DOI: 10.21917/ijsc.2021.0324

PREDICTING INSTITUTE GRADUATION RATE WITH GENETIC ALGORITHM ASSISTED REGRESSION FOR EDUCATION DATA MINING

Mala H. Mehta[1], N.C. Chauhan[2] and Anu Gokhale[3]
[1,2]*Department of Information Technology, Gujarat Technological University, India*
[2]*Technology Department, Illinois State University, United States of America*

Abstract

In recent era of Digitization, large amount of computer generated data is accumulated on servers. However, these gathered data is useful only when interesting, novel and applicable knowledge is generated out of it. In the field of education, digitized data is generated by online academic activities. Data mining is discussed with every aspect of society however its use in academia is at infancy. Data mining is used to find novel, interesting and useful knowledge out of data which directs to actionable patterns on which academicians could work to enhance the productivity of academic activities. Education data mining is focused to use educational data for knowledge discovery to attain valuable insights in education domain. In this paper, an important task of prediction of institute graduation rate is addressed. Two novel approaches are proposed in the paper for effective graduation rate prediction. The first approach is genetic algorithm assisted regression model. The second approach investigates and uses various filter methods to further enhance the results in terms of time and number of features. Three regression models – multiple linear regression, decision tree regression and support vector regression are considered for experiments and comparative results are produced. The proposed methods provide better institute graduation rate prediction.

Keywords:

Genetic Algorithm, Regression, Education Data Mining, Filter Method

1. INTRODUCTION

Education data mining is aimed to churn useful knowledge out of digitized educational data. Data mining forms the basis to find novel, interesting and actionable patterns out of large amount of data. Three of important tasks in education data mining are prediction, classification/grouping of students and subgroup discovery. Examples of prediction include student performance prediction, student dropout prediction rate, institution graduation rate and students' placement prediction etc. Examples of classification/grouping include student classification based on their learning levels, grouping elective courses etc. Sub group discovery indicates identification of groups in hierarchical fashion using combination of supervised and unsupervised learning. Education data mining, involves facets of education experiments which includes cognitive learning, group learning, individual learning, visual learning etc. For every educational institute, it is important to strive in the current competitive environment by giving good placement to college students and achieving high results in performance parameters at primary, secondary and college level. The weightage of online learning has been greatly increased. In such scenario, it is required to investigate role of data mining in education using novel approaches. In this paper, the focus is on prediction of students' graduation rates as a whole for institutions. It also allows the institutes to know the relations of parameters associated with graduation rates. The factors that influence students' study most is vital to know for institutions to receive grants from government as well for many countries. A standard dataset is considered for this purpose, and is processed rigorously to get the desired format of dataset. The dataset considered is having large number of features and the target variable is a continuous variable. Various regression techniques are applied on datasets where dependent variable is continuous. To predict graduation rate of institutions, three regression algorithms are used in this work are: support vector regression, multiple linear regression and decision tree regression. The performance of these regression models is affected whenever the dataset contains high dimensionality. Handling high dimensionality during various data mining tasks is an important task. The selection of appropriate features for the task can be treated as an optimization problem. As number of dimensions increase, it becomes important for algorithm to choose features those are really worth and input giving for the subjected task. Evolutionary methods are one of prominent ways to handle optimization problems. Evolutionary computation techniques involve genetic programming, genetic algorithm (GA), evolutionary strategies and evolutionary programming in its basic form. Out of these techniques, genetic algorithms are one which are discussed and used in all fields of optimization of engineering. Genetic algorithm mimics human genetics process for optimization. It involves three operations: selection, crossover and mutation. Each operation is having various types to use and it is a matter of research that which operator should be used in which way for what purpose.

Remaining paper is arranged in following way. Section 2 describes detailed literature survey of evolutionary algorithms in data mining and education data mining. Section 3 presents problem definition. Section 4 narrates proposed algorithms. Section 5 describes dataset, experiments and results taken for both proposed approaches. Finally, the conclusion is presented at the end.

2. LITERATURE SURVEY

2.1 EVOLUTIONARY ALGORITHMS IN DATA MINING

Evolutionary algorithms have recently attracted many researchers to investigate its applicability in data mining field. Evolutionary algorithms (EA) are basically optimization algorithms. It could be applied to maximization or minimization problems. EAs identified area of applications in data mining is: Parameter Optimization, Feature Selection and Rule mining to get best set of rules. Classification of heart disease using K-Nearest Neighbours and Genetic Algorithm [1] presents an approach for helping clinical staff to take precautionary steps for heart patients. KNN is a lazy learner classifier which in combination of GA increases classification accuracy in this paper. GA is used to prune redundant and irrelevant attributes which contributes more towards classification. Least ranked attributes are removed and classification task is carried out. Application of K-means and GA for dimension reduction by integrating Support Vector Machine for diabetes diagnosis [2] uses K-means clustering algorithm for removing noisy features. GA is used to find optimal set of features here. SVM is finally applied as classification algorithm which predicts that from given attributes of a person he/she could have diabetes in near future or not. Results show that GA in combination with SVM and clustering can improve results. A novel medical assistance system based on data mining [3] focuses on use of Improved GA to extract proper behaviour model according to target data base of hospital and develop clinical pathways. Paper highlights IGA as main module of medical assistance system. Paper focuses on conversion of GA into Apriori algorithm. Reliable Confidence Measures for Medical Diagnosis with Evolutionary Algorithms [4] highlights on conformal predictors as machine learning algorithms which can assist medical personnel in taking decision. Paper proposes a CP with evolutionary algorithm for evolved efficient rule set. Genetic algorithm with logistic regression for prediction of progression to Alzheimer’s disease [5] emphases on selecting necessary features using GA for selecting necessary number of tests to perform for progress prediction of Alzheimer’s disease. Paper concludes that GA can be efficient for searching a combination of variables for best accuracy achievement especially when search space is large. Applying genetic algorithms to improve students’ academic performance by group formation [6] presents a novel approach where Genetic Algorithm (GA) is used to find optimal group for student learning. Author has considered a special designed formula for predicting future score of students. GA is applied to find finest groups of students based on their last examination’s score. Results of paper prove that students do better in group environment. Groups created by GA was also balanced and a proper combination of good and poor students. Evolutionary algorithm approach to pupils’ pedantic accomplishment [7] proposes an approach for a broad group formation technique among pupils studying in a regular college. Main objective is to increase the sum of difference between the previous score and the predicted score. Authors have used GA to achieve optimal group formation taking inspiration from travelling salesman Problem (TSP). Contemplating crossover operators of genetic algorithm for student group formation problem [8] compares three crossover operators for genetic algorithm which is used in optimum group formation for student better performance achievement. Partially mapped crossover, order crossover and edge recombination crossover operators are compared and it is found that statistically edge recombination crossover operator is efficient. Using genetic algorithm for data mining optimization in an education web-based system [9] presents an approach for classifying students in order to predict their final grade based on features mined from logged data in an educational web-based system. A combination of multiple classifiers leads to a substantial improvement in classification performance. Through weighting feature vectors using a genetic algorithm authors have optimized the prediction accuracy. A novel hybrid wrapper-filter approach based on genetic algorithm, particle swarm optimization for feature subset selection [10] proposes algorithm which focuses on reducing complexity of calculation time and search time in feature subset selection problem. Proposed algorithm integrates filter and wrapper approach to take advantage of speed of filter approach and accuracy of wrapper approach. Improvement in terms of classification accuracy and feature subset size is measured compared to other hybrid methods.

