**Chapter 1**

**Introduction**

## Introduction

Selecting right study track for undergraduate program is very significant for every students because this is something that determines the academic and professional achievement of a student [1]. Lacking of useful information on a particular undergraduate program before choosing it may cause straight failure in the academic and professional life [2]. A lot of student often select undergraduate programs without analyzing their own capability and the level of complexity it encounters. Apart from this, there are lots of indecisive students’ who can’t decide whether they should switch from their background to Computer Science after their under graduation program or they should continue in their respected field.

Computer Science has now become a buzzword in the global community. Being one of the developing countries Bangladesh Government has already taken the challenge of outshining in the ICT department, so as the students. According to University Grants Commission of Bangladesh, Country’s 116 out of 136 universities (32 out of 41 public universities and 84 out of 95 private universities) are offering Computer Science program which indicates that a large number of students are involved with this very program. But when the question of skill, show casing talent and achievements in national and international level comes it seems significantly important percent of students are failing to do so. Many students are rushing into this program without assessing their potential in this program which is ultimately creating a massive number of unskilled and less effective graduates. Many IT industries are hiring IT professionals from India to fill the gap [3].

Purpose of this work is to build a classification model to classify which student have chance in Computer Science and which doesn’t. We evaluate their profiles before jumping into this program using their academic results, experience with ICT course, personal interests, personal experiences of current students’ with similar academic achievements and problem solving skills. To predict students’ performance in any program it is important to take the previous academic results into account [1]. Besides academic results, to make our proposed classifier more accurate in this paper we have collected data from different universities of Bangladesh which includes student’s interests, their experiences with the logical and mathematical courses they have already faced, their experience with ICT course, Online class experiences etc. Firstly, we predicted their final result and their programming skill using regression analysis. Afterwards we implemented K-Means clustering algorithm using elbow method to find the optimal number of clusters in the dataset. We further labeled the dataset and built a classification model to categories the student’s profiles into OK or Not OK profile for Computer Science Program.

## Background

Educational data mining refers to a research field which is concerned with the information produced from educational settings and further application of different machine learning and data mining techniques to these information. To analyze large amounts of data generated from educational settings, EDM has recently increased people’s interest in developing different techniques. In recent past, Students’ final result predictor and Classification model for determining students’ future was built by taking different types of feature sets like as academic record, family income, family expenditure and some other personal information [4, 5, 6, 7]. Different types of classification model were built by implementing Decision tree, Support Vector Machine and Naive Bayes Classifier algorithm on students’ academic results [8, 9]. This work investigates students previous academic results, personal interest, experience with ICT course, participation in online courses, online course experience and their skills in solving problems to build student’ profile. Afterwards we predict their final result and programming skill. Finally, we build a classification model to help students whoever want to join Computer Science Program.

## Motivation

Having ICT favorable governance in Bangladesh a lot of students are taking Computer Science as their undergraduate program to make an impact that it promises. Students are tempted to choose Computer Science as their undergraduate program due to the amount of hype it carries. Without proper analysis substantial amount of students are taking this program and eventually the performance of larger part of students’ taking Computer Science Program is deteriorating. In the end it is hurting their academic and professional life. Besides, students’ from Statistics, Mathematics and different Engineering background want to switch after their under graduation program to join Computer Science for its vast popularity and sophistication. To help students to choose the right program many researcher worked with their personal data, academic data and their family expanses [4, 5, 6, 7]. To make this even better we used few more features including students’ interest, Online Course experience, students’ performance and experience with the ICT course they faced in the higher secondary level to determine students’ chance in Computer Science Program.

## Drawbacks of the Existing Models

Researchers have used different data mining approaches so far to influence students’ decision making in positive sense. They used student’s previous academic result, family income, family expenses, medium of teaching, marital status, parent’s occupation, parent’s qualification, family size, attendance, assignment, lab work to predict their final result and to build different types of classifier models [5, 8, 9]. But they did not take students’ personal interests like online class experience, Internet browsing reasons, Participation in Mathematics or Science Olympiad, Interest in Competitive programming and students problem solving skills into consideration to predict their final results and building classification model.

## 1.5 Definition of Problem

Computer Science Program is attracting lots of students because of the amount of opportunity offers by this program. But rushing into this program is not good for someone who is most likely to feel suffocated after choosing Computer Science as an. There are also substantial amount of students who feel confused whether they should switch to undergraduate program. As, it is not a program for every student to take because the amount of complexity it encounters Computer Science or not. To help students in decision making before they choose this very program, we predict their final result and their programming skill. We further build classification model using eleven most important features from students information which was primarily consists of twenty-two features.

## Objectives

The aim of this research is to predict student’s final result, programming skill afterwards build a classification model from selected features to classify students profile in terms of failure and success before they start their journey with Computer Science program. Therefore, the objectives of this research are as follows:

* To help student by predicting their final result in Computer Science.
* To help predict student programming skill in Computer Science.
* To discover most important features which indicates student’s success in Computer Science.
* To build a classification model and classify student’s profile into its belonging class.

## 1.7 Contribution

The main contributions of this work are:

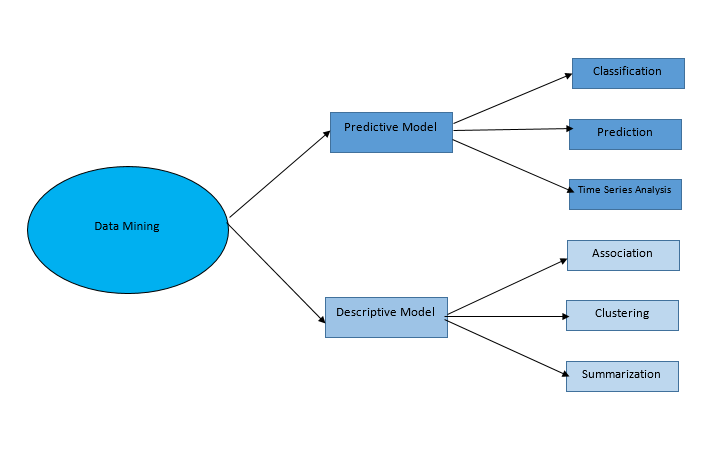
* A productive feature sets of twenty-two features is constructed.
* Predicted students’ final result and their programming skill by using Multiple Linear Regression, SVR and Decision Tree Regression C4.5 algorithm.
* Used different error evaluation techniques and k- fold cross validation technique to test the model’s ability to predict new data that were not used in estimating it.
* Eleven most influential features are shortlisted by using Gain Ratio and Ranker algorithm.
* Developed a classification model to help students’ in making good decision before they choose Computer Science.
* Measured the accuracy and F-score of the model by building a confusion matrix.

## Techniques

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

**1.8.1 Data Mining Techniques**

The term Data mining [10] is a misnomer, as it refers to extracting knowledge from large amount of data, not the extraction of data itself. Data mining techniques are also used to discover hidden patterns from large volumes of data. It is mostly applied to computer decision supporting system [11], AI, business intelligence and for information processing. The data mining techniques is featured to create model which will help to find new data using unknown data. Data mining can be basically of two types- Predictive and Descriptive [12]. Predictive techniques uses known data set for analysis and gather details information about that database. Classification, regression, time series analysis, prediction is predictive. The descriptive technique finds patterns and relations in datasets. Clustering, sequence analysis, summarizations and association rules included in descriptive techniques. **Figure 1.1** shows various data mining techniques.



**Figure 1.1** Data mining techniques

Data mining is of two types according to its class.

* Unsupervised algorithms
  + - * No supervision needed for mining
      * undirected data mining
      * Pattern analysis
      * Discovering pattern automatically
      * No information about class or class label
      * E.g.: Clustering and association rules
* Supervised algorithms
  + - * Class data is known
      * Pre-known information
      * Training data set have data of input and expected output
      * E.g.: classification

#### **Predictive Data Mining Model**

Predicting data mining builds model using neural networks, tree, rule set, support vector etc. to class of new data set as an outcome for future. This method study previous historical data and predict and forecast what is going to happen to future dataset. In below we will discuss about various predictive data mining models [13].

* + - * 1. Classification

Data mining classification model allows to have a predicted class. Goal is to predict a class for each of the instance accurately with a minimal error rate. It is known as the best modeling system of data mining techniques [14].

**1.8.1.1.2 Regression**

Statistical regression find the dependencies of values in dataset and predict the numerical value for similar dataset. It requires a set of independent attribute and a dependent variable. Regression can be linear regression, support vector regression, polynomial regression etc. [15].

* + - * 1. Time Series Analysis

It’s a measurement of occurrences of set of sequence within a sequential scale of time. It works to predict data in time intervals like weekly, daily, yearly etc. It predicts sequence of any occurrences in future time period. Any natural disaster like earthquake, flood, and cyclone can be predicted using time series analysis [16].

#### **Descriptive Data Mining Model**

Descriptive data model is a predictive model consisting of clustering, summarization, association rule etc. It finds pattern in large data and further works in intelligence system decision making. These models uses past data to predict future data. Some descriptive data models are described below-

* + - * 1. Clustering

Clustering is one of the most important descriptive predicting and data mining model. It’s a system of predicting patterns for similar type of data and divide them into clusters. Where label in unknown. It studies past data to predict future data. E.g.: K-Means. [17]

* + - * 1. Summarization

This method gives overview about data by studying. It generalizes data and also map data into various subsets. Various types of summarizations are mean, median, standard deviation, tabulation etc. Its main task is to visualize data, analysis of data and automated report generation about the dataset. [18]

* + - * 1. Association

It finds relationships between attributes. It is one of the important attribute to analyze market. It establishes statistical relation between attributes of a model. It also uses if/else this types of condition to find out the relationship between various datasets. [19]

### **KDD Process**

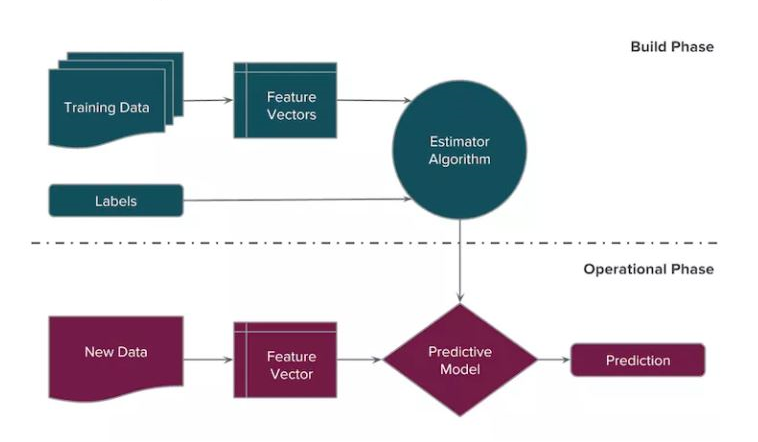
This process is a knowledge Discovery in Database or KDD process. KDD is mainly divided into mainly seven stages:

1. Identifying the goal
2. Selection of subset of dataset
3. Preprocessing and cleaning of data
4. Transformation
5. Choosing data mining algorithm
6. Evaluation
7. Interpretation of hidden patterns knowledge and using it

**Figure 1.2** shows the workflow of KDD process.

**Figure 1.2** The KDD steps

### **Machine Learning Approach**

We have used machine learning approach for the model creation. Machine learning approach has two phase one is the build phase and the other is operational phase [20]. **Figure 1.3** shows the general schema of machine learning approach.

