Final Grade Prediction of Secondary School Student using Decision Tree

Bashir Khan
PHD Student, Preston University Islamabad
System Analyst,FBISE,Islamabad

Malik Sikandar Hayat Khiyal, Ph.D
Faculty of Computer Science
Preston University, Islamabad

Muhammad Daud Khattak, Ph.D
Regional director
Allama Iqbal open university
Islamabad

ABSTRACT

Every educational institution around the globe maintain student result repository which contain information about student marks, grade in different subjects and examinations. This repository contains important hidden pattern/knowledge which can be uncovered through data mining. A decision tree classifier based on divide and conquer rules is widely used for data exploration in such repository. In this paper J48 decision tree algorithm is applied on student previous result data to build a model in the form of decision tree. This model can then predict the student final grade. This will be helpful for teacher, student and their parents to know in advance about student final predicted grade and will enable them to take preventive measure.

Keyword:

Data Mining, Educational Data Mining (EDM), Classification, Prediction, Decision Tree, J48, Data repository, Student Grade

1. INTRODUCTION

The searching of new, useful information in large repositories of data is known as data mining (DM). It is a supportive endeavor of computers and humans. Problems and goals describing knowledge of human experts combined with speed and searching capabilities of computer accelerate achieving the best results [1].

The research interest by using data mining is increasing day by day in education. This novel promising field of research is known as Educational Data Mining. Educational data mining is concern with mounting methods to determine facts from data especially unknown knowledge-driven pattern from educational repository in order to highlight the strength and weaknesses of the student [2].

In this modern era of information technology, the repository of students result contains a lot of data. The student repository contain scores and grades of courses for which student are enrolled in the institution. This data bank can be used for knowledge mining; however, many of educational institution around the world and in Pakistan do not apply any knowledge unearthing process on these students data bank. This knowledge discovery process can be applied to improve the student performance, enhance student teacher interaction, identify interdependency among subjects, uncover student failure pattern, predict student final result, and to improve the overall quality of education.

In this paper student of Secondary School Certificate (SSC) in Islamabad Capital Territory will be consider to develop a predictive model. Classification technique of data mining will be used to develop the said model. Since there are many methods for classification, here we will use decision tree method. The model will predict the student final grade based on SSC-1(part 1) marks. The methods used are as follows:

1.1 Classification

Classification is the most widely used technique of data mining, which is applied on pre-classified data record to develop a predictive model which can be used to classify the unclassified data records. This technique is based on network or decision tree algorithms. The process includes two kinds of steps: learning and classification. In the learning step training data set is analyzed by the classification algorithm. The training data set is used to approximate the precision of classification rules. When the accuracy is measure to be satisfactory, these rules can then be applied to new data sets. The pre-classified data records are used by classifier training algorithm to conclude the required parameters for proper identification/discrimination. These parameters are then used to develop a model called classifier [3].

For realization of a data classification and prediction system, first of all it is required to analyze and understand the data in hand and the problem to be solved, because the data may be incomplete, inconsistent or noisy, that is null value, outlier, therefore preprocessing of data and problem is very important for accurate classification and prediction system [4]

1.2 DECISION TREE

Decision tree technique uses tree structure to builds regression or classification models. In this technique dataset is divided into smaller subsets and at the same time an associated decision tree is incrementally developed. That result in a tree having decision and leaf nodes. A decision node is one which has two or more branches. Leaf node represents a decision or classification. The root node known as a best predictor is the topmost decision node in a tree. Decision trees handle both numerical data and categorical data [1]. Decision tree is a well-known and effective technique which builds classification models in the form of a tree. A decision tree is developed through a recursive process which breaks down the set of training data into discrete groups with the objective to maximize distance among groups. The final result is a tree with leaf nodes and decision nodes where the leaf represents a decision or classification [5].

Decision tree is commonly used in data mining to examine data, induce the tree which rules are used to make prediction. The prediction may be of categorical value or continuous value. The output of a decision tree is simple, transparent and easy for non-technical persons to understand.

1.3 DECISION TREE ALGORITHM

Quinlan [6] introduces a decision tree learning algorithm called C4.5 which is a successor of ID3 algorithm. C4.5 follows a greedy approach (non-backtracking) in which trees is constructed in a top-down recursive manner. Top-down recursive approach follows divide-and-conquer strategy to partition the training set into discrete groups until no further splitting is possible [1].

J48 decision tree algorithm considered in this research study, which is java based implementation of the learning algorithm C4.5 in the WEKA (Waikato Environment for Knowledge Analysis).

2. RELATED WORK

Numerous previous studies on data mining applications and methods in different fields have used variety of data types ranging from numeric to text to images stored in variety of data structure and data bases. Thus different data mining methods are used for discovering knowledge. Selection of data mining methods and data is an important task in knowledge/information discovery and needs area knowledge. Many researchers have carried out research to design and develop, generic data mining system but no one has found completely generic system yet. Thus it looks very difficult to design and develop a data mining system which can work for any domain dynamically [7].

