Dropout rate prediction among Indian youth using machine learning

[1]Sumit Vishwakarma, [2]Anuj Panchal, [3]Mansoorul Shaikh, [4]Kalpita Wagaskar

Department of Computer Engineering, Don Bosco Institute of Technology, Mumbai, India

[1]sumitvishwakarma6767@gmail.com, [2]panchalanuj961@gmail.com, [3]moinshaikh2556@gmail.com, 4kalpitags@gmail.com

Abstract— **India is a growing country and is one of the youngest countries with a greater number of young people. There is necessity of huge number of developments with the growth in the number of people in the country the consequent advancement of which totally depend on the young people who is taking education in the respective fields. But there is also a situation of concern as there is significant drop out rate of Student from school due to Numerous reasons which will have a huge impact on National development. It is also a reason of concern for Schools and is huge concern for the Educational Administers and Researchers. Parents send their child to school for their Child's Benefit but there are some factors which are really responsible for Dropping from school. Expectation of student towards the school is not fulfilled is one of the Reason of Termination. There are some schemes which are run by Government of Maharashtra free food (midday meal), girls fees waiver scheme, Vocational Courses (skill program), girls hostel scheme (girls safety program), Quality education (Qualified teacher or school training program), school infrastructure. In this paper we work towards predicting the drop-out rate for various features (Reason of dropout) in government school of Maharashtra Region. We also want to create awareness and schools know the reasons so that the school will take necessary actions or change their Course structure, schemes accordingly, For Prediction we are going to use Machine Learning Supervised algorithms (regression technique).**

Keywords — **Machine learning, Linear Regression supervised algorithm, correlation, prediction.**

I. INTRODUCTION

Education is critical aspect for socio-economic improvement of the individual or network. The country's Education influences employment possibilities and cap potential to stand for increasing demands of society for employment [1]. India is a young Country, Indian Youth Population is responsible for development, research and innovation of the country [2]. If youth is educated India will become developed nation. Foundation base of education is built in the school days, School education is incredibly important for every youth [3].

In India, there are various states where dropout in school is major concern for student in school. Age 14 – 16 years there has been a large dropout rate. According to data collected from the UDISE, Maharashtra dropout rate for children in school of academic year 2018-19 is 3.842, the dropout rate for children studying in Classes I to X is 3.374 for 2019-2020. Dropout from school can be dependent upon various feature. It included School Infrastructure, Teaching Quality, Lack of interest in studies, Child Labor, Sickness in the family, Sickness of the child, Poor academic performance etc.

For dropout rate concern, we require a early warning dropout rate predictive system. Predictive model system gives feature importance to prevent dropout rate for educational institute. High Dropout rates have negative effects on schools as well as Government. Government has many schemes to help prevent dropout rates [4], like Rashtriya Madhyamik Shikshan Abhiyan, RTE, Girls Hostel and Fees waiver schemes, National Schemes of Incentives and Vocational Education scheme (skill program) etc.

II. LITERATURE REVIEW

In [1], the study has been done on Dropout Students in Indian Schools, there are some predicting Machine Learning Algorithms for prediction Dropout and Non-Dropout. They Implement supervised Classification algorithms on dataset, such as Support Vector Machine, Naïve Bayes, Random Forest supervised Classification. The logistic regression showed an average precision of 0.978 and the highest of 0.93 on the validation results and the random forest algorithm showed a precision of 0.960 [1]. They have such variable which shows high collinear with the dependent variables, the teachers classroom ratio

