

**EXPLORATION AND EVALUATION OF SKILLS AND
JOB PROSPECTS OF GRADUATING STUDENTS USING
DATA MINING**

BY

**PARVEZ ZAMIL
ID: 151-15-5392**

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering.

Supervised By

**Dr. Syed Akhter Hossain
Professor and Head**

Department of Computer Science and Engineering
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

May 2019

APPROVAL

This Project/internship titled “EXPLORATION AND EVALUATION OF SKILLS AND JOB PROSPECTS OF GRADUATING STUDENTS USING DATA MINING”, submitted by Parvez Zamil, ID No: 151-15-5392 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 3th May,2019.

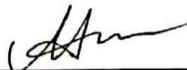
BOARD OF EXAMINERS



Dr. Syed Akhter Hossain
Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



Nazmun Nessa Moon
Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Mr. Abdus Sattar
Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Mohammad Shorif Uddin
Professor

Department of Computer Science and Engineering
Jahangirnagar University

External Examiner

DECLARATION

I hereby declare that, this project has been done by us under the supervision of **Prof. Dr. Syed Akhter Hossain, Head, Department of CSE** Daffodil International University. I also declare that neither this research nor any part of this research has been submitted elsewhere for award of any degree or diploma.

Supervised by:



Prof. Dr. Syed Akhter Hossain
Head
Department of CSE
Daffodil International University

Submitted by:



Parvez Zamil
ID: 151-15-5392
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I would like to express my heartiest thanks and gratefulness to almighty God for His divine blessing for helping us for making possible to complete the final year Research based project successfully.

I am really grateful and wish my profound indebtedness to **Prof. Dr. Syed Akhter Hossain, Head**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of “*Data Mining*” helped us to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to other faculty member for their kind help to finish my research-based project and also to the staff of CSE Department of Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

In twenty first century education is driven towards the achievement of skills. From that perspective in this thesis, I attempted exploring dynamic data set and apply data mining-based methods to explore student's insights based on characteristics related to academic, technical and interpersonal factors. This research helped in the assessment of skills including student's strength and weakness. The accuracy of analysis actually lies with the set of relevant skill parameters and factors. The research also helped teachers identifying the students who need special attention and allowed the teacher to provide appropriate counseling as well as give them a proper guideline for achieving analytical skills which leads a healthy collaboration between academia and industry. The model was tested and found performing well in constraint-based learning environment.

TABLE OF CONTENTS

CONTENS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
CHAPTER	
CHAPTER 1: Introduction	1-4
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the Study	3
1.4 Research Questions	3
1.5 Expected Output	4
1.6 Report Layout	4
CHAPTER 2: Background	5-8
2.1 Introduction	5
2.2 Related Works	5
Research Summary	8
Scope of the Problem	8
2.5 Challenges	8

Chapter 3: Research Methodology	9-16
Introduction	9
Research Subject and Instrumentation	10
Data Collection Procedure	10
Statistical Analysis	12
Implementation Requirements	16
Chapter 4: Experimental Results and Discussion	17-26
Introduction	17
Experimental Results	17
Descriptive Analysis	22
Summary	26
Chapter 5: Summary, Conclusion, Recommendation and Implication for Future Research	27-28
Summary of the Study	27
Conclusions	27
Recommendations	28
Implication for Further Study	28
APPENDIX	29-30
REFERENCES	31-32

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1.1: Work flow of the study	9
Figure3.4.1: Class distribution of the data	13
Figure 3.4.2 Problem Solving skills of CSE graduates	14
Figure 3.4.3 Professional Skills of CSE graduates	14
Figure 3.4.3 Preferable Programming Language of CSE Students	15
Figure 3.4.3 Soft Skill of CSE Students	15
Figure 4.2.1: Confusion Matrix of ID3	18
Figure 4.2.2: Confusion Matrix of CART	19
Figure 4.2.3: Confusion Matrix of Random Forest	20
Figure 4.2.4: Confusion Matrix of SVM	21
Figure 4.2.5: Confusion Matrix of MLP	22
Figure 4.3.1: Accuracy percentage	23
Figure 4.3.2: Precision & Recall chart	24
Figure 4.3.3: F-Measure scores for	25

CHAPTER 1

INTRODUCTION

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 student 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.

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].

I have used classification to analyze successful alumni data (who are currently in job field) which is collected through a survey and I predict final year student's career based on some quality attributes. I 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 4th year considering their responses on the same quality aspects as test sets. There are different classification techniques available. So I 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.

Motivation

Due to the versatility in computer science job dimension, under-graduating computer science students often get confused to choose an appropriate career path. Besides, some of them are lagged behind in terms of technical skills and other factors. So, every university must ensure an effective counseling to the students who are lagged behind or confused and also to the prospective students to make them better. But only academic factors are not enough to get a prospective career in computer science domain these days. So, universities must consider other factors and skills to evaluate their students more precisely. That's why analyzing former students profile considering those factors and skills and their current job field is important to get a clear vision of a successful or failure student's profile in Bangladeshi IT job market. And from these data of the alumni students, universities can get an early signal about their running students regarding their prospects or failure in their career which will eventually enable them to ensure a proper counseling. The ability of predicting student's career considering different academic, technical and interpersonal factors can serve these purposes very satisfactorily. Besides, it can also enable the universities to serve proper skilled manpower to the industry which will help maintaining the academia-industry collaboration.

Rationale of the Study

The aim of this study is to classify the skill level considering technical, analytical and interpersonal skills of final year students of CSE department. Compared to the other related studies, the objective of this study is unique. Besides, in this study, different interpersonal and technical factors are considered besides academic factors for the prediction task which makes it distinct amongst the previous related works. More detailed discussion has been done in chapter 2, section 2.3 regarding this issue.

The problem of knowing students' insights precisely and ensure effective counseling can be solved by the proposed approach of this study. I have discussed it elaborately in the previous section.

As I proposed a data mining-based approach through this study, this study can be further enhanced to an intelligent system.

