# STUDENT RETENTION IN HIGHER EDUCATION INSTITUTIONS


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

# Student Retention in Higher Education Institutions

By

**Junaid Aftab**

**MASTER OF SCIENCE IN COMPUTER SCIENCE**


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

# Student Retention in Higher Education Institutions

By

**Junaid Aftab**

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**

**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 Retention in Higher Education Institutions

by

Junaid Aftab

MS133004

**THESIS EXAMINING COMMITTEE**

| S No | Examiner | Name | Organization |
|---|---|---|---|
| (a) | External Examiner | Dr. Umair Abdullah | FUI, Islamabad |
| (b) | Internal Examiner | Dr. Arshad Islam | CUST, Islamabad |
| (c) | Supervisor | Dr. Nayyer Masood | CUST, Islamabad |

_____

Dr. Nayyer Masood
**Thesis Supervisor**
July, 2017

_____

Dr. Nayyer Masood
Head of Department
Department of Computer Sciences
Dated : July, 2017

_____

Dr. Muhammad Abdul Qadir
Dean
Faculty of Computing
Dated : July, 2017

# Certificate

This is to certify that Mr. Junaid Aftab has incorporated all observations, suggestions and comments made by the external evaluators as well as the internal examiners and thesis supervisor. The title of his Thesis is: Student Retention in Higher Education Institutions.

Dr. Nayyer Masood

# DEDICATION

*I dedicate all my efforts to my beloved "Parents" who supported me and to all those who prayed for my success.*

# ACKNOWLEDGEMENT

I praise Allah Almighty who helped me in every way to maintain my enthusiasm to achieve this success. Then I heartedly admire the true concern and best guidance of my respected supervisor "**Dr. Nayyar Masood**". Words are not enough to express the gratitude towards him.

His kind supervision with great support and motivation urged me to engage with my work with devotion. In my professional career, I will feel honor to follow the way he teaches and make counseling of students to raise their interest in studies while maintaining the confidence on them. I want to pay all my thanks to my **supervisor "Dr. Nayyar Masood"** who encouraged me to give my best in every circumstance. The sheer amount of his support helped me to maintain my confidence to complete my degree.

True friends are like bright shadows in the dark, who thinks you a good egg even if you are half-cracked. I want to dedicate the part of my success with my friends **"Yasir Noman Khalid"** and **"Shafiq ur Rehman"**.

# Abstract

Scientific community has been proposing variety of approaches to identify factors affecting student retention in higher education institutions, since many years. Student retention is a hot issue in higher education institutions all over the world. From university perspective it is very costly and time consuming to bring new students into system. Majority of researchers have used statistical approaches to solve this issue but from last couple of years researchers used data mining approaches which give better result as compared to statistical approaches. The widely used attributes to conduct experiments were collected from three domains which are demographic, pre-college and institutional. After comprehensive literature review we have found that GPA, ACT score, SSG, HSSG, and parent occupation attribute effect the student retention. These attributes play role behind the attrition of students. After the comprehensive literature review we have found no research work is done on this issue in higher learning institutions of Pakistan. The objective of this research work is to investigate the 1) factors behind the attrition in our local context using data mining approaches, 2) to find the most influential courses behind the attrition, 3) to check the impact of teacher methodology on student attrition, and 4) to check the impact of introducing tutorial in first semester on student performance. In order to answer above mentioned research questions, firstly we acquired undergraduate computer science student's data from registrar office of Capital University of Science and Technology followed by cleaning the data. We collect data of BS (CS) students from spring 2014 to spring 2017. In pre-processing phase first we removed irrelevant data. Then we converted whole data in to required form which was compatible to tool (WEKA). In order to ensure that experiments have been conducted thoroughly and rigorously and to eliminate any doubts about the possibility of improved results I have used variations of the Decision tree classifier. A number of interesting and worth mentioning findings have been discussed. The finding shows that CGPA and HSSC (Higher Secondary School Grade) are applicable in our context. In addition the dropout students had poor performance in CP/ITC (Computer programming) and Cal-I (Calculus) courses. However, there is clear difference between attrition rate associated with different teachers' classes that MAY be due to teachers' methodology that need to be further investigated. Furthermore it is investigated that performance of students has improved due to tutorial period. These findings are important for decision makers of higher learning institutions and researchers.

# Contents

# List of Figures

# List of Tables

CHAPTER 1

# 1 INTRODUCTION

Educational data mining (EDM) is a growing research area that emphasizes on data mining methods, tools and techniques for exploring the institutional related data. The main objective of this discipline is to analyze the institutional related data to improve the learning experience and institutional effectiveness. Research in educational data mining domain is growing very fast (Jayanthi, Surendran and Prathap, 2016). The International Educational data mining society organizes a lot of seminar and conferences in this area for sharing data, methods and techniques. To answer the educational related research questions the EDM experts collect data from educational department. To conduct research, the educational related data comes from university databases, data warehouses, learning management systems, tutoring management systems or any other resources. The hot research areas in this domain ***are Personal recommender and Learning environments, course management system, and student attrition and retention***.

## 1.1 Personal Recommender System and Learning Environments

Personal recommender system and learning environments are correlated concepts in educational data mining. To adapt student learning needs on demand the personal learning environment emphasizes on providing a number of artifacts, services, and tools. Nowadays, students prefer to learn on internet without interaction with instructor (Shah, 2014). Personal learning is an online learning technique which provides material to students on demand. Students can collect their information from internet. Personal recommender system should be altered when it is used for educational purpose because recommendations should meet with educational

objectives. Reason is that existing recommendations systems are domain dependent and may not give accurate results (Shah, 2014).

## 1.2 Course Management system and Educational data mining

Course management system is another area of interest for researcher in educational data mining. Course management system provides a great platform or workspace to facilitate sharing data and information among users in a course. A course management system allows users to generate reports, prepare test and assignments, manage e-learning and facilitate online learning through chats, files, news services. This is an on demand service. The student can log on and work anywhere, anytime. From last couple of years, to assist instructors researchers are applying various data mining tool and techniques to improve course management system (Shah, 2014).

## 1.3 Student Retention and Attrition

Student retention is a serious issue in higher education institutions. Student retention is a symbol of success for university management enrollment and faculty. Poor student retention rate create bad image for university and create financial and academic problems for university. Poor student retention badly effect on different aspects of an institution. Student retention is first priority of university management and decision makers. It is globally accepted that Retention and Graduation rates in high education institutions are key factor of effectiveness or efficiency of an institutions. The reduction in number of students from a program as the time passed is called Student Attrition. To improve student retention there is need to find out the reasons behind the attrition. From the comprehensive literature study it is found that average retention rate in colleges is approximately 55 percent which means 45% engineering students leave their program before completing degree (Nandeshwar, Menzies and Nelson, 2011). This research title is

student retention in higher education institutions. The objective of this research is to determine factors that are causes of student persistence and dropout.

## 1.4 Background

Student retention is a major problem in higher education institutions all over the world. The literature study shows that majority of students quit their study before completing their degree. According to college student retention center the average retention rate is 50%. High attrition rate creates problems for institutions and students as well. From university point of view it is very time consuming and costly to bring new students in the system. This problem also personally impact on life of students and creates financial problems for universities and students both. Research on student retention is done widely over the last couple of years. A number of theories established to find the factors behind the attrition of students. Initially researchers have used traditional statistical approaches to find the factors behind the attrition. From last of couple of years researchers used data mining approaches and tools to find the factors behind the attrition. After the comprehensive literature study it is found that the data mining approaches gives better results as compared to traditional statistical approaches.( Pal, 2012) ( Jia, Mareboyana, 2013). The researchers have identified many factors that effect on student retention. After the comprehensive literature survey it is found that student demographic, academic, social and precollege attributes effect directly or indirectly on retention status but mostly researchers have used only one category of data to conduct experiments. The literature study shows that *GPA, ACT score, SSG (Secondary School Grade), HSSG (Higher Secondary School Grade), and parent occupation* are the main factors behind the attrition of students. To conduct experiments the researchers collect data from three domain that is demographic, pre-college and institutional ( Pal, 2012) ( Jia, Mareboyana, 2013). The students enter the institution

with some background. These attributes are demographic or family background, pre-college attributes. The **Demographic** attributes are (*Age, gender, Financial status, Balance due, Permanent address, Residential address, Guardian (Brother, uncle, Self), Father qualification, Father Occupation, Full/ Part time student).*The previous research shows that demographic attributes play important role for student persistence and dropout. For example, location of living is important factor that effect on retention. Mostly students drop out because they come from long distance. Their long traveling distance disturbs their study (Djulovic, & Li, 2013). The difficulty in fee payment also impact on attrition of students.

**Pre-college Attributes** (*Secondary School Grade*, *Higher Secondary Grade*, *SAT Score, Pre-college, Pre-board, pre-program*). Different students come with different academic, demographic and social background and with their own individual perception level. Every teacher has his/her own personality, communication, way of teaching, interaction etc., which can all be accumulated under teaching methodology of a particular teacher. I have included some new attributes in dataset for example teacher methodology and first semester courses like computer programming, Calculus, Introduction to computer and English and secondly **Institutional attributes** (*CGPA, First semester courses*). The research shows that institutional attributes play important role for predicting the attrition of the students. For example, GPA is main oversees identified factor that play important role for predicting the attrition of the students.

The research gap that we identified during the literature review that majority of researchers used only one category of data from demographic, pre-college and institutional domain to conduct their experiments. ***Secondly***, majority of research has been done in oversees institutions.