2.2 EDUCATION DATA MINING

Electronic data gathered by various educational and governmental institutes raise opportunity for generating useful and novel data patterns out of it. In education, challenges are no less. Student dropout rates, student deteriorating performance, students’ placement and teaching learning challenges are just few examples. Traditional systems or online systems always give data to analyse for wide range of problems. The problems of education domain can be categorised roughly in three areas. First, Classification or prediction problems, examples are student performance prediction, student dropout rate prediction or institute graduation rate prediction. Second, clustering or grouping problems, examples are efficient student grouping for group learning, ideal student grouping for assigning projects etc. Third, Subgroup discovery problems are semi-supervised techniques which prepare groups with respect to property of interest. Here, a target variable is considered and with respect to it groups are created. Many algorithms are suggested in literature for this. Example is identifying group of students having similar characteristics for online portal. A survey of data mining approaches in performance analysis and evaluation was published by Shelke and Gadke [11]. Paper presented a detailed survey including classification algorithms used, its purpose and open source tools used. A neuro-fuzzy approach in the classification of students’ academic performance was published by Do and Chen [12]. Paper explained various classification approaches including support vector machines (SVM), naïve Bayes classifiers, neural networks (NN) and decision trees. An NFC is a multilayer feed-forward network. NFC includes input layer, membership layer along with fuzzification and defuzzification layer. Output layer gives output as classes after passed from normalization layer. Features are input for NFC architecture. Results show that NFC model can be used to classify students in to different groups based on their expected academic performance levels. Model achieved accuracy of over 90% which shows that it could be accepted as a classifier of students’ academic performance levels. Learning from student data was published by Barker et al. [13]. Paper introduces the use of SVM and NN for classifying student graduation behaviour from academic, demographic and attitudinal variables maintained about students at university of Oklahoma. Analysis of data was done by three methods, combined data, between years’ data and among years’ data. Results show that more work is necessary to improve results as misclassification error is high. Data mining in education was published by Romero and Ventura [14]. Paper introduces and reviews key milestones and the current state of affair in the field of EDM, combined with tools, applications and future in sights. Detailed background study is preceded by types of educational environment which are categorized into traditional and computer based education. Computer based systems could further be classified into learning management system, intelligent tutoring system, adaptive and intelligent hypermedia system and test and quiz systems. Stakeholders of EDM are discussed along with current topics of interest in EDM. Methods of EDM are discussed and narrated. They include predictions, clustering, outlier detection, relationship mining, social network analysis, process mining, text mining, discovery with models etc. Education data mining: a case study was published by Merceron and Yacef [15]. Authors have performed case study on Logic-ITA web based tool used at Sydney University. After doing data exploration, association rule mining, clustering and classification is applied on data and how the results could help teachers and learners are narrated. Applicability of EDM in real world is performed. A comparative analysis of techniques for predicting academic performance was published by Thai-Nighe et al. [16]. Paper compares accuracy of decision tree and Bayesian network algorithms for predicting the academic performance of undergraduate and post graduate students at two different universities. Results prove that data mining algorithms are efficient to get good accuracy even on two diverse populations of students. Results are useful for identifying failing students at one university whereas for other university it was helpful in identifying very good students who are eligible for scholarships. Decision tree was found better than Bayesian network in this paper. Predicting student academic performance in an engineering dynamics course: A comparison of four types predictive mathematical models was published by Huang and Fang [17]. Paper presents a study for prediction of student academic performance by four mathematical models. These models include multiple linear regression (MLR), SVM, multilayer perception network and radial basis function network model. Data is gathered from undergraduate students and six different sets of variables are considered as input. Paper presents a detailed study on data collection, pre-processing and predictive modelling. Paper concludes that if instructor's goal is to predict average academic performance of his/her dynamics class as a whole, the instructor should use MLR and if the instructor's goal is to predict individual student's academic performance, then the instructor should use SVM. Performance analysis and prediction in education data mining: a research travelogue was published by Thakar et al. [18]. Paper presents a comprehensive survey from 2002 to 2014 and also discusses its scope in future. Authors have divided the research in EDM into five categories and have presented a detailed Table, which includes key findings and methodology of authors. Data mining: a prediction for performance improvement using classification was published by Bhardwaj and Pal [19]. Paper presents an investigation for preparing education database and its applicability using Bayes classification. Students' academic performance has effects of many factors which are not only dependent on students' own efforts, which is the finding of the study. An analysis of education data mining in advanced education system [20] is a survey paper published by G. Rao and DLS Reddy in 2015. D. Fatima, S. Fatima and A.V. Krishna Prasad published a survey on research work in EDM [21] in 2015. A frame work for research on technology-enhanced special education [22] focuses on two projects aimed at individual needs of special students. Paper presents a frame work for dealing with this issue in education. A comprehensive study of education data mining [23] presents a survey on EDM which includes EDM process, EDM modules with its applications and tools. Assessment of student feedback from the training course and instructor performance through the combination of clustering methods and decision tree algorithms [24] addresses the subject of prediction of instructor performance by student feedback. Paper proposes a methodology which uses PCA followed by clustering and finally applying Decision tree classifier. Decision tree algorithm gives good accuracy of 0.93. A survey on predicting student dropout analysis using data mining algorithms was published by Balraj and Malini [25]. Research work suggests a model which can recognize whether student will continue study or not based on decision tree classification algorithm. Factors to predict dropout at the universities: a case of study in Ecuador by Mayra and Mauricio [26]. The research has objective to design a model to determine new factors to predict the dropout in which five dimensions of analysis are considered. SVM, Logistic Regression (LR) and Decision tree algorithms are applied to check whether new factors have effect on dropout prediction or not. Predicting the performance fluctuation of students based on long-term and short-term data [27] focuses on students' performance prediction for different duration of time. Step regression, decision trees, logistic regression and SVM Regression are used in the paper to predict the outcome. Predicting student performance using advanced learning analytics [28] focuses on problem definition of student performance prediction. Two types of models are used to predict the output. These are discriminative and generative, which includes SVM, C4.5, Classification and Regression tree (CART), Bayes Network (BN) and Naïve Bayes (NB). Authors have proposed new set of features which include personal information and family expenditure of students. Paper concludes that new proposed features have significant effect on output and SVM is found effective out of all models. Predicting student success using data generated in traditional educational environments [29] focuses on predicting students at risk early so that tutors could take appropriate action. For this, authors have gathered data from traditional education system where for years, data has been collected. Five classification models are decision tree CART, Extra tree classifier, random forest, Logistic Regression and C-support vector classification. Results prove that Logistic regression is effective the most. Data mining application on students' data was published by Buldu and Ucgun [30], paper focuses on use of Association Rule Mining in education domain. Authors focus on students' different standards' grades and analysed that if a student has failed in certain subject in some standard then can it be derived that he/she would also fail in specific subject of next coming standards. This paper derives useful rules as output by applying Apriori algorithm on students' failed subjects data. Predicting student performance by using data mining methods for classification [31] presents initial results from a data mining research project implemented at a Bulgarian university, aimed at revealing potential of data mining applications for management of university. Six classification models are used out of which decision tree classifier performs best. Evolutionary algorithms for subgroup discovery applied to e-learning data [32] presents use of subgroup discovery technique applied to e-learning data to get useful rules as output. Authors have used genetic algorithm as Evolutionary algorithm with subgroup discovery to get results. Two detailed survey papers about EDM are as follows. Education Data Mining: A Review of the state of the Art [33] gives a detailed survey about EDM dividing it in to several useful categories. Education data mining: A survey from 1995 to 2005 [34] could be regarded also a pioneer paper in the field of EDM.

3. PROBLEM DEFINITION

Given training samples $(X_i,z_i),\ldots,(X_n,z_n)$ where X_i is the feature vector for institute a, and A is the set of n institutes where $A=\{a_1,a_2,a_3,\ldots,a_n\}$. The $X_i \in R$ where R is feature set containing total number of features and z_i is academic performance status. To predict the performance of an institute, the following prediction function is proposed: $z_i = F(X)$

Learning Task: Aim is to learn a predictive function F or alternatively to predict institute performance. It is written as: $Z=F(X)$."

