Education and Information Technologies
https://doi.org/10.1007/s10639-020-10189-1

# Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student

Bashir Khan Yousafzai[1] · Maqsood Hayat[1] · Sher Afzal[1]

Received: 4 February 2020 / Accepted: 8 April 2020/Published online: 29 April 2020


## Abstract

The presented work is a student marks and grade prediction system using supervised machine learning techniques, the system is developed on the historic performance of students. The data used in this research is collected from Federal Board of Intermediate and Secondary Education Islamabad Pakistan, there are 7 regions in FBISE i.e. Punjab, Sindh, Khyber Pakhtunkhwa, Balochistan, Azad Jammu and Kashmir and overseas. The aims of this work is to analyze the education quality which is closely tightened with the sustainable development goals. The implementation of the system has produced an excess of data which must be processed suitably to gain more valuable information that can be more useful for future development and planning. Student marks and grade prediction from their historic academic data is a popular and useful application in educational data mining, so it is becoming a valuable source of information which can be used in different manners to improve the education quality in the country. Related work shows that several method for academic grade prediction are developed for the betterment of teaching and administrative staff of an educational organizational system. In our proposed methodology, the obtained data is preprocessed to improve the quality of data, the labeled student historic data (29 optimal attributes) is used to train decision tree classifier and regression model. The classification system will predict the grade while the regression model will predict the marks, finally the results obtained from both the model are analyzed. The obtain results show the effectiveness and importance of machine learning technology in predicating the students performance.

**Keywords** Data mining · Machine learning · Supervised learning · Attribute selector · Evolutionary search

✉ Bashir Khan Yousafzai
bashir2k2@yahoo.com

Extended author information available on the last page of the article

## 1 Introduction

In Sustainable Development Goals quality education plays an important role which is approved by the United Nations (U. Nations 2019), this is also an important and basic challenge for supporting sustainable development globally. A fundamental element that should be considered while working on sustainable development is the principal to provide opportunities and sharing it equally. In education field, the principal is to guarantee every single student for equal possibilities and opportunities of accessing various resources for the completion of studies (Shields and Satz 2017). In achieving higher education student dissertation is a serious issue which needs to be analyzed globally. The rate of rusticated/dropout students from the school, colleges and universities produce a waste of resources that are important and expensive in the academic sector for other actors and thus also affect the assessment and evaluation procedures of the educational institutions. As the study shows (Paura and Arhipova 2014), the drop out ratio is higher in engineering programs than all other science and arts programs. In our proposed system, a predictive analysis of intermediate students is carried out, the final grades and marks prediction system will facilitate to improve the education standard and focuses on the specific location where students are not getting good grades. The education must be prioritized to change our societies. Teaching and administrative staff must grow sustainability in understanding the development and the abilities of curriculum improvement which will provide students an expanded learning opportunity (Publishing et al. 2017). In this sense, institutions offering secondary and intermediate level education are also required to work on the development and improvement of the educational models by incorporating information and communication technologies, which may be functional as a tool for supporting social responsibilities and equal opportunities.

With such perspective, the outcomes of information and communication technologies in educational systems is imperative since it can significantly contribute to enhance the learning and teaching process, and also in the process of knowledge construction encouragement (Visvizi et al. 2018). The use of and application of information and communication technology involved in learning/teaching phases is also called as Technology based enhanced learning. The term Technology based enhanced learning describes the utilization of digital equipment which intended for enhancing the learning and experiences. Technology based enhance learning has become applicable due to the availability of a number of new technologies, it can help in the improvement of students critical thinking (Casanova et al. 2011). Technology based Enhance learning integrates a lot of emerging technologies, that includes learning management systems, smartphone learning application, virtual, augmented and mixed reality involvements, cloud computing based services for learning, social media and social networking web based applications for learning, video lectures, machine learning, data mining etc. (Daniela 2017).

According to the consequences of teaching-learning about the sustainability of secondary/intermediate education and Technology enhance learning (Casanova et al. 2011), we should be very careful to define the essential conditions for information technology which will benefit us rather than being an obstacle in learning and teaching

sessions. For example, the training of teachers and administrative staff for the development of predictive analytical capabilities are important for calculating the potential outcome of the use of computerized system (Lee and Choi 2017). All of the mentioned technologies above, that are implemented with a better impact in the educational systems to generate a huge amount of data and store it in a way that can be efficiently provided everywhere (Castro et al. 2007). The size of the data can sometime exceed the volume of processing, storing and analyzing it with conventional techniques. To perform a data analysis, new technologies should be considered such as intelligent systems, data mining, big data, and association rules mining. The combination of these new technologies will allow easy and efficient analysis of educational data, these technologies can also be used to transform the education data in some new form that can be more meaningful and useful. (Luj 2018).

Mining of educational data with deep learning is an emerging research area that allows us to process and analyze the educational data gathered from various sources. Various statistical techniques, machine learning, visualization and data mining tools are used for the analysis of educational data. The learning analytics developed from educational data aims to analyze the data collected from the educational database. The learning management system understands the data, enhances the learning techniques and learning environment in which the data occurs (Buenaño-Fernandez 2019).

There are numerous studies in the related work (Castro et al. 2007) (Buenaño-Fernandez 2019; S. Member 2010; Baker and Yacef 2009; Baker n.d.) who have developed various classification-based systems in data mining for prediction of students final grades. Among these methods, followings are the most representative techniques used for classifications: Data analysis and Visualization; Feedback based instructor support system; A Recommendations system for Students; Grade and marks prediction based Students Performance evaluation; Student Modeling; and Social media analysis system. In our proposed system, we aimed to develop a system that will be able to predict student marks and grade, such system is also known as student performance prediction system in the educational data mining, now a days it has a lot of application and gaining a lot of popularity due to its accuracy and efficiency. The objective of the proposed system is to predict and estimate an unknown entity (marks and grades), the quality of results produces by the prediction system will be dependent upon the data and robustness of the classifier or regression model. The data used in our research work consists of multiple attributes and at the same time two supervised learning techniques will work synchronously, one will predict the final grade of the student while others will predict the marks (Ren and Sweeney 2016).