Prediction and Classification are two main forms of data analysis in data mining which can be used to extract models that describe data classes or predict trends future data [1].

Classification and prediction also play a very interesting role in the field of education. Classification technique can be applied at every level of education i-e primary, secondary and higher level. If applied in a pragmatic manner from different horizons its application will surely lead to improvement in standard of education and producing quality graduates.

The Author [8] carried out a research project in data mining for the prediction of student performance which was implemented at Bulgarian University with a view to make known the potential profitability of data mining application for the management of University. Different classification techniques were applied on the research data and a comparison was made and it was found that decision tree gave the best results among all.

The authors [9] performed an experimental study on student's database by applying Bayesian classification to forecast the student division/grade based on previous year data and other influencing variables. They found that not only student own effort is necessary for academic performance but other factor like mother education, student's habits, family income also play vital role in the performance of students.

Different decision tree algorithm can be applied to predict and identify the failure risk of student at primary education level. Decision tree generates rules that are easily interpreted and understood by non-expert DM researchers. Therefore non-expert user like teacher can effectively use the output to detect students with problem and to take preventive measure to avoid possible risk of school failure [10].

[11] Conducted a research study in which they collected 1000 records of five different schools of three districts of Tamilnadu, India. They worked to identify prediction variables influencing the performance of students of higher secondary education. They identified that medium of instruction, location of school, marks obtained, type of education and area of living were the main influencing variables for the performance of student. Later a prediction model known as CHIAD was developed to forecast student performance by different decision tree algorithms implemented on the identified variables.

The authors [12] performed a research study to forecast final performance of engineering students in Indian universities from their marks of previous or first year examination. They used three different algorithms of classification ID3, C4.5, and CART, from which they found that for student classification C4.5 algorithm was the best. Later on they implemented the result of predictive study and found that prediction had helped the students at risk to show improvement in final results.

[3] Carried out a research study on student data of colleges affiliated with Awadh University, India. They used Bayesian classification technique for the prediction of performer or underperformer students. They claimed that the research had helped in the improvement of student final division / grade.

The authors [13] conducted a research study on the data of students of introductory management sciences. They used four independent variables for the development of a monitoring and predictive model to monitor the running progress of student and predict their final grade. They claimed that by using the model, the student failing ratio could be curtail by about 20%, and 22% of non-failing students could improve their predicted grades.

[14] Carried out a research study on university student of IE (Informatica Economica – Business Information System) and CIG "(Contabilitate si Informatica de Gestiune – Accounting)" department. The author developed a predictive model to forecast the students' option for continuing post-university education from their social behavior data. The author implemented J48 algorithm, implemented in the WEKA environment. The author claimed that choice of specialization really affect the interest of post-university study.

The authors [15] had carried out a research study on real data which they collected from different schools of secondary education of Kancheepuran district, India. They used naïve Bayes and decision tree algorithm for classification of students. They claimed and concluded the following:

- ⇒ Parents' occupation and not the type of school played a major role in predicting final grade.
- ⇒ Decision tree classification was the best for student modeling.
- ⇒ Final grade of higher secondary students could be predicted from students' previous data.

The authors [16] used data mining methodology on data of MCA student of India to predict the performance of student from their previous marks. They used C4.5 classification by implementing J48 algorithm for predictive model. They concluded that student and teacher can improve their performance from this early prediction.

They also claimed that J48 perform much better in term of speed, space and accuracy etc than ID3 algorithm for student modeling.

From the above literature review it is conclude that J48 work better in term of speed, space and accuracy for the prediction of student performance. Since the current research is also about area wise prediction of student performance therefore J48 will be the best option.

3. MODEL CONSTRUCTION

The steps involved in construction of proposed predictive model are shown in figure 1. The student result repository contains information about student performance, attendance, test and quiz. This repository needs to be explored in a systematic to unearth hidden information.

3.1 DATA COLLECTION

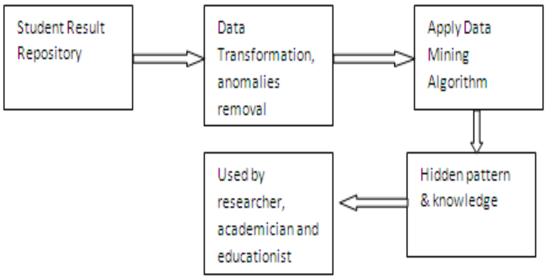
Student Result Repository
Data Transformation, anomalies removal
Apply Data Mining Algorithm
Hidden pattern & knowledge
Used by researcher, academician and educationist

Figure 1: System Diagram for knowledge Discovery

This research study needs data of Secondary School Certificate (SSC) students from Islamabad Capital Territory (ICT). FBISE (Federal Board of Intermediate and Secondary Education) is the only educational Board in Pakistan, which is conducting and regularizing Secondary Education) is the only educational Board in Pakistan which is conducting and regularizing examination in ICT. Therefore the data of students of FBISE can fulfill our requirements for the desired study. The required data collected from FBISE student database for the year 2005, 2006, 2009, 2010, and 2012 in the following format as shown in table 1.