4. PROPOSED ALGORITHMS

4.1 EVOLUTIONARY FEATURE SELECTION BASED REGRESSION MODEL FOR INSTITUTE GRADUATION RATE (IGR) PREDICTION

Above survey shows that for individual students, studies have been carried out. For institutions there is a need to design a framework to effectively predict institute graduation rate. Detailed survey depicts that evolutionary algorithms are not much applied in education data mining with regression being rarely used. In this section, a novel genetic algorithm based evolutionary feature selection with regression technique is presented for institute graduation rate prediction. Feature selection in data mining is an important task. In data mining, dimensionality reduction is considered a useful step in pre-processing of data. Datasets in real world are having many features but it is important to recognize that all features might not be really significant for problem at hand. For researchers, to find out what features should be actually considered for prediction and what features are having effect on output is vital to know. Feature selection is an optimization problem to find optimal number/set of features from available features. Mathematically, if n numbers of features are there, then there are 2^n possible sets of features which could be tested to see if it gives maximum accuracy in case of classification or minimum error in case of regression. High dimension dataset with more than 100 features is considered for current study. After integrating and pre-processing dataset, genetic algorithm is applied for optimum features selection. Features selected by GA are then applied with regression for calculating error between actual and predicted results. Detailed Pictorial view in form of flow chart for Proposed Algorithm is given in Fig.1.

4.2 A HYBRID FILTER-WRAPPER APPROACH USING EVOLUTIONARY FEATURE SELECTION MODEL FOR INSTITUTE GRADUATION RATE (IGR) PREDICTION

In proposed algorithm 1, wrapper approach of feature selection method is used for efficient prediction. Wrapper approach is a type of feature selection which takes help of learning algorithm for selecting vital features. Combined with evolutionary algorithm, wrapper approach is effective yet slow in processing. To further enhance the proposed method considering time dimension, filter approaches are investigated. Filter approach for feature selection contains single variate and multi variate analysis. Filter approaches analyze feature characteristics before selecting features. They do not need any learning algorithm for taking decision. Filter approaches are faster than wrapper approaches. Single variate filter methods concentrate on single feature at a time, analyzes it and determines rank of them. Multi variate feature methods consider interdependent relationships of features, analyze them and decides rank of feature. Hybrid Filter-Wrapper approach is proposed in Fig.2 which involves single variate methods and multi variate methods in combination.

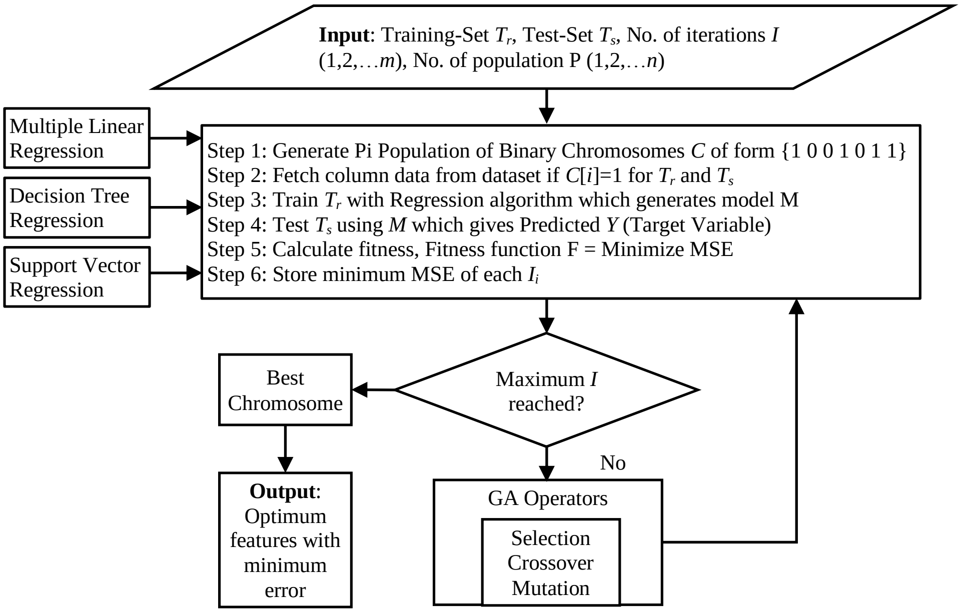
Input: Training-Set T_r, Test-Set T_s, No. of iterations I (1,2,…m), No. of population P (1,2,…n)
Multiple Linear Regression
Decision Tree Regression
Support Vector Regression
Step 1: Generate Pi Population of Binary Chromosomes C of form {1 0 0 1 0 1 1}
Step 2: Fetch column data from dataset if $C[i]$=1 for T_r and T_s
Step 3: Train T_r with Regression algorithm which generates model M
Step 4: Test T_s using M which gives Predicted Y (Target Variable)
Step 5: Calculate fitness, Fitness function F = Minimize MSE
Step 6: Store minimum MSE of each I_i
Maximum I reached?
Best Chromosome
No
Output: Optimum features with minimum error
GA Operators
Selection
Crossover
Mutation

Fig.1. Evolutionary Feature Selection Model for IGR Prediction

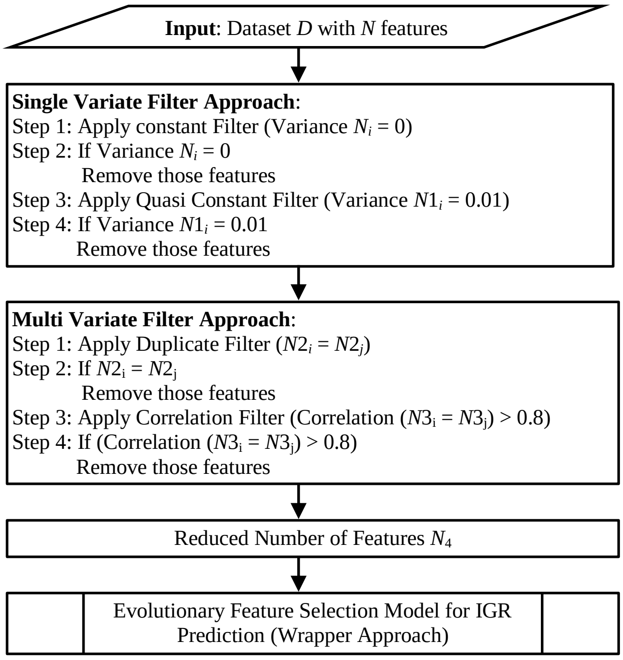
Input: Dataset D with N features
Single Variate Filter Approach:
Step 1: Apply constant Filter (Variance $N_i = 0$)
Step 2: If Variance $N_i = 0$
Remove those features
Step 3: Apply Quasi Constant Filter (Variance $N1_i = 0.01$)
Step 4: If Variance $N1_i = 0.01$
Remove those features
Multi Variate Filter Approach:
Step 1: Apply Duplicate Filter ($N2_i = N2_j$)
Step 2: If $N2_i = N2_j$
Remove those features
Step 3: Apply Correlation Filter (Correlation ($N3_i = N3_j$) > 0.8)
Step 4: If (Correlation ($N3_i = N3_j$) > 0.8)
Remove those features
Reduced Number of Features N_4
Evolutionary Feature Selection Model for IGR Prediction (Wrapper Approach)

Fig.2. Hybrid Filter-wrapper Model for IGR Prediction

5. DATASET, EXPERIMENTS AND RESULTS

5.1 DATASET USED IN STUDY

Integrated Postsecondary Education Data System (IPEDS) [35] is a system of U.S. department established for reviewing and funding education institutes of various states of U.S.A. According to rules established by government of U.S.A. all education institutes submit their data annually to IPEDS. IPEDS is a big repository containing data about different institutes of states. It contains data about institute characteristics, human resources, finance, salary, graduation rate and enrolment of institutions for more than 10 years. Different research has been carried out on IPEDS dataset [36]-[39]. Regression analysis with evolutionary feature selection for IGR prediction is a novel approach. In this study, full-time, first-time students seeking a bachelor's or equivalent degree - 2006 Bachelors sub cohort for 4-year institutions are targeted.

5.2 DATA INTEGRATION

IPEDS has many year data which is distributed in terms of tables. One cohort year involves approximately 40 tables for different facets of institution. It took a lot of efforts to understand data and then integrate it. For preparing targeted cohort dataset various types of 12 queries and selection operations were performed and integration was achieved.

5.3 FEATURE EXTRACTION

IPEDS is a large repository, and all features in every table needed a special caution. After thorough study and investigation, for concerned study, features were extracted as shown in Table.1. Thus total of 152 features were extracted out of which 16 features were merged in to 8 features. One feature was deleted as it was having NULL values completely. A final dataset of 903 rows and 143 features were constructed for our study. Out of 143 features, Graduation rate is dependent variable whereas 142 are independent variables for concerned study.