According to these principles, the student's grades and marks will be predicted, the prediction and performance evaluation will be based on student past academic records. According to the predicted results, the students with poor grades and low marks will be identified, proper attention will be given to the weak students so that they can perform well by obtaining good grades in the exams. Early warning may be given to the students who does not achieve good grades and marks according to the decision making system. This predictions system may also assist the education department, providing a summary of the annual exam of secondary and intermediate level students before the exam is

taken. The proposed can also be used for predicting the number of graduating and fail students (Buenaño-Fern and Gil 2019; Márquez-Vera et al. 2016).

The present work is analyzed according to the case study which is used to evaluate the effectiveness of our proposed methodology, the education department will obtain a machine learning based generated and validated result. Every year 8,000 to 100,000 students in the FBISE take the exam, the secondary examination is taken in two steps (Class 9 and Class 10), similarly the intermediate exam is also taken in 2 years (class 11 and class 12). Every year four class exams take place which generates a massive amount of data, this is a big data issue where we consider both the parameters volume of the data and velocity of the data (Khalifa et al. 2016). The conventional methods used for developing a predictive model for small data cannot be used for capturing, preprocessing, features extraction, classification and regression. Big data technology enables us to deal with such data (Provost and Fawcett 2013). For developing a grade and marks predictive system, it is necessary that the students' data should be correct, complete and updated. The scale of the magnitude of the analyzed data will lead us to the development of a big data-based predictive project.

The presented work is organized as follows. In Section 2 the related research work is discussed, in the related work authors' contribution, methodologies and experimental evaluation are discussed. In Section 3 proposed methodology is discussed. Similarly in section 4 detailed experimental works are carried out. In section 3.1, the data collection and preprocessing steps are discussed; in section 3.2 & 3.3 machine learning algorithms and attributes/features selection techniques are discussed; in section 4 experimental procedures, results analysis and results visualization will be discussed. In the last section, discussion and conclusion of our proposed system are discussed.

## 2 Related work

A lot of research work in the past is carried on the topic of educational data mining, and it is still a hot research area in machine learning, data mining, big data and deep learning. Different methodologies and tools are used to visualize and analyze the data, the main aim of many types of research is to develop an automatic system that predicts the grades, marks, institution rating and institution recommendation. Below, some state-of-the-art research works are discussed which has assisted us to describe our proposed methodology.

The study describes the application of big data in the field of education (Sin and Muthu 2015), big data techniques are incorporated in different ways for learning analytics i.e. performance prediction of a system, data visualization, risk detection, student skills estimation, course recommendation system, fraud detection, student grouping and collaboration among other students. The functionality is emphasized by the predictive analysis in this study that is concentered on student performance, behavior and skills prediction.

In Northern Taiwan University (Lu et al. n.d.), a study was carried out, the educational data mining and big data techniques were applied to develop a system for students final marks prediction based on the performance of the students. A principal component regression model was trained, the trained model was used to predict the academic performance of the students. Variables other than courses, such as student behavior out-of-class, quiz marks, video-viewing concentration and tutoring after School time were used as features.

The factors which affect the correctness of software are discussed in a study (Gil et al. 2018). Two types of data analysis techniques are performed on the data when working on education data mining task i.e. predictive models and descriptive models. Predictive modeling uses supervised learning algorithms that estimate the unknown values (Hong and Weiss 2001) while descriptive modeling uses unsupervised learning algorithms to identify and explain the structures of extracted data (Brooks et al. 2010).

In a study (Bydžovská 2015), collaborative filtering technique applications are identified, the objective of the system was to examine the student at the start of an academic period and predict the performance via academic record. The student approved courses are selected to represent the student learning. The historical data was collected from information system to find similar characterizes of Masaryk University students. The experimental results show that this method as effective as a support vector machine classifier.

In research work, the author proposed a new method that uses the students historical academic grade data as an input, the objective is to estimate the performance of the student (Polyzou and Karypis 2016). The research proposal was based on low-range matrix factorization and dispersed linear model. The proposed method was evaluated with the data obtained from the University of Minnesota, the dataset consists of 12.5 years of academic historical data. The proposed system shows improvement in grade prediction accuracy.

In the research work (Thai-nghe et al. 2010), the author proposed a novel approach that extracts education data using a recommendation system, the system is specially designed to predict student performance. The recommendation system is validated by comparing it with another state-of-the-art regression model such as linear and logistic regression. Another contribution of the proposed system is an application of the recommendation system, the educational context such as matrix factorization is used to predict student performance.

In the paper (Ph et al. 2015), the data from various secondary schools located in the district of Kancheepuran was collected. Two state-of-the-art classifiers i.e. decision tree and naïve Bayes were incorporated to build a student classification system, the parents occupation was also added as a features, father occupation in the dataset played an important role to improve the accuracy of the final grade prediction system. The decision tree classifier performs better in terms of accuracy than the naïve Bayes classifier.

The opportunities and facilities offered by the big data technology for education data mining have been studied in detail. The research article shows the relationship between the education environment and big data. The proposed method (F. Authors 2018) focuses on tools, techniques and big data algorithms that are used in the education

context to facilitate and provide benefit in the learning and teaching process. This paper also describes the benefits and importance of big data in the field of education data mining.

In the proposed method (Dahdouh et al. 2019), using big data a smart recommendation system is presented. Association rules mining is an unsupervised method used to find the relationship between student academic activities. Rules are extracted using the rule mining algorithm, student behavior is used to catalog the courses. Finally, a recommendation system is developed using Spark and Hadoop. The obtained results show the effectiveness of the proposed recommendation system.

## 3 Methodology

Our proposed framework for education grade and marks prediction system is described in Fig. 1. Preprocessing techniques are applied to the data, to handle records with missing values, remove the attributes that consist of students personal information, redundant data removal etc. (Table 1). Machine learning techniques are applied to develop a system for the prediction of student performance. Classification algorithms will be used to build a grade prediction system while the marks are predicted using regression model. The methods of machine learning and data mining are selected.

