Predicting Student's Career Using Data Mining Techniques

Md. Yeasin Arafath
Student
Daffodil International University
Dhaka, Bangladesh
arafath4341@diu.edu.bd

Mohd. Saifuzzaman
Lecturer
Daffodil International University
Dhaka, Bangladesh
saifakash.cse@gmail.com

Sumaiya Ahmed
Student
Daffodil International University
Dhaka, Bangladesh
sumaiya4323@diu.edu.bd

Dr. Syed Akhter Hossain
Professor
Daffodil International University
Dhaka, Bangladesh
aktarhossain@daffodilvarsity.edu.bd

***Abstract*— Recently, choosing the appropriate career by monitoring the scope and trends in computer science career paths have been a prime need for all computer science undergraduate youngsters. In this paper we develop a data mining based method where knowing student's insights by studying different academic, technical and interpersonal factors, we can be able to predict student's estimated career including student's strength and weakness. The accuracy of prediction actually depends on the set of relevant skill parameters, interpersonal and academic factors. It helps the teacher's identifying the students who need special attention and allows the teacher to provide appropriate counselling as well as give them a proper guideline for selecting a specific job sector which leads a healthy collaboration between academia and industry.**

Keywords— Accuracy; Career prediction; Data mining; Education; Student.

I. INTRODUCTION

Due to the revolutionary growth of IT industry, more and more students are moving towards computer science to assure a prospective career. As a result more and more graduates are coming out every year. Ensuring jobs for this huge number of graduates is pretty difficult. So it is a prime concern for the universities to ensure a healthy collaboration with the industry which will enhance job opportunities for the student of computer science. Having a proper idea about the running students, their interests, strengths & weaknesses and their prospects is a necessity for the universities. It's pretty difficult for the universities to keep track of each and every students individually because of the huge number of students. The ability of predicting student's career can help the universities to keep track of the students with more ease and have a better understanding about students thus maintain academia-industry collaboration. Besides, some students aren't aware of their own interests and capabilities. So it can also be helpful in a way to ensure the students a proper counseling regarding their career.

As we are living in the data age, data in educational sector is increasing rapidly. Useful information and knowledge about students which can be mined from this vast amount of data, stored in different educational databases, such as, Result Portal, Student Portal, Admission Systems, Registration Systems, Course Management Systems, Library Management Systems and so on. Alike all other sectors, decisions are being made based on data in educational sector these days.

We have used classification to analyze successful alumni data (who are currently in job field) which is collected through a survey and we predict final year student's career based on some quality attributes. We mainly looked into several academic, technical and interpersonal aspects of the alumni during their undergrad period and their current job field. The quality attributes are considered as features and their current job field is considered as class labels. The models are trained with these data and predict the career of the running students who've completed their 3rd year considering their responses on the same quality aspects as test sets. There are different classification techniques available. So we applied multiple classification techniques and did a comparative study among the classifiers regarding their performance. The performance of model is measured by different aspects, such as: accuracy, precision, recall and f-measure.

II. LITERATURE REVIEW

B. Dietz-Uhler and J. E. Hurn [1] show the importance of learning analytics in predicting and improving the student's performance showing the list of universities that used learning analytics and the available learning analytics tools.

R. Ade and P. R. Deshmukh [2] proposed an incremental learning approach to predict student's career choice using a couple of classifiers obtaining 90.8% accuracy. Training

dataset had 1333 records with 14 attributes [2]. First classifier in the pair is for generating the hypothesis and the rest one is for updating the weight. A hypothesis is generated for each of the chunks of the dataset and the final one is selected using weighted majority voting rule [2].

S. Elayidom, Dr. S. Mary Idikkula, and J. Alexander [3] showed an approach to predict job absorption rate and waiting time needed for 100% job placement using linear regression. For waiting time prediction for 100% placement, they calculated placement rate status for a certain batch for a period of every 3 months for each year. They predicted the time needed to attain 100% placement using curve fitting concept and regression modeling.

