IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

FACTORS INFLUENCING STUDENTS DROPOUT: A MACHINE LEARNING BASED STUDY

V. Lavanya , Dr.K. Santhi Sree

M.Tech Student, Professor of Computer Science and engineering
Computer Science and Engineering
School of Information Technology JNTUH, Hyderabad, India

Abstract: Dropout rates for students are increasing. We need to make certain measurements to find out the factors which are leading to the dropout of the student by using machine learning techniques with which we can choose an alternate way to select the dropout rates of the student from the course. We can make an analysis of each and every student by identifying the most related features. There are many factors which are responsible for the dropouts of the students like financial, economic, and many other problems. If the dropout rate of the students is increasing then the employment opportunities of the students will decrease gradually. Here we are going to implement some of the Machine Learning techniques to identify the students who are likely to drop out. Our main task is to apply different machine learning techniques to the selected features. A comparison is made for machine learning algorithms with each metric, and the model that gives the more appropriate result is considered the best one.

Keywords: Dropouts, Identification, Features, Machine Learning Techniques.

I. INTRODUCTION

Machine learning scrutinizes the study and disposition of algorithms that may learn from the data and frame predictions on the available data. Machine learning algorithms are determined by building a model from example inputs so as to create knowledge-driven forecasts or choices, instead of following strictly static program instructions. Machine learning techniques will automatically detect the change by the users and start to flag them without manually telling them to do so. We use machine learning to solve problems that are very complex. There are different types of machine learning systems here in this work we are going to use supervised machine learning techniques.

Being a high school graduate can open up the doors to several rewarding careers within the past generations. Higher education helps you to research your area of interest with which you will be able to reach your goals and implement your ideas efficiently.

The Covid pandemic situation has drastically changed the education system of the world. Almost all of the schools and other learning centres have moved to the online education system which has drastically changed the lifestyle of the student population which is affecting the students in the low and lower-middle-income countries. A large number of schools have moved to online mode in order to continue the education of the students. This made them face many challenges not only to the students, teachers, and also to their family members. Almost all of the schools have moved to an online education system.

Countries' productivity mainly lies in the education levels of the students. The main goal here is to identify the students who are likely to drop out and solve their problems in order to continue their studies. The task to identify the dropout rates is not an easy one. We should consider many factors of a particular student like age, gender, family support, area of interest, educational background of the family, academic record, and many others. There are many techniques involved in dropout prediction such as data gathering and pre-processing, attribute selection, classification of the data. More generally, various reasons for school dropouts, viz. failure in academics, unavailability of schools, inaccessibility of schools, harsh teaching environment, financial problems to name a few can be classified into some broad categories like school-centric, student-centric. Students start disengaging as an outcome of these reasons, and if identified well prior, dropouts can be prevented by timely actions. In the past decades, multiple studies have been undertaken to ascertain the main reasons behind dropouts in India, but none focused on identifying the students on the verge of dropping out during the course of their education. The procedure will be as follows:

1. We collect data regarding students such as academic records, financial status, family education details of the students which are the critical factors controlling dropout rates.
2. We identify the most dominant features that lead to students dropping out.
3. We apply various supervised machine learning algorithms to predict at-risk students.
4. Apply different metrics for each tested algorithm.

Data collection is a challenging step in this study as schools, a lot of times, especially in rural areas they do not maintain proper records and also because they do not make data available to the public easily. However, the Unified District Information System for Education, an initiative by the National University of Educational Planning and Administration, started student tracking in the year 2016. The system will keep a record of the academic journey of students studying in about 1.5 million government and private schools in India. The data capture will continue every year now and will move towards tracking using aadhar card numbers in the future.

II. METHODOLOGY

In the proposed methodology we are using extensive and descriptive viewpoints and include a wide scope of approaches that are officially recognized as possessing certain qualifications or meeting certain standards. It also involves practical assessment to construct validity including structural aspects.