 DOI : 10.35291/2454-9150.2022.0427

and Upper Administration can take planning of that act accordingly. In [2], Students' dropout Prediction Model through ML Techniques are used to Predict whether the student is Dropout or not. School has all the records of student which helps to construct ML Model Used 10-fold cross-validation to estimate generalization accuracy. Used Decision Tree Classifier and Naïve Bayes Classifiera single experiment and a small sample of students. Many researchers can replicate this type of study with larger sets of student data from many schools [3] includes discussion to overcome dropout rates in India School/colleges. By using advance machine learning algorithms like logistic regression, decision trees and K-Nearest Neighbours to predict whether a student will drop out or continue his/her education. Data were divided into training and testing parts with training data and then Preprocessing is happened (Delete Missing Values and Min Max Scaling) to select appropriate feature for Model. a single experiment and a small sample of students and Few algorithms are discussed [4] discussed To Prediction of student performance is the significant part in processing the educational data. Deep Neural Network is technique is used to classify multi class Classification problem [4]. Deep Neural Network architecture contains input layer hidden layer and output layer they constructed two hidden layer structure for classification. Didn't use traditional Machine Learning Model for the comparisons.

[5] analyses a dropout early warning system enables schools to pre-emptively identify students who are at risk of dropping out of school by using SMOT (Synthetic Minority Oversampling Techniques), Machine Learning methods used to detect. Data were divided into training and testing parts with training data and then Preprocessing is happened (Delete Missing Values and Min Max Scaling) to select appropriate feature for Model 10-fold tuning is used a single experiment and a small sample of students and few algorithms are discussed.

An Ensemble Predictive Model Based Prototype for Student Drop-out in Secondary Schools address to the Study Student dropout Problem in Tanzania [6]. ML Techniques are used to Predict whether the students are dropped out or not. They have used the combination of a tuned Logistic Regression and Multi-Layer Perceptron modelsa single experiment and a small sample of students. Evaluating performance of developed system

[7] addresses the problem to predicting the drop out feature of students. Academic performance in this study is measured by their cumulative grade point average (CGPA) upon the graduating. Decision tree technique to choose the best prediction and analysis. The list of students who are predicted as likely to drop out from college by data mining is then turned over to teachers and management for direct or indirect intervention. Only decision tree technique is used Random Forest algorithm can be implemented here.

Machine learning approach for reducing students' dropout rates address the methodology to study the out of school children problem. School has all the records of student which helps to construct ML Model Used 10-fold cross validation to estimate generalization accuracy. Using ML Decision Tree Classifier and Naïve Bayes Classifier and single experiment and a small sample of students [8].

To study the out of school children problem, by using clustering technique using genetic algorithm with respect to education data set, [9] Selected a cluster using some cluster of population then evaluate fitness function according to fitness rank is given. The authors apply centroid to it by generic Algorithm and generate final cluster. Small sample of students and only K-mean technique is used.

In [10], solution is provided to predict high school dropout rate. Using Machine Learning Approach to predict Using Machine Learning Approach using Naïve Bayes Classification. Only Naïve bayes Algorithm used there are other algorithms which can classify dropout more efficiently.

III. PROPOSED SYSTEM

In system architecture, first user has to login and give some questions on dropout rate. This information will be applied on ML model. In ML model, we are going to Preprocess the data.

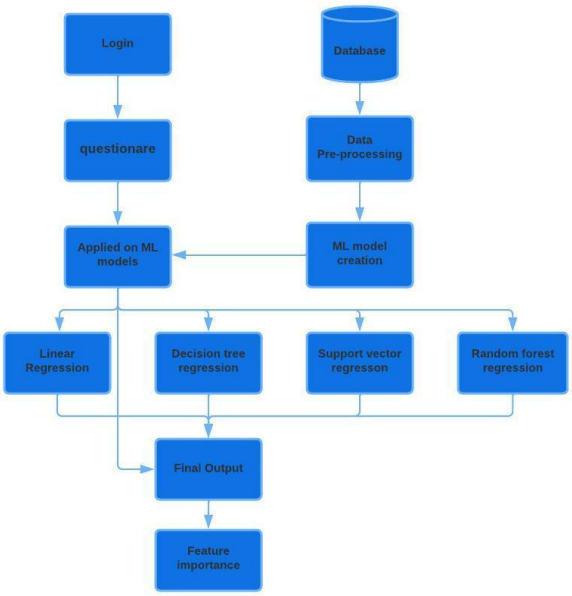

Fig. 1. Architecture Diagram

In the Preprocessing we are going to take care of missing values and find out which values are outliers. Remove such outliers which will affect our model to predict