Research Questions

In this section, the research questions are shown.

- What are the trending programming skills? What are the main factors of student's unwillingness to programming?
- Can I get a precise knowledge about the factors that are really significant points of a programmer by providing some problems to solve?
- Can I get an idea about the skill level of maximum of the students?
- Can data mining models be employed on the collected data to classify the levels of students in programming?
- How efficient the classification results would be?
- How effective will be the research to the university and students?

Expected Output

An effective data mining-based approach that is capable of learning the insights of the running CS students and predict an estimated career of them is one of the expected outcomes of this study. Besides, a comparative study of performance amongst the predictive models for my multidimensional dataset is also expected from this study.

Report Layout

In this section, I'll give a prologue of all the chapters and sections of this thesis paper.

- Chapter 1 – The basic introductory part of the study is in this chapter. It includes introduction, motivation, rationale of the study, research questions and expected outcomes of the study.
- Chapter 2 – The background and related works are discussed in this chapter. Besides, scopes and challenges for my study are also discussed in this chapter.
- Chapter 3 – The Methodology of this study is described in this chapter. It includes data collection procedure, data analysis and implementation requirements.
- Chapter 4 – In this chapter, the implementation and experimental results are shown. Besides analysis of the result is also discussed.
- Chapter 5 – This chapter includes the summary and conclusion of the research. Recommendation and implication of future study is also discussed briefly.

CHAPTER 2

BACKGROUND

Introduction

Predicting student's career is not a whole new concept in the implementation of Educational Data Mining. There have been some related studies conducted previously. In this section, I shall discuss the related works and the works that I reviewed. Then I shall give a summary of the reviewed works. Later on, I shall also discuss scope and challenges of the problem. I shall also try to differentiate between my proposed approach and other existing approaches.

Related Works

The importance of using learning analytics in predicting and improving the student's performance was shown by *Beth Dietz-Uhler&Janet E. Hurn [1]* 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 emphasize on several factors that have impact on the importance of students. Such as: interest, ability, strengths etc.

Roshani Ade & P. R. Deshmukh [2] 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 might updating. The dataset is divided into several chunks and the hypothesis is generated for each of the chunks. The final hypothesis is selected using ighted majority voting rule. They have obtained an accuracy of 90.8% for their proposed algorithm.

Elayidom et al [3] conducted a research to predict job absorption rate and waiting time needed for 100% job placement, for different engineering cmyses in 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. 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.

Katore et al [4] 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&Saurabh Pal [5] conducted a research on student's performance prediction using classification. 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.

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. 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 [7] 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. Different classification algorithms are run initially but eventually CART decision tree algorithm is selected as the classification model based on highest accuracy score.

Yadav & Pal [8] 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 most accuracy attained by the c4.5 algorithm (66.778%). The results provide steps to improve the performance of the students who are predicted to fail or promoted.

Research Summary

From the reviewed works, I see that for career prediction of the students there are some proposed approaches available. One of the studies described above used student's scores from the psychometric test as training dataset for predicting career choice [2]. Another study used personal traits for career prediction and recommendation [3]. Some academic factors of the students are used to predict student's performance [6]. Prediction of job absorption rate and waiting time needed for 100% job placement was also conducted in one of the studies [4].

On the contrary, in my study, I predicted specific job field in computer science domain of the computer science under-graduating students. I considered different academic, technical and interpersonal skills of an individual as features for prediction. The labeled data was collected through an online survey from the former computer science students from 13 different students of Bangladesh who are currently at various job fields. The data was analyzed to study the insights of the students more precisely. Then multiple data mining predictive models I employed on the data to get the prediction result. A comparative discussion amongst the classifiers was also done to evaluate the models.

Scope of the problem

This research can help the university authority to have a better understanding about their running CS undergrads regarding their strengths and weaknesses. It will eventually enable the university authority to ensure proper counseling to both the prospective students and also to those are a bit lagged behind regarding their career opportunities and practicable. Ability of predicting career of the students will also help the university to maintain good collaboration with the industry by serving them with proper skilled manpower.

Challenges

Data collection is the first and foremost challenge. After collecting the survey response, the preparation of the data is also an arduous task. The noisy and redundant instances are needed to be obviated as they can affect the prediction results miserably.

As I discussed earlier, I predicted specific job dimension for the running CS undergrads in computer science domain. The reliability and the efficiency of the research result is a big concern. That is why the feature selection is really a meticulous task. Besides the high dimensionality of data is also a challenge for the predictive models to give effective and accurate result.

CHAPTER 3

RESEARCH METHODOLOGY

Introduction

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.

Figure 3.1.1 shows the work flow of this study briefly.