## 1.5 Problem Statement

High ***attrition*** rate causes ***financial and academic*** problems to universities and students both. There is need to determine those factors in local context that cause high attrition rate. This research will be helpful to reduce the attrition rate of higher education institutions. I have addressed the following questions in our research:

1. **To check whether oversees identified factors GPA, ACT score, SSG, HSSG, and p-occupation are valid in CUST.**
2. **To find most influential subject(s) behind attrition in first semester.**
3. **To check the impact of Teachers' Methodology behind attrition in first semester.**
4. **To check the impact of introducing tutorials in first semester on student performance.**

## 1.6 Purpose

The purpose of this study is to identify the factors behind the attrition in Capital University of Science and Technology, Islamabad.

## 1.7 Scope

Various approaches are used to identify the factors that effect on student retention in higher education institutions. (1) Initially traditional statistical approaches were used to identify the factors behind attrition and now from last couple of years researchers are using data mining approaches and tools. The scope of our work is to identify the factors behind the attrition and to check whether oversees identified factors are valid in our local context.

## 1.8 Significance of the Solution

This research is significance for several reasons. First, this research will contribute the literature study related to student retention. Majority of researchers have done research on student retention in higher education institutions outside the Pakistan but there is no study conducted on this issue in Pakistan.

Second, this study is also helpful to faculty, staff and decision makers for Capital University of Science and Technology as this research gives a clear picture of factors affecting student retention. The university management can develop their policies and programs that may help them to prevent student's dropout. Finally this research is helpful for future students and their families.

## 1.9 Organization of Thesis

This thesis has been organized on five chapters. Chapter one discuss the introduction and background of the problem, purpose of the research, and its significance. Chapter two provides the state of art approaches to the study. Chapter three presents the proposed methodology of research work. Chapter four describe the results and final chapter five summaries the whole research works with conclusion and future work.

## 1.10 Definitions of terms used

**Retention**: Retention mean to keep students enrolled. How many students remain enrolled from start to end? Successful completion of students' academic goals of degree attainment

**Attrition**: The reduction in number of students attending courses as time goes by. The process of leaving study before the completion of degree is called attrition.

# CHAPTER 2

## 2 LITERATURE REVIEW

Student retention is a challenging issue in higher education institutions. Majority of researchers applied data mining approaches to predict student retention in higher education institutions. (Sherrill, Eberle and Talbert, 2011) In this paper author used machine learning technique to improve the retention rate in higher learning institutions. The main objective was to identify which newly students can be targeted from risks. If new coming students can be identified at early stage then attrition rate can be reduced. The student data was collected from university student database and then preprocessed it. Once the data is cleaned it is given for classification. The numeric attributes in dataset are like (GPA, test scores, hours) are grouped using Discretize function for minimizing the complexity. Then all numeric values were converted to nominal values using numerical to nominal function. After preprocessing they have applied classification techniques on the whole dataset. A computer science degree need 120 credit hours to complete, if a student complete 90 credit hours then they declared it as retained student.

In this research study (Sherrill, Eberle and Talbert, 2011) two classes are created of whole dataset Retention and Not Retention. Decision tree, Bayesian classifier, neural network and SVM were used for classification. Ten-fold cross validation is used for validation testing. Confusion matrix is used for performance comparison. The author also used bagging, boosting and attributes selection techniques improving accuracy but no significant accuracy they get. The students whom retention status they want to determined were grouped according to their terms. In first test, they grouped from Term 1 to Term 4. Each test set included information from the Student, High school and student term tables.

In second experiment the author perform experiments on zero and first term data and another change was needed for this data, a separation of transfer and non-transfer students. The results show that the accuracy is improved when zero term is included. Their work relies on small amount of dataset.

(Ngemu, Elisha and Bernard, 2011) In this paper data mining techniques had been applied to improve the retention rate in higher education institutions. They have used student demographic and institutional data and build a prototype to predict student persistence or dropout. The model is built using 10-k fold cross validation and 60% data is used as a training data and remaining data is used as test data. Random sampling techniques are used for extracting the data.

The results shows that student age, parent occupation, parent occupation, health of student and financial variables are most important factors that predict the student persistence and dropout. Results of classifiers were compared using accuracy level, and confusion matrices.

Their methodology was consisting of data collection, preprocessing, building classification model, using training data, and evaluation of model on test data. Data was sourced from university database having 270 instances and 14 attributes. The classifiers that were used for classification are decision tree, Navie bayes, and SVM. The raw data is provided to Weka. The dataset is divided into training set, testing and validation. After the model building the model is evaluated on test data. Based on benchmarks it is observed that J48 gives best accuracy as compared to others. Few patterns are identified from experiments which are shown below (5)

If difficult in fee payment= Yes then outcome= Dropout

If difficult in fee payment = No, student health= Good, then outcome= Persist.

If difficult in fee payment= Yes, age is < 20, parent occupation= Self-employed then outcome= Dropout

If difficult fee payment= No, Std.health = poor, parent occupation= good, then outcome= Persist

(Dagley and Young, 2016) National Science Foundation started STEM talent national program at University of Florida (USF) because of improving retention and graduation rates. USF EXCEL program annually recruits almost 200 students into a education community. The overall retention rate of EXCEL program is 43 % higher than comparison group. The Excel program was funded by NSF in the field of Science, technology, engineering and mathematics at the USF. The main objective of this program is to increase the number of students receiving degrees in the field of STEM from USF. From November to May high school seniors can apply to the EXCEL program and selection is made on rolling base.

The selection of students was based on math SAT score, intended major, math placement, ethnicity, and gender. After the selection these cohorts are divided into two groups (pre-calculus and calculus I).To appropriate assess the usefulness of EXCEL educational program a comparison group is created for similar students.

The comparison group consists of those students who have applied a STEM major at the time of application to UCF and fall within in same SAT score. Other comparison group's also created on the base of high school GPA.

The success of Excel program is measured on annual retention rates compared with comparison group. To compare a Chi-square test of association was used to find relationship between retention and Excel program. Retention rate of students participating in EXCEL program is higher as compared to comparison group. First year retention in a STEM major for

the EXCEL program is between 80% and 85 % while retention in comparison group is between 63 % and 67 %. Form 2006 data results 23% advantage in first year retention over the comparison group. Examination of 2013 retention data indicates that while STEM attrition occurs after the first year and overall retention rate is 43% higher than comparison group.

(Alkhasawneh and Hobson, 2011) In this paper data mining approaches were used to predict the factors that influence undergraduate student retention in HBCU (Historically black C College) and develop a model which can predict student attrition risk. The dataset was collected from College enrollment server. It consists of 771 instances and 12 attributes. The entire dataset was divided into two classes Retention and Not Retention. The Weka tool is used to classify the dataset by using different machine education algorithms. Decision tree, Navies Bayes, J48 algorithm has been applied for classification purpose. Their results revealed that GPA is main factor behind the attrition of students in higher learning institutions.

(Jia and Mareboyana, 2014) Data mining approaches had been used to predict student retention in the STEM (Science, Technology, Engineering and Mathematics) domain. The first model identify correlated precollege factors and also to predict incoming freshmen retention. The second model classified first coming students into three groups: at risk, intermediate and advanced students. In this research the response variable is overall GPA. The experimental dataset consist of 338 incoming freshmen from STEM discipline.

The dataset ratio is 52 percent male and 21 percent females from the (African, American, Hispanic, etc.). The independent variable in this study is overall college GPA. Two independent models were to build 1) prediction of first coming students 2) to classify into three groups. Neural network algorithm is used for model building.

Gauss newton learning method and 10 fold cross validation is used to train the network and to avoid over fitting. To build GPA model prediction the r value of this model is 0.54 and the accuracy 68% margin error is set within [-0.5, 0.5]. The accuracy of classification model is 70 percent with r value equal to 0.41. For this small data set 70 % accuracy is acceptable for predicting the absolute GPA and can be improved with large datasets and more related factors such as math performance test.

(Grier-Reed, Inman, 2016) The author collected data of 91 black women and 56 black men from African American student network and on other side take 68 women and 36 men data randomly from Black undergraduates from a Midwestern university and used an analysis of covariance to control for *ACT score* and *first term GPA average.* Their result shows a statically significant main effect for network, where African American network were retained at significant higher rate than randomly selected non network African American students.

This study took place at a large, urban, predominantly White university in the Midwest. The sample comprised a total of 251 African American undergraduates. These students were not part of a single cohort year. Of this group, 147 participated in AFAM and 104 were non-AFAM participants randomly selected from the same entering or matched cohort. Of those who participated in AFAM or the treatment group, 91 (62%) were female and 56 (38%) were male. Of the non-AFAM participants or the control group, 68 (65%) were female and 36 (35%) were male. The result shows that AFAM students were retained on average 3.7 years with a standard deviation of 1.54 and non AFAM students were retained for 3.2 years with standard deviation of 1.84. There was also significant effect for the covariate GPA although standard deviation and mean for both groups is similar. There was also more variation between sexes, female perform better. Their research is only limited for black students. Furthermore the author included only

two covariates like ACT score and first term GPA. In future psychosocial variables and other more environmental factors can be used to perform greater experiments.

(Yadav, Bharadwaj and Pal, 2012) In this paper author find factors behind undergraduate student retention using signal processing techniques. Their result shows that GPA is main factor that influence the retention rate. In this paper linear smoothing approach is used to remove the noisy data for improving student retention results. For accurate classification data is decomposed into Haar coefficients. The response variable is GPA and class attribute is Retention.

