**Abstract:**

Due to the revolutionary growth of the IT industry and the increased dependency on IT products and services, quality computer science graduates are vital demand for the industry. So, ensuring good quality computer science graduates is a great concern for the universities. Knowing student’s insights by studying different academic, technical and interpersonal factors of the students and predicting their career is very important for the universities to understand student’s prospects and ensure a proper guideline for them. Ability of predicting student’s career also enables the universities to serve proper skilled manpower to the industry thus helping them to maintain their collaboration with the industry. In this study, several academic, technical and interpersonal factors of the CS graduates were looked into to predict an estimated career dimension for them using data mining techniques.

**Introduction:**

As the IT industry is booming all over the world, 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 a difficult task for the universities. The size of the job market is much smaller compared to the number of graduates in Bangladesh. So it is a prime concern for all the universities of Bangladesh that offers Computer Science to the students to have a collaborative environment with the industry to help students getting jobs. So it is very important for the universities to have a proper idea about the running CS students, their interests, strengths & weaknesses and their prospects .This will help the university to serve proper manpower to the industry. But the number of students being pretty large, it is not an easy task for the universities to keep track of each and every students individually. The ability of predicting student’s career can serve the purpose of better understanding the students to help maintaining 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 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, and Library Management Systems and so on. So, alike all other sectors, decisions these days are being made based on data in educational sector also.

We used classification to analyze successful alumni’s data (who are currently in job field) which is collected through a survey and predicted final year student’s career based on some quality attributes. We mainly looked into different quality 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 model performance is measured by different aspects, such as: accuracy, precision, recall, f-measure, correctly classified instances and incorrectly classified instances.

**Review research:**

Beth Dietz-Uhler & Janet E. Hurn [1] show the importance of using learning analytics in predicting and improving the student’s performance showing the list of universities that used learning analytics, the learning analytics tools that are available and the way how faculty can make use of data to monitor and predict student performance.

Roshani Ade & P. R. Deshmukh [2] proposed an incremental learning approach for prediction of student’s career choice using pair of classifiers obtaining 90.8% accuracy. Students’ scores from the psychometric test have been used as training dataset and the dataset contains 1333 records with 14 attributes [2]. First classifier in the pair is for generating the hypothesis and the second one is for weight updating. The dataset is divided into several chunks and the hypothesis is generated for each of the chunks. The final hypothesis is selected using weighted majority voting rule [].

Sudheep Elayidom, Dr. Sumam Mary Idikkula, and Joseph Alexander [3] conducted a research to predict job absorption rate and waiting time needed for 100% job placement using linear regression to get the percentage of students placing in a particular branch in a certain year in the future. For waiting time prediction for 100% placement, they calculated placement rate status for a particular 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.

Lokesh S. Katore, Bhakti S. Ratnaparkhi and Dr. Jayant S. Umale [4] proposed C4.5 algorithm for career prediction and recommendation method based on personal traits. They started with 110 instances with 12 attributes. Values of the attributes are gained from the answer of questions. the C4.5 achieved the highest accuracy of 86% amongst multiple classifiers.

Brijesh Kumar Bhardwaj and Saurabh Pal [5] conducted a research on student’s performance prediction using Naïve Bayes classifier. They had 16 attributes initially. But they came up with 7 attributes (Students grade in Senior Secondary Education, Living Location, Medium of Teaching, and Mother’s Qualification, Students other Habit, Family annual income status and Students family status) after filtering attributes based on high potentiality of the variable.

Baradwaj and Pal [6] conducted a research on performance prediction of the students based on attributes: ‘Previous Semester Marks’, ‘Class Test Grades’, ‘Seminar Performance’, ‘Assignments’, ‘General Proficiency’, ‘Attendance’, ‘Lab Work’ and ‘End Semester Marks’. The initial data size was 50. Based on the passed out student’s data, they predicted the existing student’s ‘End Semester Marks’ using ID3 decision tree algorithm.

Amjad Abu Saa [7] conducted a research on performance prediction of the students using data mining. The data was collected via survey and initially 270 responses are recorded with 24 attributes. CART decision tree algorithm gave the highest accuracy amongst multiple algorithms.

