Student-Performulator: Predicting Students' Academic Performance at Secondary and Intermediate Level Using Machine Learning

Shah Hussain[1] · Muhammad Qasim Khan[1]

Abstract

Forecasting academic performance of student has been a substantial research inquest in the Educational Data-Mining that utilizes Machine-learning (ML) procedures to probe the data of educational setups. Quantifying student academic performance is challenging because academic performance of students hinges on several factors. The in hand research work focuses on students' grade and marks prediction utilizing supervised ML approaches. The data-set utilized in this research work has been obtained from the Board of Intermediate & Secondary Education (B.I.S.E) Peshawar, Khyber Pakhtunkhwa. There are 7 areas in BISEP i.e., Peshawar, FR-Peshawar, Charsadda, Khyber, Mohmand and Upper and Lower Chitral. This paper aims to examine the quality of education that is closely related to the aims of sustainability. The system has created an abundance of data which needs to be properly analyzed so that most useful information should be obtained for planning and future development. Grade and marks forecasting of students with their historical educational record is a renowned and valuable application in the EDM. It becomes an incredible information source that could be utilized in various ways to enhance the standard of education nationwide. Relevant research study reveals that numerous methods for academic performance forecasting are built to carryout improvements in administrative and teaching staff of academic organizations. In the put forwarded approach, the acquired data-set is pre-processed to purify the data quality, the labeled academic historical data of student (30 optimum attributes) is utilized to train regression model and DT-classifier. The regression will forecast marks, while grade will be forecasted by classification system, eventually analyzed the results obtained by the models. The results obtained show that machine learning technology is efficient and relevant for predicting students performance.

Keywords Supervised learning · Educational data-mining (EDM) · Machine-learning (ML) · Students' performance prediction · Learning analytics

---

Extended author information available on the last page of the article

1 Introduction

Quality Education plays an essential role in the Sustainable-Development Goals (17-SDGs), endorsed by United Nations [1]. A crucial aspect to remember while working on sustainable development goals, to provide equal opportunities and sharing it equally. Students' dissertation in attaining higher education is a serious matter, which need to be evaluated globally. The drop-out students' ratio from academic institutions causes a loss/resource wastage, which is significant and costly in the educational settings, and affects the evaluation and assessment processes of the academic organizations. The declines in the engineering programs are higher than in all other science and art disciplines, as the study shows [2]. In our aimed study, we will gain the Data Science goals to gain insights from any sort of data. As the Data Science involves developing techniques of storing, recording and analyzing data to extract useful information efficiently. Forecasting analysis of secondary and intermediate students is performed, the marks and grades prediction framework will promote and enhance the standard of education and concentrate on the particular area wherever students do not earn satisfying grades. Education should be given priority in improving our organizations. Administrative and teaching staff should improve their productivity by recognizing curriculum development and skills, which will give students an improved opportunity for learning [3]. In this context, secondary and intermediate-level institutions must also focus on developing and enhancing education models by integrating information and communication technologies (ICT), which can act as an instrument for fostering social accountability and equal opportunities. With this view, the results of ICT in education systems are important as they can make important contributions to the process of learning and teaching as well as to encouraging knowledge building [4]. The use and implementation of ICT application involved in teaching/learning is also known as enhanced learning relying on technology. The term enhanced learning, based on Technology defines the usage of digital technologies to boost learning experience. The utilization of emerging technologies has enhanced learning, it can help to improve critical thinking among students [5]. Technology-based enhancement in learning involves several leading-edge technologies comprises smartphone learning applications, learning-management frameworks, cloud-based learning tools, web-based applications of social networking and social media for education, visual-aids, machine-learning(ML) and DM etc., [6].

In accordance with the impacts of learning-teaching on the sustainability of intermediate & secondary education and tech-boosted learning [5], We must carefully describe the basic information technology requirements that will serve us instead of being a hitch in learning and teaching. For instance, the preparation of teaching and managerial personnel's for the production of predictive analytic skills as it is crucial for measuring the latent outcomes of the computer-aided framework usage [7]. In to discussed technologies earlier, that is executed with a greater impression in the academic set-ups to produce a large volume of data and save it in means that it could be efficiently presented ubiquitously [8]. The data

size can exceed the processing volume sometimes, storing and evaluating it with conventional methods. New technologies should be considered in order to perform data analysis such as data mining, intelligent systems, association rules mining, optimization based data mining [9] and big data. The bunch of these novel technologies will enable simple and effective analysis of educational-data, and can be utilized to transform the educational-data in a new shape which could be more beneficial [10–12].

With deep learning mining of educational data is a growing field for research that enable us to analyze and process the educational information collected from different roots [13]. For analyzing educational data, several statistical methods, data-mining, visualization and ML gears are utilized. The study analytics generated from academic-data intends to investigate obtained data from the institutional databases. Learning-management frameworks interprets the information, improve learning procedures and atmosphere in which the data befalls [14].

The related research includes various studies [8, 14–17] who have developed different classification-based data-mining systems to forecast the final grades of students. The following methods are the most representative of classifications techniques: Data analytics and Visualization; Instructor support system based on Feedback; A Recommendations system for Students; Social Media Analytics system, Student Modeling, and Marks and Grade forecasting Students Performance Assessment; In our aim system, we intended to make a framework in which student marks and grade can be predicted, in educational data mining such framework recognized as student performance prediction system. It has many applications nowadays and grows in reputation because of reliability and performance. The aim of the proposed framework is to estimate and forecast an unknown thing (Grade and marks), the nature of outcomes generate by the forecasting framework will be reliant on data and resilience of the regression model or the classifier. The data utilized in our study comprises several attributes and simultaneously dual methods of supervised-learning will work in synchrony, one will forecast the marks whilst another will forecast the grade [18].