Operational Phase

Build Phase

**Figure 1.3** Machine Learning approach

### **Feature Selection**

Feature selection is very important in the field of data mining. It is not very easy to run data mining algorithms in a very high dimensional data. It makes training and tests of a model very complex. To get rid of this various feature selection methods are used. These methods make data mining process very efficient [21]. There are mainly two types. One is wrapper and another is filter.

#### **Wrapper**

Wrapper uses the method of own classification model for selecting the important features. It provides better result. Because it uses the classifier it optimizes. But for bigger dimensional dataset it’s inefficient. It takes so much time to run. Because it has to compute and run each feature in classification algorithm [22].

#### **Filter**

Filter method precedes classification method. It’s simple, fast and complexity is least. First it can finish the feature selection process and then it can give input for the classification models. It’s not dependent on classification models. There are several efficient filter method. Some of them also gives the ranking of features. Correlation-based Feature Selection (CFS), Principal Component Analysis (PCA), Information gain, Markov blanket filter, Gain Ratio(GR) attribute evaluation, Chi-square Feature Evaluation, Euclidean distance, i-test, Fast Correlation-based Feature selection (FCBF) etc. are widely used filter method for feature selection and ranking features [22].

## Organization of the Thesis

The rest of the paper is designed as follows: Section 2 provides an overview of related works for predicting students’ performance from educational settings by using different data mining techniques; section 3 shows the statistical overview and description of survey questionnaire and data; section 4 describes methodology; Section 5 shows the implementation, results and discussions; Finally, section 6 outlines the conclusion and the future work.

**Chapter 2**

**Literature review**

## Overview

As we have learnt about the background, objectives and contributions of our work. It’s time to delve into literatures that relates to our work. We will be discussing about different data mining and machine learning approaches used to help students in decision making.

Over the years from various regions of the world researchers have focused on giving the right direction to the students before they select undergraduate program and to improve their performance even before they sit for the exams. The existing works can be categorized into two approaches based on:

1. Data Mining approaches for selecting study track.
2. Data Mining approaches for Analyzing and Predicting Students Performance

## Selecting Study Track Related Research Work

To determine which program will be suitable for Thai students after completing High school education, a classification model using Decision Tree was built by Waraporn [2]. GPA (Grade Point Average) in old school, their education type were taken into consideration to extract rules from the decision tree.

S. Venkata Krishna Kumar, et al. [23] proposed a recommendation system for predicting study track for student’s based on their profile. Past history of the Students who have succeed in their academic area are used. They used FCM for clustering student’s profiles and used C4.5 algorithm for classifying the students into the course and the college which they matches to them.

Cesar, et al. [24] proposed a recommendation system that used the experience of previous students with similar academic results to provide supports to Spanish students for selecting the suitable course to enroll. The main objective of the work was to recommend students who should take which courses based on historical data. They collected data of 3230 students from the Universidad de Lima and used C4.5 algorithm to find the pattern that will be effective for the recommendation system.

Qasem A. Al-Radaideh et al. [1] proposed a classification model for predicting suitable track for the students of Jordan. After completing their two-tier school majority of the students having insufficient knowledge about the undergraduate program often make mistake. To overcome this challenge they collected data from the students of class 9-10 and 11-12 and determined suitable track using Decision tree. The accuracy of their model was 87.9%.

## Students Performance Analyzing and Predicting Related Research Work

Cristobal Romer0o, et al. [25] conducted a study on students participating in on-line discussion forum and predicted students’ final performance in Spain. They collected forum interaction data such as number of messages post/read, ask and reply relationship between students. Afterwards compared them in between classification and classification via clustering approach.

Yeasmin, et al. [26] conducted an analysis of student performance using Data Mining in Bangladesh. Data was collected form the central database of Military Institute of Science Technology and the main objective was to relate student’s final result and student’s performance. Using J48 data mining algorithm predicted student’s final result and then generated a classification tree.

Zahyah Alharbi, et al [27] did a case study whether they can highlight performance problems early on and propose remedial actions using data mining techniques. They collected students data during admission and after completing their academic first year and eventually predicted good honors outcomes with reasonable accuracy by using classifying model with highlighting students that are predicted to low achievers with high probability module results.

Brijesh Kumar, et al. [8] suggested a classification model for Predicting Performance improvement on the Educational databases which contains invisible information for improvement of students’ performance. They collected 300(74 females, 226 males) students record from Dr. R. M. L. Awadh University, Fazibad India and used Bayes classification approach. They further conducted a research by taking students class test, assignments, attendance, lab work and seminars into consideration and analyzed students’ performance in the semester final examination. They used decision tree for predicting students’ performance [9].

Ali Daud, et al [5] conducted an study where using Educational Data mining approach they considered students family expenditure, personal information and predicted whether he will be able to complete his degree or not. They used WEKA tool to classify 100 students record with 23 features each from different universities of Pakistan.

Edin Osmanbegović and Mirza Suljić [4] predicted favorable outcome in a course of University of Tuzla, department of economics. Academic results of first year students of 2010-2011 were collected and special attention was given on socio-demographic factors. They used WEKA tool for the analysis and implemented in Java.

Goga et al. [28] proposed a tool by using .Net framework which takes students various information as input and predicts students’ grade. They first collected the student’s enrollment records from Babcock University, Nigeria and then built models using classification trees and a multilayer perceptron learning algorithms operating on WEKA. In the domain of this study random tree adopted as the best algorithm and served as a building block of this generic system.

Alfiani et al. [29] used K-means clustering algorithm to map students. This approach helped to reveal the hidden pattern and to classify students successfully. Data of 300 students Faculty of Industrial Engineering Department of Industrial Technology, Islamic University of Indonesia were collected. Specially, many demographic factors were taken into account such as sex, origin, CGPA etc. Using SPSS16 they found four clusters based on six variables.

Al-Radaideh, et al [30] conducted a study in Yarmouk University, Jordan to predict final score of C++ course. Different classification algorithms were implemented to build the best model where decision tree showed better result.

John M. Mativo and Shaobo Huang [31] used Support Vector Machine and Multiple Linear Regression algorithm on 48 students dataset enrolled in Engineering to predict their academic performance. SVM provided better accuracy to identify the students having low grades where MLR model more often over-estimated the performance and failed in number of cases.

Pauziah Mohd Arsad [32] conducted a study to predict student’s final result at the Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM), Malaysia. They showed Artificial Neural Network model served as a vital to predict students result. The performance was measured by calculating coefficient of Correlation R and Mean Square Error. It was revealed from the study that Students’ first and thirds semesters fundamental course results reflected their final result. Therefore, fundamental subjects should be fully understood because without clearing fundamental knowledge it is very difficult to face advance courses.

Shaobo Huang and Ning Fang [33] proposed a mathematical model for making early prediction of students’ final score and Engineering dynamic courses using the results of ﬁrst semester and validated using next three semester data. They collected data from 1900 engineering students of four semesters. Four different mathematical modeling approaches: radial basis function neural networks, multilayer perceptron neural networks, multivariate linear regression, multilayer perceptron neural networks, and support vector machines were applied and experimental results showed anyone can be used to predict the desired result.

Hidden patterns can be found using data mining algorithm on students database which size is increasing every day. Sequential pattern mining algorithm was proposed to find hidden patterns from student’s data by Priyanka Anandrao Patil and R. V. Mane [34]. By using FP tree algorithm they constructed a tree using these patterns to predict students’ performance so that students having higher failure risk can get necessary help.

Pratiyush Guleria et al. [35] calculated the entropy of the students attributes after preprocessing students dataset and to split it further, attributes having highest information gain is taken as the root node. This study revealed that to analyze and predict the result of the class, two most important factors are sessional marks and class performance. Generated result using data mining will assist faculty member to take special care of students getting poor marks.

Kamal Bunkar et al. [36] collected data from Vikram University, Ujjain of course B.A. first year student and applied data mining classification techniques to improve the quality of existing educational system. After building the classifier they developed an automated system which uses DT and generates rules to predict students’ final grades.

Hijazi, et al. [37] collected data from 300 students of Panjab University of Pakistan and stated in their hypothesis that student’s class attendance, hour spent in study, parent’s age and education are strongly correlated with their academic performance.

Pandey, et al. [38] collected data from 600 students from different colleges of India and by using Bayes Classification on academic results and language predicted whether the newcomer a performer or Underperformer.

* 1. **Summary**

Researchers from different regions found that study on forecasting student’s final performance and selecting appropriate program for under graduation program is very important. They considered students’ academic result, parents data, hours spent in study, activity in online discussion forum, marital status & student’s class attendance for forecasting their final performance and building classification model which can help them take decision which track they should follow for the under graduation program. Different from the literature, we considered only Computer Science program as our purpose of study for its immeasurable popularity and the kind of challenges it encounters. Especially, for someone who is not ready to take those challenges is sure to suffocate in CS. To get rid of this we discovered two different clusters of students profile from our dataset and father used different classification algorithm to classify these  profiles into OK or Not OK profile based on 22 features including their academic results, experience with ICT course, personal interests, personal experiences of current students’ with similar academic achievements and problem solving skills.

**Chapter 3**

**Methodology**

## Overview

In the previous chapter we have discussed about some basic topic and definitions related to this thesis work. Now we will go through the whole methodology here. From section 4.2 onwards we will be discussing each step of the proposed model in detail.

## Proposed method

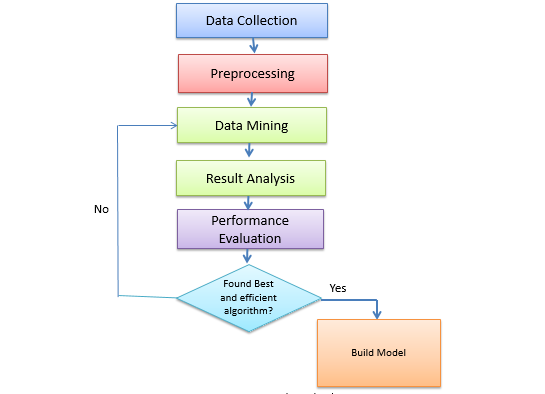
As a machine learning approach our proposed method has mainly two phase. One is the “Build Phase” and the other is “Operational Phase”

### **Build Phase**

Following KDD process build phase of our proposed method has seven stages.

1. Data collection
2. Pre-processing
3. Data cleaning
4. Transformation
5. Integration
6. Standardization
7. Feature selection
8. Data mining and model generation using various algorithms
9. Performance measurement of algorithms
10. Finally we will get a model to use.