5.4 DATA PRE-PROCESSING

After understanding whole dataset and extracting features according to interest of study, it was necessary to understand data and pre-process it. Many features were having null data which were substituted by mean value of that feature. To tackle outliers and skewness of data, visual techniques of data mining were used. To handle skewness of data, histograms were created for each feature. Data was scaled down to normalize range to handle skewness.

5.5 PERFORMANCE PARAMETERS

To assess performance of proposed work, four parameters were considered. All parameters were decided according to prediction technique considered at hand that is regression.

- *Mean Squared Error:* It measures the average of the squares of the errors—that is, the average squared difference between the predicted values$\hat{y}$ and the actual values y. MSE is a risk function, corresponding to the expected value of the squared error loss. Total data samples in formula are represented by n as given in Eq.(1).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y-\hat{y}\right)^{2} \quad (1)$$

Table.1. Number of Features Extracted: Table and Survey wise

Survey Component	Number of Tables Selected	Number of features selected
Graduation rates	2	5
Institutional Characteristics	4	83
Student Financial Aid	1	17

Library	1	3
HR	5	27
Finance	1	9
Enrolment	1	8

- *Root Mean Squared Error:* It represents the square root of MSE. It shows a perfect fit of a model. Lower values of RMSE indicate perfect fit of model for data.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y-\hat{y}\right)^2} \tag{2}$$

- R^2*:* It represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. $\bar{y}$ is mean of actual values y.

$$R^2 = 1 - \frac{\sum\left(y-\hat{y}\right)^2}{\sum\left(y-\bar{y}\right)^2} \tag{3}$$

- *Adjusted* R^2*:* is an adapted version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared rises only if the new term improves the model more than would be expected by chance. It declines when a predictor improves the model by less than expected by chance.

$$\textit{Adjusted } R^2 = 1-\left[\left(\frac{n-1}{n-k-1}\right)*\left(1-R^2\right)\right] \tag{4}$$

where *n* is number of observations and *k* is number of independent variables.

5.6 ALGORITHM ENVIRONMENT

Graduation rate is a continuous variable to predict as a dependent variable. Whenever, target variable is continuous, regression technique should be applied. For this study, three regression algorithms are considered: Multiple Linear Regression, Decision tree regression and Support Vector Regression. Genetic algorithm parameters are decided as follows: Tournament selection size=10, Mutation probability=0.2 and crossover probability=0.5.Training and Testing size of dataset is having 80-20% ratio.

5.7 RESULTS

5.7.1 Evolutionary Feature Selection Based Regression Model for Institute Graduation Rate (IGR) Prediction:

Proposed algorithm for feature selection is applied on final prepared dataset for measuring performance parameters. Features obtained by GA are then given as input to regression algorithms. Various experiments have been performed for different generations and populations of GA.

- **Multiple Linear Regressions**

Table.2. Multiple Linear Regressions without GA Feature Selection

MSE	RMSE	R^2	Adjusted R^2	Features
5.14	2.26	-2.28	-14.54	142

The Table.2 shows results of MLR before applying proposed algorithm on test set. The Table.3 shows results of MLR after proposed algorithm is applied. To check the performance, three runs were taken; iterations were fixed at 60 with varying population. The Table.4 shows results of MLR after proposed algorithm is applied with fixed population and varying iterations, three runs were taken for each instance. Among three runs, best result for each run is put in table.

Table.3. MLR with GA Feature Selection Iterations=60

Population	Iterations	MSE	RMSE	R^2	Adj. R^2	Features Selected
50	**60**	**0.86**	**0.93**	**0.45**	**0.07**	**72**
60	60	0.90	0.94	0.42	0.09	66
70	60	0.91	0.95	0.41	0.02	73

Table.4. MLR with GA Feature Selection Population=40

Population	Iterations	MSE	RMSE	R^2	Adj. R^2	Features Selected
40	**40**	**0.93**	**0.96**	**0.40**	**0.11**	**58**
40	50	0.93	0.96	0.40	0.02	70
40	60	0.88	0.94	0.43	0.08	69

The above results show that MLR with GA gives minimum MSE of 0.86 with 0.45 R^2which is significantly improved in comparison of results before applying proposed algorithm MSE 5.14 and R^2-2.28. It should be noted that features are reduced from 142 to 72. MLR with GA gives lowest features 58 with MSE 0.93.

- **Decision Tree Regression**

The Table.5 shows results of DTR before applying proposed algorithm on test set. The Table.6 shows results of DTR after proposed algorithm is applied. To check the performance three runs were taken, iterations were fixed at 60 with varying population. The Table.7 shows results of DTR after proposed algorithm is applied with fixed population and varying iterations, three runs are taken here for each instance. Among three runs, best result for each run is put in table.

Results show that DTR with GA gives minimum MSE of 0.57 with 0.63 R^2 which is greatly improved in comparison of results before applying proposed algorithm MSE 1.69 and R^2 - 0.08. It should be noted that features are reduced from 142 to 76. DTR with GA gives lowest features 62 with MSE 0.84.

Table.5. Decision Tree Regression without GA Feature Selection

MSE	RMSE	R2	Adjusted R2	Features
1.69	1.30	-0.08	-4.13	142

Table.6. DTR with GA Feature Selection Iterations=60

Population	Iterations	MSE	RMSE	R^2	Adj. R^2	Features Selected
50	60	0.77	0.87	0.50	0.14	76
60	**60**	**0.57**	**0.75**	**0.63**	**0.36**	**76**
70	60	0.75	0.86	0.51	0.18	74

Table.7. DTR with GA Feature Selection Population=40

Population	**Iterations**	**MSE**	**RMSE**	**R^2**	**Adj. R^2**	**Features Selected**
40	40	0.71	0.84	0.54	0.25	69
40	**50**	**0.84**	**0.91**	**0.46**	**0.18**	**62**
40	60	0.68	0.82	0.56	0.30	67

5.7.1.3 Support Vector Regression:

The Table.8 shows results of SVR before applying proposed algorithm on test set. The Table.9 shows results of SVR after proposed algorithm is applied. To check the performance three runs were taken, iterations were fixed at 60 with varying population. The Table.10 shows results of SVR after proposed algorithm is applied with fixed population and varying iterations, three runs are taken here for each instance. Among three runs, best result for each run is put in table.

Table.8. Support Vector Regression without GA Feature Selection

MSE	**RMSE**	**R^2**	**Adj. R^2**	**Features**
1.09	1.04	0.30	-2.29	142

Table.9. SVR with GA Feature Selection Iterations=60

Population	**Iterations**	**MSE**	**RMSE**	**R^2**	**Adj. R^2**	**Features Selected**
50	60	0.73	0.8569	0.53	0.31	57
60	60	0.76	0.8753	0.51	0.27	60
70	60	0.74	0.8639	0.52	0.29	59

Table.10. SVR with GA Feature Selection Population=40

Population	**Iterations**	**MSE**	**RMSE**	**R^2**	**Adj. R^2**	**Features Selected**
40	**40**	**0.70**	**0.84**	**0.54**	**0.36**	**51**
40	50	0.78	0.88	0.49	0.28	54
40	**60**	**0.72**	**0.85**	**0.53**	**0.36**	**49**

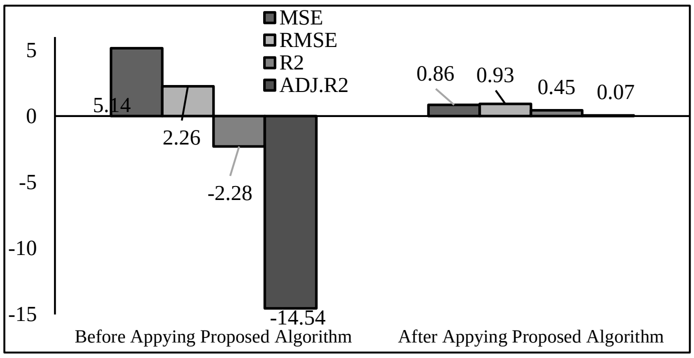

Fig.3. MLR performance after and before applying proposed algorithm

Above results show that SVR with GA gives minimum MSE of 0.70 with 0.54 R^2 which is visibly improved in comparison of results before applying proposed algorithm MSE 1.09 and $R^2$0.30. It should be noted that features are reduced from 142 to 51. SVR with GA gives lowest features 49 with MSE 0.72. Below a comparative graph summary for performance of proposed algorithm with three regression algorithms before applying GA and after applying GA is presented.