 DOI : 10.35291/2454-9150.2022.0427

incorrectly. Then we are going to choose few columns (Independent Columns) which are highly correlated to our dropout (dependent column). We split the data into training and testing after that we create Supervised machine Learning Regression Models (Linear Regression, Decision Tree, Support Vectors regression, Random Forest) to predict Dropout Rate. Figure 1 shows the flow of the system.

Let us go in depth of each to see what it exactly

1. *User Authentication*: This login-authenticate system involves all the user authentication processes right from account creation to logging into the website and creating a database of all the users. The highlight of this part is the use of google firebase to create an option for the user to use google sign in if required. Also, we have used MySQL for the DBMS part.
2. *Data Preprocessing*: After getting a Dataset we going to create Machine Learning Model that prepare a data in such a way that Model will become ready. For that we are going to do pre-processing on a Data
   - *Remove Missing Values*: Python pandas libraries to read the data as well as to remove missing values because they create not working of our model, these missing values are also called as NaN values
   - *Conversion of Categorical Feature*: In the Dataset, may encounter some categorical values like let's say you have a column of places, Places name are considered as string but the problem is Machine Learning Algorithm only work on Number and not string, but here convert the text data to Number
   - Removing Outliers: In the dataset, feature column having outlier can affect the model to not predict effectively, such outliers have to be removed from feature column using stats library.
   - Correlation: In the dataset, which feature column (independent column) factors highly correlated with dependent column (dropout rate) to show correlation matrix.
3. *Question-are:* After taking various questionnaire on features information for dropout rate knowledge which is responsible for dropout rate. It gets questions are information and applied to machine learning model.
4. *Algorithm*: It involves the prediction of dropout rate using regression supervised ML (Linear regression, Support vector machine, Random forest regressor, Decision tree regressor) technique and also gives feature importance graph for dropout rate, so institutes can change their pattern school. Here we had check suitable algorithm (algorithm studied from related work) by testing on our dataset and after testing highly accurate algorithm is used to create a prediction model. Using extra tree regressor feature importance graph will be displayed based on model fitting, data generation technique.

IV. METHODOLOGY

A. Ground work survey and define the problem

To define the dropout rate problem in government schools of the Maharashtra region, we had done ground work on some schools by physically going to school. We had target two government schools in Ulhasnagar, Maharashtra region to identify what factors can be responsible for dropout rate and also, we had done a case study on dropout students by interactive talk. From that information, we can define our problem statement. Using UDISE data collection (raw data), we had worked on those factors and found rate dependency for features with dropout rate, also gives feature importance.

The survey conducted and sample data collected on the region, which is shown in Table 1:

Region	School address
Mharal , Thane , Maharashtra 421103	Zilla Parishad School Mharal pada
Ulhasnagar, Thane , Maharashtra 421002	Shahad Vibhag High School, Shahad

Table 1: Table showing the survey region

The sample dropout data is taken from these region survey in Table 2:

Region school	Dropout size
Zilla Parishad School Mharal pada	39
Shahad Vibhag High School, Shahad	15

Table 2: Table showing the survey data size

Sample data of these schools contains Reasons given by dropout students to school administration, shown in the below graph (Figure 2):

