

Student Performance Prediction and Teacher Recommender System

By

Muthara-Tul-Ain

MASTER OF SCIENCE IN COMPUTER SCIENCE



**DEPARTMENT OF COMPUTER SCIENCE
CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY
ISLAMABAD
2017**

Student Performance Prediction and Teacher Recommender System

By

Muthara-Tul-Ain

A research thesis submitted to the Department of Computer
Science, Capital University of Science and Technology, Islamabad
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE



**DEPARTMENT OF COMPUTER SCIENCE
CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY
ISLAMABAD
2017**

Copyright ©2017 by CUST Student

All rights reserved. Reproduction in whole or in part in any form requires the prior written permission of Muthara-Tul-Ain (MS133033) or designated representative



**CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY
ISLAMABAD**

Islamabad Expressway, Kahuta Road, Zone-V, Islamabad

Phone: +92 51 111 555 666, Fax: 92 51 4486705

Email: info@cust.edu.pk, Website: http://www.cust.edu.pk

CERTIFICATE OF APPROVAL

**Student Performance Prediction and Teacher Recommender
System**

by

Muthara-Tul-Ain

MS133033

THESIS EXAMINING COMMITTEE

S No	Examiner	Name	Organization
(a)	External Examiner	Dr. Ehsan Munir	CIIT, Wah Cantt
(b)	Internal Examiner	Dr. Muhammad Arshad Islam	CUST, Islamabad
(c)	Supervisor	Dr. Nayyer Masood	CUST, Islamabad

Dr. Nayyer Masood
Thesis Supervisor

November, 2017

Dr. Nayyer Masood
Head

Department of Computer Science

Dated : November, 2017

Dr. Muhammad Abdul Qadir
Dean

Faculty of Computing

Dated : November, 2017

DEDICATED

TO

MY

RESPECTED

ALLAH (SWT) & PROPHET MUHAMMAD (PBUH)

&

PARENTS, BROTHER, SISTER



FOR

THEIR CONSTANT

AFFECTION AND SUPPORT

ACKNOWLEDGMENT

All praise and exaltation is due to ALLAH (S.W.T) The creator and sustainer of all seen and unseen worlds. First and foremost I would like to express my gratitude and thanks giving to Him for providing me the boundaries and blessings to complete this work. Secondly, I would like to express my sincerest appreciation to my supervisor **Dr Nayyar Masood** for his directions, assistance, and guidance. I sincerely thanked for his support, encouragement and technical advice in the research area. I am heartily thankful to him from the final level, as he enabled me to develop an understanding of the subject. He has taught me, both consciously and unconsciously, how good experimental work is carried out. Sir you will always be remembered in my prayers.

I am highly indebted to my parents and my family, for their expectations, assistance, support and encouragement throughout the completion of this Master of Science degree. They form the most important part of my life. After ALLAH (S.W.T) they are the sole source of my being in this world. No words can ever be sufficient for the gratitude I have for my parents and for my family. A special thanks to my employing company for their support and encouragement to complete this Master of Science degree.

I pray to ALLAH (S.W.T) that may He bestow me with true success in all fields in both worlds and shower His blessed knowledge upon me for the betterment of all Muslims and whole Mankind.

AAMEEN

Muthara-Tul-Ain

DECLARATION

It is declared that this is an original piece of my own work, except where otherwise acknowledged in text and references. This work has not been submitted in any form for another degree or diploma at any university or other institution for tertiary education and shall not be submitted by me in future for obtaining any degree from this or any other University or Institution.

Muthara-Tul-Ain

Reg. No, MS133033

November, 2017

ABSTRACT

Mining data and extracting information from huge databases has become an interesting research area for the researchers. The idea to extract information with the help of data mining techniques came into being since a couple of decades ago. Initially, researchers were supposed to apply classification and clustering techniques to partite the dataset and analyze the intrinsic features. On the basis of such features, they make reasonable predictions. These predictions have taken place in the field of educational data mining for many purposes such as; predict the performance of students on the basis of factors associated with them, to enable them suitable courses and appropriate teachers. These purposes have been derived from the area of student retention and attrition. Our research aims to achieve these purposes under the roof of student attrition and retention. Moreover, we have identified such exiting factors which are beneficial for predicting the performance to students, recommend them best suitable teachers and help them to select the courses. We have applied classification algorithms with respect to the nature of data which have been collected from the Capital University of Science and Technology (CUST), ISB. In this study, the GPA of first semester on the basis of Midterm and previous academic grades have been tried to predict. For the second semester, we have predicted the CGPA of same students by using their complete proceeding academic record with the help of hybrid approach. The hybrid approach consists of combination of factors that have been evaluated against our research questions. Moreover, we tried to improve their performance by recommending those suitable courses and teachers whose performance is better amongst others comparatively. The reason of collecting the data from CUST is to validate the exiting factors in local context (Pakistan). On the basis of classification algorithms such as; Naïve Bayes and J48, we have become able to build the recommender system. Then the factors which contributed well to validate the exiting factors in local context have been measured. In the last, appropriate teacher allocation has been measured by two ways; Statistical and Prediction. In statistical experimentations, average performance of teachers, Z- test and ANOVA test has been applied. In prediction experimentations, one subject teacher with other

subject teacher's name attribute, and one subject teacher without other subject teacher's name attribute and overall performance of teachers have been computed with respect to each subject independently. With the help of our research we might have become able to provide a way to educational institutions to reduce the attrition and increase the retention rate.

Table of Contents

Abstract.....	vii
Introduction.....	1
1.1 Background Of Research	2
1.2 Problem Statement.....	6
1.3 Research Questions.....	6
1.4 Scope.....	6
1.5 Application Of Proposed Approach.....	6
1.6 Significance Of Solution.....	7
1.7 Organization Of Thesis	8
1.8 Definitions And Frequently Used Terms.....	8
Literature Review	9
2.1 Educational Data Mining	10
2.2 Student Performance Prediction	14
2.3 Commonly Used Approaches In Edm	16
2.5 Commonly Used Attributes For Student Performance Prediction.	19
2.5 Literature Review Summary	21
Research Methodology	22
3.1 Data Collection And Pre-Processing	24
3.2 Classification.....	27
3.3 Algorithms And Techniques	29
3.4 Evaluation Of Research Questions	30
Results & Evaluations.....	35
4.2 Final Attribute Selection	38
4.3 Result Of Evaluation Of Research Questions	38
Conclusion And Future Work	67
5.1 Future Work	70
References.....	71

List of Tables

Table 2.1: Literature Review Summary.....	21
Table 3.2: Summary Of Session-Wise Attrition.....	24
Table 3.3: Summary Of Session-Wise Attrition.....	25
Table 3.4: Selection Of Attributes.....	28
Table 3.5: Attributes Selection To Dataset	31
Table 3.6: Pattern Of Combination Of Attributes.....	32
Table 4.1: Occurrence Of Missing Values	36
Table 4.2: Handling Of Missing Values	37
Table 4.3: Attributes Table.....	39
Table 4.4: Classification Of Attributes And Algorithms	40
Table 4.5 Ranked Attributes.....	42
Table 4.6 GPA PREDICTION FOR THE FIRST SEMESTER	42
Table 4.7: Cgpa Prediction For The Second Semester	44
Table 4.8: Pre-Qualification Vs Maths Courses	47
Table 4.9: Average Of Cal-LI Teachers.....	48
Table 4.10: Average Of Cal-I Teachers	49
Table 4.11: Overall Average Of Maths Teachers	50
Table 4.12: Anova Analysis On Cal-1	52
Table 4.13: Two Samples For Mean With Respect To Teacher “E”.....	53
Table 4.14: Two Samples For Means With Respect To Teacher “B”	54
Table 4.15: Teacher B With Same Number Of Students.....	55
Table 4.16: Teacher B With Different Number Of Students	55
Table 4.17: Comparison Of All Programming Courses	56
Table 4.18: Comparison Of Actual And Predicted Grades Of Itc/Itp.....	57
Table 4.19: Comparison Of Actual And Predicted Performance Of Oop/Cp Teachers.....	58
Table 4.20: Overall Performance Of Oop Teachers.....	59

List of Figures

Figure 3.1: DATA FLOW DIAGRAM OF PROPOSED METHODOLOGY	23
Figure 4.1: Formats Of Data.....	36
Figure 4.2: comparison Of Local And Existing Attributes.....	41
Figure 4.3: Gpa Prediction For The 1st Semester	44
Figure 4.4: Cgpa Prediction For The Second Semester.....	45
Figure 4.5: Pre-Qualification Vs Maths Courses	48
Figure 4.6: Average Of Cal-LI Teachers	49
Figure 4.7: Overall Average Of Maths Teachers	50
Figure 4.8: Comparison Of Itc/Itp Teacher.....	51
Figure 4.9: Comparison Of Actual And Predicted Grades Of Itc/Itp.....	56
Figure 4.10: Comparison Of Actual And Predicted Performance Of Oop/Cp Teachers.....	58
Figure 4.11: Overall Performance Of Oop Teachers	59
Figure 4.12: Overall Performance Of Oop Teachers	60

List of Abbreviations

CUST: Capital University of Science and Technology

WEKA: Waikato Environment for Knowledge Analysis

MOOCs: Massive Open Online Courses

CSV: Comma-separated values (*Comma-delimited*)

ARFF: Attribute-Relation File Format.

GPA: Grade point average

CGPA: Cumulative Grade Point Average

Cal-I: Calculus-I

Cal-II: Calculus-II

ITC: Information Technology Computing

ITP: Information Technology Programming

CP: Computer Programming

OOP: Object Oriented Programming

SEM: Semester

Chapter 1

INTRODUCTION

Researchers have been working in the area of educational data mining for a decade. Educational Data Mining (EDM) has become a broadened field in which researchers are conducting their experiments to extract useful information from the data belonging to the educational sectors for many purposes. The purposes include identifying student attrition and retention rate, students' performance prediction, building a course recommender system, and teacher recommender system, etc. In the field of Educational Data Mining (EDM), researchers are busy in exploring effectiveness or role of different types of variables to measure and predict the performance of students. Among those variables, student's age, academic record and biography of students are involved. EDM is a growing research field, which carries data mining techniques in educational system (Romero, C., & Ventura, S. 2007). In previous study, (Romero, C., & Ventura, S. 2013) has explored the phenomenon of making the student's outcome better with the help of data mining approaches. According to them, huge data from the institutions have problems associated with it. Therefore, data mining approaches are not supposed to be applied directly on this data. Therefore, knowledge discovery process has been implemented. In educational institutions, data mining is playing a pivotal role on datasets after preprocessing. Most widely used data mining approaches involved classification and clustering, outlier detection, association rule mining, pattern mining and text mining. Researchers are trying to extract useful, novel and interesting information with the help of data mining approaches (Romero, C., & Ventura, S. 2010). EDM process is used to convert raw educational data into useful information.

In the field of EDM, researchers are active in three broad categories; Personal recommender system and learning environment, course management system, and student attrition and retention. Our research will cover all these mentioned areas in different contexts such as identification of students with low academic performance, building prediction model to predict the student's performance by

using their historical data, improvement in raising the confidence level among them, assist them to choose their courses, building a model that will assign appropriate teachers to the students.

1.1 Background of research

As discussed in above section, EDM is subfield of data mining which is applied in educational institutions to help them to maintain their prestige. Educational data mining is all related to propose and develop novel approaches to improve the performance of students on a large scale. Educational data mining is an active research areas in which researchers are working actively in three main areas of educational data mining are:

a) Personal recommender system and learning environment

Researchers are using this concept in hybrid fashion in the context of educational data mining. They are constantly using various data mining approaches to make the personal recommender system better to produce maximum accurate results. In a research (Miller, B. 2004), personal recommender system uses collaborative filtering to reduce noise from the data. A limitation of personal recommender system is found to be its un-portability as they are operational only on large computers that are connected to the internet. Another limitation was found to be its trustworthy relationship between the owner of recommender system and its user.

b) Course management system and educational data mining

Researchers are associated with this branch of EDM to provide the universities and educational institutions a platform that can help them in management of courses in the department. For this purpose, different data mining approaches have been carried out (Peña-Ayala, A. 2014). The concept has been derived from e-learning system “Moodle”. Moodle is a learning management system that helps educators to build up effective online learning communications through the analysis of information generated by the users (Aher, S. B, 2012).

c) Student attrition and retention

Student attrition means the number of students who are leave their courses without completing because of certain reasons. Those reasons may include poor choices of courses, consistent undesirable result, poor performance, financially instability and immature selection of courses etc whereas student retention means the number of students who complete their course and acquire degree despite of any kind of circumstance and gain good grades in the transcript. In educational institutions, rate of student attrition and retention has reached up to considerable unit (Stallone, M. N. 2011).

Student attrition and retention not only affect the performance of education institution or department but also affect the faculty positions and raise the financial problems for the parents eventually. To analyze the pattern of student retention and attrition, one has to find out the reasons of this cause then take steps to resolve this problem. Our research topic falls into the type student attrition and retention which is further explained in the following subsection in detail. For this research collected dataset from Capital University of science and Technology, Department of Computer Science. Let's consider first semester and second semester courses because student attrition rate is high in first and second semester. These semesters are helpful attrition of students and predicted the student performance.

Student performance prediction

To recognize the efforts of students in academia, there are some known parameters carried out by universities or educational institutions (Romero, C., & Ventura, S, 2007). Among those parameters, student's grades, attendance, grades per courses are commonly evaluated. Student performance prediction has been the growing research problem for the researchers since a decade. To improve the institution's prestige and efficiency with respect to student performance improvement, there are several techniques used by different researchers such as course specific regression, personalized linear multi regression, classification and clustering. These approaches not only work for educational institutions those are in existence physically but also for virtual campuses too i-e; LMS and Moore of Stanford University (Elbadrawy, et al,

2016). EDM is a broad research field in which researchers are exploring a lot of problems including student's performance, attrition and retention, course selection, and particular teacher's selection. The widely covered research problems in the domain of educational data mining are known as student's future grade prediction, performance prediction and course enrollment recommender system.

This research is based on different aspects. One aspect is to identify the number of those students who had to leave the institution because of their poor performance in studies or financial problem. Then the number of those students have been found who completed their degree belong to the spring semesters of three years. Expectedly this research will help the students to improve their results in the courses which have been taken for the experiments by recommending them appropriate teachers and courses based on their associated factors.

This is helpful not only for students to raise their grades but also for parents not to suffer from financial problems. Another aspect which is being considered is to retain maximum percentage of students in the university and fulfill their degree requirements till its completion. Ultimately, if the grades of students would be good their retention will become stronger in the institution. This research might reduce the attrition rate because of the provision of teacher recommender system, course recommender system and student performance prediction. The outcome of our research has been discussed in detail as follows.

Students Performance Prediction is the research problem which has been identified by many researchers with the help of different approaches such as Matrix Factorization, classification and clustering algorithms (Polyzou, A., & Karypis, G, 2016). Before applying these approaches, students variables are supposed to be extracted from the collected dataset that are unique registration number, result of previous two terms, GPA (Grade Point Average), number of family members, family welfare, gender, usage of social networking and use of technology and internet (Oskouei, R. J., & Askari, M, 2014). After collection, data preprocessing is applied in which positive and negative effects of dataset

are removed and then classification and clustering are applied. By using such techniques, data can be further divided into clusters and classes to analyze useful and hidden patterns from it (Kabakchieva, D, 2013).

With respect to performance prediction of students, there are many cases such as student's next term grade prediction, presence performance based on assessment in the current courses. To evaluate this problem, researchers have used some of the characteristics of students such as admission records, High school scores, SAT/ACT scores and grades of previously completed courses (Elbadrawy, et al, 2016). More characteristics such as class test, seminar attendance and marks, assignment marks (Baradwaj, B. K., & Pal, S. 2012).

Course enrollment recommender system It helps to determine the performance of students by recommending them courses up to their performance. For this purpose researchers have proposed the recommender `in which he used 67 templates of active students. This recommender system not only helps to predict the grades of students, it also helps to recommend the courses to the students by considering their timetable (Bydžovská, H, 2013).

Proposed attributes of students affects their performance in various ways. It not only helps in identifying the students with low academic performance but also paves the ways to improve their grades and enables recommender systems of universities to predict the grades of students. For this purpose, different classification techniques of data mining have been studied in this research. With the help of this study, student carrier can be affected by means of building up and students can easily pave their way towards their desired destination. This study will also help to construct the course recommender system which will allow the students to select their courses that are recommended by the proposed system. Another advantage of this study is highly beneficial economically for parents. With the help of proposed system, students may not waste their money in selection of courses to repeat. Moreover, student's academic record can be improved and efficiency of department management can be raised.

The factors used in the research have already been applied in countries other than Pakistan. Those countries are Canada, USA, England, and Nigeria. All these factors have been applied in the colleges and universities of narrated countries and now will be validated in the context of Pakistan.

1.2 Problem Statement

A major problem that the academic institutions are facing worldwide is poor performance of students in academics. This causes high attrition rate that is a loss to students, parents and institutions. Student's performance prediction can help to reduce attrition rate as it raises an early alarm for students not performing well and are likely to leave the institution/studies. Most of the work on performance prediction has been done in foreign institutions; this has to be done in the context of Pakistan for more accurate results.

1.3 Research Questions

Given below is the proposed problem statement

RQ1. Whether existing identified factors for grade prediction are valid in our local context (Pakistan)?

RQ2. Which factors help in accurate prediction of students' GPA of first semester?

RQ3. Is it possible to improve the performance of a student by allocating of appropriate teacher for a subject?

1.4 Scope

The application of this research is versatile. The experiments of this research have been conducted on the dataset that has been collected from Capital University of Science and Technology, Islamabad (Pakistan). Such factors can be applied on the data of Government institutions as well as schools and colleges.

1.5 Application of Proposed Approach

Our proposed research will be beneficial in one of the following ways in the domain of educational data mining.

To make the university management system efficient

This research might be highly beneficial for the university management system to keep check on the performance of student's belong to different departments. This will not only help to predict and improve student's performance but also supports the parents economically.

To enable Course recommender system pro-active

The selection of courses is major problem for the students. To build such recommender system that helps students in selection and registration of the courses will be quite beneficial for the institutions.

Helpful in improving teacher recommender system

In the proposed solution this recommender system will help department to recommend specific teachers to particular students based on previous 0239 academic record and inclination of selected courses in which they pursued good grades.

1.6 Significance of solution

This research is a fine contribution for the educational institutions of Pakistan. With the help of this research, the factors that have been used by the researchers from the foreign universities are considered in local context. These factors are planned to apply over the data set of Capital University of Science and Technology of Pakistan. These factors have been utilized with the help of WEKA for the student's performance prediction. Universities and other educational institutions may capable of predicting the student's next term grades.

Moreover, with the help of the contribution of this research there are chances to propose teacher recommender system that will help to recommend suitable teachers to the students by considering their current performance in the semester and their relative interest in the courses that might improve their grades.

1.7 Organization of thesis

This thesis comprises of five chapters. The first chapter states the introduction of the proposed research. Second chapter is the literature review in which the work and research contribution in chosen area by former researchers has been discussed. Moreover, overall literature summary has been presented in summarized way. Third chapter contains the methodology diagram that has been adopted to perform the experiments and how experiments will be performed. The results of experiments by using selected tools have been presented in the chapter four. Last chapter contains the conclusion and future work in which summary of whole thesis has been presented.

1.8 Definitions and frequently used terms

Performance Prediction

A student's grade is used to acknowledge his/her related performance in academia. It is now possible to predict the student's performance with scientific methods in which quantitative approaches have been used by the researchers (Bhardwaj, B. K, 2012) (Pal, S. 2012).

Recommender System

Usually recommender system is built to assign a particular entity to required entity. In the field of data mining, researchers have proposed some recommender systems in which teacher recommender system and course recommender systems are involved (Bozo, J., 2010) (Alarcón, R.,2010)(Iribarra, S., 2010).

Chapter 2

LITERATURE REVIEW

The initiative of online course learning system is based on e-commerce venture. With the growth of this venture, data from web resources started collecting and storing in the excel that contains customer and product information and order information (Kokina, J, 2017). E-commerce is a term that refers to use as online business through internet. There are various websites that are working for this purpose such as e-bay, Alibaba, and Amazon etc. It can be said that data storage in excel from the resources has been derived from e-commerce. Predicting student's performance by using data mining techniques to extract information from the academic dataset of universities has become state of the art research in the scientific society. Universities are confronting with some challenges now a day to analyze the performance of their students. That's why researchers are focusing on student's profiles and characteristics to make the university management aware of student's performance and overall academic result (Kabakchieva, D, 2013). There is another dimension of student's performance that is the dependence of student retention upon student student's performance. To minimize the problem of student retention cases in the universities, different researchers have proposed different methods to predict the performance of students in their future semester based on the performance of previous one.

To predict the courses of next term grades, four parameters have been considered in this study such as; admission records, High school scores, SAT/ACT scores and grades of previously completed courses. Based upon these parameters, recommender system can be trained to predict the grades of students accurately in any of the educational institution. Historical information about the course has also been considered in this study such as which course is taught by which teacher and information about contents of the course. Many researchers have used LMS and Moore to predict the successive chances of success and failure of students. In this research, regression based methods such as course specific regression (CSPR) and personalized linear multi regression

(PLMR) has been used. Another method known as matrix factorization based methods in which standard matrix factorization (MF) has been used for the grade prediction of students (Elbadrawy, A, et al, 2015).

In this chapter, the background of educational data mining and its branches will be discussed in detail. Also the student's performance prediction and data mining approaches that are commonly used by researchers in the literature are being discussed.

2.1 Educational Data Mining

Data mining is a process of sorting data and extracting information from existing databases.¹ With the help of pattern mining and data analysis, hidden information can be obtained from huge datasets. The strategy of data mining is now applied in the field of education by researchers. They are busy in exploiting a lot of dimensions in education sector. This is now known as educational data mining. Data mining is applying in educational sector by considering the performance of students and finding the position of students by using their academic records. Educational dataset is being collected from various resources such as interactive learning systems, computer-supported collaborative systems, and administrative datasets of school, colleges and universities². Data mining methods are now implemented in well known universities to analyze the patterns of student performance from the dataset through which information can be extract and decision making may become easier for the management of institutions (Kabakchieva, D. 2013).

With the incremental growth in the use of technology everywhere, educational institutions are now busy in finding hidden trends and patterns in their larger datasets (Scheuer, O. 2012). (Merceron, A. 2005). Datasets that are used for experimental purposes in educational institutions have become possibly available because of web based educational systems in which LMS, Moore and Portal system have become common. With the help of these sources, dataset can easily be collected if authorization is accessed. One purpose of extracting information from its own dataset is to make its prestige among other

¹ <http://searchsqlserver.techtarget.com/definition/data-mining>

² <http://www.educationaldatamining.org/>

educational institutions stronger. Another purpose is to build the student career by covering its each and every aspect such as improvement of their grades if lacking, overall performance booster, support them financially, make them enable to select the courses suggested by course recommender system, assign them appropriate teachers based on their inclination of interest in course selection and many more.

However, these are some common areas in which researchers are producing their research by using different data mining techniques. In our research, same areas are being covered under the dataset of Pakistani students. Three factors are being applied on students those factors have been discussed briefly in section 2.4. In current section, our focus will remain on key areas of EDM, in which researchers are engaged.

a) Personal recommender and learning environment

Web based learning environments have become common right now. It has become powerful medium to act as a bridge between learner and instructor and provides interesting learning mechanisms to both (Nam, C. S, 2007) (Osmar, R, 2002) has elaborated the concept and working of web based learning which is also known as e-learning. A recommender system helps to recommend the actions to a learner by using his previous actions with the help of intelligent software agents. The use of such recommender systems were first initiated in e-commerce and now carrying out by researchers in the domain of e-learning. Osmar has implemented this recommender system in assisting the offline web miners who are responsible to find hidden trends in the data and online course navigation. With the help of recommender system, educational institutions are now strengthening their management and departmental progress. Researchers have proposed several recommender systems among which course recommender system and teacher recommender system are commonly known.

b) Course management system

By means of course management, researchers have explored different sources to find the best possible use of course management system. In this context, Moodle is found to be open source course management system that helps

educators to make their courses available on the internet. All the data on the Moodle is managed by Moodle team. Moodle is now facilitating educational institutions, community colleges, and schools to create online teaching system (Dougiamas, M., 2003). It delivers courses online, unlike traditional classrooms. Data of courses and students is stored in its specified database which is further used by researchers that carries out for mining purposes to extract useful information from it. Performance of Moodle has raise higher in virtual environment. Other online course management system such as web portals and learning management systems are serving for the same purpose now days.

With the use of internet in classrooms and institutions, those institutions have launched their proper websites or web pages, through which students can register themselves into particular courses (Kaminski, J. 2005). In the area of online course learning, researchers have presented many research articles. In this domain LMS and Moore have contributed a lot. They have been offering different courses on their websites and thousands of students from all over the world registered themselves. This is how, huge amount of dataset is collected from LMS and Moore and researchers have performed analysis on these dataset to evaluate the performance of students who enrolled (Kizilcec, R. F, 2017) (Pérez-Sanagustín, M., 2017) (Maldonado, J. J. 2017). Self regulated learning (SRL) is a term that has been stated by narrated researchers. According to them, students having strong SRL are good enough in planning, managing and controlling as compared to the students having weak SRL. In this regard, MOOCs have been providing support for the learners with the help of different levels of SRL.

c) Student attrition and retention

With the passage of time, growth of private educational institutions has been increased up to the remarkable extend. These institutions have become source of higher learning and business entity. Therefore, maximum number of student's enrollment is its lifeline. For the survival of private institutions, profitability, proper management and alignment are mandatory. In this respect, student retention until the completion of degree is quite necessary. That's why

institutions are finding that factor that ultimately causes student attrition (Azarcon Jr et al, 2014). After analyzing those factors, it is important for educational institutions to make strategic adjustments accordingly to improve student retention in institutions. In Gaviria, Colombia, people are closely belongs to social mobility that is a causing the academic performance of the students and student attrition rate may increase in educational institutions. Researchers have used three drivers the student performance analytics. The first driver is the volume of data that is being collected from learning management system and student information system, second one is the e-learning and third one is political concerns (Guarín, C. E. L., 2015)(Guzmán, E. L.,2015)(González, F. A. 2015). The application of data mining in the field of educational data mining has been emerged in different areas and researchers are exploiting these areas with various dimensions.

There is another supplement for student learning through web that is known as MOOCs. In online course learning and management system, high rate of dropout of students have been identified by researchers. This problem has been enlightened by the YANG in his research that is students are dropping at considerable level in Massive Open Online Courses (MOOCs). To control the student attrition problem from the coarser classes, researchers have proposed a model in their research. This model is a helping hand to determine such influential factors that are causing student drop out. So the predictors have been tried to propose in this research that will determine the factors related to student's behavior and social position in the discussion in which they participate in the forums (Yang, D.et al, 2013).

The problem of student attrition and retention is not new for the educational institutions. It has been enlightened by the researchers from the fields of data mining and information visualization. Now it has become very common research problem for the researchers. Student attrition and retention problem has been observed by the researchers when this problem was raised up to the ratio of 50% on the colleges of Ontario (Drea, C. 2004). To reduce attrition rates, institutions should focus on student retention. Researchers have analyzed the factors that causes student attrition and in the research, Drea has addressed

both elements; student attrition and retention. To retain the persistence of institutions, two theories have been formulated in this context. The first one is student integration model by Tinto and student attrition model by Bean. Both models work almost similar (Cabrera, A. F. 1993).

There are numerous reasons for what student attrition has become research problem for the researchers. In those reasons, personal disappointments, financial setbacks, and lowering of career and life goals are considerable. Therefore scientists have carried out the retention and persistent in their research to resolve this societal problem (Ramist, L.1981). This research focus on student performance and also find a teacher methodology has positive impact on student grade prediction and has reduced student retention rate.

2.2 Student performance prediction

Student performance prediction is the research problem which has been considered by many researchers from the field of educational data mining (Kabakchieva, D, 2013). With respect to performance prediction of students, there are many cases such as student's next term grade prediction, presence performance based on assessment in the current courses. To evaluate this problem, researchers have used some of the characteristics of students such as admission records,

High school scores, SAT/ACT scores and grades of previously completed courses (Elbadrawy, et al, 2016). More characteristics such as class test, seminar attendance and marks, assignment marks (Baradwaj, B. K., & Pal, S. 2012). For the sake of performance prediction, data that has been stored in open polytechnic student management system belong to 2006-2009 had been used. For the experiments to be applied over the dataset, feature selection and clustering techniques were applied on the dataset which are also known as data mining techniques.

In the field of education, predicting the performance of students has become considerably important for the university management and departments. Because the prestige of any educational institution is ultimately depends upon the performance of its students and if performance of maximum number of

students degrade up to the mark able level, it should be considered by the department management and find out the factors that affects the performance of students. Under consideration of this problem, researchers have evaluated their studies against the parameters that affects the overall performance of students, data mining techniques and data mining tools (Kaur, G, 2016)(Singh, W, 2016). In such parameters, psychological, personal and environmental factors are involved. For the experiments, they have used Naïve Based and J48 techniques through WEKA.

a) Future grades prediction

To predict the future grades of students is the considerable research problem that has been identified by many researchers with the help of different approaches such as Matrix Factorization, classification and clustering algorithms (Polyzou, A., & Karypis, G, 2016). Before applying these approaches, students variables are supposed to be extracted from the collected dataset that are unique registration number, result of previous two terms, GPI (Grade Point Average), number of family members, family welfare, gender, usage of social networking and use of technology and internet (Oskouei, R. J., & Askari, M, 2014). After collection, data preprocessing is applied in which positive and negative effects of dataset are removed and then classification and clustering are applied. By using such techniques, data can be further divided into clusters and classes to analyze useful and hidden patterns from it. To predict the performance of students and predict their grades, researchers have used socio-demographic factors of the students such as age, gender, ethnicity, education, work, status, and disability. In addition, researchers have used environment may also affect the student's performance that causes their dropout or retention throughout the course accomplishment (Kovacic, Z. 2010).

b) Teacher recommender system

When students retain their degree the end of their duration, it encompasses the quality education of any educational institution that refers to the term of student retention. According to the national statistics in a research, fifty percent of the students from higher educational institutions get attired. To (Bozo, J., 2010)

(Alarcón, R., 2010) (Iribarra, S., 2010). As online course enrollment system has taken the place globally because of internet. In this variation in technology, there are several web pages and websites that are busy in providing best at their own but still contain some missing elements in their services. Therefore, researchers have worked upon this problem and proposed solution regardingly. To make every course learnable creatively, researchers has focused on teacher recommendation according to the course respectively. The new recommender system is known as A³. With the help of this recommender system, not only accurate teacher will be associated to the relative course contents but also best contents of the course will be available on that site after some time (Tewari, A. S, 2015).

c) Course recommender system

A course recommender system helps to determine the performance of students by recommending them courses up to their performance. For this purpose researchers have proposed the recommender in which he used 67 templates of active students. This recommender system not only helps to predict the grades of students, it also helps to recommend the courses to the students by considering their timetable (Bydžovská, H, 2013).

