**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.

Data Mining is a technique for finding useful patterns and mining knowledge from large amounts of data. Its popularity in the educational sector is much renowned.

Educational Data Mining is defined as an emerging discipline which is concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students and the settings which they learn in [10].

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. The main goal of research in this field is to discover useful knowledge to help both the administrative bodies regarding managerial purposes and students to get better in what they are pursuing.

Classification is a popular data mining technique to classify and predict class labels of variables. We used this technique 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 show the importance of using learning analytics in predicting and improving the student’s performance from a faculty perspective. They show 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[]. They emphasizes on several factors that have impact on the importance of students. Such as: interest, ability, strengths etc.

Roshani Ade & P. R. Deshmukh proposed an incremental learning approach for prediction of student’s career choice using pair of classifiers. Students’ scores from the psychometric test have been used as training dataset and the dataset contains 1333 records with 14 attributes []. The proposed incremental algorithm is an ensemble of a pair of classifiers. 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 []. They have obtained an accuracy of 90.8% for their proposed algorithm [].

Sudheep Elayidom, Dr. Sumam Mary Idikkula, and Joseph Alexander conducted a research to predict job absorption rate and waiting time needed for 100% job placement, for different engineering courses in India. They obtained the data about passed out students from NTMIS (National technical manpower information system) NODAL center in Kochi, India. The attributes extracted from the data are Roll no of the candidate; month and year he joined the company. They used linear regression technique to figure out the percentage of students that will be placed in a particular branch in a particular 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. From this data, with the help of curve fitting concept and regression modeling, they predicted the time needed to attain 100% placement for the given batch. The purpose of job absorption rate prediction is to reduce the troubles of those who are responsible for displaying the statistics and also the students seeking for colleges which can guarantee them a secure future. Waiting time prediction is useful in the sense that more the waiting time for a branch, more will it indicates that intake for the coming years should be reduced.

Lokesh S. Katore, Bhakti S. Ratnaparkhi and Dr. Jayant S. Umale proposed C4.5 algorithm for career prediction and recommendation method based on personal traits. The dataset is collected via questionnaires answered by the students. They started with 110 instances with 12 attributes. Values of the attributes are gained from the answer of questions. They tried several algorithms (Simple Cart, K Star, Naïve Bayes and C4.5) for classification but the C4.5 achieved the highest accuracy of 86%. The aim of the research is to analyze the psychological condition of the students and recommend them career [].

Brijesh Kumar Bhardwaj and Saurabh Pal conducted a research on student’s performance prediction using classification. Predicting a student’s performance is very important in educational environments. Students’ academic performance is based upon diverse factors like personal, social, psychological and other environmental variables. They collected the data of passed out students from different degree colleges and institutions affiliated with Dr. R. M. L. Awadh University, Faizabad, India. 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. They used Naïve Bayes algorithm for classification.

Nikita Gorad, Ishani Zalte, Aishwarya Nandi & Deepali Nayak conducted a research on career counseling using data mining. The purpose of the research was to develop a system that helps a student studying in high school selecting a course for his/her career based on three factors: personality trait, interest and capacity. They collected the data via survey questions performed on the students studying in different courses and achieved the values of the three factors from the answers. They used C5 decision tree algorithm on the dataset to derive the decision tree for different courses. Based on the values of the three factors, the system helps the high school students to choose a course for their career using data mining algorithm.

Baradwaj and Pal 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’. They collected the data from VBS Purvanchal University, Jaunpur (Uttar Pradesh) on the sampling method of computer Applications department of course MCA (Master of Computer Applications) from session 2007 to 2010[]. 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. According to them, predicting student’s performance will help identifying those students which needed special attention to reduce fail ration and taking appropriate action for the next semester examination[].

Amjad Abu Saa conducted a research on performance prediction of the students using data mining. The objective of this study is to predict performance of the students in the upcoming semesters by discovering the relations between students’ personal and social factors, and their educational performance in the previous semester using data mining tasks[]. The data was collected via survey and initially 270 responses are recorded. From the data, 24 attributes are extracted. The model technique used is Classification. Different classification algorithms were run initially but eventually CART decision tree algorithm is selected as the classification model based on highest accuracy score.

Surjeet Kumar Yadav & Saurabh Pal conducted a research on prediction for performance Improvement of Engineering Students using classification []. Three different classification techniques (C4.5, ID3 and CART) are used. The outcome will be the number of students who are likely to pass, fail or promoted to next year. The dataset used for this survey is collected from the enrollment form filled up by the students at the time of admission from VBS Purvanchal University, Jaunpur. The most accuracy attained by the c4.5 algorithm (66.778%). The results provide steps to improve the performance of the students who were predicted to fail or promoted [].

