A Machine Learning Approach to Identify the Students at the Risk of Dropping Out of Secondary Education in India

Sagarika Nangia[1], Jhilmil Anurag[2], and Ishani Gambhir[3]

[1]Oracle India Pvt. Ltd,Bengaluru
[2]TATA Motors, Pune, Maharashtra
[3]TATA Motors, Lucknow, U.P

***Abstract*— Having a significant number of student dropping studies at higher education limits their possibilities of having better employment opportunities. According to a UNESCO report, India is fifty years behind in achieving the goal of universal secondary education. There is almost a 30% drop in GER(Gross Enrolment Ratio) from secondary education to higher education. In this paper, we use machine learning models to identify the students who are likely to drop out in the ongoing session. We demonstrate, to the best of our knowledge, the first results on a secondary level education in a developing nation like India to classify students as potential dropouts. We collect data from secondary level institutions for the academic session 2016-17, which captures the variation in ethnic, social and financial background of students as well as their academic performances. We do feature analysis to find out the most dominant attributes in the dataset and use them in different machine learning algorithms on the task of binary classification for potential drop out. Our best algorithm is able to classify correctly with least number of false negatives.**

I. Introduction

Education is basic to human development. It not only impacts an individual but is directly related to the growth of the entire nation. India's School Education Vision 2030 aims at transforming the poor quality outcomes of education system into one where all children get high quality education. It expects to see an increase in school enrollments from 25 crore in 2010 to 30 crore in 2030. In the past couple of years India has seen an increase in its school enrolment ratios. However, there has been no decrease in the number of students who dropout every year which is still at a high rate of 17% in the secondary level compared to 4% at primary level. Thus, even after witnessing an increase in its enrolment ratio, India is still far from its goal of achieving universal education. Various reasons like the economic status of the family, lack of self motivation and unforeseen life circumstances can cause students to dropout from school. This impediment of Indian educational system of retaining students leads to high illiteracy rates, unemployment, poor standards of living and slow GDP growth.

More generally, various reasons of school dropouts viz. failure in academics, unavailability of schools, inaccessibility of schools, harsh teaching environment, financial problems to name a few, can be classified in to some broad categories like school-centric, student-centric and parent-centric [1]. Students start disengaging as an outcome of these reasons and if identified well a 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 paper contributes as follows :

1) We collect data such as academic records, financial status, social category, medical records of the students which are the critical factors controlling drop out rates.
2) We identify most dominant features that lead to students dropping out.
3) We test various machine learning algorithms to predict at-risk student during the ongoing session.

Data collection is a challenging step in this study as schools, a lot of times, especially in rural settings, don't maintain proper records and also, because they don't make data available to the public easily. However, UDISE, an initiative by National University of Educational Planning and Administration started student tracking in the year 2016. The system will keep record of academic journey of student 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 number in the future.

II. Literature Review

Student attrition at secondary school level across the world remains a matter of concern, since the level of education is de facto for the growth of a nation and its individuals. Several studies have been conducted so far to predict the students who have a higher probability of dropping out of their educational program, whether part-time, full time or in a distance learning program. The studies are focused on determining the factors, whether demographic or behavioural common to the students dropping out. The reasons for dropping out differ in developing and developed countries. In developed countries causes of dropout are more student-centric like inability to pay their own expenses, not being able to manage school and work together [2],[3] . However in developing countries like India, Pakistan, Bangladesh, factors like poor financial conditions of parents, daily wage responsibilities are more predominant [4]. The studies conducted in the US, and countries such as Greece, Mexico focus on both school

sagarika.nangia16@gmail.com
jhilmilanurag@gmail.com
ishanigambhir.26@gmail.com

and university student dropouts, somewhat indicate GPA as one of the most dominant deciding factor affecting student attrition [5].