2.3 Commonly used approaches in EDM

From the literature, a lot of data mining approaches that have been carried out by the researchers to evaluate their studies. Most commonly used approaches have been discussed in detail in subsections of section.

a) Classification and Clustering

To improve the performance of weak students in the class, with the help of data mining techniques, collected data of 1100 students have been transformed in Weka. Researchers have used freeware software such as Weka, Clementine and Rapid-Miner in this work. Different classifiers such as Naïve Based, C4.5, Neural networks and random forest have been used. The classifiers which have been used are Adaboost, Bagging and boosting. It is found that some factors have same effects on both countries and some have different. Moreover, it has

been found that male students suffer with stress more as compared to female. The performance of male students in Mathematics and formal SCIENCE is better whereas performance of female students aroused better in Literature and mnemonic SCIENCE (Oskouei, R. J.,2014)(Askari, M. 2014).In this field of research, there exist wide varieties of benchmark which are used to evaluate the performance and accuracy of experiments conducted by using machine learning approaches. Different researchers have used various types of educational datasets and each dataset is unique in its attributes (Garcia-Saiz, D., 2011) (Zorrilla, M. E, 2011). Therefore, in the study researchers have proposed meta algorithm to preprocess dataset. Various data mining models have been studied in this research to find the most accurate one with the help of Meta algorithm.

Educational Data Mining approach is emerging in providing feasibility to the educational institutions to improve their teaching and learning methodologies. EDM is the process of applying data mining techniques in educational institutions. In educational institutions, the data of student's performance, teacher's evaluation, student's enrollment data, and gender differences are to be stored in the database. On this data, Classification, Clustering and association rules are applied. These rules have been applied on mandatory courses taught in IT department of King Saud University. To extend this approach, these rules can be applied on elective courses as well to predict the grades of students in all courses as well as final GPA. Furthermore, neural network and clustering can also make these rules effective for prediction (Al-Barrak, M. A., & Al-Razgan, M. 2016). Researchers have considered more attributes of students such as their attendance, class test, seminar attendance and marks, assignment marks and constructed classification tree in Weka. According to those researchers, teachers can minimize the failure ratio of students in the semester by adopting this approach and attributes (Baradwaj, B. K., & Pal, S. 2012).

b) Association rule mining

In the study, some factors which affects the grades of students of Iran and India has been observed. In such factors, their respective gender, family background, education level of their parents and their lifestyle has been encountered by

asking questions from them (Oskouei, R. J., & Askari, M, 2014). To evaluate the performance of students and improve the management of educational institutions, researchers have introduced pre university characteristics of students which are known as student's profile and place of secondary school, final secondary education score, total admission score, and score achieved that exams. In this research data of University of National and World Economy (Bulgaria) has been collected and data mining techniques has been applied. In such data mining techniques naïve bayes and bayes net, nearest neighbor algorithm, and two rule learners One R and JRIP has been applied on the dataset (Kabakchieva, D. 2013).

c) Pattern mining

As discussed earlier, it is now clear knowledge is gathered about online teaching and learning system. Universities through internet are now big source of student-teacher interaction and source of learning and training. This technology has made place in the field of educational data mining in which knowledge diffusion has become research icon for the researchers who belong to field of data mining, web mining and graph algorithms. The data of online teaching system is serving best to the researchers who keen to extract knowledge from big datasets. Researchers are analyzing patterns of online learning behavior of students and to draw outcomes from these sets of data, they have been using machine learning algorithms and various data mining approaches. In a study, researchers used 19,934 servers to identify the behavior of students in Taiwan who had registered online courses and with the help of this data, they drawn conclusion after predicting their performance too (Hung, J. L., & Zhang, K. 2008). In addition, data mining techniques were proven to be helpful for the course developers, online trainers, and instructional designers. In their study, they used WEKA and KNIME tools to perform analysis of descriptive and artificial intelligence. Moreover, for data visualization and statistical analysis, they used

SPSS. In another study, data mining techniques such as classification, clustering, relationship mining, prediction and social area networking have been applied in educational data. (Sachin, R. B., & Vijay, M. S. 2012). Well all

of the mentioned approaches are ultimately works to represent the results after applying over the data. The concept of data visualization has been derived from visual reasoning (Inoue, S et al, 2017).

d) Text mining

In the past decade, there are number of tools such as WEKA, RapidMiner, R, KEEL, and SNAPP that have been used to extract text (Useful information) from the datasets in the field of educational data mining (Baker, R. S., & Inventado, P. S. 2014). In other research, data mining is also known as knowledge discovery from databases (Anand, S. S. et al, 1996). In this process, database techniques are bind with mathematical and artificial intelligence techniques.

2.4 Commonly used attributes for student performance prediction.

While review of papers from the literature, many data mining approaches have been found that are applied in academic datasets of different educational institutions for various purposes. Such approaches have been applied on the attributes that are considered after analyzing the factors. Those factors are briefly discussed in following passage.

I. Demographic attributes

From the literature, the importance of demographic attributes upon student's performance is inaugurated. According to the researchers, they condemn that living standard of any student ultimately affect their studies and variation their performance. Demographic factors include Age, gender, Financial status, Balance due, Permanent address, Residential address, Guardian (Brother, uncle, Self), Father qualification, Father Occupation, Full/ Part time student. Demographic factors equally supports the researchers whether to find the student attrition or student's retention. Student's retention is an obvious success of educational institutions. In order to sustain from student retention and keep the performance of students up to the mark, educational institutions better need to recruit the limited students as per its availability. In a study, the student's

general weighted average, School's Radial Distance, and school ownership are taken (Abaya, S. A., 2013) (Gerardo, B. D., 2013)

II. Pre-university attributes

Another factor which has been found during literature survey is *pre-university attributes*. These attributes have quite great impact on student's performance in any educational institution. Pre-university attributes will be fruitful with respect to accommodate the student's interest to map in course recommendation system. Pre-university attributes include *Secondary School Grade, Higher Secondary Grade, SAT Score, Pre-college, Pre-board, pre-program*. Researchers have used the historical data of students as well to find the attrition rate of students from academia through cross validation process under the classification and naïve based methods Guarín, C. E. L., 2015)(Guzmán, E. L.,2015)(González, F. A. 2015).

III. Institutional attributes

Last factor that has been extracted from literature to improve the performance of students is institutional attributes. These attributes will support in modeling the teacher recommender system and future course prediction model in our study. In institutional factors, *Term 1 GPA, Term 2 GPA, Term 3 GPA, and Term 4 GPA, CGPA, Total Credit hours taken, Total courses taken and Initial major, current major, current enrollment status* are included.

Proposed attributes of students affects their performance in various ways. It not only helps in identifying the students with low academic performance but also paves ways to improve their grades and enables recommender systems of universities to predict the grades of students. With the help of this study, student carrier can be effected by means of building up and students can easily pave their way towards their desired destination. This study will also help to construct the course recommender system which will enable students to select their courses that are recommended by the proposed system. Another advantage of this study is highly beneficial economically for parents. With the help of proposed system, students may not waste their money in selection of courses to

repeat. Moreover, student's academic record can be improved and efficiency of department management can be raised.

2.5 Literature review summary

Table 2.1: Literature Review Summary

Reference	Relevant Techniques	Tools	Used Attributes		
			Demographic	Pre-university attributes	Institutional attributes
(Guarín, C. E. L., Guzmán, E. L., González, F. A. 2015)	Decision Trees, Bayesian Classification	10 fold cross validation model, Cost sensitive model		√	
(Abaya, S. A. & Gerardo, B. D., 2013)	Classification through C4.5	Irecruit Application in UNIX	√		
(Kovacic, Z. 2010).	Clustering, Feature selection thorough cross validation, and CART classification	CHAID Model, Gian chart	√		
Kabakchieva, D. (2013).	CRISP-DM, Neural networking, Nearest neighbor classifier, Rule learner, Decision tree classifier	WEKA		√	
(Al-Barrak, M. A., & Al-Razgan, M. 2016).	Data Visualization, Classification	E-learning Web Miner, WEKA			√
(Baradwaj, B. K., & Pal, S. 2012)	Classification and Decision trees	Manual			√
(Elbadrawy et al, 2016)	Matrix factorization based methods, regression based methods	Recommend er system based personal analytics			√
(Bydžovská, H. 2013).	EDM, SNA and Collaborative filtering	WEKA and R	√		
(Kaur, G., & Singh, W. 2016)	Naïve based and J48 Decision trees	WEKA	√		
(Kokina, J., Pachamanova, D., & Corbett, A., 2017).	Predictive Modeling, Data Visualization	Excel and Tableau			√

Chapter 3

RESEARCH METHODOLOGY

This chapter presents the methodology adopted to address the research questions presented in chapter 1. The focus of the research is accurate grade prediction of critical courses of first semester of BS CS students of CUST. Prediction will help to take appropriate measures to control the attrition rate and hence will be beneficial for students, parents and university. Targeting the same objective, appropriate faculty members are also being recommended based on the results of previous semester. The factors under consideration for the proposed research are classified into three categories; Demographic, Pre-university and Institutional. Such researches in the field of educational data mining have been conducted in foreign earlier. Experiments have been conducted in local context, Pakistan. For this purpose, all the data have been collected from Capital University of Science and Technology, Islamabad Pakistan. After data collection and its pre-processing, some data mining techniques are being selected like SVM, Linear regression and Non-linear regression in WEKA.

This study will not only be useful in improving the overall performance of students but also reduces the attrition rate. Now it is easy to do such things on the basis of early prediction. Field of data mining has been emerged with the linkage of natural language processing, artificial intelligence, visual data analytics, and social data analysis etc (Romero, C., & Ventura, S. 2013)(Zhang, Y.et al, 2010). This early prediction is computed by using term marks and mid-term marks of the students along with their pre-university; Intermediate/O-levels marks and Matriculation marks and demographic factors; gender and city. This methodology will be discussed briefly in the current chapter. The related work of this research task has been discussed in detail in chapter 2. Chapter 3 contains detailed discussion of proposed methodology along with the data flow diagram that has been given as 3.1

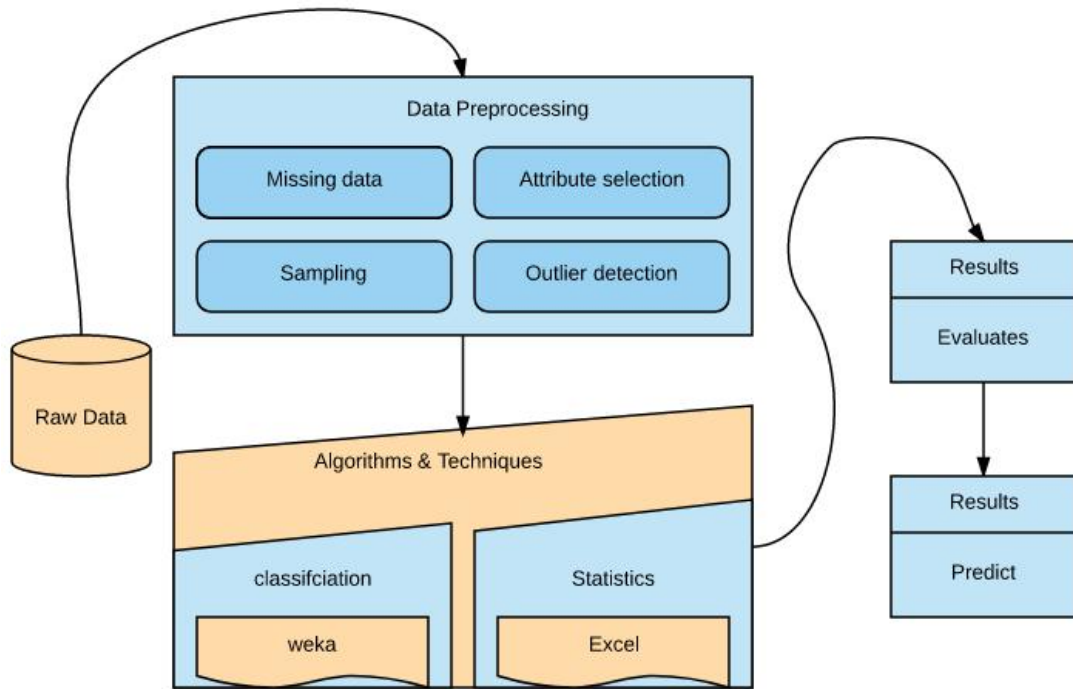




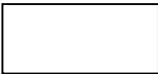


Figure 3.1: Data Flow Diagram of Proposed Methodology

Proposed data flow diagram is constructed on the basis of following set of symbols.

Table 3.1: Description about Data Flow Diagram

	<p>This cylindrical shape represents the raw data that has been collected for the research and need to be pre-processed.</p>
	<p>This filled arrow is used to represent the flow of tasks that has been carried out for the experiments.</p>
	<p>This symbol represents the working of pre-processing steps in our data flow diagram.</p>
	<p>This oval shaped rectangle represents the algorithms and techniques that are supposed to apply on the data set for the experiments.</p>
	<p>This shape represents the final results that include the result prediction and evaluation of results.</p>

To conduct the experiments, first of all the data is collected from the CUST (Capital University of Science and Technology), Islamabad Pakistan. The data set contains the data of six semesters that belongs to the BSCS program. On the basis adopted factors, the GPA of 1st semester as well as CGPA of 2nd semester has been predicted. In addition, performance of teachers against their respective courses has been measured through ANOVA analysis. The experiments have been conducted through WEKA. Each research question has been answered with the help of proposed methodology and experiments.

3.1 Data collection and Pre-processing

As discussed earlier, domain of computer science is being selected for the experiments and CUST has provided the facility to collect the data from its sources. The academia of CUST is semester based and there are two semesters per year. One semester corresponds to the spring semester which starts from February and ends in June. And the fall intake starts from mid of September and ends in the end of January. The data initially contained 5 courses of first semester, such as English-1, Physics, Calculus-1, Pakistan Studies & Islamiat and ITC/CP.

Table 3.2: Dataset Summary

Term	Registered Students
Spring 2014	74
Fall 2014	155
Spring 2015	115
Fall 2015	102
Spring 2016	62
Fall 2016	101
Spring 2017	58
Total	667

However, the information has been extracted from a research (Junaid. 2017), that most crucial courses with respect to attrition or students' performance are ITP/CP and Cal-I. So results of only these two semesters are considered and excluded the rest three.

Collected data is related to the students of first semesters and seven semesters starting from Spring 2014 (Term 141) till Spring 2017 (Term 171), and also of second semester students of six semesters starting from Fall 2014 (Term 143) to Spring 2017 (Term 173). The courses of first semester, as mentioned earlier, are Cal-I and ITC/CP, and the second semester courses are Cal-II and CP/OOP from each spring new intakes that belongs to the semester no 1. The initiative of this research is to identify the factors which contribute to keep the students retained till the end of the semesters as well those factors that affect the performance of the students. Following table 3.2 shows the information of the students comprehensively.

Table 3.2: Summary of Session-wise Attrition

Term	1st Sem	2nd Sem	3rd Sem	4th Sem	5th Sem	6th Sem	7th Sem	8th Sem
141	64%	20%	8%	8%	0%	0%	0%	0%
143	29%	35%	25%	12%	0%	0%	0%	0%
151	60%	30%	9%	0%	0%	0%	0%	0%
153	71%	25%	4%	0%	0%	0%	0%	0%
161	94%	6%	0%	0%	0%	0%	0%	0%
163	100%	0%	0%	0%	0%	0%	0%	0%
Attrition %	70%	19%	8%	3%	0%	0%	0%	0%

The dataset contains the institutional factors; Registration Number, Name, GPA, Internal marks like midterm, and term works marks etc and overall information of the students such as Student's registration number, GPA, Registered courses, Matriculation and Intermediate results. To conduct the experiments, we have used demographic and pre-university factors of the students apart from institutional factors. In the data set, the data of term 141, 143, 151, 153, 161, 163, and term 171 are included. The terms of the spring semesters are associated with the registration numbers of the students such as the term 141 means spring 2014. Similarly, the data of spring 2015 and spring 2016 are collected and on the basis of three years record of the students. The file of dataset is initially imported into WEKA after converted into CSV and ARFF format for the pre-processing and further experimentations have been conducted respectively.

Missing Data Handling

The problem of missing data generally arises due to absence of data in an observation for any variable during experiments. Missing data problem also arises when no information is provided or unavailable for the variables. The reason of missing data may involve the problem in collection of data by the researchers or the sources have not provided it completely³.

Missing data has been handled by two ways. Firstly, null values of courses have been found from other semesters and may be read in late semester in which students have passed that course. In case of consistent absence of values, the grades of such course have been supposed by taking the average result of other courses in respective semester. The missing data can be handled through different filters in WEKA.

³ <https://measuringu.com/handle-missing-data/>

Noisy Data Handling

While data collection, irrelevant data is called as noise. As the data of BSCS students of spring 2014 from spring 2016 is collected, therefore during the collection, data belongs to Software engineering and bioinformatics was present and then eliminated through filters in WEKA. Noisy data occurred in the form of errors such as; GPA = “-0.4” and Intermediate or Matriculation marks = “-455, 544.8” which were handled in Pre-processing of data.

Attribute Selection

After missing data handling and outlier detection phase, final attributes have been selected on which overall experiments and results depend. Such attributes have been narrated in table 3.1 that belongs to all three types of factors; demographic, pre-university, and institutional. One more factor has been considered to answer our research question that is teacher performance. After completion of data pre-processing, classification algorithms have been applied. Attributes are then finalized for the experiments that have been evaluated to answer the research questions. In the current section, we have discussed adopted classifiers in detail.

3.2 Classification

Before applying algorithms, factors have been grouped that are earlier used by foreign researchers. These factors have been compared against the factors used by the researchers in local context (Pakistan). Moreover, we have expanded our research work by making the variations of different combinations attributes and then compared. The comprehensive detail about these factors has been mentioned in table 3.4.

Table 3.3: Selection of Attributes

Sr.No.	Factors	Attribute Name
1.	DEMOGRAPHIC	Gender
2.		Age
3.		Residence
4.		Location
5.		Race
6.		Father's qualification
7.		Father's occupation
8	PRE-COLLEGE	Secondary school grade
9		Higher secondary grade
10		Pre-college
11		Pre-program
12		SAT score
13	INSTITUTIONAL	GPA (Term1, Term2,Term3, Term4)
14		CGPA
15		Financial status
16		Total credit hours taken
17		Total courses taken
18		Initial Major
19		Current Major
20		Current enrollment status
21		Teacher methodology
22	TEAHCER	Instructor grading
23		Instructor feedback
24		Instructor teaching methodology
25		Teacher name

3.3 Algorithms and Techniques

To evaluate our research by using different classification algorithms; Naïve Based and J48 has been compared. These classifiers are chosen on the basis of different reasons i-e; these classifiers supports Categorical and Nominal. For this purpose WEKA 3.8 tool has been used. The tool is open source available on the web and is generally used for machine learning algorithms, classification, training and testing of data. The data is then can be visualize in the form of graphs as well. In present research, the formed combinations of attributes have been evaluated to answer our research questions. For this purpose the data has been converted into csv format to import in the WEKA and the results from various filters and classifiers have been accumulated.

a) Naïve Based

Naïve based algorithm is comparatively fast algorithm in terms of classification. It works faster on huge datasets by using Bayes algorithm of probability. Bayes algorithm generally used to predict the class of unknown dataset⁴. Naïve based algorithm works on assumptions to label an item whose features are known but name is unknown. For example; a fruit is labeled as an apple if it is round and red in color and its size is 3 inches in diameter. These features of apple will raise the probability of this fruit that it is an apple.

b) J48

J48 decision tree is used to predict the target variable of new dataset. If dataset contains predictors or independent variables and set of target or dependent variables, then this algorithm is applied to extract the target variable of new dataset⁵.

c) Linear Regression

Generally linear regression classification algorithm is an approach to identify the relationship between dependent and independent variable. It is generally used for predictive analysis and has two main points. One is to check whether predictor variable does a good job in predicting the expected outcome variable.

⁴ <https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/>

⁵ <http://data-mining.business-intelligence.uoc.edu/home/j48-decision-tree>

Second main thing that linear regression does is the identification of variable that are significant predictors of dependent variables. At the end, the regression equation is used which helps to determine the set of predictor which are used to predict the outcome. In this research, algorithms are being used to compare the trend and pattern of the factors with other approaches like non-linear regression and SMO.

3.4 Evaluation of Research Questions

For our research, we have proposed three research questions that have been evaluated in multiple dimensions. Such dimensions have been explored in following ways respectively.

RQ1: Whether exiting identified factors for grade prediction are valid in our local context (Pakistan)?

Our research question number 1 determines the validity of factors in local context which have been carried out in foreign context by the researchers. The main reason that distinguishes this research in local context from the foreign context is that the norms, traditions and culture of every country that varies. Therefore it is needed to be evaluated that whether the existing factors that work in the colleges and universities of Canada, USA, England, and Nigeria etc, are applicable in Pakistan or not. After pre-processing of whole data, classification algorithms have been applied to analyze the behavior of demographic, pre-university and institutional factors. According to the results, it has come to known that these factors are also valid in Pakistan because of nearest difference in accuracies. Detailed results have been discussed in chapter no 4.

For our research, data being collected from term 141 to term 171 which consist of 667 students. Then three types of exiting attributes to predict the performance of students have been taken. First type of attribute is demographic; Gender, Age, Residence, Location, Race, Father's Qualification and have considered only Gender and Age of the student. Second type of attribute is Institutional; GPA(Term1, Term2, Term3, Term4, CGPA, Financial status, Total credit hours taken, Total courses taken, Initial Major, Current

Major, Current enrollment status, Teacher methodology and Midterm was considered, Term Marks for the prediction of final CGPA.

Third type of attribute is Pre-qualification; Secondary school grade, higher secondary grade, Pre-college, Pre-program, SAT score and Marks of FSC and MATRIC are considered. Among these attributes, specific attributes

During pre-processing, it has been found the total number of students who could not continue their program and left the course incomplete. After pre-processing, there were total 660 remained and 7 out of 667 who left the course. Two types of errors occurred in this dataset. The first was missing data and second was incomplete data which has been corrected in the phase of pre-processing.

In Table 3.5, Attributes selection to dataset for experiments has been shown

Table 3.5: Attributes selection to Dataset

Sr. No.	Gender	City	Matric	F.Sc	Mid ITC	Mid Eng-1	Mid Cal-1	Mid Physics	Studies	GPA
1	Male	Rawalpindi	51	64	10	10	16	9	16	2.51
2	Male	Islamabad	61	56	9	9	8	8	8	2.22
3	Male	Rawalpindi	51	64	7	7	13	17	13	3.15
4	Male	Islamabad	48	55	9	9	12	7	12	3.26
5	Male	Islamabad	77	58	12	12	17	18	17	3.07
6	Male	Islamabad	74	60	9	9	16	18	16	2.74
7	Male	Rawalpindi	84	70	6	6	12	7	12	2.86
8	Male	Islamabad	64	51	7	7	12	6	12	1.9
9	Female	Islamabad	61	56	8	8	11	6	11	1.83

RQ2: Which factors help in accurate prediction of student’s GPA of first semester?

Some specific attributes form all three factors; demographic, institutional and pre-university have been selected. The combinations of those attributes are constructed and then filters have been applied respectively. There are total 21 unique combinations of attributes presented in table 3.6.

Table 3.6: Pattern of Combination of Attributes

Sr. No.	ATTRIBUTES
1	Demographic, Pre-Qualification, Institutional (Result all subject of 1st semester)
2	Demographic, Pre-Qualification, Institutional (Result Eng-1, Cal-1, ITC, subject of 1st semester)
3	Demographic, Pre-Qualification, Institutional (Result Eng-1, Cal-1 subject of 1st semester)
4	Demographic, Pre-Qualification, Institutional (Result Eng-1 subject of 1st semester)
5	Demographic, Pre-Qualification, Institutional (Cal-1 subject of 1st semester)
6	Demographic, Pre-Qualification, Institutional (Result ITC, subject of 1st semester)
7	Demographic, Pre-Qualification, Institutional (Result Cal-1, ITC subject of 1st semester)
8	Demographic, Pre-Qualification, Institutional (Result Eng-1, ITC subject of 1st semester)
9	Demographic, Pre-Qualification, Institutional(GPA)
10	Demographic, Pre-Qualification
11	Demographic
12	Pre-Qualification
13	Pre-Qualification, Institutional (Result all subject of 1st semester)
14	Demographic, Institutional (Result all subject of 1st semester)
15	Institutional (Result Eng-1, Cal-1, ITC, subject of 1st semester)
16	Institutional (Result Eng-1 subject of 1st semester)
17	Institutional (Result Cal-1 subject of 1st semester)
18	Institutional (Result ITC, subject of 1st semester)
19	Institutional (Result Eng-1, Cal-1 subject of 1st semester)
20	Institutional (Result Cal-1, ITC subject of 1st semester)
21	Institutional (Result Eng-1, ITC subject of 1st semester)

RQ3. Is it possible to improve the performance of a student by allocating of appropriate teacher for a subject?

There is a fact that is normally faced during the research period is; the personality of teacher affects the performance of the students. If the background of the teacher is already known with the help of prediction system, it might become feasible for the department to recommend the best suitable teacher to the student up to his level of interest. And this phenomenon can surely boost the performance of the student. To answer the question number 3, experiments have been performed using two approaches have been used. And these approaches Statistical experiments and Predictive experiments.

a) The research has used combination of techniques comprising simple mean and the prediction based on teacher. The focus of this question is to find the most suitable teachers for Cal-I and ITP as these two subjects are considered the first semester with respect to the academic performance or attrition rate. Following calculations have been performed for the Cal-I teachers with respect to the prediction experiments:

1. Average of prequalification is taken
2. Average of actual Cal-I marks based on the Cal-I teacher
3. Average of actual Cal-II marks based on the Cal-I teacher
4. Prediction of the Cal-I and ITC/ITP is taken on the basis of prequalification
5. Predict Cal-II marks on the basis Cal-I and average is taken

b) ANOVA data analysis

ANOVA is a statistical technique which is used to measure the difference in a scale level dependent variable by a nominal variable which comprises of two or more types of categories. Normally ANOVA test is applied to find out the significant difference among groups. Similarly we will be applying this test to find the difference between the performances of teacher's subject wise in our

case of experiments⁶.

c) **Z-score data analysis**

Unlike ANOVA (analysis of variance), Z-score is usually applied on three or more means. A Z-score is a type of hypothesis test which is a way to find whether the results obtained from a test are valid or need to be repeated. For example, if someone said they had found a new drug to cure the cancer, one would want to be sure it was probably true. Similarly, in our research, we have applied Z-test to compare the performance of teachers particularly with respect to their subject. Z – Test will exploit the likelihood that the obtained results are true or not. A Z-test is generally used when the data is approximately normally distributed in the form of pairs⁷.

⁶ <http://www.statisticssolutions.com/manova-analysis-anova/>

⁷ <http://www.statisticshowto.com/z-test/>

Chapter 4

RESULTS & EVALUATIONS

Researchers have made their significant contributions to introduce such factors in the field of educational data mining that are quite beneficial for the educational institutions to mine hidden patterns from the academic database. Educational data mining is broadened field in which Online learning management system as well as physical institution's datasets are used. One limitation of this approach is to get the access of the dataset from the administration in order to perform analysis and conduct research. The dataset have been collected from Capital University of Science and Technology, Islamabad and applied existing factors on this dataset to answer our proposed research questions. Following sections of this chapter will present the overall picture of obtained outcomes in detail.

Data collection and pre-processing are the initial steps towards the analysis of the research. For this purpose, the demographic data has been collected from the registrar office of the CUST which contains student's gender and city. With the co-operation of administrative resources of CUST, the data related to pre-university and institutional factors has been collected. Institutional and pre-university factors contain the data about mid-term marks, term-marks, GPA, intermediate marks and matriculation mark. University portal provided privilege to access this data of the students to gather the sufficient dataset for our experiments.

The gathered data came into many formats such as PDF, word and excel. For the pre-processing, firstly this data set brought into the single file; Excel. Figure 4.1 represents the format of files and transformation of all files into single one. The collected data set arises in three formats; word, PDF and excel. After collection of dataset, it became mandatory to transfer the data into single file from all the files. The purpose of excel file is to import it into the WEKA to apply filters for the pre-processing. The final versions of dataset in the form of

excel is then converted into CSV format and imported in to the WEKA. Such filters have been applied that are discussed in subsections of section 4.1.

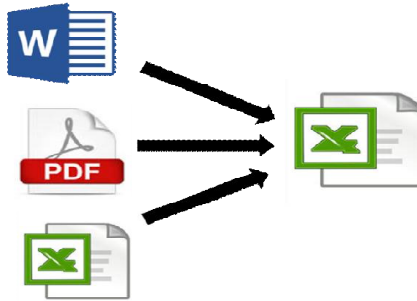


Figure 4.1: Formats of Data

In context of missing data, lacking attribute values, lacking certain attributes of interest, or containing only aggregate data instead of complete information are found. Incomplete information may lead the research towards in accurate outcomes that might affect the overall experiments. Therefore, the missing data problem has been handled in a way that no attribute remained null or incomplete.