The dataset was collected from HBCU database from fall 2006 to 2011. After pre-processing the entire data was grouped into two files Retention and Non Retention. Linear smoothing technique was used for removing the noise in GPA data. Finally applied Haar transform approach to the GPA data and discussed the average GPA for the HBCU undergraduate student retention. Then tested the Haar average retention GPA using test data and compared the results with Navie bayes mean value. Results reveals Haar based classification is better than Navies bayes. Smoothing the data removed the highest and lowest GPA values from both retention and non-retention data. Smoothing technique filtered out the noise and made it pure. From Haar transform results the author identify that the average GPA for the HBCU undergraduate student retention should be 2.8597 and average difference should be 0.023307.

(Fike, 2013) The author used data mining techniques to predict the factors behind the attrition of students on freshmen student data. The main aim of this research was to found the factors behind the attrition so that the university can improve their retention rate. After the analysis of data the author find out most appropriate variables to construct the student retention

prediction model. For the analysis of data the most well know data mining algorithms have been conducted. The result revealed that student GPA and financial variables play big impact on student's retention.

The data has been collected from university database from 2006 to 2011 to conduct the research. For accurate prediction of retention the author included pre college academic attributes like residency, gender, SAT scores, GPA, amount of balance due and living from campus also included in dataset. In addition the author also include Retained variable to denote if the student is retained then it is set to 1 else it is set to 0. From 7800 instances 12% of them have missing values, for biasness these values are removed from dataset. For better classification model numeric variables are converted into categorical attributes based on domain knowledge. To determine the importance of each variable Information Gain, Gain ratio, Chi-squared and correlation analysis are adopted.

The result shows that student CGPA and GPA are the main factors that effect on performance of students especially spring and winter season. After applying attribute selection technique the results revealed that financial balance has a big impact on student retention. The author used Weka tool for classification, C4.5 algorithm builds a binary tree. The author set the confidence threshold 0.25 for pruning. The result shows that CGPA and GPA are the main factors that affect the retention of the students. This attribute is more important in determining the target variable Retained. And second spring balance is also an important attribute for detection of Retained students which is selected this algorithm.

(Mamiseishvili, Deggs, 2013) The author applied data mining techniques to develop a predictive model for prediction of attrition student in the first year engineering. The model can predict the correct list of new incoming at risk students. The experiment data reveals that machine learning techniques can build effective predictive model on existing attrition data. In this paper, the classification approaches ID3, C4.5, CART and ADT decision tree is used for analyses of the previous student drop out data. The attributes of students like Family income, grades in previous high school and secondary, Guardian qualification etc. are collected from student enrolment form. The engineering student dataset is collected form university database enrollment form filled by students from 2006 to 2011. The engineering student dataset consist of 1650 records.

In first step, only those attributes are extracted which are necessary for data mining process. The variables related to student which are collected is Sex, Cat, HSG, SSG, Atype, Med, LLoc, FAIn, FQual, MQual, FOcc, MOcc, and Dropout. For Implementation WEKA toolkit is used. To use this tool the extracted dataset is first prepared and converted into (arff) file format because WEKA tool is compatible with this format.

Four classification algorithms have been applied using ten cross fold validation. The algorithms which are used in conducting experiments are C4.5, Cart, ID3 and ADT. Then applied preprocessing and preparation techniques and then analyzed the results visually. The ID3 algorithm generated very deep tree starting with HSSG attribute it means this attribute is very effective play important role for predicting the attrition. Other hints that can be identified from the tree is HSSG= A or O are continue their study.

The second technique was C4.5 which also indicated that HSG variable is most important attribute. Then applied CART algorithm which is started with SSG. Other hint was observed

form the tree that students which have “F” or “E” grade in SSG they dropout. In this paper a number of attributes have been investigated and some of them are very effective. HSG is found very effective as compared to SSG with little effect of MOcc and FAIn. The medium variable did not play any role but Sex and category play little impact in some experiments [15]

(Pal, 2012) The author developed a predictive model using data mining techniques that is capable to predict incoming retention students. If factors behind the attrition are known then decision makers can take preemptive action, so student retention rate can be increased. To conduct experiments, the data is collected from university enrollment database having 432 records consists of twelve years having basic information. The results revealed 398 students continued their study after their first year while 34 students were dropped out by the end of year.

In data pre-processing phase only those attributes were selected which are necessary for data mining. A few attributes were selected after pre-processing which are. Sex, GSS, GMSS, GS, GOG, MED, CL, AType, RET. WEKA software is used for implementation that consists of a number of data mining algorithms. The selected techniques were ID3, C4.5, and ADT under the test of ten-fold cross validation. The model was developed in the form of decision tree which is capable to predict new incoming students whether they will continue their study or not. Another interesting pattern is found from the retention dataset that GS (Graduation Stream) is most appropriate factor. The accuracy of C4.5 algorithm is highest as compared to ID3 and ADT. The precision value of ADT algorithm is highest found, having 83 percent precision and 11.4 percent recall rate without Over fitting.

(Kabakchieva, 2012) The author found the factors behind the student retention from fall to fall and to spring first time in in college students. This study was conducted in community

college USA. Almost 10000 students data was collected who enrolled in fall 2001 to 2004. Two dependent variables were comprised for analysis. The independent variables were age, ethnicity, gender, financial status, enrollment in online courses, status of development courses writing, reading and mathematical, credit hours taken in first semester and dropped and parent background.

To preprocess and cleaning Statistical Package for social sciences were used. For analysis and prediction data mining approaches Chi square analysis, Bivariate correlation coefficients, Point biserial correlation coefficients, phi correlation coefficients and multivariate regression models were used for prediction of odd student retention. The student data statistics were 56 % was female and 66 % was white with the median age 19. Student semester hours enrollment percentage is 12 and 99 % were enrolled in less than 20 TCH.

The retention status of fall to spring was differing by year about a third of students who enrolled in fall did not enrolled in spring in the same institution. Fall to fall retention rate is from 45% to 49% which is low as compared to fall to spring. The strongest predictors are successful completion of development subjects especially in writing and mathematics are difficult to continue in further semesters. Limitations in study are lot of missing values was present and some of them data were self-reported.

| Attributes | [1] | [2] | [3] | [4] | [5] | [6] | [7] | [8] | [9] | [10] | [11] | [12] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GPA | ✓ | ✓ | | | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| T1, T2, T3, GPA | | | | | | ✓ | | | ✓ | ✓ | F/P.Time | |
| Grade | | | | ✓ | | | | | | | | |
| F. Aid | ✓ | | | | | ✓ | | | | | | |
| HSG | | | ✓ | | | ✓ | | | ✓ | ✓ | | |
| SSG | | | ✓ | | | | | | | | | |
| Gender | ✓ | | ✓ | ✓ | | ✓ | | ✓ | ✓ | ✓ | ✓ | |
| Age | ✓ | | | ✓ | | ✓ | | ✓ | ✓ | | | ✓ |
| Race | ✓ | | | | ✓ | | | ✓ | ✓ | | | ✓ |
| Minority Status | | | | | | | | | | | | |
| TCH | ✓ | | | | | | | | | | | |
| Course Taken | | | | ✓ | | | | | | | ✓ | |
| Intial Major | | | | | | | | | | ✓ | | |
| Current Major | | | | | | | | | | ✓ | | |
| C. enrollment Status | | | | | | | | | | ✓ | | |
| Course Award | | | | | | | | ✓ | | | | |
| Library Info | | | | | | | | ✓ | | | | |
| School | ✓ | | | | | | | | | | | |
| Plan/Program | ✓ | | ✓ | | | | | ✓ | | | ✓ | |
| Distance | ✓ | | | | | ✓ | | | | | | |
| Difficult in Fee. Payment | | | | ✓ | | | | | | | | |
| SAT score | ✓ | | | | ✓ | ✓ | | | ✓ | ✓ | ✓ | |
| Retention | ✓ | | | | | | | | | | | |
| Med of Teaching | | | ✓ | | | | | | | | | |
| Living Loc of Std | | | ✓ | | | | | | | | | |
| SP/Guardian | | | | ✓ | | | | | | | | |
| M. Qualification | | | ✓ | ✓ | | | | | | | | |
| F. Qualification | | | ✓ | ✓ | | | | | | | | |
| F. Occupation | | | ✓ | ✓ | | | | | | | | |
| M. Occupation | | | ✓ | ✓ | | | | | | | | |
| Health | | | | ✓ | | | | | | | | |
| Class | | | | | | | | | | | | |

**Table 1: List of Attributes**

This chapter presented the state of the art approaches on student retention in higher learning institutions. The literature review shows that majority of researchers have used limited data set to conduct experiments. Secondly, no research is done in higher learning institutions of Pakistan. The next chapter is presented a detail description of proposed methodology used in thesis.

# CHAPTER 3

# 3 METHODOLOGY

This chapter provides detail description of research methodology. The main objective of this study is to identify the factors behind the attrition and use this information to improve the retention rate. For this purpose, I performed a comprehensive literature review and identified certain factors that have been established as main cause of attrition in foreign universities. One of my motivations is to check whether oversees identified factors behind attrition are valid in CUST and to what extent. Moreover, I included some new factors in our experiments to evaluate their role in student's attrition. To conduct experiments I collected data from Department of Computer Science of Capital University of Science and Technology (CUST), Islamabad. I have performed experiments over the data of three years of BS(CS) program. After data collection, I cleaned and prepared data through preprocessing phase and then performed different classification techniques. The results obtained have been analyzed and discussed in the next chapter. The structure of this chapter is as follows: In section 3.1 Data collection is described in detail and in section 3.2 block diagram of propose methodology is shown and then in section 3.3 Pre-processing of data is defined and in last proposed techniques and tool is described.

## 3.1 Strategy to Address Research Questions

There are five different research questions that I will be able to answer at the end of this research task. This will be research contribution in this particular research activity which falls in the domain of student retention in Capital University of science and technology.