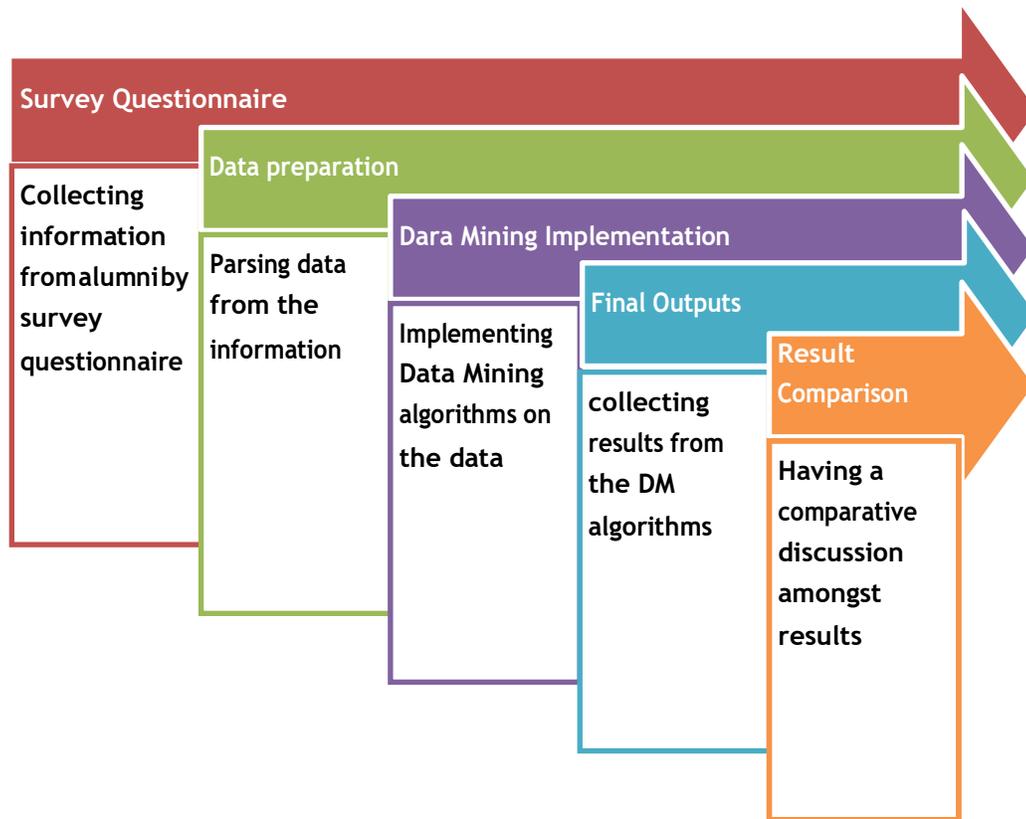


Figure 3.1.1: Work flow of the study

Research Subject and Instrumentation

In this section, I shall talk about my research subject and instrumentation elaborately. My research subject is ‘graduated student’s data considering different their Skills’. I have taken the data from the present & former computer science graduates based on some factors, analyzed those data, implemented data mining models on those data and tried to retrieve information from those actions. So the data of alumni & current student is the research subject.

And the research instrumentation is the survey that I conducted. I conveyed an online survey form to the students to collect data from them. The questionnaire was designed in such way that I could get information regarding their academic, technical and interpersonal factors. I collected the responses and processed them to prepare my dataset.

Data Collection Procedure

A. Dataset preparation

The dataset used in this study is collected from the computer science students of different universities of Bangladesh who are currently serving the industry via online survey using Google forms. Initially the dataset has 250 records.

B. Data description

In this section, I only shown the resultant features after pre-processing and these features are ready to be used for the data mining process. The dataset has 9 variables (20 feature variable and 5 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: Description of the features

Variable	Description	Possible Values with numerical equivalents
C Programming	Problem solving skill	Below Average(1) ,Average(2) Good(3) Very Good(4) Excellent(5)
Java Programming	Professional skill	Below Average(1) Average(2) Good(3) Very Good(4) Excellent(5)
Python Programming	Problem solving skill	Below Average(1) Average(2) Good(3) Very Good(4) Excellent(5)
JavaScript	Problem solving skill	Below Average(1) ,Average(2) Good(3) Very Good(4) Excellent(5)
PHP	Problem solving skill	Below Average(1) Average(2) Good(3) Very Good(4) Excellent(5)
Professional Skills		Below Average(1) Average(2) Good(3) Very Good(4) Excellent(5)

Here is some detailed description of the attributes given in the table:

©Daffodil International University

- **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, SI, CS etc. The possible values of PS are: Application Development (web / Mobile / Desktop), Computer Networking, Database Administration, Designing, System Administration, Competitive programming, Cyber security, Games Developing, Data analysis / Big data management / Data Mining, Artificial Intelligence / Machine Learning / Deep Learning, IT support and None.

Statistical Analysis

In this section, I shall analyze out data statistically. I shall visualize my data by different bar charts and I shall focus on the percentage of programming skill of CSE graduating students.

First of all, I shall have a look at the class distribution of the data which is shown at Figure 3.4.1.

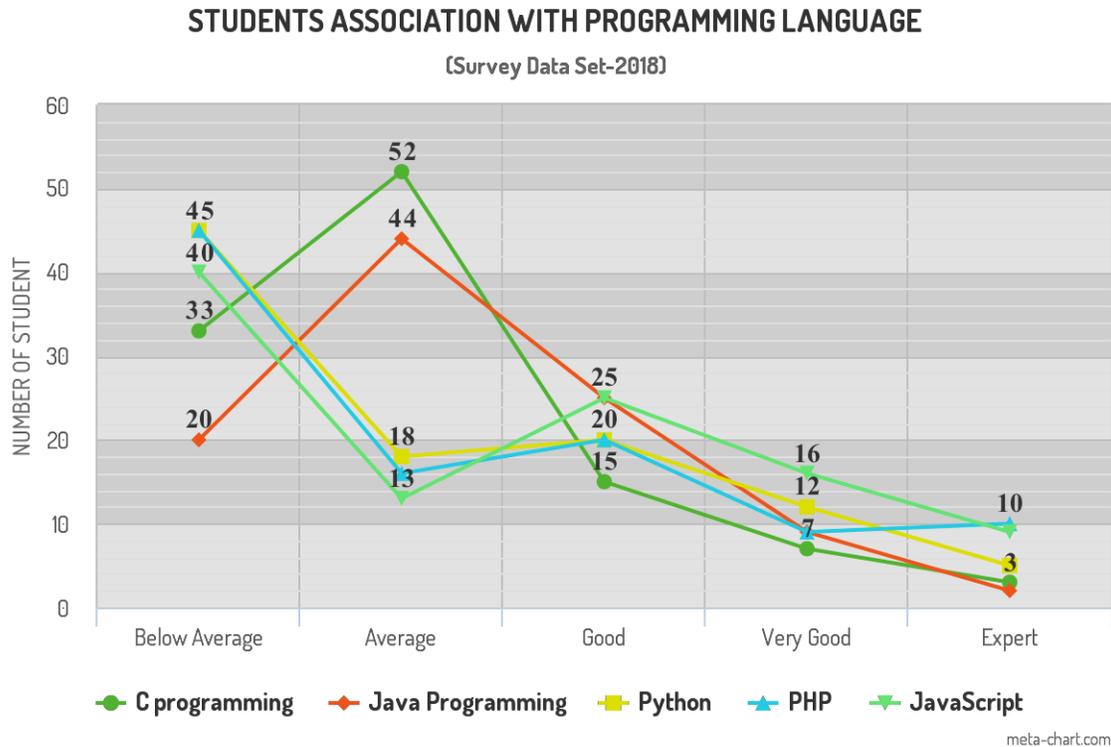


Figure 3.4.1: Class distribution of the data

From the figure I see that the percentage of students pursuing various programming language. Around 40 % of students are with below average skills of programming where only 2-3% students are in expert level of programming. 20-25 % students have Good skill in programming languages.