Studies have been done to predict student dropout in Distance Learning programs because it allows the university to impart education to a diverse and larger group of students, without the requirement of any physical resources. Also, in European countries and in the US, these studies are used as a method of evaluating the quality of education provided by the institutions [6], [7]. In Greece it has been found that distance learning programs have a greater student attrition rate as compared to conventional learning programs. It experiences a dropout rates of 25-40% , while traditional educational courses have a dropout rate of 10-20% [6], [7]. In India, studies undertaken on student enrolment and retention have challenged universalizing of education and have examined factors leading to dropouts. The major reasons cited by the households for the dropout were children were not interested in studies, cost was too much, children were required for household work and also for work outside to contribute to family income [4]. Also, as expected, the gender differentials were still persisting in school education. The dropout was higher among girls (15 percent) than boys (11 percent) [8]. With regard to rural-urban differences, more girls dropped out in rural areas (17 percent) than in urban areas [4] . A similar study undertaken in the state of Kerala aimed to predict the factors responsible for student dropout using educational data mining techniques. It concluded that children aged below 15, failing in subjects namely Physics, English and Maths have a higher chance of dropping out, hence indicating marks are a pivotal factor for more accurate prediction results [9]. However, no study based on secondary education in India has successfully identified the students at risk of dropping out during the academic session by continuous update of the feature values like examination marks, attendance and class response to identify students at a greater risk of dropping out and suggesting necessary measures to prevent it.

III. Methodology

In this section, we outline our approach to predict the 'at-risk students'.

- **Data Procurement** : This step involves gathering data on secondary school students from U-DISE (Unified District Information System), a database of information about schools in India created by National Institute of Educational Planning and Administration [8].
- **Data Analysis**: Data cleaning, transformation and manipulation is done before feeding the data into the models followed by data visualization to understand patterns and correlations in dataset.
- **Deploying Machine learning**: Different machine learning algorithms are applied on the dataset. The results obtained from each are compared in table II with the highest weighted accuracy (7).

TABLE I
Number of Schools in Bodwad District

Name of cluster	Bodwad	Engaon	Nadgoan	Shelwad	Surwade
No. of Schools	24	12	15	16	8

A. Data Procurement

The final objective of this project is to identify potential dropouts, which requires comprehensive data of all the enrolled students. This comprehensive data containing the demographic details of the students was procured from the U-DISE .The system keeps record of the academic journey of every student studying in about 1.5 million Government and Private schools in India [8] .The dataset used for this study is from block BODWAD of the state of Maharashtra for the classes 10th and 11th. Bodwad is a Taluka in Jalgaon District of Maharashtra State, India with its headquarters in Bodwad town [10]

The dataset has 20000 entries with information captured according to the Student Data Capture Format of U-DISE. (Annexure I - First page of the DCF) [8]. The variables used to design predictive model are as follows:

1) Date of Birth
2) Date of Admission
3) Gender
4) Disadvantaged*
5) Social Category (General, Scheduled Caste, Scheduled Tribe, Other Backward classes) *
6) Religion (Hindu,Muslim, Christian, Sikh) *
7) Below Poverty Line (BPL) *
8) Physical Disability
9) Free Education Recipient *
10) Attendance Records
11) Examination marks

*These variables were partially modified to diversify the students data.

Since U-DISE started collecting student data only in the year 2016-17, attendance records and examination were not available. Hence, random values between 0 and 100% were inserted and dummy data was created. For the purpose of this study, 17% students out of the database were picked based on classifiers identified from literature review with some degree of randomness and marked as output class 'dropped-out' to train our models.

B. Data Analysis

1) Data Cleanup: When a huge amount of data is collected, some errors such as missing information, duplicate rows and wrong attribute values are inevitable. These errors are handled using the following ways:

- The duplicate rows (same feature set) were deleted from the dataset.

- If a row exists with three or less features having null or wrong values, random values specific to these features were assigned. For more than 3 attributes containing null or wrong information, the row was removed from the dataset.

The data cleansing resulted in a dataset of 17359 students.

2) Feature Analysis: For prediction of student attrition, it is important to analyze the features which contribute most to the students dropping their education. This helps in eliminating the noise in the dataset and thereby increasing the efficiency of the models. A new feature 'Age' was developed using the 'Date of Birth' attribute. For each feature in the dataset, it's correlation with the output label was found out as shown in figure 1, 2, 3, 4, 5 and 6.