**RQ1**. **To check whether oversees identified factors *GPA, SSG, HSSG, ACT score and p-occupation are valid in our context:*** Every country or society has its own social, cultural, financial and educational norms and environment. Most of the studies on student attrition have been performed in societies/countries USA, UK, India, Kenya, Australia that are quite different from Pakistan. It then becomes a valid question whether those factors are applicable in local context and if they are then what extent. For this purpose, data has been collected from CUST Islamabad, Pakistan. I pre-processed the whole data in our required form. I include oversees identified and some new attributes in data set to conduct experiments. The collected attributes were from the student demographic, pre-college and academic domain.

In the following, research questions targeted in this question and the strategy adopted to answer each of them have been described.

**RQ2**.**To finds most influential subject(s) behind attrition in local context:** To address the question to find the most influential subject behind the attrition in our local context. Some students come in computer science from different domain so they cannot survive after one or two semesters. In BS (CS) program there are two major streams of subjects. One is Computer programming and other one is Mathematics. There are almost eight courses from mathematics domain. If a student is weak in programming or mathematics and cannot get good grade in his/her first course then he cannot survive in remaining courses, then finally dropout out from

program. So to check which subject play major role behind the attrition I have collected data of CP (computer programming) and Cal-I (Calculus-I) from term 141 to term 163. I have prepared data of computer programming and Calculus in required form. We have calculated CGPA of every student in their respective subject and then converted into their grade. After pre-processing we have forward it to tool.

**RQ3. To check the impact of Teachers' Methodology behind attrition:** Different students come with different academic, demographic and social background and with their own individual perception level. Since there are multiple sections of same subjects in CUST, so it is quite common that same subject is being taught by different teachers in different sections (of course with the same course outline). Every teacher has his/her own personality, communication, way of teaching, interaction etc., which can all be accumulated under teaching methodology of a particular teacher. Now there is a good chance that the teaching methodology of some teachers is (naturally) more appropriate for students with specific background. There might be different approaches to address this question that whether the teaching methodology has some impact on learning. But in this research we addressed this problem by exploring the existence of any pattern(s) between attrition and a particular teacher who taught influential subject(s). We divided students with different academic background into different groups and then evaluated their performance in influential subjects and attrition considering the teaching methodology.

**RQ4**. **To check the impact of introducing tutorials in first semester on student performance**: Based on the continuous feedback of teachers and students, the Department of Computer Science at CUST took the initiative of introducing tutorial for the subjects of

"Introduction to Programming" and "Calculus-I" from Fall 2016 semester on the regular basis. The frequency of each tutorial was one 1.5 hour session per week. Students were encouraged to attend the tutorials, rather it had been declared for the students to attend the tutorials. After completion of semester and declaration of results, the attrition rate and results performance in both subjects were compared with previous semester. The difference, if any, will be attributed to tutorial as it is the only difference between previous semesters and the Fall 2016.

The strategy adopted to address the research questions has been mentioned in the above. The steps taken to implement the above strategy have been mentioned in the following.

## 3.2 Data Collection

This study examines the retention status of undergraduate students of BS (CS) students of Capital University of Science and Technology (CUST). The admission process in CUST is performed on semester basis. There are two semesters in one year, named as Spring semester and Fall semester. The Spring semester starts approximately in mid-February and continues till end of June, whereas Fall semester spans from mid-September to end of January. A semester is referred by a term number comprising three digits; first two digits of the term describe the year of admission and last digit means Spring (1) or Fall (3).Fall or Spring. So, the term 141 means Spring semester of 2014 and term 143 means Fall semester of 2014. The three years data that I collected range the students admitted in Spring 2014 till those admitted in Fall 2016 and they are identified with the terms 141, 143, 151, 153, 161 and 163 terms. In first step, I identified the number of students that are registered in 141 terms through ITC (Introduction To Computing) course which is a compulsory course.

We collected total registered students 141 term through ITC course and then I collected CP (Computer Programming), Calculus, English, and ITP courses data of 143, 151, 153, 161 and 163 term. There are two reasons behind the collection of CP and Cal-I courses data because CP and Cal-I are two major streams of courses that are offered in BS CS program. First I have collected CP and Calculus data to check the impact of these two major streams of courses on attrition of students. If a student have poor grade in CP and Calculus, which is first courses, then he cannot survive in remaining courses and finally dropout. Secondly, I have manually identified dropout students from these two major streams of courses after each semester. Then I manually identified the students who were dropped out of each session, starting right from session 141 to 163. The university has no proper data about students' attrition. We established an indirect way of identifying the dropout students which is based on the registration data. I checked the registration data of a particular session (like 141) and from there we highlighted the students registered in first semester, and then for the same students we see the registration data of the next session (for 141, next session will be 143). The students of a particular session who are missing in the registration of next one are possibly dropout students. Likewise, I analyzed the registration data of every session to identify the missing students who are probably dropped. The missing student's data is further explored through portal to verify that these students have actually dropped. For example, 74 students took admission in session 141 and by their sixth semester (163) 26 of them were dropped out. I further confirmed from university portal that these students were actually dropped out. The same procedure we adopted for the students of all remaining semester. After identifying dropout students then in next step we collected student demographic and pre-college data from registrar office of the university. The demographic and academic data of students were not in proper form, data was in different files we prepared the demographic

attributes in our required format. Then in next step I collected academic data like CGPA of students this data was also in different files and then I found out CGPA of each student according to their university registration number. The total instances were 608 from them 166 students were dropout before completion of their degree. The detail description of registered and dropout students is provided in below Table (2) below.

| **Term** | **Registered Students** | **1st Semester** | **%** | **2nd Semester** | **3rd Semester** | **4th Semester** | **5th Semester** | **6th Semester** | **Total** | **%** |
|---|---|---|---|---|---|---|---|---|---|---|
| **141** | 74 | **16** | 21 | 5 | 2 | 2 | 0 | - | **25** | 33.78 |
| **143** | 155 | **15** | 10 | 18 | 13 | 6 | 0 | - | **52** | 33.54 |
| **151** | 115 | **20** | 17 | 10 | 3 | 0 | - | - | **33** | 28.69 |
| **153** | 101 | **20** | 19 | 7 | 1 | | - | - | **28** | 27.722 |
| **161** | 62 | **15** | 24 | 1 | - | - | - | - | **16** | 25.80 |
| **163** | 101 | **12** | 11 | **-** | **-** | **-** | **-** | **-** | **12** | 11.88 |
| **Sum** | **608** | 98 | | 41 | 19 | 8 | 0 | **-** | **166** | **28%** |
| **Attrition %** | | **59%** | | **25%** | **11%** | **5%** | **0%** | **0%** | | |

**Table 2: Summary of session-wise registration and attrition**

The primary goal of this research is to identify factors behind attrition before the completion of degree as the overall attrition ration (34%) is quite high and university would definitely want to reduce this percentage not only for the benefit of itself but also for the betterment of students.

After carefully analyzing we have found majority of our students drop out because of academic performance. The graph 3.1 shows that majority of our students are dropout which have low grade and mostly our students retained which have good grade. This shows that students with poor academic performance cannot continue their study. In figure (1) x-axis shows the student's grade earned and Y-axis shows the percentage of student's dropout and retention.

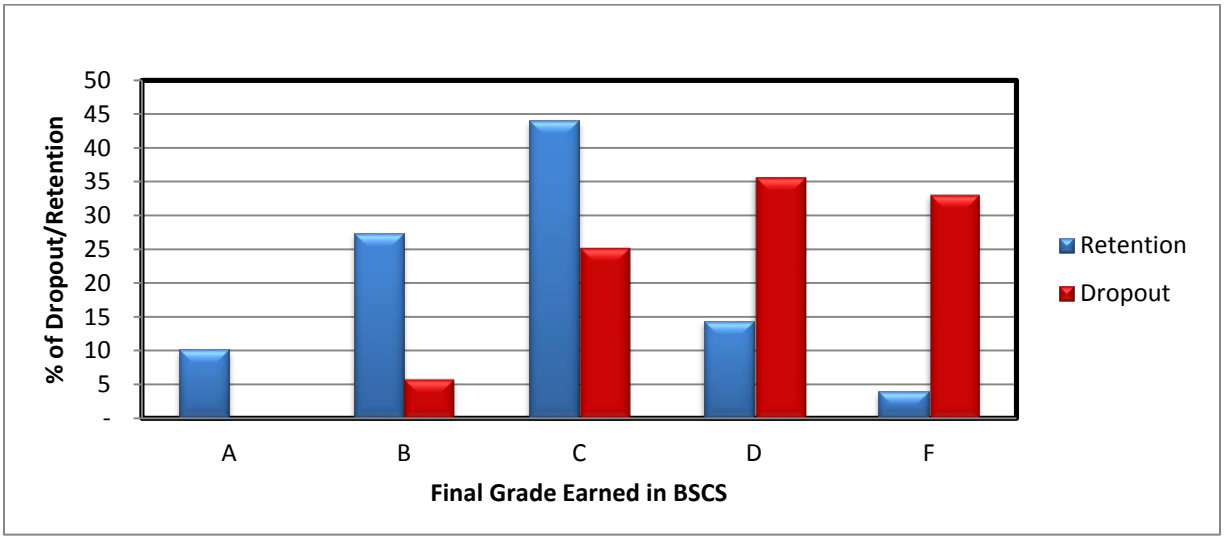


**Figure 1: Student Academic Performance**

## 3.3 Data Mining Technique

Selection of appropriate data mining technique is the next task. For this purpose we had to evaluate our objective which was basically to predict the factors behind the attrition especially in the context of Pakistan. This will help us to identify the potential students who have more chances to leave the study and university can take some special measures for these students to stop or reduce the attrition rate. Most of the techniques that we reviewed in relevant literature used tree based classification techniques for this purpose so we also selected the same. The purpose of tree based classification is that these techniques give us results in the form of rules as well which are another easy to understand way of interpretation and through which we can

predict the attributes behind the attrition. We have used all the variations of tree based classification techniques for validation of our results.