L. S. Katore, B. S. Ratnaparkhi, and Dr. J. S. Umale [4] proposed C4.5 algorithm to recommend and predict career based on personal traits. They started with 110 instances with 12 attributes. The C4.5 obtained best accuracy of 86% amongst multiple classifiers.

B. K. Bhardwaj and S. Pal [5] proposed Naïve Bayes classifier for student's performance prediction. They had 16 attributes initially. But they came up with 7 after filtering attributes based on high potentiality of the variable.

B. K. Bhardwaj and S. Pal [6] proposed ID3 to predict end semester marks of the students based on attributes: 'Previous Semester Marks', 'Class Test Grades', 'Seminar Performance', 'Assignments', 'General Proficiency', 'Attendance', and 'Lab Work'.

A. A. Saa [7] did a study to predict the student performance using classification. The data was collected via survey and initially 270 responses are recorded with 24 attributes. CART decision tree algorithm gave the best accuracy amongst the rest of the classifiers.

S. K. Yadav and S. Pal [8] did a study to predict student's result whether they're going to pass, fail or promoted to next year using classification. Three different classification techniques (C4.5, ID3 and CART) are used. The most accuracy attained by the c4.5 algorithm (66.778%).

III. Proposed Methodology

This study aims at predicting an estimated career of the running CS student's by analyzing successful alumni data considering different important parameters. These important parameters mostly emphasize on professional skill, interpersonal skill and academic records to ensure an effective prediction. The data then analyzed using classification techniques to predict student's career. Fig.1. describes the entire working flow as shown below.

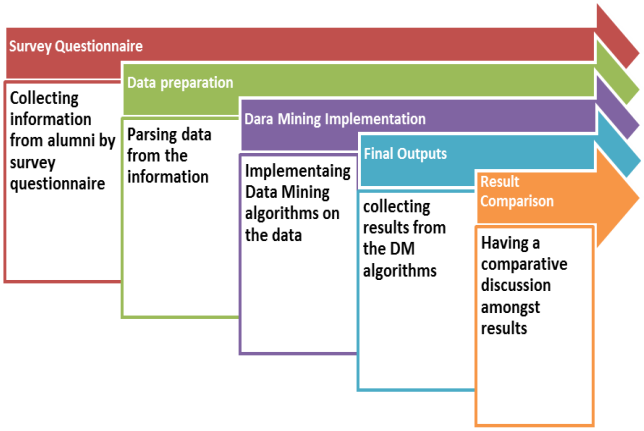

Fig.1. The steps of extracting knowledge from data

A. *Dataset preparation*

The dataset for this study is collected by an online survey from the former computer science students of 13 different universities of Bangladesh who are currently serving the industry. The dataset contains information about their academic, technical and interpersonal factors. Initially the dataset has 506 records.

B. *Data description*

In this section, we only showed the final features after pre-processing and these features are ready to be used for the data mining process. The dataset has 9 variables (8 feature variable and 1 class). Table I describes the features with their description of the dataset. The feature values are encoded with numeric values to help them fit into all models.

Table I: Attributes with their possible values

Variable	**Description**	**Possible Values with numerical equivalents**
PSS	Problem solving skill	Good (2), Medium (1), Poor (0)
PS	Professional skill	Application Development (Web / Mobile / Desktop) (1), Computer Networking (2), Database Administration (3), Designing (4), System Administration (5), Competitive programming (6), Cyber security (7), Game Developing (8), Data analysis / Big data management / Data Mining (9), Artificial Intelligence / Machine Learning / Deep Learning (10), IT support (11), None (0).