### 3.1 Dataset

The dataset used in our proposed methodology is obtained from the Federal Board of Intermediate and Secondary Education Islamabad. The dataset consists of secondary

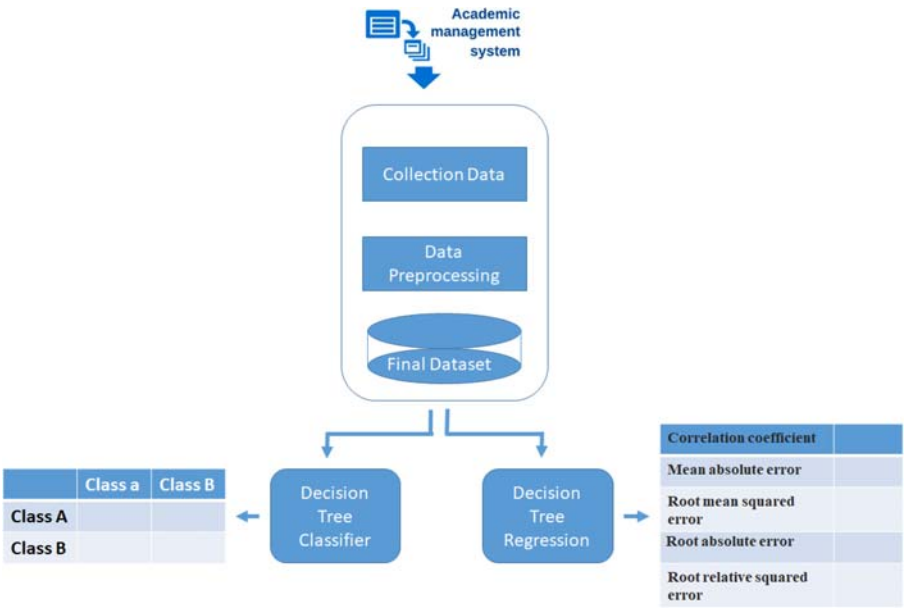


**Fig. 1** Proposed framework

**Table 1** List of attributes in dataset

| reg_no | Dbo_result.roll_no | exam_code | Exam | Session | Year |
|---|---|---|---|---|---|
| Sex | Status | SSC-I | SSC-II | HSSC-I | HSSC-II |
| Grade | PassFail | Remarks | Area | ARB:I | ARB:II |
| U-S | A-S:II | BIO:I | BIO:II | B-M | BMS:I |
| BMS:II | BNK | B-S | C-G | CHE:I | CHE:II |
| CIV:I | CIV:II | CNM | CPS | C-S:I | C-S:II |
| CST | D-H:I | D-H:II | E-A:1 | E-A:II | E-C:I |
| E-C:II | ECO:I | ECO:II | ECP | EDU:I | EDU:II |
| E-E:I | E-E:II | EMT | F-A:I | F-A:II | GEO:I |
| GEO:II | HBB | HOP-I | HOP_II | HPE:I | HPE:II |
| I-H:1 | IH1:II | IH2:II | I-S:I | I-S:II | ISL |
| L-S-I | L-S-II | M-B:I | M-B:II | MTH:I | MTH:II |
| OHE:I | OHE:II | OPT:I | OPT:II | OTT:I | OTT:II |
| P-A:I | P-A:II | P-C | PCL:I | PCL:II | P-E |
| PER:I | PER:II | PHI:I | PHY:II | PST | PSY:I |
| PSY:II | P-T:I | P-T:II | R-T:I | R-T:II | SND |
| SOC:I | SOC:II | STS:I | STS:II | U-C:I | U-C:II |
| U-E:I | U-E:II | A-S:I | PHI:II | | |

and intermediate student academic historic, the secondary examination consists of 9th and 10th class while the intermediate examination consists of 11th and 12th class. There are seven regions in Pakistan, each student taking an exam in FBISE must be from one of the regions. The regions are: Punjab, Balochistan, Sindh, Khyber Pakhtunkhwa, Gilgit and Baltistan, Azad Jammu & Kashmir and Overseas. The dataset consists of the following attributes:

The attributes consisting of students' personal information are removed from the dataset, irrelevant attributes such as Regisitration_No, Roll_No, Exam_Code and Directorate_Code are also removed from the dataset. This obtained dataset consists of 80,000 historical students' data, as the data consists of Grade and Marks_Obtain, so using educational data mining techniques two supervised learning model can be used to predict the grade and marks of the students.

## 3.2 Selection of machine learning techniques

### 3.2.1 Decision tree

In this research, data mining and machine learning techniques are used to develop an accurate and efficient student's grades prediction system for the historical dataset. Supervised learning algorithms are applied for the development of a predictive model which would lay the fundamentals to develop a

recommendation system for the students. Grades and marks based performance prediction systems are considered one of the most challenging and time consuming problem, at the same time achieving higher accuracy in education data mining is also a complex task. Supervised learning is a machine learning technique that is applied to the labeled dataset, the supervised learning approaches include classification and regression. Classification is applied to discrete data and the type of class label is nominal while the regression model works on continuous data and the labels are in numerical format. A decision tree is a state-of-the-art classifier used in data mining, computer vision and bioinformatics. Decision tree classifier has two phases: training and testing (Kami and Jakubczyk 2017), in the training phase classifier is trained by providing features and labels, upon completion of the training phase. The trained classifier predicts a class label is used to predict the class label for the test data. One of the benefits of using the decision tree model for classification and regression is that the predicated results can easily be explained and interpreted, it also has a nice graphical representation to visualize the tree and classification rules.