2.1 Data Acquisition

Data used for predicting the students' dropout rate is being collected from various online sources. The implementation of the system is done by gathering information about students from two different schools in Uttar Pradesh. We have gathered the data regarding students' information, their academic performance, and their relationship with their family members which mainly influence the students to continue their higher education. It consists of the details of the 10th studying students of two different schools.

2.2 Data Cleanup

The data collected may contain some missing values, duplicate rows, and some unrelated data. The errors can be handled by:

- Deleting the duplicate rows from the dataset.
- If a row exists with fewer features having null or wrong values then the mean of the values is taken and it is filled in those columns.

2.3 Data Pre-processing

The process is to identify the required independent variables for predicting the dropout of students' and to predict the binary dependent variable 'PASSED' using the independent variables in data pre-processing. The dataset here we are using for predicting students dropout contains information about their academic performance and their relationship with their family members which mainly influence the students to continue their higher education.

To predict the dropout of students, the dataset is split into training and testing. At any point in time splitting has an 80% training rate and a 20% testing rate. The predicted value will be 1 if the student is likely to drop out of school for higher education and 0 if continued.

2.4 Feature Analysis

In our dataset, there will be many features for a particular student. All of these may not be needed to predict the correct result, so we are going for the feature selection. Here we are going to analyze each feature and only the features which are mostly affecting the result are going to be considered.

Feature analysis has been performed on the data to improve the following aspects:

- **Reduces Overfitting**: It removes the unnecessary data to make decisions i.e. if the data is having any noise it will be removed and the decision is made in accordance with it.
- **Improves Accuracy**: The data provided should not mislead by which accuracy of the model increases.
- **Reduces Training Time**: If we are providing less amount of data the algorithm will be able to train faster.

Feature analysis has been resulted in selecting the following features:

- Gender
- Age
- Family size
- Mother education
- Father education
- Mother job
- Father job
- Travel time
- Study time
- Failures
- Health
- Absences

III. IMPLEMENTING ALGORITHMS

The main purpose of deploying our model is that we can make the predictions from a trained machine learning model available to others. Model deployment refers to the arrangement and interaction of software components within a system to achieve a predefined goal.

The algorithms we are using for this task are K Nearest neighbour, Logistic regression, Decision tree, Random forest, Bernoulli naive Bayes.

3.1 K-Nearest Neighbour

In the K-Nearest Neighbour algorithm, the predictions for the new data points are made by looking at their neighbours. K-Nearest Neighbour has no learning process i.e, initially, we need not train the algorithm; it will automatically learn and make predictions based its neighbours. We will be calling it a Lazy learning algorithm. It is one of the simplest machine learning algorithms we have ever used. It tries to estimate the distribution of y given x, and classify given observation to the class with the highest estimated probability. First thing is to plot the training dataset and then locate the new test instance. Calculate distance from all train data points and then sort the distance list in ascending order from that list choose the first 'k' distances from the sorted list. The next step is to calculate the test instance distance from all nearest distances. Here we are considering a classification problem so for the new instance we need to assign the mode.

Here we have created two models by changing the weights of the algorithms. The value of 'k' range is set between 6 to 30. The main goal is to improve the metrics. To create the elbow curve we will train the model with different values of 'n' neighbours and finally pick the value which gives the least error. We need to create a function that defines the elbow curve and create the empty list in order to save the error generated for each value of 'k'. To this function we input 'k' and for every value, in 'k' we create an instance in KNN and using that value as 'k' neighbours we train the model to make the predictions on the dataset and we calculate the metrics and finally, we calculate the error by subtracting the metric from 1 since higher the score would be the better model. The model is created for every alternative value.

3.2 Logistic Regression

It is a classification algorithm. In this algorithm, we are going to draw a line and measure whether the line will be fitting into the data or not. Whenever we are going to add the new data items a new point comes which will be changing the model line to a curve it is called a sigmoid curve which will classify the data items which will be ranging from 0 and 1.