Students who have good skills in core programming languages have good command in all other programming languages. But the data says that the number of students in core programming field is very few where more students are involved in web development based programming

First of all, I shall look into their technical aspects of them. Figure 3.4.2 shows the problem-solving skill of them.

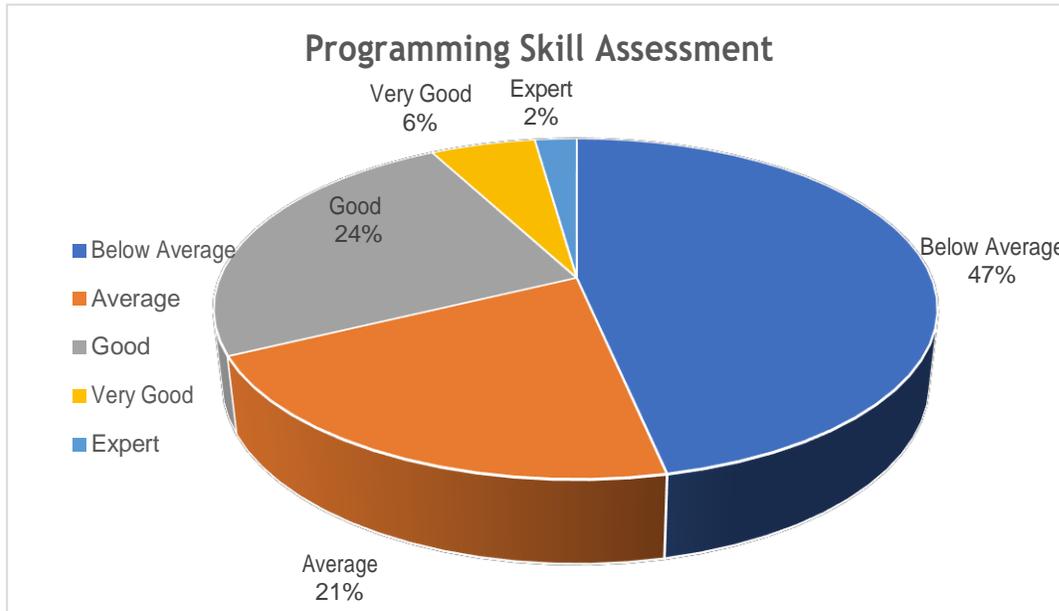


Figure 3.4.2: PSS of CSE graduates



Figure 3.4.3: PS of CSE graduates

I can clearly see that majority (47%) of them had poor problem solving skill. Only 2% of them had expert level skill. Now, I shall have a look at another distribution in Figure 3.4.3, which shows the professional skill set of the unemployed graduates.

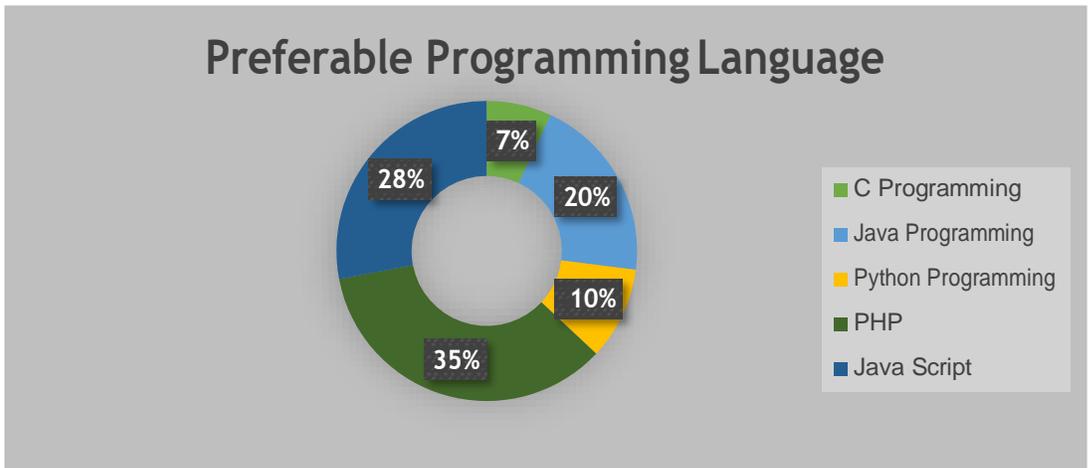


Figure 3.4.4: Preferable Programming Language of students

24% of them had no technical skill. 59% of them had some application development skill. Other 17% had some hands on computer networking. I did not mention other skill sets as there are not any response on those skill sets.

Now, I shall move onto some soft skill sets like 'teamwork ability' and 'communication & networking skill'. Figure 3.4.4 shows both the 'teamwork ability' and the 'communication & networking skill' distribution amongst the unemployed graduates.