Surjeet Kumar Yadav & Saurabh Pal [8] conducted a research to predict the number of students who are likely 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%).

**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.

1. *Dataset preparation*

The dataset used in this study is collected from the former computer science students of 13 different universities of Bangladesh who are currently serving the industry via online survey using Google forms. Initially the dataset has 500 records.

1. *Data description*

In this section, we only showed the resultant 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 1 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:

|  |  |  |
| --- | --- | --- |
| 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 detailed description of the attributes given in the table:

* **PSS**: ‘Problem Solving Skill’ is measured by the competitive programming background of the. PSS is considered to be ‘Poor’ if the no. of programming contests attended and no. of programming problems solved are less than 2 and 50 respectively. For value ‘Medium’: No. of contests is between 1 and 5 and solves are between 50 and 200. Anything better than ‘Medium’ is considered to be ‘Good’.
* **PS**: ‘Professional Skill’ is mainly the skill that an undergrad IT student can possibly obtain during his/her academic period in Bangladesh. Values of ‘PS’ are set by researching the university course curriculum of different IT courses, such as: CSE, SWE, CS etc.
* **En:** ‘Enthusiasm’ is measured 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 is less than 2. ‘Medium’ if the no. of projects is between 1 and 4. Anything better than ‘Medium’ is considered as ‘Good’.
* **RB:** In this study, RB or Research Background is basically to check the research background of the students. Research Background is highly related with some of the class values. 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. In this study, TA is measured by the number of projects done with team by the student during his undergrad period.
* **CS:** In this study, Communication Skill is measured by analyzing student’s involvement into different extra-curricular activities and clubs/organizations.
* **CGPA:** In this study, student’s CGPA is binned it into 3 possible values: ‘High’, ‘Medium’ and ‘Low’. The value is ‘High’ if CGPA is greater than or equal 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.

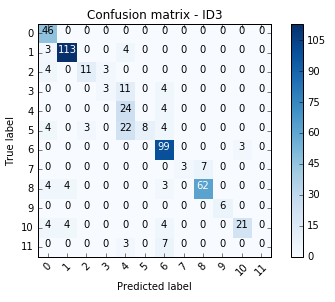
1. *Implementation of Data Mining model*

In this section, we ran multiple classification models on our dataset to predict student’s estimated career in this study. The reason we did it is to have a deeper look at the final output. This also enabled us to have a comparative study amongst the predictive models. We measured the outcomes of different models based on these 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].

1. **ID3**: ID3 worked fine with our dataset as all the variables in 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 has been 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)

We generated the following confusion matrix after running ID3 on our dataset. The matrix is generated using Sklearn library and plotted using Matplotlib library of python [17].

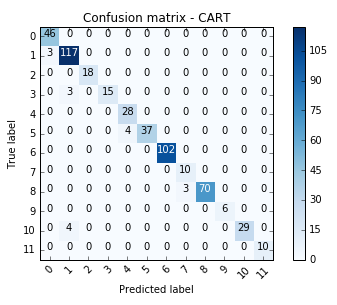


Index 0 to 11 refers to the numeric equivalents (given in Table: I) of classes.

1. **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 has been used as a splitting criterion. And to measure the quality of the split, 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.)

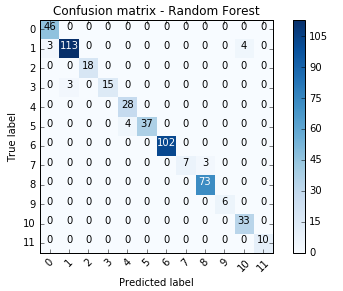
Following confusion matrix was generated after running CART on out dataset:



1. **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 Random Forest Classifier for this study:

* n\_estimators = 50 (No. of trees in the forest is 50.)
* Criterion = gini (Gini Impurity has been used as a splitting criterion. And to measure the quality of the split, gini function is used.)
* 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.)
* Bootstrap = True (Bootstrap aggregating were used in building tree.)

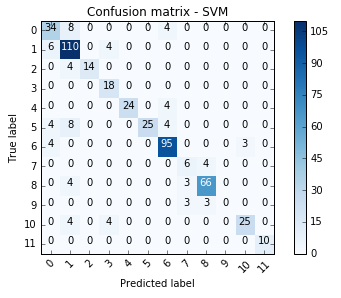
Following confusion matrix was generated after running Random Forest on our dataset.