A decision tree classifier can be constructed using **recursive partitioning,** it starts from the parent node also known as the root node. The nodes can be split and become parent node, as we start from the root node, the features are split which results in the largest information gain. As this is an iterative process, the splitting procedure is repeated until all the leaves are pure at the child node.

$$\text{Entropy is defined as : } H(X) = -\sum_{i=1}^{j} p_i \log 2\, p_i \tag{1}$$

$$\text{Info Gain is defined as : } (Y|X) = \; H(Y) - (Y|X) \tag{2}$$

**Decision tree classifier algorithm**

A decision tree can be created from the training tuples of the different data partition, *D*.

1. A single attribute is selected for dividing the given data.
2. Based on the selected attribute, the data is divided.
3. For every single set that is created in the above steps - repeat step No 1 and step No 2 until the leaf node is not found in the three branches.

– End.

### 3.2.2 K-nearest neighbor

In machine learning, the K-NN is a non-parametric technique (Akhil et al. 2013) that is widely used for solving classification and regression problems. In case of classification and regression, the input data consists of k-closet training examples in the given features space. The output always depends on the

application whether it is used for classification or regression. In supervised learning, the classification output is a class membership. An instance is classified using the plurality vote of neighbors instances, in assigning the predicator to the instance, the most used and common method for assigning a class is the k-nearest selection method ($k$ represent the number nearest neighbor which is a positive integer). If $k = 3$, then the instance is assigned to the class of three nearest neighbors. In case of solving a regression problem using K-NN, the prediction is the property value for the testing instance. The predicated value is obtained by averaging the values of K-nearest neighbors. $k$-NN is also known as lazy learning or instance-based learning technique, in this method there is no training phase and the function is approximated locally. Both for classification and regression, a useful technique can be to assign weights to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. The neighbors in a features space selected from a set of instances for which the class-label in classification or the instance property value in K-NN regression. This can be considered as a training-set for the k-NN algorithm, however the k-NN is a lazy classifier that simply means that this classifier does not perform the training procedures Figs. 2 and 3.

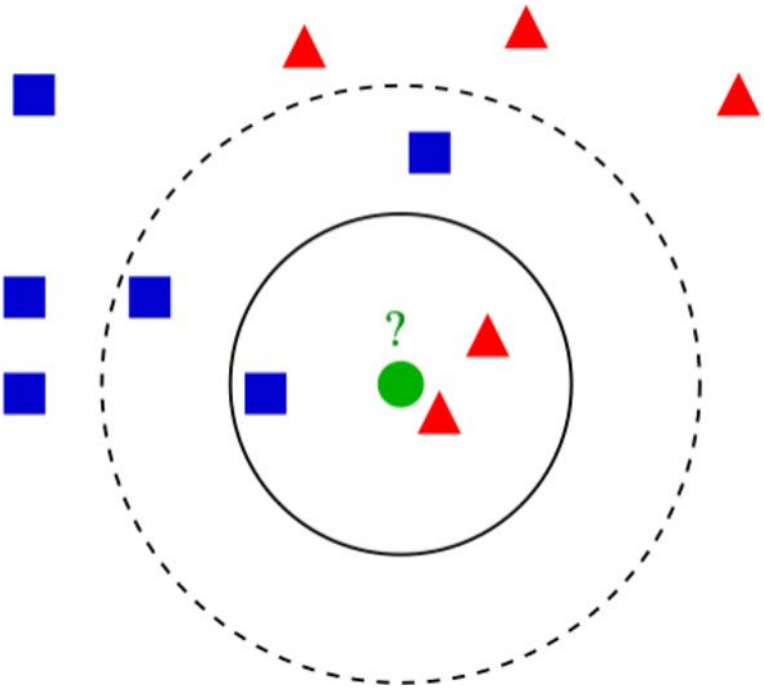


**Fig. 2** Example of $k$-nn classifier, the green circle is a test instance that should be classified into either red triangles or Blue Square. In case of $k = 3$, the test data consists of 3 neighbors, the majority of the 3 neighbors are red triangles so it classified as red triangle. If we select $k = 5$ for the above problem them majority neighbors are from blue square class so the green circle which is test data will be assigned blue square class

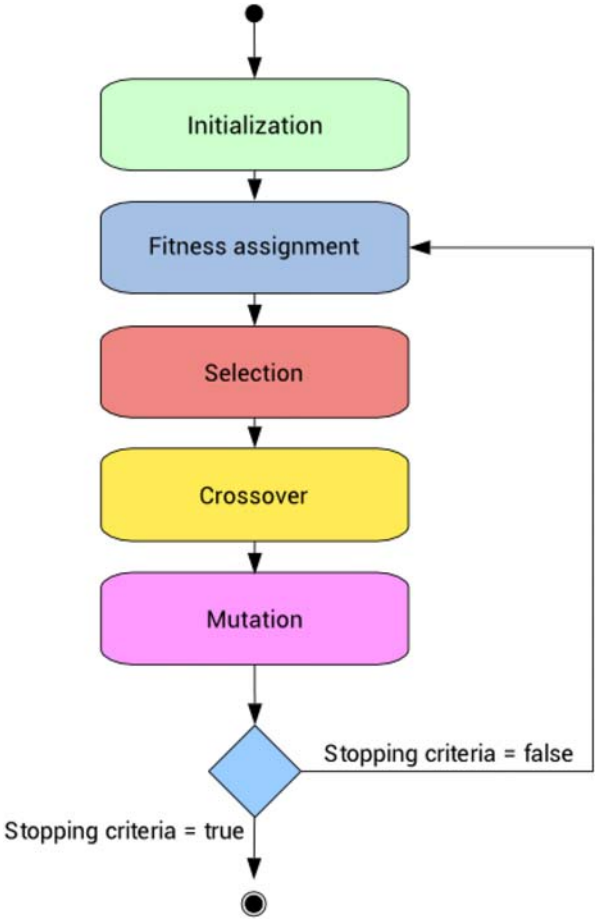


**Fig. 3** State diagram of GA training

### 3.3 Genetic algorithm based attributes selection

The genetic algorithm is a state-of-the-art optimization method (Kuncheva 1993) that has been extensively used in data mining researches, the GA model is an evolutionary search method that copies the working mechanism of selection, mutation and crossover from the nature. GA is a metaheuristic features selection method, it starts the searching and finds a number of solution to the problem, GA is an optimizer which means it will select the best solution from a list of solutions to the problem. In our case of predicting the performance of the students, we have more than 100 attributes that may affect the training accuracy and makes the classification system more inefficient. In our proposed system, at the first step, a random uniform population is generated, the probability mutation and crossover for each generation are 0.3. The model starts and creates an initial random population, the chromosomes consists of a number of gene features where every gene has a real number. The representation of chromosomes is shown in the equation below