According to these concepts, marks and grades of the student's will be forecasted, the forecast and performance assessment will be reliant on previous educational history of the student. Based on forecasted results' the student with week grade and low marks will be recognized, appropriate care and devotion to the weak students will be given in order to achieve good performance in examinations. Advanced advice can be served via the decision-making system to students' who are at risk and do not attain satisfying marks and grades. This predictions framework will enable education department to sum up the annual examination of secondary & intermediate students before the examination takes place. This will also be used to forecast/predict the number of students who are graduating or failing in the ongoing uncertain situation of Covid-19 [19, 20].

The present research is evaluated according to the case-studies utilized to determine the efficacy of our proposed approach. The department of education will attain a validated results reliant on machine learning. Every year 180,000 to 200,000 students sit for the BISEP exam, the secondary exam (Class 9 and Class 10), is taken in two steps likewise the intermediate examination (class 11 and class 12) is taken in

two years. Every year there are four class examinations take place which produces a huge amount of data, It's a big data challenge [21]. The typical approaches used to build a small data predictive model cannot be used for clutching, pre-processing, feature extraction, regression and classification. Big-data technology empowers us to cope with the challenge [22]. The data of the students must be accurate, complete, and updated to make marks and grade forecasting framework.

This research study is structured as: Segment-2.o addresses the associated study, methodologies, experimental evaluations and the contribution of the authors of the specific work. In segment-3 proposed methodology is outlined. Segment-3.1 approaches the processing of data-collection and its pre-processing steps. Segment-3.2 and 3.3 explains ML algorithms and selection techniques for attributes/ features. Similarly in segment-4 comprehensive experimental techniques, result analytics, and visualization of results are outlined. Lastly, the conclusion and work are summed-up.

2 Related Work

The educational data-mining has been researched extensively in the past and remains a hot area of research in data-mining (DM), machine-learning (ML), deep-learning, and big-data. The aim of various forms of study is the development of a predictive framework, which will forecast the marks, grades, institutional ranking and institutional recommendations. Different tools and techniques are used to analyze and visualize the data. Below are some leading-edge work that helped us to explain our proposed methodology.

The research explains the use of large-data application in education [23], Big-data methods for learning analysis are used in different ways e.g., system performance prediction, visualization of data, student skills estimation, risk detection, fraud detection, system of course recommendation, grouping of students and collaboration with other students. The predictive analysis of this study focuses on student achievements, behavior, and skill prediction enhances the usefulness of this work.

During the forecast of student performance [24] data-mining procedures were utilized to build a predictor framework for the final-marks reliant on students' achievements. A key element regression-model trained, and used to forecast the academic achievement of students. As features, non-courses variables, for instance out-of-class student conduct, who notes, focus in video watching and post-school tutoring were used.

The factors which influence the validity of application are examined in a study [25] while working on EDM, two sort methods for data-analysis are used, that is, descriptive and predictive models. Descriptive model utilizes unsupervised-learning techniques to explain and recognize the structure of mined-data while Predictive model utilizes supervised-learning techniques which determine unknown-values.

The objective of a study [26] was to evaluate students at the beginning of an academic session and to forecast achievements through academic history, using collaborative filtering technique. The authorized courses reflecting the student's learning are chosen. The information system gathered historical data to identify similar

characteristics for students at the University of Masaryk. The research analysis indicates that this approach is efficient as a SVM-classifier.

In a study [27] the author has approached a novel technique by utilizing historical academic data of students as an input, in order to evaluate students' performance. The study was relying on a factorization of low-range-matrixes and dispersed linear model. The data-set acquired from the University, about 12 years academic data was evaluated. The system proposed has improved the accuracy of grade prediction. In the study[28], the author suggested a new method that mines educational-data utilizing a recommendation-system, which is uniquely build to forecast performance and validated through comparing to another leading-edge regression-model, suchlike logistic and linear regression.

Data from different secondary schools in the Kanchi district were acquired in the paper [29]. To develop a student classification system two leading-edge classifiers i.e. naïve Bayes and decision tree were used, the occupation of parents was also included as features, which is played a significant role in the data-set to enhance the correctness of the grade-prediction framework. The classifier of the DT achieves better than the naïve Bayes Classification in accuracy terms.

Big-data technology scenarios and deployments for EDM have been analyzed in depth. The paper reveals the connection between big-data and educational data. The suggested technique [30] emphases on methods, strategies and algorithms' of big-data utilized in educational setting to make learning and teaching processes easier and fairer for the future. The study also reveals the advantages and significance of large/complex-data in EDM.

Using big-data, the suggested approach [31] provides a smart recommendation framework. The mining of association-rule is an unsupervised approach applied to uncover the connection linking the academic activities of students. Rules are mind applying an algorithm for rule-mining and the conduct of students is utilized to index the subjects. Spark and Hadoop are eventually used to build a recommendation framework. The results obtained show that the proposed recommendation framework is successful.