Table 4.1: Occurrence of Missing Values

Sr. No.	Gender	ITC	Total	Weighted	Final Exam	Mid Term	Term Work	Teacher
1	Male	B-	75.95	76	25.2	15	21.67	D
2	Male	D	50.27	50	17.6	7.25	14.8	C
3	Male	C	66.93	67	22.8	8.25	21.65	A
4	Male							
5	Male	B-	76.43	76	21.6	13.75	27.27	D
6	Male	D	50.19	50	14.4	10.25	10.95	B
7	Male	A	89.93	90	34.8	14.25	30.57	D

The methods of data mining behave differently in the way that they treat missing values. Normally, they ignore the missing values, or exempt those records which contain missing values or either replace missing values with the mean, or conclude missing values from existing values. Missing Values Replacement Policies include the number of strategies such as; ignore the records with missing values. Following example shows the occurrence of missing values in the dataset. After the handling of missing data by taking the average of last three records of students and filled the missing one, complete dataset is gained. Further, this dataset has been used for experiments to evaluate all research questions.

Table 4.2: Handling of Missing Values

Sr. No.	Gender	ITC	Total	Weighted	Final Exam	Mid Term	Term Work	Teacher
1	Male	B-	75.95	76	25.2	15	21.67	D
2	Male	D	50.27	50	17.6	7.25	14.8	C
3	Male	C	66.93	67	22.8	8.25	21.65	A
4	Male	C+	72.91	73	19.2	18	27.75	A
5	Male	B-	76.43	76	21.6	13.75	27.27	D
6	Male	D	50.19	50	14.4	10.25	10.95	B
7	Male	A	89.93	90	34.8	14.25	30.57	D

Outliers are an observation point that is found to be distant from other observation points. An outlier may occur due to the change in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. After handling noise from the dataset, following result is obtained.

Attribute Selection

Appropriate selection of attributes is very important task in the process of pre-processing. It ultimately affects the performance of results and may lead to inaccurate results, if not chosen carefully. That's why we have classified the attributes into two types; Nominal and Categorical. On these two types of attributes researcher have selected the attributes that best suit to the nature of these two types of attributes.

Sampling

Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points in order to identify patterns and trends in the larger data set being examined. For example: term 141,143,151,153,161,163 dataset has been used as a training data and term 171 is test data because prediction of result of corresponding students is required to be computed.

4.1 Final Attribute Selection

From the literature, there are plenty of attributes found and some of them have been stated in the following table. This table belongs to the table presented in chapter no 3 but used in the current chapter with the addition of column named as selected attributes. These attributes have been marked to distinguish from each other. Then final attributes have been acquired for the experiments to fulfill the requirements of our results.

4.2 Result of Evaluation of Research Questions

With the help of adopted algorithms for the experiments in WEKA, the proposed research questions have been tried to answer. These answers make this research methodology validate and complete. Classification and linear regression has been selected to measure the analysis. The results of experiments have been discussed in detail in this section for each research question.

RQ1: Whether existing identified factors for grade prediction are valid in our local context (Pakistan)?

As existing factors along with their attributes have been stated in above sections, we have drawn the combinations of local factors with the existing factors. We found a slight variation between both types

of attributes. The comprehensive results of such variation have been arranged as the answer to each of three questions. The format of attributes has been presented in table 4.3 which are used in experiments.

Table 4.3: Attributes table

Sr. No.	Gender	City	Matric	FSc	Mid ITC	Mid Eng-1	Mid Cal-1	Mid Physics	Mid Pak-Studies	GPA
1	Male	Rawalpindi	51	64	10	10	16	9	16	2.51
2	Male	Islamabad	61	56	9	9	8	8	8	2.22
3	Male	Rawalpindi	51	64	7	7	13	17	13	3.15
4	Male	Islamabad	48	55	9	9	12	7	12	3.26
5	Male	Islamabad	77	58	12	12	17	18	17	3.07
6	Male	Islamabad	74	60	9	9	16	18	16	2.74
7	Male	Rawalpindi	84	70	6	6	12	7	12	2.86
8	Male	Islamabad	64	51	7	7	12	6	12	1.9
9	Female	Islamabad	61	56	8	8	11	6	11	1.83

As an input we have imported the .CSV into the WEKA that contains information about each attribute against every student. The input data contains the data of students of BSCS department from term 141 to term 163. The

dataset of CUST (Capital University of Science and Technology) has been used for this purpose. Among those factors, attributes like **Pervious** Course Marks/Grade, GPA, SSG, HSSG, Gender, City has been used. Afterwards, we have applied well-renown classification algorithms i-e; Naïve Bayes and J48 on the dataset.

Table 4.4: Classification of attributes and algorithms

Reference	Attributes	Class Label	Naïve Bayes	J48
Data mining approach for predicting student performance (2012)	Student Demographic, Per-Qualification, 1st semester 5 Subjects Grade	CGPA Grade	69.12%	72.39%
	Student Demographic, High School Background, Scholarship, Social network interaction	CGPA Grade	76%	73%
Comparison of Classification Techniques for predicting the performance of Students Academic Environment(2014)	Midterm marks all subjects	GPA Grade	65.6%	80.98%
	Internal assessment, Extra-Curricular activities	GPA Grade	66%	73%
Predicting Student Performance: A Statistical and Data Mining Approach(2013)	Student Demographic, Per-Qualification,	CGPA Grade	49.90%	63.19%
	Student Demographic, High School Background	CGPA Grade	50%	65%

In table 4.4, we have summarized the attributes of each factor with corresponding to their accuracies. These attributes have been assigned with class label and the reference of each attribute has been given as well. During the literature review, we have found the accuracy of each attribute in their respective scientific research publication and compare the accuracy of our

results. The inferences show the slight variations which is quite considerable with the correspondence to our research. The existing attributes have been shown in black color in the summary table and attributes in red color are local attributes.

The summarized results have been shown in the figure 4.2 in which bars of different colors on X-axis shows the presence of attributes and y-axis shows the interval of accuracy.

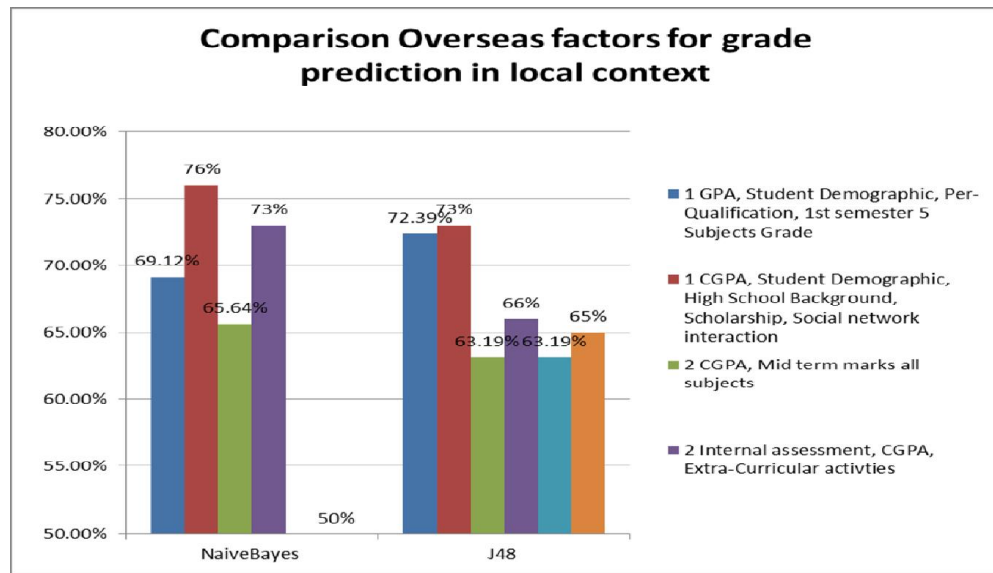


Figure 4.2 Comparison of local and Existing attributes

Our Findings narrate that existing identified factors for grade prediction are considered to be comparable with the local factors because of low variation in the result of their accuracies. According to the results, Pervious Grade, Internal Marks, HSSC were found to be effective for grade prediction.

RQ2. Which factors help in accurate prediction of student’s GPA of first semester?

Our second research question is the extraction of those factors that helps in accurate prediction of GPA of students belong to 1st semester. To answer the research question, we have transformed this question into following two ways.

1. Based on Mid term marks

In this phase of experiment, we have combined the attributes including marks of mid term to predict the GPA of the first semester. our prediction leads to the positive focal point where our proposed combinations worked well.

2. Based on Pervious Grade

In this phase of experiment, we have interpreted our attributes to predict the CGPA of the students. The CGPA of 2nd semester has been computed on the basis of previous academic record.

After compilation of attributes, their accuracies have been ranked that is presented in table 4.5 is descending order.

Information Gain Ranking Filter

Table 4.5: Ranked Attributes

Ranked Attributes	
0.2277	Cal-I MID
0.1903	Cal-II MID
0.1778	ITC MID
0.1684	Physics MID
0.1548	Pak-Studies MID
0.1514	FSC
0.1006	City
0.0992	Matric
0.0791	Gender

According to results, picture states the prediction of CAL-I MID and CAL-II MID remained better in terms of accuracy. Then the accuracies gradually decreases from ITC MID to GENDER. Based on the corresponding accuracies of attributes, we may inference the importance of acquired ranking information gain of all the instances.

1- First Semester GPA Prediction

We have expanded our answer for the research question number 2. With the help of “GPA Prediction for the first semester” we became able to evaluate our results through WEKA. We covered the list of attributes narrated in table 4.6 in which the combination of attributes are defined in column 1. There are four combinations of attributes which belong to the three factors; Pre-Qualification, Demographic and Institutional. The performance of J48 is considerably better than Naïve Bayes that has been shown in table as well as presented in figure 4.3.

Table 4.6: GPA Prediction for the first semester

Attributes	Class Label	Naïve Bayes	J48
Pre-Qualification, Institutional (Internal marks only Midterm) all subject of 1st semester	GPA Grade	64.62%	81.39%
Institutional (Internal marks only Midterm all subject of 1st semester)	GPA Grade	65.64%	80.98%
Demographic, Pre-Qualification, Institutional (Internal marks only Midterm all subject of 1st semester)	GPA Grade	66.46%	79.55%
Demographic, Pre-Qualification, Institutional (Internal marks only Midterm Eng-1, Cal-1 ,ITC subject of 1st semester)	GPA Grade	69.33%	78.12%

According to the results, the performance of Demographic, Pre-Qualification, Institutional (Internal marks only Midterm Eng-1, Cal-1, ITC subject of 1st semester) remained better as compared to the other three combinations of factors when Naïve Bayes classifier applied. On applying J48 classifier, the performance of Pre-Qualification, Institutional (Internal marks only Midterm) all subjects of 1st semester remained better comparatively to predict the grades of first semester.

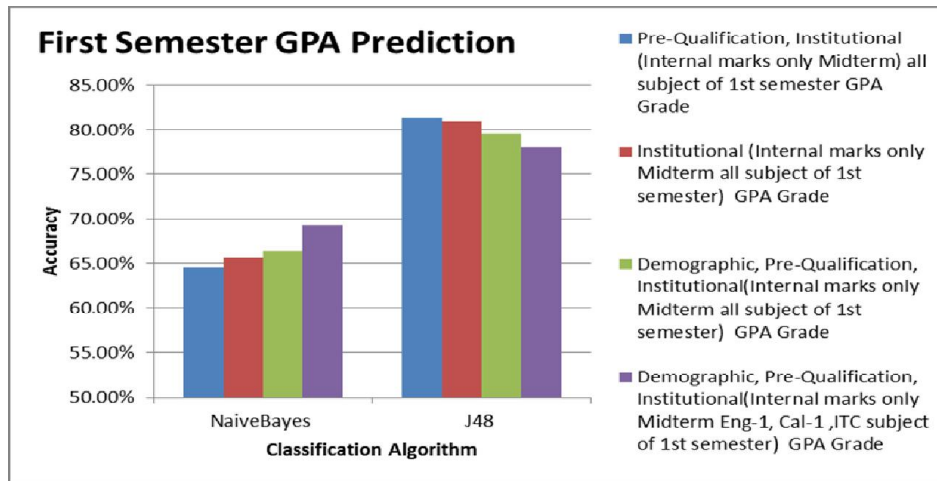


Figure 4.3: GPA Prediction for the 1st semester

2- CGPA of second Semester

To predict the CGPA of 2nd semester, we have defined the following narrated set of experiments as summarized in table 4.7. We have merged all the grades of previous semester along with the adopted factors. These combinations and their accuracies on the basis of Naïve Bayes and J48 have been presented in table 4.4.

Table 4.7: CGPA Prediction for the second semester

Attributes	Class Label	Naïve Bayes	J48
Demographic, Pre-Qualification, Institutional (Subject Grade Cal-1 subject of 1st semester)	CGPA Grade	67.08%	74.44%
Demographic, Pre-Qualification, Institutional (Subject Grade ITC subject of 1st semester)	CGPA Grade	68.10%	74.44%
Demographic, Pre-Qualification, Institutional (Subject Grade Cal-1, ITC subject of 1st semester)	CGPA Grade	70.55%	73.82%
Demographic, Pre-Qualification, Institutional (Subject Grade Eng-1, Cal-1 subject of 1st semester)	CGPA Grade	65.44%	72.60%

These tabular results have been presented as figure 4.4. Like 1st semester GPA prediction, we have used same classifiers for the CGPA prediction for 2nd semester. The only change lies in the combinations of attributes. According to the accuracy of each combination, we reached to the point that, the accuracy of Demographic, Pre-Qualification, Institutional (Subject Grade Cal-1, ITC subject of 1st semester) was high whereas when J48 classifier was applied, the prediction of Demographic, Pre-Qualification, Institutional(Subject Grade Cal-1 subject of 1st semester) and Demographic, Pre-Qualification, Institutional(Subject Grade ITC subject of 1st semester) remained better than other two combinations with the minor difference i-e; Demographic, Pre-Qualification, Institutional (Subject Grade Cal-1, ITC subject of 1st semester) and Demographic, Pre-Qualification, Institutional (Subject Grade Eng-1, Cal-1 subject of 1st semester).

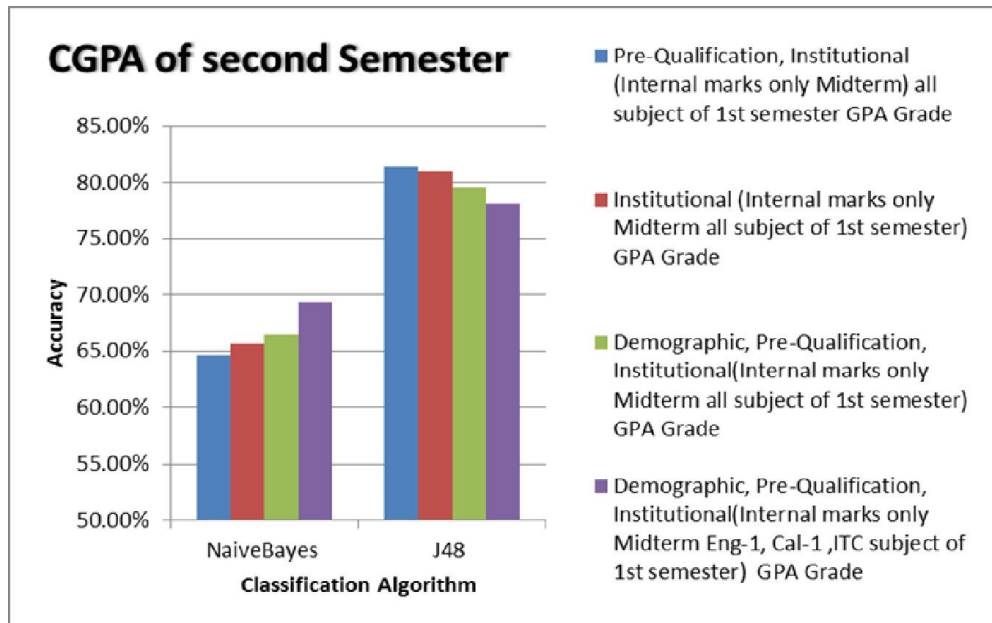


Figure 4.4: CGPA Prediction for the second semester

Our Findings enlighten the following assumptions.

- For Grade Predictions, the Attribute Evaluator has computed the Information Gain and ranked the accuracy of each attribute accordingly.

Then it is found that the subjects of First semester i-e; Cal-1, Eng-1, ITC have high value of information gain and played an effective role in order to predict the accurate grades of students.

- As we came to know from the situation of dataset that the Attrition rate is high in first semester. The purpose of the research is to reduce the attrition rate, so we have considered these subjects to predict the attrition rate of students as well as student's performance.
- Results states that Demographic attributes does not play significant role in grade prediction of students.
- In addition, internal marks like Mid Terms are used for early grade prediction that has high accuracy and also helps to minimize the attrition rate.

RQ3. Is it possible to improve the performance of a student by allocating of appropriate teacher for a subject?

The RQ3 has been addressed with a combination of techniques comprising simple mean and the prediction based on teacher. As it has been discussed in the precious chapter that the focus of this question is to find the most suitable teachers for Cal-I and ITP as these two subjects are considered to be critical for the first semester with respect to the academic performance or attrition rate. To evaluate the research question no 3, we have adopted the following set of experiments. Initially, the dataset has been imported into WEKA that contains the information of students from term 141 to term 163. The set of attributes consist of the Pervious Core and Elective Course Internal Marks, GPA, SSG, HSSG, Gender, City. After finalizing the set of attributes, we have applied following proposed techniques.

- **Predictive Approach:** Classification algorithm: Linear Regression
- **Statistical Approach:** Simple average , Z-Score analysis , ANOVA data analysis

The experiments have been conducted subject wise for the individual course grade prediction. On the basis of course grades, accumulated teacher

performance has been computed. Furthermore, following calculations have been performed for the Cal-I teachers:

1. Average of prequalification is taken
2. Average of actual Cal-I marks based on the Cal-I teacher
3. Average of actual Cal-II marks based on the Cal-I teacher
4. Prediction of the Cal-I is taken on the basis of prequalification
5. Predict Cal-II marks on the basis Cal-I and average is taken

Pre-Qualification vs Maths Courses

The comparison of Maths course and Pre-qualification has been made in this phase of experiment. The teachers have been labeled alphabetically and the overall average of the students has been given against every teacher in table 4.8.

Table 4.8: Pre-Qualification VS Maths Courses

Teacher	Pre-Qualification average	Cal-I average	Cal-II average
A	58.37142857	58.7429	64.4269
B	61.70231214	64.8324	68.03595
C	59.92207792	57.91	66.5098
D	60.17073171	56.1951	64.18938
E	58.96428571	73	72.75
F	58.95588235	57.2059	54.42167
G	60.73512748	61.8491	63.43205
H	59.82369942	61.3628	68.46664

The result shows that the performance of teacher “B, D and E” is found to be better in Pre-qualification combination. With respect to the Cal –I performance of teacher “B, G and H remained better. On using the Cal-II attribute, the performance of teacher “B, C, and D” found to be better than other teachers. These results have been shown graphically in figure 4.5.

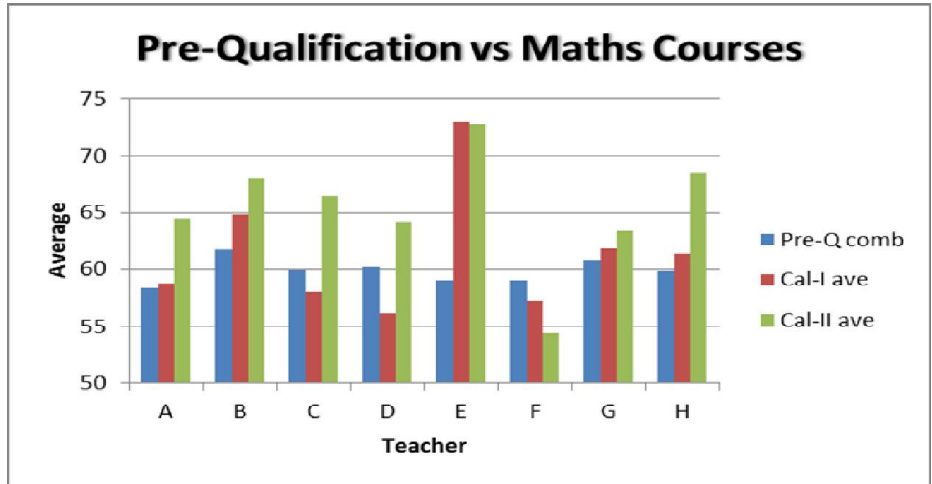


Figure 4.5: Pre-Qualification vs Maths Courses

Average of Cal-II Teachers

In this case, we have predicted the the performance of teahcers who teach Cal-II and compared with the actual performance of the teachers in Cal-II. According to the results, we have found significantly unexpected outcomes. There is a minor variation between actual Cal-II and predicted Cal-II. The results have been narrated in table 4.9 and presented in figure 4.6.

Table 4.9: Average of Cal-II teachers

Teacher	Cal-II ave	Pred-Cal-II ave
A	64.42689655	62.16017221
B	68.03594595	69.39823039
C	66.50980392	63.70893212
D	64.189375	64.42028569
E	72.75	71.255395
F	54.42166667	56.0806705
G	63.43205128	67.83774562
H	68.46663551	67.88178679

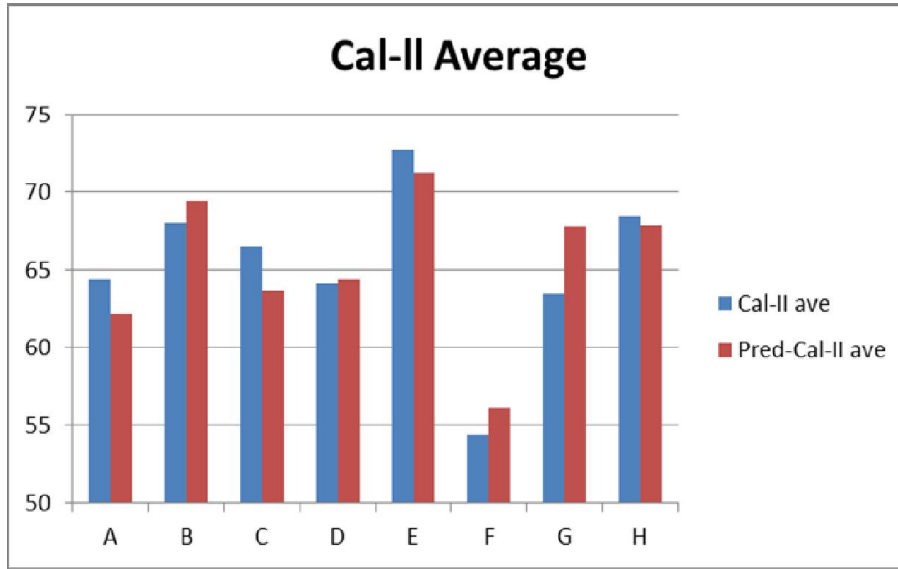


Figure 4.6: Average of Cal-II teachers

Average of Cal-I teachers

The average of Cal-I teachers has been compared with the results of actual and predicted values. There is another attribute “Pre-qualification” that has been used in this case. The obtained results have been given in the table 4.10 as well as presented in figure 4.7.

Table 4.10: Average of Cal-I teachers

Teacher	Pre-Q comb	Pred-C1-Preq	Cal-I ave
A	58.37142857	60.6696821	58.7429
B	61.70231214	61.50918169	64.8324
C	59.92207792	61.91539229	57.91
D	60.17073171	59.51882925	56.1951
E	58.96428571	60.379367	73
F	58.95588235	59.785871	57.2059
G	60.73512748	63.12452977	61.8491
H	59.82369942	61.70135392	61.3628

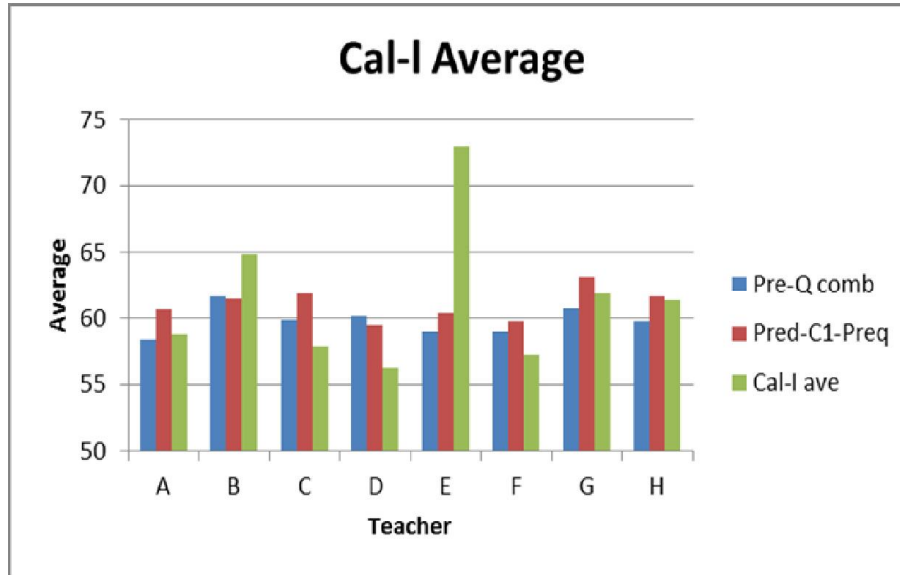


Figure 4.7: Average of Cal-I teachers

Average of Maths Teachers

Like cal-I and Cal-II the overall performance of math’s teachers have been computed against each teacher who teaches Cal-I and Cal-II. The information in table 4.11 narrates that the performance of teacher “B and E” remained better than other teachers who teach same subjects. This information has been presented in figure 4.8.

Table 4.11: Overall Average of Maths Teachers

Teacher	Over All Performance
A	60.87421589
B	65.09561403
C	61.99324125
D	60.89886433
E	67.26980954
F	57.2899981
G	63.39571083
H	63.84725513

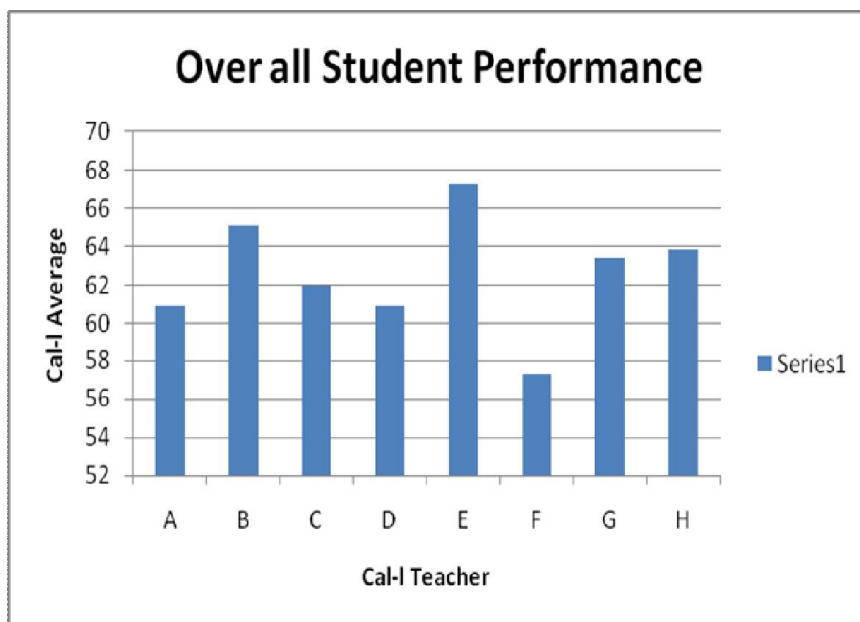


Figure 4.8: Overall Average of Maths Teachers

After the computation of results of Cal-1 and Cal-II, we became able to find the teacher “E” whose performance found to be better and able to recommend teaching Cal-I and Cal-II. Similarly, we will compute the results of other subjects which are to be taught in the 1st semester and then computed similarly. Firstly, researchers have computed the performance of every teacher of each subject which is being taught in first and second semester and gained the results. The results have been interpreted for every subject with respect to its related teacher. These results have been shown significant average of all students in Cal-1 & Cal-2 with Cal-1 teacher name. After the computation of results for Cal-1 and Cal-II, it became clear to find that the teacher “E” whose performance found to be better and able to recommend teaching Cal-I and Cal-II then teacher B have better performances and further teacher B and H performs better among all other teachers.

Further the average of these teachers will be checked whether it is significant or not significant by applying statistical test “ANOVA”.

ANOVA data analysis on Cal-I

ANOVA analysis is used for group data we have many teacher for Cal-1 & Cal-2 subjects. The prediction shows that teacher average performance is

significant. To prove these results, we applied ANOVA analysis on this dataset.

We have applied ANOVA (Analysis of Variance) test on teacher's average. By this way, results indicate the performance of all Cal-I teacher on the basis of group.

Table 4.12: ANOVA Analysis on Cal-1

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Row 1	5	303.5967	60.71935	15.95621		
Row 2	5	330.7894	66.15789	3.27359		
Row 3	5	313.1846	62.63691	19.02353		
Row 4	5	292.9438	58.58876	27.08998		
Row 5	5	357.9515	71.5903	2.338546		
Row 6	5	274.268	54.85359	20.50589		
Row 7	5	313.5665	62.7133	1.073095		
Row 8	5	318.981	63.79621	18.29478		
ANOVA						
Source of Variation	SS	Df	MS	F	P-value	F crit
Between Groups	872.7731	7	124.6819	9.273853	3.14E-06	2.312741
Within Groups	430.2225	32	13.44445			
Total	1302.996	39				

In this case $F > F \text{ crit}$ has been shown as $9.273853 > 2.312741$. Researchers reject the null hypothesis and concluded that there are significant differences between the methods because all 11 methods don't have the same mean.

Z-Score Data Analysis

Now Z-score test is applied to evaluate the performance of every pair of combination of teacher's subject wise. Generally, Z-test is applied on two populations to compare its proportion. In this case, we are supposed to compare the performance of teachers of the same course which is to recommend to the student. Researchers have classified each Z-test with respect to the subject and then have considered the results of Z-score with respect to Cal-1. All the teachers who teach Cal-I have been indicated with alphabetic characters i-e; A, B, C, D, E, and F. Pairs such as teacher A with other teachers of Cal-1 are made for experiment. Moreover, for the comparison, teacher's data has been classified as same number of students and different number of students.