**Figure 3.1** shows the overview of build phase of our proposed model.



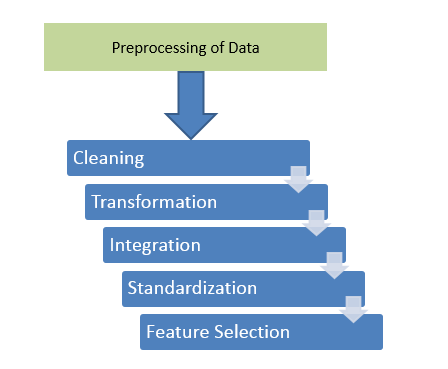
**Figure 3.1** Build Phase

#### **3.2.1.1 Data collection**

Data collection is the most hardest and important process of the process. Data is collected through a questionnaire using various tools i.e.: IBM SPSS Statistics. After this process all the available data is fetched in a data set and saved in a recognizable format. Training data is collected from existing students of Computer Science.

#### **3.2.1.2 Preprocessing**

Data preprocessing is the process of creating data ready for the machine. Often collected data is not understandable, inconsistent, lacking in important criteria or can contain various errors. Preprocessing makes data understandable by various process and solve those issues. Data preprocessing is an important step of the build phase. **Figure 3.2** shows the steps of data preprocessing.



**Figure 3.2 Data Preprocessing**

Some steps of preprocessing are described below-

* + - * 1. Data cleaning

To improve the quality of data, data clearing is important. Sometimes redundant data take places in dataset. Or there can be inconsistence data. [39]

* + - * 1. Transformation

Data is aggregated by various method. Also normalized and generalized to use the data efficiently.

* + - * 1. Integration

There can be conflict between data in different place of data set. These problem should be solved. It’s known as data integration process.

* + - * 1. Standardization

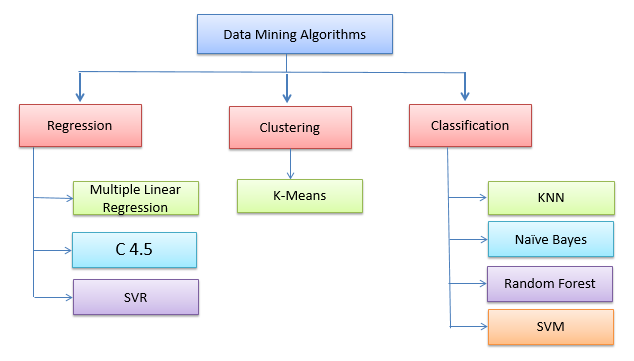
Standardization is the process to bring dataset into a common format which is needed for cross checking, research and large-scale analytics.

* + - * 1. Feature selection

Not all of the attributes of dataset is important. We have to find out the key feature on which our prediction and data mining model depends most. For this various method is used. It’s a process from which we get a subset of original dataset. Feature selection algorithm uses training set of data to evaluate the features and fit those onto model.

**3.2.1.3 Data Mining and Model Generation Using Various Algorithms**

After preprocessing pure and error free data is ready for mining and further processing to create model and predict. Our proposed algorithm is for regression is linear regression support vector regression. Then K-Means for clustering and KNN, Naive Bayes, Random forest, SVM for classification. **Figure 3.3** shows proposed algorithms for data mining process.



**Figure 3.3** Proposed Data mining algorithms

**3.2.1.4 Performance Evaluation of Algorithms**

Two types of performance measurement used. One type is for measuring performance of regression and another type is for measuring performance of classification process.

3.2.1.4.1 Performance Evaluation for Regression

Two standard evaluation metrics (MAE, RMSE) are used to evaluate the performance of our prediction. These performance evaluation parameters are defined as:

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

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

**3.2.1.4.2 Performance Measurement of Classification**

Performance of classification models usually evaluated by confusion matrix. Confusion matrix contains information about original and predicted classification done by a classifier. **Table 3.2** shows the context of confusion matrix-

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Actual | |
| Positive | Negative |
| Predicted | Positive | TP | FP |
| Negative | FN | TN |

**Table 3.1** Confusion Matrix

Here,

**TP** = True Positive = Predicted as positive and originally member of positive class

**FP** = False Positive = Predicted as positive but originally member of negative class

**FN** = False Negative = Predicted as negative but originally member of positive class

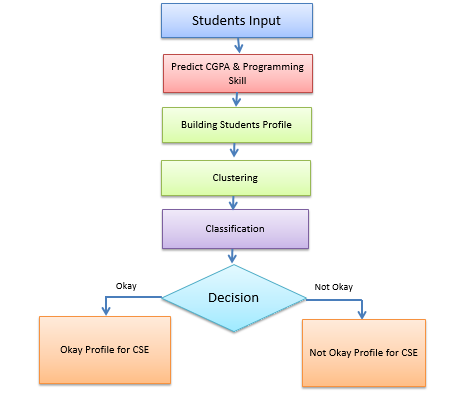
**TN** = True Negative = Predicted as negative and originally member of negative class

Several standard terms for evaluating by confusion matrix for two class-

|  |  |
| --- | --- |
| **Accuracy** | ACC = (TP + TN) / (P + N) |
| **Sensitivity or Recall or True Positive rate** | TPR = TP / (TP + FN) |
| **Specificity or True Negative rate** | SPC = TN / (FP + TN) |
| **False Positive Rate** | FPR = FP / (FP + TN) |
| **False Negative Rate** | FNR = FN / (FN + TP) |
| **Precision** | PPV = TP / (TP + FP) |
| **F1 Score or F-Measure** | F1 = 2TP / (2TP + FP + FN) |

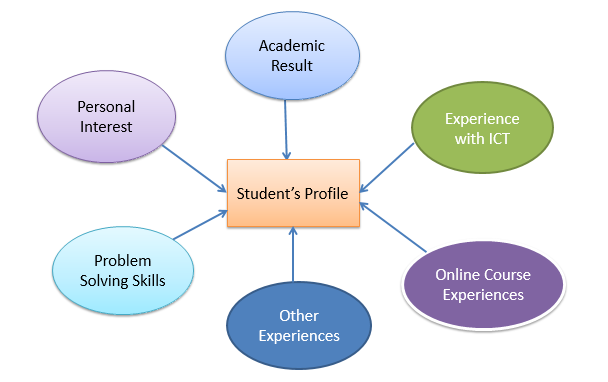
### **Operational Phase**

In the operational phase our model first predict the data which are not available in the new data i.e.: students CGPA, Programming skill. Then it builds student profile to use the model which should be created on the build phase. **Figure 3.4 shows the** overview of operational phase.



**Figure 3.4** Operational Phase

In operational phase building students profile is the most important task. For building students profile we have selected several criteria by using our domain knowledge i.e.: Personal information, experience with ICT etc. As we said earlier we are taking training data in the build phase from current student. But in this phase new data is collected from students whoever want to get themselves admitted into CS. As some of attributes is not available for new data we planned to predict and then build students profile. Figure **3.5** shows the overview of building students profile.



**Figure 3.5** Overview the Build-up process of Students’ Profile

* 1. **Summary**

As our proposed model is a machine learning approach we must follow the two phase of machine learning approach. And the first phase which is build phase follows KDD process. Our proposed method stores the final output and gain knowledge for discovering next level of hidden patterns. For this purpose our proposed method is very well planned by following an efficient planning process by documentation.

**Chapter 4**

**Survey Data Analysis**

**4.1 Overview**

In chapter 4 we have seen description of various tools and techniques. We tried to give a short but efficient description to describe them. Here, we will describe how we have collected our data and how our attributes related to each other with the help of various statistical measurement.

We have collected survey data from 3rd and 4th year students of Computer Science Department in different universities of Bangladesh. We have created the questionnaire having 21 features of different categories. Afterwards gathered and fetched these data using IBM SPSS Statistical tool. We assigned numeric values to these questions responses for data analysis and research purpose. This questionnaire for students are sectioned into six different sections including Personal Information, Academic Result, Experiences with ICT, Personal interest, Personal Experience and Problem solving skills and others. **Table-4.1** Shows the options with correspondent values and short term of the survey questionnaire.

**Table-4.1** Options with correspondence values and short terms

|  |  |  |
| --- | --- | --- |
| Options | Short Term | Values |
| Strongly Agree/Very Good/CGPA 3.5 and above /5:00P/Many /Very Interesting /9 to 10 ratings/Social media and Entertainment | SA/VG/C3.8AA/5P/MT/VI/9-10R/FLP/SMAE | 5 |
| Agree/Good/CGPA 3.40 and above/4.50P and above/More than 5/Interesting/7 to 8 ratings/Study and Social Media | A/G/C3.4AA/4.5PAA/MT5/I/7-8R | 4 |
| Neutral/Satisfactory/CGPA 3.00 and above/4.00P and above/More than Once/5-6 ratings/, Reading Blogs only | N/S/C3.0AA/4.0PAA/MTO/5-6R | 3 |
| Disagree/Less than satisfactory/CGPA 2.50 and above/3.50P and above/Once/3-4 ratings/Online Courses only | D/LTS/C2.5 AA/3.5PAA/O/3-4R | 2 |
| Strongly disagree/Poor/CGPA Less than 2.25/Less than 3.50P/Never/1-2 ratings/Online Course and Reading Blogs | SD/P/CLT2.5/LT3.5P/Ne/1-2R | 1 |

**4.2 Questionnaire for Students**

**Section A- Personal Information**

A1- What is your name?

A2- Gender?

A3 -What is the Location of your University?

**Section B- Academic result**

B1- What is your SSC GPA out of 5.00?

B2- What is your HSC GPA?

B3- What is your current CGPA?

**Section C- Experiences with ICT Course**

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

C2- How was your academic result in ICT?

**Section D- Personal Interest**

D1-Why do you browse internet mostly?

D2- Participated in number of math or science Olympiad?

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

D4- Rate your interest in Competitive Programming?

**Section E- Personal Experience**

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

E2- “Having Patience help in Computer Science”:

E3- “Knowing PC configuration helps in reading Computer Science”:

E4- ‘Computer gamer have better chance in CS’:

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

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

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

**Section F- Problem solving skills and others**

F1- Rate your programming skill:

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

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

F4- Rate your patience out of 10:

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

**4.2.1 Personal Information**

In **Table-4.2**, shows the mean and standard deviation values of personal information of

Students including name, gender, location of University.

**Table-4.2:** Frequencies of Personal information

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Name | Gender | Location of University |
| N | Valid | 394 | 394 | 394 |
| Missing | 0 | 0 | 0 |
| Mean | | - | 1.334 | 2.023 |
| Std. Deviation | | - | 0.473 | 0.577 |

**Frequency Table**

In this section, Frequency tables will show the frequency, percent valid percent and cumulative percent of the student’s Personal information.