In data mining domain, for prediction of attributes there must be two types of variable. Dependent and Independent variables; the dependent variable is output variable in which the researchers are interested to see during monitoring whether it is affected or not. It is also called responding variable, measure variable etc. The variables which we believe that may impact on dependent variables are called independent variables. These are sometime also called controlled variables, manipulated or explanatory variables.

The dependent variable in our study was attrition and it is measured from the registration data of Department of Computer Science, Capital University of Science and Technology from Spring 2014to Fall 2016.

The independent variables were collected from the categories of student demographic, pre-schooling and academic data. We have collected data from these three categories because mostly researchers have collected data from these three categories to conduct their experiments [4][5][6]. The independent attributes Matric grade, matric percentage, HSSC percentage, HSSC grade, ITC grade, ITC teacher name, ITC mid marks, CP mid, CP grade, CP teacher name, English mid, English teacher name, English grade, gender, city and student CGPA. From selected attributes mostly attributes were used in previous research papers and some of them are new proposed.

For example CGPA, HSSC grade, gender, city, were used in previous research papers and play important role in prediction so that is why we have selected these attributes to check whether these attributes are valid in Pakistan. The new proposed attributes are teaching

methodology, critical courses' grade, courses mid marks details, and pre-program. The description of selected attributes is shown in Table (3).

| Sr.# | Name of Attribute | Description | Category of Attributes | |
|---|---|---|---|---|
| 1 | **Gender** | M and F means male and female | Demographic | Old |
| 2 | **City** | Rawalpindi (R), Islamabad (I), Hostel (H) | Demographic | Old |
| 3 | **Matgrd** | Matric Grade | Pre-college | Old |
| 4 | **MatPerc** | Matric marks percentage | Pre-college | Old |
| 5 | **Pre-Program** | Intermediate program (Pre-engineering, Pre-medical and Diploma) | Pre-college | New |
| 6 | **HSSCgrd** | Intermediate grade | Pre-college | Old |
| 7 | **ITC/CP result** | Introduction to computing the first course of the term/ their grade mid marks and teacher name | Academic | New |
| 8 | **Teacher Methodology** | The name of teacher that teaches CP and Cal-I | Academic | New |
| 9 | **Cal-I result** | Calculus course their teacher name and mid marks | Academic | New |
| 10 | **Tutorial** | Data of students which attempted tutorial | Academic | New |
| 11 | **CGPA** | CGPA of students | Academic | Old |
| 12 | **Class** | Retention and Attrition | | |

**Table 3: List of Selected Attributes**

The new included attributes are Teacher methodology, CP course grade/mid marks/teacher name and Cal-I grade/mid marks/grade. Teacher methodology attribute play

important role on attrition of students. In our education system every teacher have different teaching methodology so want to check the impact of this attribute on attrition of students we have included in our dataset.

The CP and Cal-I course grade/ mid marks/ teacher attribute also play very important role behind the attrition of the students. These two major streams of courses play important role behind the attrition of the students. These are compulsory and tough courses in computer science program so if student cannot get good grade in initial course then he/she cannot survive in remaining courses.

## Block Diagram of the Proposed Methodology

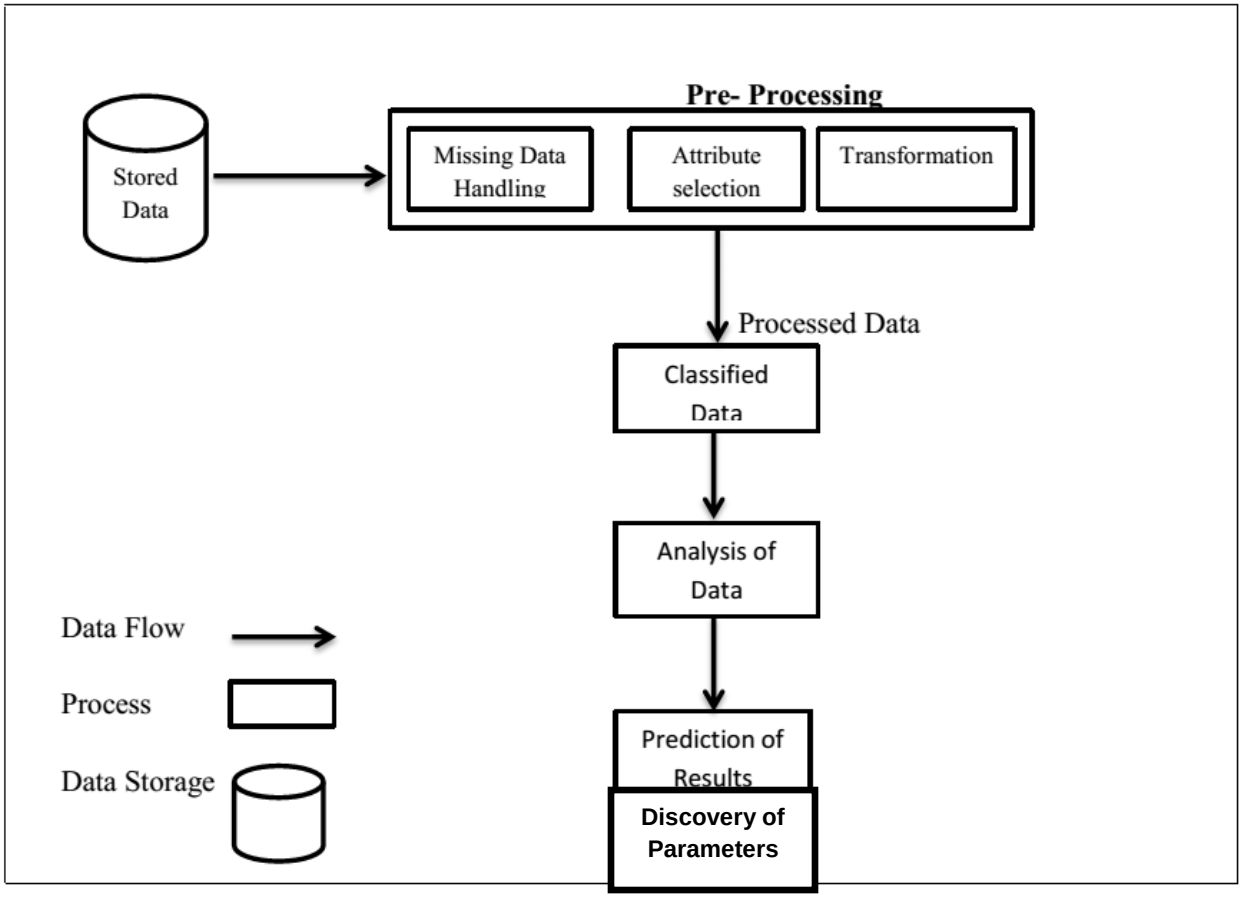


**Figure 2: Proposed Methodology**

In figure (2) the container symbol contains the student data that is collected and next is rectangle symbol which shows pre-processing that contain missing data handling, attributes and data transformation phases. After pre-processing the data was given to tool then I applied tree based classification approaches. After processing the data the rectangle symbol comes where results were analyzed and in last results were presented.

## 3.4 Pre-Processing

For our experiments, we have acquired a data set of undergraduate Computer Science department students BS (CS).The collected data was from term 141 to 163 term. The dataset attributes were collected from Student demographic, pre-schooling and academic categories.

For the task of preprocessing, the collected data must be prepared in ARFF format which is compatible with the selected tool. The collected data were in raw form. We have prepared the entire data set in our required form. We pre-processed the entire data.

In pre-processing stage first irrelevant attributes were eliminated from whole data set. Bio-informatics and BSE students were excluded from dataset because we want to conduct experiments on only BSCS students. In next step Registration number, name of the students, father name and NIC number were not required for analysis then we have eliminated all these irrelevant attributes from entire data.

The included attributes were matric grade, matric percentage, pre-program, HSSC percentage, HSSC grade, gender, city, ITC/grade/mid-term marks/Teacher, CP/grade/mid-term marks/Teacher, Cal-I/mid-term marks /Teacher name and student CGPA.

For tree based classifiers experimentation we prepared entire data in specific format. Following steps were performed to acquire required data in specific format.

- The HSSC marks of different programs offered by Department of Computer Science were mixed in different files, so first of all the marks of students of BS (CS) were separated from those files.
- The HSSC data is stored in the form of "total marks" and "obtained marks". Since the total marks of different boards and HSSC programs are different, they were converted into percentage.
- Since classifiers perform better for categorical data, the percentages were converted into grades as per the ranges given in table (3-3) below.

| Total HSSC | Obtained HSSC |
|---|---|
| 1100 | 641 |
| 1100 | 635 |
| 1100 | 702 |
| 1100 | 604 |
| 1100 | 634 |
| 1100 | 660 |
| 1100 | 768 |
| 1100 | 565 |
| 1100 | 620 |
| 1100 | 583 |
| 1100 | 741 |
| 1100 | 658 |

**Table 4: Before Pre-Processing**

After pre-processing the prepared categorical data is shown Figure (3-4):

| HSSC % | HSSCGrade |
|---|---|
| 58 | C |
| 58 | C |
| 64 | B |
| 55 | C |
| 58 | C |
| 60 | B |
| 70 | B |
| 51 | C |
| 56 | C |
| 53 | C |
| 67 | B |
| 60 | B |

**Table 5: After Pre-Processing**

## 3.5 Selection of Classifiers

For analysis of data, we have used Weka3.8 (Waikato Environment for Knowledge Analysis), a tool used for data preprocessing and data analysis tasks. WEKA is open source software for data mining and machine learning. It contains large number of state of the art machine learning algorithms. This tool contains classification, regression, association rules, visualization and pre-processing techniques. The next task in this research activity is the selection of the most suitable classifier(s) for the purpose of identification of most relevant attributes behind the dropout of students before the completion of degree. It is one of the most important aspects of the identification of most relevant attributes behind attrition because the usage of the relevant classifier plays a vital role in obtaining the useful results. In order to ensure the best possible effectiveness and efficiency for prediction of most relevant attributes, we have to select the most relevant classifiers with respect to our task.