Religion	Correlation
Hindu	0.21
Christian	0.22
Muslim	0.23
Sikh	0.18

Social category	Correlation
General	0.17
SC	0.28
ST	0.30
OBC	0.30

Fig. 1. Correlation of Religion *(left)* and Social Category (right) with output class.

Gender	Correlation
Male	0.18
Female	0.26

Disability	Correlation
Yes	0.21
No	0.21

Fig. 2. Correlation of Gender (left) and Disability *(right)* with output class.

Disadvantaged	Correlation
Yes	0.27
No	0.21

Free Education	Correlation
Yes	0.21
No	0.39

Fig. 3. Correlation of Disadvantaged *(left)* and Free Education Recipient *(right)* with output class.

Homelessness	Correlation
Yes	0.48
No	0.21

BPL	Correlation
Yes	0.41
No	0.20

Fig. 4. Correlation of Homelessness *(left)* and Below Poverty Line *(right)* with output class.

Attendance	Correlation
Below 35%	0.64
35%-60%	0.24
60%-75%	0.15
Above 75%	0.03

Fig. 5. Correlation of Attendance with output class.

Exam Marks	Correlation
Below 40%	0.56
40%-70%	0.16
Above 70%	0.05

Age	Correlation
14	0.22
15	0.22
16	0.21

Fig. 6. Correlation of Exam Marks *(left)* and Age*(right)* with output class.

It can be seen that the age and disability gave almost same correlation for their respective values with output class, so these features were eliminated.

Data is further analyzed by visualizing in the form of bar graphs and histograms as shown in figures 7, 8, 9, 10, 11, 12, 13, 15 and 14.

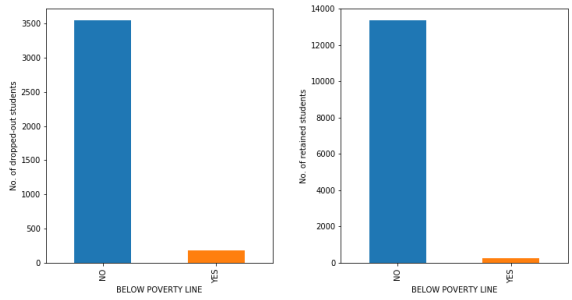

Fig. 7. Comparison of Dropped out and Retained students on the basis of Below Poverty Line.

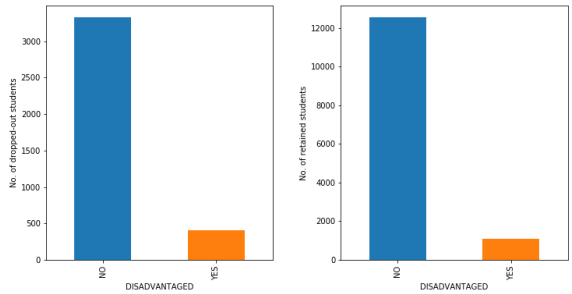

Fig. 8. Comparison of Dropped out and Retained students on the basis of being Disadvantaged.

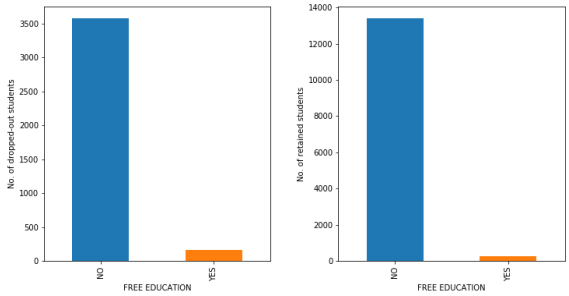

Fig. 9. Comparison of Dropped out and Retained students on the basis of Free Education.

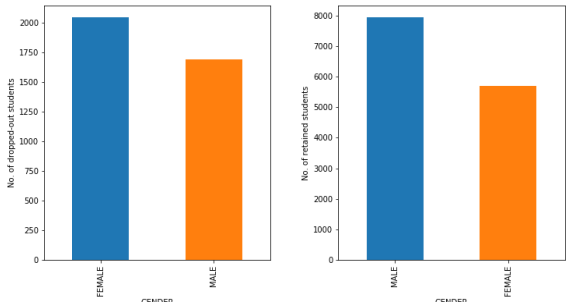

Fig. 10. Comparison of Dropped out and Retained students on the basis of Gender.