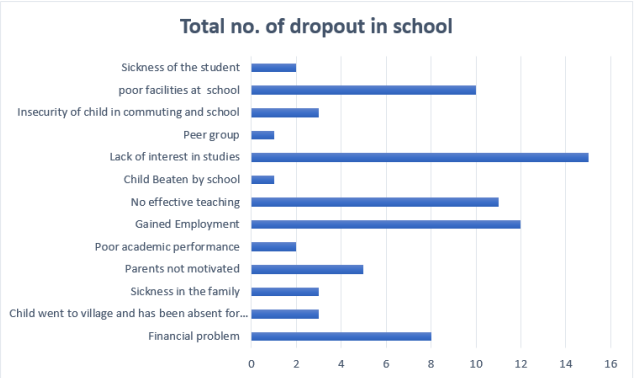

Fig.2. Number of dropout students in school w.r.t. reasons

Some responsible Factors of student case study:

1. *Peer Group*: The school environment and the

 DOI : 10.35291/2454-9150.2022.0427

company that the student life is a major often role in case of dropout. Some students had dropout out because their friend left the school, and by taking it as motivation, others also leave the school.

Case study 1: Peer group

Pratik, Kartik and Shivam three friends who live in Dhobighat, khemani Ulhasnagar area walk about 20 min to reach school. One of the friends in their group Kartik started working at an early age to earn money and left school. By taking as motivation the other friends of their group also start working in early age to make money and leave school.

2. *Infrastructure Facilities in the School*: Poor facilities in school like unclean toilet (Figure 3), Not separate operational toilet facility for girls and boys (poor door facility), not sanitization of all rooms, non-working fans, absence of safe drinking water (Figure 4) and library book reading facility, lack of seating arrangement, absence of functional electricity and others nonfunctional infrastructure can be major cause of dropout. Non infrastructure school cannot provide facility and interest in study for student. Hence lack of interest in studies increases for student.

Case study 2: Infrastructure Facilities in the School

Kajal and her three friends who live in Ganesh chawl colony, Mharal area walks about 10 min to reach school. Due to poor facility and non-functional separate toilet facility, she has a medical issue by which she has to go toilet frequently. Hence, she did not come to school and after her friends also started absent. From that due to their high absent, she had a lack of interest in her study. She had left the school.

Fig.3. Images of school toilet facility

Fig.4. Images of school water drinking facility

3. *Insecurity of child*: Even in the place of residence safety is a major issue for girls. Insecurity of student commuting in alone to school, especially in slums bad locality can be factor for insecurity.

Case study 3: Insecurity of the child

Shanti who lives in Mharal gaon walks about 15 minutes to reach school. She used to go with her friend but because of some family disputes her friends had leave the school, after that she has to go school alone. Four – Five boys were following and teasing her sometime. She had told these things to her parents; they did not allow her to continue school. As the reason may be educational attainment of parents.

4. *Bad teaching quality*: In government schools, non-strictness related to teacher attendance and evaluation of quality of teaching, also some region it is found teachers had a poor knowledge, low quality information and also, they were not interested to teach someone. They came to school for income perspective. Evaluation of teaching quality is required for prevent dropout of student.

B. Input (dataset using UDISE data)

1. ***Dataset:*** We had collected real time governmental survey data from UDISE (Unified District Information System for Education**)** site for Maharashtra region of governmental school (as there are no financial constraint for student, like fees). Dataset contain 19 feature column (input independent column), output (dropout rate) (dependent column) and 777 records having real time data of dropout rate and feature column.

Dataset having 19 feature column and output dropout rate:

2. **ML model**

After pre-processing, we had split the dataset in training and testing size ratio 0.75. We had tested four regression algorithms (Linear Regression, Random Forest Regressor, Support Vector Regression, Decision tree

 DOI : 10.35291/2454-9150.2022.0427

regressor) on our dataset as supervised technique. Regressor is name of any variable of regression model from using that prediction of response variable is taken.

Linear regression has a task create best fit line predict the dependent variable (Y) by input independent column (X). It will create regression model which is a linear relationship between input (X) and output (Y). The best fit line is a regression line.

Decision tree regression is nothing but a decision-making tool which uses a tree like structure for decision, also for continuous output prediction. For continuous output prediction decision tree observe the objects and train a model in a tree structure which having decision condition. Decision tree act like a set of binary rules for target value and every individual tree having leaves, nodes, branches.