$$C = \{\sim ki | \sim ki \in [0, 1]\} n\ i = 1 \tag{3}$$

#### 3.3.1 Pseudocode of proposed features selection method using genetic algorithm

---

*C,T,fbest* ← *Ø* →Initialize

*P0* ← gaussian random distribution with $\sigma = 0.15$ *and* $\mu = 0.3$

All genes are converted to binary discrete

such that {*0, pi* < *0.5*}

{*1, otherwise*}

while t ≤ T do

*t* ← *t* + *1*

GA(*Pt*)

if *argmaxt (Pt)* ≥ *then fbest* ← *argmaxt(Pt)*

end while

return *best p* →Return the best individual

---

### 3.4 Classification and regression performance calculation

Cross-validation is a state-of-the-art statistical technique used for the estimation of machine learning skills. This technique is used in machine learning for optimizing parameters and the selection of the best model, it is easy to understand and implement. Three state-of-the-art cross validation methods are used to validate the classifier and regression. K-Fold Cross-Validation methods will be used in our experimental works.

## 4 Results

In the academic grade predictions system, the dataset consists of regular, private, science group and arts group students. The dataset consists of more than 100 attributes (Table 1), some of these attributes consists of students personal information and other irrelevant attributes for classification are removed from the data. For the prediction of students' grades the classification approach is used, the grade attribute in the dataset is selected as the class label while all other remaining attributes are selected as features (Table 4). Similarly, for predicting the marks of the student's regression model is used, the HSSC-II attributes are selected as predictor, while the remaining are selected as features (Table 5). The description of the dataset is given in the Table 2 below:

The dataset consists of SSC-1, SSC-2, HSSC-1 and HSSC-2 marks. Suppose, a student took an exam of SSC-1 in the year 2012 he/she has taken the exam of SSC-2 in 2013, HSSC-1 in 2014 and HSSC-2 in 2015. Based on the performance of the past

**Table 2** List of optimal attributes

| | | | | |
|---|---|---|---|---|
| U-S | C-S:I | CST | D-H:I | D-H:II |
| C-S:II | CPS | C-G | Grade | CHE:II |
| CIV:I | SSC-I | E-A:I | E-A:II | E-C:I |
| E-C-II | E-E:II | SSC-II | F-A:I | E-E:I |
| EDU:II | EDU:I | ECO:I | HSSC-I | ECP |
| CHE:I | B-S | U-E:II | HSSC-II | |

3 years (SSC-1, SSC-2 and HSSC-1), the proposed system will predict the grade and marks of the students for HSSC-2 Table 3.

### 4.1 Attributes optimization using GA

Tournament selection is a state-of-the-art Genetic algorithm optimization technique that selects a single/multi tokens from thousands of tokens using genetic algorithm. In this method several Tournaments is run, the chromes (Tokens) selection can be random from a population of tokens. Each tournament has a winner, the winner is one which best fits and winner is selected for crossover. In case of larger tournament size, the weak words in vocabulary have less chance to be selected.

### 4.2 Decision tree results

Decision tree is applied on original data consisting 106 attributes and also on the new dataset containing 29 attributes selected by genetic algorithm (Table 7). The optimal attributes help in creating small size tree that reduces the number of rules created by the decision tree. The classification and regression model are evaluated using K-Fold cross-validation technique by selecting the value of k = 10. In k-fold cross-validation, all the

**Table 3** Marks, grade and percentage distribution for SSC and HSSC examination system

| Class | Marks | Grade | Percentage | Division |
|---|---|---|---|---|
| SSC | >= 840 | A-1 | 80% | 1st |
| SSC | >= 735 & <840 | A | 70 | 1st |
| SSC | >= 630 & <735 | B | 60 | 1st |
| SSC | >= 525 & <630 | C | 50 | 2nd |
| SSC | >= 420 & <525 | D | 40 | 3rd |
| SSC | <420 | E | 39 | Fail |
| HSSC | >= 880 | A-1 | 80% | 1st |
| HSSC | >= 770 & <880 | A | 70 | 1st |
| HSSC | >= 660 & <770 | B | 60 | 1st |
| HSSC | >= 550 & <660 | C | 50 | 2nd |
| HSSC | >= 440 & <550 | D | 40 | 3rd |
| HSSC | <440 | E | 39 | Fail |

**Table 4** Distribution of the dataset into features set and labels for grade prediction

| SSC-1 Grade | SSC-2 Grade | HSSC-1 Grade | HSSC-2 Grade (Actual) | HSSC-2 Grade (Predicted) |
|---|---|---|---|---|
| Annual 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 |

**Table 5** Distribution of the dataset into features set and labels for marks prediction

| SSC-1 Marks | SSC-2 Marks | HSSC-1 Marks | HSSC-2 Marks (Actual) | HSSC-2 Marks (Predicted) |
|---|---|---|---|---|
| Annual 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 |

instances in the dataset are involved in training and testing phases, among other cross-validation techniques k-fold method is considered to be the most reliable method in performance evaluation of the trained model.

The above Table 8 demonstrate the performance of decision tree classifier using the k-fold cross validation method. In case of k = 10, the data is divided into 10 trainset and 10 testset, all the instances in the dataset are involved in training and testing phase. The decision tree classifier using 10-folds cross validation achieved an average accuracy of 94.39%.