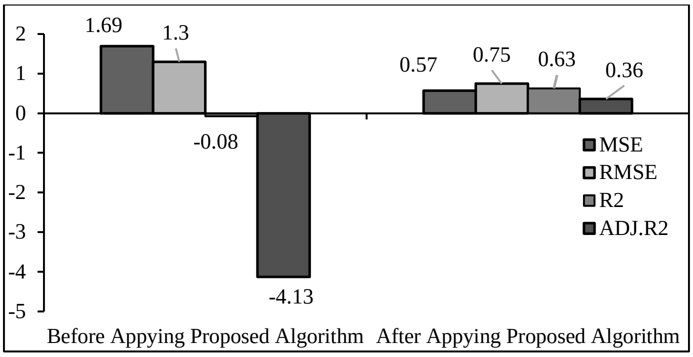

Fig.4. DTR performance after and before applying proposed algorithm

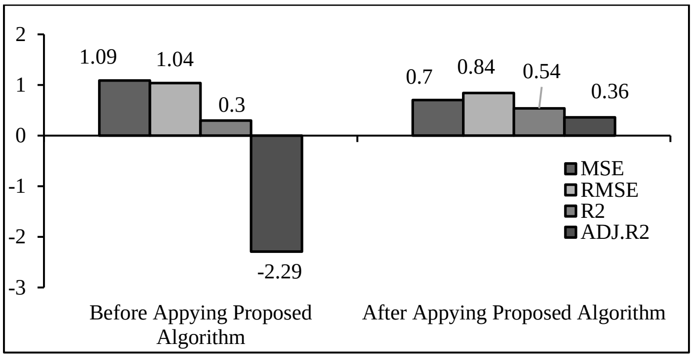

Fig.5. SVR performance after and before applying proposed Algorithm

It is evident to note that Fig.3 – Fig.5 clearly show that MSE and RMSE have reduced in all three cases whereas $R^2$and adjusted R^2have increased, which is an indication that proposed algorithm definitely have improved the performance of Three regression models. Comparative analysis for three regression algorithms' performance with proposed algorithm is produced in Fig.6. The Fig.6 shows undoubtedly DTR with Proposed Algorithm produces optimum results. SVR with GA performs better than Multiple Linear Regression with GA but poorer than Decision tree regression algorithm with GA. Above results undeniably prove that proposed algorithm enhances results and is successful enough to form an efficient model for institution graduation rate prediction.

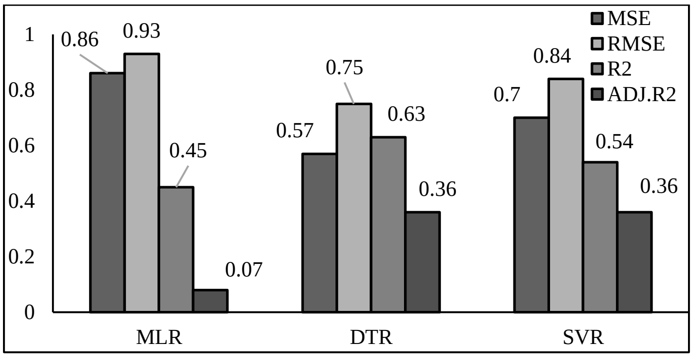

Fig.6. Comparative Analysis

5.7.2 A Hybrid Filter-Wrapper Approach using Evolutionary Feature Selection Model for Institute Graduation Rate (IGR) Prediction:

Proposed algorithm 2 is applied as Pre-processing step on Final prepared dataset and then it is input to Proposed Algorithm 1. Main goal of developing hybrid filter-wrapper model was to reduce time complexity inherent in wrapper approach used in proposed work 1.

- **Multiple Linear Regressions**

The Table.11 shows results of MLR after proposed hybrid algorithm is applied. To check the performance three runs were taken, iterations were fixed at 60 with varying population. The Table.12 shows results of MLR after proposed hybrid algorithm is applied with fixed population and varying iterations, three runs are taken here for each instance. Out of three runs, minimum result for each run is put in table.

MLR with hybrid Approach gives minimum MSE of 0.91 with 0.41 R^2 which is comparable with results of proposed approach 1 MSE 0.86 and R^2 0.45. Features are reduced from 72 to 31, which is noteworthy. MLR with hybrid approach gives lowest features 31 with MSE 0.91.

Table.11. Hybrid Approach with MLR with GA Iterations=60

Population	**Iterations**	**MSE**	**RMSE**	**R^2**	**Adj. R^2**	**Features Selected**
50	60	0.91	0.95	0.41	0.26	38
60	60	0.91	0.95	0.41	0.28	34
70	**60**	**0.91**	**0.95**	**0.41**	**0.29**	**31**

Table.12. Hybrid Approach with MLR with GA Population=40

Population	**Iterations**	**MSE**	**RMSE**	**R^2**	**Adj. R^2**	**Features Selected**
40	40	0.93	0.96	0.4	0.26	34
40	50	0.93	0.96	0.4	0.27	32
40	60	0.91	0.95	0.41	0.28	33

5.7.2.2 Decision Tree Regression:

Table.13 shows results of DTR after hybrid approach is applied. To check the performance three runs were taken, iterations were fixed at 60 with varying population. The Table.14 shows results of DTR after hybrid approach is applied with fixed population and varying iterations, three runs are taken here for each instance. Among three runs, best result for each run is put in table. DTR with hybrid approach gives minimum MSE of 0.62 with 0.59 R^2 which is comparable in performance with proposed algorithm 1 MSE 0.57 and R^2 0.63. Features are reduced from 76 to 40. It gives lowest features 34 with MSE 0.77.

Table.13. Hybrid Approach with DTR with GA Iterations=60

Population	**Iterations**	**MSE**	**RMSE**	**R^2**	**Adj. R^2**	**Features Selected**
50	**60**	**0.77**	**0.87**	**0.5**	**0.39**	**34**
60	60	0.8	0.89	0.48	0.35	35
70	60	0.76	0.87	0.51	0.37	39

Table.14. Hybrid Approach with DTR with GA Population=40

Population	**Iterations**	**MSE**	**RMSE**	**R^2**	**Adj. R^2**	**Features Selected**
40	40	0.78	0.88	0.49	0.38	34
40	50	0.78	0.88	0.49	0.37	35
40	**60**	**0.62**	**0.79**	**0.59**	**0.48**	**40**

- **Support Vector Regression**

The Table.15 shows results of SVR after proposed hybrid algorithm is applied. To check the performance three runs were taken, iterations were fixed at 60 with varying population. The Table.16 shows results of SVR after proposed hybrid algorithm is applied with fixed population and varying iterations, three runs are taken here for each instance. Out of three runs, minimum result for each run is put in table.

SVR with hybrid Approach gives minimum MSE of 0.70 with 0.54 R^2 which is as good as results of proposed approach 1 MSE 0.72 and R^2 0.54. Features are reduced from 51 to 28. SVR with hybrid approach gives lowest features 17 with MSE 0.73.

Table.15. Hybrid Approach with SVR with GA Iterations=60

Population	**Iterations**	**MSE**	**RMSE**	**R^2**	**Adj. R^2**	**Features Selected**
50	60	0.74	0.86	0.52	0.42	30
60	60	0.71	0.84	0.54	0.46	27
70	**60**	**0.73**	**0.85**	**0.53**	**0.48**	**17**

Table.16. Hybrid Approach with SVR with GA Population=40

Population	**Iterations**	**MSE**	**RMSE**	**R^2**	**Adj. R^2**	**Features Selected**
40	40	0.74	0.86	0.52	0.43	29
40	50	0.74	0.86	0.52	0.43	29
40	**60**	**0.7**	**0.83**	**0.54**	**0.46**	**28**

Results show that hybrid filter-wrapper approach gives satisfactory performance in all three cases as far as MSE parameter is considered. However, it shows great potential in reducing number of features. DTR with hybrid approach gives lowest MSE. SVR with hybrid approach gives minimum no. of features.