Support vector Regression is based on SVM. SVR is use to predict continuous value which is used to find the best fit line. The straight line which is required to fit the data is hyperplane. Hyperplane are decision boundary having a large data point where data points are either side of hyperplane is support vector.

Random forest regression also known as collection of "Randomly created decision tree" uses quality feature by randomly creating decision tree i.e. Creates multiple decision tree to make decision. Random forest regression uses a multiple tree-based structure for decision. One single decision tree can't make decision well.

Root mean square error describes the error which is a difference between the actual value and predicted value. RMSE is the standard deviation of all error. R2 score or r – squared describes the performance of regression model. It is also a "coefficient of determination".

Random Forest Regression is suitable regression model on our dataset having less RMSE and high R2-score. By importing pickle module of python dump the Random Forest regressor variable by convert into object file, so that it can be used for prediction anywhere by loading on pickle. We used regressor object file (.sav file) on flask web app.

3. Feature Importance

Feature importance describes the evaluation of feature on dataset in regression task. By data generation and model fitting technique, importance of feature can be evaluated and show in bar graph. Extra Tree Regressor actually makes a meta estimator which is fits a randomized decision tree to improve predictive accuracy and control the overfitting. Error bar graph based on inter variability of feature column.

V. RESULT

1. From Government School survey we had found some informative and analytical data cause of dropout. Due to Poor infrastructure and facility, lack of interest in study, no effective teaching no. of dropout student in that school is 10, 15, 11, which is highest among the collected data (Figure 5).

Reason	Total no. of dropout in school
Financial problem	8
Child went to village and has been absent for substantial period of time and the school cut off the name and did not allow re entry to the class	3
Sickness in the family	3
Parents not motivated	5
Poor academic performance	2
Gained Employment	12
No effective teaching	11
Child Beaten by school	1
Lack of interest in studies	15
Peer group	1
Insecurity of child in commuting and school	3
poor facilities at school	10
Sickness of the student	2

Fig.5. Number of dropout student w.r.t. cause

2. Regression model (Linear regression, Decision tree regression, Support Vector regression, Random Forest regression) had fitted and created a regressor variable. Among all regressor, Random Forest regressor having highest R2-score – 94.20 and less RMSE - 0.8968 (Figure 6).

	Real Values	Predicted Values
10	17.378	16.28876
462	3.514	3.50756
743	0.532	0.51736
188	5.538	5.53864
397	3.894	3.90292
...	...	...
700	1.602	1.60208
507	3.684	4.52014
771	6.740	4.14474
412	6.392	3.69014
189	5.528	5.53190

Fig.6. Real value and model predicted dropout rate

3. For feature importance of dataset (Figure 7), data generation (x, y) and using fitted attribute feature_importance of extra tree regressor, feature importance graph will be display having inter-variability of feature column.

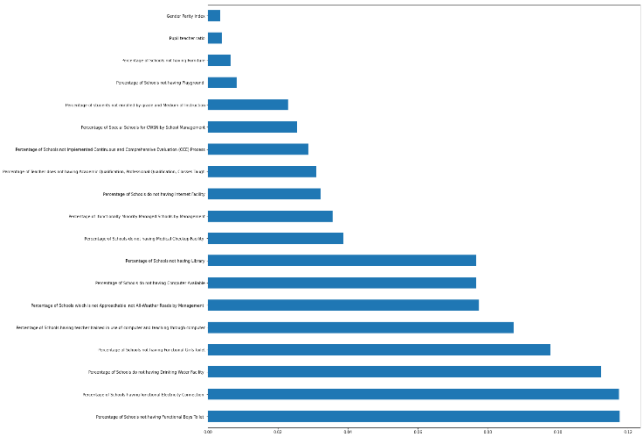

Fig.7. Feature importance image

4. Web application created (using flash web framework

 DOI : 10.35291/2454-9150.2022.0427

of python) for Early Dropout rate predictive system taking feature rate as input and flash dropout rate on prediction page. It gives the feature importance of the model.