**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 1:

|  |  |  |
| --- | --- | --- |
| 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**: PSS refers to Problem Solving Skill. It is basically measured by the competitive programming background of the student which includes no. of programming contests attended and number of programming problems solved by the individual. In Bangladesh, software companies requires passionate and diplomatic person for their team, more precisely the person with good problem solving skill. In fact, every company related or non-related to IT wants people with good problem solving skill. In undergrad level, students with good competitive programming skill are considered to be more diplomatic and passionate. Possible values of PSS are: Good, Medium and Poor. In this paper, 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**: PS or Professional Skill is mainly the skill that an undergrad IT student can possibly obtain during his/her academic period in Bangladesh. PS has 11 possible values. These values are set by researching the university course curriculum of different IT courses, such as: CSE, SWE, CS etc. The possible values of PS are: Application Development (Web / Mobile / Desktop), Computer Networking, Database Administration, Designing, System Administration, Competitive programming, Cyber security, Game Developing, Data analysis / Big data management / Data Mining, Artificial Intelligence / Machine Learning / Deep Learning, IT support and None.
* **En:** En refers to Enthusiasm. Possible values of En are: Good, Medium and Poor. 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. The possible values are: ‘Yes’ and ‘No’. ‘Yes’ is considered if no. of research paper publication is at least 1. Else RB is considered to be ‘No’.
* **FYPT:** FYPT refers to Final Year Project Type. FYTP has 3 possible values: ‘Thesis’, ‘Project’ and ‘Intern’. These values are selected by the direct response of the student.
* **TA:** TA refers to Teamwork Ability. Industries these days requires person with good teamwork ability. Students with good teamwork ability have better chances to get recruited. However, in this study, TA is measured by the number of projects done with team by the student during his undergrad period. Possible values of TA are: ‘Good’ and ‘Not good’.
* **CS:** CS refers to Communication Skill. It is equally important to have a good communication skill besides other skills. Because every team wants a member who is easy to work with. However, undergrad students can acquire the basics of this skill by taking part into different extracurricular activities (debating, public speaking, performing etc.) and getting involved into different clubs and organizations. In this study, CS is measured by analyzing student’s involvement into different extra-curricular activities and clubs/organizations. Possible values of CS are: ‘Good’ and ‘Not Good’.
* **CGPA:** Cumulative Grade Point Average is the basic criteria to evaluate academic record. In this study, student’s CGPA is discretized and 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. Possible values for JF are: ‘Software Engineer or Developer (Web / mobile / Desktop)’, ‘Programmer’, ‘Database Admin’, ‘Network Admin / Engineer’, ‘System Admin / System Engineer / devOps Engineer’, ‘IT Support Engineer / IT Management’, ‘Data Scientist / Analyst / Researcher’, ‘Teaching Profession’, ‘Obtained Scholarship for Higher Studies’, ‘Non-technical field’, ‘UI/UX Designer’ and ‘Haven't found any job yet’.

1. *Implementation of Data Mining model*

There are various techniques of discovering knowledge from databases. Some of the well-known techniques are: Association Rule Mining, Classification, Clustering, Regression Analysis, Anomaly or Outlier Detection etc.

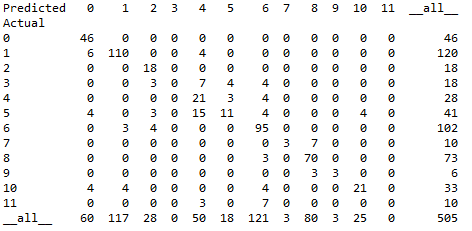
Classification is a very popular and useful technique for Data Analysis which mainly predicts categorical (discrete, unordered) class level [11]. More precisely, classification models predict classes for unknown values learnt from the training dataset with values of known classes. There are plenty of classification models available. Some of them are: K-Nearest Neighbors, Logistic Regression, Decision Tree, Random Forest, Neural Network etc.

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.

1. **ID3**: ID3 (Iterative Dichotomiser 3) is a decision tree algorithm based on hunt’s algorithm, introduced by Quinlan Ross on 1986 []. In ID3, the decision tree is built by splitting the attributes. Information gain is calculated to decide which attribute to split. The splitting process is stopped if a pure subset is found. Only categorical attributes are allowed in building tree models with ID3. ID3 can’t handle noisy data. So preprocessing of data is required before working with ID3. For the tree building process, information gain is calculated for each and every attribute and the attribute with most information gain measure is selected. Continuous attributes must be discretized to be used in ID3. 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 pandas\_ml library of python [17].

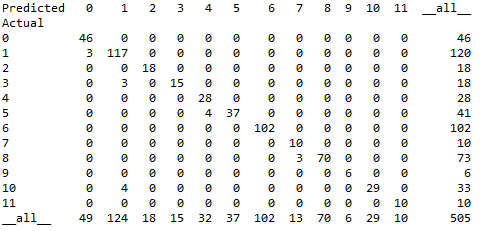


Index 0 to 11 refers to the numeric equivalents (given in Table: I) of classes.