1. **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 in this study are following:

* Kernel = ‘rbf’ (Radical Basis Function has been used as kernel type)
* Gamma = Auto (Kernel coefficient 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).)

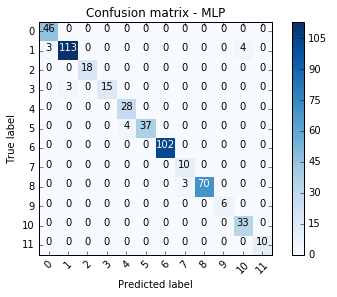
Following confusion matrix was generated after running Support Vector Machine:



1. **Neural Networks**: In this study, 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 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)

Confusion matrix generated for MLP is following:



**Result Analysis:**

In this section, we did a comparative study between the classifiers regarding their results. 4 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 every necessary data to calculate the performance measures.

Accuracy of a classifier is the percentage of test samples that are correctly classified by a classifier on a given test set [19]. The calculation of model accuracy for a model M is,

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.

As we can see Classification and Regression Tree (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 Machine give accuracy of 75.46% and 80.41% respectively.

Precision is another performance measure for classifiers. Precision of any classifier is the ability of that classifier of not to label an actual negative labeled sample as positive [17]. In other words, it is the measure to determine how exact our model is [19]. The best possible value for precision is 1 and the worst possible value is 0 [17]. We calculate precision as following:

On the other hand, Recall is the measure to determine the completeness [19]. More precisely, it is the percentages of the actual positive samples that are labeled as positive [19]. Best and worst values for recall are same as precision. The calculation for the recall is:

We calculated the precision and recall scores using sckit-learn library of python [17] and ploted the chart using Microsoft Excel 2010.

As we can see in the chart, CART, Random Forest and MLP gives the highest precision and recall score (almost 1).

Now, we have both precision and recall measures. Actually 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. measure is basically the weighted harmonic mean of precision and recall which assigns times as much weight to recall as precision[19]. We calculate F-beta as following:

However in this problem we want equal importance to the precision and recall. So, we have to assign So the equation becomes the simple harmonic mean of the precision and recall.

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

**Conclusion:**

The aim of this study is to help the university authority to have a better 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. The ability of predicting student’s career will eventually help the university authority to maintain their collaboration with the industry by serving proper skilled CS engineers to the industry and also serve the purpose of ensuring proper counseling and training sessions for both the prospective students and the ones who are not aware of their career. A survey was performed to collect data from the alumni who are currently serving inside or outside the industry based on the different factors considering these factors as features and their current job position as the target. Different predictive models were implemented on the data to obtain the result. Five classification models were implemented on the data and interesting predictions were found. Then we did a comparative study 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 aspects of the student. Rather it also depends on different interpersonal and social skills.

**References:**

1. UD Beth, HE Janet, “Using Learning Analytics to Predict (and Improve) Student Success: A Faculty Perspective”, Journal of Interactive Online Learning 2013; 12:17-26.
2. Roshani 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. Sudheep Elayidom, Dr. Sumam Mary Idikkula, and Joseph Alexander, “Applying Data mining using Statistical Techniques for Career Selection”, International Journal of Recent Trends in Engineering, Vol. 1, No. 1, May 2009.
4. Lokesh S. Katore, Bhakti S. Ratnaparkhi and Dr. Jayant 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. Brijesh Kumar Bhardwaj and Saurabh Pal, “ Mining Educational Data to Analyze Student’s Performance”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.
7. Amjad Abu Saa, “Educational Data Mining & Students’ Performance Prediction”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5, 2016.
8. Surjeet Kumar Yadav & Saurabh, “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. Ryan S.J.D. Baker & Kalina 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. Han, Kamber, 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.](http://www.amazon.com/exec/obidos/ASIN/0471257087/ref=nosim/ericstreasuretro) 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
18. “8.5 Model Evaluation and Selection.” *Data Mining: Concepts and Techniques*, by Jiawei Han and Micheline Kamber, Elsevier, 2012.