The number of male and female shown in **Table 4.3.**

**Table 4.3** Gender of Students

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Male  Female  Total | 261  133  394 | 66.24  33.76  100.0 | 66.2  33.7  100.0 | 66.2  100.0 |

261 of 395 students are male and 133 is female. That is 66.24% students are male and 33.76% students are female.

The location of university shown in **Table 4.4**

**Table 4.4** location of university

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Dhaka  Chittagong  Other divisional City  District City  Rural area  Total | 40  322  19  9  4  394 | 10.15  81.73  4.82  2.28  1.02  100 | 10.2  81.8  4.8  2.3  1.0  100 | 10.2  92.0  96.7  99.0  100.0 |

Here most of the student’s university location is Chittagong. 10.15% is from Dhaka. 4.82% is from other divisional city, 2.28% from District city and 1% from rural area.

**4.2.2 Academic Results**

In this section, Frequency tables will show the frequency, percent valid percent and cumulative percent of the student’s Academic result

In **Table-4.5**, shows the mean and standard deviation values of Academic Result of

Students including SSC GPA, HSC GPA and Current CGPA.

**Table 4.5 Academic Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | B1 | B2 | B3 |
| N | Valid | 394 | 394 | 394 |
| Missing | 0 | 0 | 0 |
| Mean | | 4.259 | 4.254 | 3.388 |
| Std. Deviation | | 0.917 | 0.791 | 0.395 |

Now we analyze B2 question of questionnaire is shown at **Table 4.6.**

**Table 4.6** Frequency table of B2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative  Percent |
| Valid | Less than 3.50  3.50 and above  4.00 and above  4.50 and above  5.00  Total | 116  119  110  43  6  394 | 29.44  30.20  27.92  10.91  1.5  100.0 | 29.4  30.2  27.9  10.9  1.5  100.0 | 29.4  59.6  87.6  98.5  91.0 |

We can see only 1.5% students have 5 GPA in HSC. Most of the students have GPA less than 4.00 in HSC. Approximately 60% students have less than 4.00 GPA in HSC.

**4.2.3 Experiences with ICT**

In this section, Frequency tables will show the frequency, percent valid percent and cumulative percent of the student’s Experiences with ICT

In **Table-4.7**, shows the mean and standard deviation values of Experiences with ICT

**Table 4.7** Frequency of Experiences with ICT

|  |  |  |  |
| --- | --- | --- | --- |
|  | | C1 | C2 |
| N | Valid | 394 | 394 |
| Missing | 0 | 0 |
| Mean | | 4.25 | 4.22 |
| Std. Deviation | | .792 | .806 |

Now we analyze C2 question of questionnaire is shown at **Table 4.8**

**Table 4.8** Frequency table of C2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Poor | 3 | .8 | .8 | .8 |
| Less than satisfactory | 10 | 2.5 | 2.5 | 3.3 |
| Satisfactory | 46 | 11.7 | 11.7 | 15.0 |
| Good | 174 | 44.2 | 44.2 | 59.1 |
| Very Good | 161 | 40.9 | 40.9 | 100.0 |
| Total | 394 | 100.0 | 100.0 |  |

Most of the students having very good and good results in ICT of intermediate. Only few students having poor result which is less than 1%.

**4.2.4 Personal Interest**

In this section, Frequency tables will show the frequency, percent valid percent and cumulative percent of the student’s Personal Interest

In **Table-4.9**, shows the mean and standard deviation values of Experiences with ICT

**Table 4.9** Frequency of Personal Interest

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | D1 | D2 | D3 | D4 |
| N | Valid | 394 | 394 | 394 | 394 |
| Missing | 0 | 0 | 0 | 0 |
| Mean | | 2.26 | 2.26 | 2.91 | 3.50 |
| Std. Deviation | | 1.191 | 1.191 | 1.415 | 1.331 |

**Frequency Table**

In this section, Frequency tables will show the frequency, percent valid percent and cumulative percent of the student’s Personal Interest

Now we analyze D2 question of questionnaire is shown at **Table 4.10**

**Table 4.10** Frequency table of D2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Never | 80 | 20.3 | 20.3 | 20.3 |
| Once | 97 | 24.6 | 24.6 | 44.9 |
| More than once | 72 | 18.3 | 18.3 | 63.2 |
| More Than 5 | 69 | 17.5 | 17.5 | 80.7 |
| Many | 76 | 19.3 | 19.3 | 100.0 |
| Total | 394 | 100.0 | 100.0 |  |

Here we are seeing average results in all category. Question was how many science or math Olympiad did they participated. Only 20.3% students never participated and 24.6% student participated once.

**4.2.5 Personal Experiences**

In this section, Frequency tables will show the frequency, percent valid percent and cumulative percent of the student’s Personal Experiences

In **Table-4.11**, shows the mean and standard deviation values of Personal Experiences Student’s questionnaire E1, E2, E3, E4, E5, E6, E7.

**Table 4.11** Frequency of Personal Experience

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | E1 | E2 | E3 | E4 | E5 | E6 | E7 |
| N | Valid | 394 | 394 | 394 | 394 | 394 | 394 | 394 |
| Missing | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Mean | |  | 4.26 | 4.26 | 4.05 | 3.48 | 4.05 | 3.93 |
| Std. Deviation | |  | .721 | .734 | .850 | .981 | .921 | .813 |

**Frequency Table**

In this section, Frequency tables will show the frequency, percent valid percent and cumulative percent of the student’s Personal Interest Now we analyze E7 question of questionnaire is shown at **Table 4.12**

**Table 4.12** Frequency table of E7

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Very difficult | 21 | 5.3 | 5.3 | 5.3 |
| difficult | 26 | 6.6 | 6.6 | 11.9 |
| Neutral | 77 | 19.5 | 19.5 | 31.5 |
| interesting | 179 | 45.4 | 45.4 | 76.9 |
| Very interesting | 91 | 23.1 | 23.1 | 100.0 |
|  | Total | 394 | 100.0 | 100.0 |  |

Here the question was how many of students have previous knowledge of CS before admitted into CS and how did they find the subject. Most of them (45.4%) found this interesting and 23.1% students found this subject very interesting. 19.5% have no comments. 6.6% and 5.3% student found CS difficult and very difficult.

* + 1. **Problem Solving Skills And Others**

In this section, Frequency tables will show the frequency, percent valid percent and cumulative percent of the student’s Problem solving skills and others

In **Table-4.13**, shows the mean and standard deviation values of Personal Experiences Student’s questionnaire F1, F2, F3, F4, F5, F6, F7.

**Table-4.13** Frequency of Problem solving skills and others

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | F1 | F2 | F3 | F4 | F5 |
| N | Valid | 394 | 394 | 394 | 394 | 394 |
| Missing | 0 | 0 | 0 | 0 | 0 |
| Mean | |  | 2.3528 | 2.1269 | 2.7843 | 2.7614 |
| Std. Deviation | |  | .75490 | .87028 | 1.21343 | 1.16302 |

**Frequency Table**

In this section, Frequency tables will show the frequency, percent valid percent and cumulative percent of the student’s Personal Interest

Now we analyze F1 question of questionnaire is shown at **Table 4.14**

**Table 4.14** Frequency table of F1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | 1 to 2 | 21 | 5.3 | 5.3 | 5.3 |
| 3 to 4 | 69 | 17.5 | 17.5 | 22.8 |
| 5 to 6 | 89 | 22.6 | 22.6 | 45.4 |
| 7 to 8 | 90 | 22.8 | 22.8 | 68.3 |
| 9 to 10 | 125 | 31.7 | 31.7 | 100.0 |
| Total | 394 | 100.0 | 100.0 |  |

We can see 31.7% students think they are excellent programmer (rating 9-10). 22.8% rated themselves 7to8, 22.6% rated themselves 5to6 and only 5.3% rated themselves as low quality programmer that is 1to 2.

## Selecting Features

In our model we have selected important attributes by Gain Ratio (GR) setting the threshold at 2.0 algorithm and searching by ranker algorithm[40].

**Gain Ratio Calculation:**

* Amount of information gained by knowing the value of the attribute
* Gain Ratio for a attribute P =

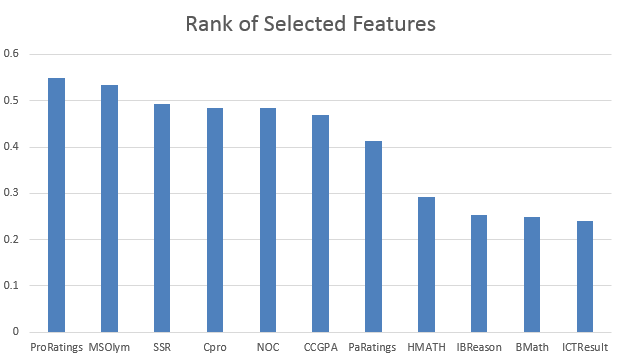
Where ,

* Gain=(Entropy of distribution before the split)–(entropy of distribution after it)
* Entropy(P1,P2,..,Pn)=−P1log(P1)−P2log(P2)−⋯−Pnlog(Pn) = -
* Split Info(A) = , where A = Attribute

Selected features are shown in **Table 4.15.** We can see we have shorten our attributes by selection 22 to 11 which is very efficient for our model.

**Table 4.15:** List of selected features

|  |  |
| --- | --- |
| **Selected Features** | **Gain Ratio** |
| ProRatings | 0.5486 |
| MSOlym | 0.5344 |
| SSR | 0.4933 |
| Cpro | 0.4839 |
| NOC | 0.4833 |
| CCGPA | 0.4701 |
| PaRatings | 0.4125 |
| HMATH | 0.2911 |
| IBReason | 0.2528 |
| BMath | 0.248 |
| ICTResult | 0.2396 |

**Figure 4.4** shows the ranking of selected features-

**Figure 4.1** Rank of selected features

## Summary

Here we have discussed our questionnaire, shorted variable list for the transformation and preprocessing of data. We have found that attributes in the section Personal Experience is mostly same for all. And there is no class difference between male and female for various attributes. From the frequency table we came to know that for some variable values are almost same for all. Also figured out the important attributes by feature selection process with the help of gain ratio.

**Chapter 5**

**Implementation and Results**

## 5.1 Overview

In the previous chapter we have showed various statistical relation between attributes in our dataset. In this chapter will discuss and show the implementation results of various algorithms. Considering CGPA and Programming skill as most important attributes for the students to determine whether they will do well or not in Computer Science Program we have used MLR, SVR and C4.5 to predict CGPA and programming skill of a student. After that we used Gain-Ratio and Ranker algorithm to find out eleven most important features from student’s datasets. We further implemented different classification algorithm by using these features. Afterwards we measured accuracy and F1-score for different classifiers.

## 5.2 Tools

To implement and verify our proposed work Spyder which is available in Anaconda and Weka is employed. These tools are widely used as a predicting platform of intelligence and decision making. These tools contains a large number of data mining and machine learning advanced level algorithm which helps us to predict and make good decisions. We also used IBM SPSS Statics to take data and process statistical analysis of data taken from students.

