
6. T. Sasada, Z. Liu, T. Baba, K. Hatano, and Y. Kimura, ``A resampling method for imbalanced datasets considering noise and overlap,'' *Proc. Comput. Sci.*, vol. 176, pp. 420_429, Jan. 2020.
7. A. J. Fernández-García, L. Iribarne, A. Corral, J. Criado, and J. Z. Wang, ``A recommender system for component-based applications using machine learning techniques,'' *Knowl.-Based Syst.*, vol. 164, pp. 68_84, Jan. 2019.
8. F. Pérez, J. Aldás, R. Aragón, and I. Zaera, Eds., *Indicadores Sintéticos de las Universidades Espanolas*. San Luis Potosí, Mexico: Fundación BBVA, 2019.
9. J. Shana, and T. Venkatacalam, "A framework for dynamic Faculty Support System to analyse student course data", International Journal of Emerging Technology and Advanced Engineering, Vol. 2, No. 7, 2012, pp.478-482.
10. Quality Assurance Division (2010). InMinds: Intelligent Mining and Decision Support System. University of Malaysia Sarawak, UNIMAS [Online]. Available:
11. http://www.qad.unimas.my/Function/ICTCompliance/inminds.html [4]. SPA (2013). What is SPA Standard? SPA Student Performance Analyser [Online]. Available
12. V. Kumar, and A. Chadha, "An empirical study of the applications of data mining techniques in higher education", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 2, No..3, 2011, pp. 80-84.