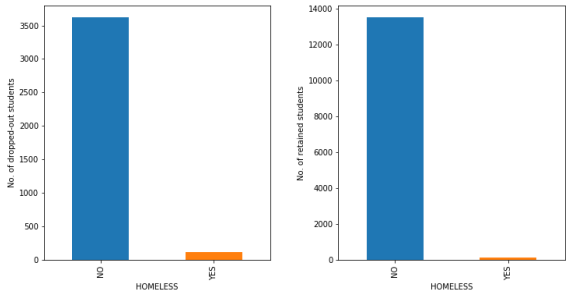

Fig. 11. Comparison of Dropped out and Retained students on the basis of Homelessness.

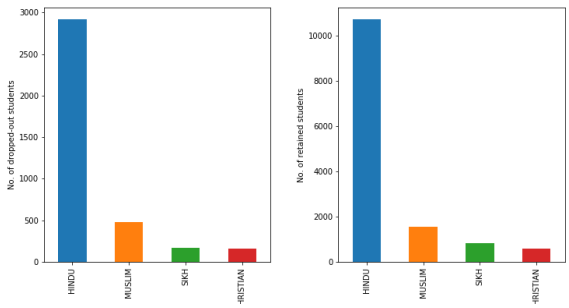

Fig. 12. Comparison of Dropped out and Retained students on the basis of Religion

.

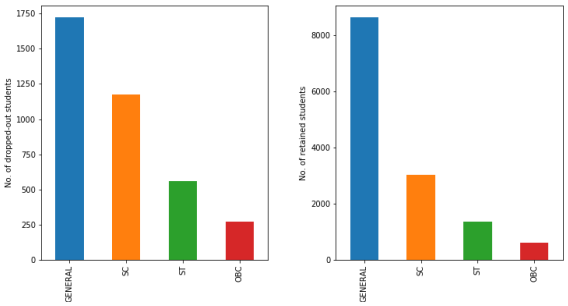

Fig. 13. Comparison of Dropped out and Retained students on the basis of Social Category.

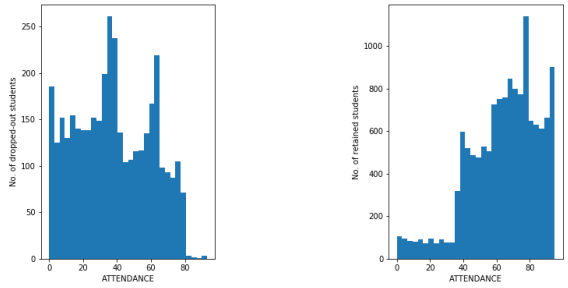

Fig. 14. Comparison of Dropped out and Retained students on the basis of Attendance.

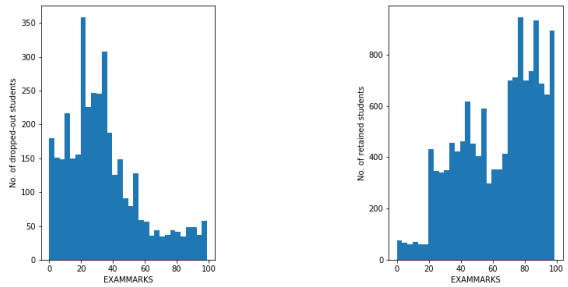

Fig. 15. Comparison of Dropped out and Retained students on the basis of Examination Marks

.

It is evident from the above frequency plots that low attendance and examination marks are the major factors that lead to students dropping out. Besides these, below poverty line, homelessness and female students are at a higher risk of withdrawing from education.

The feature analysis has resulted in selecting 9 features as below:

1) Attendance
2) Examination Marks
3) Gender
4) Homelessness
5) Below Poverty Line
6) Religion

7) Social Category
8) Disadvantaged
9) Free Education Recipient

3) Data Pre-Processing: The data procured is pre processed to make it suitable for training the prediction model. Gender, Social Category and Religion, being categorical in nature were one-hot encoded. Below Poverty Line and Free Education Recipient were mapped to binary 1 and 0 corresponding to YES and NO. Z-Score normalization was used for Attendance and Examination marks to bring all feature values to same scale. This standardization of input values ensures faster convergence of model as equal weight is given to all the features [11].