In a study [32] the author has proposed a new intelligent machine learning framework for predicting the result of the game played by targeting to discover the influential features set that affects the results using historical data. Several ML techniques that utilize various learning structures to derive the models, including Decision Tree, artificial neural network, and Naïve Bayes are chosen. Based on the results analysis, the DRB (defensive rebounds) feature was selected and was considered as the most weighty factor influencing the outcomes of the game played.

3 Methodology

Our proposed system for marks and grade forecasting is explained (See Fig. 1). Pre-processing techniques are utilized to process the data, manage instances with missing values, and pull out attributes that comprise student personal info, removal of redundant-data etc. (See Table 1). Machine-learning (ML) approaches are used to build a prediction framework of students' performance. Regression model and
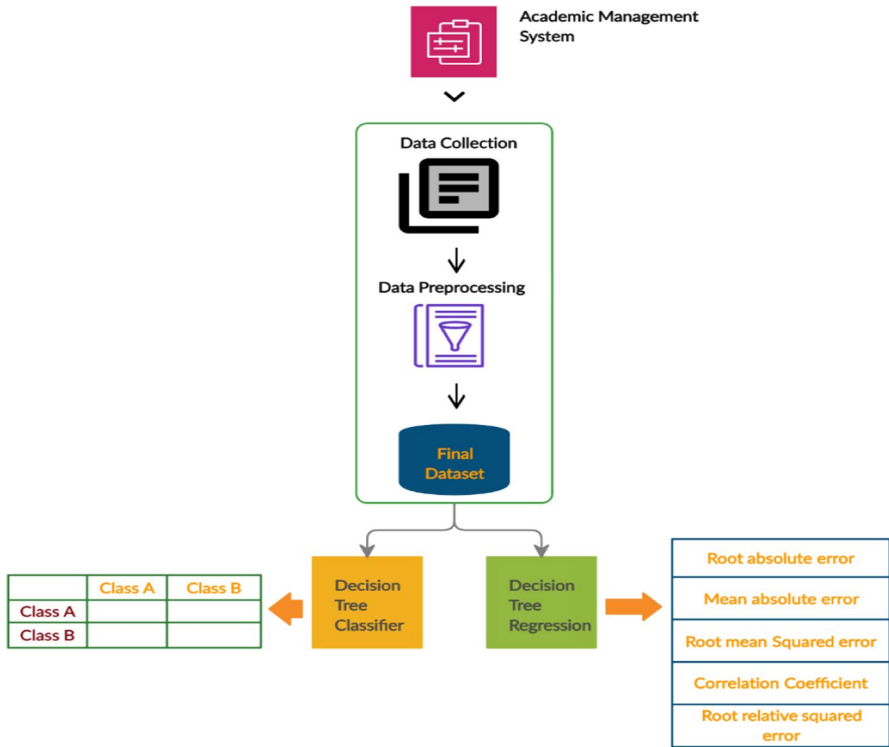
Fig. 1 Proposed framework

classification-algorithms are utilized in order to build a framework for marks/grade prediction. The techniques of machine learning are chosen.

3.1 Data-set Acquisition

The data-set utilized in the proposed system is acquired from the Khyber Pakhtunkhwa Board of Intermediate & Secondary Education Peshawar. The data-set composes intermediate and secondary students' academic records, the secondary school examination comprising 9th and 10th classes whereas the intermediate examination comprising 11th and 12th classes. In Khyber Pakhtunkwha there are seven zones within the jurisdiction of BISEP, Every BISEP student must belong to the region, i.e., Peshawar, FR-Peshawar, Charsadda, Khyber, Mohmand Upper and Lower Chitral. The data-set comprises the following attributes:

The attributes contains students' confidential information and inapplicable attributes i.e. Student Address, Cell Number, Religion, Roll_No; Registration_No, Institute_Code etc. are ousted from the data-set. This acquired data-set comprises 90,000 historical data of students'. The data comprises Obtain_marks and Grade, so using

EDM two supervised-learning models can be utilized to forecast the students' marks and grades.

3.2 Machine-Learning Techniques Selection

3.2.1 Attributes Selection based on Genetic Algorithm

The GA is a leading-edge optimizing technique[33] that has been broadly utilized for the DM study, the GAs model is a novel search approach also called an evolutionary search technique.When it comes to predicting student performance, we have over 100 attributes which could impact the training accurateness as well increases the inefficiency of the classification framework. A random uniform population is created in our proposed framework at the first phase, for each generation, the probability mutation and crossover represent 0.3. The model begins and generate a random population, the chromosomes comprises several genes, and each genes holds an actual-number. The chromosome portrayal is presented in the following equation

$$C = \{\sim ki| \sim ki \in [0, 1]\}n \quad i = 1$$

### 3.2.2 Pseudo-code of Proposed Features Selection Method Utilizing GA

$C,T,fbest \leftarrow \varnothing \rightarrow$ Initialize.
   $Po \leftarrow$ Gauss random distribution with $\sigma = 0.15$ and $\mu = 0.3$
   Each genes is converted to binary discrete.
   Such that $\begin{Bmatrix} 0, & pi < 0.5 \\ 1, & otherwise \end{Bmatrix}$.
   while $t \leq T$ **do.**
   $t \leftarrow t+1$
   GA (Pt).
   if $argmaxt (Pt) \geq$ then $fbest \leftarrow argmaxt(Pt)$.
   end while
   return $best\ p \rightarrow$ Return the best individual.
   See Fig. 2.