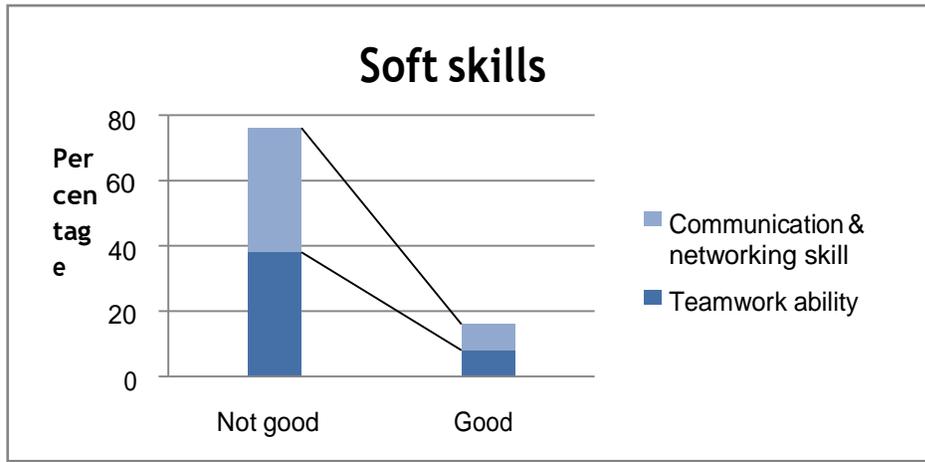


Figure 3.4.5: Soft skills of the unemployed graduates.

As I clearly see in the bar chart, unemployed graduates had mostly poor soft skills. Now, I'll look at their CGPA. Figure 3.4.5 shows the CGPA distribution of the unemployed graduates.

Implementation Requirements

In this section, I discussed the tools that I used for the implementation. First of all, the data was collected through an online survey. In this purpose, Google Forms, a free online survey service of Google is used. Then the data was received as a CSV file (Comma Separated Values). I processed the responses and prepared the dataset using Microsoft Excel 2010. Then I implemented the data mining models on the dataset using Sckit Learn library (a package of machine learning algorithms) of Python [16]. I also used 'Pandas', 'Numpy' and 'Matplotlib' library of python during implementation period. I evaluated the models by calculating different quality measures like accuracy, precision, recall, f-measure using sklearn. Finally I created different bar charts and pi charts for visualizing my data and evaluating the models using Microsoft Excel 2010. All the implementation and procedure was done in a 64 bit, Windows 10 machine.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

Introduction

In this section, I implemented different data mining models on my dataset and found interesting results. Later on, I did a comparative study amongst the models to get better results. 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.

Experimental Results

I ran multiple classification models on my dataset to predict student's estimated career in this study. I did it is to have a better look at the final output. This also enabled me to have a comparative study amongst the predictive models. I measured the outcomes of different models based on these criteria: Accuracy, Precision, F-measure and Recall. I verified the accuracy using an efficient model evaluation technique named 10 Fold Cross Validation.

I. **ID3**: ID3 (Iterative Dichotomiser 3) is a decision tree algorithm based on Hunt's algorithm, introduced by Quinlan Ross on 1986 [12]. 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 (I didn't use repeating features)
- Prune=True (I pruned the tree)

Figure 4.2.1 is the confusion matrix, generated after running ID3 on my dataset. The matrix is generated using Matplotlib library of python.

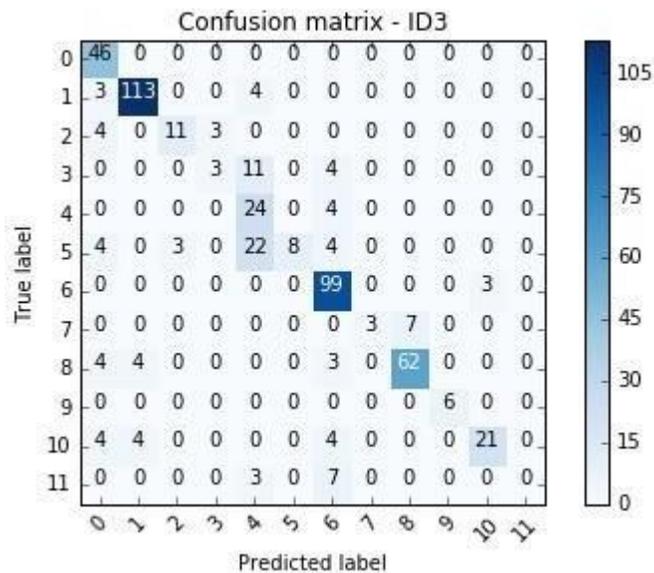


Figure 4.2.1: Confusion Matrix of ID3

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

II. **CART:** CART (Classification and Regression Tree) is also a decision tree algorithm introduced by Breiman [13]. 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.)

Figure 4.2.2 is the confusion matrix generated after running CART on out dataset:

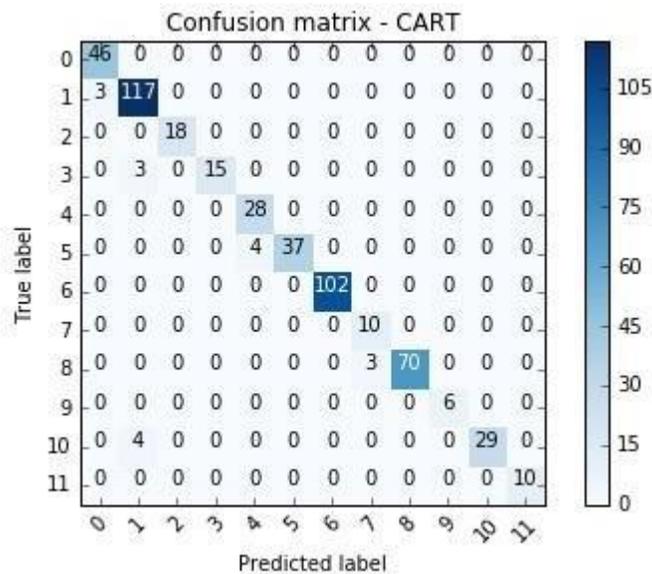


Figure 4.2.2: Confusion Matrix of CART

III. Random Forest: Random Forest Classifier is a supervised machine learning algorithm. 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 [14]. 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