### **Anaconda**

Anaconda has gained its reputation as a must use data mining tool for the people who are working as a data scientists. It’s an open source application with a numerous options and techniques for data mining and machine learning (i.e. processing of data, scientific analysis and predictive analysis of data). It mainly works with python and R. Anyone can build package in anaconda using its conda build command. We can install any open source packages and programs using conda install or pip install**.** Anaconda version 2 mainly deals with the python version 2.7 and Anaconda version 3 uses python 3.6. But we can create any environment using any version of python or use any environment which was created by anaconda in any version of python programming. Anaconda is available in windows, mac and ubuntu [41].

#### **5.2.1.1 Anaconda Navigator**

To compute, various scientific packages rely on particular versions of various packages. Data scientists use various versions of various packages, and use numerous environments to isolate these individual versions.

It has a command manager which is actually a combination of package manager and environment manager. It works to help who works as data scientists to do their works without any problem in versions of various package.

The following applications are available by default in Navigator:

* JupyterLab
* Jupyter Notebook
* QTConsole
* Spyder
* VSCode
* Glueviz
* Orange 3 App
* Rodeo
* RStudio
  + - 1. Spyder

Normal mathematical operations can be done with python plane raw code. But to implement complex and bigger methods for bigger purpose we need some add-on packages. For this now a days and open source IDE like Spyder is a must use thing. Spyder is purposefully pe to look like Matlab. Spyder integrates with a huge number of efficient packages for scientific python tools. It inclues NumPy, Matplotlib , IPython, SciPy,pandas, Cython and SymPy and also all open source packages.In 2009 it was first developed by Pierre Raybaut. From then a team of developer and community is maintaining and improving continuously Spyder.[42]

* + 1. **Weka**

Weka is a widely used machine learning tool developed in java for desktop. It gives a lot of opportunity to use various types of machine learning algorithm [43].

Advantages of Weka include:

* It’s a free software available to download
* It’s implemented in java and it has portability
* Suitable in any platform
* A numerous collection of data mining tools and techniques
* Easy to use

It supports different data mining tasks including Data preprocessing, various classiﬁcation, automatic and manual clustering, feature selection, regression analysis and visualization, and

Normally WEKA works with numeric and nominal data in a single file. But others data type is also supported. We can connect java database and sql database by Weka. For multi relational data, weka is not useful.

* + 1. **IBM SPSS Statistics**

SPSS Statistics is mainly used for batched and various statistical analysis. It’s a widely used tool for measurement of statistical overview of data. IBM gained it back in 2009. It is now used by government, social scientists, marketing analyst, educational researcher and data miner. Descriptive data mining model like summarization can be used to produce automated report on dataset [44].

It contains

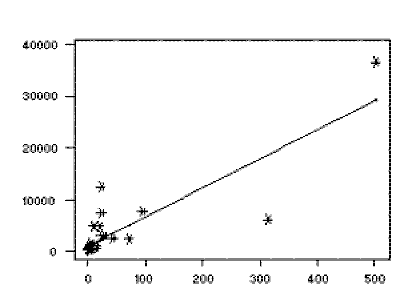
* Descriptive statically overview
* Predictive tools
* Cross validation
* Identifying groups i.e. k-means clustering
* Regression analysis and other numerical outcome
  + 1. **Prediction of CGPA and Programming Skill**

For predicting CGPA and programming skill we have used several algorithm. Here is a short brief-

**Multiple Linear Regression**

Linear regression uses a set of independent variable to predict a dependent variable to fit a model by the equation y = c + b\*x,

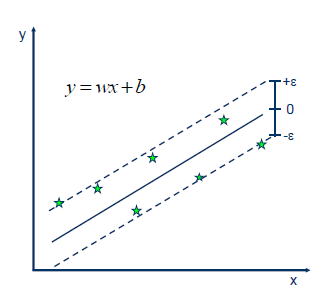
Where, y = score of dependent variable, c = constant, b = co-efficient, and x = independent variable’s score. We often use multiple linear regression by y=Xβ [45].**Figure 5.1** shows the concept of linear regression.



**Figure 5.1** Linear regression

**Support Vector Regression**

The regression which uses support vector machine [46, 47] for the regression process is known as SVR. It maintains the same principle of SVM. The process is more complex from SVM but idea is same i.e.: using the hyper plane to reduce the error [48]. Two type of SVR is mainly available. E.g.: linear and nonlinear. The figure **5.2** below shows the idea.



**Figure 5.2** Idea of SVR

**C 4.5**

It’s a decision tree [49] based algorithm for classification both numeric and nominal classes. It was written by J. R. Quinlan. [50]

**5.3.1.1 Performance Measurement of Regression**

Error evaluation of regression for predicting CGPA of students shown in **Table 5.1**

**Table 5.1** Error evaluation for Regression algorithms (CGPA prediction)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Evaluation Criteria** | **Multiple Linear Regression** | **C4.5** | **SVR (Linear kernel)** | **SVR (Poly kernel)** |
| Mean Absolute Error | 0.1180 | 0.1555 | 0.1123 | 0.1409 |
| Root Mean Square Error(RMSE) | 0.1712 | 0.2517 | 0.1604 | 0.2117 |
| Cross-Predicted Accuracy: | 0.8187 | 0.60840088 | 0.8411 | 0.7229 |

Cross validation for regression for predicting CGPA of students shown in **Table 5.2**

**Table 5.2** Cross Validation for Regression Algorithms (CGPA Prediction)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Evaluation Criteria** | **Multiple Linear Regression** | **C4.5** | **SVR (Linear kernel)** | **SVR (Poly kernel)** |
| Score | 0.8560 | 0.8367 | 0.8615 | 0.8488 |
| Cross-predicted Accuracy {k(10)-Fold} | 0.7734 | 0.6019 | 0.7806 | 0.6009 |

Error evaluation of regression for predicting Programming Skill of students shown in **Table 5.3**

**Table 5.3** Error Evaluation for Regression Algorithms (Programming Skill Prediction)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Evaluation Criteria** | **Multiple Linear Regression** | **C4.5** | **SVR (Linear kernel)** | **SVR (Poly kernel)** |
| Mean Absolute Error | 0.5226 | 0.4242 | 0.5036 | 0.4596 |
| Root Mean Square Error(RMSE) | 0.6899 | 0.7881 | 0.6880 | 0.6584 |
| Cross-predicted accuracy | 0.718556981243 | 0.83956 | 0.63407 | 0.6648 |

10-fold Cross validation for regression for predicting Programming Skill of students shown in **Table 5.4**

**Table 5.4** Cross Validation for regression algorithms (Programming Skill Prediction)

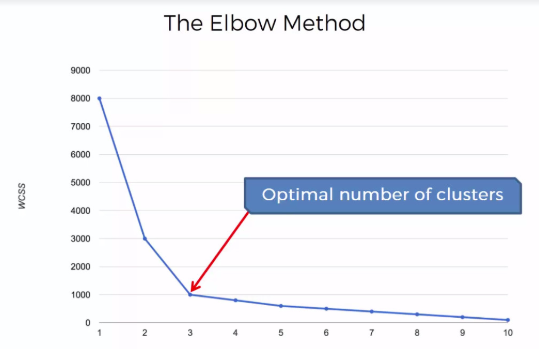
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Evaluation Criteria** | **Multiple Linear Regression** | **C4.5** | **SVR (Linear kernel)** | **SVR (Poly kernel)** |
| Score | 0.71856 | 0.83956 | 0.6950 | 8835 |
| Cross-validated Accuracy {k(10)-Fold} | 0.5476 | 0.5121 | 0.5293 | 0.4969 |

**5.4 Clustering**

For clustering student’s profile we have used K-Means clustering. Here is a short brief about K-Means algorithm.

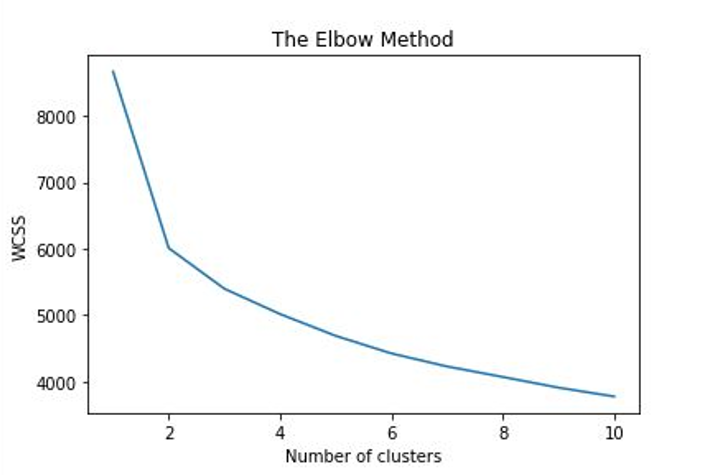
**5.4.1 K-Means Clustering algorithm**

K-Means is an iterative and simplest process between clustering [51, 52]. Target of K-means is to create a separate partition between instances. For this K-Means uses within-cluster sum of squares (WCSS) with elbow method. The **Figure 4.7** below shows the elbow method [51].



**Figure 5.3 Elbow** method

To find the appropriate number of clusters in our dataset we have used Elbow method with WCSS produced two clusters and we have labeled one as "OK" and another one as "Not OK". For this labeling we have used our domain knowledge. The result from elbow shown in figure



**Figure 5.4** Result from Elbow Method

**5.4.1.1 Labeling Clusters**

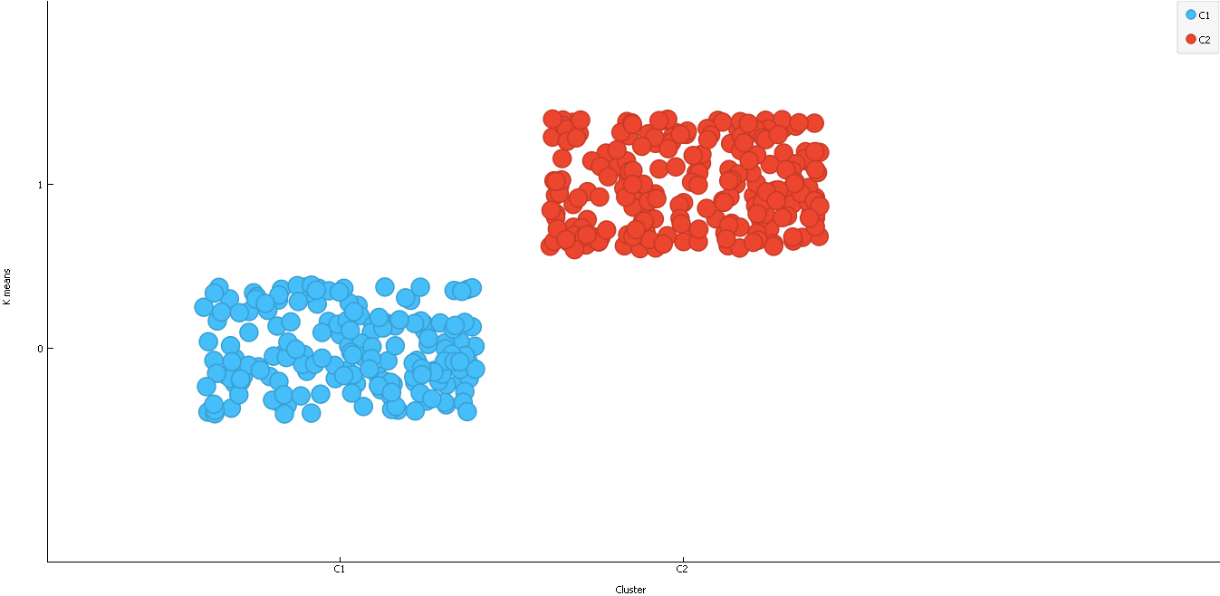
When creating model using these clusters we have found clear statistical difference between instances of two clusters. **Table 6.5** shows difference of selected attributes between two newly labeled clusters by statistical derivations.