Table 4.13: Two Samples for Mean with respect to teacher “E”

z-Test: Two Sample for Means		
	<i>Variable</i> <i>1</i>	<i>Variable</i> <i>2</i>
Mean	74.94286	58.74286
Known Variance	102.1429	333.391
Observations	15	35
Hypothesized Mean Difference	0	
Z	4.008258	
P(Z<=z) one-tail	3.06E-05	
z Critical one-tail	1.644854	
P(Z<=z) two-tail	6.12E-05	
z Critical two-tail	1.959964	

Z-score data Analysis applied on the Cal-1 teacher “E” with Teacher “B” on the basis different number of student. Value of Z-critical to tail < Z concluded that the performance of teacher “E” is not significant.

Then teacher “B” has been paired with teacher “A” on the basis same number of students. The value of Z-critical > Z concluded that significant performance

of teacher “A”. The results have been shown in table 19. Z-test has been applied in both cases and results have been compared either the performance of teachers remain same or vary. By this way, results indicate the performance of every teacher of Cal-I on the basis of combinations. The result of z-test will evolve around this statement “If $z < z$ critical two tail then performance difference is not significant”. If $z > z$ critical two tail then performance difference is significant.

Table 4.14: Two samples for means with respect to teacher “B”

z-Test: Two Sample for Means		
	<i>Variable</i> <i>1</i>	<i>Variable</i> <i>2</i>
Mean	74.94286	58.74286
Known Variance	102.1429	333.391
Observations	15	35
Hypothesized Mean Difference	0	
Z	4.008258	
P(Z<=z) one-tail	3.06E-05	
z Critical one-tail	1.644854	
P(Z<=z) two-tail	6.12E-05	
z Critical two-tail	1.959964	

Here, researchers have applied Z-score data Analysis on the Cal-1 teacher “B” with Teacher “A” on the basis different number of students who enrolled in different terms from term 141 to term 171. In this case, value of Z-critical two tails is greater than the value of Z which shows that performance of teacher “B” is not significant.

Then teacher “B” has been paired with teacher “A” on the basis same number of students and Z-test has been applied. The results shows that value of z-critical shows the significant performance of teacher “A” as it is greater than the value of Z.

Teacher B with different and same number of Students

Table 4.15: Teacher B with same number of Students

Cal-1 Teacher	Z	z Critical two-tail
A	4.008258	1.959964
C	3.370917	1.959964
D	5.288815	1.959964
F	4.463791	1.959964
G	4.119857	1.959964
H	2.997759	1.959964
E	3.630363	1.959964

Z-score data Analysis applied on the Cal-1 teacher “B” with all Teachers on the basis different number of student. The value of Z-critical tow tail $> Z$ conclude that performance of teacher “E” is not significant with Teacher “A”, “C”, “D”, “F”, “G”, “H”, “E”

Table 4.15: Teacher B with different number of Students

Cal-1 Teacher	Z	z Critical two-tail
A	3.687091	1.959964
C	1.235961	1.959964
D	3.480026	1.959964
F	2.829936	1.959964
G	3.218075	1.959964
H	1.61421	1.959964
E	1.271677	1.959964

Z-score data Analysis applied on the Cal-1 teacher “B” with all Teachers on the basis same number of student. The value of Z-critical tow tail $> Z$ conclude that performance of teacher “B” is not significant with Teacher “A”, “B”, “C”, “D”, “F”, and significant with Teacher “G”, “H”. The ANOVA test has been applied on the cal-I only. The results have been computed for the rest of the courses similarly.

Pre-Qualification vs Programming Courses

In table 4.17, we have narrated the results of all programming courses which are taught in 1st and 2nd semester. These averages of courses have been compared with the Pre-Qualification.

Table 4.17: Comparison of all programming courses

Teacher	Pre-Q comb	ITC /ITP Simple Average	OOP/CP Simple Average
A	57.170543	61.48148148	65.27777778
B	61.4620758	53.7037037	62.80851064
C	60.99346693	65.21428571	69.5
D	61.04028149	61.81578947	72.39215686
E	60.28896102	71.40625	67.91666667
F	59.72643733	77	68.58333333
G	59.72643733	66.1871345	66.73484848
H	61.887987	66.9375	59.07142857
I	59.69578266	64.16129032	68.7826087
J	59.48424926	62.55882353	70.68421053
K	58.0787338	67.575	63.72222222
L	60.18534971	67.44736842	68.3125

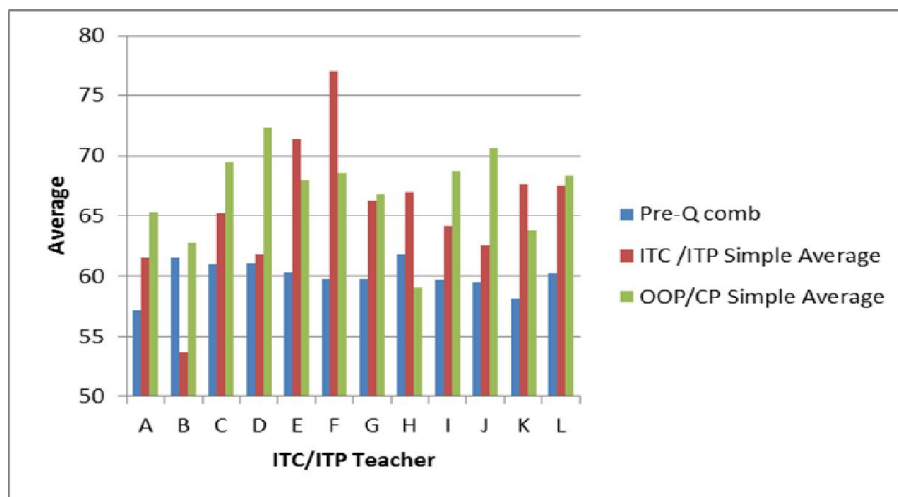


Figure 4.9: Comparison of ITC/ITP teacher

According to the figure 4.8, the performance of all of the teachers has been

computed. The performance of teachers with respect to courses average changes.

Average of ITC/ITP & CP/OOP Student

The average of students who studies ITC/ITP has been computed against every teacher who teaches these subjects. Then these results have been shown in table 4.18 and presented in figure 4.10.

Table 4.18: Comparison of Actual and Predicted Grades of ITC/ITP

Teacher	Pre-Qualification average	Predict ITC /ITP average	ITC /ITP Simple average
A	57.170543	60.90678396	61.48148148
B	61.4620758	63.21581404	53.7037037
C	60.99346693	61.1117056	65.21428571
D	61.04028149	68.16813274	61.81578947
E	60.28896102	70.68016263	71.40625
F	59.72643733	69.13777079	77
G	59.72643733	63.65621403	66.1871345
H	61.887987	72.60489569	66.9375
I	59.69578266	65.2340331	64.16129032
J	59.48424926	70.26083656	62.55882353
K	58.0787338	61.80150053	67.575
L	60.18534971	65.17485658	67.44736842

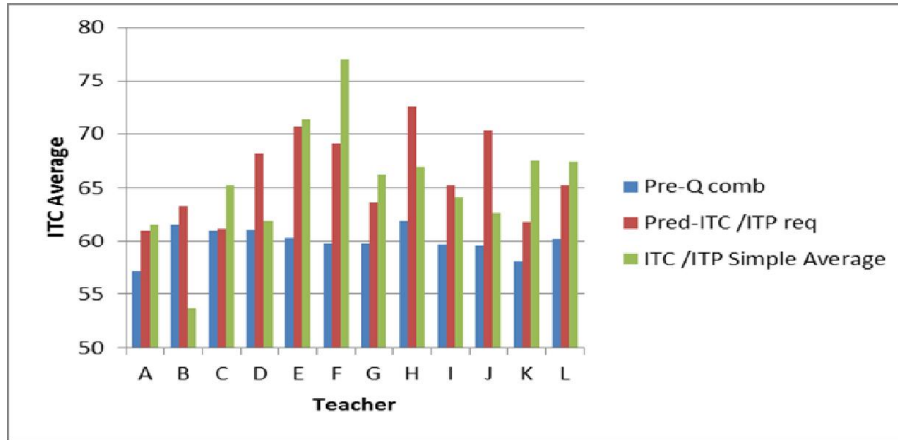


Figure 4.10: Comparison of Actual and Predicted Grades of ITC/ITP

Then the actual and predicted OOP/CP grades have been predicted against every teacher. The results shows that there is a slight variation in the results that has been stated in table 4.19 and presented in figure 4.11

Table 4.19: Comparison of Actual and Predicted Performance of OOP/CP Teachers

Teacher	OOP/CP ave	Pred- OOP ave
A	65.27777778	65.99605885
B	62.80851064	58.81069353
C	69.5	67.8661911
D	72.39215686	66.23364694
E	67.91666667	72.16535104
F	68.58333333	73.23220667
G	66.73484848	67.4231563
H	59.07142857	65.26328164
I	68.7826087	67.14102048
J	70.68421053	67.37969774
K	63.72222222	66.97798275
L	68.3125	66.77153434

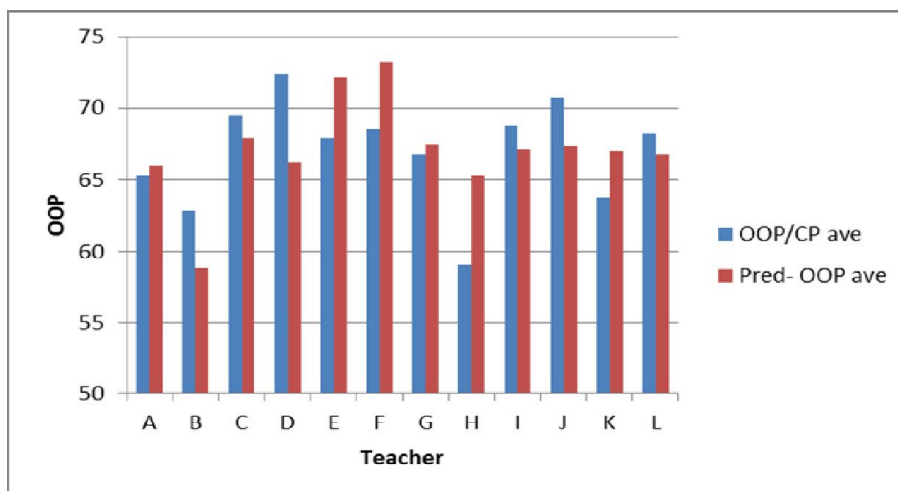


Figure 4.11: Comparison of Actual and Predicted Performance of OOP/CP Teachers

The overall performance of OOP teachers have been presented in table 4.20 as well in figure 4.12.

Table 4.20: Overall Performance of OOP Teachers

Teacher	Over All Performance
A	62.16653
B	60.00016
C	64.93713
D	65.93
E	68.49148
F	69.53595
G	64.74556
H	65.15302
I	65.00295
J	66.07356
K	63.63109
L	65.57832

The results show that the performance of teacher “E & F” is considerably high and better than other teachers. Therefore, based on the acquired results, we may recommend these two teachers to teach the courses in respective semesters.

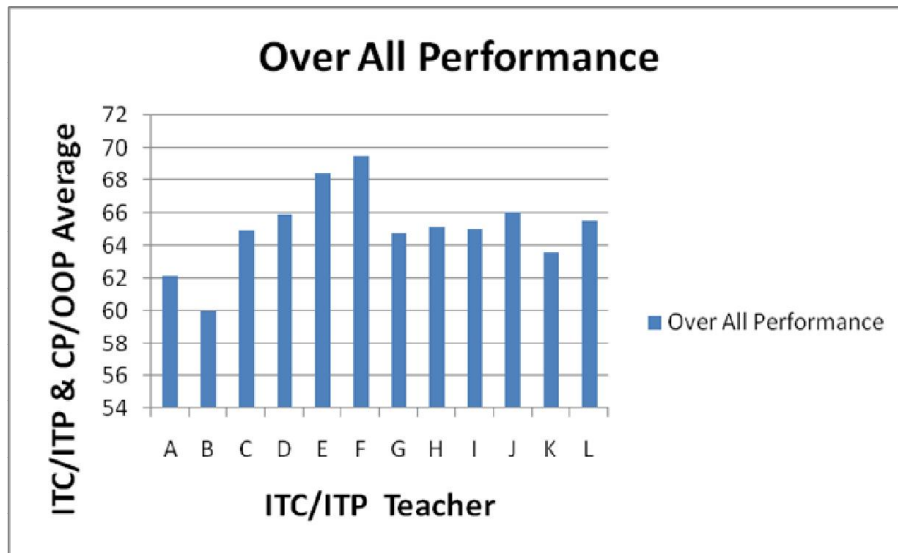


Figure 4.12: Overall performance of OOP Teachers

These results have been shown significant average of all students in ITC/ITP & CP/OOP with ITC/ITP teacher name. During the computation of results of ITC/ITP and CP/OOP, we have found teacher “F” performance better and able to recommend teaching ITC/ITP and CP/OOP.

ANOVA data analysis on ITC/ITP

ANOVs analysis is used for group data we have many teacher for Cal-1 & Cal-2 subjects. The average prediction shows that teacher average performance is significant. To prove this result applies ANOVA analysis on this dataset.

If $F > F_{crit}$, we reject the null hypothesis. This is the case $8.542703 > 1.952212$. Therefore, we reject the null hypothesis. The means of the 11 populations are not all equal. At least one of the means is different. Hence we conclude that there are significant differences between the methods (i.e. all 11 methods don't have the same mean). However, the ANOVA does not tell you where the difference lies. Therefore it is preferred T-Test, Z-Score test each pair of means.

Table 4.21: ANOVA Data Analysis on ITC/ITP

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Row 1	6	370.458	61.743	9.09211		
Row 2	6	346.3743	57.72906	16.59501		
Row 3	6	404.7156	67.4526	4.601526		
Row 4	6	397.287	66.21451	25.45386		
Row 5	6	421.2417	70.20696	3.492003		
Row 6	6	441.7835	73.63059	19.18218		
Row 7	6	400.4038	66.73397	0.377307		
Row 8	6	382.6283	63.77138	24.15625		
Row 9	6	394.865	65.81084	6.082654		
Row 10	6	392.0845	65.34741	12.9729		
Row 11	6	382.7968	63.79946	15.06479		
Row 12	6	407.9563	67.99271	0.222976		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1075.12	11	97.73818	8.542703	7.87E-09	1.952212
Within Groups	686.4678	60	11.44113			
Total	1761.588	71				

Z-Score data analysis on ITC/ITP

As discussed earlier, Z-test is a measure to analyze the performance of two attributes which have either equal or different number of population. Here, we

have applied Z-test on the combinations of teachers who teach ITC/ITP. We have made combinations similarly as we made for the teachers of Cal-1 and Cal-II. In table ..., Z-test has been applied on each pair of combination and tested on two cases. First is to compare the performance of teachers on same number of students and second is to compare the performance of teachers on different number of students. Z-score data Analysis ITC/ITP has been applied on the pair in which performance of teacher “F” with Teacher “A” on the basis different number of students has been compared. The results show that there is significant performance of teacher “F” with teacher “A”.

Table 4.22: Z-Score Data Analysis Teacher F with A different number of Students

z-Test: Two Sample for Means		
	<i>Variable 1</i>	<i>Variable 2</i>
Mean	77	61.48148
Known Variance	260.2985	278.7682
Observations	67	27
Hypothesized Mean Difference	0	
Z	4.116767	
P(Z<=z) one-tail	1.92E-05	
z Critical one-tail	1.644854	
P(Z<=z) two-tail	3.84E-05	
z Critical two-tail	1.959964	

Z-score data Analysis has been applied on the pair of ITC/ITP teacher “F” with Teacher “A” on the basis same number of students and results show that there is a significant performance between teacher “F” and teacher “A”.

Table 4.236: Z-Score Data Analysis Teacher F with A same number of Students

z-Test: Two Sample for Means		
	<i>Variable 1</i>	<i>Variable 2</i>
Mean	76.81481	61.48148
Known Variance	283.7064	278.7682
Observations	27	27
Hypothesized Mean Difference	0	
Z	3.359441	
P(Z<=z) one-tail	0.000391	
z Critical one-tail	1.644854	
P(Z<=z) two-tail	0.000781	
z Critical two-tail	1.959964	

As computed earlier, Z-score data Analysis ITC/ITP teacher “F” with other teachers who teaches same subjects on the basis different number of students has been computed. According to the summarized results, there is significant performance among teacher “F” with all of the teachers.

As computed earlier, Z-score data Analysis ITC/ITP teacher “F” with other teachers who teaches same subjects on the basis of same number of students has been computed. According to the summarized results, there is significant performance among teacher “F” with all of the teachers.

Table 4.24: Z-Score Data Analysis Teacher F with other different number of Students

Teacher F	Z	Z Critical two-tail
A	4.1168	1.96
B	7.0827	1.96
C	3.0686	1.96
D	5.4561	1.96
E	3.1293	1.96
G	4.7683	1.96
H	3.8317	1.96
I	3.8953	1.96
J	3.4076	1.96
K	3.3785	1.96
L	3.4421	1.96

Table 4.25: Z-Score Data Analysis Teacher F with other same number of Students

Teacher F	Z	z Critical two-tail
A	3.3594	1.96
B	5.6999	1.96
C	2.4668	1.96
D	5.2634	1.96
E	1.323	1.96
G	3.9718	1.96
H	2.4788	1.96
I	3.1042	1.96
J	2.8519	1.96
K	2.6697	1.96
L	2.3868	1.96

Summary

To prove our point, whether student's performance or grades are dependent on teacher's performance or not, we have taken an example to demonstrate. We have selected the course cal-1 to justify our analysis. For this task, we have taken courses; Cal-I & Cal-II and ITC/ITP & CP/OOP. Using different approaches for on the basis these results, we have ended with following assumptions.

There is no significant difference between actual and predicted average of student's grades on the basis of teacher's combinations we discussed.

When Cal-1 and ITC/ITP is taught by different teacher and get average marks of students from session 141 to 163 in Cal-1 and ITC/ITP teacher-wise using five different experiments, like : Average of actual Cal-I marks based on the Cal-I teacher, Prediction of the Cal-I marks excluding the teacher name and then taking the average teacher-wise, Prediction of the Cal-I marks including the teacher and then taking the average teacher-wise , Average of actual Cal-II marks based on the Cal-I teacher , Prediction of Cal-II marks and then computed the average of these marks based on the Cal-I teacher, our predicted results were higher than the actual results or nearly lesser than actual ones. The result has showed that all teachers have different average in these subjects and performances of teachers in this subject are significant.

For proved these result have used ANOVA analysis .The performance of teacher in CAL-I and ITC/ITP have different and average of teachers have significant between all teacher .The teacher whose performance is better might have a reason that he is lenient in giving grades to the students and the teachers whose performance is comparatively low might be strict in giving grades to the students. Z-Score have proved these results and have showed that significant result in CAL-I and ITC/ITP according to each teacher. With the help of these result, we have recommended for CAL-I teacher E is a very suitable teacher and ITC/ITP teacher F can improve the performance of students.

We have proposed a recommender system, which will recommend teachers to the students based on their interest level and understanding and communication

with their teachers. With the help of recommender system, university department may become able to identify the student's performance and enable them to make their performance better. To evaluate our prediction model and recommender system, we have produced the answers of research questions. We have found following strong observations during the course of our experiments.

- In context of CUST university Cal-I and ITC are critical subjects in semester-1
- Recommendation of appropriate teacher for Cal-I is B and G
- Recommendation of appropriate teacher for OOP is E and F

Chapter 5

CONCLUSION AND FUTURE WORK

- Prediction is a helpful activity to reduce the attrition
- Existing factor is applicable to the certain level of accuracy and a set of attributes is identified which is applicable in context of our university
- In context of CUST university Cal-I and ITC are critical subjects in semester-I
- Recommendation of appropriate teacher for Cal-I and ITC

Our proposed research has been distributed in five chapters. First chapter consist of introduction of the topic and purpose of this research. Second chapter contains literature review in which the related work of the other researchers has been presently in quite understandable manner. Third chapter contains methodology and chapter no 4 contains detailed discussion of results and experiments.

In fist chapter, we have briefly explained the background of our research topic along with its scope, significance and applications. Moreover, we have discussed research question, which we have constructed after critically reviewing of literature. We have explored various data mining techniques that were proposed and used by many researchers. The area of data mining is quite versatile. Data mining means to explore intrinsic information from the bulk of datasets. These datasets might belong to educational institutions or internet or any organization. With the help of mining techniques such as classification, we can make data analysis for prediction and find the accuracy. Researchers have worked a lot in the field of education in data mining in which they have used student's academic record for specific reasons. They have used the records for grade prediction, evaluation of student's performance and teacher's performance most commonly.

In this regard, from literature we have obtained exiting factors such as institutional factors and demographic factors. These factors are known as exiting because they have applied in the western universities. We have applied these factors in our local context. We have collected data from the Capital University of Science and Technology, Pakistan. We have used the data of BSCS program from six terms of spring, which covers the data of 3 years. In this dataset, we have used the data of students of first semester and their final grade has been predicted based on their midterm marks, term marks, gender, age, Matric grades and inter grade. These attributes cover demographic, institutional and pre-university attributes of the students. In addition, researchers have used only two factors; Demographic and Pre-university. In our experiments, we have considered another factor which is institutional in which midterm and term marks are combined to predict the students final GPA/Grade. This is how we have proposed prediction model on the basis students associated attributes.

After discussion of literature review critically, we have identified different classification approaches by the researches, so we used in our experimental work as well. Afterwards, we have discussed chapter number 3 in which diagram of proposed methodology have been mentioned and each step of methodology has been discussed in detail. Our chapter number 4 contains relative experiments and results that we have carried out through the help of WEKA tool. WEKA is specifically designed to perform analysis in the field of educational data mining. In the part of conclusion we have summarize the results against our each research question.

We have proposed a recommender system, which will recommend teachers to the students based on their interest level and understanding and communication with their teachers. With the help of recommender system, university department may become able to identify the student's performance and enable them to make their performance better. To evaluate our prediction model and recommender system, we have produced the answers of our three research questions. Those research questions and their answers have been summarized as follows:

RQ1. Whether existing identified factors for grade prediction are valid in our local context (Pakistan)?

The answer to this research question is yes because we have gained quite considerable accuracy from all data analysis measures. These analyses have been applied on categorical and nominal data. Results in both categories are same, which shows the validity of, existing factors in local context. All three factors contributed well in order to increase the accuracy through WEKA tool when ARPP format of data was given to it and classification algorithms; SMO and Linear regression were applied.

RQ2. Which factors help in accurate prediction of student's GPA of first semester?

We have given answer to this research question in quite interesting fashion. We have make combinations of all attributes that belong to these three categories of factors. When we combined term marks and mid-term grades along with inter and matriculation marks as well as with their gender and age; accuracy of algorithms were boost up to the mark able level. Therefore the answer to this research question is that demographic, pre-university and institutional factors (GPA, Internal marks, #of Credits Hours, Program). Based on these results, we have proposed a prediction model that will help to predict the final GPA/Grade of the students and then can be evaluated in other departments of the university too. There is not a single attribute that is affecting the performance of students but a combination of attributes that also occurs in terms of hybrid approach.

RQ3. Is it possible to improve the performance of a student by allocating of appropriate teacher for a subject?

To answer the research question no 3, we have drawn inferences by two ways. One way is statistical experiments and other way is prediction experiments. Both types of experiments have sub experiments associated with them. by statistical experiments, we have computed ANOVA test on the group of attributes to extract the information of such teachers having significant performance. But ANOVA has a limitation; as it is applied in group of attributes; therefore we are unable to identify the teacher having significant performance. To overcome this shortcoming of inferences, we have applied Z-test through which we have obtained the teachers having significant

performance with the pairing of each teacher belong to that particular course. By this way, we became able to bring out the teachers whose performance is better than other teachers who teach same course. Then we computed overall average performance of the teacher based on the overall average grade of students whom they taught.

With the help of prediction experiments, we have drawn prediction of grades inclusively and exclusively teacher's name who taught the courses. Then predicted results and actual results have been computed and compared. By this way we became able to get the entire set of results in which our analysis shows that predicted and actual results remained close at some point and better at some point too. In the same time, predicted were bit lower than actual ones as well which can be seen in chapter 4.

At the end conclusion and future work has been discussed.

5.1 Future Work

- The experimental work can be applied in other departments of the university.
- This data set is relatively small. We can carry out same experiments in other field of department and large datasets.

Moreover, these experiments can be carried out in government institution such as NADRA.

REFERENCES

- Abaya, S. A., & Gerardo, B. D. (2013, September). An education data mining tool for marketing based on C4. 5 classification technique. In e-Learning and e-Technologies in Education (ICEEE), 2013 Second International Conference on (pp. 289-293). IEEE.
- Aher, S. B., & Lobo, L. M. R. J. (2012). A comparative study of association rule algorithms for course recommender system in e-learning. *International Journal of Computer Applications*, 39(1), 48-52.
- Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting student's final GPA using decision trees: a case study. *International Journal of Information and Education Technology*, 6(7), 528.
- Anand, S. S., Bell, D. A., & Hughes, J. G. (1996). EDM: A general framework for data mining based on evidence theory. *Data & Knowledge Engineering*, 18(3), 189-223.
- Azarcon Jr, D. E., Gallardo, C. D., Anacin, C. G., & Velasco, E. (2014). Attrition and retention in higher education institution: A conjoint analysis of consumer behavior in higher education. *Asia Pacific Journal of Education, Arts and SCIENCE*, 1(5), 107-118.
- Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *learning analytics* (pp. 61-75). Springer New York.
- Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. arXiv preprint arXiv: 1201.3417.
- Bydžovská, H. (2013). Course Enrolment Recommender System (Doctoral dissertation, Masarykova univerzita, Fakulta informatiky).

Cabrera, A. F., Nora, A., & Castaneda, M. B. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. *The journal of higher education*, 64(2), 123-139.

Dougiamas, M., & Taylor, P. (2003). Moodle: Using learning communities to create an open source course management system.

Drea, C. (2004). Student Attrition and Retention in Ontario's Colleges. *College Quarterly*, 7(2), n2.

Elbadrawy, A., Polyzou, A., Ren, Z., Sweeney, M., Karypis, G., & Rangwala, H. (2016). Predicting Student Performance Using Personalized Analytics. *Computer*, 49(4), 61-69.

Garcia-Saiz, D., & Zorrilla, M. E. (2011, November). Comparing classification methods for predicting distance students' performance. In *WAPA* (pp. 26-32).

Guarín, C. E. L., Guzmán, E. L., & González, F. A. (2015). A model to predict low academic performance at a specific enrollment using data mining. *IEEE Revista Iberoamericana de Tecnologías del Aprendizaje*, 10(3), 119-125.

Hung, J. L., & Zhang, K. (2008). Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching. *MERLOT Journal of Online Learning and Teaching*.

Inoue, S., Rodgers, P. A., Tennant, A., & Spencer, N. (2017). Reducing Information to Stimulate Design Imagination. In *Design Computing and Cognition'16* (pp. 3-21). Springer, Cham.

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies*, 13(1), 61-72.

Kaminski, J. (2005). Moodle—a User-Friendly, Open Source Course Management System. *Online Journal of Nursing Informatics*, 9(1).

Kaur, G., & Singh, W. (2016). Prediction Of Student Performance Using Weka Tool.

Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Computers & Education*, 104, 18-33.

Kokina, J., Pachamanova, D., & Corbett, A. (2017). The role of data visualization and analytics in performance management: Guiding entrepreneurial growth decisions. *Journal of Accounting Education*.

Kovacic, Z. (2010). Early prediction of student success: Mining students' enrolment data.

Merceron, A., & Yacef, K. (2005, May). Educational Data Mining: a Case Study. In *AIED* (pp. 467-474).

Miller, B. N., Konstan, J. A., & Riedl, J. (2004). PocketLens: Toward a personal recommender system. *ACM Transactions on Information Systems (TOIS)*, 22(3), 437-476.

Nam, C. S., & Smith-Jackson, T. L. (2007). Web-based learning environment: A theory-based design process for development and evaluation. *Journal of information technology education*, 6.

Oskouei, R. J., & Askari, M. (2014). Predicting Academic Performance with Applying Data Mining Techniques (Generalizing the results of two

Different Case Studies). *Computer Engineering and Applications Journal*, 3(2), 79-88.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41(4), 1432-1462.

Polyzou, A., & Karypis, G. (2016). Grade prediction with models specific to students and courses. *International Journal of Data Science and Analytics*, 2(3-4), 159-171.

Ramist, L. (1981). College student attrition and retention. Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1), 135-146.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6), 601-618.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 12-27.

Sachin, R. B., & Vijay, M. S. (2012, January). A survey and future vision of data mining in educational field. In *Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on* (pp. 96-100). IEEE.

Scheuer, O., & McLaren, B. M. (2012). Educational data mining. In *Encyclopedia of the SCIENCE of Learning* (pp. 1075-1079). Springer US.

Tewari, A. S., Saroj, A., & Barman, A. G. (2015). e-Learning Recommender System for Teachers using Opinion Mining. In *Information Science and Applications* (pp. 1021-1029). Springer Berlin Heidelberg.

Tinto, V. (1987). *Leaving college: Rethinking the causes and cures of student attrition*. University of Chicago Press, 5801 S. Ellis Avenue, Chicago, IL 60637.

Yang, D., Sinha, T., Adamson, D., & Rosé, C. P. (2013, December). Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-driven education workshop*(Vol. 11, p. 14).

Zaiane, O. R. (2002, December). Building a recommender agent for e-learning systems. In *Computers in Education, 2002. Proceedings. International Conference on* (pp. 55-59). IEEE.

Stallone, M. N. (2011). Factors associated with student attrition and retention in an educational leadership doctoral program. *Journal of College Teaching & Learning (TLC)*, 1(6).

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 12-27.

Zhang, Y., Oussena, S., Clark, T., & Hyensook, K. (2010). Using data mining to improve student retention in HE: a case study.