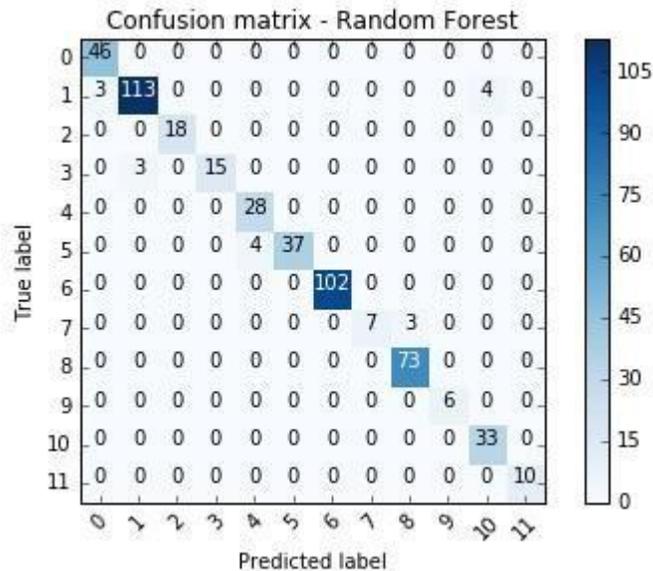
even without the pruning as the number of trees increases [14]. I can ignore overfitting as a problem because of the Strong Law of Large Numbers [15].

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 Ire used in building tree.)

Figure 4.2.3 is the confusion matrix generated after running Random Forest on my dataset.

Figure 4.2.3: Confusion Matrix of Random Forest Classifier



IV. Support Vector Machines: Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection [16]. SVMs mainly aim at determining the decision boundary that separates the classes optimally [17]. 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

[17]. 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 [17]. The closest data points from the hyperplane are called ‘support vectors’ and they are always small in number [17]. 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).)

Figure 4.2.4 is the confusion matrix that was generated after running Support Vector Machine:

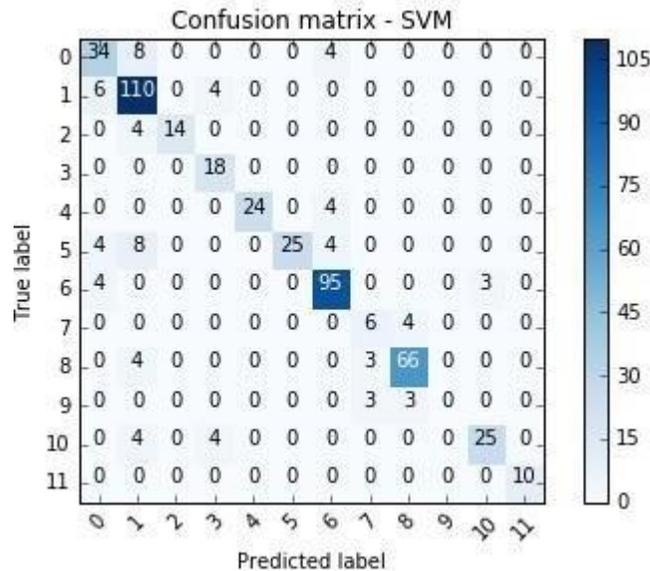


Figure 4.2.4: Confusion Matrix of SVM

V. Neural Networks: In this study, I 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 ($x_1, x_2,$

x_3, \dots, x_m) where each values are multiplied with weights ($w_1, w_2, w_3, w_4, \dots, w_m$), $\sum_{i=1}^m w_i x_i = 0$. Then the nodes use non-linear activation function, $\sigma(\cdot)$: $\mathbb{R} \rightarrow \mathbb{R}$ to produce the output. The final output is calculated by taking the values from the last hidden layer by the output layer. 'Limited-memory BFGS (lbfgs)' function is used as weight optimizer. That's why learning rate isn't necessary. The following settings are used for MLP in this study:

- Hidden_layer_sizes = (100,) (I stayed with the default: 100 hidden units with one hidden layer)
- Activation = 'relu' (The Rectified Linear Unit function which returns $\sigma(x) = \max(0, x)$ is used as the activation function for the hidden layers.)
- Solver = 'lbfgs' (The solver for weight optimization) Figure 4.2.5 is the Confusion matrix generated for MLP:

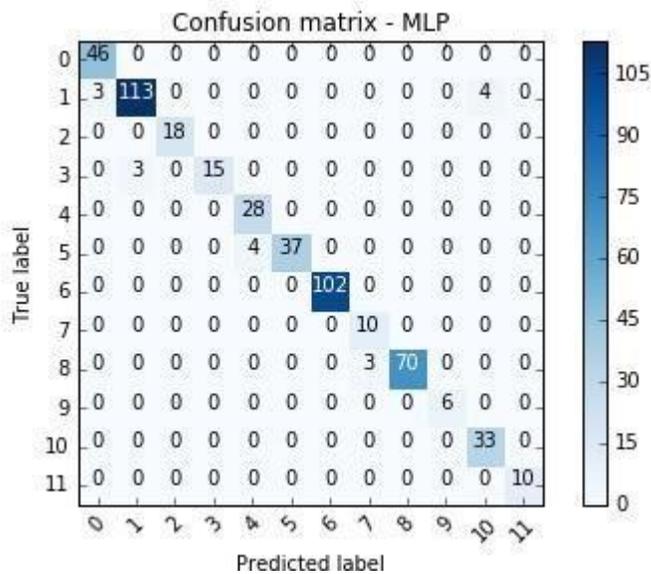


Figure 4.2.5: Confusion Matrix for MLP

Descriptive Analysis

In this section, I did a comparative study between the classifiers regarding their results. 4 performance measures are selected to evaluate the classifiers. Such as: Model Accuracy, Precision, Recall and F-measure. As I calculated the confusion matrix for each classifier, I 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 [18]. Eq. (4.2.1) is the calculation of model accuracy for a model M,

$$\text{Accuracy}(M) = \frac{TN+TP}{N+P+N+P} \dots\dots\dots (4.3.1).$$

Here, TP, TN, FP and FN are True Positive, True Negative, False Positive and False Negative respectively. I ran K-Fold Cross Validation (K=10) on the data to find out the model accuracy. Figure 4.3.1 shows the accuracy percentage of the models.