En	Enthusiasm	Good (2), Medium (1), Poor (0)
RB	Research Background	Yes (1), No (0)
FYPT	Final Year Project Type	Thesis(0), Project(1), Intern(2)
TA	Teamwork ability	Good(1), Not Good(0)
CS	Communication skill	Good(1), Not Good(0)
CGPA	Cumulative Grade Point Average	High(2), Medium(1), Low(0)
JF	Current Job Field	Software Engineer or Developer (Web / mobile / Desktop)(1), Programmer(2), Database Admin(3), Network Admin / Engineer(4), System Admin / System Engineer / devOps Engineer(5), IT Support Engineer / IT Management(6), Data Scientist / Analyst / Researcher(7), Teaching Profession(8), Obtained Scholarship for Higher Studies(9), Non-technical field(10), UI/UX Designer(11), haven't found any job yet(0)

Here is some precise description of the attributes:

- **PSS**: 'Problem Solving Skill' is estimated by the competitive programming background of a student. PSS is considered to be 'Poor' if, No. of programming contests attended < 2 and no. of programming problems solved < 50. For value 'Medium': No. of contests is between 1 and 5 and solves are between 50 and 200. Anything that is better than 'Medium' is considered to be 'Good'.
- **PS**: 'Professional Skill' is the skill that an undergrad IT student can possibly obtain during his/her academic period. Values of 'PS' are set by researching the university course curriculum of different IT courses, such as: CSE, SWE, CS etc.
- **En:** 'Enthusiasm' is described by the number of projects done by the student on his/her professional skill. In this paper, the value is 'Poor' if the no. of projects < 2. 'Medium' if the no. of projects is between 1 and 4. Anything that better than 'Medium' is considered as 'Good'.
- **RB:** 'Research Background' is estimated by student's involvement in research. Value 'Yes' is considered if no. of research paper publication is at least 1. Else RB is considered to be 'No'.
- **FYPT:** Final Year Project Type values are selected by the direct response of the student.
- **TA:** TA refers to Teamwork Ability. TA is estimated by the number of projects done with team by the student during his undergrad period.
- **CS:** Communication Skill is measured by analyzing student's involvement into different extra-curricular activities and clubs/organizations.
- **CGPA:** student's CGPA is binned it into 3 possible values: 'High', 'Medium' and 'Low'. The value is 'High' if CGPA >= 3.5. Value is 'Medium' if CGPA is between 3 and 3.5. Anything bellow 'Medium' is considered 'Low'.
- **JF:** JF is the class. It stands for Job Field of the alumni. The values are set by researching the current industry situation for the IT graduates.

C. Implementation of data mining model

In this section, we ran multiple classification algorithms on our dataset and predicted an estimated career of the student. Running multiple classifiers has enabled us to have a comparative discussion amongst the performances of the predictive models. We measured the outcomes of different models with these performance measuring criterions: Accuracy, Precision, F-Measure and Recall. We verified the accuracy using an efficient model evaluation technique named 10 Fold Cross Validation. We used the Sckit Learn library of python to implement data mining algorithms [17].

I. **ID3**: ID3 worked fine with our dataset as all the variables of our dataset are categorical. And we preprocessed our data and cleaned all the noisy instances from our data before running ID3 as ID3 can't handle noisy data. Parameters set for the ID3 for this study are following:

- Gain_ratio = True (Gain Ratio is used as splitting criterion.)
- Min_samples_split = 2 (Minimum no. of samples to split on is 2)
- Is_repeating = False (We didn't use repeating features)
- Prune=True (We pruned the tree)

Fig.2. shows the confusion matrix that we generated after running ID3 on our data using Sklearn library and plotted using Matplotlib library of python [17].

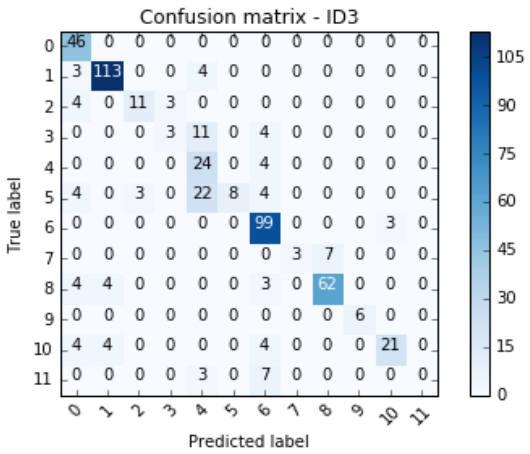

Fig.2. Confusion matrix - ID3

Index 0 to 11 refers to the numeric equivalents (given in Table: I) of classes.