In order to achieve this objective, we performed a comprehensive literature review of the student retention, student attrition techniques with respect to the classifiers used in those techniques for the purpose of the identifying the factors behind the attrition. During our literature

review, we found out that the classifier that has proved to be the most useful with respect to the student retention task is the tree based classifiers like J48, Decision tree, REP tree, Support Vector machine [5][10][16] . Tree based classifiers are very fruitful for prediction of attributes. These classifiers predict attributes in the form of rules. In this research problem we have needed attributes in the form of rules on which we can check which attributes play important role behind attrition and retention.

Therefore, we decided to use all of the available variations of the Decision tree classifier for the purpose of evaluation. The reason of the usage of all of the available variations of the Decision tree classifier is to ensure that our experiments have been conducted thoroughly and rigorously and to eliminate any doubts about the possibility of improved results using any other variations of the Decision tree classifier. Secondly we want to predict attributes in the form of rules.

## 3.6 Evaluation

The evaluations of our experiments have been carried out by performing a number of distinct tasks. We shall begin the evaluation by loading the data set in Weka (Waikato environment for knowledge analysis), a tool used for data preprocessing and data analysis tasks. The important task in this classification process is the selection of the most suitable test option from the entire given test options. For this purpose, we will select the k-fold cross-validation test option with 10-folds, sometimes known as the 10-k fold cross-validation option.

The 10-k fold cross-validation test option works by dividing the original sample into10 equal sized subsamples. Out of the 10 equal sized subsamples, a single subsample is kept as the validation data for testing purposes, and the remaining 9 equal sized subsamples are used as

training data. The cross-validation process is performed 10 times i.e., equal to the number of folds, with each of the 10 subsamples used exactly once as the training data. The 10 results from the folds can then be averaged or sometimes combined to produce a single estimate. The advantage of this particular test option in comparison to the other test options such as repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once.

After the selection of the 10-k fold cross-validation test option, the next option will be to select the most suitable classifier(s) for the purpose of the prediction of attributes behind the attrition. As mentioned previously in this chapter, we performed a comprehensive literature review in order to identify the most suitable classifier(s) for the purpose of the prediction of attributes behind the attrition of students in higher education institutions and from the literature review, we found out that the Tree based classifier and its different variations have mostly been used in the experiments of this particular type, i.e. for the purpose of prediction of the factors.

For this purpose, we decided to use all of the available variations of the Decision tree classifier for conducting our research. The reason for the usage of all of the available variations of the Decision tree classifier is to eliminate any doubts about the possibility of improved results using any other variations of the Decision tree classifier and to ensure that our research has been conducted in a thorough manner.

In order to achieve this particular research objective, we have used all the variations of the Decision tree classifiers (J48,CART, REP and BFD) that are available in the Weka preprocessing and data analysis tool. The method to conduct the experiments on all of these classifiers is to conduct the experiments on each of these classifiers, one by one, and to record

the results from each classifier in order to perform the comparison of the results from each classifier after the results have been obtained.

Our final results will contain the information such as the tree visualization, prediction attribute, percentage accuracy, correctly classified instances, incorrectly classified instances, precision and recall. Our main focus will be towards the discovery of the attributes that are caused of the attrition of the students from the institution. I also perform a comparison of the results achieved from the usage of different classifiers for the prediction of attributes for the student retention purpose.

This evaluation will help us in deriving our research contribution in the domain of the student retention in higher education institutions in Pakistan. This research will also help the research community and university management and decision makers to actively play their part in this particular domain by identifying the limitations in their admission process and academic procedure by proposing newer and more refined techniques that can provide even better results and reduced dropout rate.

**Summary of Chapter**

This chapter provides detail description of research methodology of whole thesis. In Starting we briefly introduced the motivation of the research and objective. Then next we comprehensively describe the steps how answer our research questions. Further I provide the detail of how data is collected and characteristics of attributes. Then we draw a block diagram of proposed methodology. After that we comprehensively describe step by step how we pre-processed our data to bring into to required form. Section 3.5 describes the background of proposed techniques and in last section we elaborate the evaluation of whole work.

CHAPTER 4

# 4 RESULTS & EVALUATIONS

This chapter explains the results obtained by implementing the methodology that has been described in the previous chapter. Our methodology has been used for the discovery of attributes behind the attrition by applying certain techniques and test options on a particular data set. The primary objective of this research is to identify the attributes of that are cause of the attrition of students.

In order to achieve this particular research objective, we carried out a number of tasks such as conducting a thorough literature review, proposing a unique attribute prediction classification technique, acquiring a relevant data set for the purpose of the prediction of attributes and then performing a number of different experiments in order to determine the attributes for improving the retention rate in higher education institutions. For the experimental purpose, our selection of the classifiers and the test options has been based on the literature review. In order to perform the experiments comprehensively, we have used all of the available variations of the classifier that has been used most frequently for this particular task.

## 4.1 Research Questions

There are four research questions that I will answer after performing the experiments. This will be our research contribution in this particular research activity. The four research questions are given below:

**RQ.1: To check whether oversees identified factors are valid in our local context**

The six most common overseas identified factors are CGPA, ACT score, SSG, HSSG, Financial status, p-occupation. The attributes that we used as input are CGPA, matric grade, Intermediate grade, gender, and city. The tool that we used is WEKA which is a data mining tool. We prepared dataset in .csv format which is compatible with WEKA. The techniques that we used for prediction are J48 decision tree, REP tree, Best first decision tree and CART decision tree. The results for each classifier have been shown below;

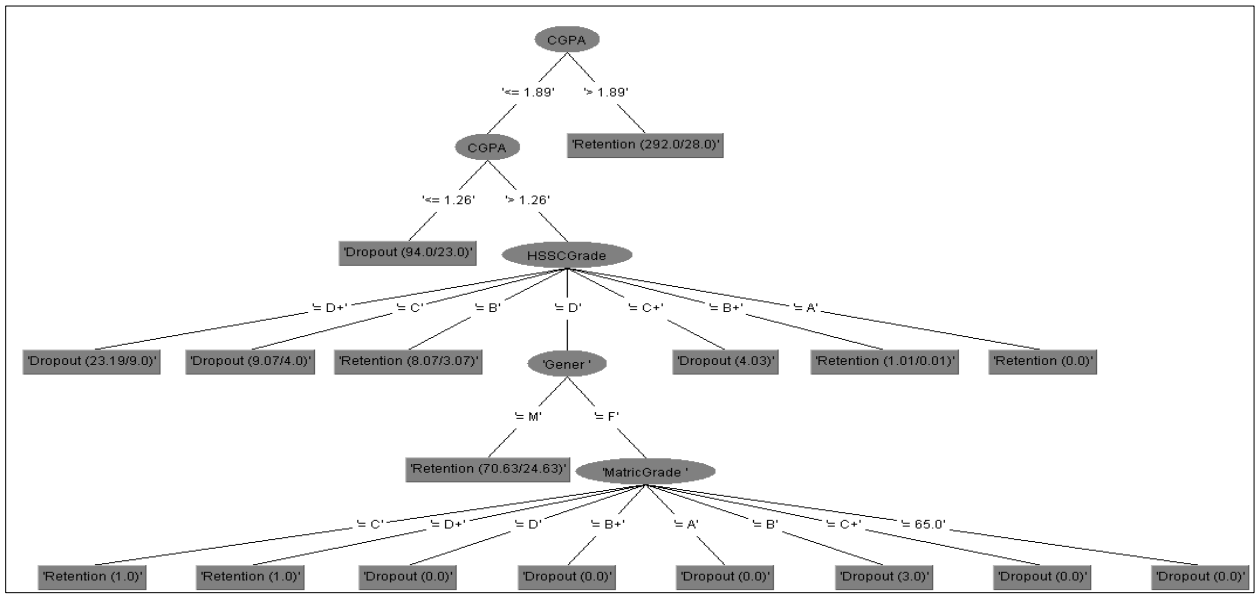


**Figure 3: J48 Decision Tree**

First we applied J48 algorithm on input data. The specialty of J48 classifier is that it produces results in the form of tree and also in the form of rules. It calculates the information gain of each attribute in the data set. The attribute which has highest information gain is on the top of the tree followed by the attributes with minimum information gain at the intermediate and leaf nodes. At each leaf node there are two numeric values that show weight of instances to reach the leaf node and after slash value shows the weight of misclassified instances. The decision tree

generated by J48 algorithm (Figure 3) started with CGPA attribute which has highest information gain and makes it starting node it means the CGPA attribute is most effective in determining attrition. The attributes that participated at lower levels of decision tree are HSSC grade, SSC grade and gender. The first rule that is established from the tree is CGPA and HSSC attribute play effective role for the prediction of attrition. Next important rule established by this experiment is that students which have CGPA greater than 1.89, they mostly (90%) continue their study and students which have grade A, B+ and even B grade in HSSC marks they have less chances of dropout. Another important pattern identified by the classifiers is that gender does not play a significant role in the attrition or retention of the student's even with the weak pre-qualification performance. The accuracy of J48 on the overseas data experiment was almost 90 percent [5] and accuracy achieved by our classifiers were 79 percent. This difference may be due to the unavailability of two attributes (ACT score, occupation) in our local dataset.

