Factors that contribute programming skill and CGPA as a CS graduate: Mining Educational Data

Md Aref BILLAH, Sheikh Arif AHMED, Shahidul Islam KHAN
Department of Computer Science and Engineering (CSE)
International Islamic University Chittagong (IIUC)
Chittagong, Bangladesh
mdarefbillah818@gmail.com, sheikharif1993@gmail.com, nayeemkh@gmail.com

Computer Science (CS) has become one of the most popular under graduate program in last few years. According to UGC roughly 116 universities out of 136 are offering computer science program which indicates a massive number of students are choosing this program as their undergraduate program. But statistically significant number of students are failing to become skilled and effective CS graduate because many students are taking CS without accessing their chance in this program. Success in academic and professional life require to choose right under graduate program. Considering CGPA and Programming Skill as two of the most significant factors to determine student's success in CS, we have predicted these two by taking students personal interest, academic results, analytical skill and problem solving skill into account. We also extracted most significant features of a prospective CS student by using gain ratio.

***Keywords**: Computer Science Student, Predicting Performance, Machine Learning Techniques, Data Mining, Programming skill, CGPA*

1 Introduction

Computer Science has now become a buzzword in the global community. Being one of the developing countries Bangladesh Government has already taken the challenge of outshining in the ICT department, so as the students. But when the question of skill, show casing talent and achievements in national and international level comes it seems significantly important percent of students are failing to do so. Without proper analysis, substantial amount of students are taking this program and eventually the performance of larger part of students of this Program is performing poorly which is hurting their

Academic and professional life. Though massive number of students are rushing into this program in Bangladesh, many IT industries are still hiring IT professionals from India because of limited number of skilled graduates [1].

So, to predicting the performance of prospective CS graduates before they start is what they need. If they can know the factors on what their performance as a CS graduate depends, they can decide whether they are going to take CS or not. There is a possibility that they can change themselves as the demands to be a good CS graduate.

In recent past, Students' final result predictor and Classification model for determining students' future was built by taking different types of feature sets like as previous academic result, family income, family expenses, medium of teaching, marital status, parent's occupation, parent's qualification, family size, attendance, assignment, lab work: [2],[3],[4],[5],[6],[7]. Different types of classification model were built by implementing Decision tree, Support Vector Machine and Naive Bayes Classifier algorithm on students' academic results: [2], [3].

Purpose of this work is to build a model to predict students' final result or we can say CGPA of their CS program and also the programming skill which is very good indicator of a good CS graduate before they start.

To predict students' performance in any program it is important to take the previous academic results into account [8]. Studies show that CS students' need mathematical skill for increasing programming and other skills like problem solving and analytical skills: [9], [10]. So besides academic results, to make our proposed model more accurate in this paper we used their previous academic results, their experience with ICT course at HSC level, online class experience, Internet browsing reasons, Participation in Mathematics or Science Olympiad, Interest in Competitive programming and students problem solving skills etc. To train our model we have used current CS student's data and when our model is ready we predicted both CGPA and programming skill. Where CGPA tells about student's theoretical knowledge and programming skill tells about their practical knowledge.

2 Literature Review

Many researcher worked for predicting the performance of students.

Romero et al. (2013) conducted a study on students participating in on-line discussion forum and predicted students' final performance in Spain. They collected forum interaction data such as number of messages post/read, ask and reply relationship between students. Afterwards compared them in between classification and classification via clustering approach. [11]

Alharbi et al. (2016) did a case study whether they can highlight performance problems early on and propose remedial actions using data mining techniques. They collected students data during admission and after completing their academic first year and eventually predicted good honors outcomes with reasonable accuracy by using classifying model with highlighting students that are predicted to low achievers with high probability module results [12].

Baradwaj & Pal (2012) suggested a classification model for Predicting Performance improvement on the Educational databases which contains invisible information for improvement of students' performance. They collected 300(74 females, 226 males) students record from Dr. R. M. L. Awadh University, Fazibad India and used Bayes classification approach. They further conducted a research by taking students class test, assignments, attendance, lab work and seminars into consideration and analyzed students' performance in the semester final examination. They used decision tree for predicting students' performance: [2], [3].