I. **CART**: We ran CART (Classification and Regression Tree) on our data pretty comfortably. As our data is already preprocessed and the variables are categorized, CART didn't need any extra effort to clean noisy data and deal with continuous variables which CART is capable of. Parameters set for CART for this study are following:

- Criterion = Gini (Gini Impurity is used as a splitting criterion. And to measure split quality, gini function is used.)
- Splitter = Best (The best split is chosen at each node.)
- Min_samples_split = 2 (Minimum no. of samples to split on is 2)
- Min_samples_leaf = 1 (Minimum number of samples to be at leaf node.)

Fig.3. shows the confusion matrix of CART.

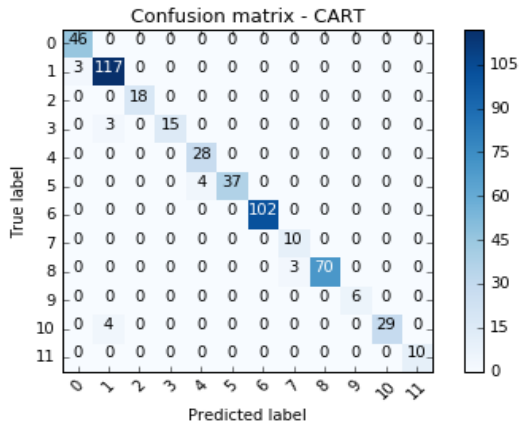

Fig.3. Confusion matrix – CART

II. **Random Forest**: In our Random Forest, we kept 50 tree classifiers. Gini Index technique was used to measure attributes by each tree classifier. Following values of parameters are set for RF:

- n_estimators = 50 (No. of trees in the forest is 50.)
- Criterion = gini (Gini Impurity is used as a splitting criterion. And to measure the split quality, gini function is used.)
- min_samples_split = 2 (Minimum no. of samples to split on is 2.)
- Min_samples_leaf = 1 (Minimum no. of samples at leaf node.)
- Bootstrap = True (Bootstrap aggregating were used in tree building.)

Fig.4. shows the confusion matrix that is generated after running Random Forest on our dataset.

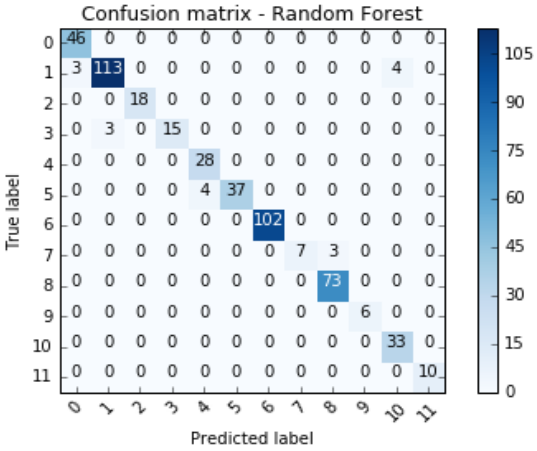

Fig.4. Confusion matrix - RF

III. **Support Vector Machines**: We used Support Vector Classifier on our data. As our dataset is multiclass, 'One vs Rest' method was used. Values set for the parameters for the classifier are following:

- Kernel = 'rbf' (Radical Basis Function is used as kernel type)
- Gamma = Auto (Kernel coefficient used for 'rbf' is 1/n_features if 'auto' is selected)
- Shrinking = True (shrinking heuristic is used.)
- decision_function_shape = 'ovr' (returns a one-vs-rest ('ovr') decision function of shape (n_samples, n_classes).)