3.3 Performance Calculation of Regression and Classification

Leading-edge technique i.e. Cross-validation is utilized for estimation of machine-learning This method can easily be understood and implemented for machine learning to optimize parameters and to choose the best model. To test the classifier and regression three leading-edge cross-validation techniques are used. In our experimental study, Ķ-Fold Cross-Validation techniques are used.

Fig. 2 State diagram of GA training
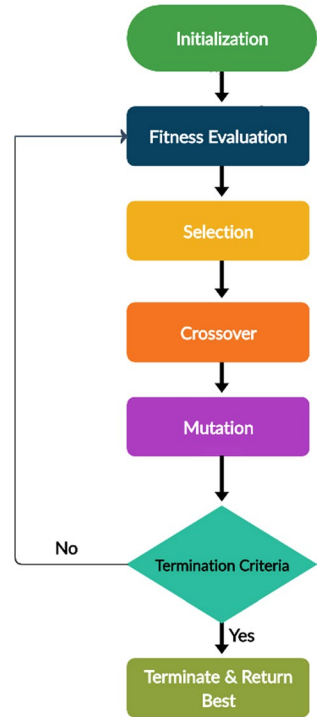
4 Results

In hand educational grade prediction framework the data-set comprises private, regular, arts and science group students. The data-set (See Table 1) comprises about 126 attributes, certain attributes contained confidential info of students' and inapplicable towards classification has been drop-out. The regression method is utilized for students' marks prediction, the attributes of "HSSC:II" have chosen as predictor, (See Table 5) and rest of as features. Likewise the classification method is utilized for students' grades, prediction, in the data-set the grade attribute is chosen as class label (See Table 4), and rest of them chosen as features. The data-set is described below (See Table 2).

The data-set comprises SSC:I, SSC:II, HSSC:I and HSSC:II marks. Assume a student was examined in the year 2012 with SSC:I, in 2013 with SSC:II, in 2014 with HSSC:I and in 2015 with HSSC:II. The proposed framework will forecast the marks and grades of HSSC:II students, reliant on the results over the last three years (SSC 1,2 and HSSC 1). The Table 2 contains the optimum attributes those are auto-selected via GA. Table 3 shows the information of marks distribution and grading scheme for SSC and HSSC examination.

The attributes of the data-set are grouped in Tables 4 and 5 utilizing optimized attributes to form a features vector and label vector. The grade prediction labels are the grades obtained earlier. Likewise, for marks prediction the predictors are the marks obtained in the past exam.

Table 1 Attributes in the data-set

Dbo	Enroll-No	Exam_code	Exam	Session	Year
Remarks	Status	Sex	Pass/Fail	Marks	Grade
E-I	Area	SSC1	SSC2	HSSC1	HSSC2
IC-I	E-II	U-I	U-II	AR-I	AR-II
PH-I	IC-II	M-I	M-II	CH-I	CH-II
ECW-I	PH-II	PER-II	BIO-II	CS-I	CS-II
PS	ECW-II	ETH-I	ETH-II	IS-I	IS-II
CIV-I	PA-I	PA-II	ART-I	ART-II	Pak.C
EDU-I	CIV-II	PLM-I	PLM-II	HHA-I	HHA-II
GEO-II	EDU-II	IH-I	IH-II	HIP-I	GEO-I
GM-II	EHE	MBH-I	MBH-II	P/H	GM-I
HPE-II	UL-I	UL-II	EL-I	EL-II	HPE-I
PSY-I	EHE-I	EHE-II	ECO-I	F&N	ECO-II
PHLO-I	PSY-II	ST-I	ST-II	OHECO-I	OHECO-II
SOC-I	PHLO-II	LS-I	LS-II	EAdv:-I	EAdv:-II
PER-1	CSI	Arch-1	GS-I	GS-II	AGRi-II
OPTH-1	BIO-I	UA-1	SRG-II	RGR-I	CHD&FL-I
APL:Sc-I	OPTH-II	PHYSOT-I	PHYSOT-II	SRG-I	UA-II
M.BIO-I	APL:Sc-II	BMS-I	BMS-II	DNT-I	DNT-II
GTD-I	M.BIO-II	Chem.Pth	IH&C	Arch-II	AGRi-I
SOC-II	GTD-II	RGR-II	CHD&FL-II	Clt&Txt-I	Clt&Txt-II

Table 2 Optimal attributes

PS	CS-I	IH-I	IH-II	M-I
CS-II	AR-II	AR-I	Grade	CH-II
CIV-I	CIV-II	SSC-I	M-II	ECO-I
ECO-II	ST-II	SSC-II	CH-I	MBH-I
EDU-I	EDU-II	IS-II	HSSC-I	MBH-II
IS-I	ST-I	EL-II	HSSC-II	EL-II

4.1 Attributes Optimization Using Genetic Algorithm

The leading-edge Genetic Algorithm optimization method, "Tournament Selection" that picks one or more tokens from thousands of tokens using GA. Herein many Tournaments is executed, the tokens (chromes) selection of tokens from a population can be random. The winner of each tournament is one who suits best and is chosen as the winner for a crossover. If the tournament size is larger, there is less possibility of selecting weak words in vocabulary.