3.3 Decision Tree

The Decision tree is a supervised machine learning technique used for producing a decision tree from training data. A decision tree is the most relevant model which is a mapping from observations about an item to conclusions about its target value. It is a simple yet powerful way of knowledge representation. A decision tree is a flow-chart like structure, where each internal node is denoted by rectangles, and leaf nodes are denoted by ovals. All internal nodes have two or more child nodes that contain splits, that test the value of an expression of the attributes.

The decision tree classifier has two phases:

- Growing phase
- Shrinking phase

In the growing phase, the tree is built by continuously splitting the training dataset by applying certain criteria until all of the specified criteria is met to label the data. The tree may overfit the data. The shrinking phase handles the problem of overfitting the data in the decision tree. The accuracy of the classifier increases in the shrinking phase. In the shrinking phase, we can access only the fully grown tree. The growth phase requires multiple moves over the training data. The time needed for shrinking the decision tree is very less compared to building the decision tree.

3.4 Random Forest

Random forest is one of the machine learning algorithms which works on the concept of bagging. The working of the algorithm is as follows. From the given dataset multiple bootstrap samples depend upon the number of models we want to build now. On each and every bootstrap sample we build a decision tree model. Each decision tree model is built on a different subset of data, now each tree thus formed are combined to predict the final output. So effectively we are combining multiple trees to get the final output and hence it is called a forest but why we are calling it a random forest means we use random bootstrap samples whether it is partially correct along with a random sampling of data points or rows of random forest performs random sampling of features. So every tree is generated with a different set of data points sampling of the data is being carried out at node level for every tree.

3.5 Bernoulli Naive Bayes

Bernoulli Naive Bayes come under the category of a classification algorithm. The main concept of Bernoulli Naive Bayes is that it accepts features only as binary values like true or false, yes or no, success or failure, 0 or 1 and so on. So, when there is a binary classification problem we can go for Bernoulli naive Bayes. The decision rule is P(xi/y)=P(i/y)xi+(1-P(i/y))(1-xi). The results generated by this are very fast as compared with other models. Here we are going to treat each feature as independent of another. This model gives very appropriate data when we are working with smaller datasets. Real-time predictions can be handled easily by considering all the relevant features. By observing the results we can easily understand that the model is self-explanatory.

IV. RESULTS

To assess the factors which are mainly responsible for predicting the dropout rates of the student, the prescribed model should be able to generalize well on all types of data. In the below **Fig.1** we have compared each and every model with its accuracy on the training data, accuracy on testing data, along with them we have a result of the metrics like precision, recall, f1-score and AUC curve. In **Fig.2** we have shown the graphical representation of training data accuracy for each of the algorithms among the applied algorithms Decision tree classifier and random forest classifier has performed best. **Fig.3** shows the testing data accuracy graphical representation on test data Bernoulli naive Bayes algorithm has performed well. We are going to assess the result based on the model's performance for each metric applied. We have applied metrics like accuracy, precision, recall, f1 score, AUC curve. The results generated are in **Fig.4**, **Fig.5**, **Fig.6**, **Fig.7. Figure.8** shows that students whose health condition is not good, who are having a shortage of attendance and number of failures are more than in three subjects are the people who are more likely to drop out of their education.

	Models Name	Models Train Accuracy	Models Test Accuracy	Models Precission	Models Recall	Models f1 score	Models AUC
2	BernoulliNB	0.7230	0.7071	0.728395	0.893939	0.802721	0.613636
6	DecisionTreeClassifier	1.0000	0.6869	0.753623	0.787879	0.770370	0.636364
3	GaussianNB	0.7297	0.6768	0.702381	0.893939	0.786667	0.568182
0	RandomForestClassifier	1.0000	0.6667	0.685393	0.924242	0.787097	0.537879
1	LogisticRegressionCV	0.7331	0.6667	0.681319	0.939394	0.789809	0.530303
5	SVC	0.6723	0.6667	0.666667	1.000000	0.800000	0.500000
4	KNeighborsClassifier	0.7601	0.5859	0.676056	0.727273	0.700730	0.515152