The Z-Score normalization is done using following expression:

$$x_{new} = \frac{x - \mu}{\sigma} \tag{1}$$

C. Deploying Machine Learning Algorithms

The supervised machine learning algorithms have been deployed for the binary classification of students into the categories of likely to 'drop-out' or retain. Supervised algorithms work by learning the mapping function of input variables to output variables . This means that the algorithm learns the relationship between input and output variables and when the new data comes in it iteratively predicts the output till the error is reduced to an acceptable limit [12]. The dataset acquired from U-DISE is randomly divided into training and testing set in 3:2 ratio. The algorithms used for the above task are Logistic Regression, K-Nearest Neighbours, Support Vector Machine, Neural Networks and Gradient Boost classifier.

1) K-Nearest Neighbor (K-NN): In k-NN classification, a student in the testing set is classified by calculating it's k (a predefined positive integer) nearest neighbours from the training data. The neighbors are identified by measuring the distance between the testing sample and training data. The output class of testing sample is then decided by taking votes from the k nearest neighbors. For instance, If k = 1, then the student is simply assigned to the class of that single nearest neighbor [13].

2) Logistic Regression: It is a statistical model which predicts the output label by calculating the probability that a given student belongs to a certain class. This probability is calculated using the sigmoid function [14].

Let X, β be training set and feature weight vector respectively. For binary classification, probability of ith student can be estimated as:

$$P(y_i = 1|X, \beta) = \frac{1}{1 + e^{-\beta^T X}} \tag{2}$$

$$P(y_i = 0|X, \beta) = 1 - P(y_i = 1|X, \beta) \tag{3}$$

The training is done by adjusting the feature weights using Stochastic Average Gradient Descent which converges faster than the conventional Stochastic gradient descent as it incorporates the previous gradient descent values [15].

3) Support Vector Machine: It performs classification by finding a hyperplane that maximizes the margin between the two classes. The nearest vectors to this hyperplane are called support vectors [16]. This is done by choosing a suitable kernel which transforms the data into required form and provides the most optimal results [17]. In this study, linear kernel is chosen after doing 10-fold cross validation on the entire dataset.

4) Decision Trees and Gradient Boost: A Decision Trees classifies by forming a graphical tree structure, where at each node a decision is made recursively until a leaf node is encountered. [18]. Leaves represent class names while other nodes represent attributes of the dataset with a branch for each possible outcome. During the testing phase, we begin at the root and at each node take the branch according to the feature set. On arriving at leaf, the student is classified as dropped out or retained. Decision trees are known to handle both numeric and categorical data without the need of one hot encoding or label encoding.

Let A be an attribute that can be branched into $A_1, A_2, ...A_v$ nodes. If P and N are two classes, then information impurity (entropy) of node A is calculated using the following:

$$I(p, n) = -\frac{p}{p+n} log_2 \frac{p}{p+n} - \frac{n}{p+n} log_2 \frac{n}{p+n} \tag{4}$$

where:

$p =$ No. of entries belonging to class P
$n =$ No. of entries belonging to class N

Let p_i and n_i be the number of entries belonging to ith value of attribute A in class P and N respectively. The entropy can be then used to calculate the information gain.

$$Gain = I(p, n) - \sum_{i=1}^{i=v} \frac{p_i + n_i}{p+n} I(p_i, n_i) \tag{5}$$

At every level of tree, the information gain is calculated for each attribute and the one with the highest attribute is chosen to be the child note until the node becomes entirely pure [18].

Gradient Boost is a greedy algorithm that works by minimizing the loss of the model by adding weak classifiers to the model [19]. Decision trees are used as weak classifiers and they are added sequentially while keeping the existing classifiers in the ensemble frozen. The loss function is minimized using gradient descent while adding the trees.