Table 3 Grade, marks, and percentage distribution for SSC/HSSC examination

Class	Marks ratio	Marks percentage	Grade	Division
SSC	840 And Above	80	A-ONE	1st
	735 To 840	70	A	1st
	630 To 735	60	B	1st
	525 To 630	50	C	2nd
	420 To 525	40	D	3rd
	419 And Below	39	E	Fail
HSSC	880 And Above	80	A-ONE	1st
	770 To 880	70	A	1st
	660 To 770	60	B	1st
	550 To 660	50	C	2nd
	440 To 550	40	D	3rd
	440 And Below	39	E	Fail

Table 4 For grade prediction, allocation of the data-set into features-set and labels

SSC:I (Grade)	SSC:II (Grade)	HSSC:I (Grade)	HSSC:II (Actual Grade)	HSSC:II (Predicted Grade)
2012-Annual	2013-Annual	2014-Annual	2015-Annual	2015-Annual
2013-Annual	2014-Annual	2015-Annual	2016-Annual	2016-Annual
2014-Annual	2015-Annual	2016-Annual	2017-Annual	2017-Annual
2015-Annual	2016-Annual	2017-Annual	2018-Annual	2018-Annual
2016-Annual	2017-Annual	2018-Annual	2019-Annual	2019-Annual

Table 5 For Marks Prediction, Allocation of the data-set into features-set and labels

SSC:I (Marks)	SSC:II (Marks)	HSSC:I (Marks)	HSSC:II (Marks Actual)	HSSC:II (Predicted Marks)
2012-Annual	2013-Annual	2014-Annual	2015-Annual	2015-Annual
2013-Annual	2014-Annual	2015-Annual	2016-Annual	2016-Annual
2014-Annual	2015-Annual	2016-Annual	2017-Annual	2017-Annual
2015-Annual	2016-Annual	2017-Annual	2018-Annual	2018-Annual
2016-Annual	2017-Annual	2018-Annual	2019-Annual	2019-Annual

The parameters of Genetic algorithm are shown in the Table 6. The Table 7 consists of information about the total attributes of the data-set and the selected optimal attributes that are used in our experimental work for grade and marks predication system.

Table 6 Selection of optimal attributes, Genetic Algorithm parameters

Parameters	Values/Types
Initializing operator	Random
Size of population	20
Operator replacement	Generational
Frequency report	20
Seed	1
Operator selection	Tournament Selection
Probability of cross-over	0.6
Generations	20
Mutation probability	Bit Flip
Mutation probability	1.1

Table 7 Genetic Algorithm selected attributes

Total attributes for marks prediction	126
Total attributes for grade classification	126
Optimal-attributes chosen for marks prediction	30
Optimal-attributes chosen for classification of grade	30

4.2 Decision-Tree Results

The DT is used for the actual data-set comprising of 126 attributes as well as on the GA optimized data-set comprising 30 attributes (See Table 7). The optimized attributes helps in making a tiny size tree which minimizes the set of rules generated by DT. A Ķ-Fold Cross-validation method evaluates the regression and classification models by choosing $K = 10$ value. All the instances in the data-set take part in training and testing stages during k-fold cross-validation, the Ķ-Fold approach is deemed the most effective tool in the efficiency valuation of the trained model amongst other cross-validation methods.

The aforementioned Table 8 shows the output of DT classifier utilizing Ķ-Fold Cross-validation technique. If $K = 10$, the data is split in ten test-set and train-set. Entire data-set instances are involved in testing and training stage. DT-classifier utilizing 10-folds cross-validation yielded 94.39% average accuracy.

GA attribute selector is utilized to catch the optimum attributes in the data-set. Genetic Algorithm approach ranks the attributes i.e. high or low, the low ones attributes are expelled from the data-set. For outcome evaluation and classifier training

Table 8 Confusion-Matrix of DT classifier on 10-Ҟ Cross-validation

Instances Classified Correctly 299,000	94.39%				
Instances Classified Incorrectly 17,750	5.60%				
~ Confusion-Matrix ~					
	A1	A	B	C	D	E
---	---	---	---	---	---	---
A1	40,125	675	800	350	325	425
A	800	52,525	450	525	925	25
B	475	800	74,025	925	275	150
C	1050	450	725	84,625	1250	365
D	325	425	925	1475	44,975	1025
E	175	275	150	225	25	2725

the high ranked/optimal attributes are used. The Table 9 shows the achieved accuracy using the subject classifier.

Table 9 Confusion-Matrix of GA reliant DT-classifier on 10-Ҟ Cross-validation

Instances Classified Correctly	306,125	96.64%			
Instances Classified Incorrectly	10,625	3.35%			
~ Confusion -Matrix ~					
	A1	A	B	C	D	E
---	---	---	---	---	---	---
A1	41,675	175	300	175	150	225
A	350	54,025	200	100	300	275
B	700	425	74,750	325	650	225
C	725	775	350	86,525	550	1000
D	300	475	125	600	46,075	500
E	100	175	0	100	125	3075

Regression model of the Decision tree is trained utilizing educational data to predict the students' marks. Utilizing 10-folds cross-validation technique the performance of the model is assessed. Standard matrices for performance evaluation like mean absolute error, root mean square error, correlation coefficient etc. are applied for evaluation of

Table 10 Results of the decision tree regression model for marks prediction

~ Summary ~	
Correlation coefficient	0.92
Mean absolute error	16.37
Root-мean-squared error(RMSE)	8.23
Root-absolute error	8.67
Root relative squared error(RAE)	11.93

regression model. A RMSE of 8.23 (see Table 10) attained by a simple DT's regression-model (Table 11).