**Table 5.5** Statistical overview of two newly labeled clusters

|  |  |  |  |
| --- | --- | --- | --- |
|  | Cluster | | Statistic Derivation |
| Attribute | 1 | 2 |
| ProRatings | 4.1415 | 3.335 | Mean |
| 0.7325 | 0.8036 | Std. dev |
| MSOlym | 3.8821 | 1.7747 | Mean |
| 1.086 | 0.7478 | Std. dev |
| SSR | 4.3774 | 2.7637 | Mean |
| 0.598 | 0.9342 | Std. dev |
| Cpro | 4.1415 | 2.3901 | Mean |
| 0.8002 | 0.9356 | Std. dev |
| NOC | 4.3255 | 2.5385 | Mean |
| 0.6887 | 1.2475 | Std. dev |
| CCGPA | 3.6442 | 3.0897 | Mean |
| 0.2572 | 0.3101 | Std. dev |
| PaRatings | 4.2877 | 2.8516 | Mean |
| 0.671 | 0.9972 | Std. dev |
| HMATH | 4.5047 | 3.5385 | Mean |
| 0.6332 | 0.7384 | Std. dev |
| IBReason | 2.783 | 1.6429 | Mean |
| 1.2886 | 0.6538 | Std. dev |
| BMath | 4.7264 | 3.9176 | Mean |
| 0.5414 | 0.7329 | Std. dev |
| ICTResult | 4.6038 | 3.7692 | Mean |
| 0.5776 | 0.7994 | Std. dev |

By using our domain knowledge we have labeled our clusters as OK and Not OK. **Table 5.6** shows the percentages.

**Figure 5.5** shows the scatter of two newly found clusters.



**Figure 5.5 newly found Clusters**

**Table 5.6** Labeling of Clusters

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster | Label | Student | Percentage (By K-Means) |
| 1 | OK | 212 | 54% |
| 2 | Not OK | 182 | 46% |

**5.5 Classification**

We have used several classification algorithm and then evaluated their performance to find best algorithm. Here a short brief about used classification algorithms.

* + 1. **KNN**

K-nearest neighbor is the simplest supervised learning algorithm [53]. It’s a not parametric and lazy learning algorithm [54]. Normally datasets are separated in several classes and the work of KNN is to learn from these training dataset and predict future data. This algorithm simply finds the closed point in dataset to classify a point and predict. It saves the full training sets.

* + 1. **Naive Bayes**

Naive Bayes lies in probabilistic class algorithm [55, 56]. It works with simple feature. Suppose there is 10 independent variable in a model. Naive Bayes takes in account only one variable at a time. It’s not only an algorithm but it refers to a full set of algorithm. For some model Naive Bayes is very efficient.

* + 1. **Random Fores**t

Random forest [57] is an assemble method. It works with multitude of decision tree [50]. It uses the decision tree algorithm and the tree bagging but the difference is they uses overlearning. For this accuracy of random forest is sometimes very accurate. It corrects the over fitting for training set for decision tree. The **Figure 5.6** below shows sample figure output of random forest tree.

Instance

Class A

Class B

Class A

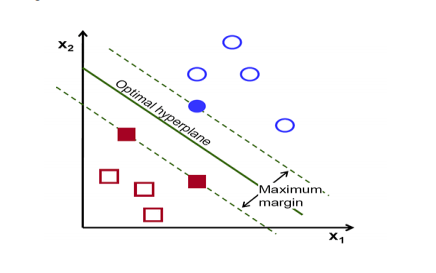
Majority Voting

Final Class

**Figure 5.6** Random forest

**5.5.4 Support Vector Machine**

Support vector machine is the supervised algorithm for using as a classifier or repressor for pattern, nested problems or mining of texts [47]. It uses a hyper plane to partition two different classes where support vectors are those which instances are used as the margin [48]. There is a gap between two classes which is known as marginal gap. **Figure 5.7** below shows the support vector machine.



**Figure 5.7** Support Vector Machine

**5.5.5 Performance Evaluation of Classification Algorithms**

**Table 5.7** shows confusion matrix of different classifier algorithms.

**Table 5.7** Comparison of Confusion Matrix for different Classifier

|  |  |  |  |
| --- | --- | --- | --- |
| **Classifier** | OK | Not OK |  |
| KNN | 206 | 6 | OK |
| 4 | 178 | Not OK |
| Naive Bayes | 205 | 7 | OK |
|  | 2 | 180 | Not OK |
| Random forest tree | 208 | 4 | OK |
| 2 | 180 | Not OK |
| SVM (linear) | 211 | 1 | OK |
| 4 | 178 | Not OK |
| SVM (RBF) | 210 | 2 | OK |
| 3 | 179 | Not OK |

Performance difference between SMV kernels is shown in **table 5.8**-

**Table 5.8** Comparison of different kernel of SVM Models confusion matrix

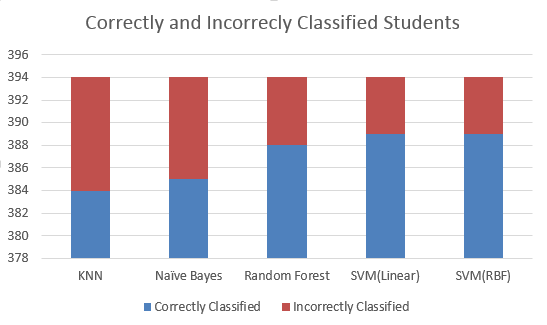
|  |  |  |  |
| --- | --- | --- | --- |
| **Kernel of SVM** | OK | Not OK |  |
| Linear | 211 | 1 | OK |
| 4 | 178 | Not OK |
| RBF | 210 | 2 | OK |
|  | 3 | 179 | Not OK |

**Table 5.9** shows number of students correctly and incorrectly classifiedstudents.

**Table 5.9** Correctly and incorrectly classified students

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| EVALUATION CRIETERIA | Classifiers | | | | |
| **KNN** | **Naive Bayes** | **Random Forest** | **SVM** | |
| **Linear** | **RBF** |
| Correctly Classified | 384 | 385 | 388 | 389 | 389 |
| Incorrectly Classified | 10 | 9 | 6 | 5 | 5 |

Graphical representation of students correctly and incorrectly classified is shown in **Figure 5.8**-



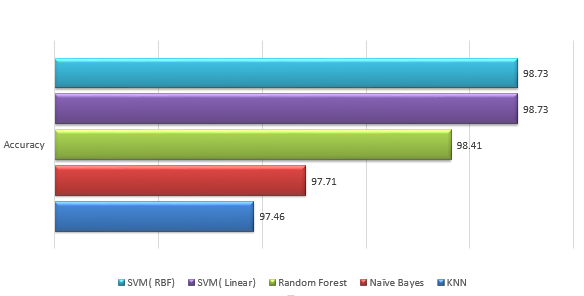
**Figure 5.8** Correctly and incorrectly classified student’s number

**Accuracy measurement of various algorithm shown in table 5.10.**

**Table 5.10** Accuracy rate of various classifier algorithm

|  |  |
| --- | --- |
| **Algorithm** | **Accuracy** |
| KNN | 97.46 |
| Naïve Bayes | 97.71 |
| Random Forest | 98.41 |
| SVM( Linear) | 98.73 |
| SVM( RBF) | 98.73 |

**Figure 5.9** shows percentage of accuracy of classifier algorithms



**Fig 5.9** Percentage of accurately classified students

**Table 5.11** showsPerformance of different classification model by various factor of confusion matrix

**Table 5.11** Performance of different classification model

|  |  |  |  |
| --- | --- | --- | --- |
| **Classifier** | **Precision** | **Recall** | **F-Measure** |
| **KNN** | 0.9717 | 0.9810 | 0.9763 |
| **Naive Bayes** | 0.9670 | 0.9903 | 0.9785 |
| **Random forest tree** | 0.9811 | 0.9905 | 0.9858 |
| **SVM (linear)** | 0.9953 | 0.9814 | 0.9883 |

**5.6 SUMMARY**

We have used several algorithm to build our proposed model. For predicting students final result SVR with linear kernel performed better than others where for predicting programming skill C4.5 decision tree performed better than others. For clustering we used K-Means clustering algorithm with Elbow method to find appropriate number of clusters in our datasets. We further built our proposed classification model using elven most significant features to classify new student’s profile which accuracy and F1-score is very impressive.

**Chapter 6**

**Conclusion and Future Work**

## **6.1** Conclusion

By building confusion matrix we can see that SVM Classification model is more accurate with accuracy rate of 98.73%. Though SVM with RBF kernel and Linear Kernel both produces 98.73% accuracy but SVM with RBF kernel produces more number of false positive event which is not convenient for our model. As our aim is to forbade students who have least chance to become successful in this program. Having SVM (linear) with better F1 score we can say it is the best Classifier Model for our dataset.

## 6.2 Contribution

For predicting student’s success in Computer Science program we first have built students’ profile having twenty-one features. Students’ academic result, personal interest, experience with ICT, online course experience, problem solving skills and other experiences are included in Students’ profile. We have used MLR, SVR, and Regression Analysis with Decision Tree for predicting students’ final result and programing skill. Afterwards we have used K-means clustering algorithm to find out the number of appropriate clusters in our dataset. Later we have labeled students profile by using domain knowledge. We further have shortlisted most important features and implemented different classifier algorithm E.g. KNN, Naive Bayes, Random Forest Tree, SVM. We have also built confusion matrix to evaluate the accuracy and efficiency of our proposed model.