5) Neural Networks: It is a model in machine learning that tries to mimic the functionality of human brain [11]. The neural networks are organized into layers where each layer consists of one or more neurons. The pattern is presented at the input layer which gets forwarded to the hidden layers where the processing of data takes place. For binary classification problem, the output layer consists of one neuron that provides the appropriate class label using an activation function. The error is calculated and using propagation, the parameters at each layer are adjusted. This way, multiple

batches of training data are sent to the network and after some finite number of epochs the network gets trained [11]. Non-linear approximators can be easily hypothesized using neural networks.

IV. Experiments and Results

Generally, the model which is able to generalize well on all types of data provide the most accurate results is considered to be the best model. However, when skewed classes are present such as in our case, the model tends to become biased towards the most frequent class giving over-optimistic results [20]. A better metric for performance evaluation is balanced accuracy [20].

Actual \ Predicted	−	+	Total
−	TN	FP	N
+	FN	TP	P
Total	N′	P′	Testing Set

Confusion Matrix

Based on the confusion matrix above,

$$Balanced\,Accuracy = \frac{1}{2}(\frac{TP}{P} + \frac{TN}{N}) \tag{6}$$

Since, a student who is likely to drop out classified as 'retained' is worse than vice-versa, the agenda is to select the model that provides sufficiently low false negatives with reasonable number of false positives. We use weighted accuracy in which the cost of True Positive Rate is higher than that of True Negative Rate.

$$Weighted\,Accuracy = 0.7(\frac{TP}{P}) + 0.3(\frac{TN}{N}) \tag{7}$$

1) KNN: We have calculated the neighbors using Euclidean Distance between the feature vectors of testing sample and the training data.

$$Euclidean\,Distance = \sqrt{\sum_{i=1}^{i=n}(x_i - y_i)^2} \tag{8}$$

where:

x_i = ith attribute of student A
y_i = ith attribute of student B
n = number of attributes of each student

The model gave the best weighted accuracy for k=7.

Actual \ Predicted	−	+	Total
−	5157	319	5476
+	653	815	1468
Total	5810	1134	6944

KNN Confusion Matrix

$$Weight\,Accuracy = 67\% \tag{9}$$

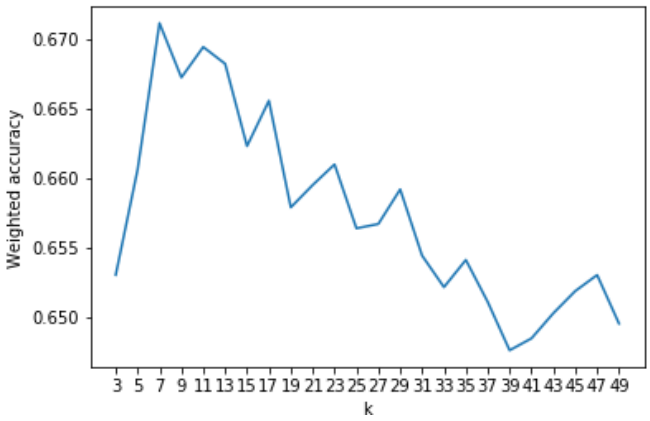

Fig. 16. Plot of Weighted Accuracy over a range of values of k

2) Logistic Regression: It produces the following output:

Actual \ Predicted	−	+	Total
−	5195	281	5476
+	745	723	1468
Total	5940	1004	6944

Logisitic Regression Confusion Matrix

$$Weight\,Accuracy = 63\% \tag{10}$$

3) SVM: The 10-fold cross validation is done on the entire dataset to find the most suitable kernel amongst Radial Bias Function, Gaussian, Linear and Polynomial functions. The linear kernel is found to give the best weighted accuracy.

Actual \ Predicted	−	+	Total
−	5192	284	5476
+	744	724	1468
Total	5936	1008	6944

SVM Confusion Matrix

$$Weight\,Accuracy = 63\% \tag{11}$$

4) Gradient Boost: We used an ensemble of 50 decision trees with a unity learning rate.