Table 11 Results of genetic algorithm based decision tree regression model for marks prediction

~ Summary ~	
Correlation coefficient	0.96
Mean absolute error	11.41
Root mean squared error(RMSE)	5.34
Root absolute error(RAE)	5.21
Root relative squared error	6.44

A tenfold cross validation technique is applied on GA base decision-tree regression model training. Root-mean-square error can be seen reduced by utilizing optimal attributes. The RMSE attained 5.34 utilizing GA based DT regression model which demonstrate performance improvement.

4.3 K-NN Results

The aforementioned Table 12 shows the output of K-NN classifier utilizing K-fold cross-validation technique. In the case of K = 10, the data is divided in 10 test-set and 10 train-set. All the data-set instances are involved in testing and training stage.

Table 12 Error-matrix of K-NN classifier on 10-K cross-validation

Correctly Classified Instances 271600	85.74%				
Incorrectly Classified Instances 45150	14.25%				
~ Confusion -Matrix ~					
	A1	A	B	C	D	E
A1	35825	1125	1725	1175	1075	1625
A	1675	46100	2200	1525	2075	1675
B	300	2225	67875	540	1300	1775
C	1600	2275	1950	79150	2825	2125
D	2050	800	2275	1925	40350	675
E	350	75	275	400	250	2300

The DT-classifier utilizing 10-folds cross-validation obtained 85.74% average accuracy (Table 13)

The GA processed 126 attributes of the educational data-set and ranked the attrib-

Table 13 Confusion-matrix of genetic algorithm based K-NN classifier on 10-K crosss-validation

Instances Classified Correctly 284850		89.92%			
Instances Classified Incorrectly 37900		10.07%			
~ Confusion -Matrix ~					
	A1	A	B	C	D	E
A1	37975	1025	1175	850	700	975
A	1350	49075	950	1150	1550	1175
B	1675	1850	71025	825	1100	750
C	2300	1325	1650	81,550	1200	1900
D	1025	1200	1075	975	42,725	1200
E	275	100	3500	225	175	2625

utes into optimum and non-optimal. Just 30 optimal attributes are chosen amongst 126 attributes, the rest of non-optimum attributes are discarded. Keeping in view the precisions of the two models (Genetic Algorithm based K-Nearest Neighbor classifier and simply K-Nearest Neighbor classifier) GA based K-Nearest Neighbor gained higher accuracy.

Using educational info, the K-NN regression model is trained to forecast students' marks. The model's efficiency is measured by 10 folds cross-validation pattern. The regression model is assessed through quality performance assessment matrices like; root mean square error; means absolute error and correlation coefficient. The Table 14 consists of the experimental results which are obtained using

Table 14 Results of the K-NN regression model for marks Prediction

~ Summary ~	
Correlation coefficient	0.85
Mean absolute error	20.10
Root-mean-squared error(RMSE)	27.66
Root-absolute error(RAE)	21.13
Root relative squared error	26.41

different regression performance evaluation parameters.

The Ḳ-NN Regression Model relying on genetic algorithm is skilled by tenfold cross-validation technique. By integrating high ranked attributes, it may be noted that root-mean-square error is decreased. The R.M.S.E obtained 24.31 (Table 15)

Table 15 Results of the GA based K-NN regression model for marks prediction

~ Summary ~	
Correlation coefficient	0.86
Mean absolute error(MAE)	18.92
Root-mean-squared error(RMSE)	24.31
Root-absolute error	17.16
Root relative squared error	19.51

by incorporating Ķ-NN regression model relying on Genetic Algorithm that demonstrates performance change.

4.4 K-NN and Decision-Tree Models Performance Appraisal

Table 16 indicates the average accuracy attained by grade predicting models, two leading-edge classifier and one leading-edge attribute chooser technique is included

Table 16 Accuracy achieved by different classification algorithms

Classifier	Accuracy
Decision-Tree (DT)	94.39
K-NN	85.74
GA+Decision-Tree	96.64
GA+K-NN	89.92

in the experiential study of grade forecasting framework. The accuracy of genetic-classification models is higher than the accuracy of normal classification models. The Genetic reliant DT-classification model outperforms the Genetic reliant Ķ-NN model for grade forecasting framework.

Table 17 shows the errors gained by marks prediction models, two leading-edge classifiers and one leading-edge attribute chooser technique included in this innovative activity of grade prediction. The error attained with GA reliant regression models are higher than the normal models of regression. The Genetic reliant decision-tree regression model outperforms the Genetic reliant Ķ-NN model for marks prediction framework.

Table 17 Root–mean-square error calculated by various regression models

Regression Model	RMSE
Decision-Tree (DT)	8.23
K-NN	27.66
GA + Decision-Tree	5.34
GA + K-NN	24.31