Daud et al. (2017) conducted a study where using Educational Data mining approach they considered student's family expenditure, personal information and predicted whether he will be able to complete his degree or not. They used WEKA tool to classify 100 students record with 23 features each from different universities of Pakistan [4]

Goga et al. (2015) proposed a tool by using .Net framework which takes students various information as input and predicts students' grade. They first collected the student's enrollment records from Babcock University, Nigeria and then built models using classification trees and a multi-layer perception learning algorithms operating on WEKA. In the domain of this study random tree adopted as the best algorithm and served as a building block of this generic system [13]

Arsad & Buniyamin (2013) conducted a study to predict student's final result at the Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM), Malaysia. They showed Artificial Neural Network model served as a vital to predict students result. They found that Students' first and thirds semesters fundamental course results reflected their final result. Therefore, fundamental subjects should be fully understood because without clearing fundamental knowledge it is very difficult to face advance courses [14].

Huang & Fang (2012) proposed a mathematical model for making early prediction of students' final score and Engineering dynamic courses using the results of first semester and validated using next three semester data. They collected data from 1900 engineering students of four semesters. Four different mathematical modeling approaches: radial basis function neural networks, multi-layer perception neural networks, multivariate linear regression, multi-layer perception neural networks, and support vector machines were applied and experimental results showed anyone can be used to predict the desired result [15].

3 Techniques

We have used few ***data miming techniques*** and existing algorithms and tested their efficiency through various techniques. Here is a short description of these tools and techniques

The term Data mining is a misnomer, as it refers to extracting knowledge from large amount of data [16]. The data mining techniques is featured to create model which will help to find new data using unknown data [17]. Data mining can be basically of two types, Predictive and Descriptive [18].

Predictive Data Mining Model

This method studies previous historical data and predict and forecast what is going to happen to future data set e.g.: classification, regression, time series analysis etc. [19].

C 4.5: It's a decision tree based algorithm for classification both numeric and nominal classes. It was written by J. R. Quinlan [20].

Support Vector Machine: Support vector machine is the supervised algorithm for using as a classifier or regression algorithm for pattern, nested problems or mining of texts [21]. It uses a hyper plane to partition two different classes where support vectors are those which instances are used as the margin [22].

Descriptive Data Mining Model

Descriptive data model is a predictive model consisting of clustering, summarization, association rule etc. It finds pattern in large data and further works in intelligence system decision-making.

Data mining is of two types according to its class i.e. supervised and unsupervised learning algorithms.

Unsupervised algorithms: No information about class or class label. E.g.: Clustering and association rules etc.

Supervised algorithms: Class data is known. E.g.: classification, regression etc.

4 Methodology

Researchers from different regions found that study on forecasting student's final result or CGPA for under graduation program is very important. They considered students' academic result, parents data, hours spent in study, activity in online discussion forum, marital status & student's class attendance for forecasting their final performance and building classification model which can help them take decision which track they should follow for the under graduation program. Different from the literature, we considered only Computer Science program as our purpose of study for its immeasurable popularity and the kind of challenges it encounters. Especially, for someone who is not ready to take those challenges is sure to suffocate in CS. For CS programming skill measurement is also very important. So we build a model to predict CGPA and programming skill using different regression algorithm to build a prediction model based on 23 features including their academic results, experience with ICT course, personal interests, personal experiences of current students' with similar academic achievements and problem solving skills.

Our proposed method has few stages-

1. Data collection
2. Pre-processing
   - Data cleaning
   - Transformation
   - Integration
   - Standardization
   - Feature selection
3. Data mining, model generation and Performance measurement of algorithms
4. Finally we will get a model to use.

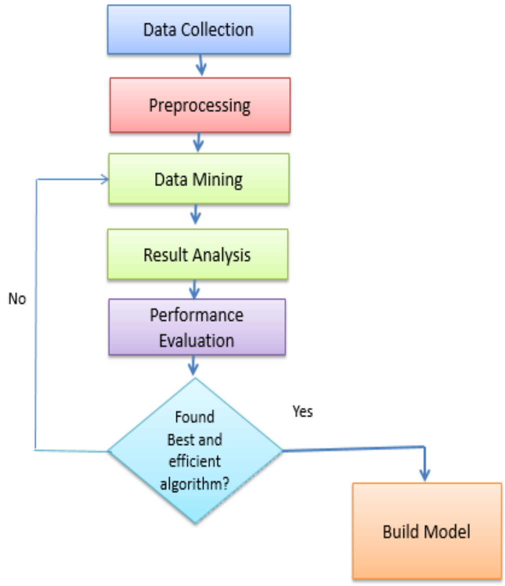

Fig. 1. Overview of Model. It shows the whole process of building our model step by step. Starts with data collection and ends with building a model.