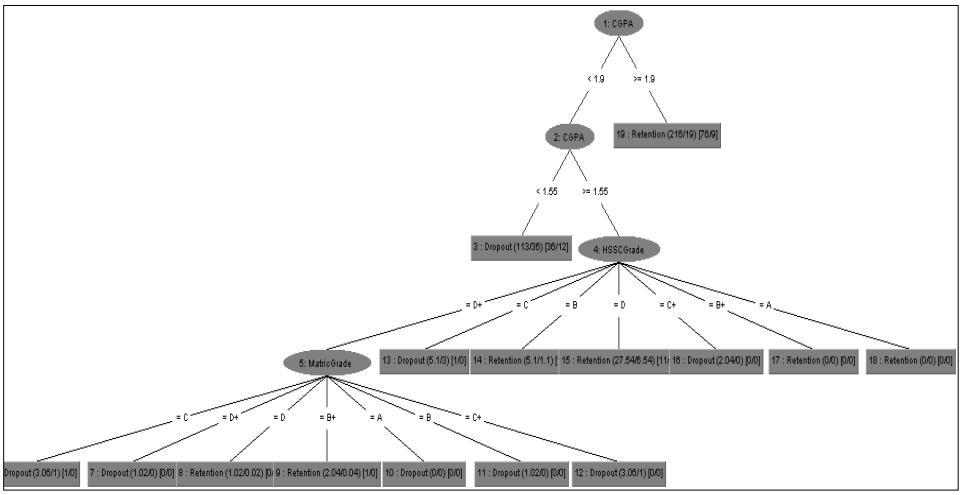


**Figure 4: REP Decision Tree**

Next we applied REP classifier on the input data. The principle of the REP (reduces error pruning) is that it computes the information gain with entropy and minimize the error which arise from variance. The REP algorithm also generates a tree based on their information gain. We can

identify the attributes from the tree against their class label. The tree generated by REP algorithm is shown in (Figure 4). The output of both J48 and REP is almost same, as both place CGPA at the top and most of the rest attributes/rules are same. However, there is one slight difference that J48 considers Gender after HSSC marks and then the matric marks; REP on the other hand does not consider Gender. The precision achieved by REP is 0.778.

```
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

Best-First Decision Tree

CGPA < 1.57: Dropout(104.0/51.0)
CGPA >= 1.57: Retention(303.0/49.0)

Size of the Tree: 3

Number of Leaf Nodes: 2

Time taken to build model: 0.03 seconds
```

**Figure 5: BFD Decision Tree**

Then we applied Best First Decision tree algorithm on input data. The result of BFD is shown in (Figure 4-3) and is similar to that of J48 with a precision of 0.795

```
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

CART Decision Tree

CGPA < 1.57: Dropout(104.0/51.0)
CGPA >= 1.57
|   CGPA < 2.025
|   |   HSSCGrade=(B+)|(C+)|(D+)
|   |   |   MatricGrade =(D+)|(D)|(A): Dropout(4.04/0.0)
|   |   |   MatricGrade !=(D+)|(D)|(A)
|   |   |   |   CGPA < 1.84: Dropout(6.19/2.0)
|   |   |   |   CGPA >= 1.84: Retention(8.0/2.0)
|   |   HSSCGrade!=(B+)|(C+)|(D+): Retention(56.0/16.76)
|   CGPA >= 2.025: Retention(237.0/20.0)

Number of Leaf Nodes: 6
```

**Figure 6: CART Decision Tree**

Then we applied CART (classification and regression tree) algorithm. This classifier handles both continuous and categorical attributes to build a decision tree. It also handles

missing values. This algorithm select attributes using Gini index to build a decision tree. The CART algorithm uses pruning to remove the unreliable branches from the tree to improve the accuracy. The tree generated by CART algorithm started from CGPA attribute. This shows that CGPA is very important attribute to determining the attrition of the students.

```
CGPA='(0.575-1.505]' 127 ==> Class=Dropout 82    conf:(0.65)
Gener =M City=R HSSCGrade=D 134 ==> Class=Retention 86    conf:(0.64)
Gener =M CGPA='(0.575-1.505]' 121 ==> Class=Dropout 76    conf:(0.63)
HSSCGrade=D CGPA='(0.575-1.505]' 85 ==> Class=Dropout 52    conf:(0.61)
```

**Figure 7: Rule generated by Aprori**

The table (6) describes the accuracy, precision, recall and f-measure of applied techniques on above data using 10-fold cross validation.

| Techniques | Accuracy | Precision | Recall | F-measure | Overseas Accuracy | References |
|---|---|---|---|---|---|---|
| **BFD** | 79 | 0.795 | 0.797 | 0.796 | 85% | [16] |
| **J48** | 79 | 0.787 | 0.791 | 0.789 | 85% | [5][17] |
| **REP** | 78 | 0.785 | 0.789 | 0.787 | 84% | [17] |
| **CART** | 79 | 0.795 | 0.797 | 0.796 | 85% | [16] |

**Table 6: Classifiers Accuracy**

**Summary:** To answer first research question the selected attributes were CGPA, SSS grade, HSSC grade, gender and city. We prepared data in .csv format and input to the tool to check whether overseas identified factors are valid in our local context. We applied Best first decision tree, CART algorithm, J48 algorithm and REP algorithm on whole dataset. The result shows that CGPA, and HSSC grade are the strongest attributes behind the attrition of the students. The result shows that oversee identified factors are valid in our context with relatively less accuracy which can be contributed to the unavailability of two overseas features in our local data and secondly the variations in size of data sets. For example,[16] paper have used rich data

set in horizontally as well as vertically. They have used 3000 records of students to conduct experiments. Furthermore, they have used 13 attributes (Branch, Sex, Cat, HSG, SSG, Atype, Med, Lloc, FAIn, Fqual, Mqual, Focc, Mocc) horizontally. The difference of accuracy is due to variations in size of dataset.

**RQ.2: To find most influential subjects behind the attrition in first semester.**

In the second question, our focus was to highlight the particular subjects that have more impact on attrition. Since the maximum attrition rate is in first semester as shown in table (3-1), we used subjects of first semester along with the pre-qualification to answer this question. Once again, we used the same dataset of BS (CS) students from semester 141 to 171. The input attributes were SSG, HSSG, Introduction to Computing grade or Computer Programming grade (ITCgrd /CPgrd), grade in English-I course (ENGgrd) and grade in Calculus-I course (CALgrd). The proposed tool was WEKA which is data mining tool. We prepared dataset in .csv format which is compatible with WEKA. The techniques used were J48 decision tree, REP tree, Best first decision tree, and CART decision tree. The result for each approach is shown below;

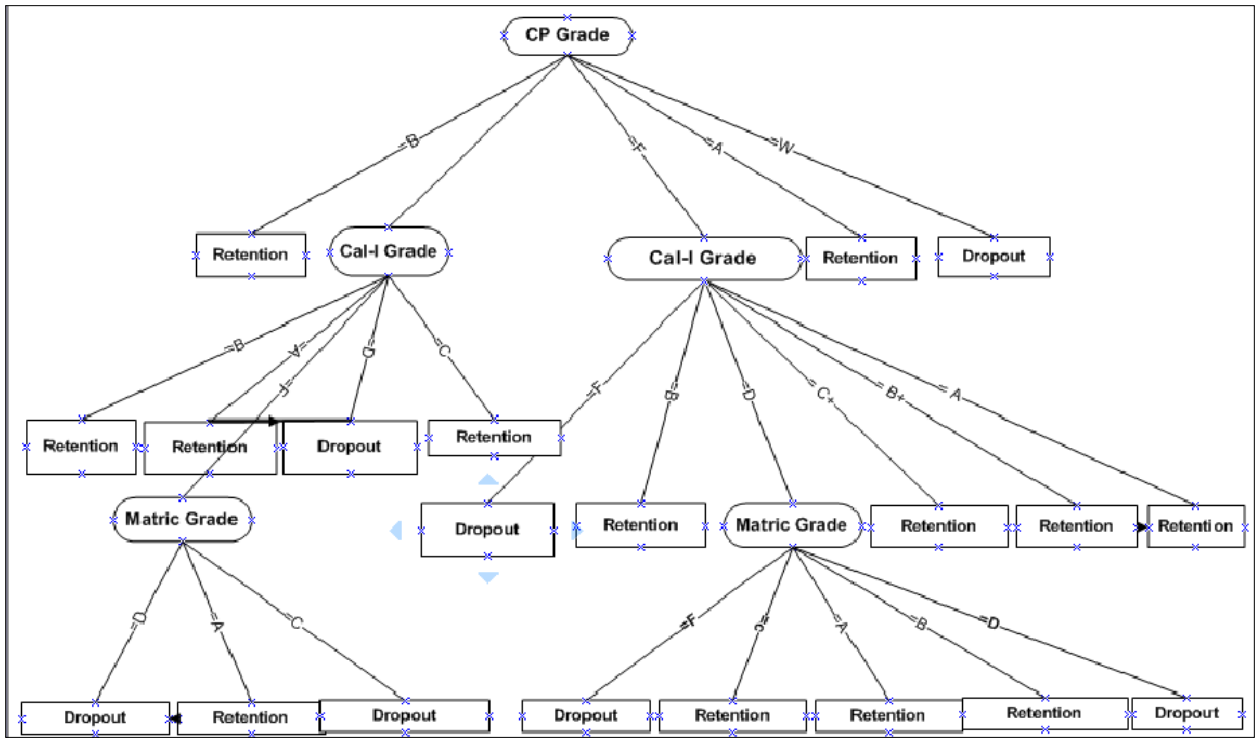


**Figure 8 : J48 Decision Tree**

First we applied J48 algorithm on input data. The output is shown in Figure (8). The tree started from CP grd it means this attributes has highest information gain and it is most influential attribute in determining the attrition of the students. Other attributes which participate in the tree are ITC grd and CALgrd. The rule which is established from the tree is that Calculus and ITC or CP grades play effective role on attrition of the students. So the findings are if the department wants to reduce attrition rate then they will have to pay special attention to these two subjects. The accuracy achieved by J48 and REP classifiers is 79%.