## 6.3 Future Work

In future Classification Models can be built for other important under graduation Program like Business, medicine, engineering etc. Many important attributes like family background, economical condition of students etc. might be considered to enhance the study. Many upcoming modern techniques can also be used in future work.

|  |  |
| --- | --- |
|  | **REFERENCES** |
| [1] | Al-Radaideh, Qasem A., Emad M. Al-Shawakfa, and Mustafa I. Al-Najjar. "Mining student data using decision trees." In *International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan*. 2006. |
| [2] | Jirapanthong, Waraporn. "Classification model for selecting undergraduate programs." In *Natural Language Processing, 2009. SNLP'09. Eighth International Symposium on*, pp. 89-95. IEEE, 2009. |
| [3] | Daily Industry News, “Bangladesh becomes 4th largest remittance source for India”, 2018 |
| [4] | Osmanbegović, Edin, and Mirza Suljić. "Data mining approach for predicting student performance." *Economic Review* 10, no. 1 (2012): 3-12. |
| [5] | Daud, Ali, Naif Radi Aljohani, Rabeeh Ayaz Abbasi, Miltiadis D. Lytras, Farhat Abbas, and Jalal S. Alowibdi. "Predicting student performance using advanced learning analytics." In *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 415-421. International World Wide Web Conferences Steering Committee, 2017. |
| [6] | Ramaswami, M., and R. Bhaskaran. "A CHAID based performance prediction model in educational data mining." *arXiv preprint arXiv:1002.1144* (2010). |
| [7] | Tair, Mohammed M. Abu, and Alaa M. El-Halees. "Mining educational data to improve students' performance: a case study." *International Journal of Information* 2, no. 2 (2012): 140-146. |
| [8] | Bhardwaj, Brijesh Kumar, and Saurabh Pal. "Data Mining: A prediction for performance improvement using classification." *arXiv preprint arXiv:1201.3418* (2012). |
| [9] | Baradwaj, Brijesh Kumar, and Saurabh Pal. "Mining educational data to analyze students' performance." *arXiv preprint arXiv:1201.3417* (2012). |
| [10] | Zhao, Yijun. "Data mining techniques." (2015). |
| [11] | Ali, Sayyed Muzammil, and Ms RR Tuteja. "Data Mining Techniques." (2014). |
| [12] | Hand, David J. "Principles of data mining." *Drug safety* 30, no. 7 (2007): 621-622. |
| [13] | Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011. |
| [14] | Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." Emerging artificial intelligence applications in computer engineering 160 (2007): 3-24. |
| [15] | Ostrom, Charles W. Time series analysis: Regression techniques. Vol. 9. Sage, 1990. |
| [16] | Hamilton, James Douglas. Time series analysis. Vol. 2. Princeton, NJ: Princeton university press, 1994. |
| [17] | Steinbach, Michael, George Karypis, and Vipin Kumar. "A comparison of document clustering techniques." In *KDD workshop on text mining*, vol. 400, no. 1, pp. 525-526. 2000. |
| [18] | Efroymson, M., “Multiple regression analysis, Mathematical Methods for Digital Computers”, Vol. 1, pp. 191-203 (1960) |
| [19] | Han, Jiawei, Jian Pei, and Micheline Kamber. “Data mining: concepts and techniques.” Elsevier, 2011. |
| [20] | Nasrabadi, Nasser M. "Pattern recognition and machine learning." *Journal of electronic imaging* 16, no. 4 (2007): 049901. |
| [21] | Mining, What Is Data. "Data Mining: Concepts and Techniques." Morgan Kaufinann (2006). |
| [22] | Karegowda, Asha Gowda, A. S. Manjunath, and M. A. Jayaram. "Comparative study of attribute selection using gain ratio and correlation based feature selection." International Journal of Information Technology and Knowledge Management 2.2 (2010): 271-277. |
| [23] | Kumar, S. Venkata Krishna, and S. Padmapriya. "An efficient recommender system for predicting study track to students using data mining techniques." *International Journal of Advanced Research in Computer and Communication Engineering* 3, no. 9 (2014): 7996-9. |
| [24] | Vialardi, Cesar, Javier Bravo, Leila Shafti, and Alvaro Ortigosa. "Recommendation in Higher Education Using Data Mining Techniques." *International Working Group on Educational Data Mining* (2009) |
| [25] | Romero, Cristóbal, Manuel-Ignacio López, Jose-María Luna, and Sebastián Ventura. "Predicting students' final performance from participation in on-line discussion forums." *Computers & Education* 68 (2013): 458-472. |
| [26] | Yeasmin, Suraiya, Rubayat Jinnah, and Atoshi Islam. "Analysis of Student Performance using Data Mining." PhD diss., Department of Computer Science and Engineering, Military Institute of Science and Technology, 2014. |
| [27] | Alharbi, Zahyah, James Cornford, Liam Dolder, and Beatriz De La Iglesia. "Using data mining techniques to predict students at risk of poor performance." (2016). |
| [28] | Goga, Maria, Shade Kuyoro, and Nicolae Goga. "A recommender for improving the student academic performance." *Procedia-Social and Behavioral Sciences* 180 (2015): 1481-1488. |
| [29] | Alfiani, Ardita Permata, and Febriana Ayu Wulandari. "Mapping student's performance based on data mining approach (a case study)." *Agriculture and Agricultural Science Procedia* 3 (2015): 173-177. |
| [30] | Al-Radaideh, Ahmad Al Ananbeh, and Emad M. Al-Shawakfa, “A classification model for predicting the suitable study track for school students,” International Journal for Research and Reviews in Applied Science, 2011. |
| [31] | Mativo, John M., and Shaobo Huang. "Prediction of students' academic performance: Adapt a methodology of predictive modeling for a small sample size." In *Frontiers in Education Conference (FIE), 2014 IEEE*, pp. 1-3. IEEE, 2014. |
| [32] | Arsad, Pauziah Mohd, and Norlida Buniyamin. "A neural network students' performance prediction model (NNSPPM)." In *Smart Instrumentation, Measurement and Applications (ICSIMA), 2013 IEEE International Conference on*, pp. 1-5. IEEE, 2013. |
| [33] | Huang, Shaobo, and Ning Fang. "Work in progress: Early prediction of students' academic performance in an introductory engineering course through different mathematical modeling techniques." In *Frontiers in Education Conference (FIE), 2012*, pp. 1-2. IEEE, 2012. |
| [34] | Patil, Priyanka Anandrao, and R. V. Mane. "Prediction of Students Performance Using Frequent Pattern Tree." In *Computational Intelligence and Communication Networks (CICN), 2014 International Conference on*, pp. 1078-1082. IEEE, 2014. |
| [35] | Guleria, Pratiyush, Niveditta Thakur, and Manu Sood. "Predicting student performance using decision tree classifiers and information gain." In *Parallel, Distributed and Grid Computing (PDGC), 2014 International Conference on*, pp. 126-129. IEEE, 2014. |
| [36] | Bunkar, Kamal, Umesh Kumar Singh, Bhupendra Pandya, and Rajesh Bunkar. "Data mining: Prediction for performance improvement of graduate students using classification." In *Wireless and Optical Communications Networks (WOCN), 2012 Ninth International Conference on*, pp. 1-5. IEEE, 2012. |
| [37] | Hijazi, S. T., and Naqvi, R.S.M.M., “Factors Affecting Student’s Performance: A Case of Private Colleges”, Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006. |
| [38] | Pandey, Umesh Kumar, and Saurabh Pal. "Data Mining: A prediction of performer or underperformer using classification." *arXiv preprint arXiv:1104.4163* (2011). |
| [39] | Rahm, Erhard, and Hong Hai Do. "Data cleaning: Problems and current approaches." *IEEE Data Eng. Bull.* 23, no. 4 (2000): 3-13. |
| [40] | Karegowda, Asha Gowda, A. S. Manjunath, and M. A. Jayaram. "Comparative study of attribute selection using gain ratio and correlation based feature selection." *International Journal of Information Technology and Knowledge Management* 2, no. 2 (2010): 271-277. |
| [41] | Chellapilla, Kumar, and David B. Fogel. "Anaconda defeats hoyle 6-0: A case study competing an evolved checkers program against commercially available software." Evolutionary Computation, 2000. Proceedings of the 2000 Congress on. Vol. 2. IEEE, 2000. |
| [42] | Langtangen, Hans Petter, and Hans Petter Langtangen. A primer on scientific programming with Python. Vol. 2. Berlin, Germany: Springer, 2009. |
| [43] | Singhal, Swasti, and Monika Jena. "A study on WEKA tool for data preprocessing, classification and clustering." International Journal of Innovative technology and exploring engineering (IJItee) 2.6 (2013): 250-253. |
| [44] | Field, Andy. Discovering statistics using IBM SPSS statistics. sage, 2013. |
| [45] | Hastie, T., Tibshirani, R., and Friedman, J., “The Elements of Statistical Learning: Data Mining, Inference, and Prediction”, Springer, 2 edition (2009). |
| [46] | Han, Jiawe, and Micheline Kamber. "Data mining concepts and techniques San Francisco Moraga Kaufman." (2001). |
| [47] | Hearst, Marti A., et al. "Support vector machines." IEEE Intelligent Systems and their applications 13.4 (1998): 18-28. |
| [48] | Gunn, Steve R. "Support vector machines for classification and regression." ISIS technical report 14.1 (1998): 5-16. |
| [49] | J. R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, Inc, 1992. |
| [50] | J. R. Quinlan, "Introduction of decision tree", Journal of Machine learning", pp. 81-106, 1986 |
| [51] | Alsabti, Khaled, Sanjay Ranka, and Vineet Singh. "An efficient k-means clustering algorithm." (1997). |
| [52] | Kanungo, Tapas, et al. "An efficient k-means clustering algorithm: Analysis and implementation." IEEE Transactions on Pattern Analysis & Machine Intelligence 7 (2002): 881-892. |
| [53] | Zhang, Min-Ling, and Zhi-Hua Zhou. "A k-nearest neighbor based algorithm for multi-label classification." Granular Computing, 2005 IEEE International Conference on. Vol. 2. IEEE, 2005. |
| [54] | Zhang, Min-Ling, and Zhi-Hua Zhou. "ML-KNN: A lazy learning approach to multi-label learning." Pattern recognition 40.7 (2007): 2038-2048. |
| [55] | Murphy, Kevin P. "Naive bayes classifiers." University of British Columbia 18 (2006). |
| [56] | Keogh, Eamonn. "Naive bayes classifier." Accessed: Nov 5 (2006): 2017. |
| [57] | Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22. |