5 Implementation and Result

Primary data was collected by doing a survey on current students of Computer Science. We collected CS student's data having 23 features in these six following categories. i.e.: Personal information, personal Interest, academic results, their experience with ICT course in HSC level, level of their problem solving skill and their online course experience. We assigned numeric values to these questions responses for data analysis and research purpose. We have removed inconsistency from our data set and transformed them into the proper format to use them for analysis. **Table 1** Shows the options with correspondent values and short term of the survey questionnaire for students and here is the questionnaire-

Section A- Personal Information

A1- What is your name?

A2- Gender?

A3 -What is the Location of your University?

Section B- Academic result

B1- What is your SSC GPA out of 5.00?

B2- What is your HSC GPA?

Section C- Experiences with ICT Course

C1- How did you find ICT course (Your experience with ICT)?

C2- How was your academic result in ICT?

Section D- Personal Interest

D1-Why do you browse internet mostly?

D2- Participated in number of math or science Olympiad?

D3- Online courses you've followed related to your study?

D4- Rate your interest in Competitive Programming?

Section E- Personal Experience

E1- "Tendency of using online resources help a lot in Computer Science":

E2- "Having Patience help in Computer Science":

E3- "Knowing PC configuration helps in reading Computer Science":

E4- 'Computer gamer have better chance in CS':

E5- "Capability of Self-study makes significant difference in Computer Science":

E6- "HSC ICT Course result reflects once potential in Computer Science":

E7- Knowledge (about the CS program) you had before starting your Program?

Section F- Problem solving skills and others

F1- Rate your programming skill:

F2- Your skill in Basic mathematics (SSC level):

F3- Your skill in Higher Level Mathematics (HSC level):

F4- Rate your patience out of 10:

F5- Rate your capability of self-study out of 10:

Table 1. Options with correspondence values and short terms

Options	Short Term	Values
Strongly Agree/Very Good /5:00P/Many /Very Interesting /9 to 10 ratings/Social media and Entertainment	SA/VG/C3.8AA/5P/MT/VI/9-10R/FLP/SMAE	5

Agree/Good /4.50P and above/More than 5/Interesting/7 to 8 ratings/Study and Social Media	A/G/C3.4AA/4.5PAA/MT5/I/7-8R	4
Neutral/Satisfactory/ /4.00P and above/More than Once/5-6 ratings/, Reading Blogs only	N/S/C3.0AA/4.0PAA/MTO/5-6R	3
Disagree/Less than satisfactory /3.50P and above/Once/3-4 ratings/Online Courses only	D/LTS/C2.5 AA/3.5PAA/O/3-4R	2
Strongly disagree/Poor/ Less than 3.50P/Never/1-2 ratings/Online Course and Reading Blogs	SD/P/CLT2.5/LT3.5P/Ne/1-2R	1

Table 2. Students related variables that illustrates the questions we asked the students and probable answers.

Variables	Description	Possible Values
Gender	Students Gender	{Male, Female}
UniLoc	University Location	{Dhaka, Chittagong, Other City, Outside City}
GSSC	Students grade in SSC	{5.00 >4.50 and <5.00, >4.00 and <4.50, >3.5 and <4.00, >3.5}
GHSC	Students grade in HSC	{5.00 >4.50 and <5.00, >4.00 and <4.50, >3.5 and <4.00, >3.5}
ICTResult	Students ICT result in HSC	{Very Good, Good, Satisfactory, Less than satisfactory, poor}
ProRatings	Students programming skill ratings	{ 9 to 10,7 to 8, 5 to 6,3 to 4,1 to 2}
Bmath	Students skill in basic mathematics	{Very Good, Good, Satisfactory, Less than satisfactory, poor}
HMath	Students skill in Higher level mathematic	{Very Good, Good, Satisfactory, Less than satisfactory, poor}
PaRatings	Students Patience's ratings	{ 9 to 10,7 to 8, 5 to 6,3 to 4,1 to 2}
ICTExp	Students experience with ICT	{Very Interesting, Interesting, Neutral, Difficult, Very Difficult }
IBReason	Most important reason behind browsing internet	{Online courses and Reading Blogs, Online Courses only, Reading Blogs only, Study and Social Media, Social media and Entertainment}
MSOlym	Student participation in number of math or science	{many , less than 5, more than once, once, never}