**Table 6** Genetic algorithm parameters for the selection of optimal attributes

| Parameter | Type/Values |
|---|---|
| Cross-Over Probability | 0.6 |
| Generations | 20 |
| Initialization Operators | Random |
| Mutation Probability | Bit-Flip |
| Mutation Probability | 0.1 |
| Population Size | 20 |
| Replacement Operator | Generational |
| Report Frequency | 20 |
| Seed | 1 |
| Selection Operator | Tournament Selection |

**Table 7** List of attributes selected by GA

| | |
|---|---|
| Total Attributes for Grade Classification | 106 |
| Total Attributes for Marks Prediction | 106 |
| Optimal Attributes Selected for Grade Classification | 29 |
| Optimal Attributes Selected for Marks Prediction | 29 |

**Table 8** Confusion matrix of decision tree classifier on 10-k cross-validation

| | | | | | | |
|---|---|---|---|---|---|---|
| Correctly Classified Instances 299,000 94.39% | | | | | | |
| Incorrectly Classified Instances 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 |

Genetic algorithm attribute selector parameters (Table 6) is used to find the optimal attributes in the dataset, GA method rank the attributes into low and high rank, the low rank attributes are dropped from the dataset. The optimal attributes or highly ranked attributes are selected for the classifier training and performance evaluation. From the accuracies of both the models i.e. simply decision tree and GA based decision tree, the

**Table 9** Confusion matrix of genetic algorithm based decision tree classifier on 10-k cross-validation

| | | | | | | |
|---|---|---|---|---|---|---|
| Correctly Classified Instances 306,125 96.64% | | | | | | |
| Incorrectly Classified Instances 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 |

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

| ===Summery=== | |
|---|---|
| Correlation coefficient | 0.92 |
| Mean absolute error | 16.37 |
| Root mean squared error | 8.23 |
| Root absolute error | 8.67 |
| Root relative squared error | 11.93 |

accuracy obtained using GA based decision tree (Table 9) is higher than the simple decision tree classification method.

Decision tree regression model is trained using the educational data to predict the marks of the students. The performance of the model is evaluated using 10-folds cross validation scheme. Standard performance evaluation matrices such as correlation coefficient, mean absolute error, root mean square error etc. are used to evaluate the regression model. Simple decision tree based regression model achieved a Root mean square error of 8.23 (Table 10).

Genetic algorithm based decision tree regression model is trained using 10-fold cross validation method. By incorporating optimal attributes, it can be seen that the root mean square error is reduced. The RMSE achieved using the GA based decision tree regression model (Table 11) is 5.34 which shows improvement in the performance.

### 4.3 K-nearest neighbor results

The above Table 12 demonstrate the performance of K-nearest neighbor classifier using the k-fold cross validation method. In case of k = 10, the data is divided into 10 trainset and 10 test set, all the instances in the dataset are involved in training and testing phase. The decision tree classifier using 10-folds cross validation achieved an average accuracy of 85.74%.

The educational dataset consists of 106 attributes which are processed by the genetic algorithm to rank the attributes into optimal and non-optimal. Among the 106 attributes only 29 attributes are selected as optimal while the non-optimal attributes are dropped from the dataset. From the accuracies of both the models i.e. simply K-NN classifier and GA based K-NN classifier, the accuracy obtained using GA based K-NN (Table 13) is higher than the simple K-NN classification method.

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

| ===Summery=== | |
|---|---|
| Correlation coefficient | 0.96 |
| Mean absolute error | 11.41 |
| Root mean squared error | 5.34 |
| Root absolute error | 5.21 |
| Root relative squared error | 6.44 |

**Table 12** Confusion matrix of *K*-NN classifier on 10-k cross-validation

| Correctly Classified Instances 271,600 85.74% | | | | | | |
|---|---|---|---|---|---|---|
| Incorrectly Classified Instances 45,150 14.25% | | | | | | |
| === Confusion Matrix === | | | | | | |
| | A1 | A | B | C | D | E |
| A1 | 35,825 | 1125 | 1725 | 1175 | 1075 | 1625 |
| A | 1675 | 46,100 | 2200 | 1525 | 2075 | 1675 |
| B | 300 | 2225 | 67,875 | 540 | 1300 | 1775 |
| C | 1600 | 2275 | 1950 | 79,150 | 2825 | 2125 |
| D | 2050 | 800 | 2275 | 1925 | 40,350 | 675 |
| E | 350 | 75 | 275 | 400 | 250 | 2300 |

K-nearest neighbor regression model is trained using the educational data to predict the marks of the students. The performance of the model is evaluated using 10-folds cross validation scheme. Standard performance evaluation matrices such as correlation coefficient, mean absolute error, root mean square error etc. are used to evaluate the regression model. Simple K-NN based regression model achieved (Table 14) a Root mean square error of 27.66.

Genetic algorithm based K-NN regression model is trained using 10-fold cross validation method. By incorporating optimal attributes, it can be seen that the root mean square error is reduced. The RMSE achieved using the GA based K-NN regression model is 24.31(Table 15) which shows improvement in the performance.

**Table 13** Confusion matrix of genetic algorithm based K-NN classifier on 10-k cross- validation

| Correctly Classified Instances 284,850 89.92% | | | | | | |
|---|---|---|---|---|---|---|
| Incorrectly Classified Instances 31,900 10.07% | | | | | | |
| === Confusion Matrix === | | | | | | |
| | A1 | A | B | C | D | E |
| A1 | 37,975 | 1025 | 1175 | 850 | 700 | 975 |
| A | 1350 | 49,075 | 950 | 1150 | 1550 | 1175 |
| B | 1675 | 1850 | 71,025 | 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 |

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

| ===Summery=== | |
|---|---|
| Correlation coefficient | 0.85 |
| Mean absolute error | 20.10 |
| Root mean squared error | 27.66 |
| Root absolute error | 21.13 |
| Root relative squared error | 26.41 |

**Table 15** Results of the genetic algorithm based *K*-NN regression model for marks prediction

| ===Summery=== | |
|---|---|
| Correlation coefficient | 0.86 |
| Mean absolute error | 18.92 |
| Root mean squared error | 24.31 |
| Root absolute error | 17.16 |
| Root relative squared error | 19.51 |

### 4.4 Decision tree and *K*-NN models performance comparison

The Table 16 show the average accuracies achieved by grade predication models, two state-of-the-art classifier and one state-of-the-art attribute selection method is incorporated in the experimental work of the grade predication system. The accuracies achieved using genetic based classification models are better than the accuracies achieved using simple classification models. The GA based decision tree classification model outperform the GA based K-NN model for grade predication system.