Figure 1 Performance comparison of each model with metrics

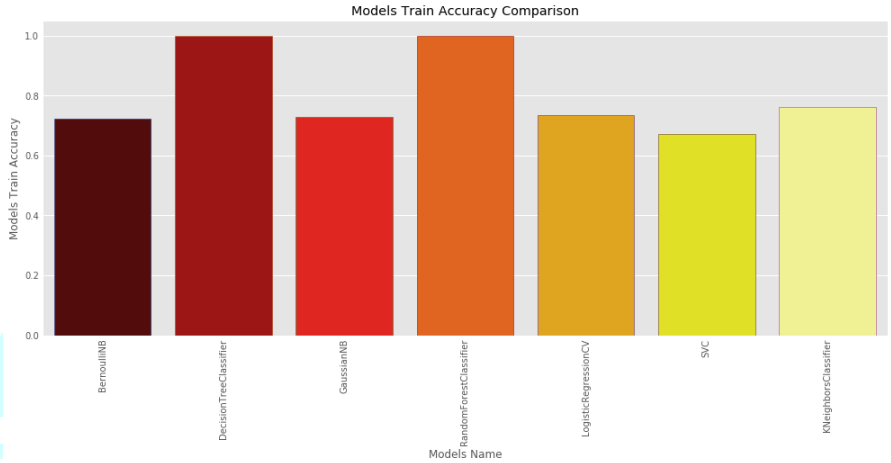

Figure 2 Train accuracy comparison

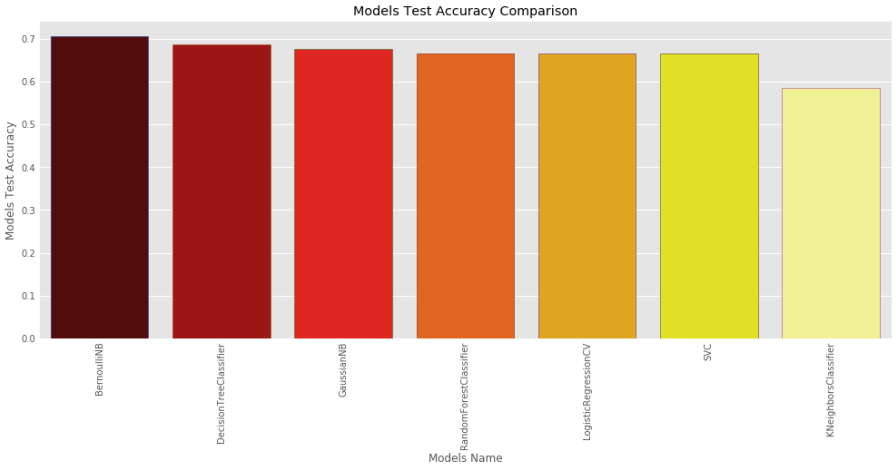

Figure 3 Model test accuracy comparison

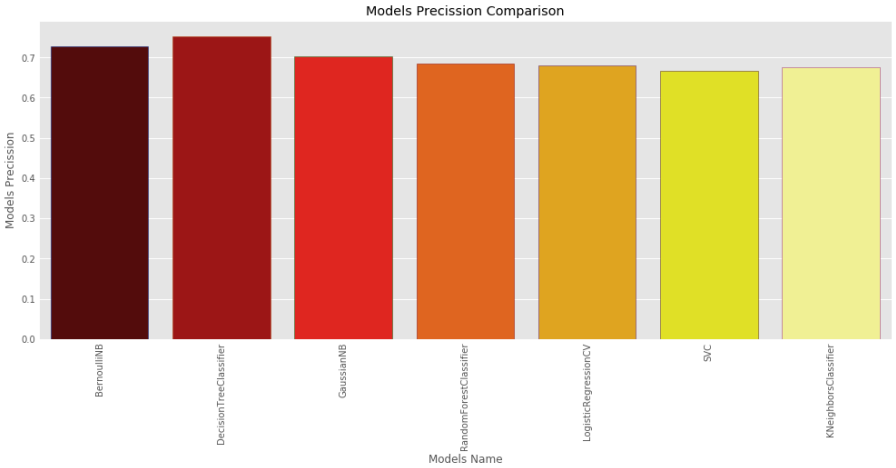

Figure 4 Comparison of all algorithms using precision

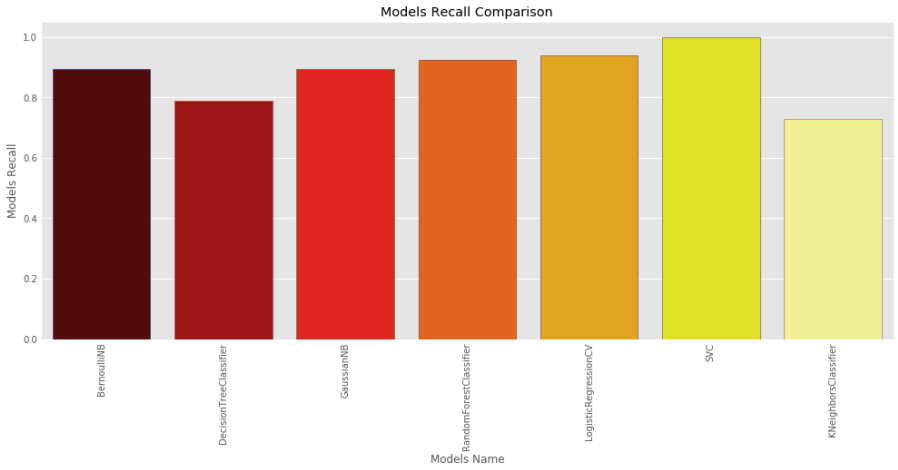

Figure 5 Comparison of all algorithms using recall

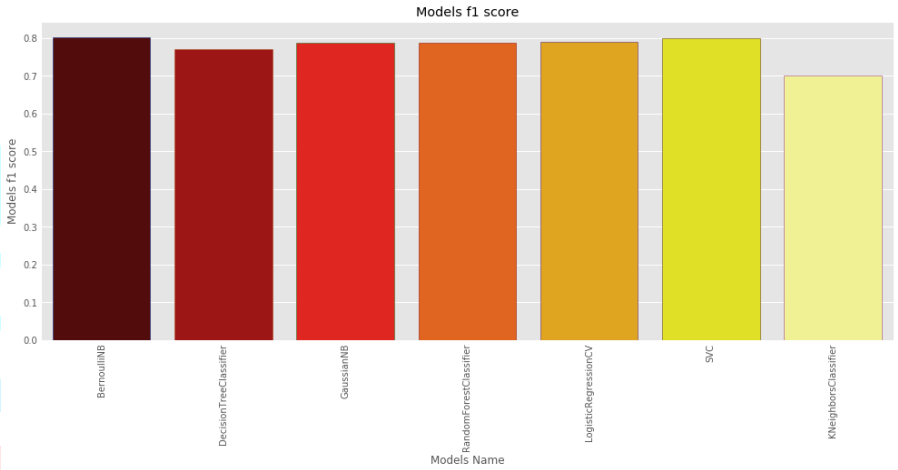

Figure 6 Comparison of all algorithms using F1 score

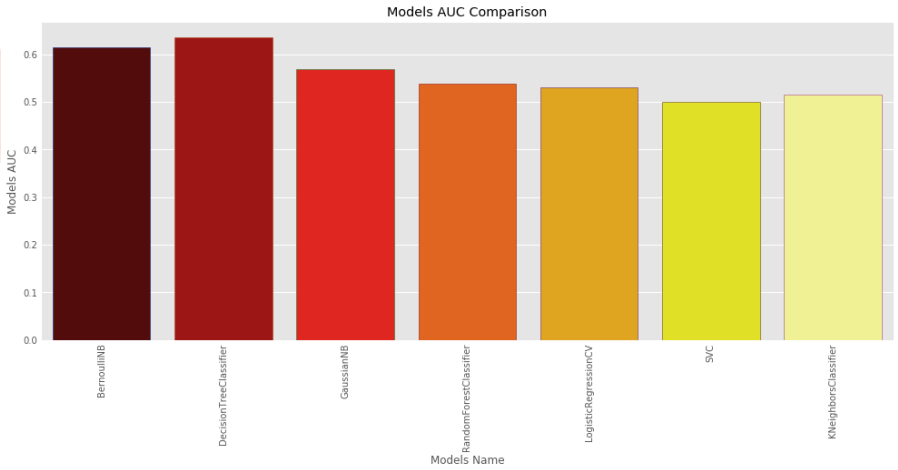

Figure 7 Comparison of algorithms using AUC

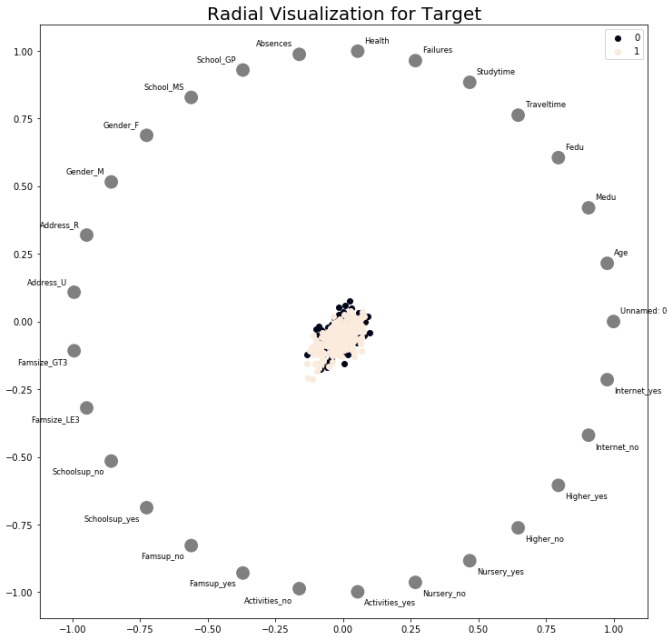

Figure 8 Radial visualization of all features

V. Conclusion