	Olympiad	
NOC	Number of online courses followed	{many, less than 5, more than once, once, never}
Cpro	Students competitive programming Interest ratings	{ 9 to 10,7to 8, 5 to 6,3to 4,1 to 2}
SSR	Students capability of self-study ratings	{ 9 to 10,7to 8, 5 to 6, 3to 4,1 to 2}
ORH	Tendency of using online resources helps	{Strongly agree, agree, neutral, disagree, strongly disagree}
PH	Having Patience helps	{Strongly agree, agree, neutral, disagree, strongly disagree}
SSH	Self-study helps	{Strongly agree, agree, neutral, disagree, strongly disagree}
CGC	Computer gamers chance in Computer Science	{Strongly agree, agree, neutral, disagree, strongly disagree}
ICTRR	ICT result reflects success in CSE	{Strongly agree, agree, neutral, disagree, strongly disagree}
PC	Knowing about PC configuration helps in reading Computer Science?	{Strongly agree, agree, neutral, disagree, strongly disagree}

We have collected data from current CS student of various university from various city of Bangladesh. **Table 3** shows the frequency, percent valid percent and cumulative percent of dataset in respect to Gender

Table 3. Frequency Table (Gender of Students). 329 of 501 students are male and 172 is female. That is 65.7% students are male and 34.3% students are female.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	329	65.7	65.7	65.7
	Female	172	34.3	34.3	100.0
	Total	501	100.0	100.0	

Table 4. Programming skill * CGPA cross tabulation. We can see most of the good programmers (with rating 4 and 5) have CGPA 3.5-4.00 and out of 90 students with programming skill 5, 72 students have CGPA 3.7-4.0.

Programming skill * CGPA Cross tabulation							
		CGPA					Total
		2.49-bellow	2.5-2.99	3.0-3.49	3.5-3.69	3.7-4.00	
Programming skill	1	5	9	10	0	0	24
	2	0	36	63	15	3	117
	3	0	15	62	21	12	110
	4	0	0	42	**60**	**58**	160
	5	1	2	1	**14**	**72**	90
Total		6	62	178	110	145	501

5.1 Preprocessing

Often collected data is not understandable, inconsistent, lacking in important criteria or can contain various errors. Preprocessing makes data understandable by various process and solve those issues. Hence we have removed inconsistency from our data set and transformed them into the proper format to use them for analysis. Also selected important features to reduce the complexity of our process. For this we have used a filter method which is feature selection by gain ratio. Table-3 shows the comparison selected features-

Table 5. List of selected features. Here we can see for CGPA programming knowledge is an important factor. For programming skill self-study, patience and skill of higher level mathematics is most important.

Selected Features (For CGPA)	Gain Ratio	Selected Features (For CGPA)	Gain Ratio
ProRatings	0.474	SSR	0.279
PaRatings	0.465	PaRatings	0.268
GHSC	0.457	HMATH	0.255
SSR	0.452	Cpro	0.246
Cpro	0.448	MSOlym	0.224
GSSC	0.439	NOC	0.212
HMATH	0.439	CGPA	0.211
MSOlym	0.410	BMath	0.183
NOC	0.407	ICTResult	0.182
IBReason	0.404	IBReason	0.172
ICTResult	0.389	GHSC	0.168

5.2 Data Mining and Model Generation Using Various Algorithms

For predicting the CGPA and programming skill we have used several algorithms i.e. SVR, C4.5 etc. While predicting two standard evaluation metrics (MAE, RMSE) are used as cost function to evaluate the o performance of our prediction.