The Table 17 show the errors achieved by marks predication models, two state-of-the-art classifier and one state-of-the-art attribute selection method is incorporated in the experimental work of the grade predication system. The error achieved using genetic based regression models are better than the error achieved using simple regression models. The GA based decision tree regression model outperform the GA based K-NN model for marks predication system.

**Table 16** Accuracy achieved by different classification algorithms

| Classifier | Accuracy |
|---|---|
| Decision Tree | 94.39 |
| *K*-NN | 85.74 |
| GA + Decision Tree | 96.64 |
| GA + *K*-NN | 89.92 |

**Table 17** Root mean square error calculated by various Regression Models

| Regression Model | RMSE |
|---|---|
| Decision Tress | 8.23 |
| *K*-NN | 27.66 |
| GA + Decision Tree | 5.34 |
| GA + *K*-NN | 24.31 |

## 5 Conclusion

In this manuscript, a methodology is proposed for monitoring and predicting the students' grades and marks automatically. The aim of this research work is to achieve higher accuracy for the classification system and low Root mean square error. This study also led us to make groups of students who have common education historic record, for example, students have taken the same subjects in the same academic period. This task is not easy, because secondary and intermediate level students have not the same behaviors when studying in the same group. So to achieve good prediction results it is important to select students from the same group and same academic section. In this research, an analysis of student grades and marks was carried out by knowledge areas. It can be justified that a grade from one subject can be used to predict from the grade of a student who took the exam in the previous academic. The proposed genetic algorithm based decision tree classier and regression achieve impressive results, as shown in Figs. 4 and 5 the classification accuracy for grade prediction is 96.64% while regression based marks predicting system has an RMSE of 5.34 respectively.

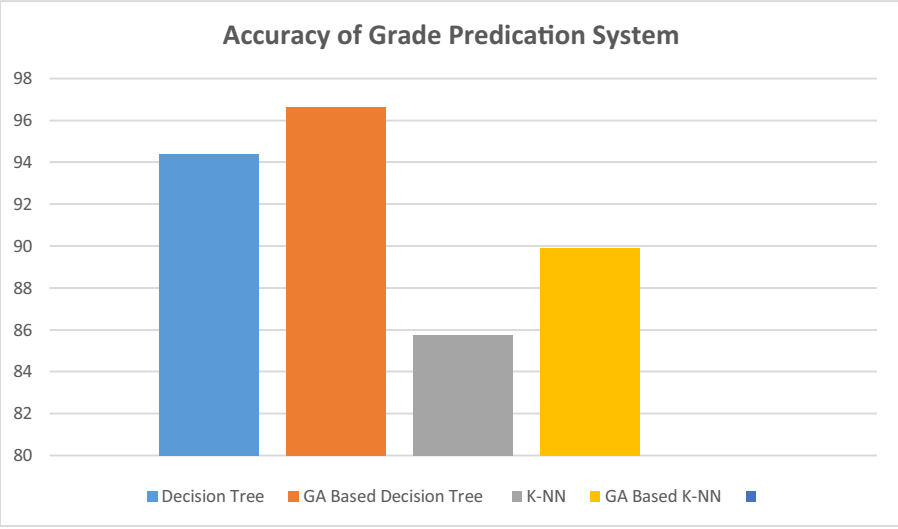


**Fig. 4** Performance comparison of decision tree and KNN classifier for grade prediction

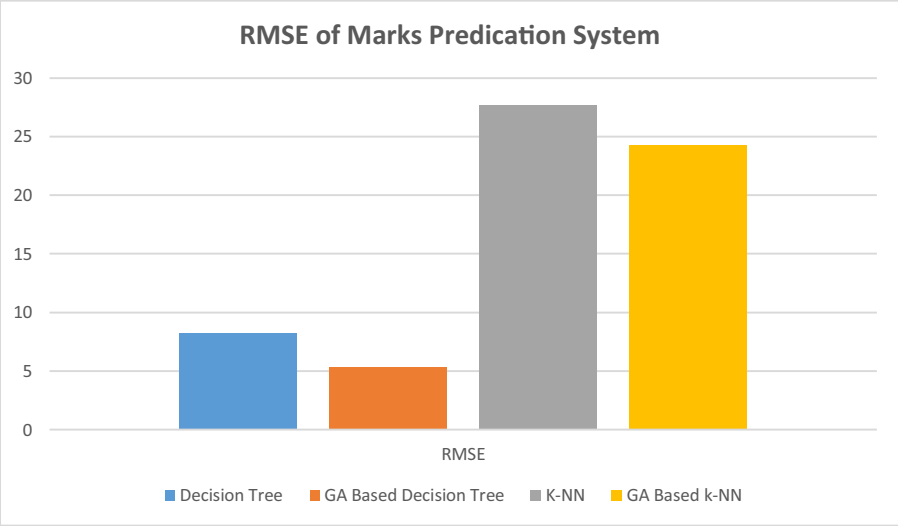


**Fig. 5** Performance comparison of decision tree and KNN regression for marks prediction

## 6 Future work

The proposed systems predict the academic performance (Final grade & marks) of the student with higher accuracy. In this research oriented project, 5 years of student data were collected from the Federal Board of Pakistan. In the future, 10 years of data of SSC and HSSC will be collected from more than 5 Boards. As the size of the data will increase, it is becoming a big data problem due to the volume of the data, machine learning algorithms do not perform well on big data. Deep-learning based classifiers and regression models will be incorporated for the performance prediction of the students. Recurrent Neural Networks are Long-Short-Term-Memory supervised learning approach that uses long term dependencies to train and are also more suitable for big data applications, LSTM can solve both i.e. classification and regression problem. RNN uses an optimization technique that adaptively updates the learning rate, updating the learning rate using optimization techniques tries to obtained the best weights for the features to achieve higher accuracy in the performance prediction system.