In our study Decision tree classifier is proved to be the best classifier. The results obtained portray that if sufficient data is available in the upcoming years of the data captured by the Unified district information system for education, the Decision tree classifier will provide accurate predictions of at-risk students in advance. Though there are immense students who are attentive in higher education, the primary discernment for dropping out came to be health issues, and attendance is yielded to be the considerable factors causing students dropouts.

REFERENCES

[1] UNESCO, Education For All Monitoring Report 2008, Net Enrollment Rate in primary education.
[2] Revision of the International Standard Classification of Education (ISCED), retrieved 05-04-2012.
[3]Https://www.vistacollege.edu/blog/resources/high-er education-in-the-21st-centuary/
[4] Altaher, A., BaRukab, O.: Prediction of student's academic performance based on adaptive neuro-fuzzy inference. Int. J. Comput. Sci. Netw. Secur. 17(1), 165–169 (2014)
[5] Acharya, A., Sinha, D.: Early prediction of student performance using machine learning techniques. Int. J. Comput. Appl. 107(1), 37–43 (2014)
[6] Diseth, A., Martinsen, O. (2003). Approaches to learning, cognitive style, and motives as predictors of academic achievement. Educational Psychology, 23 (2), 195-207
[7] F. Araque, C. Roldán, and A. Salguero, "Factors influencing university dropout rates," Computer education.
[8] Bharadwaj B.K. and Pal S. "Mining Educational Data to Analyze Students" Performance", International Journal of Advance Computer Science and Applications (IJACSA), Vol. 2, No. 6, pp. 63-69.