Mean Absolute Error (MAE): measures the difference between two continuous values. It uses absolute values and gives intuition (the “average error”).

Root Mean Square Error (RMSE): refers to the standard deviation of the prediction errors. By using RMSE we can tell how concentrated the data is around the line of best fit.

Now we will see the results of regression algorithms for predicting the CGPA and programming skill in respect to these cost functions.

Table 6. Error evaluation for Regression (CGPA prediction)

Evaluation Criteria	**Multiple Linear Regression**	**C4.5**	**SVR (Linear kernel)**	**SVR (Poly kernel)**
MAE	0.1180	0.1555	0.1123	0.1409
RMSE	0.1712	0.2517	0.1604	0.2117

Table 7. Error Evaluation for Regression (Programming Skill Prediction)

Evaluation Criteria	**Multiple Linear Regression**	**C4.5**	**SVR (Linear kernel)**	**SVR (Poly kernel)**
MAE	0.5226	0.4242	0.5036	0.4596
RMSE	0.6899	0.7881	0.6880	0.6584

Taking all these results to account we can say MLR is best for predicting CGPA and SVR with linear kernel is best for predicting programming skill.

Suppose 'X' and 'Y' trying to figure out their eligibility in CS. **Table 8 shows** their data by important features from both programming skill and CGPA.

Table 8. Example of CGPA and Programming Skill prediction

Features	**Values**		
Name (Real Name not used)	X	Y	Z
Rating of programming skill (out of 10)	?	?	?
Number of participated math or science Olympiad	Never	More than once	Less than 5
Capability of self-study (out of 10)	3 to 4	5 to 6	9 to 10
Interest in Competitive Programming	1 to 2	3 to 4	5 to 6
Number of online courses followed	Never	Once	Many
CGPA	?	?	?
Rating of patience (out of 10)	3 to 4	3 to 4	9 to 10
Skill in Higher Level Mathematics	Poor	Satisfactory	Good
Reason behind internet browsing	Social Media Only	Social Media only	Online Study and Social Media
Skill in Basic mathematics	Satisfactory	Good	Very Good
Result in ICT	Satisfactory	Good	Good

Table 9. Predicted CGPA and programming skill of example data

Student	**CGPA**	**Programming Skill**
X	2.8	1 to 2
Y	3.3	3 to 4
Z	3.8	9 to 10

6 Conclusions

We can see CGPA and programming skill are connected with each other for most of the case. While selecting important features we have found few rules like students with good CGPA have good programming skill. For good programming skill students need patient, self-study, good skill in higher level mathematics etc. For predicting the CGPA multiple linear regression is the efficient one because it has the lowest error rate for both RMSE and MAE. And for predicting programming skill Support vector regression is more efficient than other algorithms. C4.5 gave less error but it was found over fitted for our dataset.

Contribution

The main contributions of this work are:

- We have discovered eleven most influential features to get success in CSE
- We have predicted student's final result (CGPA) and programming skill.
- We have discovered factors behind a good CGPA and Programming Skill

Future work

Other factors like course number of a specific course can be taken to predict the final result for the student of other discipline. Technology is changing day by day. Also the educational system. A lot more improvement can be done to this study. Study can be done taking data of other students from other important discipline like Bachelor of Medicine, pharmacy, Electronic engineering etc.

References

[1] Daily Industry Bangladesh becomes 4th largest remittance source for India. Retrieved from http://www.dailyindustry.news/bangladesh-becomes-4th-largest-remittance-source-india, 2 July 2018

[2] B.K. Baradwaj, P. Saurabh "Mining educational data to analyze students' performance." *arXiv preprint arXiv:1201.3417*, 2012

[3] T. Beaubouef "Why computer science students need math.", *ACM SIGCSE Bulletin*, 34, no. 4, 2002, pp. 57-59.

[4] A. Daud, A. Naif Radi, R. Ayaz Abbasi, M. D. Lytras, F. Abbas, and J. S. Alowibdi. "Predicting student performance using advanced learning analytics.", *Proceedings of the 26th International Conference on World Wide Web Companion,* pp. 415-421. International World Wide Web Conferences Steering Committee, 2017.