Witten, I. H., Frank, E., Trigg, L. E., Hall, M. A., Holmes, G., & Cunningham, S. J. (1999). *Weka: Practical machine learning tools and techniques with Java implementations*.

Student Performance Prediction and Teacher Recommender System

By

Muthara-Tul-Ain

A research thesis submitted to the Department of Computer
Science, Capital University of Science and Technology, Islamabad
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE



**DEPARTMENT OF COMPUTER SCIENCE
CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY
ISLAMABAD
2017**

Copyright ©2017 by CUST Student

All rights reserved. Reproduction in whole or in part in any form requires the prior written permission of Muthara-Tul-Ain (MS133033) or designated representative



**CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY
ISLAMABAD**

Islamabad Expressway, Kahuta Road, Zone-V, Islamabad
Phone: +92 51 111 555 666, Fax: 92 51 4486705
Email: info@cust.edu.pk, Website: http://www.cust.edu.pk

CERTIFICATE OF APPROVAL

**Student Performance Prediction and Teacher Recommender
System**

by

Muthara-Tul-Ain

MS133033

THESIS EXAMINING COMMITTEE

S No	Examiner	Name	Organization
(a)	External Examiner	Dr. Ehsan Munir	CIIT, Wah Cantt
(b)	Internal Examiner	Dr. Muhammad Arshad Islam	CUST, Islamabad
(c)	Supervisor	Dr. Nayyer Masood	CUST, Islamabad

Dr. Nayyer Masood
Thesis Supervisor

November, 2017

Dr. Nayyer Masood
Head

Department of Computer Science

Dated : November, 2017

Dr. Muhammad Abdul Qadir
Dean

Faculty of Computing

Dated : November, 2017

DEDICATED

TO

MY

RESPECTED

ALLAH (SWT) & PROPHET MUHAMMAD (PBUH)

&

PARENTS, BROTHER, SISTER



FOR

THEIR CONSTANT

AFFECTION AND SUPPORT

ACKNOWLEDGMENT

All praise and exaltation is due to ALLAH (S.W.T) The creator and sustainer of all seen and unseen worlds. First and foremost I would like to express my gratitude and thanks giving to Him for providing me the boundaries and blessings to complete this work. Secondly, I would like to express my sincerest appreciation to my supervisor **Dr Nayyar Masood** for his directions, assistance, and guidance. I sincerely thanked for his support, encouragement and technical advice in the research area. I am heartily thankful to him from the final level, as he enabled me to develop an understanding of the subject. He has taught me, both consciously and unconsciously, how good experimental work is carried out. Sir you will always be remembered in my prayers.

I am highly indebted to my parents and my family, for their expectations, assistance, support and encouragement throughout the completion of this Master of Science degree. They form the most important part of my life. After ALLAH (S.W.T) they are the sole source of my being in this world. No words can ever be sufficient for the gratitude I have for my parents and for my family. A special thanks to my employing company for their support and encouragement to complete this Master of Science degree.

I pray to ALLAH (S.W.T) that may He bestow me with true success in all fields in both worlds and shower His blessed knowledge upon me for the betterment of all Muslims and whole Mankind.

AAMEEN

Muthara-Tul-Ain

DECLARATION

It is declared that this is an original piece of my own work, except where otherwise acknowledged in text and references. This work has not been submitted in any form for another degree or diploma at any university or other institution for tertiary education and shall not be submitted by me in future for obtaining any degree from this or any other University or Institution.

Muthara-Tul-Ain

Reg. No, MS133033

November, 2017

ABSTRACT

Mining data and extracting information from huge databases has become an interesting research area for the researchers. The idea to extract information with the help of data mining techniques came into being since a couple of decades ago. Initially, researchers were supposed to apply classification and clustering techniques to partite the dataset and analyze the intrinsic features. On the basis of such features, they make reasonable predictions. These predictions have taken place in the field of educational data mining for many purposes such as; predict the performance of students on the basis of factors associated with them, to enable them suitable courses and appropriate teachers. These purposes have been derived from the area of student retention and attrition. Our research aims to achieve these purposes under the roof of student attrition and retention. Moreover, we have identified such exiting factors which are beneficial for predicting the performance to students, recommend them best suitable teachers and help them to select the courses. We have applied classification algorithms with respect to the nature of data which have been collected from the Capital University of Science and Technology (CUST), ISB. In this study, the GPA of first semester on the basis of Midterm and previous academic grades have been tried to predict. For the second semester, we have predicted the CGPA of same students by using their complete proceeding academic record with the help of hybrid approach. The hybrid approach consists of combination of factors that have been evaluated against our research questions. Moreover, we tried to improve their performance by recommending those suitable courses and teachers whose performance is better amongst others comparatively. The reason of collecting the data from CUST is to validate the exiting factors in local context (Pakistan). On the basis of classification algorithms such as; Naïve Bayes and J48, we have become able to build the recommender system. Then the factors which contributed well to validate the exiting factors in local context have been measured. In the last, appropriate teacher allocation has been measured by two ways; Statistical and Prediction. In statistical experimentations, average performance of teachers, Z- test and ANOVA test has been applied. In prediction experimentations, one subject teacher with other

subject teacher's name attribute, and one subject teacher without other subject teacher's name attribute and overall performance of teachers have been computed with respect to each subject independently. With the help of our research we might have become able to provide a way to educational institutions to reduce the attrition and increase the retention rate.

Table of Contents

Abstract.....	vii
Introduction.....	1
1.1 Background Of Research	2
1.2 Problem Statement.....	6
1.3 Research Questions.....	6
1.4 Scope.....	6
1.5 Application Of Proposed Approach.....	6
1.6 Significance Of Solution.....	7
1.7 Organization Of Thesis	8
1.8 Definitions And Frequently Used Terms.....	8
Literature Review	9
2.1 Educational Data Mining	10
2.2 Student Performance Prediction	14
2.3 Commonly Used Approaches In Edm	16
2.5 Commonly Used Attributes For Student Performance Prediction.	19
2.5 Literature Review Summary	21
Research Methodology	22
3.1 Data Collection And Pre-Processing	24
3.2 Classification.....	27
3.3 Algorithms And Techniques	29
3.4 Evaluation Of Research Questions	30
Results & Evaluations.....	35
4.2 Final Attribute Selection	38
4.3 Result Of Evaluation Of Research Questions	38
Conclusion And Future Work	67
5.1 Future Work	70
References.....	71

List of Tables

Table 2.1: Literature Review Summary.....	21
Table 3.2: Summary Of Session-Wise Attrition.....	24
Table 3.3: Summary Of Session-Wise Attrition.....	25
Table 3.4: Selection Of Attributes.....	28
Table 3.5: Attributes Selection To Dataset	31
Table 3.6: Pattern Of Combination Of Attributes.....	32
Table 4.1: Occurrence Of Missing Values	36
Table 4.2: Handling Of Missing Values	37
Table 4.3: Attributes Table.....	39
Table 4.4: Classification Of Attributes And Algorithms	40
Table 4.5 Ranked Attributes.....	42
Table 4.6 GPA PREDICTION FOR THE FIRST SEMESTER	42
Table 4.7: Cgpa Prediction For The Second Semester	44
Table 4.8: Pre-Qualification Vs Maths Courses	47
Table 4.9: Average Of Cal-LI Teachers.....	48
Table 4.10: Average Of Cal-I Teachers	49
Table 4.11: Overall Average Of Maths Teachers	50
Table 4.12: Anova Analysis On Cal-1	52
Table 4.13: Two Samples For Mean With Respect To Teacher “E”.....	53
Table 4.14: Two Samples For Means With Respect To Teacher “B”	54
Table 4.15: Teacher B With Same Number Of Students.....	55
Table 4.16: Teacher B With Different Number Of Students	55
Table 4.17: Comparison Of All Programming Courses	56
Table 4.18: Comparison Of Actual And Predicted Grades Of Itc/Itp.....	57
Table 4.19: Comparison Of Actual And Predicted Performance Of Oop/Cp Teachers.....	58
Table 4.20: Overall Performance Of Oop Teachers.....	59

List of Figures

Figure 3.1: DATA FLOW DIAGRAM OF PROPOSED METHODOLOGY	23
Figure 4.1: Formats Of Data.....	36
Figure 4.2: comparison Of Local And Existing Attributes.....	41
Figure 4.3: Gpa Prediction For The 1st Semester	44
Figure 4.4: Cgpa Prediction For The Second Semester.....	45
Figure 4.5: Pre-Qualification Vs Maths Courses	48
Figure 4.6: Average Of Cal-LI Teachers	49
Figure 4.7: Overall Average Of Maths Teachers	50
Figure 4.8: Comparison Of Itc/Itp Teacher.....	51
Figure 4.9: Comparison Of Actual And Predicted Grades Of Itc/Itp.....	56
Figure 4.10: Comparison Of Actual And Predicted Performance Of Oop/Cp Teachers.....	58
Figure 4.11: Overall Performance Of Oop Teachers	59
Figure 4.12: Overall Performance Of Oop Teachers	60

List of Abbreviations

CUST: Capital University of Science and Technology

WEKA: Waikato Environment for Knowledge Analysis

MOOCs: Massive Open Online Courses

CSV: Comma-separated values (*Comma-delimited*)

ARFF: Attribute-Relation File Format.

GPA: Grade point average

CGPA: Cumulative Grade Point Average

Cal-I: Calculus-I

Cal-II: Calculus-II

ITC: Information Technology Computing

ITP: Information Technology Programming

CP: Computer Programming

OOP: Object Oriented Programming

SEM: Semester

Chapter 1

INTRODUCTION

Researchers have been working in the area of educational data mining for a decade. Educational Data Mining (EDM) has become a broadened field in which researchers are conducting their experiments to extract useful information from the data belonging to the educational sectors for many purposes. The purposes include identifying student attrition and retention rate, students' performance prediction, building a course recommender system, and teacher recommender system, etc. In the field of Educational Data Mining (EDM), researchers are busy in exploring effectiveness or role of different types of variables to measure and predict the performance of students. Among those variables, student's age, academic record and biography of students are involved. EDM is a growing research field, which carries data mining techniques in educational system (Romero, C., & Ventura, S. 2007). In previous study, (Romero, C., & Ventura, S. 2013) has explored the phenomenon of making the student's outcome better with the help of data mining approaches. According to them, huge data from the institutions have problems associated with it. Therefore, data mining approaches are not supposed to be applied directly on this data. Therefore, knowledge discovery process has been implemented. In educational institutions, data mining is playing a pivotal role on datasets after preprocessing. Most widely used data mining approaches involved classification and clustering, outlier detection, association rule mining, pattern mining and text mining. Researchers are trying to extract useful, novel and interesting information with the help of data mining approaches (Romero, C., & Ventura, S. 2010). EDM process is used to convert raw educational data into useful information.

In the field of EDM, researchers are active in three broad categories; Personal recommender system and learning environment, course management system, and student attrition and retention. Our research will cover all these mentioned areas in different contexts such as identification of students with low academic performance, building prediction model to predict the student's performance by

using their historical data, improvement in raising the confidence level among them, assist them to choose their courses, building a model that will assign appropriate teachers to the students.

1.1 Background of research

As discussed in above section, EDM is subfield of data mining which is applied in educational institutions to help them to maintain their prestige. Educational data mining is all related to propose and develop novel approaches to improve the performance of students on a large scale. Educational data mining is an active research areas in which researchers are working actively in three main areas of educational data mining are:

a) Personal recommender system and learning environment

Researchers are using this concept in hybrid fashion in the context of educational data mining. They are constantly using various data mining approaches to make the personal recommender system better to produce maximum accurate results. In a research (Miller, B. 2004), personal recommender system uses collaborative filtering to reduce noise from the data. A limitation of personal recommender system is found to be its un-portability as they are operational only on large computers that are connected to the internet. Another limitation was found to be its trustworthy relationship between the owner of recommender system and its user.

b) Course management system and educational data mining

Researchers are associated with this branch of EDM to provide the universities and educational institutions a platform that can help them in management of courses in the department. For this purpose, different data mining approaches have been carried out (Peña-Ayala, A. 2014). The concept has been derived from e-learning system “Moodle”. Moodle is a learning management system that helps educators to build up effective online learning communications through the analysis of information generated by the users (Aher, S. B, 2012).

c) Student attrition and retention

Student attrition means the number of students who are leave their courses without completing because of certain reasons. Those reasons may include poor choices of courses, consistent undesirable result, poor performance, financially instability and immature selection of courses etc whereas student retention means the number of students who complete their course and acquire degree despite of any kind of circumstance and gain good grades in the transcript. In educational institutions, rate of student attrition and retention has reached up to considerable unit (Stallone, M. N. 2011).

Student attrition and retention not only affect the performance of education institution or department but also affect the faculty positions and raise the financial problems for the parents eventually. To analyze the pattern of student retention and attrition, one has to find out the reasons of this cause then take steps to resolve this problem. Our research topic falls into the type student attrition and retention which is further explained in the following subsection in detail. For this research collected dataset from Capital University of science and Technology, Department of Computer Science. Let's consider first semester and second semester courses because student attrition rate is high in first and second semester. These semesters are helpful attrition of students and predicted the student performance.

Student performance prediction

To recognize the efforts of students in academia, there are some known parameters carried out by universities or educational institutions (Romero, C., & Ventura, S, 2007). Among those parameters, student's grades, attendance, grades per courses are commonly evaluated. Student performance prediction has been the growing research problem for the researchers since a decade. To improve the institution's prestige and efficiency with respect to student performance improvement, there are several techniques used by different researchers such as course specific regression, personalized linear multi regression, classification and clustering. These approaches not only work for educational institutions those are in existence physically but also for virtual campuses too i-e; LMS and Moore of Stanford University (Elbadrawy, et al,

2016). EDM is a broad research field in which researchers are exploring a lot of problems including student's performance, attrition and retention, course selection, and particular teacher's selection. The widely covered research problems in the domain of educational data mining are known as student's future grade prediction, performance prediction and course enrollment recommender system.

This research is based on different aspects. One aspect is to identify the number of those students who had to leave the institution because of their poor performance in studies or financial problem. Then the number of those students have been found who completed their degree belong to the spring semesters of three years. Expectedly this research will help the students to improve their results in the courses which have been taken for the experiments by recommending them appropriate teachers and courses based on their associated factors.

This is helpful not only for students to raise their grades but also for parents not to suffer from financial problems. Another aspect which is being considered is to retain maximum percentage of students in the university and fulfill their degree requirements till its completion. Ultimately, if the grades of students would be good their retention will become stronger in the institution. This research might reduce the attrition rate because of the provision of teacher recommender system, course recommender system and student performance prediction. The outcome of our research has been discussed in detail as follows.

Students Performance Prediction is the research problem which has been identified by many researchers with the help of different approaches such as Matrix Factorization, classification and clustering algorithms (Polyzou, A., & Karypis, G, 2016). Before applying these approaches, students variables are supposed to be extracted from the collected dataset that are unique registration number, result of previous two terms, GPA (Grade Point Average), number of family members, family welfare, gender, usage of social networking and use of technology and internet (Oskouei, R. J., & Askari, M, 2014). After collection, data preprocessing is applied in which positive and negative effects of dataset

are removed and then classification and clustering are applied. By using such techniques, data can be further divided into clusters and classes to analyze useful and hidden patterns from it (Kabakchieva, D, 2013).

With respect to performance prediction of students, there are many cases such as student's next term grade prediction, presence performance based on assessment in the current courses. To evaluate this problem, researchers have used some of the characteristics of students such as admission records, High school scores, SAT/ACT scores and grades of previously completed courses (Elbadrawy, et al, 2016). More characteristics such as class test, seminar attendance and marks, assignment marks (Baradwaj, B. K., & Pal, S. 2012).

Course enrollment recommender system It helps to determine the performance of students by recommending them courses up to their performance. For this purpose researchers have proposed the recommender `in which he used 67 templates of active students. This recommender system not only helps to predict the grades of students, it also helps to recommend the courses to the students by considering their timetable (Bydžovská, H, 2013).

Proposed attributes of students affects their performance in various ways. It not only helps in identifying the students with low academic performance but also paves the ways to improve their grades and enables recommender systems of universities to predict the grades of students. For this purpose, different classification techniques of data mining have been studied in this research. With the help of this study, student carrier can be affected by means of building up and students can easily pave their way towards their desired destination. This study will also help to construct the course recommender system which will allow the students to select their courses that are recommended by the proposed system. Another advantage of this study is highly beneficial economically for parents. With the help of proposed system, students may not waste their money in selection of courses to repeat. Moreover, student's academic record can be improved and efficiency of department management can be raised.

The factors used in the research have already been applied in countries other than Pakistan. Those countries are Canada, USA, England, and Nigeria. All these factors have been applied in the colleges and universities of narrated countries and now will be validated in the context of Pakistan.

1.2 Problem Statement

A major problem that the academic institutions are facing worldwide is poor performance of students in academics. This causes high attrition rate that is a loss to students, parents and institutions. Student's performance prediction can help to reduce attrition rate as it raises an early alarm for students not performing well and are likely to leave the institution/studies. Most of the work on performance prediction has been done in foreign institutions; this has to be done in the context of Pakistan for more accurate results.

1.3 Research Questions

Given below is the proposed problem statement

RQ1. Whether existing identified factors for grade prediction are valid in our local context (Pakistan)?

RQ2. Which factors help in accurate prediction of students' GPA of first semester?

RQ3. Is it possible to improve the performance of a student by allocating of appropriate teacher for a subject?

1.4 Scope

The application of this research is versatile. The experiments of this research have been conducted on the dataset that has been collected from Capital University of Science and Technology, Islamabad (Pakistan). Such factors can be applied on the data of Government institutions as well as schools and colleges.

1.5 Application of Proposed Approach

Our proposed research will be beneficial in one of the following ways in the domain of educational data mining.

To make the university management system efficient

This research might be highly beneficial for the university management system to keep check on the performance of student's belong to different departments. This will not only help to predict and improve student's performance but also supports the parents economically.

To enable Course recommender system pro-active

The selection of courses is major problem for the students. To build such recommender system that helps students in selection and registration of the courses will be quite beneficial for the institutions.

Helpful in improving teacher recommender system

In the proposed solution this recommender system will help department to recommend specific teachers to particular students based on previous 0239 academic record and inclination of selected courses in which they pursued good grades.

1.6 Significance of solution

This research is a fine contribution for the educational institutions of Pakistan. With the help of this research, the factors that have been used by the researchers from the foreign universities are considered in local context. These factors are planned to apply over the data set of Capital University of Science and Technology of Pakistan. These factors have been utilized with the help of WEKA for the student's performance prediction. Universities and other educational institutions may capable of predicting the student's next term grades.

Moreover, with the help of the contribution of this research there are chances to propose teacher recommender system that will help to recommend suitable teachers to the students by considering their current performance in the semester and their relative interest in the courses that might improve their grades.

1.7 Organization of thesis

This thesis comprises of five chapters. The first chapter states the introduction of the proposed research. Second chapter is the literature review in which the work and research contribution in chosen area by former researchers has been discussed. Moreover, overall literature summary has been presented in summarized way. Third chapter contains the methodology diagram that has been adopted to perform the experiments and how experiments will be performed. The results of experiments by using selected tools have been presented in the chapter four. Last chapter contains the conclusion and future work in which summary of whole thesis has been presented.

1.8 Definitions and frequently used terms

Performance Prediction

A student's grade is used to acknowledge his/her related performance in academia. It is now possible to predict the student's performance with scientific methods in which quantitative approaches have been used by the researchers (Bhardwaj, B. K, 2012) (Pal, S. 2012).

Recommender System

Usually recommender system is built to assign a particular entity to required entity. In the field of data mining, researchers have proposed some recommender systems in which teacher recommender system and course recommender systems are involved (Bozo, J., 2010) (Alarcón, R.,2010)(Iribarra, S., 2010).

Chapter 2

LITERATURE REVIEW

The initiative of online course learning system is based on e-commerce venture. With the growth of this venture, data from web resources started collecting and storing in the excel that contains customer and product information and order information (Kokina, J, 2017). E-commerce is a term that refers to use as online business through internet. There are various websites that are working for this purpose such as e-bay, Alibaba, and Amazon etc. It can be said that data storage in excel from the resources has been derived from e-commerce. Predicting student's performance by using data mining techniques to extract information from the academic dataset of universities has become state of the art research in the scientific society. Universities are confronting with some challenges now a day to analyze the performance of their students. That's why researchers are focusing on student's profiles and characteristics to make the university management aware of student's performance and overall academic result (Kabakchieva, D, 2013). There is another dimension of student's performance that is the dependence of student retention upon student student's performance. To minimize the problem of student retention cases in the universities, different researchers have proposed different methods to predict the performance of students in their future semester based on the performance of previous one.

To predict the courses of next term grades, four parameters have been considered in this study such as; admission records, High school scores, SAT/ACT scores and grades of previously completed courses. Based upon these parameters, recommender system can be trained to predict the grades of students accurately in any of the educational institution. Historical information about the course has also been considered in this study such as which course is taught by which teacher and information about contents of the course. Many researchers have used LMS and Moore to predict the successive chances of success and failure of students. In this research, regression based methods such as course specific regression (CSPR) and personalized linear multi regression

(PLMR) has been used. Another method known as matrix factorization based methods in which standard matrix factorization (MF) has been used for the grade prediction of students (Elbadrawy, A, et al, 2015).

In this chapter, the background of educational data mining and its branches will be discussed in detail. Also the student's performance prediction and data mining approaches that are commonly used by researchers in the literature are being discussed.

2.1 Educational Data Mining

Data mining is a process of sorting data and extracting information from existing databases.¹ With the help of pattern mining and data analysis, hidden information can be obtained from huge datasets. The strategy of data mining is now applied in the field of education by researchers. They are busy in exploiting a lot of dimensions in education sector. This is now known as educational data mining. Data mining is applying in educational sector by considering the performance of students and finding the position of students by using their academic records. Educational dataset is being collected from various resources such as interactive learning systems, computer-supported collaborative systems, and administrative datasets of school, colleges and universities². Data mining methods are now implemented in well known universities to analyze the patterns of student performance from the dataset through which information can be extract and decision making may become easier for the management of institutions (Kabakchieva, D. 2013).

With the incremental growth in the use of technology everywhere, educational institutions are now busy in finding hidden trends and patterns in their larger datasets (Scheuer, O. 2012). (Merceron, A. 2005). Datasets that are used for experimental purposes in educational institutions have become possibly available because of web based educational systems in which LMS, Moore and Portal system have become common. With the help of these sources, dataset can easily be collected if authorization is accessed. One purpose of extracting information from its own dataset is to make its prestige among other

¹ <http://searchsqlserver.techtarget.com/definition/data-mining>

² <http://www.educationaldatamining.org/>

educational institutions stronger. Another purpose is to build the student career by covering its each and every aspect such as improvement of their grades if lacking, overall performance booster, support them financially, make them enable to select the courses suggested by course recommender system, assign them appropriate teachers based on their inclination of interest in course selection and many more.

However, these are some common areas in which researchers are producing their research by using different data mining techniques. In our research, same areas are being covered under the dataset of Pakistani students. Three factors are being applied on students those factors have been discussed briefly in section 2.4. In current section, our focus will remain on key areas of EDM, in which researchers are engaged.

a) Personal recommender and learning environment

Web based learning environments have become common right now. It has become powerful medium to act as a bridge between learner and instructor and provides interesting learning mechanisms to both (Nam, C. S, 2007) (Osmar, R, 2002) has elaborated the concept and working of web based learning which is also known as e-learning. A recommender system helps to recommend the actions to a learner by using his previous actions with the help of intelligent software agents. The use of such recommender systems were first initiated in e-commerce and now carrying out by researchers in the domain of e-learning. Osmar has implemented this recommender system in assisting the offline web miners who are responsible to find hidden trends in the data and online course navigation. With the help of recommender system, educational institutions are now strengthening their management and departmental progress. Researchers have proposed several recommender systems among which course recommender system and teacher recommender system are commonly known.

b) Course management system

By means of course management, researchers have explored different sources to find the best possible use of course management system. In this context, Moodle is found to be open source course management system that helps

educators to make their courses available on the internet. All the data on the Moodle is managed by Moodle team. Moodle is now facilitating educational institutions, community colleges, and schools to create online teaching system (Dougiamas, M., 2003). It delivers courses online, unlike traditional classrooms. Data of courses and students is stored in its specified database which is further used by researchers that carries out for mining purposes to extract useful information from it. Performance of Moodle has raise higher in virtual environment. Other online course management system such as web portals and learning management systems are serving for the same purpose now days.

With the use of internet in classrooms and institutions, those institutions have launched their proper websites or web pages, through which students can register themselves into particular courses (Kaminski, J. 2005). In the area of online course learning, researchers have presented many research articles. In this domain LMS and Moore have contributed a lot. They have been offering different courses on their websites and thousands of students from all over the world registered themselves. This is how, huge amount of dataset is collected from LMS and Moore and researchers have performed analysis on these dataset to evaluate the performance of students who enrolled (Kizilcec, R. F, 2017) (Pérez-Sanagustín, M., 2017) (Maldonado, J. J. 2017). Self regulated learning (SRL) is a term that has been stated by narrated researchers. According to them, students having strong SRL are good enough in planning, managing and controlling as compared to the students having weak SRL. In this regard, MOOCs have been providing support for the learners with the help of different levels of SRL.

c) Student attrition and retention

With the passage of time, growth of private educational institutions has been increased up to the remarkable extend. These institutions have become source of higher learning and business entity. Therefore, maximum number of student's enrollment is its lifeline. For the survival of private institutions, profitability, proper management and alignment are mandatory. In this respect, student retention until the completion of degree is quite necessary. That's why

institutions are finding that factor that ultimately causes student attrition (Azarcon Jr et al, 2014). After analyzing those factors, it is important for educational institutions to make strategic adjustments accordingly to improve student retention in institutions. In Gaviria, Colombia, people are closely belongs to social mobility that is a causing the academic performance of the students and student attrition rate may increase in educational institutions. Researchers have used three drivers the student performance analytics. The first driver is the volume of data that is being collected from learning management system and student information system, second one is the e-learning and third one is political concerns (Guarín, C. E. L., 2015)(Guzmán, E. L.,2015)(González, F. A. 2015). The application of data mining in the field of educational data mining has been emerged in different areas and researchers are exploiting these areas with various dimensions.

There is another supplement for student learning through web that is known as MOOCs. In online course learning and management system, high rate of dropout of students have been identified by researchers. This problem has been enlightened by the YANG in his research that is students are dropping at considerable level in Massive Open Online Courses (MOOCs). To control the student attrition problem from the coarser classes, researchers have proposed a model in their research. This model is a helping hand to determine such influential factors that are causing student drop out. So the predictors have been tried to propose in this research that will determine the factors related to student's behavior and social position in the discussion in which they participate in the forums (Yang, D.et al, 2013).

The problem of student attrition and retention is not new for the educational institutions. It has been enlightened by the researchers from the fields of data mining and information visualization. Now it has become very common research problem for the researchers. Student attrition and retention problem has been observed by the researchers when this problem was raised up to the ratio of 50% on the colleges of Ontario (Drea, C. 2004). To reduce attrition rates, institutions should focus on student retention. Researchers have analyzed the factors that causes student attrition and in the research, Drea has addressed

both elements; student attrition and retention. To retain the persistence of institutions, two theories have been formulated in this context. The first one is student integration model by Tinto and student attrition model by Bean. Both models work almost similar (Cabrera, A. F. 1993).

There are numerous reasons for what student attrition has become research problem for the researchers. In those reasons, personal disappointments, financial setbacks, and lowering of career and life goals are considerable. Therefore scientists have carried out the retention and persistent in their research to resolve this societal problem (Ramist, L.1981). This research focus on student performance and also find a teacher methodology has positive impact on student grade prediction and has reduced student retention rate.

2.2 Student performance prediction

Student performance prediction is the research problem which has been considered by many researchers from the field of educational data mining (Kabakchieva, D, 2013). With respect to performance prediction of students, there are many cases such as student's next term grade prediction, presence performance based on assessment in the current courses. To evaluate this problem, researchers have used some of the characteristics of students such as admission records,

High school scores, SAT/ACT scores and grades of previously completed courses (Elbadrawy, et al, 2016). More characteristics such as class test, seminar attendance and marks, assignment marks (Baradwaj, B. K., & Pal, S. 2012). For the sake of performance prediction, data that has been stored in open polytechnic student management system belong to 2006-2009 had been used. For the experiments to be applied over the dataset, feature selection and clustering techniques were applied on the dataset which are also known as data mining techniques.

In the field of education, predicting the performance of students has become considerably important for the university management and departments. Because the prestige of any educational institution is ultimately depends upon the performance of its students and if performance of maximum number of

students degrade up to the mark able level, it should be considered by the department management and find out the factors that affects the performance of students. Under consideration of this problem, researchers have evaluated their studies against the parameters that affects the overall performance of students, data mining techniques and data mining tools (Kaur, G, 2016)(Singh, W, 2016). In such parameters, psychological, personal and environmental factors are involved. For the experiments, they have used Naïve Based and J48 techniques through WEKA.

a) Future grades prediction

To predict the future grades of students is the considerable research problem that has been identified by many researchers with the help of different approaches such as Matrix Factorization, classification and clustering algorithms (Polyzou, A., & Karypis, G, 2016). Before applying these approaches, students variables are supposed to be extracted from the collected dataset that are unique registration number, result of previous two terms, GPI (Grade Point Average), number of family members, family welfare, gender, usage of social networking and use of technology and internet (Oskouei, R. J., & Askari, M, 2014). After collection, data preprocessing is applied in which positive and negative effects of dataset are removed and then classification and clustering are applied. By using such techniques, data can be further divided into clusters and classes to analyze useful and hidden patterns from it. To predict the performance of students and predict their grades, researchers have used socio-demographic factors of the students such as age, gender, ethnicity, education, work, status, and disability. In addition, researchers have used environment may also affect the student's performance that causes their dropout or retention throughout the course accomplishment (Kovacic, Z. 2010).

b) Teacher recommender system

When students retain their degree the end of their duration, it encompasses the quality education of any educational institution that refers to the term of student retention. According to the national statistics in a research, fifty percent of the students from higher educational institutions get attired. To (Bozo, J., 2010)

(Alarcón, R., 2010) (Iribarra, S., 2010). As online course enrollment system has taken the place globally because of internet. In this variation in technology, there are several web pages and websites that are busy in providing best at their own but still contain some missing elements in their services. Therefore, researchers have worked upon this problem and proposed solution regardingly. To make every course learnable creatively, researchers has focused on teacher recommendation according to the course respectively. The new recommender system is known as A³. With the help of this recommender system, not only accurate teacher will be associated to the relative course contents but also best contents of the course will be available on that site after some time (Tewari, A. S, 2015).

c) Course recommender system

A course recommender system helps to determine the performance of students by recommending them courses up to their performance. For this purpose researchers have proposed the recommender in which he used 67 templates of active students. This recommender system not only helps to predict the grades of students, it also helps to recommend the courses to the students by considering their timetable (Bydžovská, H, 2013).