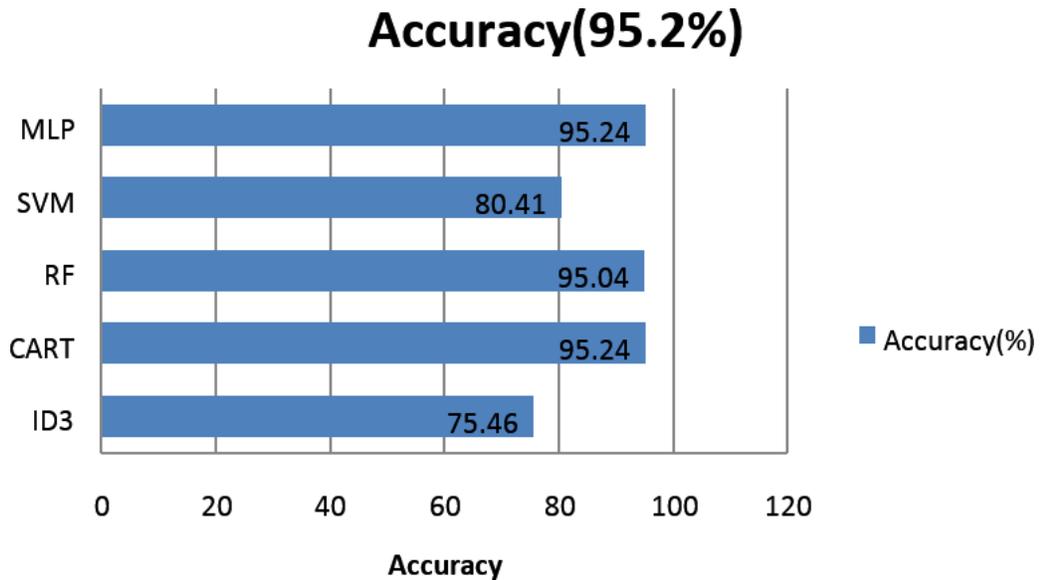


Figure 4.3.1: Accuracy Percentage

As I 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 [16]. In other words, it is the measure to determine how exact my model is [18]. The best possible value for precision is 1 and the worst possible value is 0 [16].Eq. (4.2.2) shows the calculation for precision:

$$precision = \frac{TP}{TP + FP} \dots\dots\dots (4.3.2)$$

On the other hand, Recall is the measure to determine the completeness [18]. More precisely, it is the percentages of the actual positive samples that are labeled as positive [18]. Best and worst values for recall are same as precision. Eq. (4.2.3) shows the calculation for the recall:

$$recall = \frac{TP}{TP + FN} \dots\dots\dots (4.3.3)$$

I calculated the precision and recall scores using scikit-learn library of python [16] and plotted the chart using Microsoft Excel 2010 which is shown in Figure 4.3.2.

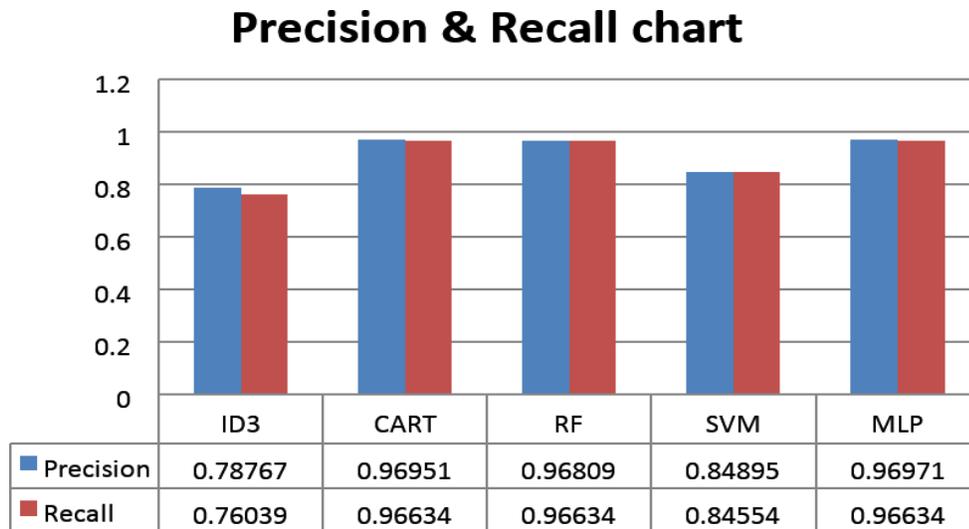


Figure 4.3.2: Precision & Recall chart

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

Now, I have both precision and recall measures. Actually, I 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 basically the weighted harmonic mean of precision and recall which assigns β^2 times as much weight to recall as precision [18]. Eq. (4.3.3) shows the equation for F-beta measure:

$$F_\beta = \frac{(1 + \beta^2) * precision * recall}{precision + recall} \dots\dots\dots (4.3.3)$$

However in this problem I want equal importance to the precision and recall. So, I have to assign $\beta^2 = 1$. So the equation becomes the simple harmonic mean of the precision and recall as shown in Eq. (4.3.4).

$$F_1 = \frac{2 * precision * recall}{precision + recall} \dots\dots\dots (4.3.4)$$

Figure 4.3.3 shows the F-Measure scores of the models:

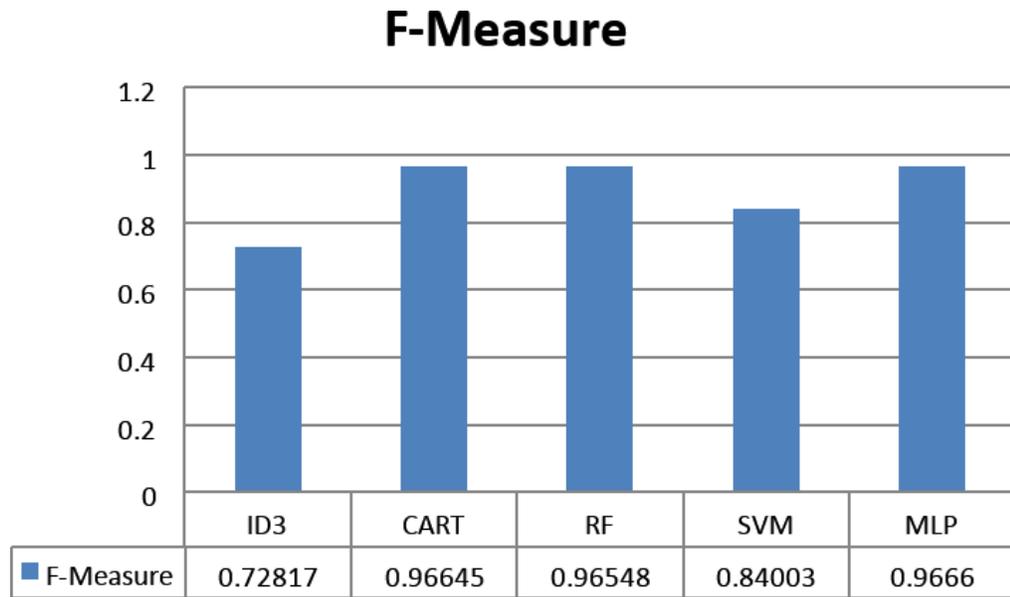


Figure 4.3.3: F-Measure scores

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

Summary

In chapter 4, I have ran multiple classifiers on my dataset. The Scikit-Learn API of python was used to employ the models. I showed the parameters of models that are used for classification in this study. The confusion matrix of each classifier was also generated using Matplotlib library of python after running the classifiers on my data. Later on, I had a comparative discussion amongst the classifiers I ran regarding their performances based on some quality measures (Accuracy, Precision, Recall and F-measure). I found out that CART and MLP gives us the best classification performance for my dataset considering all of my quality measures.

CHAPTER 5

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

Summary of the Study

Studying different academic, technical and interpersonal factors of the students to understand their insights and predicting their estimated career was the main objective of this study. The dataset for the research was collected by survey from the students who are currently at various job positions. I studied different academic, technical and interpersonal factors of them as features for career and their current job sector as label. Five different classifiers I applied on the dataset to get the prediction of the test set. To have a good career in CSE Field students must have technical and interpersonal skill. It's harder to sustain with only technical skills. Having both skills opens more opportunity to the students. This proposed approach of Skill Assessment is expected to help the university authority to have a clear idea of the skills of CS undergrads, provide enough guidelines & maintain good collaboration with the industry by creating skilled individuals.

Conclusion

The mission of this research was to have a clear understanding of technical quality of CSE graduating student by studying different academic, technical and interpersonal factors of the students and categories the students on the basic of their skills. The assessment of the skills of the students will help the authority of have a strong overview of the students. The research is also expected to ensure proper counseling and training sessions for the students who are on a poor skill level. Final result of the research is generated by implementing various algorithms and statistical techniques. Students who had taken their initial stages of programming seriously have shined in almost every other sector. Knowledge of core programming helps a lot to sustain in other technical areas. And technical knowledge with interpersonal skills leads to a balanced career.

Recommendation

To enhance the effectiveness, reliability and efficiency of the study, further acquisition of data is needed. The more the data is, the more reliable the results are. Besides a validation set is also needed to reduce the over-fitting of the models. More advance models can be applied on the data to explore further.

Implication for Further Study

Based on further acquisition of data set, further exploration will be performed for real time data mining and apply enhanced algorithms to make it more efficient and effective. Further with these huge dataset association rules will be developed to explore interesting patterns which can be able to improve the performance. In future this research can be enhanced into an intelligent system.

APPENDIX A

EXPLORATION AND EVALUATION OF SKILLS AND JOB PROSPECTS OF GRADUATING STUDENTS USING DATA MINING

Questionnaire

Please answer the following questions. There are open ended questions, yes-no questions and multiple category questions.

```
<script type="text/javascript">
<!--
function validateEmail()
{
var emailID = document.myForm.Email.value;
atpos = emailID.indexOf("@");
dotpos = emailID.lastIndexOf(".");

if (atpos < 1 || (dotpos - atpos < 2 ))
{
alert("Please enter correct email ID");
document.myForm.Email.focus();
return false;
}
return( true );
}
//-->
</script>
```

- email validation
- email send
- email pup up
- email activation

Question 17: What will be result of following javascript function? *

```
<script type="text/javascript">
var name1 = "WCF quiz";

function DisplayName () {
var name2 = "ASP.NET MCQs";
if(name1 != "")
document.write(name1);
else
document.write(name2);
}
</script>
```

- WCF quiz
- quiz WCF
- Name 1

Question 9: What is the output of the following code? *

```
number = 5.0
try:
    x = 10/number
    print(x)
except:
    print("Oops! Error occurred.")
```

- 2.0
- Oops! Error occurred
- 10.0
- Error

Question 10: What does the following code do? *

```
try:
    # code that can raise error
    pass

except (TypeError, ZeroDivisionError):
    print("Two")
```

- Prints Two if TypeError or ZeroDivisionError exception occurs.
- Prints Two if TypeError
- Prints ZeroDivisionError
- all of the above

Question 11: What is the output of the following program? *

```
def outerFunction():
    global a
    a = 20
    def innerFunction():
        global a
        a = 30
        print('a =', a)
a = 10
```

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, Jerous 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
- [18]. “8.5 Model Evaluation and Selection.” *Data Mining: Concepts and Techniques*, by Jiali Han and Micheline Kamber, Elsevier, 2012.