Fig.5. shows the confusion matrix that is generated after running Support Vector Machine:

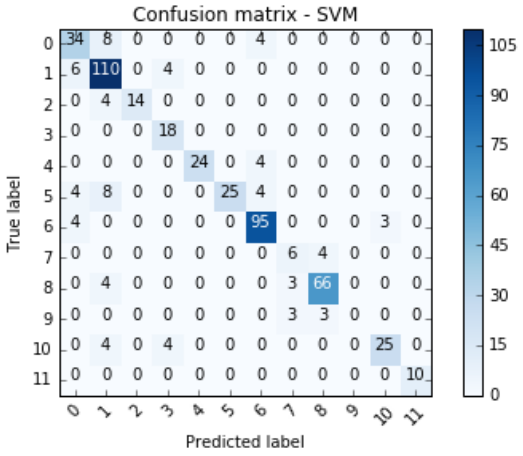

Fig.5. Confusion matrix – SVM

IV. **Neural Networks**: we used Multilayer Perceptron (MLP), a form of feed-forward artificial neural network with a minimum of one hidden layer of nodes besides the input and output layers on our dataset. The following settings are used for MLP in this study:

- Hidden_layer_sizes = (100,) (We stayed with the default: 100 hidden units with one hidden layer)
- Activation = 'relu' (The Rectified Linear Unit function which returns $f(x) = \max(0, X)$ is used as the activation function for the hidden layers.)
- Solver = 'lbfgs' (The solver for weight optimization)
- Learning_rate = 0.001 (We used a constant learning rate of 0.001)

Fig.6. shows the confusion matrix for MLP:

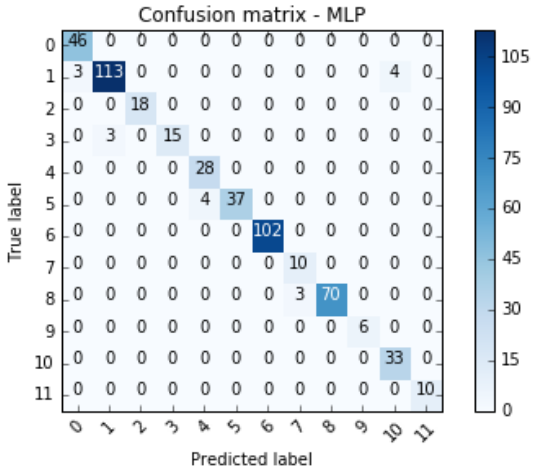

Fig.6. Confusion matrix - MLP

D. Result Analysis

In this section, we did a comparative discussion between the classifiers regarding their results. Four performance measures were selected to evaluate the classifiers. Such as: Model Accuracy, Precision, Recall and F-Measure. As we calculated the confusion matrix for each classifier, we have all the necessary data to get the performance measures values.
Accuracy is very general and common performance measure. The calculation of model accuracy for a model M is,

$$A(M) = \frac{TN+TP}{TN+FP+FN+TP}$$

Here, TP, TN, FP and FN are True Positive, True Negative, False Positive and False Negative respectively. We ran K-Fold Cross Validation (K=10) on the data to find out the model accuracy. Fig.6. shows the accuracy percentage of the models.