5 Conclusion

An approach is proposed in this article for observing and forecasting the students marks and grades in an automated way. This research study aims to gain better accuracy for the classification and low root means square error. This study also led us to make groups of students who have same education historic record, for instance, students have taken the same subjects in the same academic session. This job is not simple and easy, fact that intermediate & secondary grade students do not have the same conduct while studying in the same group. Thus to attain reliable forecasting outcomes it is essential to choose students' of the same academic section and group. The student marks and grade were analyzed in this study by knowledge areas. It can be justified that a grade from one subject can be utilized to predict from the grade of a student who took the exam in the previous academic session. The proposed GA based decision tree classifier and regression achieved remarkable results, as presented in Figs 3 and 4, for grade prediction, the classification accuracy is 96.64% whereas marks predicting system based on regression has an Root Means Square Error of 5.34.
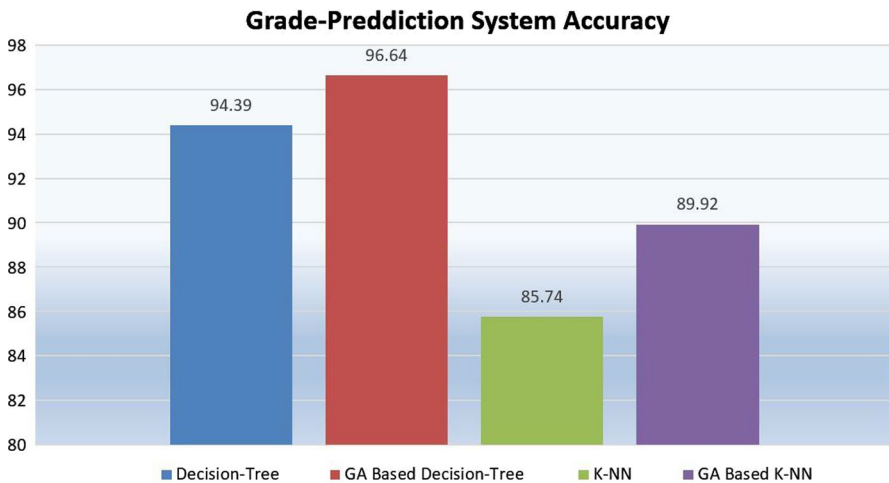

Fig. 3 Performance comparison of K-NN Classifier and GA-Decision-Tree for Grade-Prediction
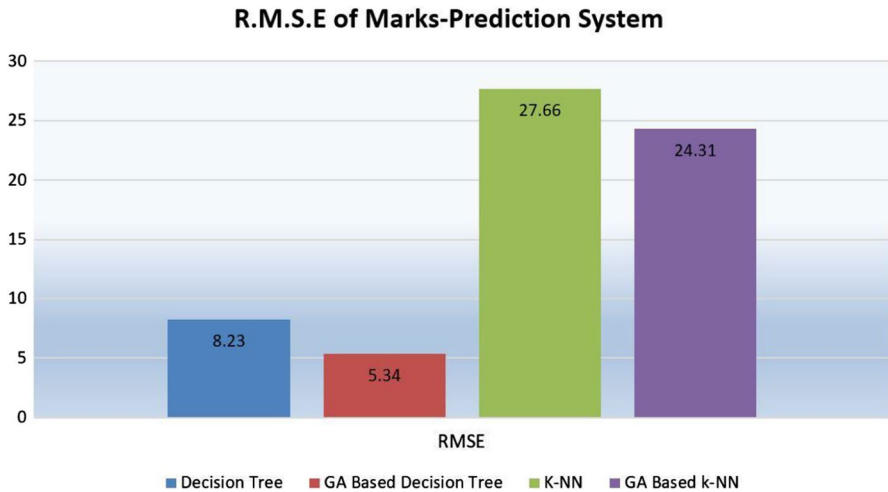
Fig. 4 Performance comparison of Decision-Tree and Ķ-NN Regression for Marks Prediction

6 Future Work

The proposed system have a higher accuracy of the student's academic achievements i.e. final marks & grade. Five sessions students' academic historical data-set were obtained from the BISE Peshawar in this research-driven project, for future ten sessions data-set of SSC & HSSC will be acquired exceeding 6 Boards. Since the data volume increases, it is fitting a Big-Data issue, algorithms of Machine-Learning doesn't play adequately. Classifiers based on Deep-learning and regression models will be integrated for students' performance prediction. Recurrent neural networks uses an optimization approach, which adaptably up-dates the learning rate; aims to get best weight for the features in order to obtain maximum precision in prediction framework.

Ethical Statement

It is to certify that the manuscript "Student-Performulator: Predicting Students' Academic Performance at Secondary and Intermediate Level Using Machine Learning" fulfilled the following:

1) This material is the authors' own original work.
2) The paper reflects the authors' own research and analysis in a truthful and complete manner.
3) The paper properly credits the meaningful contributions of co-author.
4) The results are appropriately placed in the context of prior and existing research.
5) All sources used are properly disclosed/cited.
6) Authors have been personally and actively involved in substantial work leading to the paper, and will take public responsibility for its content.

Conflict of interest The authors declare no conflict of interest.

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Code availability The code generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Authors' contribution All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Shah Hussain and Muhammad Qasim Khan. The first draft of the manuscript was written by Shah Hussain and the 2nd author commented on previous versions of the manuscript. Authors read and approved the final manuscript.