Table.17. Execution time (seconds) performance analysis

Regression Algorithm Used	**Proposed Approach 1**	**Proposed Approach 2**
Multiple Linear Regression	107.05 sec	78.19 sec
Decision Tree Regression	150.43	85.67 sec
Support Vector Regression	223.06 sec	141.17 sec

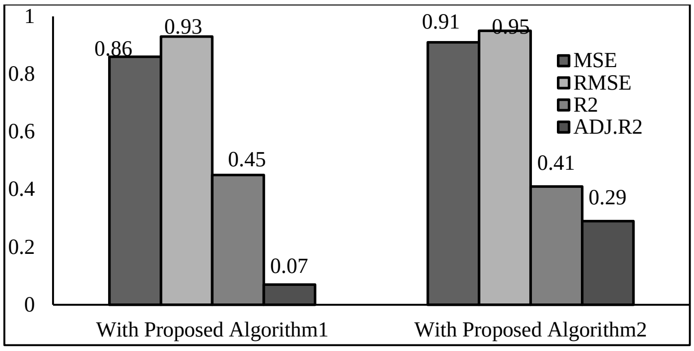

Fig.7. MLR performance comparison with proposed approach 1 and 2

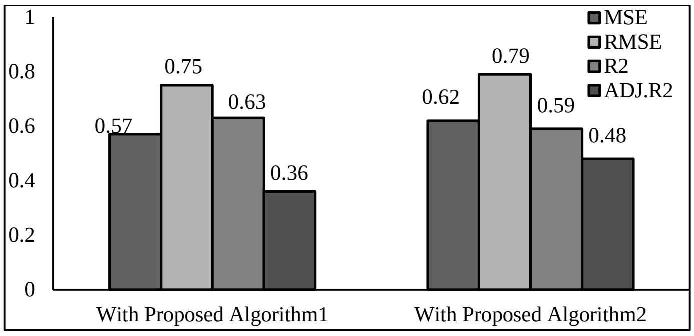

Fig.8. DTR performance comparison with proposed approach 1 and 2

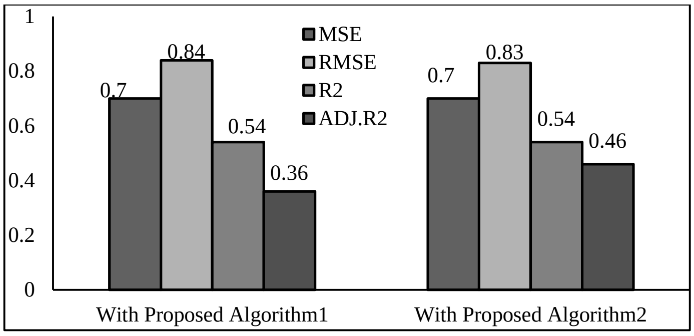

Fig.9. SVR performance comparison with proposed approach 1 and 2

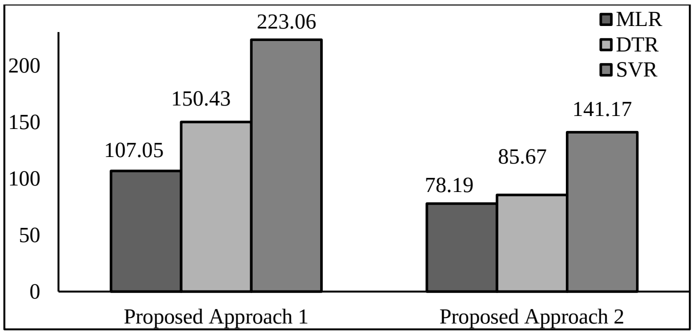

Fig.10. Execution time performance analysis

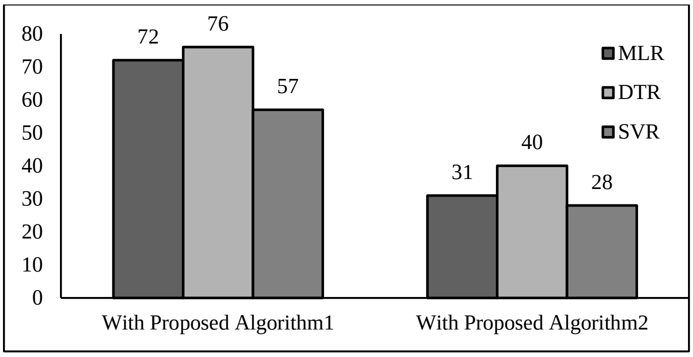

Fig.11. Number of features selected by both algorithms

The Table.17 shows execution time in seconds of both proposed algorithms for similar population and iteration size considered. Hybrid filter-wrapper approach's aims to reduce complexity in terms of running time of proposed approach 1. Above results show that hybrid approach was successful enough in reducing time as compared to proposed approach 1. Thus, it can be derived that both proposed approaches are impactful enough in prediction of IGR on various parameters considered. The Fig.7-Fig.11 compare performances of hybrid approaches with proposed approach 1.

Careful investigation and analysis of selected features are fundamental to do for this study. Considering Decision tree regression combined with Hybrid filter wrapper approach, 40 features are selected for Institute Graduation Rate Prediction with minimum error. It is important to analyze from which category these features belong to. The Fig.12 shows category or academic survey components effective in prediction of IGR.

Below distribution shows that for institutes to have good graduation rate prediction, they should focus on Institute characteristics with prime focus followed by human resource and student financial aid components in the institute. The Table.18 shows features selected for institute graduation rate prediction from various survey components. It also describes feature name from the original IPEDS dataset and every selected feature's description. Output shows that features related to institute characteristics have maximum influence in IGR prediction followed by Human resource, graduation rate and student financial aid.