VI. CONCLUSION

In this paper, we present our work what we aim for. The goal of this work is to create a web application for an early dropout rate predictive system and also give importance to responsible features (responsible features cause of dropout). We had collected two (nearby our location) government school survey data some informative and correlative to describe these dropout rate issues. From this survey, we conclude that poor facilities of school teaching quality can be a major cause of dropouts Its rate should be decreased, what we will provide is the dropout rate and feature importance of our model on a dataset. By entering the previous academic causes rate on our application, it will get the dropout rate, it will act as an early rate predictive system and school institute, government education organizations can act accordingly (as our prediction model tested on real-time data). After testing of different regression model, we got highest R2-score (accuracy / performance) on Random Forest regression model, we used these models. For feature importance of dataset, we fitted extra tree regressor model on data generation (x,y) and provided fitted attribute feature_importance. Bar graph has displayed by taking the feature importance inter-variability value of the feature.

VII. FUTURE SCOPE

In the future, major work can be done to target small regions or schools with details surveys and case study data through an interactive talk with students individually. AI agent method can use for predictive systems (like chatbot), by taking input agent will predict the rate and this technology can be used in Government School organization site, so that it will easily use.

REFERENCES

[1] Malik Mubasher Hassan,Tabasum Mirza "PREDICTION OF SCHOOL DROP OUTS WITH THE HELP OF MACHINE LEARNING ALGORITHMS," GIS SCIENCE JOURNAL, ISSN NO : 1869-9391, July 2020

[2] Jay S. Gil , Allemar Jhone P. Delima , Ramcis N. Vilchez,"Predicting Students' Dropout Indicators in Public School using Data Mining Approaches ",International Journal of Advanced Trends in Computer Science and Engineering,Volume 9 No.1 ISSN 2278-3091 January – February 2020

[3] Sunbok Lee and Jae Young Chung, " MACHINE LEARNING ALGORITHMS TO PREDICT POTENTIAL DROPOUT IN HIGH-SCHOOL," in 2016 EasyChair, Preprint No.3920, July 21, 2020 .

[4] V. Vijayalakshmi, K. Venkatachalapathy, "Deep Neural Network for Multi-Class Prediction of Student Performance in Educational Data," International Journal of Recent Technology and Engineering (IJRTE) , ISSN: 2277-3878, Volume-8 Issue-2, July 2019

[5] Sunbok Lee and Jae Young Chung, "The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction," Applied Sciences doi:10.3390, 2019.

[6] Neema Mduma , Khamisi Kalegele , Dina Machuve ,"An Ensemble Predictive Model Based Prototype for Student Drop-out in Secondary Schools," Journal of Information Systems Engineering & Management,ISSN: 2468-4376, August 22, 2019.

[7] Mr. M. N. Quadri Dr. N.V. Kalyankar, "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques," Global Journal of Computer Science and Technology , Vol. 10 Issue 2 (Ver 1.0), April 2010

[8] N. Mduma, K. Kalegele, and D. Machuve, "Machine learning approach for reducing students dropout rates," Int. J. Adv. Comput. Res., vol. 9, no. 42, pp. 156–169,2019, doi: 10.19101/ijacr.2018.839045.

[9] Astha Pareek, Amita Sharma, Manish Gupta, "A Study of Out of School Children Problem in Rajasthan using K-means clustering with Genetic Algorithm," in International Journal of Computer Applications (0975 – 8887), Volume 144 – No.8, June 2016.

[10] Vinayak Hegde, Prageeth P P, "Higher Education Student Dropout Prediction and Analysis through Educational Data Mining," e Second International Conference on Inventive Systems and Control (ICISC 2018), ISBN:978-1-5386-0806-7 2018.

 DOI : 10.35291/2454-9150.2022.0427