2.3 Commonly used approaches in EDM

From the literature, a lot of data mining approaches that have been carried out by the researchers to evaluate their studies. Most commonly used approaches have been discussed in detail in subsections of section.

a) Classification and Clustering

To improve the performance of weak students in the class, with the help of data mining techniques, collected data of 1100 students have been transformed in Weka. Researchers have used freeware software such as Weka, Clementine and Rapid-Miner in this work. Different classifiers such as Naïve Based, C4.5, Neural networks and random forest have been used. The classifiers which have been used are Adaboost, Bagging and boosting. It is found that some factors have same effects on both countries and some have different. Moreover, it has

been found that male students suffer with stress more as compared to female. The performance of male students in Mathematics and formal SCIENCE is better whereas performance of female students aroused better in Literature and mnemonic SCIENCE (Oskouei, R. J.,2014)(Askari, M. 2014).In this field of research, there exist wide varieties of benchmark which are used to evaluate the performance and accuracy of experiments conducted by using machine learning approaches. Different researchers have used various types of educational datasets and each dataset is unique in its attributes (Garcia-Saiz, D., 2011) (Zorrilla, M. E, 2011). Therefore, in the study researchers have proposed meta algorithm to preprocess dataset. Various data mining models have been studied in this research to find the most accurate one with the help of Meta algorithm.

Educational Data Mining approach is emerging in providing feasibility to the educational institutions to improve their teaching and learning methodologies. EDM is the process of applying data mining techniques in educational institutions. In educational institutions, the data of student's performance, teacher's evaluation, student's enrollment data, and gender differences are to be stored in the database. On this data, Classification, Clustering and association rules are applied. These rules have been applied on mandatory courses taught in IT department of King Saud University. To extend this approach, these rules can be applied on elective courses as well to predict the grades of students in all courses as well as final GPA. Furthermore, neural network and clustering can also make these rules effective for prediction (Al-Barrak, M. A., & Al-Razgan, M. 2016). Researchers have considered more attributes of students such as their attendance, class test, seminar attendance and marks, assignment marks and constructed classification tree in Weka. According to those researchers, teachers can minimize the failure ratio of students in the semester by adopting this approach and attributes (Baradwaj, B. K., & Pal, S. 2012).

b) Association rule mining

In the study, some factors which affects the grades of students of Iran and India has been observed. In such factors, their respective gender, family background, education level of their parents and their lifestyle has been encountered by

asking questions from them (Oskouei, R. J., & Askari, M, 2014). To evaluate the performance of students and improve the management of educational institutions, researchers have introduced pre university characteristics of students which are known as student's profile and place of secondary school, final secondary education score, total admission score, and score achieved that exams. In this research data of University of National and World Economy (Bulgaria) has been collected and data mining techniques has been applied. In such data mining techniques naïve bayes and bayes net, nearest neighbor algorithm, and two rule learners One R and JRIP has been applied on the dataset (Kabakchieva, D. 2013).

c) Pattern mining

As discussed earlier, it is now clear knowledge is gathered about online teaching and learning system. Universities through internet are now big source of student-teacher interaction and source of learning and training. This technology has made place in the field of educational data mining in which knowledge diffusion has become research icon for the researchers who belong to field of data mining, web mining and graph algorithms. The data of online teaching system is serving best to the researchers who keen to extract knowledge from big datasets. Researchers are analyzing patterns of online learning behavior of students and to draw outcomes from these sets of data, they have been using machine learning algorithms and various data mining approaches. In a study, researchers used 19,934 servers to identify the behavior of students in Taiwan who had registered online courses and with the help of this data, they drawn conclusion after predicting their performance too (Hung, J. L., & Zhang, K. 2008). In addition, data mining techniques were proven to be helpful for the course developers, online trainers, and instructional designers. In their study, they used WEKA and KNIME tools to perform analysis of descriptive and artificial intelligence. Moreover, for data visualization and statistical analysis, they used

SPSS. In another study, data mining techniques such as classification, clustering, relationship mining, prediction and social area networking have been applied in educational data. (Sachin, R. B., & Vijay, M. S. 2012). Well all

of the mentioned approaches are ultimately works to represent the results after applying over the data. The concept of data visualization has been derived from visual reasoning (Inoue, S et al, 2017).

d) Text mining

In the past decade, there are number of tools such as WEKA, RapidMiner, R, KEEL, and SNAPP that have been used to extract text (Useful information) from the datasets in the field of educational data mining (Baker, R. S., & Inventado, P. S. 2014). In other research, data mining is also known as knowledge discovery from databases (Anand, S. S. et al, 1996). In this process, database techniques are bind with mathematical and artificial intelligence techniques.

2.4 Commonly used attributes for student performance prediction.

While review of papers from the literature, many data mining approaches have been found that are applied in academic datasets of different educational institutions for various purposes. Such approaches have been applied on the attributes that are considered after analyzing the factors. Those factors are briefly discussed in following passage.

I. Demographic attributes

From the literature, the importance of demographic attributes upon student's performance is inaugurated. According to the researchers, they condemn that living standard of any student ultimately affect their studies and variation their performance. Demographic factors include Age, gender, Financial status, Balance due, Permanent address, Residential address, Guardian (Brother, uncle, Self), Father qualification, Father Occupation, Full/ Part time student. Demographic factors equally supports the researchers whether to find the student attrition or student's retention. Student's retention is an obvious success of educational institutions. In order to sustain from student retention and keep the performance of students up to the mark, educational institutions better need to recruit the limited students as per its availability. In a study, the student's

general weighted average, School's Radial Distance, and school ownership are taken (Abaya, S. A., 2013) (Gerardo, B. D., 2013)

II. Pre-university attributes

Another factor which has been found during literature survey is *pre-university attributes*. These attributes have quite great impact on student's performance in any educational institution. Pre-university attributes will be fruitful with respect to accommodate the student's interest to map in course recommendation system. Pre-university attributes include *Secondary School Grade, Higher Secondary Grade, SAT Score, Pre-college, Pre-board, pre-program*. Researchers have used the historical data of students as well to find the attrition rate of students from academia through cross validation process under the classification and naïve based methods Guarín, C. E. L., 2015)(Guzmán, E. L.,2015)(González, F. A. 2015).

III. Institutional attributes

Last factor that has been extracted from literature to improve the performance of students is institutional attributes. These attributes will support in modeling the teacher recommender system and future course prediction model in our study. In institutional factors, *Term 1 GPA, Term 2 GPA, Term 3 GPA, and Term 4 GPA, CGPA, Total Credit hours taken, Total courses taken and Initial major, current major, current enrollment status* are included.

Proposed attributes of students affects their performance in various ways. It not only helps in identifying the students with low academic performance but also paves ways to improve their grades and enables recommender systems of universities to predict the grades of students. With the help of this study, student carrier can be effected by means of building up and students can easily pave their way towards their desired destination. This study will also help to construct the course recommender system which will enable students to select their courses that are recommended by the proposed system. Another advantage of this study is highly beneficial economically for parents. With the help of proposed system, students may not waste their money in selection of courses to

repeat. Moreover, student’s academic record can be improved and efficiency of department management can be raised.

2.5 Literature review summary

Table 2.1: Literature Review Summary

Reference	Relevant Techniques	Tools	Used Attributes		
			Demographic	Pre-university attributes	Institutional attributes
(Guarín, C. E. L., Guzmán, E. L., González, F. A. 2015)	Decision Trees, Bayesian Classification	10 fold cross validation model, Cost sensitive model		√	
(Abaya, S. A.& Gerardo, B. D., 2013)	Classification through C4.5	Irecruit Application in UNIX	√		
(Kovacic, Z. 2010).	Clustering, Feature selection thorough cross validation, and CART classification	CHAID Model, Gian chart	√		
Kabakchieva, D. (2013).	CRISP-DM, Neural networking, Nearest neighbor classifier, Rule learner, Decision tree classifier	WEKA		√	
(Al-Barrak, M. A., & Al-Razgan, M. 2016).	Data Visualization, Classification	E-learning Web Miner, WEKA			√
(Baradwaj, B. K., & Pal, S. 2012)	Classification and Decision trees	Manual			√
(Elbadrawy et al, 2016)	Matrix factorization based methods, regression based methods	Recommend er system based personal analytics			√
(Bydžovská, H. 2013).	EDM, SNA and Collaborative filtering	WEKA and R	√		
(Kaur, G., & Singh, W. 2016)	Naïve based and J48 Decision trees	WEKA	√		
(Kokina, J., Pachamanova, D., & Corbett, A., 2017).	Predictive Modeling, Data Visualization	Excel and Tableau			√

Chapter 3

RESEARCH METHODOLOGY

This chapter presents the methodology adopted to address the research questions presented in chapter 1. The focus of the research is accurate grade prediction of critical courses of first semester of BS CS students of CUST. Prediction will help to take appropriate measures to control the attrition rate and hence will be beneficial for students, parents and university. Targeting the same objective, appropriate faculty members are also being recommended based on the results of previous semester. The factors under consideration for the proposed research are classified into three categories; Demographic, Pre-university and Institutional. Such researches in the field of educational data mining have been conducted in foreign earlier. Experiments have been conducted in local context, Pakistan. For this purpose, all the data have been collected from Capital University of Science and Technology, Islamabad Pakistan. After data collection and its pre-processing, some data mining techniques are being selected like SVM, Linear regression and Non-linear regression in WEKA.

This study will not only be useful in improving the overall performance of students but also reduces the attrition rate. Now it is easy to do such things on the basis of early prediction. Field of data mining has been emerged with the linkage of natural language processing, artificial intelligence, visual data analytics, and social data analysis etc (Romero, C., & Ventura, S. 2013)(Zhang, Y.et al, 2010). This early prediction is computed by using term marks and mid-term marks of the students along with their pre-university; Intermediate/O-levels marks and Matriculation marks and demographic factors; gender and city. This methodology will be discussed briefly in the current chapter. The related work of this research task has been discussed in detail in chapter 2. Chapter 3 contains detailed discussion of proposed methodology along with the data flow diagram that has been given as 3.1

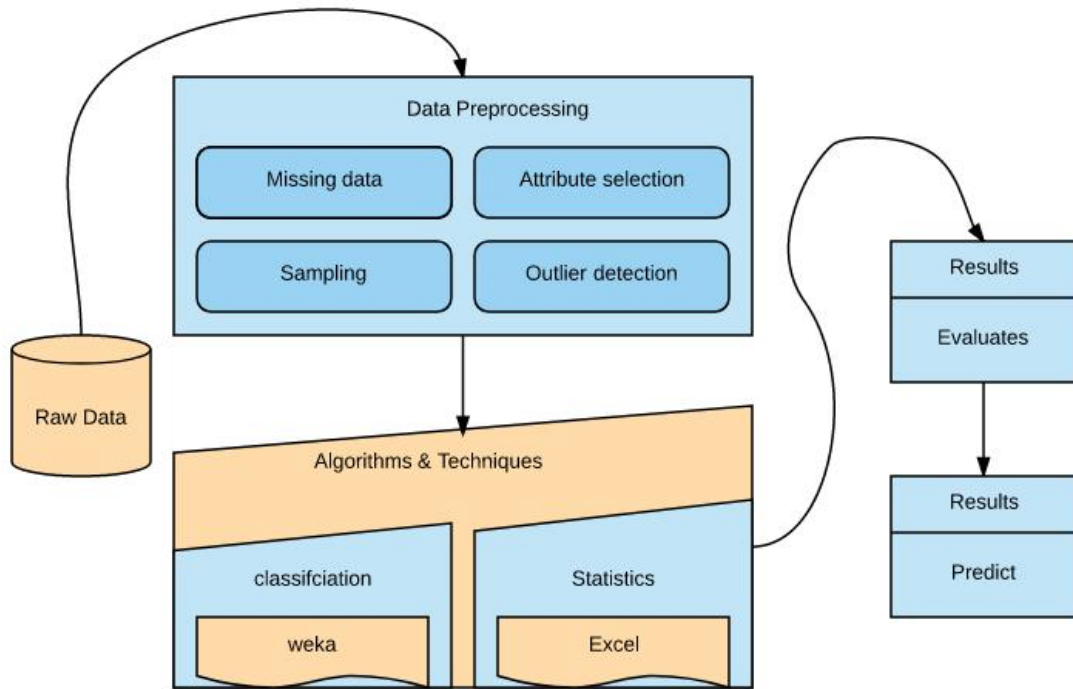




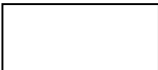


Figure 3.1: Data Flow Diagram of Proposed Methodology

Proposed data flow diagram is constructed on the basis of following set of symbols.

Table 3.1: Description about Data Flow Diagram

	<p>This cylindrical shape represents the raw data that has been collected for the research and need to be pre-processed.</p>
	<p>This filled arrow is used to represent the flow of tasks that has been carried out for the experiments.</p>
	<p>This symbol represents the working of pre-processing steps in our data flow diagram.</p>
	<p>This oval shaped rectangle represents the algorithms and techniques that are supposed to apply on the data set for the experiments.</p>
	<p>This shape represents the final results that include the result prediction and evaluation of results.</p>

To conduct the experiments, first of all the data is collected from the CUST (Capital University of Science and Technology), Islamabad Pakistan. The data set contains the data of six semesters that belongs to the BSCS program. On the basis adopted factors, the GPA of 1st semester as well as CGPA of 2nd semester has been predicted. In addition, performance of teachers against their respective courses has been measured through ANOVA analysis. The experiments have been conducted through WEKA. Each research question has been answered with the help of proposed methodology and experiments.

3.1 Data collection and Pre-processing

As discussed earlier, domain of computer science is being selected for the experiments and CUST has provided the facility to collect the data from its sources. The academia of CUST is semester based and there are two semesters per year. One semester corresponds to the spring semester which starts from February and ends in June. And the fall intake starts from mid of September and ends in the end of January. The data initially contained 5 courses of first semester, such as English-1, Physics, Calculus-1, Pakistan Studies & Islamiat and ITC/CP.

Table 3.2: Dataset Summary

Term	Registered Students
Spring 2014	74
Fall 2014	155
Spring 2015	115
Fall 2015	102
Spring 2016	62
Fall 2016	101
Spring 2017	58
Total	667

However, the information has been extracted from a research (Junaid. 2017), that most crucial courses with respect to attrition or students' performance are ITP/CP and Cal-I. So results of only these two semesters are considered and excluded the rest three.

Collected data is related to the students of first semesters and seven semesters starting from Spring 2014 (Term 141) till Spring 2017 (Term 171), and also of second semester students of six semesters starting from Fall 2014 (Term 143) to Spring 2017 (Term 173). The courses of first semester, as mentioned earlier, are Cal-I and ITC/CP, and the second semester courses are Cal-II and CP/OOP from each spring new intakes that belongs to the semester no 1. The initiative of this research is to identify the factors which contribute to keep the students retained till the end of the semesters as well those factors that affect the performance of the students. Following table 3.2 shows the information of the students comprehensively.

Table 3.2: Summary of Session-wise Attrition

Term	1st Sem	2nd Sem	3rd Sem	4th Sem	5th Sem	6th Sem	7th Sem	8th Sem
141	64%	20%	8%	8%	0%	0%	0%	0%
143	29%	35%	25%	12%	0%	0%	0%	0%
151	60%	30%	9%	0%	0%	0%	0%	0%
153	71%	25%	4%	0%	0%	0%	0%	0%
161	94%	6%	0%	0%	0%	0%	0%	0%
163	100%	0%	0%	0%	0%	0%	0%	0%
Attrition %	70%	19%	8%	3%	0%	0%	0%	0%

The dataset contains the institutional factors; Registration Number, Name, GPA, Internal marks like midterm, and term works marks etc and overall information of the students such as Student's registration number, GPA, Registered courses, Matriculation and Intermediate results. To conduct the experiments, we have used demographic and pre-university factors of the students apart from institutional factors. In the data set, the data of term 141, 143, 151, 153, 161, 163, and term 171 are included. The terms of the spring semesters are associated with the registration numbers of the students such as the term 141 means spring 2014. Similarly, the data of spring 2015 and spring 2016 are collected and on the basis of three years record of the students. The file of dataset is initially imported into WEKA after converted into CSV and ARFF format for the pre-processing and further experimentations have been conducted respectively.

Missing Data Handling

The problem of missing data generally arises due to absence of data in an observation for any variable during experiments. Missing data problem also arises when no information is provided or unavailable for the variables. The reason of missing data may involve the problem in collection of data by the researchers or the sources have not provided it completely³.

Missing data has been handled by two ways. Firstly, null values of courses have been found from other semesters and may be read in late semester in which students have passed that course. In case of consistent absence of values, the grades of such course have been supposed by taking the average result of other courses in respective semester. The missing data can be handled through different filters in WEKA.

³ <https://measuringu.com/handle-missing-data/>

Noisy Data Handling

While data collection, irrelevant data is called as noise. As the data of BSCS students of spring 2014 from spring 2016 is collected, therefore during the collection, data belongs to Software engineering and bioinformatics was present and then eliminated through filters in WEKA. Noisy data occurred in the form of errors such as; GPA = “-0.4” and Intermediate or Matriculation marks = “-455, 544.8” which were handled in Pre-processing of data.

Attribute Selection

After missing data handling and outlier detection phase, final attributes have been selected on which overall experiments and results depend. Such attributes have been narrated in table 3.1 that belongs to all three types of factors; demographic, pre-university, and institutional. One more factor has been considered to answer our research question that is teacher performance. After completion of data pre-processing, classification algorithms have been applied. Attributes are then finalized for the experiments that have been evaluated to answer the research questions. In the current section, we have discussed adopted classifiers in detail.

3.2 Classification

Before applying algorithms, factors have been grouped that are earlier used by foreign researchers. These factors have been compared against the factors used by the researchers in local context (Pakistan). Moreover, we have expanded our research work by making the variations of different combinations attributes and then compared. The comprehensive detail about these factors has been mentioned in table 3.4.

Table 3.3: Selection of Attributes

Sr.No.	Factors	Attribute Name
1.	DEMOGRAPHIC	Gender
2.		Age
3.		Residence
4.		Location
5.		Race
6.		Father's qualification
7.		Father's occupation
8	PRE-COLLEGE	Secondary school grade
9		Higher secondary grade
10		Pre-college
11		Pre-program
12		SAT score
13	INSTITUTIONAL	GPA (Term1, Term2,Term3, Term4)
14		CGPA
15		Financial status
16		Total credit hours taken
17		Total courses taken
18		Initial Major
19		Current Major
20		Current enrollment status
21		Teacher methodology
22	TEAHCER	Instructor grading
23		Instructor feedback
24		Instructor teaching methodology
25		Teacher name

3.3 Algorithms and Techniques

To evaluate our research by using different classification algorithms; Naïve Based and J48 has been compared. These classifiers are chosen on the basis of different reasons i-e; these classifiers supports Categorical and Nominal. For this purpose WEKA 3.8 tool has been used. The tool is open source available on the web and is generally used for machine learning algorithms, classification, training and testing of data. The data is then can be visualize in the form of graphs as well. In present research, the formed combinations of attributes have been evaluated to answer our research questions. For this purpose the data has been converted into csv format to import in the WEKA and the results from various filters and classifiers have been accumulated.

a) Naïve Based

Naïve based algorithm is comparatively fast algorithm in terms of classification. It works faster on huge datasets by using Bayes algorithm of probability. Bayes algorithm generally used to predict the class of unknown dataset⁴. Naïve based algorithm works on assumptions to label an item whose features are known but name is unknown. For example; a fruit is labeled as an apple if it is round and red in color and its size is 3 inches in diameter. These features of apple will raise the probability of this fruit that it is an apple.

b) J48

J48 decision tree is used to predict the target variable of new dataset. If dataset contains predictors or independent variables and set of target or dependent variables, then this algorithm is applied to extract the target variable of new dataset⁵.

c) Linear Regression

Generally linear regression classification algorithm is an approach to identify the relationship between dependent and independent variable. It is generally used for predictive analysis and has two main points. One is to check whether predictor variable does a good job in predicting the expected outcome variable.

⁴ <https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/>

⁵ <http://data-mining.business-intelligence.uoc.edu/home/j48-decision-tree>

Second main thing that linear regression does is the identification of variable that are significant predictors of dependent variables. At the end, the regression equation is used which helps to determine the set of predictor which are used to predict the outcome. In this research, algorithms are being used to compare the trend and pattern of the factors with other approaches like non-linear regression and SMO.

3.4 Evaluation of Research Questions

For our research, we have proposed three research questions that have been evaluated in multiple dimensions. Such dimensions have been explored in following ways respectively.

RQ1: Whether exiting identified factors for grade prediction are valid in our local context (Pakistan)?

Our research question number 1 determines the validity of factors in local context which have been carried out in foreign context by the researchers. The main reason that distinguishes this research in local context from the foreign context is that the norms, traditions and culture of every country that varies. Therefore it is needed to be evaluated that whether the existing factors that work in the colleges and universities of Canada, USA, England, and Nigeria etc, are applicable in Pakistan or not. After pre-processing of whole data, classification algorithms have been applied to analyze the behavior of demographic, pre-university and institutional factors. According to the results, it has come to known that these factors are also valid in Pakistan because of nearest difference in accuracies. Detailed results have been discussed in chapter no 4.

For our research, data being collected from term 141 to term 171 which consist of 667 students. Then three types of exiting attributes to predict the performance of students have been taken. First type of attribute is demographic; Gender, Age, Residence, Location, Race, Father's Qualification and have considered only Gender and Age of the student. Second type of attribute is Institutional; GPA(Term1, Term2, Term3, Term4, CGPA, Financial status, Total credit hours taken, Total courses taken, Initial Major, Current

Major, Current enrollment status, Teacher methodology and Midterm was considered, Term Marks for the prediction of final CGPA.

Third type of attribute is Pre-qualification; Secondary school grade, higher secondary grade, Pre-college, Pre-program, SAT score and Marks of FSC and MATRIC are considered. Among these attributes, specific attributes

During pre-processing, it has been found the total number of students who could not continue their program and left the course incomplete. After pre-processing, there were total 660 remained and 7 out of 667 who left the course. Two types of errors occurred in this dataset. The first was missing data and second was incomplete data which has been corrected in the phase of pre-processing.

In Table 3.5, Attributes selection to dataset for experiments has been shown

Table 3.5: Attributes selection to Dataset

Sr. No.	Gender	City	Matric	F.Sc	Mid ITC	Mid Eng-1	Mid Cal-1	Mid Physics	Studies	GPA
1	Male	Rawalpindi	51	64	10	10	16	9	16	2.51
2	Male	Islamabad	61	56	9	9	8	8	8	2.22
3	Male	Rawalpindi	51	64	7	7	13	17	13	3.15
4	Male	Islamabad	48	55	9	9	12	7	12	3.26
5	Male	Islamabad	77	58	12	12	17	18	17	3.07
6	Male	Islamabad	74	60	9	9	16	18	16	2.74
7	Male	Rawalpindi	84	70	6	6	12	7	12	2.86
8	Male	Islamabad	64	51	7	7	12	6	12	1.9
9	Female	Islamabad	61	56	8	8	11	6	11	1.83

RQ2: Which factors help in accurate prediction of student’s GPA of first semester?

Some specific attributes form all three factors; demographic, institutional and pre-university have been selected. The combinations of those attributes are constructed and then filters have been applied respectively. There are total 21 unique combinations of attributes presented in table 3.6.

Table 3.6: Pattern of Combination of Attributes

Sr. No.	ATTRIBUTES
1	Demographic, Pre-Qualification, Institutional (Result all subject of 1st semester)
2	Demographic, Pre-Qualification, Institutional (Result Eng-1, Cal-1, ITC, subject of 1st semester)
3	Demographic, Pre-Qualification, Institutional (Result Eng-1, Cal-1 subject of 1st semester)
4	Demographic, Pre-Qualification, Institutional (Result Eng-1 subject of 1st semester)
5	Demographic, Pre-Qualification, Institutional (Cal-1 subject of 1st semester)
6	Demographic, Pre-Qualification, Institutional (Result ITC, subject of 1st semester)
7	Demographic, Pre-Qualification, Institutional (Result Cal-1, ITC subject of 1st semester)
8	Demographic, Pre-Qualification, Institutional (Result Eng-1, ITC subject of 1st semester)
9	Demographic, Pre-Qualification, Institutional(GPA)
10	Demographic, Pre-Qualification
11	Demographic
12	Pre-Qualification
13	Pre-Qualification, Institutional (Result all subject of 1st semester)
14	Demographic, Institutional (Result all subject of 1st semester)
15	Institutional (Result Eng-1, Cal-1, ITC, subject of 1st semester)
16	Institutional (Result Eng-1 subject of 1st semester)
17	Institutional (Result Cal-1 subject of 1st semester)
18	Institutional (Result ITC, subject of 1st semester)
19	Institutional (Result Eng-1, Cal-1 subject of 1st semester)
20	Institutional (Result Cal-1, ITC subject of 1st semester)
21	Institutional (Result Eng-1, ITC subject of 1st semester)

RQ3. Is it possible to improve the performance of a student by allocating of appropriate teacher for a subject?

There is a fact that is normally faced during the research period is; the personality of teacher affects the performance of the students. If the background of the teacher is already known with the help of prediction system, it might become feasible for the department to recommend the best suitable teacher to the student up to his level of interest. And this phenomenon can surely boost the performance of the student. To answer the question number 3, experiments have been performed using two approaches have been used. And these approaches Statistical experiments and Predictive experiments.

a) The research has used combination of techniques comprising simple mean and the prediction based on teacher. The focus of this question is to find the most suitable teachers for Cal-I and ITP as these two subjects are considered the first semester with respect to the academic performance or attrition rate. Following calculations have been performed for the Cal-I teachers with respect to the prediction experiments:

1. Average of prequalification is taken
2. Average of actual Cal-I marks based on the Cal-I teacher
3. Average of actual Cal-II marks based on the Cal-I teacher
4. Prediction of the Cal-I and ITC/ITP is taken on the basis of prequalification
5. Predict Cal-II marks on the basis Cal-I and average is taken

b) ANOVA data analysis

ANOVA is a statistical technique which is used to measure the difference in a scale level dependent variable by a nominal variable which comprises of two or more types of categories. Normally ANOVA test is applied to find out the significant difference among groups. Similarly we will be applying this test to find the difference between the performances of teacher's subject wise in our

case of experiments⁶.

c) **Z-score data analysis**

Unlike ANOVA (analysis of variance), Z-score is usually applied on three or more means. A Z-score is a type of hypothesis test which is a way to find whether the results obtained from a test are valid or need to be repeated. For example, if someone said they had found a new drug to cure the cancer, one would want to be sure it was probably true. Similarly, in our research, we have applied Z-test to compare the performance of teachers particularly with respect to their subject. Z – Test will exploit the likelihood that the obtained results are true or not. A Z-test is generally used when the data is approximately normally distributed in the form of pairs⁷.

⁶ <http://www.statisticssolutions.com/manova-analysis-anova/>

⁷ <http://www.statisticshowto.com/z-test/>

Chapter 4

RESULTS & EVALUATIONS

Researchers have made their significant contributions to introduce such factors in the field of educational data mining that are quite beneficial for the educational institutions to mine hidden patterns from the academic database. Educational data mining is broadened field in which Online learning management system as well as physical institution's datasets are used. One limitation of this approach is to get the access of the dataset from the administration in order to perform analysis and conduct research. The dataset have been collected from Capital University of Science and Technology, Islamabad and applied existing factors on this dataset to answer our proposed research questions. Following sections of this chapter will present the overall picture of obtained outcomes in detail.

Data collection and pre-processing are the initial steps towards the analysis of the research. For this purpose, the demographic data has been collected from the registrar office of the CUST which contains student's gender and city. With the co-operation of administrative resources of CUST, the data related to pre-university and institutional factors has been collected. Institutional and pre-university factors contain the data about mid-term marks, term-marks, GPA, intermediate marks and matriculation mark. University portal provided privilege to access this data of the students to gather the sufficient dataset for our experiments.

The gathered data came into many formats such as PDF, word and excel. For the pre-processing, firstly this data set brought into the single file; Excel. Figure 4.1 represents the format of files and transformation of all files into single one. The collected data set arises in three formats; word, PDF and excel. After collection of dataset, it became mandatory to transfer the data into single file from all the files. The purpose of excel file is to import it into the WEKA to apply filters for the pre-processing. The final versions of dataset in the form of

excel is then converted into CSV format and imported in to the WEKA. Such filters have been applied that are discussed in subsections of section 4.1.

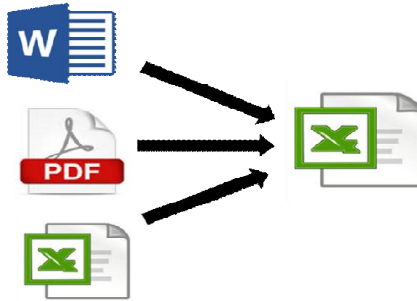


Figure 4.1: Formats of Data

In context of missing data, lacking attribute values, lacking certain attributes of interest, or containing only aggregate data instead of complete information are found. Incomplete information may lead the research towards in accurate outcomes that might affect the overall experiments. Therefore, the missing data problem has been handled in a way that no attribute remained null or incomplete.