**Appendix-I**

Student related Variables

**Table A 1.1:** Student related Variables

|  |  |  |
| --- | --- | --- |
| Variables | Description | Possible Values |
| Gender | Students Gender | {Male, Female} |
| UniLoc | University Location | {Dhaka, Chittagong, Other City, Outside City} |
| GSSC | Students grade in SSC | {5.00  >4.50 and <5.00,  >4.00 and <4.50,  >3.5 and <4.00  >3.5} |
| GHSC | Students grade in HSC | {5.00  >4.50 and <5.00,  >4.00 and <4.50,  >3.5 and <4.00,  >3.5} |
| ICTResult | Students ICT result in HSC | {Very Good, Good, Satisfactory, Less than satisfactory, poor} |
| CCGPA | Students current CGPA | { >3.80,  >3.40 and <3.80,  >3.00 and <3.40,  >2.5 and <3.00,  <2.5} |
| ProRatings | Students programming skill ratings | { 9 to 10,  7 to 8,  5 to 6,  3 to 4,  1 to 2} |
| Bmath | Students skill in basic mathematics | {Very Good, Good, Satisfactory, Less than satisfactory, poor} |
| HMath | Students skill in Higher level mathematic | {Very Good, Good, Satisfactory, Less than satisfactory, poor} |
| PaRatings | Students Patience’s ratings | { 9 to 10,  7 to 8,  5 to 6,  3 to 4,  1 to 2} |
| ICTExp | Students experience with ICT | {Very Interesting, Interesting, Neutral, Difficult, Very Difficult } |
| IBReason | Most important reason behind browsing internet | {Online courses and Reading Blogs, Online Courses only, Reading Blogs only, Study and Social Media, Social media and Entertainment} |
| MSOlym | Student participation in number of math or science Olympiad | {many , less than 5, more than once, once, never} |
| NOC | Number of online courses followed | {many, less than 5, more than once, once, never} |
| Cpro | Students competitive programming Interest ratings | { 9 to 10,  7 to 8,  5 to 6,  3 to 4,  1 to 2} |
| SSR | Students capability of self-study ratings | { 9 to 10,  7 to 8,  5 to 6,  3 to 4,  1 to 2} |
| ORH | Tendency of using online resources helps | {Strongly agree, agree, neutral, disagree, strongly disagree} |
| PH | Having Patience helps | {Strongly agree, agree, neutral, disagree, strongly disagree} |
| SSH | Self-study helps | {Strongly agree, agree, neutral, disagree, strongly disagree} |
| CGC | Computer gamers chance in Computer Science | {Strongly agree, agree, neutral, disagree, strongly disagree} |
| ICTRR | ICT result reflects success in CSE | {Strongly agree, agree, neutral, disagree, strongly disagree} |
| PC | Knowing about PC configuration helps in reading Computer Science? | {Strongly agree, agree, neutral, disagree, strongly disagree} |

**Appendix-II**

**Frequency Tables of Survey Data**

**Frequency Table**

In this section, Frequency tables will show the frequency, percent valid percent and cumulative percent of the student’s Academic Result.

Now we analyze B1 question of questionnaire is shown at **Table A2.1**

**Table A2.1** Frequency table of B1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Less than 3.50  3.50 and above  4.00 and above  4.50 and above  5.00  Total | 203  211  65  9  6  394 | 51.52  28.17  16.50  2.28  1.52  100.0 | 51.5  28.2  16.5  2.3  1.5  100.0 | 51.5  79.7  96.2  98.5  100.0 |

Now we analyze C1 question of questionnaire is shown at **A2.2**

**Table A2.2** Frequency table of C1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Very difficult | 3 | .8 | .8 | .8 |
| difficult | 14 | 3.6 | 3.6 | 4.3 |
| Neutral | 26 | 6.6 | 6.6 | 10.9 |
| interesting | 188 | 47.7 | 47.7 | 58.6 |
| Very interesting | 163 | 41.4 | 41.4 | 100.0 |
| Total | 394 | 100.0 | 100.0 |  |

Now we analyze D1 question of questionnaire is shown at **Table A2.3**

**Table A2.3** Frequency table of D1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Social media and Entertainment | 83 | 21.1 | 21.1 | 21.1 |
| Online Study and Social Media | 239 | 60.7 | 60.7 | 81.7 |
| Reading Blogs only | 6 | 1.5 | 1.5 | 83.2 |
| Online Study and Social Media | 20 | 5.1 | 5.1 | 88.3 |
| Social media and Entertainment | 46 | 11.7 | 11.7 | 100.0 |
| Total | 394 | 100.0 | 100.0 |  |

Now we analyze D3 question of questionnaire is shown at **Table A2.4**

**Table A2.4** Frequency table of D3

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Never | 34 | 8.6 | 8.6 | 8.6 |
| Once | 76 | 19.3 | 19.3 | 27.9 |
| More than once | 64 | 16.2 | 16.2 | 44.2 |
| More Than 5 | 99 | 25.1 | 25.1 | 69.3 |
| Many | 121 | 30.7 | 30.7 | 100.0 |
| Total | 394 | 100.0 | 100.0 |  |

Now we analyze D4 question of questionnaire is shown at **Table A2.5**

**Table A.2.5** Frequency table of D4

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | 1 to 2 | 25 | 6.3 | 6.3 | 6.3 |
| 3 to 4 | 89 | 22.6 | 22.6 | 28.9 |
| 5 to 6 | 100 | 25.4 | 25.4 | 54.3 |
| 7 to 8 | 90 | 22.8 | 22.8 | 77.2 |
| 9 to 10 | 90 | 22.8 | 22.8 | 100.0 |
| Total | 394 | 100.0 | 100.0 |  |

Now we analyze E1 question of questionnaire is shown at **Table A2.6**

**Table A2.6** Frequency table of E1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Strongly Disagree | 2 | .5 | .5 | .5 |
| Disagree | 9 | 2.3 | 2.3 | 2.8 |
| Neutral | 25 | 6.3 | 6.3 | 9.1 |
| Agree | 206 | 52.3 | 52.3 | 61.4 |
| Strongly Agree | 152 | 38.6 | 38.6 | 100.0 |
| Total | 394 | 100.0 | 100.0 |  |

Now we analyze E2 question of questionnaire is shown at **Table A2.7**

**Table A2.7** Frequency table of E2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Strongly Disagree | 2 | .5 | .5 | .5 |
| Disagree | 20 | 5.1 | 5.1 | 5.6 |
| Neutral | 60 | 15.2 | 15.2 | 20.8 |
| Agree | 185 | 47.0 | 47.0 | 67.8 |
| Strongly Agree | 127 | 32.2 | 32.2 | 100.0 |
| Total | 394 | 100.0 | 100.0 |  |

Now we analyze E3 question of questionnaire is shown at **Table A2.8**

**Table A2.8** Frequency table of E3

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Strongly Disagree | 2 | .5 | .5 | .5 |
| Disagree | 20 | 5.1 | 5.1 | 5.6 |
| Neutral | 60 | 15.2 | 15.2 | 20.8 |
| Agree | 185 | 47.0 | 47.0 | 67.8 |
| Strongly Agree | 127 | 32.2 | 32.2 | 100.0 |
| Total | 394 | 100.0 | 100.0 |  |

Now we analyze E4 question of questionnaire is shown at **Table A2.9**

**Table A2.9** Frequency table of E4

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Strongly disagree | 15 | 3.8 | 3.8 | 3.8 |
| Disagree | 44 | 11.2 | 11.2 | 15.0 |
| Neutral | 122 | 31.0 | 31.0 | 45.9 |
| Agree | 162 | 41.1 | 41.1 | 87.1 |
| Strongly Agree | 51 | 12.9 | 12.9 | 100.0 |
| Total | 394 | 100.0 | 100.0 |  |

Now we analyze E5 question of questionnaire is shown at **Table A2.10**

**Table A2.10** Frequency table of E5

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Strongly Disagree | 3 | .8 | .8 | .8 |
| Disagree | 30 | 7.6 | 7.6 | 8.4 |
| Neutral | 50 | 12.7 | 12.7 | 21.1 |
| Agree | 174 | 44.2 | 44.2 | 65.2 |
| Strongly Agree | 137 | 34.8 | 34.8 | 100.0 |
| Total | 394 | 100.0 | 100.0 |  |

Now we analyze E6 question of questionnaire is shown at **Table A2.11**

**Table A2.11** Frequency table of E6

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Strongly Disagree | 3 | .8 | .8 | .8 |
| Disagree | 16 | 4.1 | 4.1 | 4.8 |
| Neutral | 79 | 20.1 | 20.1 | 24.9 |
| Agree | 204 | 51.8 | 51.8 | 76.6 |
| Strongly Agree | 92 | 23.4 | 23.4 | 100.0 |
| Total | 394 | 100.0 | 100.0 |  |

Now we analyze E7 question of questionnaire is shown at **Table A2.12**

**Table A2.12** Frequency table of E7

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Very difficult | 21 | 5.3 | 5.3 | 5.3 |
| difficult | 26 | 6.6 | 6.6 | 11.9 |
| Neutral | 77 | 19.5 | 19.5 | 31.5 |
| interesting | 179 | 45.4 | 45.4 | 76.9 |
| Very interesting | 91 | 23.1 | 23.1 | 100.0 |
|  | Total | 394 | 100.0 | 100.0 |  |

Now we analyze F2 question of questionnaire is shown at **Table A2.13**

**Table A2.13** Frequency table of F2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Poor | 0 | 0 | 0 | 0 |
| Less than satisfactory | 5 | 1.3 | 1.3 | 1.3 |
| Satisfactory | 52 | 13.2 | 13.2 | 14.5 |
| Good | 136 | 34.5 | 34.5 | 49.0 |
| Very Good | 201 | 51.0 | 51.0 | 100.0 |
| Total | 394 | 100.0 | 100.0 |  |

Now we analyze F3 question of questionnaire is shown at **Table A2.14**

**Table A2.14** Frequency table of F3

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Poor | 0 | 0 | 0 | 0 |
| Less than satisfactory | 15 | 3.8 | 3.8 | 3.8 |
| Satisfactory | 82 | 20.8 | 20.8 | 24.6 |
| Good | 162 | 41.1 | 41.1 | 65.7 |
| Very Good | 135 | 34.3 | 34.3 | 100.0 |
|  | Total | 394 | 100.0 | 100.0 |  |

Now we analyze F4 question of questionnaire is shown at **Table A2.15**

**Table A2.15** Frequency table of F4

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
|  | 1 to 2 | 12 | 3.0 | 3.0 | 3.0 |
|  | 3 to 4 | 66 | 16.8 | 16.8 | 19.8 |
| Valid | 5 to 6 | 71 | 18.0 | 18.0 | 37.8 |
|  | 7 to 8 | 154 | 39.1 | 39.1 | 76.9 |
|  | 9 to 10 | 91 | 23.1 | 23.1 | 100.0 |
|  | Total | 394 | 100.0 | 100.0 |  |

Now we analyze F5 question of questionnaire is shown at **Table A2.16**

**Table A2.16** Frequency table of F5

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | 1 to 2 | 7 | 1.8 | 1.8 | 1.8 |
| 3 to 4 | 78 | 19.8 | 19.8 | 21.6 |
| 5 to 6 | 68 | 17.3 | 17.3 | 38.8 |
| 7 to 8 | 141 | 35.8 | 35.8 | 74.6 |
| 9 to 10 | 100 | 25.4 | 25.4 | 100.0 |
| Total | 394 | 100.0 | 100.0 |  |