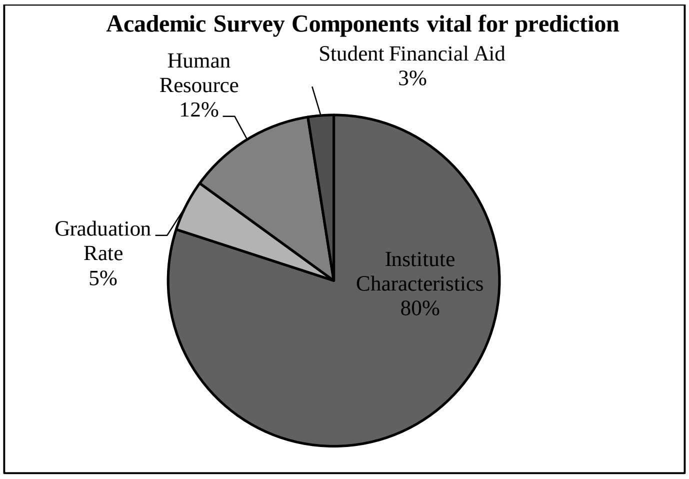

Fig.12. Features from survey components' contribution in prediction

Table.18. Features selected for IGR Prediction by Hybrid Approach

Feature Selected	**Feature description**	**Survey Component of feature selected**
Chg1ay3	Published in-district tuition and fees	Institute Characteristics
Chg5ay3	On campus, room and board	
Chg9ay3	Off campus (with family), other expenses	
ADMSSN	Admissions total	
PUBPRIME	Primary public control	
PUBSECON	Secondary public control	
ENRLFTW	Enrolled full time women	
ENRLPTW	Enrolled part time women	
SATPCT	Percent of first-time degree/certificate-seeking students submitting SAT scores	
SLO3	Distance learning opportunities	
SLO52	ROTC: Navy	
SLO7	Weekend/evening college	
SLO53	ROTC: Air Force	
STUSRV1	Remedial services	
STUSRV2	Academic/career counselling service	
STUSRV4	Placement services for completers	
STUSRV8	On-campus day care for students' children	
LIBAC	Library facilities at institution	
ASSOC1	Member of National Collegiate Athletic Association (NCAA)	Institutional Characteristics
ASSOC5	Member of National Christian College Athletic Association (NCCAA)	
SPORT1	NCAA/NAIA member for football	
CONFNO1	NCAA/NAIA conference number football	
SPORT2	NCAA/NAIA member for basketball	
CONFNO2	NCAA/NAIA member for basketball	
SPORT3	NCAA/NAIA member for baseball	
BOARDAMT	Typical board charge for academic year	
ENRLPT	Enrolled part time total	
PT_FP	Part-time first-professional students are enrolled	
ALLONCAM	Full-time, first-time degree/certificate-seeking students required to live on campus	
CINDON	Total price for in-district students living on campus	
CINSFAM	Total price for in-state students living off campus (with family)	
COTSFAM	Total price for out-of-state students living off campus (with family)	
GRTOTLT	Grant total	Graduation rates
GRTOTLM	Total men	
SGRNT_N	Number receiving state/local grant aid	Student financial aid
FTPT	Full and part-time status (Employees)	Human Resource
FSTAT	Faculty Status	
OPRFSTF	Other professional FTE staff	
BENTYPE	Fringe benefits	
EMPCNTT	Number of full time instructional staff total	

The Table.18 highlights the point that institutes should pay more attention on giving special learning opportunities (SLO) to students. Institutes should also not ignore expenses those students have to bear. Library, distance learning facilities and placement services are also important. On campus day care for children of students could help students focus better on their studies. Institutes should also provide sports infrastructure and environment for learners. Financial aid and human resources components such as number of full and part time staff, fringe benefits provided to staff also influence IGR prediction.

5.8 PROPOSED ALGORITHMS' PERFORMANCE ON UCI DATASET

The proposed Algorithms' performance was also measured with bench mark dataset. For this, UCI [40] machine learning repository's Communities and crime, un-normalized dataset was considered.

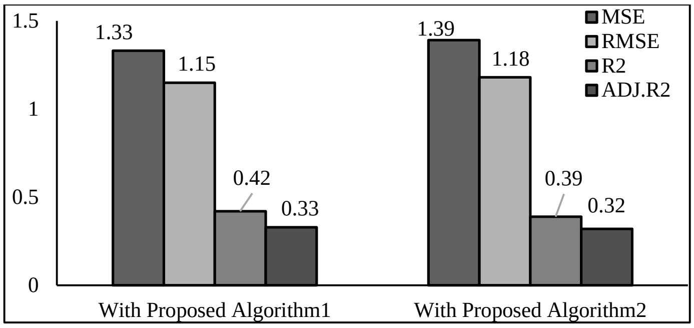

Fig.13. MLR performance with proposed approach 1 and 2 on UCI Dataset

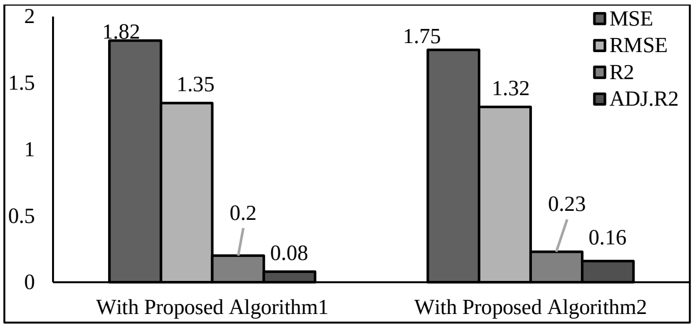

Fig.14. DTR performance with proposed approach 1 and 2 on UCI Dataset

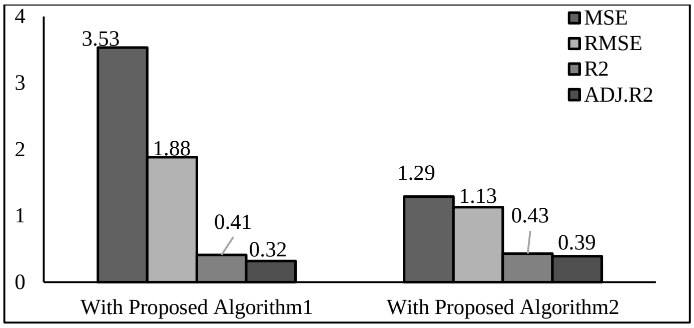

Fig.15. SVR performance with proposed approach 1 and 2 on UCI Dataset

The proposed algorithm was applied on the dataset with identical steps and experimental setup. Number of iterations and population size were fixed to 10, 20 and 30 respectively.

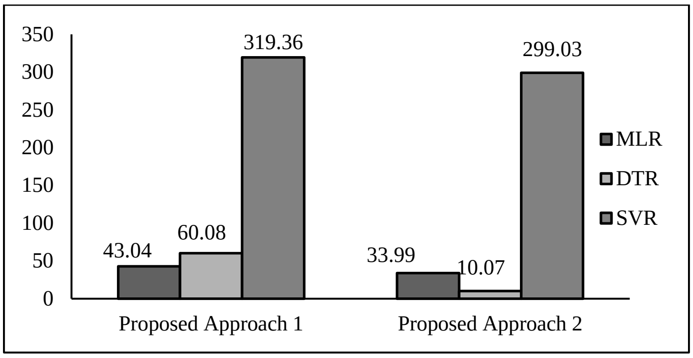

Fig.16. Execution Time Performance of proposed approach 1 and 2 on UCI dataset

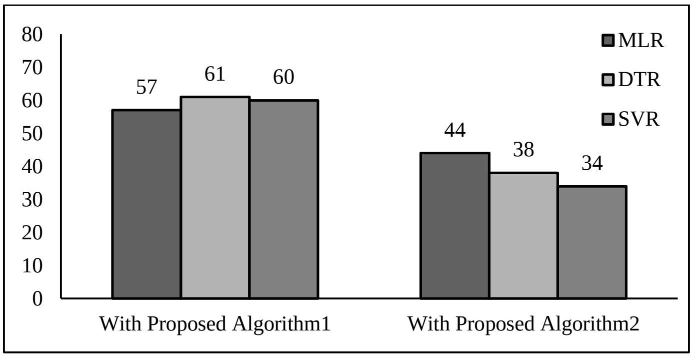

Fig.17. Number of Features selected by proposed approach 1 and 2 on UCI dataset

6. CONCLUSION

Data mining is a much researched area but in education domain still there is a scope of its real applicability to create actionable knowledge. To realize power of data mining, in this paper, feature selection is explored in evolutionary way. Aim of this study is to determine effectiveness of GA in field of education for finding optimal parameters with the least prediction error. Proposed algorithms prove that evolutionary algorithm can play noteworthy role in education data mining. Decision tree regression performs better than two other regression techniques, support vector and multiple linear regressions in combination with proposed algorithms. To improve quality of education and to aid education domain, such novel techniques should be considered. This is the first effort of its kind which proposes an innovative algorithmic framework to improve institute graduation rate prediction in the field of education data mining. Proposed algorithms are also applied on UCI dataset. Results prove that proposed algorithms are effective in yielding minimum error with reduction in number of features and execution time. Proposed algorithms could be generalized on other domains to get effective results as well. In future, proposed algorithms' would be further enhanced by doing effective parameter tuning. Other evolutionary algorithms' core ideas can also be merged with genetic algorithm to get better results.

REFERENCES

[1] B.L. Deekshatulu and Priti Chandra, “Classification of Heart Disease using K- Nearest Neighbor and Genetic Algorithm”, *Proceedings of International Conference on Computational Intelligence: Modelling Techniques and Applications*, pp. 85-93, 2013.

[2] T. Santhanam and M.S Padmavathi, “Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis”, *Procedia Computer Science*, Vol. 47, pp. 76-83, 2015.

[3] Jingyuan Dai, “A Novel Medical Assistance System Based on Data Mining”, *Proceedings of IEEE Workshop on Advanced Research and Technology in Industry Applications*, pp.264-273, 2014.