Table 4.1: Occurrence of Missing Values

Sr. No.	Gender	ITC	Total	Weighted	Final Exam	Mid Term	Term Work	Teacher
1	Male	B-	75.95	76	25.2	15	21.67	D
2	Male	D	50.27	50	17.6	7.25	14.8	C
3	Male	C	66.93	67	22.8	8.25	21.65	A
4	Male							
5	Male	B-	76.43	76	21.6	13.75	27.27	D
6	Male	D	50.19	50	14.4	10.25	10.95	B
7	Male	A	89.93	90	34.8	14.25	30.57	D

The methods of data mining behave differently in the way that they treat missing values. Normally, they ignore the missing values, or exempt those records which contain missing values or either replace missing values with the mean, or conclude missing values from existing values. Missing Values Replacement Policies include the number of strategies such as; ignore the records with missing values. Following example shows the occurrence of missing values in the dataset. After the handling of missing data by taking the average of last three records of students and filled the missing one, complete dataset is gained. Further, this dataset has been used for experiments to evaluate all research questions.

Table 4.2: Handling of Missing Values

Sr. No.	Gender	ITC	Total	Weighted	Final Exam	Mid Term	Term Work	Teacher
1	Male	B-	75.95	76	25.2	15	21.67	D
2	Male	D	50.27	50	17.6	7.25	14.8	C
3	Male	C	66.93	67	22.8	8.25	21.65	A
4	Male	C+	72.91	73	19.2	18	27.75	A
5	Male	B-	76.43	76	21.6	13.75	27.27	D
6	Male	D	50.19	50	14.4	10.25	10.95	B
7	Male	A	89.93	90	34.8	14.25	30.57	D

Outliers are an observation point that is found to be distant from other observation points. An outlier may occur due to the change in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. After handling noise from the dataset, following result is obtained.

Attribute Selection

Appropriate selection of attributes is very important task in the process of pre-processing. It ultimately affects the performance of results and may lead to inaccurate results, if not chosen carefully. That's why we have classified the attributes into two types; Nominal and Categorical. On these two types of attributes researcher have selected the attributes that best suit to the nature of these two types of attributes.

Sampling

Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points in order to identify patterns and trends in the larger data set being examined. For example: term 141,143,151,153,161,163 dataset has been used as a training data and term 171 is test data because prediction of result of corresponding students is required to be computed.

4.1 Final Attribute Selection

From the literature, there are plenty of attributes found and some of them have been stated in the following table. This table belongs to the table presented in chapter no 3 but used in the current chapter with the addition of column named as selected attributes. These attributes have been marked to distinguish from each other. Then final attributes have been acquired for the experiments to fulfill the requirements of our results.

4.2 Result of Evaluation of Research Questions

With the help of adopted algorithms for the experiments in WEKA, the proposed research questions have been tried to answer. These answers make this research methodology validate and complete. Classification and linear regression has been selected to measure the analysis. The results of experiments have been discussed in detail in this section for each research question.

RQ1: Whether existing identified factors for grade prediction are valid in our local context (Pakistan)?

As existing factors along with their attributes have been stated in above sections, we have drawn the combinations of local factors with the existing factors. We found a slight variation between both types

of attributes. The comprehensive results of such variation have been arranged as the answer to each of three questions. The format of attributes has been presented in table 4.3 which are used in experiments.

Table 4.3: Attributes table

Sr. No.	Gender	City	Matric	FSc	Mid ITC	Mid Eng-1	Mid Cal-1	Mid Physics	Mid Pak-Studies	GPA
1	Male	Rawalpindi	51	64	10	10	16	9	16	2.51
2	Male	Islamabad	61	56	9	9	8	8	8	2.22
3	Male	Rawalpindi	51	64	7	7	13	17	13	3.15
4	Male	Islamabad	48	55	9	9	12	7	12	3.26
5	Male	Islamabad	77	58	12	12	17	18	17	3.07
6	Male	Islamabad	74	60	9	9	16	18	16	2.74
7	Male	Rawalpindi	84	70	6	6	12	7	12	2.86
8	Male	Islamabad	64	51	7	7	12	6	12	1.9
9	Female	Islamabad	61	56	8	8	11	6	11	1.83

As an input we have imported the .CSV into the WEKA that contains information about each attribute against every student. The input data contains the data of students of BSCS department from term 141 to term 163. The

dataset of CUST (Capital University of Science and Technology) has been used for this purpose. Among those factors, attributes like **Pervious** Course Marks/Grade, GPA, SSG, HSSG, Gender, City has been used. Afterwards, we have applied well-renown classification algorithms i-e; Naïve Bayes and J48 on the dataset.

Table 4.4: Classification of attributes and algorithms

Reference	Attributes	Class Label	Naïve Bayes	J48
Data mining approach for predicting student performance (2012)	Student Demographic, Per-Qualification, 1st semester 5 Subjects Grade	CGPA Grade	69.12%	72.39%
	Student Demographic, High School Background, Scholarship, Social network interaction	CGPA Grade	76%	73%
Comparison of Classification Techniques for predicting the performance of Students Academic Environment(2014)	Midterm marks all subjects	GPA Grade	65.6%	80.98%
	Internal assessment, Extra-Curricular activities	GPA Grade	66%	73%
Predicting Student Performance: A Statistical and Data Mining Approach(2013)	Student Demographic, Per-Qualification,	CGPA Grade	49.90%	63.19%
	Student Demographic, High School Background	CGPA Grade	50%	65%

In table 4.4, we have summarized the attributes of each factor with corresponding to their accuracies. These attributes have been assigned with class label and the reference of each attribute has been given as well. During the literature review, we have found the accuracy of each attribute in their respective scientific research publication and compare the accuracy of our

results. The inferences show the slight variations which is quite considerable with the correspondence to our research. The existing attributes have been shown in black color in the summary table and attributes in red color are local attributes.

The summarized results have been shown in the figure 4.2 in which bars of different colors on X-axis shows the presence of attributes and y-axis shows the interval of accuracy.

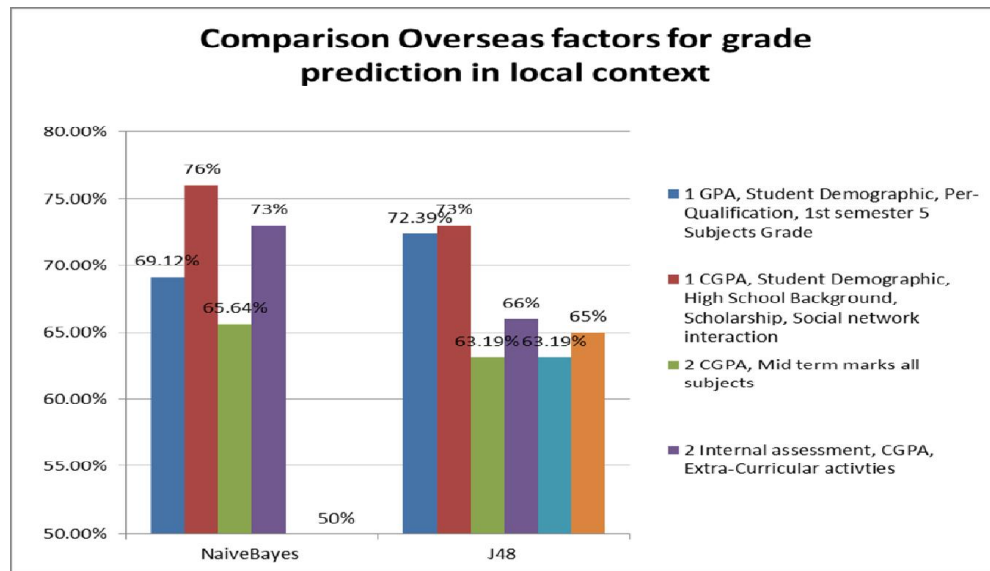


Figure 4.2 Comparison of local and Existing attributes

Our Findings narrate that existing identified factors for grade prediction are considered to be comparable with the local factors because of low variation in the result of their accuracies. According to the results, Previous Grade, Internal Marks, HSSC were found to be effective for grade prediction.

RQ2. Which factors help in accurate prediction of student’s GPA of first semester?

Our second research question is the extraction of those factors that helps in accurate prediction of GPA of students belong to 1st semester. To answer the research question, we have transformed this question into following two ways.

1. Based on Mid term marks

In this phase of experiment, we have combined the attributes including marks of mid term to predict the GPA of the first semester. our prediction leads to the positive focal point where our proposed combinations worked well.

2. Based on Pervious Grade

In this phase of experiment, we have interpreted our attributes to predict the CGPA of the students. The CGPA of 2nd semester has been computed on the basis of previous academic record.

After compilation of attributes, their accuracies have been ranked that is presented in table 4.5 is descending order.

Information Gain Ranking Filter

Table 4.5: Ranked Attributes

Ranked Attributes	
0.2277	Cal-I MID
0.1903	Cal-II MID
0.1778	ITC MID
0.1684	Physics MID
0.1548	Pak-Studies MID
0.1514	FSC
0.1006	City
0.0992	Matric
0.0791	Gender

According to results, picture states the prediction of CAL-I MID and CAL-II MID remained better in terms of accuracy. Then the accuracies gradually decreases from ITC MID to GENDER. Based on the corresponding accuracies of attributes, we may inference the importance of acquired ranking information gain of all the instances.

1- First Semester GPA Prediction

We have expanded our answer for the research question number 2. With the help of “GPA Prediction for the first semester” we became able to evaluate our results through WEKA. We covered the list of attributes narrated in table 4.6 in which the combination of attributes are defined in column 1. There are four combinations of attributes which belong to the three factors; Pre-Qualification, Demographic and Institutional. The performance of J48 is considerably better than Naïve Bayes that has been shown in table as well as presented in figure 4.3.

Table 4.6: GPA Prediction for the first semester

Attributes	Class Label	Naïve Bayes	J48
Pre-Qualification, Institutional (Internal marks only Midterm) all subject of 1st semester	GPA Grade	64.62%	81.39%
Institutional (Internal marks only Midterm all subject of 1st semester)	GPA Grade	65.64%	80.98%
Demographic, Pre-Qualification, Institutional (Internal marks only Midterm all subject of 1st semester)	GPA Grade	66.46%	79.55%
Demographic, Pre-Qualification, Institutional (Internal marks only Midterm Eng-1, Cal-1 ,ITC subject of 1st semester)	GPA Grade	69.33%	78.12%

According to the results, the performance of Demographic, Pre-Qualification, Institutional (Internal marks only Midterm Eng-1, Cal-1, ITC subject of 1st semester) remained better as compared to the other three combinations of factors when Naïve Bayes classifier applied. On applying J48 classifier, the performance of Pre-Qualification, Institutional (Internal marks only Midterm) all subjects of 1st semester remained better comparatively to predict the grades of first semester.

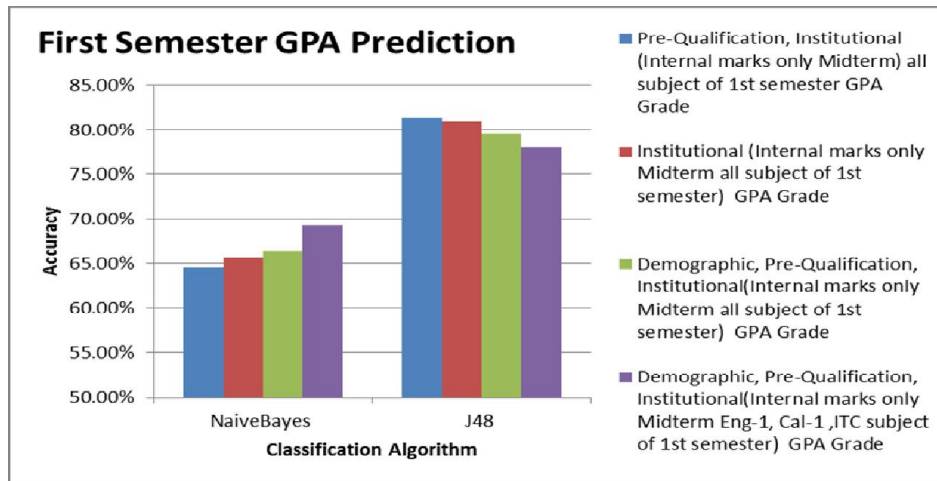


Figure 4.3: GPA Prediction for the 1st semester

2- CGPA of second Semester

To predict the CGPA of 2nd semester, we have defined the following narrated set of experiments as summarized in table 4.7. We have merged all the grades of previous semester along with the adopted factors. These combinations and their accuracies on the basis of Naïve Bayes and J48 have been presented in table 4.4.

Table 4.7: CGPA Prediction for the second semester

Attributes	Class Label	Naïve Bayes	J48
Demographic, Pre-Qualification, Institutional (Subject Grade Cal-1 subject of 1st semester)	CGPA Grade	67.08%	74.44%
Demographic, Pre-Qualification, Institutional (Subject Grade ITC subject of 1st semester)	CGPA Grade	68.10%	74.44%
Demographic, Pre-Qualification, Institutional (Subject Grade Cal-1, ITC subject of 1st semester)	CGPA Grade	70.55%	73.82%
Demographic, Pre-Qualification, Institutional (Subject Grade Eng-1, Cal-1 subject of 1st semester)	CGPA Grade	65.44%	72.60%

These tabular results have been presented as figure 4.4. Like 1st semester GPA prediction, we have used same classifiers for the CGPA prediction for 2nd semester. The only change lies in the combinations of attributes. According to the accuracy of each combination, we reached to the point that, the accuracy of Demographic, Pre-Qualification, Institutional (Subject Grade Cal-1, ITC subject of 1st semester) was high whereas when J48 classifier was applied, the prediction of Demographic, Pre-Qualification, Institutional(Subject Grade Cal-1 subject of 1st semester) and Demographic, Pre-Qualification, Institutional(Subject Grade ITC subject of 1st semester) remained better than other two combinations with the minor difference i-e; Demographic, Pre-Qualification, Institutional (Subject Grade Cal-1, ITC subject of 1st semester) and Demographic, Pre-Qualification, Institutional (Subject Grade Eng-1, Cal-1 subject of 1st semester).

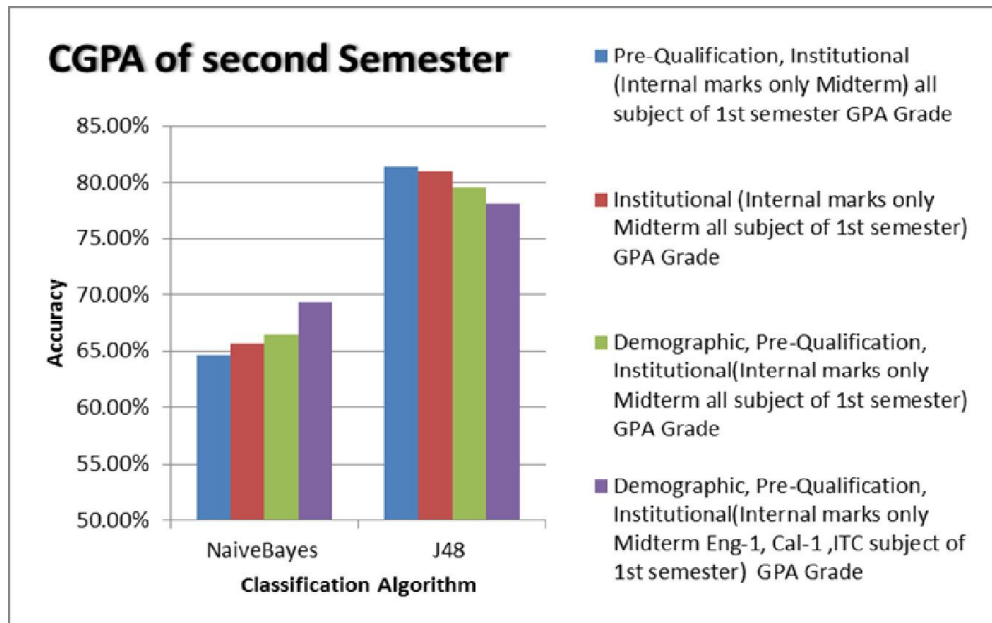


Figure 4.4: CGPA Prediction for the second semester

Our Findings enlighten the following assumptions.

- For Grade Predictions, the Attribute Evaluator has computed the Information Gain and ranked the accuracy of each attribute accordingly.

Then it is found that the subjects of First semester i-e; Cal-1, Eng-1, ITC have high value of information gain and played an effective role in order to predict the accurate grades of students.

- As we came to know from the situation of dataset that the Attrition rate is high in first semester. The purpose of the research is to reduce the attrition rate, so we have considered these subjects to predict the attrition rate of students as well as student's performance.
- Results states that Demographic attributes does not play significant role in grade prediction of students.
- In addition, internal marks like Mid Terms are used for early grade prediction that has high accuracy and also helps to minimize the attrition rate.

RQ3. Is it possible to improve the performance of a student by allocating of appropriate teacher for a subject?

The RQ3 has been addressed with a combination of techniques comprising simple mean and the prediction based on teacher. As it has been discussed in the precious chapter that the focus of this question is to find the most suitable teachers for Cal-I and ITP as these two subjects are considered to be critical for the first semester with respect to the academic performance or attrition rate. To evaluate the research question no 3, we have adopted the following set of experiments. Initially, the dataset has been imported into WEKA that contains the information of students from term 141 to term 163. The set of attributes consist of the Pervious Core and Elective Course Internal Marks, GPA, SSG, HSSG, Gender, City. After finalizing the set of attributes, we have applied following proposed techniques.

- **Predictive Approach:** Classification algorithm: Linear Regression
- **Statistical Approach:** Simple average , Z-Score analysis , ANOVA data analysis

The experiments have been conducted subject wise for the individual course grade prediction. On the basis of course grades, accumulated teacher

performance has been computed. Furthermore, following calculations have been performed for the Cal-I teachers:

1. Average of prequalification is taken
2. Average of actual Cal-I marks based on the Cal-I teacher
3. Average of actual Cal-II marks based on the Cal-I teacher
4. Prediction of the Cal-I is taken on the basis of prequalification
5. Predict Cal-II marks on the basis Cal-I and average is taken

Pre-Qualification vs Maths Courses

The comparison of Maths course and Pre-qualification has been made in this phase of experiment. The teachers have been labeled alphabetically and the overall average of the students has been given against every teacher in table 4.8.

Table 4.8: Pre-Qualification VS Maths Courses

Teacher	Pre-Qualification average	Cal-I average	Cal-II average
A	58.37142857	58.7429	64.4269
B	61.70231214	64.8324	68.03595
C	59.92207792	57.91	66.5098
D	60.17073171	56.1951	64.18938
E	58.96428571	73	72.75
F	58.95588235	57.2059	54.42167
G	60.73512748	61.8491	63.43205
H	59.82369942	61.3628	68.46664

The result shows that the performance of teacher “B, D and E” is found to be better in Pre-qualification combination. With respect to the Cal –I performance of teacher “B, G and H remained better. On using the Cal-II attribute, the performance of teacher “B, C, and D” found to be better than other teachers. These results have been shown graphically in figure 4.5.

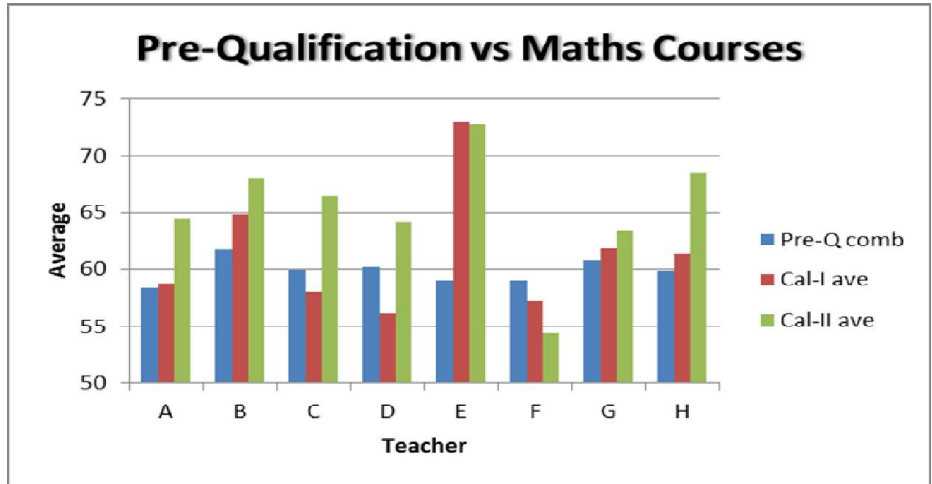


Figure 4.5: Pre-Qualification vs Maths Courses

Average of Cal-II Teachers

In this case, we have predicted the the performance of teahcers who teach Cal-II and compared with the actual performance of the teachers in Cal-II. According to the results, we have found significantly unexpected outcomes. There is a minor variation between actual Cal-II and predicted Cal-II. The results have been narrated in table 4.9 and presented in figure 4.6.

Table 4.9: Average of Cal-II teachers

Teacher	Cal-II ave	Pred-Cal-II ave
A	64.42689655	62.16017221
B	68.03594595	69.39823039
C	66.50980392	63.70893212
D	64.189375	64.42028569
E	72.75	71.255395
F	54.42166667	56.0806705
G	63.43205128	67.83774562
H	68.46663551	67.88178679

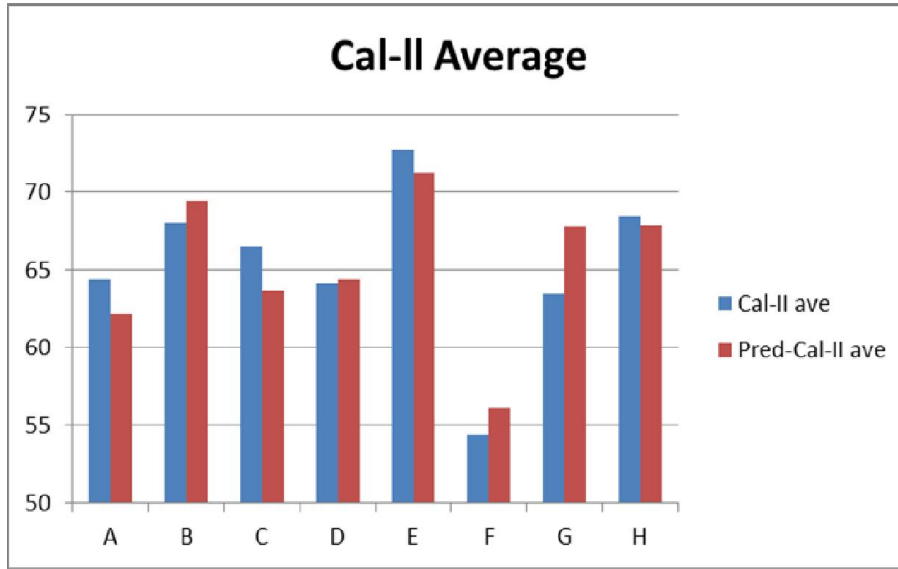


Figure 4.6: Average of Cal-II teachers

Average of Cal-I teachers

The average of Cal-I teachers has been compared with the results of actual and predicted values. There is another attribute “Pre-qualification” that has been used in this case. The obtained results have been given in the table 4.10 as well as presented in figure 4.7.

Table 4.10: Average of Cal-I teachers

Teacher	Pre-Q comb	Pred-C1-Preq	Cal-I ave
A	58.37142857	60.6696821	58.7429
B	61.70231214	61.50918169	64.8324
C	59.92207792	61.91539229	57.91
D	60.17073171	59.51882925	56.1951
E	58.96428571	60.379367	73
F	58.95588235	59.785871	57.2059
G	60.73512748	63.12452977	61.8491
H	59.82369942	61.70135392	61.3628

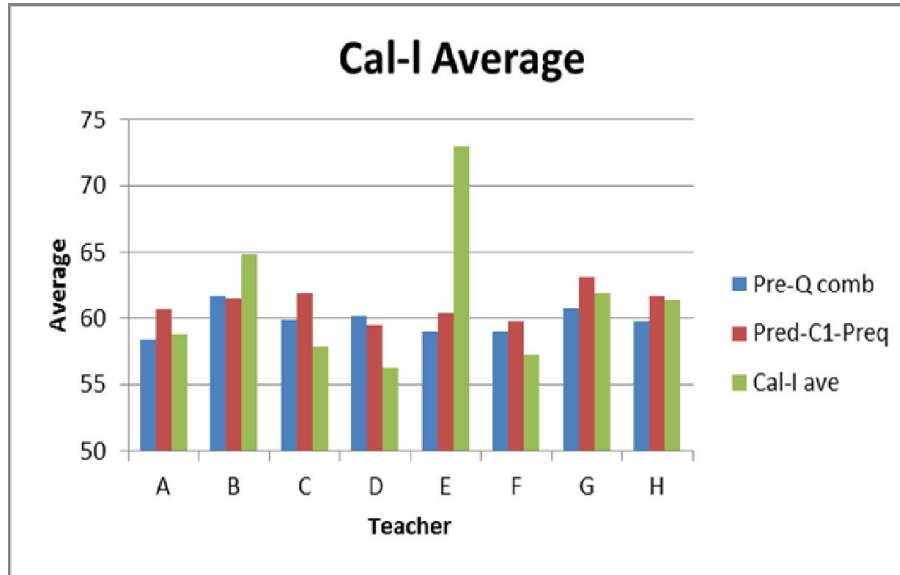


Figure 4.7: Average of Cal-I teachers

Average of Maths Teachers

Like cal-I and Cal-II the overall performance of math’s teachers have been computed against each teacher who teaches Cal-I and Cal-II. The information in table 4.11 narrates that the performance of teacher “B and E” remained better than other teachers who teach same subjects. This information has been presented in figure 4.8.

Table 4.11: Overall Average of Maths Teachers

Teacher	Over All Performance
A	60.87421589
B	65.09561403
C	61.99324125
D	60.89886433
E	67.26980954
F	57.2899981
G	63.39571083
H	63.84725513

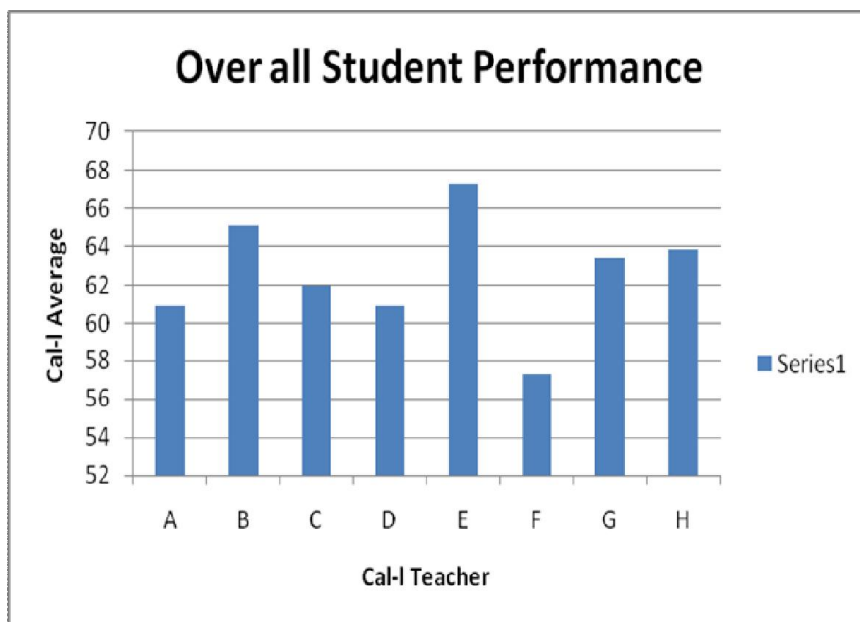


Figure 4.8: Overall Average of Maths Teachers

After the computation of results of Cal-1 and Cal-II, we became able to find the teacher “E” whose performance found to be better and able to recommend teaching Cal-I and Cal-II. Similarly, we will compute the results of other subjects which are to be taught in the 1st semester and then computed similarly. Firstly, researchers have computed the performance of every teacher of each subject which is being taught in first and second semester and gained the results. The results have been interpreted for every subject with respect to its related teacher. These results have been shown significant average of all students in Cal-1 & Cal-2 with Cal-1 teacher name. After the computation of results for Cal-1 and Cal-II, it became clear to find that the teacher “E” whose performance found to be better and able to recommend teaching Cal-I and Cal-II then teacher B have better performances and further teacher B and H performs better among all other teachers.

Further the average of these teachers will be checked whether it is significant or not significant by applying statistical test “ANOVA”.

ANOVA data analysis on Cal-I

ANOVA analysis is used for group data we have many teacher for Cal-1 & Cal-2 subjects. The prediction shows that teacher average performance is

significant. To prove these results, we applied ANOVA analysis on this dataset.

We have applied ANOVA (Analysis of Variance) test on teacher's average. By this way, results indicate the performance of all Cal-I teacher on the basis of group.

Table 4.12: ANOVA Analysis on Cal-1

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Row 1	5	303.5967	60.71935	15.95621		
Row 2	5	330.7894	66.15789	3.27359		
Row 3	5	313.1846	62.63691	19.02353		
Row 4	5	292.9438	58.58876	27.08998		
Row 5	5	357.9515	71.5903	2.338546		
Row 6	5	274.268	54.85359	20.50589		
Row 7	5	313.5665	62.7133	1.073095		
Row 8	5	318.981	63.79621	18.29478		
ANOVA						
Source of Variation	SS	Df	MS	F	P-value	F crit
Between Groups	872.7731	7	124.6819	9.273853	3.14E-06	2.312741
Within Groups	430.2225	32	13.44445			
Total	1302.996	39				

In this case $F > F \text{ crit}$ has been shown as $9.273853 > 2.312741$. Researchers reject the null hypothesis and concluded that there are significant differences between the methods because all 11 methods don't have the same mean.

Z-Score Data Analysis

Now Z-score test is applied to evaluate the performance of every pair of combination of teacher's subject wise. Generally, Z-test is applied on two populations to compare its proportion. In this case, we are supposed to compare the performance of teachers of the same course which is to recommend to the student. Researchers have classified each Z-test with respect to the subject and then have considered the results of Z-score with respect to Cal-1. All the teachers who teach Cal-I have been indicated with alphabetic characters i-e; A, B, C, D, E, and F. Pairs such as teacher A with other teachers of Cal-1 are made for experiment. Moreover, for the comparison, teacher's data has been classified as same number of students and different number of students.