## References

Akhil, M., Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using K- nearest neighbor and genetic algorithm. *Procedia Technology, 10*, 85–94.

Baker, R. S. J. (n.d.). "Data Mining for Education Data Mining for Education Advantages Relative to Traditional Educational Research Paradigms."

Baker, R. S. J. D. and Yacef, K. (2009). "The State of Educational Data Mining in 2009 : A Review and Future Visions," vol. 1, no. 1, pp. 3–16.

Brooks, C., Thompson, C., Ri, H. D., Hgxfdwlrqdo, D., and Prghoolqj, S. (2010). "Chapter 5 : Predictive Modelling in Teaching and Learning," pp. 61–68.

Buenaño-Fern, D. and Gil, D. (2019). "Application of Machine Learning in Predicting Performance for Computer Engineering Students : A Case Study," pp. 1–18.

Buenaño-Fernandez, D. (2019). "The use of tools of data mining to decision making in engineering education — A systematic mapping study," no. December 2018.
Bydžovská, H. (2015). "Are Collaborative Filtering Methods Suitable for Student Performance Prediction ?," pp. 425–430.
Casanova, D., Moreira, A., Costa, N. (2011). "Procedia Social and Technology Enhanced Learning in Higher Education : results from the design of a quality evaluation framework," vol. 29.
Castro, F., Vellido, A., Nebot, À., and Mugica, F. (2007). "Applying Data Mining Techniques to e-Learning Problems," 221, 183–221.
Dahdouh, K., Dakkak, A., Oughdir, L., and Ibriz, A. (2019). "Large - scale e - learning recommender system based on spark and Hadoop," *Journal of Big Data*.
Daniela, L. (2017). "An Overview on Effectiveness of Technology Enhanced Learning (TEL)," 8(1), 79–91.
F. Authors. (2018). "Understand , develop and enhance the learning process with big data".
Gil, D., Fernández-Alemán, J. L., Trujillo, J., García-Mateos, G., Luján-Mora, S., and Toval, A. (2018). "The Effect of Green Software : A Study of Impact Factors on the Correctness of Software," pp. 1–19.
Hong, S. J. and Weiss, S. M. (2001). "Advances in predictive models for data mining," vol. 22, pp. 55–61.
Kami, B. and Jakubczyk, M. (2017). "A framework for sensitivity analysis of decision trees".
Khalifa, S., Elshater, Y., Sundaravarathan, K., and Bhat, A. (2016). "The Six Pillars for Building Big Data Analytics Ecosystems," vol. 49, no. 2, pp. 1–36.
Kuncheva, L. (1993). "Genetic algorithm for feature selection for parallel classifiers," vol. 16, pp. 163–168.
Lee, J., and Choi, H. (2017). "What affects learner's higher-order thinking in technology-enhanced learning environments? The effects of learner factors," *Computers & Education*.
Lu, O. H. T., Huang, A. Y. Q., Huang, J. C. H., Lin, A. J. Q., and Yang, S. J. H. (n.d.). "Applying Learning Analytics for the Early Prediction of Students ' Academic Performance in Blended Learning."
Luj, S. (2018). "Big Data , the Next Step in the Evolution of Educational Data Analysis," vol. 1, no. Icits.
Márquez-Vera, C. et al. (2016). "A rticle," vol. 33, no. 1, pp. 107–124.
Paura, L., & Arhipova, I. (2014). Cause analysis of students' dropout rate in higher education study program. *Procedia - Social and Behavioral Sciences, 109*, 1282–1286.
Ph, D., Daud, M., and Ph, D. (2015). "Final Grade Prediction of Secondary School Student using Decision Tree," vol. 115, no. 21, pp. 32–36.
Polyzou, A., & Karypis, G. (2016). Grade prediction with models specific to students and courses. *International Journal of Data Science and Analytics, 2*(3), 159–171.
Provost, F. and Fawcett, T. (2013). "Data science and its relationship data-driven decision making," vol. 1, no. 1, pp. 51–59.
Publishing, E. G., E. Group, Limited, P., Reserved, A. R., Url, O., & Uri, E. (2017). Catalysing Change in Higher Education for Sustainable Development : A Review of Professional Development Initiatives for University Educators. *International Journal of Sustainability in Higher Education, 18*(5), 798 Development : A Review of Professional Development.
Ren, Z. and Sweeney, M. (2016). "Predicting Student Performance Using Personalized Analytics," no. April, pp. 61–69.
S. Member. (2010). "Educational Data Mining : A Review of the State of the Art," vol. 40, no. 6, pp. 601–618.
Shields, L. D., Satz, N. A. 2017. "Equality of Educational Opportunity," *Stanford Encyclopedia of Philosophy*, California USA.
Sin, K. and Muthu, L. (2015). "Application of big data in education data mining and learning analytics – A literature review," vol. 6956, no. July, pp. 1035–1049.
Thai-nghe, N., Drumond, L., Krohn-grimberghe, A., & Schmidt-thieme, L. (2010). Recommender system for predicting student performance. *Procedia Computer Science, 1*(2), 2811–2819.
U. Nations. 2019). "*Sustainable development goals*".
Visvizi, A., Lytras, M. D., and Daniela, L. 2018. "The Future of Innovation and Technology in Education : Policies and Practices for Teaching and Learning Education, Innovation and the Prospect of Sustainable Growth and Development".

## Affiliations

**Bashir Khan Yousafzai [1] · Maqsood Hayat [1] · Sher Afzal [1]**

Maqsood Hayat
m.hayat@awkum.edu.pk

Sher Afzal
skhan.afzal@gmail.com

[1] Department of Computer Science, Abdul Wali Khan University, Mardan, Pakistan