[4] Antonis Lambrou, Harris Papadopoulos, and Alex Gammerman, “Reliable Confidence Measures for Medical Diagnosis with Evolutionary Algorithms”, *IEEE Transaction on Information Technology in Biomedicine*, Vol. 15, No. 1, pp. 1-23, 2011.

[5] P. Johnson, “Genetic Algorithm with Logistic Regression for Prediction of Progression to Alzheimer’s Disease”, *BMC Bioinformatics*, Vol. 15, No. 16, pp. 1-14, 2014.

[6] N. Devasenathipathi and Nilesh Modi, “Applying GA to Improve Students’ Academic Performance by Group Formation”, *International Journal of Data Warehousing and Mining*, Vol. 1, No. 2, pp. 142-146, 2011.

[7] N. Devasenathipathi and Nilesh Modi, “Evolutionary Algorithm Approach to Pupils’ Pedantic Accomplishment”, *Proceedings of International Conference on Frontiers of Intelligent Computing: Theory and Applications*, pp. 415-423, 2013.

[8] N. Devasenathipathi and Nilesh Modi, “Contemplating Crossover Operators of Genetic Algorithm for Student Group Formation Problem”, *International Journal of Emerging Technology and Advanced Engineering*, Vol. 2, No. 2, pp. 192-197, 2012.

[9] Behrouz Minaei-Bidgoli and William F. Punch, “Using Genetic Algorithm for Data Mining Optimization in an Educational Web-based System”, *Proceedings of International Conference on Genetic and Evolutionary Computation*, pp. 2252-2258, 2003.

[10] Fateme Moslehi and Abdorrahman Haeri, “A Novel Hybrid Wrapper–Filter”, *Ambient Intelligence and Humanized Computing*, Vol. 11, No. 3, pp.1105-1127, 2020.

[11] N. Shelke and S. Gadage, “A Survey of Data Mining Approaches in Performance Analysis and Evaluation”, *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 5, No. 2, pp. 456-459, 2015.

[12] Q. Hung Do and Jeng Fung Chen, “A Neuro-Fuzzy Approach in the Classification of Students’ Academic Performance”, *Computational Intelligence and Neuroscience*, Vol. 2013, pp. 1-8, 2013.

[13] K. Barker, T.B. Trafalis and T. Reed, “Learning from Student Data”, *Proceedings of International Conference on Systems and Information Engineering Design*, pp. 79-84, 2004.

[14] C. Romero and S. Ventura, “Data Mining in Education”, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 3, No. 1, pp. 12-27, 2013.

[15] A. Merceron and K. Yacef, “Education Data Mining: A Case Study”, *Proceedings of Conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology*, pp. 467-474, 2005.

[16] N. Thai-Nighe, Paul Janecek and P. Haddawy, “A Comparative Analysis of Techniques for Predicting Academic Performance”, *Proceedings of International Conference on Frontiers in Education*, pp. 1-7, 2007.

[17] S. Huang and N. Fang, “Predicting Student Academic Performance in an Engineering Dynamics Course: A Comparison of Four Types Predictive Mathematical Models”, *Computers and Education*, Vol. 61, No. 1, pp. 133-145, 2013.

[18] P. Thakar, A. Mehta and Manisha, “Performance Analysis and Prediction in Education Data Mining: A Research Travelogue”, *International Journal of Computer Applications*, Vol. 110, No. 15, pp. 60-68, 2015.

[19] B. Bhardwaj and S. Pal, “Data Mining: A Prediction for Performance Improvement using Classification”, *International Journal of Computer Science and Information Security*, Vol. 9, No. 4, pp. 1-5, 2011.

[20] G. Rao and D.L.S. Reddy, “An Analysis of Education Data Mining in Advanced Education System”, *International Journal of Science and Research*, Vol. 4, No. 12, pp. 2149-2153, 2015.

[21] D. Fatima, S. Fatima and A.V. Krishna Prasad, “A Survey on Research Work in EDM”, *IOSR Journal of Computer Engineering*, Vol. 17, No. 2, pp. 43-49, 2015.

[22] E. Jormanainen, “A Frame Work for Research on Technology-Enhanced Special Education”, *Proceedings of 7th IEEE International Conference on Advanced Learning Technologies*, pp. 1-9, 2007.

[23] J. Kumar, “A Comprehensive Study of Education Data Mining”, *International Journal of Electrical Electronics and Computer Science Engineering*, Vol. 4, No. 2, pp. 58-63, 2015.

[24] E. Taherifar and T. Banirostam, “Assessment of Student Feedback from the Training Course and Instructor Performance through the Combination of Clustering Methods and Decision Tree Algorithms” *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 6, No. 2, pp. 1-5, 2016.

[25] E. Balraj and D. Malini, “A Survey on Predicting Student Dropout Analysis using Data Mining Algorithms”, *International Journal of Pure and Applied Mathematics*, Vol. 18, No. 8, pp. 621-626, 2018.

[26] A. Mayra and D. Mauricio, “Factors to Predict Dropout at the Universities: A Case of Study in Ecuador”, *Proceedings of IEEE International Conference on Global Engineering Education*, pp. 1238-1243, 2018.

[27] Tao Zhang, “Predicting the Performance Fluctuation of Students based on Long-Term and Short-Term Data”, *Proceedings of International Conference of Educational Innovation Through Technology*, pp. 126-134, 2017.

[28] A. Daud, “Predicting Student Performance using Advanced Learning Analytics”, *Proceedings of International*

Conference on World Wide Web Companion, pp. 415-423, 2017.

[29] M. Bucos and B. Dragulescu, "Predicting Student Success using Data Generated in Traditional Educational Environments", *TEM Journal*, Vol. 7, No. 3, pp. 617-625, 2018.

[30] Ali Buldu and Kerem Ucgun, "Data Mining Application on Students' Data", *Procedia-Social and Behavioral Sciences*, Vol. 2, No. 2, pp. 5251-5259, 2010.

[31] Dorina Kabakchieva, "Predicting Student Performance by using Data Mining Methods for Classification", *Cybernetics and Information Technologies*, Vol. 13, No. 1, pp. 61-72, 2013.

[32] C.J. Carmona, P. Gonzalez, M.J. del Jesus, C. Romero and S. Ventura, "Evolutionary Algorithms for Subgroup Discovery Applied to E-Learning Data", *Proceedings of IEEE International Conference on Education Engineering*, pp. 1-7, 2010.

[33] Cristobal Romero and Sebastian Ventura, "Education Data Mining: A Review of the State of the Art", *IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews*, Vol. 40, No. 6, pp. 1-21, 2010.

[34] C. Romero and S. Ventura, "Educational Data Mining: A Survey from 1995 to 2005", *Expert Systems with Applications*, Vol. 33, No. 1, pp. 135-146, 2007.

[35] NCES, "National Center for Education Statistics", Available at: https://nces.ed.gov/ipeds/trendgenerator/, Accessed at 2019.

[36] Brenda L. Bailey, "Let the Data Talk: Developing Models to Explain IPEDS Graduation Rates' Data", *Wiley Inter Science-Special Issue: Data Mining in Action: Case Studies of Enrollment Management*, Vol. 2006, No. 131, pp. 101-115, 2006.

[37] Angela E. Henderson and William F. Punch, "Predicting U.S. News and World Report Ranking of Regional Universities in the South using Public Data", Ph.D. Dissertation, School of Education, Colorado State University, pp. 1-126, 2017.

[38] Abby Miller, Sue Clery and Amy Topper, "Assessing the Capacity of IPEDS to Collect Transfer Student Data", Project Report, NPEC, Coffey Consulting, 2018.

[39] Joshua Lee Whitlock, "Using Data Science and Predictive Analytics to Understand 4-Year University Student Churn", Ph.D. Dissertation, Department of Educational Leadership and Policy Analysis, East Tennessee State University, pp. 1-157, 2018.

[40] UCI, "University of California Irvine-Machine Learning Repository". Available at: https://archive.ics.uci.edu/ml/index.php, Accessed at 2018.