Table 4.13: Two Samples for Mean with respect to teacher “E”

z-Test: Two Sample for Means		
	<i>Variable</i> <i>1</i>	<i>Variable</i> <i>2</i>
Mean	74.94286	58.74286
Known Variance	102.1429	333.391
Observations	15	35
Hypothesized Mean Difference	0	
Z	4.008258	
P(Z<=z) one-tail	3.06E-05	
z Critical one-tail	1.644854	
P(Z<=z) two-tail	6.12E-05	
z Critical two-tail	1.959964	

Z-score data Analysis applied on the Cal-1 teacher “E” with Teacher “B” on the basis different number of student. Value of Z-critical to tail < Z concluded that the performance of teacher “E” is not significant.

Then teacher “B” has been paired with teacher “A” on the basis same number of students. The value of Z-critical > Z concluded that significant performance

of teacher “A”. The results have been shown in table 19. Z-test has been applied in both cases and results have been compared either the performance of teachers remain same or vary. By this way, results indicate the performance of every teacher of Cal-I on the basis of combinations. The result of z-test will evolve around this statement “If $z < z$ critical two tail then performance difference is not significant”. If $z > z$ critical two tail then performance difference is significant.

Table 4.14: Two samples for means with respect to teacher “B”

z-Test: Two Sample for Means		
	<i>Variable</i> <i>1</i>	<i>Variable</i> <i>2</i>
Mean	74.94286	58.74286
Known Variance	102.1429	333.391
Observations	15	35
Hypothesized Mean Difference	0	
Z	4.008258	
P(Z<=z) one-tail	3.06E-05	
z Critical one-tail	1.644854	
P(Z<=z) two-tail	6.12E-05	
z Critical two-tail	1.959964	

Here, researchers have applied Z-score data Analysis on the Cal-1 teacher “B” with Teacher “A” on the basis different number of students who enrolled in different terms from term 141 to term 171. In this case, value of Z-critical two tails is greater than the value of Z which shows that performance of teacher “B” is not significant.

Then teacher “B” has been paired with teacher “A” on the basis same number of students and Z-test has been applied. The results shows that value of z-critical shows the significant performance of teacher “A” as it is greater than the value of Z.

Teacher B with different and same number of Students

Table 4.15: Teacher B with same number of Students

Cal-1 Teacher	Z	z Critical two-tail
A	4.008258	1.959964
C	3.370917	1.959964
D	5.288815	1.959964
F	4.463791	1.959964
G	4.119857	1.959964
H	2.997759	1.959964
E	3.630363	1.959964

Z-score data Analysis applied on the Cal-1 teacher “B” with all Teachers on the basis different number of student. The value of Z-critical tow tail $> Z$ conclude that performance of teacher “E” is not significant with Teacher “A”, “C”, “D”, “F”, “G”, “H”, “E”

Table 4.15: Teacher B with different number of Students

Cal-1 Teacher	Z	z Critical two-tail
A	3.687091	1.959964
C	1.235961	1.959964
D	3.480026	1.959964
F	2.829936	1.959964
G	3.218075	1.959964
H	1.61421	1.959964
E	1.271677	1.959964

Z-score data Analysis applied on the Cal-1 teacher “B” with all Teachers on the basis same number of student. The value of Z-critical tow tail $> Z$ conclude that performance of teacher “B” is not significant with Teacher “A”, “B”, “C”, “D”, “F”, and significant with Teacher “G”, “H”. The ANOVA test has been applied on the cal-I only. The results have been computed for the rest of the courses similarly.

Pre-Qualification vs Programming Courses

In table 4.17, we have narrated the results of all programming courses which are taught in 1st and 2nd semester. These averages of courses have been compared with the Pre-Qualification.

Table 4.17: Comparison of all programming courses

Teacher	Pre-Q comb	ITC /ITP Simple Average	OOP/CP Simple Average
A	57.170543	61.48148148	65.27777778
B	61.4620758	53.7037037	62.80851064
C	60.99346693	65.21428571	69.5
D	61.04028149	61.81578947	72.39215686
E	60.28896102	71.40625	67.91666667
F	59.72643733	77	68.58333333
G	59.72643733	66.1871345	66.73484848
H	61.887987	66.9375	59.07142857
I	59.69578266	64.16129032	68.7826087
J	59.48424926	62.55882353	70.68421053
K	58.0787338	67.575	63.72222222
L	60.18534971	67.44736842	68.3125

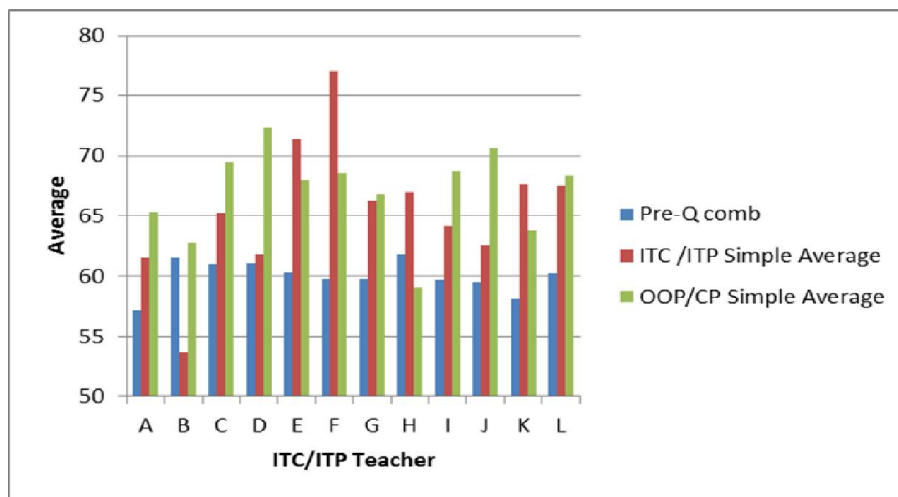


Figure 4.9: Comparison of ITC/ITP teacher

According to the figure 4.8, the performance of all of the teachers has been

computed. The performance of teachers with respect to courses average changes.

Average of ITC/ITP & CP/OOP Student

The average of students who studies ITC/ITP has been computed against every teacher who teaches these subjects. Then these results have been shown in table 4.18 and presented in figure 4.10.

Table 4.18: Comparison of Actual and Predicted Grades of ITC/ITP

Teacher	Pre-Qualification average	Predict ITC /ITP average	ITC /ITP Simple average
A	57.170543	60.90678396	61.48148148
B	61.4620758	63.21581404	53.7037037
C	60.99346693	61.1117056	65.21428571
D	61.04028149	68.16813274	61.81578947
E	60.28896102	70.68016263	71.40625
F	59.72643733	69.13777079	77
G	59.72643733	63.65621403	66.1871345
H	61.887987	72.60489569	66.9375
I	59.69578266	65.2340331	64.16129032
J	59.48424926	70.26083656	62.55882353
K	58.0787338	61.80150053	67.575
L	60.18534971	65.17485658	67.44736842

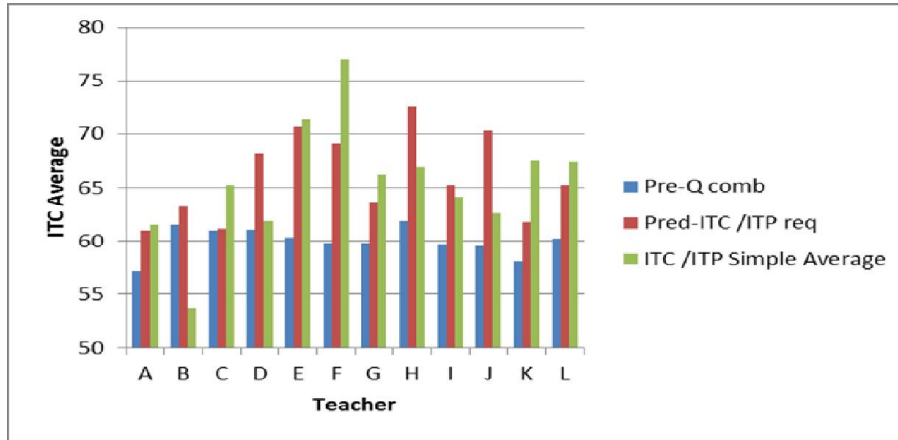


Figure 4.10: Comparison of Actual and Predicted Grades of ITC/ITP

Then the actual and predicted OOP/CP grades have been predicted against every teacher. The results shows that there is a slight variation in the results that has been stated in table 4.19 and presented in figure 4.11

Table 4.19: Comparison of Actual and Predicted Performance of OOP/CP Teachers

Teacher	OOP/CP ave	Pred- OOP ave
A	65.27777778	65.99605885
B	62.80851064	58.81069353
C	69.5	67.8661911
D	72.39215686	66.23364694
E	67.91666667	72.16535104
F	68.58333333	73.23220667
G	66.73484848	67.4231563
H	59.07142857	65.26328164
I	68.7826087	67.14102048
J	70.68421053	67.37969774
K	63.72222222	66.97798275
L	68.3125	66.77153434

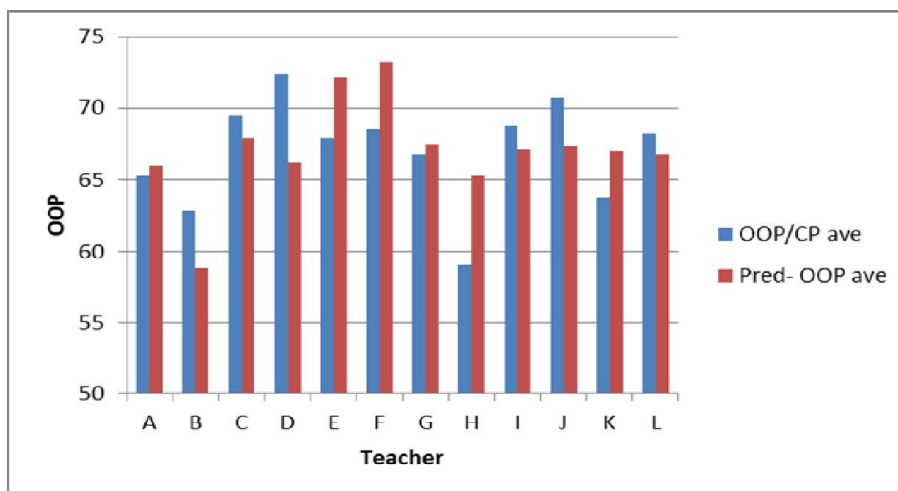


Figure 4.11: Comparison of Actual and Predicted Performance of OOP/CP Teachers

The overall performance of OOP teachers have been presented in table 4.20 as well in figure 4.12.

Table 4.20: Overall Performance of OOP Teachers

Teacher	Over All Performance
A	62.16653
B	60.00016
C	64.93713
D	65.93
E	68.49148
F	69.53595
G	64.74556
H	65.15302
I	65.00295
J	66.07356
K	63.63109
L	65.57832

The results show that the performance of teacher “E & F” is considerably high and better than other teachers. Therefore, based on the acquired results, we may recommend these two teachers to teach the courses in respective semesters.

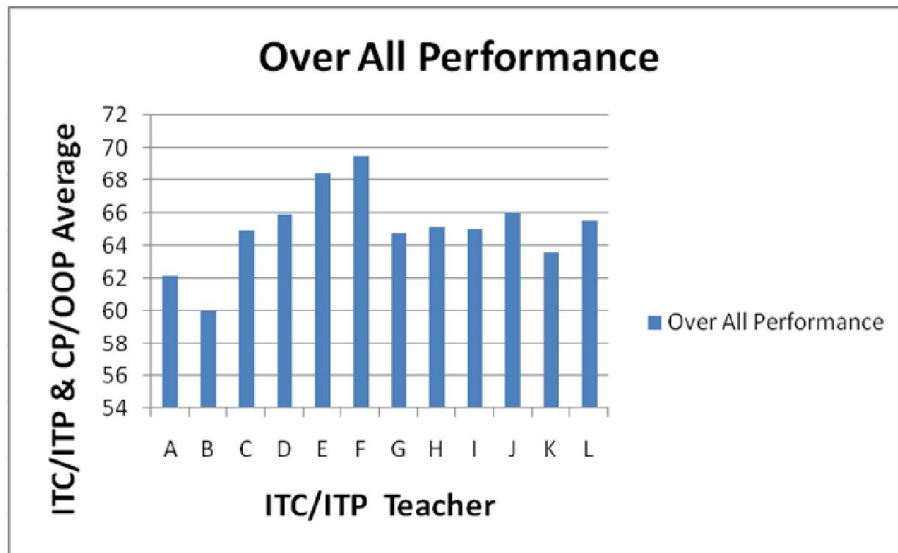


Figure 4.12: Overall performance of OOP Teachers

These results have been shown significant average of all students in ITC/ITP & CP/OOP with ITC/ITP teacher name. During the computation of results of ITC/ITP and CP/OOP, we have found teacher “F” performance better and able to recommend teaching ITC/ITP and CP/OOP.

ANOVA data analysis on ITC/ITP

ANOVs analysis is used for group data we have many teacher for Cal-1 & Cal-2 subjects. The average prediction shows that teacher average performance is significant. To prove this result applies ANOVA analysis on this dataset.

If $F > F_{crit}$, we reject the null hypothesis. This is the case $8.542703 > 1.952212$. Therefore, we reject the null hypothesis. The means of the 11 populations are not all equal. At least one of the means is different. Hence we conclude that there are significant differences between the methods (i.e. all 11 methods don't have the same mean). However, the ANOVA does not tell you where the difference lies. Therefore it is preferred T-Test, Z-Score test each pair of means.

Table 4.21: ANOVA Data Analysis on ITC/ITP

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Row 1	6	370.458	61.743	9.09211		
Row 2	6	346.3743	57.72906	16.59501		
Row 3	6	404.7156	67.4526	4.601526		
Row 4	6	397.287	66.21451	25.45386		
Row 5	6	421.2417	70.20696	3.492003		
Row 6	6	441.7835	73.63059	19.18218		
Row 7	6	400.4038	66.73397	0.377307		
Row 8	6	382.6283	63.77138	24.15625		
Row 9	6	394.865	65.81084	6.082654		
Row 10	6	392.0845	65.34741	12.9729		
Row 11	6	382.7968	63.79946	15.06479		
Row 12	6	407.9563	67.99271	0.222976		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1075.12	11	97.73818	8.542703	7.87E-09	1.952212
Within Groups	686.4678	60	11.44113			
Total	1761.588	71				

Z-Score data analysis on ITC/ITP

As discussed earlier, Z-test is a measure to analyze the performance of two attributes which have either equal or different number of population. Here, we

have applied Z-test on the combinations of teachers who teach ITC/ITP. We have made combinations similarly as we made for the teachers of Cal-1 and Cal-II. In table ..., Z-test has been applied on each pair of combination and tested on two cases. First is to compare the performance of teachers on same number of students and second is to compare the performance of teachers on different number of students. Z-score data Analysis ITC/ITP has been applied on the pair in which performance of teacher “F” with Teacher “A” on the basis different number of students has been compared. The results show that there is significant performance of teacher “F” with teacher “A”.

Table 4.22: Z-Score Data Analysis Teacher F with A different number of Students

z-Test: Two Sample for Means		
	<i>Variable 1</i>	<i>Variable 2</i>
Mean	77	61.48148
Known Variance	260.2985	278.7682
Observations	67	27
Hypothesized Mean Difference	0	
Z	4.116767	
P(Z<=z) one-tail	1.92E-05	
z Critical one-tail	1.644854	
P(Z<=z) two-tail	3.84E-05	
z Critical two-tail	1.959964	

Z-score data Analysis has been applied on the pair of ITC/ITP teacher “F” with Teacher “A” on the basis same number of students and results show that there is a significant performance between teacher “F” and teacher “A”.

Table 4.236: Z-Score Data Analysis Teacher F with A same number of Students

z-Test: Two Sample for Means		
	<i>Variable 1</i>	<i>Variable 2</i>
Mean	76.81481	61.48148
Known Variance	283.7064	278.7682
Observations	27	27
Hypothesized Mean Difference	0	
Z	3.359441	
P(Z<=z) one-tail	0.000391	
z Critical one-tail	1.644854	
P(Z<=z) two-tail	0.000781	
z Critical two-tail	1.959964	

As computed earlier, Z-score data Analysis ITC/ITP teacher “F” with other teachers who teaches same subjects on the basis different number of students has been computed. According to the summarized results, there is significant performance among teacher “F” with all of the teachers.

As computed earlier, Z-score data Analysis ITC/ITP teacher “F” with other teachers who teaches same subjects on the basis of same number of students has been computed. According to the summarized results, there is significant performance among teacher “F” with all of the teachers.

Table 4.24: Z-Score Data Analysis Teacher F with other different number of Students

Teacher F	Z	Z Critical two-tail
A	4.1168	1.96
B	7.0827	1.96
C	3.0686	1.96
D	5.4561	1.96
E	3.1293	1.96
G	4.7683	1.96
H	3.8317	1.96
I	3.8953	1.96
J	3.4076	1.96
K	3.3785	1.96
L	3.4421	1.96

Table 4.25: Z-Score Data Analysis Teacher F with other same number of Students

Teacher F	Z	z Critical two-tail
A	3.3594	1.96
B	5.6999	1.96
C	2.4668	1.96
D	5.2634	1.96
E	1.323	1.96
G	3.9718	1.96
H	2.4788	1.96
I	3.1042	1.96
J	2.8519	1.96
K	2.6697	1.96
L	2.3868	1.96

Summary

To prove our point, whether student's performance or grades are dependent on teacher's performance or not, we have taken an example to demonstrate. We have selected the course cal-1 to justify our analysis. For this task, we have taken courses; Cal-I & Cal-II and ITC/ITP & CP/OOP. Using different approaches for on the basis these results, we have ended with following assumptions.

There is no significant difference between actual and predicted average of student's grades on the basis of teacher's combinations we discussed.

When Cal-1 and ITC/ITP is taught by different teacher and get average marks of students from session 141 to 163 in Cal-1 and ITC/ITP teacher-wise using five different experiments, like : Average of actual Cal-I marks based on the Cal-I teacher, Prediction of the Cal-I marks excluding the teacher name and then taking the average teacher-wise, Prediction of the Cal-I marks including the teacher and then taking the average teacher-wise , Average of actual Cal-II marks based on the Cal-I teacher , Prediction of Cal-II marks and then computed the average of these marks based on the Cal-I teacher, our predicted results were higher than the actual results or nearly lesser than actual ones. The result has showed that all teachers have different average in these subjects and performances of teachers in this subject are significant.

For proved these result have used ANOVA analysis .The performance of teacher in CAL-I and ITC/ITP have different and average of teachers have significant between all teacher .The teacher whose performance is better might have a reason that he is lenient in giving grades to the students and the teachers whose performance is comparatively low might be strict in giving grades to the students. Z-Score have proved these results and have showed that significant result in CAL-I and ITC/ITP according to each teacher. With the help of these result, we have recommended for CAL-I teacher E is a very suitable teacher and ITC/ITP teacher F can improve the performance of students.

We have proposed a recommender system, which will recommend teachers to the students based on their interest level and understanding and communication

with their teachers. With the help of recommender system, university department may become able to identify the student's performance and enable them to make their performance better. To evaluate our prediction model and recommender system, we have produced the answers of research questions. We have found following strong observations during the course of our experiments.

- In context of CUST university Cal-I and ITC are critical subjects in semester-1
- Recommendation of appropriate teacher for Cal-I is B and G
- Recommendation of appropriate teacher for OOP is E and F

Chapter 5

CONCLUSION AND FUTURE WORK

- Prediction is a helpful activity to reduce the attrition
- Existing factor is applicable to the certain level of accuracy and a set of attributes is identified which is applicable in context of our university
- In context of CUST university Cal-I and ITC are critical subjects in semester-I
- Recommendation of appropriate teacher for Cal-I and ITC

Our proposed research has been distributed in five chapters. First chapter consist of introduction of the topic and purpose of this research. Second chapter contains literature review in which the related work of the other researchers has been presently in quite understandable manner. Third chapter contains methodology and chapter no 4 contains detailed discussion of results and experiments.

In fist chapter, we have briefly explained the background of our research topic along with its scope, significance and applications. Moreover, we have discussed research question, which we have constructed after critically reviewing of literature. We have explored various data mining techniques that were proposed and used by many researchers. The area of data mining is quite versatile. Data mining means to explore intrinsic information from the bulk of datasets. These datasets might belong to educational institutions or internet or any organization. With the help of mining techniques such as classification, we can make data analysis for prediction and find the accuracy. Researchers have worked a lot in the field of education in data mining in which they have used student's academic record for specific reasons. They have used the records for grade prediction, evaluation of student's performance and teacher's performance most commonly.

In this regard, from literature we have obtained exiting factors such as institutional factors and demographic factors. These factors are known as exiting because they have applied in the western universities. We have applied these factors in our local context. We have collected data from the Capital University of Science and Technology, Pakistan. We have used the data of BSCS program from six terms of spring, which covers the data of 3 years. In this dataset, we have used the data of students of first semester and their final grade has been predicted based on their midterm marks, term marks, gender, age, Matric grades and inter grade. These attributes cover demographic, institutional and pre-university attributes of the students. In addition, researchers have used only two factors; Demographic and Pre-university. In our experiments, we have considered another factor which is institutional in which midterm and term marks are combined to predict the students final GPA/Grade. This is how we have proposed prediction model on the basis students associated attributes.

After discussion of literature review critically, we have identified different classification approaches by the researches, so we used in our experimental work as well. Afterwards, we have discussed chapter number 3 in which diagram of proposed methodology have been mentioned and each step of methodology has been discussed in detail. Our chapter number 4 contains relative experiments and results that we have carried out through the help of WEKA tool. WEKA is specifically designed to perform analysis in the field of educational data mining. In the part of conclusion we have summarize the results against our each research question.

We have proposed a recommender system, which will recommend teachers to the students based on their interest level and understanding and communication with their teachers. With the help of recommender system, university department may become able to identify the student's performance and enable them to make their performance better. To evaluate our prediction model and recommender system, we have produced the answers of our three research questions. Those research questions and their answers have been summarized as follows:

RQ1. Whether existing identified factors for grade prediction are valid in our local context (Pakistan)?

The answer to this research question is yes because we have gained quite considerable accuracy from all data analysis measures. These analyses have been applied on categorical and nominal data. Results in both categories are same, which shows the validity of, existing factors in local context. All three factors contributed well in order to increase the accuracy through WEKA tool when ARPP format of data was given to it and classification algorithms; SMO and Linear regression were applied.

RQ2. Which factors help in accurate prediction of student's GPA of first semester?

We have given answer to this research question in quite interesting fashion. We have make combinations of all attributes that belong to these three categories of factors. When we combined term marks and mid-term grades along with inter and matriculation marks as well as with their gender and age; accuracy of algorithms were boost up to the mark able level. Therefore the answer to this research question is that demographic, pre-university and institutional factors (GPA, Internal marks, #of Credits Hours, Program). Based on these results, we have proposed a prediction model that will help to predict the final GPA/Grade of the students and then can be evaluated in other departments of the university too. There is not a single attribute that is affecting the performance of students but a combination of attributes that also occurs in terms of hybrid approach.

RQ3. Is it possible to improve the performance of a student by allocating of appropriate teacher for a subject?

To answer the research question no 3, we have drawn inferences by two ways. One way is statistical experiments and other way is prediction experiments. Both types of experiments have sub experiments associated with them. by statistical experiments, we have computed ANOVA test on the group of attributes to extract the information of such teachers having significant performance. But ANOVA has a limitation; as it is applied in group of attributes; therefore we are unable to identify the teacher having significant performance. To overcome this shortcoming of inferences, we have applied Z-test through which we have obtained the teachers having significant

performance with the pairing of each teacher belong to that particular course. By this way, we became able to bring out the teachers whose performance is better than other teachers who teach same course. Then we computed overall average performance of the teacher based on the overall average grade of students whom they taught.

With the help of prediction experiments, we have drawn prediction of grades inclusively and exclusively teacher's name who taught the courses. Then predicted results and actual results have been computed and compared. By this way we became able to get the entire set of results in which our analysis shows that predicted and actual results remained close at some point and better at some point too. In the same time, predicted were bit lower than actual ones as well which can be seen in chapter 4.

At the end conclusion and future work has been discussed.

5.1 Future Work

- The experimental work can be applied in other departments of the university.
- This data set is relatively small. We can carry out same experiments in other field of department and large datasets.

Moreover, these experiments can be carried out in government institution such as NADRA.

REFERENCES

- Abaya, S. A., & Gerardo, B. D. (2013, September). An education data mining tool for marketing based on C4. 5 classification technique. In e-Learning and e-Technologies in Education (ICEEE), 2013 Second International Conference on (pp. 289-293). IEEE.
- Aher, S. B., & Lobo, L. M. R. J. (2012). A comparative study of association rule algorithms for course recommender system in e-learning. *International Journal of Computer Applications*, 39(1), 48-52.
- Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting student's final GPA using decision trees: a case study. *International Journal of Information and Education Technology*, 6(7), 528.
- Anand, S. S., Bell, D. A., & Hughes, J. G. (1996). EDM: A general framework for data mining based on evidence theory. *Data & Knowledge Engineering*, 18(3), 189-223.
- Azarcon Jr, D. E., Gallardo, C. D., Anacin, C. G., & Velasco, E. (2014). Attrition and retention in higher education institution: A conjoint analysis of consumer behavior in higher education. *Asia Pacific Journal of Education, Arts and SCIENCE*, 1(5), 107-118.
- Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *learning analytics* (pp. 61-75). Springer New York.
- Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. arXiv preprint arXiv: 1201.3417.
- Bydžovská, H. (2013). Course Enrolment Recommender System (Doctoral dissertation, Masarykova univerzita, Fakulta informatiky).

Cabrera, A. F., Nora, A., & Castaneda, M. B. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. *The journal of higher education*, 64(2), 123-139.

Dougiamas, M., & Taylor, P. (2003). Moodle: Using learning communities to create an open source course management system.

Drea, C. (2004). Student Attrition and Retention in Ontario's Colleges. *College Quarterly*, 7(2), n2.

Elbadrawy, A., Polyzou, A., Ren, Z., Sweeney, M., Karypis, G., & Rangwala, H. (2016). Predicting Student Performance Using Personalized Analytics. *Computer*, 49(4), 61-69.

Garcia-Saiz, D., & Zorrilla, M. E. (2011, November). Comparing classification methods for predicting distance students' performance. In *WAPA* (pp. 26-32).

Guarín, C. E. L., Guzmán, E. L., & González, F. A. (2015). A model to predict low academic performance at a specific enrollment using data mining. *IEEE Revista Iberoamericana de Tecnologías del Aprendizaje*, 10(3), 119-125.

Hung, J. L., & Zhang, K. (2008). Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching. *MERLOT Journal of Online Learning and Teaching*.

Inoue, S., Rodgers, P. A., Tennant, A., & Spencer, N. (2017). Reducing Information to Stimulate Design Imagination. In *Design Computing and Cognition'16* (pp. 3-21). Springer, Cham.

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies*, 13(1), 61-72.

Kaminski, J. (2005). Moodle—a User-Friendly, Open Source Course Management System. *Online Journal of Nursing Informatics*, 9(1).

Kaur, G., & Singh, W. (2016). Prediction Of Student Performance Using Weka Tool.

Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Computers & Education*, 104, 18-33.

Kokina, J., Pachamanova, D., & Corbett, A. (2017). The role of data visualization and analytics in performance management: Guiding entrepreneurial growth decisions. *Journal of Accounting Education*.

Kovacic, Z. (2010). Early prediction of student success: Mining students' enrolment data.

Merceron, A., & Yacef, K. (2005, May). Educational Data Mining: a Case Study. In *AIED* (pp. 467-474).

Miller, B. N., Konstan, J. A., & Riedl, J. (2004). PocketLens: Toward a personal recommender system. *ACM Transactions on Information Systems (TOIS)*, 22(3), 437-476.

Nam, C. S., & Smith-Jackson, T. L. (2007). Web-based learning environment: A theory-based design process for development and evaluation. *Journal of information technology education*, 6.

Oskouei, R. J., & Askari, M. (2014). Predicting Academic Performance with Applying Data Mining Techniques (Generalizing the results of two

Different Case Studies). *Computer Engineering and Applications Journal*, 3(2), 79-88.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41(4), 1432-1462.

Polyzou, A., & Karypis, G. (2016). Grade prediction with models specific to students and courses. *International Journal of Data Science and Analytics*, 2(3-4), 159-171.

Ramist, L. (1981). College student attrition and retention. Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1), 135-146.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6), 601-618.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 12-27.

Sachin, R. B., & Vijay, M. S. (2012, January). A survey and future vision of data mining in educational field. In *Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on* (pp. 96-100). IEEE.

Scheuer, O., & McLaren, B. M. (2012). Educational data mining. In *Encyclopedia of the SCIENCE of Learning* (pp. 1075-1079). Springer US.

Tewari, A. S., Saroj, A., & Barman, A. G. (2015). e-Learning Recommender System for Teachers using Opinion Mining. In *Information Science and Applications* (pp. 1021-1029). Springer Berlin Heidelberg.

Tinto, V. (1987). *Leaving college: Rethinking the causes and cures of student attrition*. University of Chicago Press, 5801 S. Ellis Avenue, Chicago, IL 60637.

Yang, D., Sinha, T., Adamson, D., & Rosé, C. P. (2013, December). Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-driven education workshop*(Vol. 11, p. 14).

Zaiane, O. R. (2002, December). Building a recommender agent for e-learning systems. In *Computers in Education, 2002. Proceedings. International Conference on* (pp. 55-59). IEEE.

Stallone, M. N. (2011). Factors associated with student attrition and retention in an educational leadership doctoral program. *Journal of College Teaching & Learning (TLC)*, 1(6).

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 12-27.

Zhang, Y., Oussena, S., Clark, T., & Hyensook, K. (2010). Using data mining to improve student retention in HE: a case study.

Witten, I. H., Frank, E., Trigg, L. E., Hall, M. A., Holmes, G., & Cunningham, S. J. (1999). *Weka: Practical machine learning tools and techniques with Java implementations*.