[5] M.Ramaswami, R. Bhaskaran. "A CHAID based performance prediction model in educational data mining." *arXiv preprint arXiv:1002.1144,* 2010

[6] M. Tair, M. Abu, A. M. El-Halees. "Mining educational data to improve students' performance: a case study.", *International Journal of Information*, 2, no. 2 , 2012, pp.140-146.

[7] E. Osmanbegović, M. Suljić. "Data mining approach for predicting student performance." *Economic Review*, 10, no. 1, 2012, pp. 3-12.

[8] Q.A. Al-Radaideh, E. M. Al-Shawakfa, M. I. Al-Najjar. "Mining student data using decision trees.", *International Arab Conference on Information Technology (ACIT'2006),* Yarmouk University, Jordan. 2006.

[9] T. Beaubouef, R. Lucas, J. Howatt. "The UNLOCK system: enhancing problem solving skills in CS-1 students." *ACM SIGCSE Bulletin*, 33, no. 2, 2001, pp. 43-46.

[10] S. Kumar, Venkata Krishna, S. Padmapriya. "An efficient recommender system for predicting study track to students using data mining techniques." *International Journal of Advanced Research in Computer and Communication Engineering*, 3, no. 9, 2014, pp.7996-7999.

[11] C. Romero, M.I. López, J-M. Luna, S. Ventura. "Predicting students' final performance from participation in

on-line discussion forums." *Computers & Education*, 68, 2013, pp.458-472.

[12] Z. Alharbi, J. Cornford, L. Dolder, B. De La Iglesia. "Using data mining techniques to predict students at risk of poor performance.", *SAI Computing Conference (SAI)* , 2016.

[13] M. Goga, S. Kuyoro, N. Goga. "A recommender for improving the student academic performance." *Procedia-Social and Behavioral Sciences* 180, 2015, pp.1481-1488.

[14] P.M. Arsad, N. Buniyamin. "A neural network students' performance prediction model (NNSPPM).", *Smart Instrumentation, Measurement and Applications (ICSIMA), 2013 IEEE International Conference*, pp. 1-5. IEEE, 2013.

[15] S. Huang, N. Fang. "Work in progress: Early prediction of students' academic performance in an introductory engineering course through different mathematical modeling techniques.", *Frontiers in Education Conference* (FIE), 2012, pp. 1-2. IEEE, 2012.

[16] Y. Zhao, "*Data mining techniques*.", 2015.

[17] S. M.Ali, M. R. Tuteja, "*Data Mining Techniques"*, 2014.

[18] D.J. Hand, "Principles of data mining." *Drug safety*, 30, no. 7, 2007, pp. 621-622.

[19] J. Han, J. Pei, M. Kamber, "*Data mining: concepts and techniques*". Elsevier, 2011.

[20] J. R. Quinlan, "Induction of decision trees." *Machine learning*, 1, no. 1, 1986, pp.81-106.

[21] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, B. Scholkopf, "Support vector machines." *IEEE Intelligent Systems and their applications*, 13, no. 4, 1998, pp.18-28.

[22] S. R. Gunn, "Support vector machines for classification and regression." *ISIS technical report*, 14, no. 1, 1998, pp.5-16.

Shahidul Islam Khan obtained his B.Sc. and M.Sc. Engineering Degree in Computer Science and Engineering (CSE) from Ahsanullah University of Science and Technology (AUST) and Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh in 2003 and 2011. He is now a Ph.D. Candidate in the Department of CSE, BUET, which is the highest ranked technical university of Bangladesh. His current fields of research are Data Science, Database Systems, Machine Learning, Data Security & Privacy, and Health Informatics. He has more than thirty published papers in refereed journals and conferences. He is also an Associate Professor in the Dept. of CSE, International Islamic University Chittagong (IIUC), Bangladesh.

Sheikh Arif Ahmed obtained his B. Sc. Engineering Degree in Computer Science and Engineering (CSE) from International Islamic University Chittagong, Bangladesh in 2018. He also worked as an undergraduate Teaching Assistant (TA) in International Islamic University Chittagong. His current research fields are Data Science, Machine Learning, Data mining, Security and Privacy etc. He has two papers in International Conferences.