Actual \ Predicted	−	+	Total
−	4161	1312	5473
+	318	1153	1471
Total	4479	2465	6944

Gradient Boost Confusion Matrix

$$Weight\,Accuracy = 78\% \tag{12}$$

5) Neural Networks: During training of network, a batch of 200 training examples is sent to the network for each epoch . The training data is shuffled after every epoch as the model learns fastest from the most unexpected dataset [11]. The network gave the best results with two hidden layers of nine neurons each. The sigmoid(2) activation function has been used in hidden as well as output layers.

Actual \ Predicted	−	+	Total
−	5177	299	5476
+	641	827	1468
Total	5818	1126	6944

Neural Networks Confusion Matrix

$$Weight\ Accuracy = 70\% \tag{13}$$

TABLE II
COMPARISON OF PERFORMANCE OF CLASSIFICATION MODELS

Model	KNN	Logistic Regression	SVM	**Gradient Boost**	Neural Networks
Weighted Accuracy	67%	63%	63%	**78%**	70%

V. CONCLUSIONS

In our study, Gradient Boost proved to be the best classifier followed by neural networks. The results obtained portray that if sufficient data is available in the forthcoming years of data capture by U-DISE, the deployed machine learning algorithms can provide accurate predictions of 'at-risk' students in advance. The main reason for dropping out came to be academic performance and attendance of the students which is in accordance with researches and studies done in the past to determine the major factors behind drop outs. The availability of such information can help us take specific preventive actions and help students from dropping out of schools. Rather, it can also provide input for policy recommendations to contain drop out at the secondary school level and improve completion rates. A more detailed analysis is possible by conducting focused field research and taking into account the behavioural factors, and educational and financial background of the student's family. Such additional input variables and their relative weightage in explaining the phenomenon of drop out can provide greater insights into the fact whether it is school related, personal or socio-cultural factors.

REFERENCES

[1] R. Basumatary, "School dropout across indian states and uts: An econometric study," *International Research Journal of Social Sciences*, vol. 1, no. 4, pp. 28–35, 2012.

[2] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting students drop out: A case study." *International Working Group on Educational Data Mining*, 2009.

[3] G. Zhang, T. J. Anderson, M. W. Ohland, and B. R. Thorndyke, "Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study," *Journal of Engineering education*, vol. 93, no. 4, pp. 313–320, 2004.

[4] S. Gouda and T. Sekher, "Factors leading to school dropouts in india: An analysis of national family health survey-3 data," *IOSR Journal of Research and Method in Education*, vol. 4, no. 6, pp. 75–83, 2014.

[5] A. Nandeshwar, T. Menzies, and A. Nelson, "Learning patterns of university student retention," *Expert Systems with Applications*, vol. 38, no. 12, pp. 14 984–14 996, 2011.

[6] S. B. Kotsiantis, C. Pierrakeas, and P. E. Pintelas, "Preventing student dropout in distance learning using machine learning techniques," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2003, pp. 267–274.

[7] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques," *Computers & Education*, vol. 53, no. 3, pp. 950–965, 2009.

[8] "Student data collection in sync with u-dise." [Online]. Available: https://student.udise.in

[9] A. Pradeep, S. Das, and J. J. Kizhekkethottam, "Students dropout factor prediction using edm techniques," in *Soft-Computing and Networks Security (ICSNS), 2015 International Conference on*. IEEE, 2015, pp. 1–7.

[10] "Onefivenine.com." [Online]. Available: http://www.onefivenine.com/india/villag/Jalgaon/Bodwad

[11] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 9–48.

[12] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, pp. 3–24, 2007.

[13] Wikipedia contributors, "K-nearest neighbors algorithm — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=K-nearest_neighbors_algorithm&oldid=867125760, 2018, [Online; accessed 29-November-2018].

[14] P. McCullagh and J. A. Nelder, *Generalized linear models*. CRC press, 1989, vol. 37.

[15] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Mathematical Programming*, vol. 162, no. 1-2, pp. 83–112, 2017.

[16] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992, pp. 144–152.

[17] B. Schölkopf, A. J. Smola, F. Bach, *et al.*, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[18] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[19] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[20] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *Pattern recognition (ICPR), 2010 20th international conference on*. IEEE, 2010, pp. 3121–3124.