References

1. U. Nations (2019) Sustainable Development Goals
2. Mula I, Tilbury D, Ryan A, Mader M, Dlouha J, Mader C, Benayas J, Dlouhý J, Alba D (2017) Catalysing change in higher education for sustainable development : a review of professional development initiatives for university educators. Int J Sustain High Educ 18(5):798
3. Visvizi A, Lytras MD, Daniela L (2018) The future of innovation and technology in education: policies and practices for teaching and learning education. Innov Prospect Sustain Growth Dev. https://doi.org/10.1108/978-1-78756-555-520181015
4. Casanova D, Moreira A, Costa N (2011) Technology Enhanced Learning in Higher Education: results from the design of a quality evaluation framework. Proc Soc Behav Sci. https://doi.org/10.1016/j.sbspro.2011.11.319
5. Daniela L (2017) An overview on effectiveness of technology enhanced learning (TEL). Int J Knowl Soc Res 8(1):79–91. https://doi.org/10.4018/IJKSR.2017010105
6. Lee J, Choi H (2017) What affects learner's higher-order thinking in technology-enhanced learning environments? The effects of learner factors. Comput Educ. https://doi.org/10.1016/j.compedu.2017.06.015
7. Castro F, Vellido A, Nebot À, Mugica F (2007) Applying data mining techniques to e-learning problems. In: Eevolution of teaching and learning paradigms in intelligent environment, vol 221, pp 183–221
8. Luj S (2018) Big data, the next step in the evolution of educational data analysis. Icits. https://doi.org/10.1007/978-3-319-73450-7
9. Shi Y, Tian Y, Kou G, Peng Y, Li J (2011) Optimization based data mining: theory and applications. J Chem Inf Model 53(9):1689–1699
10. Buenaño-fernandez D (2019) The use of tools of data mining to decision making in engineering education: a systematic mapping study. Comput Appl Eng Edu. https://doi.org/10.1002/cae.22100
11. Tien JM (2017) Internet of things, real-time decision making, and artificial intelligence. Ann Data Sci 4(2):149–178. https://doi.org/10.1007/s40745-017-0112-5
12. Olson DL, Shi Y (2006) Introduction to business data mining, p 389340
13. Ahmed M, Najmul Islam AKM (2019) Deep learning: hope or hype. Ann Data Sci 7(3):427–432. https://doi.org/10.1007/s40745-019-00237-0
14. Member S (2010) Educational data mining: a review of the state of the Art. IEEE Trans Syst Man Cybern C 40(6):601–618
15. Baker RSJD, Yacef K (2009) The state of educational data mining in 2009: a review and future visions. J Edu Data Mining 1(1):3–16
16. Baker RSJ (2020) Data mining for education data mining for education advantages relative to traditional educational research paradigms
17. Ren Z, Sweeney M (2016) Predicting student performance using personalized analytics, pp 61–69
18. Buenaño-fern D, Gil D (2019) Application of machine learning in predicting performance for computer engineering students : a case study, pp 1–18
19. Márquez-vera C et al (2016) Early dropout prediction using data mining: a case study with high school students. Exp Syst 33(1):107–124. https://doi.org/10.1111/exsy.12135
20. Khalifa S, Elshater Y, Sundaravarathan K, Bhat A (2016) The Six Pillars for Building Big Data Analytics Ecosystems. ACM Comput Surv 49(2):1–36

21. Provost F, Fawcett T (2013) Data science and its relationship to big data and data-driven decision making. Big Data 1(1):51–59. https://doi.org/10.1089/big.2013.1508
22. Sin K, Muthu L (2015) Application of big data in education data mining and learning analytics: a literature review. ICTACT J Soft Comput 6956:1035–1049. https://doi.org/10.21917/ijsc.2015.0145
23. Gil D, Fernández-alemán JL, Trujillo J, García-mateos G, Luján-mora S, Toval A (2018) The effect of green software: a study of impact factors on the correctness of software. Sustainability. https://doi.org/10.3390/su10103471
24. Lu OHT, Huang AYQ, Huang JCH, Lin AJQ, Yang SJH (2018) Applying learning analytics for the early prediction of students ' academic performance in blended learning. Edu Technol Soc 21(2):220–232
25. Brooks C, Thompson C, Ri HD, Hgxfdwlrqdo D, Prghoolqj S (2010) Chapter 5: predictive modelling in teaching and learning, pp 61–68. doi: https://doi.org/10.18608/hla17.005.
26. Bydžovská H (2015) Are collaborative filtering methods suitable for student performance prediction? pp 425–430. doi: https://doi.org/10.1007/978-3-319-23485-4.
27. Polyzou A, Karypis G (2016) Grade prediction with models specific to students and courses. Int J Data Sci Anal 2(3):159–171. https://doi.org/10.1007/s41060-016-0024-z
28. Thai-nghe N, Drumond L, Krohn-grimberghe A, Schmidt-thieme L (2010) Recommender system for predicting student performance. Proc Comput Sci 1(2):2811–2819. https://doi.org/10.1016/j.procs.2010.08.006
29. Khan B, Sikandar Hayat Khiyal M, Daud Khattak M (2015) Final grade prediction of secondary school student using decision tree. Int J Comput Appl 115(21):32–36. https://doi.org/10.5120/20278-2712
30. Authors F (2018) Understand, develop and enhance the learning process with big data. Inf Discov Deliv. https://doi.org/10.1108/IDD-09-2018-0043
31. Dahdouh K, Dakkak A, Oughdir L, Ibriz A (2019) Large - scale e-learning recommender system based on Spark and Hadoop. J Big Data. https://doi.org/10.1186/s40537-019-0169-4
32. Thabtah F, Zhang L, Abdelhamid N (2019) NBA game result prediction using feature analysis and machine learning. Ann Data Sci 6(1):103–116. https://doi.org/10.1007/s40745-018-00189-x
33. Kuncheva L (1993) Genetic algorithm for feature selection for parallel classifiers. Inf Process Lett 16:163–168

Authors and Affiliations

Shah Hussain[1] · **Muhammad Qasim Khan[1]**

✉ Shah Hussain
  shahhussainit@gmail.com

  Muhammad Qasim Khan
  mqasimkhan21@hotmail.com

[1] Department of Computer Science, Iqra National University, Peshawar, Pakistan