Table 1 Sample of student data

Student Marks in SSC-I	**Final Grade in SSC-II**	**Number of Students**	**Total (No. of students * No. of Years)**
Marks obtained in previous year	A1	50	50 * 5 = 250
	A	50	50 * 5 = 250
	B	50	50 * 5 = 250
	C	50	50 * 5 = 250
	D	50	50 * 5 = 250
	E	50	50 * 5 = 250
	Total Record for 05 years		**1500/-**

The students' final grades represented in table 1 are:

Grade	**Meaning**
A1	80% and Above of marks
A	70% to 80% of marks
B	60% to 70% of marks
C	50% to 60% of marks
D	40% to 50% of marks
E	minimum passing marks (33%) to 40%

3.2 IMPLEMENTATION OF ALGORIHTM

The WEKA is recognized as a most widely used tool for research in Data Mining and has achieved widespread acceptance in the academia and business [17]. WEKA provides excellent implementation of classification algorithm in data mining with a graphical user interface. Since this research also use data mining classification, therefore, WEKA is the best available option to use.

In order to utilize WEKA software, the data needs to be transformed into a format compatible with WEKA which is ARFF format. The following steps were being taken to bring out the data in desired format:

a. Data was collected for five year through SQL query from student database.

b. This data was then saved in CSV format in Excel i-e studentmodel.csv.

c. WEKA environment provides the facility to convert data from CSV format into ARFF format. This facility was used and studentmodel.csv was converted into studentmodel.arff format.

studentmodel.arff file was loaded into WEKA explorer. The Classify option in WEKA enables us to apply classification algorithm on the dataset, to obtain estimated accuracy of the model, show the confusion matrix. We choose "J48" decision tree algorithm for classification. From the "Test Option" we select 10-fold cross-validation because we have not separated training data set and was therefore required to get a logical idea of accuracy. The attribute "Grade" was selected for prediction. The resultant model is generated in the Form of decision tree.

3.3. RESULTS AND FINDINGS

The model in the Form of decision tree generated from sudentmodel.arff is shown in figure 2. The accuracy of the predictive model obtained is 84.53% which mean that 1268 student out of 1500 are correctly classified. The high accuracy of the model mean that student previous marks is most important factor in predicting student final grade.

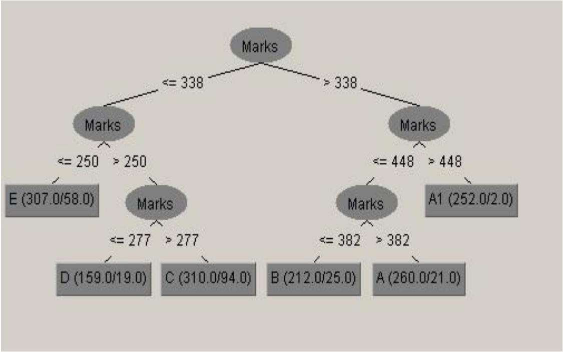

Figure 2: Decision Tree for student final grade prediction

The rules generated from the above decision tree are as below where P_Marks represent the student previous marks or SSC1 or 9th marks.

```
IF  P_Marks > 338 then
        IF  P_Marks > 448
                then Grade = "A1"
        ELSE
                IF P_Marks > 382
                        then Grade = "A"
                ELSE Grade = "B"
ELSE
        IF  P_Marks <= 250
                then Grade = "E"
        ELSE
                IF P_Marks <=  277
                        then Grade = "D"
        Else Grade = "C"
```

It is obvious from the confusion matrix shown in table 2 that the model is ideal in the prediction of grade 'A1' and grade 'E' because out of 250 students it has been classified 100 % correctly and therefore it is similarly in grade 'A' 233 students out of 250 has been classified correctly, in grade 'B' 184 students out of 250 has been classified correctly, in grade 'C' 217 students out of 250 has been classified correctly and in grade 'D' 135 students out of 250 has been classified correctly.

Table 2 Confusion Matrix for each grade prediction

Grades	Predicted As					
	A1	A	B	C	D	E
A1	250	0	0	0	0	0
A	2	233	15	0	0	0
B	0	21	184	45	0	0
C	0	0	15	217	18	0
D	0	1	0	56	135	58
E	0	0	0	0	0	250

As shown in accuracy table 3, TP (True Positive) rate and FP (False Positive) rate confirm the discussion on confusion matrix (table 2) that model is ideal in prediction of A1 and E Grade because their TP rate is 1. Similarly the model is good in predicting A, B, and C grade but not very efficient in prediction of D grade. The results presented in table 3 are also shown in graphical Form in figure 3.