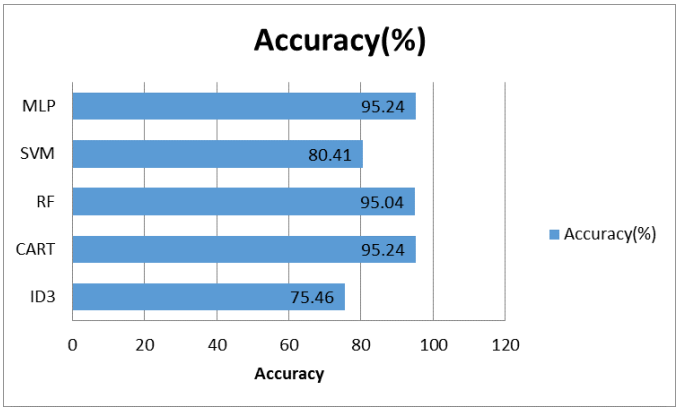

Fig.6. Accuracy percentage

As we can see CART and Multi-Layer Perceptron (MLP) gives us the highest prediction accuracy of 95.24%. Random Forest (RF), the second best classifier gives an accuracy of 95.04%. Other two algorithms, ID3 and Support Vector Classifier give accuracy of 75.46% and 80.41% respectively.

Precision of any classifier is mainly the ability of that classifier of not to predict an actual negative labeled sample as positive [17]. In a word, it is the measure to determine how exact our model is [10]. The best possible value for precision is 1 and the worst possible value is 0 [17]. We calculate precision as following:

$$precision = \frac{TP}{TP + FP}$$

Recall is the measure to determine the completeness [10]. More precisely, it is the percentages of the actual positive samples which are labeled as positive [10]. Best and worst values for recall are same as precision. The calculation for the recall is:

$$recall = \frac{TP}{TP + FN}$$

We calculated the precision scores and the recall scores using sckit-learn library of python [17] and plotted the chart using Microsoft Excel 2010. Fig.7. is the chart for precision-recall measure.

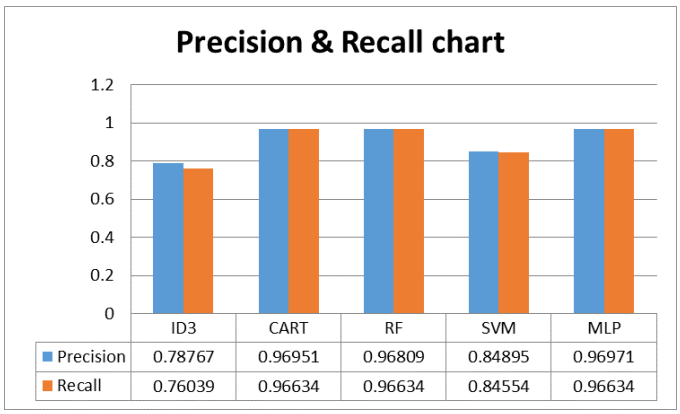

Fig.7. Precision & Recall chart

From the chart, we see that CART, Random Forest and MLP gives the highest precision and recall score (almost 1).
Now, we can do a little bit better with the help of F-beta measure by using both precision and recall scores of a model to do a better comparison amongst the models. F_β measure is calculated using both the P (precision) and R (recall) scores which assigns β times as much weight to recall as precision[19]. We calculate F-beta as following:

$$F_\beta = \frac{(1+\beta^2) * precision * recall}{precision + recall}$$

However in this problem we want equal weight or importance to the precision and recall. So, we have to assign $\beta = 1$. So the equation becomes the simple harmonic mean of the P and R.

$$F_1 = \frac{2 * precision * recall}{precision + recall}$$

Fig.8. shows the f-measure of the models.

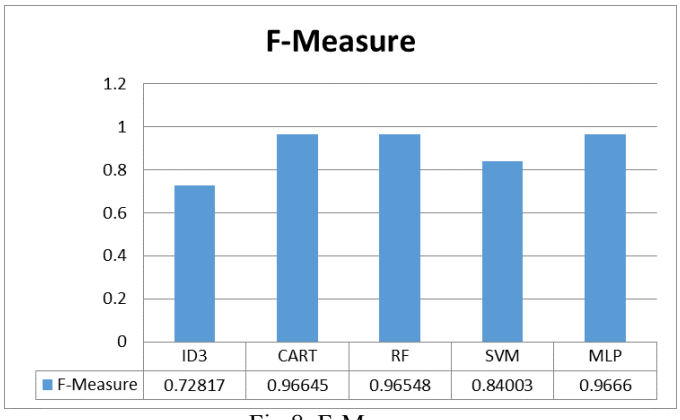

Fig.8. F-Measure

As we can see, CART, RF (Random Forest) and MLP has the highest and almost the same F-measure score.