1. **CART**: CART (Classification and Regression Tree) is also a decision tree algorithm introduced by Breiman []. It is also based on Hunt’s algorithm. It selects the attributes to split based on Gini Index measure. CART can handle both categorical and continuous attributes. It also handles missing values. CART produces binary tree as it performs binary split. To avoid over-fitting and eliminating the unnecessary branches from the decision tree, CART performs cost complexity pruning. 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
* Min\_samples\_leaf = 1 (Minimum number of samples to be at leaf node.)

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

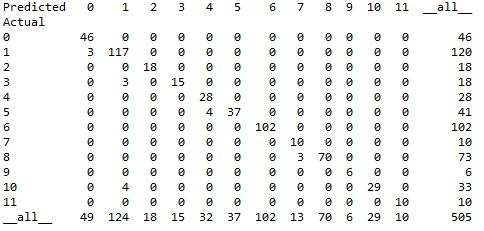


1. **Random Forest**: Random Forest Classifier is a supervised machine learning algorithm that is capable of both classification and regression. Like the name, it is a forest or combination of decision tree classifiers. Each tree classifier is generated using a random vector of inputs which is sampled independently from the input vector [15]. Each tree classifies the input individually which is counted as vote. Forest chooses the classification having most votes. In Random Forest, trees follow the Gini Index technique to measure attributes. Each tree is grown to the maximum depth and there is no pruning. The generalization error gets converged even without the pruning as the number of trees increases [15]. We can ignore over-fitting as a problem because of the Strong Law of Large Numbers [16].

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
* min\_samples\_split = 2
* Min\_samples\_leaf = 1
* Bootstrap = True (Bootstrap aggregating were used in building tree.)

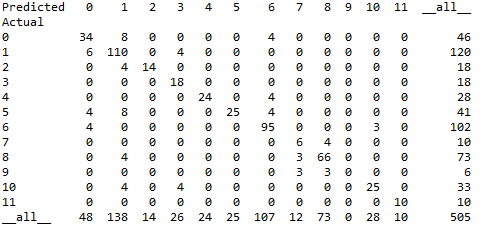
Following confusion matrix was generated after running Random Forest on our dataset.



1. **Support Vector Machines**: Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection [17]. SVMs mainly aim at determining the decision boundary that separates the classes optimally [18]. In a binary classification problem, the SVMs select the linear decision boundary based on the greatest distance between the two classes. The sum of the distances between the hyperplane and the closest points of the two classes is considered to be the margin [18]. To select the optimal decision boundary we need to maximize the margin. To maximize the margin, standard Quadratic Programming (QP) optimization techniques can be used. While dealing with multiple classes, multi-class methods like ‘one against one’ and the ‘one against the rest’ are used for the multi-class problems [18]. The closest data points from the hyperplane are called ‘support vectors’ and they are always small in number [18]. 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).)

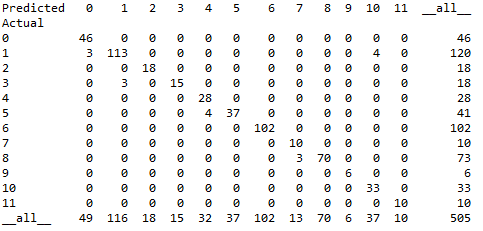
Following confusion matrix was generated after running Support Vector Machine:



1. **Neural Networks**: In this study, we used Multilayer Perceptron (MLP). MLP is a form of feed-forward artificial neural network with a minimum of one hidden layer of nodes besides the input and output layers. Nodes / Neurons of the input layers represent the inputs. Each node of the hidden layer sums the values from the previous layer (x1, x2, x3,……xm) where each values are multiplied with weights (w1, w2, w3, w4,……wm), . Then the nodes use non-linear activation function, to produce the output. The final output is calculated by taking the values from the last hidden layer by the output layer. 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. Than 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. Nikita Gorad, Ishani Zalte, Aishwarya Nandi & Deepali Nayak, “Career Counseling Using Data Mining”, International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 5, Issue 4, April 2017.
7. 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.
8. Amjad Abu Saa, “Educational Data Mining & Students’ Performance Prediction”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5, 2016.
9. 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.
10. 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.
11. Han, Kamber, Pei. *Data Mining: Concepts and Techniques*. Waltham: Morgan Kaufmann, 2012. Print.
12. J. R. Quinlan. *C4.5: Programs for machine learning.* San Francisco: Morgan Kaufmann, 1993.
13. Quinlan, J. R. *Induction of Decision Trees*. Machine. Learning. 1, (Mar. 1986), 81–106
14. Breiman, Leo, Jerome Friedman, R. Olshen and C. Stone (1984). *Classification and Regression Trees*. Belmont, California: Wadsworth.
15. L. Breiman. *Random forests*. Machine learning, 45(1):5–32, 2001.
16. 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
17. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
18. CORTES, C. AND VAPNIK, V. 1995. Support-vector network. Mach. Learn. 20, 273–297
19. “8.5 Model Evaluation and Selection.” *Data Mining: Concepts and Techniques*, by Jiawei Han and Micheline Kamber, Elsevier, 2012.