Table 3 Grade wise TP rate and FP rate

Grade	TP Rate	FP Rate
A1	1.0	0.002
A	0.932	0.018
B	0.736	0.024
C	0.868	0.081
D	0.54	0.014
E	1.0	0.002

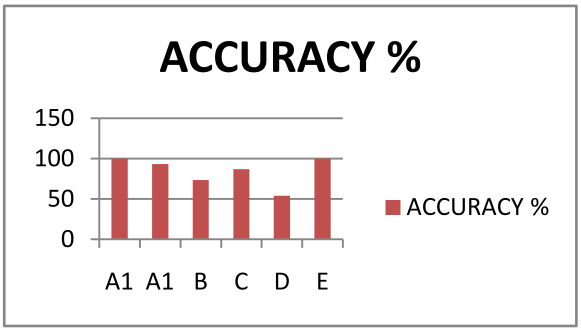

Figure 3: Grade Wise Prediction Accuracy

4. CONCLUSION

This research study show that student previous marks can be use to generate a predictive model using decision tree algorithm, which can be used for predicting the final grade of Secondary School Student. The accuracy of the model is 84.53% which means that the model is successfully predicting the final grade of student i-e 1268 out of 1500 students has been successfully classified. The teacher, students and their parents can improve the result of student who are likely to pass in low grade through proper counseling.

5. FUTURE WORK

The current student model predicts the final grade of SSC student for ICT student only. Future research is needed to develop a similar system for HSSC students and for other regions of Pakistan. For future, the research can also be extended to investigate subject wise performance of SSC and HSSC students which will be helpful in targeting fine quality graduate of specific region in Pakistan.

6. REFERENCES

[1] Han, J., and Kamber. M, (2012) Data Mining: Concepts and Techniques, San Francisco, Morgan Kaufmann.

[2] Alcala, J., Sanchaz, L., Garcia, S., Del Jesus, M. ct. (2007). KEEL :A software tool to assess Evolutionary Algorithms to Data Mining problems. Soft comput, 10.1007/s00500-008-0323y.

[3] Panday, U. K., & Pal, S. (2011). Data mining: A prediction of performance or underperformer using classification. International Journal of Computer Science and information technology, 2(2), pp. 686-690.

[4] Ozekes, S., & Camurcu, A. Y. (2002), classification and prediction in a data mining application, journal of marmara for pure and applied science, 18, pp. 159-174.

[5] Danso, S. O. (2006). An Exploration of Classification prediction techniques in data mining: the insurance domain. [Thesis]. Bournemouth: Bournemouth univeristy.

[6] Quinlan, J. R. (1993), C4.5: Programs for Machine Learning, San Mateo, CA:Morgan Kaufmann.

[7] Deshpande, S. P., & Thakare, V.M. (2010), Data mining system and applications: A review, international journal of distributed and parallel system, 1(1), pp. 22-44.

[8] Kabachieva, D. (2013), Predicting Student Performance by Using Data Mining Methods for Classification, Cybernetics and information technologies, 13(1), PP.61-72

[9] Bhardwaj, B. K., & Pal, S. (2011), Data mining: A prediction for performance improvement using classification, International journal of computer science and information security, 9 (4), pp. 86-91

[10] Marquez-Vera, C., Romero, C., & Ventura, S. (2011). Predicting School Failure Using Data Mining, Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, The Netherlands, PP. 271-276

[11] Ramaswami, M., & Bhaskaran, R. (2010), A Chaid based performance prediction model in educational data mining, international journal of computer science issue, 7(1), pp. 10-18

[12] Yadav, S. K., Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification, World of computer science and information technology journal, 2(2), pp. 51-56.

[13] D'Souza, K. A., & Maheshwari (2011). Predicting and Monitoring student performance in the introductory management science course, Academy of Educational Leadership Journal, 15, PP. 69-80

[14] Bresfelean, V. P. (2007). Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment, Proceeding of the ITI 2007 29th international conference on information technology interfaces, June 25-28, Cavatt, Croutia. pp. 51-56

[15] Ramesh, V., Parkavi, P., & Ramar, K. (2013). Predictive student performance: A statistical and data mining approach. International Journal of computer application, 63(8), pp. 35-43.

[16] Prasad, G. N. R., & Babu, A.V. (2013), Mining previous marks data to predict students performance in their final year examination, international journal of engineering research & technology, 2(2), pp. 1-4.

[17] Frank, E., and Whitten,I.H., (2005) Data Mining: Practical Machine learning and technique, San Francisco, Elsevier & Morgan Kaufmann.