IV. Conclusion

The aim of this research is to help the university authority to have a more precise understanding about their CS under-graduating students by studying different academic, technical and interpersonal factors of the students and predicting an estimated career of them. Ability of predicting student's career will eventually help the university authority in maintaining their collaboration with the industry by serving proper skilled CS engineers to the industry. It'll also enable to ensure proper counseling and training sessions for both the prospective students and the ones who are unaware of their career. A survey was performed on the alumni to collect academic, technical, interpersonal, and job information from them who are currently serving the industry. Different predictive models were implemented on the data to get the result. Five classification algorithms were implemented on the data and interesting predictions were found. Then we did a comparative discussion amongst the classifiers to evaluate their performances. However, from this study, we see that the prospective career of CS graduates doesn't depend only on the academic or technical quality of the student. Rather it also depends on different interpersonal and social skills.

V. Future Work

We need enormous amount of data for real time data mining and apply more algorithms in future to make it more efficient and effective.
By ensuring a bigger dataset, we could also apply association rules as well as clustering to find out interesting patterns which can be able to improve the performance.
In future this research can be enhanced into an intelligent system.

References

[1] B. Dietz-Uhler, J.E. Hurn, "Using Learning Analytics to Predict (and Improve) Student Success: A Faculty Perspective," Journal of Interactive Online Learning 2013; 12:17-26.

[2] R. Ade and P. R. Deshmukh, "Efficient Knowledge Transformation System Using Pair of Classifiers for Prediction of Students Career Choice," International Conference on Information and Communication Technologies (ICICT 2014).

[3] S. Elayidom, Dr. S. M. Idikkula, and J. Alexander, "Applying Data mining using Statistical Techniques for Career Selection," International Journal of Recent Trends in Engineering, Vol. 1, No. 1, May 2009.

[4] L. S. Katore, B. S. Ratnaparkhi, and Dr. J. S. Umale, "Novel Professional Career prediction and recommendation method for individual through analytics on personal Traits using C4.5 Algorithm," 2015 Global Conference on Communication Technology (GCCT 2015).

[5] Brijesh Kumar Bhardwaj and Saurabh Pal, "Data Mining: A prediction for performance improvement using classification," (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4, April 2011.

[6] B. K. Bhardwaj and S. Pal, " Mining Educational Data to Analyze Student's Performance," (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[7] A. A. Saa, “Educational Data Mining & Students’ Performance Prediction,” (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5, 2016.

[8] S. K. Yadav and S. Pal, “Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification,” World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 2, 51-56, 2012.

[9] R. S.J.D. Baker & K. Yacef, “The State of Educational Data Mining in 2009: A Review and Future Visions,” Journal of Educational Data Mining, Article 1, Vol 1, No 1, Fall 2009.

[10] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, Waltham: Morgan Kaufmann, 2012. Print.

[11] J. R. Quinlan. C4.5: Programs for machine learning. San Francisco: Morgan Kaufmann, 1993.

[12] Quinlan, J. R. Induction of Decision Trees. Machine. Learning. 1, (Mar. 1986), 81–106

[13] Breiman, Leo, Jerome Friedman, R. Olshen and C. Stone (1984). Classification and Regression Trees. Belmont, California: Wadsworth.

[14] L. Breiman. Random forests. Machine learning, 45(1):5–32, 2001.

[15] Feller, W. "The Strong Law of Large Numbers." §10.7 in An Introduction to Probability Theory and Its Applications, Vol. 1, 3rd ed. New York: Wiley, pp. 243-245, 1968

[16] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[17] Cortes, C. and Vapnik, V. 1995. Support-vector network. Mach. Learn. 20, 273–297