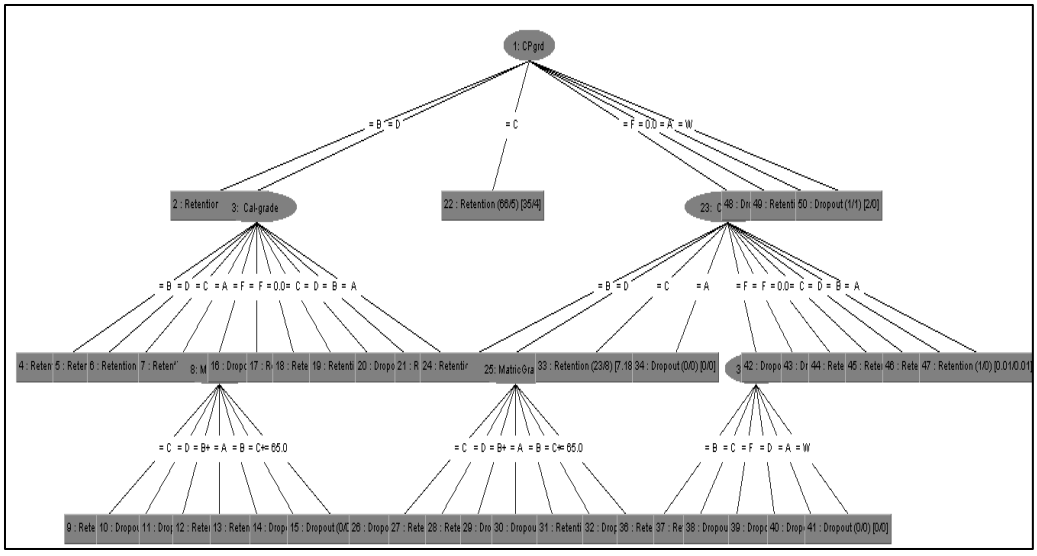


**Figure 9: REP Decision Tree**

Figure (10) and figure (11) shows the outputs of three other classifiers that we used to answer the second research question. The findings through these three classifiers are similar to the first one (decision tree induction), that is, most significant subject in attrition is computer programming (CP) and Calculus is the next one. The accuracy achieved by these both classifiers is 78%.

```
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

Best-First Decision Tree

CPgrd=(0.0)|(W)|(F)|(D+)|(D)
|   Cal-grade =(A)|(0.0)|( F )|( B- )|(F)|(D)|(A-)|( A ): Dropout(103.11/44.0)
|   Cal-grade !=(A)|(0.0)|( F )|( B- )|(F)|(D)|(A-)|( A ): Retention(86.0/32.89)
CPgrd!=(0.0)|(W)|(F)|(D+)|(D): Retention(224.0/17.0)

Size of the Tree: 5

Number of Leaf Nodes: 3

Time taken to build model: 0.78 seconds
```

**Figure 10: BFD Decision Tree**

```
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

CART Decision Tree

CPgrd=(0.0)|(W)|(F)|(D+)|(D)
|   Cal-grade =(A)|(0.0)|( F )|( B- )|(F)|(D)|(A-)|( A ): Dropout(103.1/44.0)
|   Cal-grade !=(A)|(0.0)|( F )|( B- )|(F)|(D)|(A-)|( A ): Retention(86.0/32.89)
CPgrd!=(0.0)|(W)|(F)|(D+)|(D): Retention(224.0/17.0)

Number of Leaf Nodes: 3

Size of the Tree: 5
```

**Figure 11: CART Decision Tree**

| Techniques | Accuracy | Precision | Recall | F-measure | Overseas Accuracy | References |
|---|---|---|---|---|---|---|
| **BFD** | 79 | 0.795 | 0.797 | 0.796 | 85% | [16] |
| **J48** | 79 | 0.787 | 0.791 | 0.789 | 85% | [5][17] |
| **REP** | 78 | 0.785 | 0.789 | 0.787 | 84% | [17] |
| **CART** | 79 | 0.795 | 0.797 | 0.796 | 85% | [16] |

**Table 7: Accuracy of Classifiers**

**Summary:** To answer second research question the selected attributes were SSG, HSSG, ITC/ CPgrd, ENGgrd, CAL-Igrd. We have prepared data in .csv format and input to the tool to check whether overseas identified factors are valid in our local context. We applied J48 algorithm, REP decision tree, Best First Decision Tree and CART algorithm. The result shows **CP-grade and Cal-I grade** are main attributes behind the attrition of the students. The accuracy of our classifiers is relatively less than oversees classifiers; for variations in size of data sets and difference of attributes.

**RQ.3: To check the impact of Teacher methodology behind the attrition**

To answer third research question, we prepared dataset in excel. To conduct first experiment we have collect BSCS student data from 141 to 171. The input attributes were: SSG, HSSG, ITC teacher, CP teacher, English teacher, and Calculus teacher. The proposed tool was WEKA which is data mining tool. We have prepared dataset in .csv format which is compatible with WEKA. The technique used was REP decision tree. Best First Decision tree and some statistical approaches. The results are shown as below;

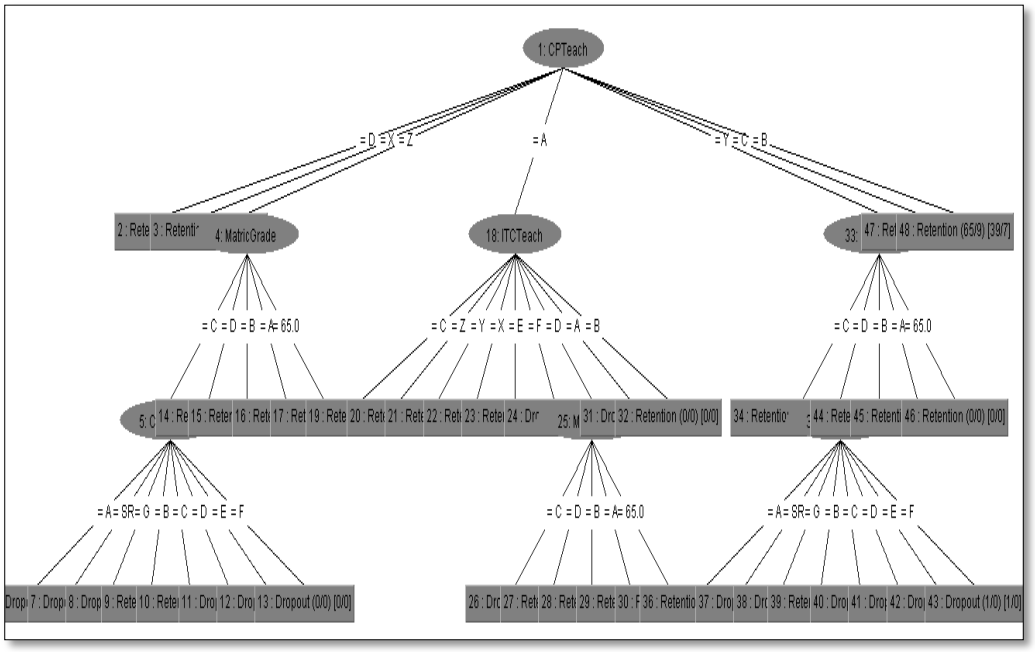


**Figure 12: REP Decision Tree**

The tree started from CP teacher attribute which shows it is very effective attribute. It means that CP teacher attribute plays important role for determining the prediction of attrition of students. Other attribute which participated in the tree are ITC teacher, Cal-I teacher and Matric grade. The tree indicated that all attributes have effect on attrition of students but the most

effective attributes are CP/ITC teacher and Cal-I teacher. The accuracy achieved by REP classifier is 70 %.

```
Best-First Decision Tree

CPTeach=(A)|(C)|(X)|(Z)|(Y)
|  EngTeach=(X)
|  |  CalTeacher =(G)|(E)|(SR)|(B)|(C)|(D)
|  |  |  HSSCGrade=(B)|(D)|(A)
|  |  |  |  ITCTeach=(F)|(A)|(Y): Dropout(15.0/1.0)
|  |  |  |  ITCTeach!=(F)|(A)|(Y)
|  |  |  |  |  MatricGrade =(C)|(A)|(65.0)
|  |  |  |  |  |  CPTeach=(X)|(D)|(A)|(Y)|(C)|(B): Dropout(7.0/1.0)
|  |  |  |  |  |  CPTeach!=(X)|(D)|(A)|(Y)|(C)|(B): Retention(2.0/2.0)
|  |  |  |  |  MatricGrade !=(C)|(A)|(65.0): Retention(4.0/2.0)
|  |  |  HSSCGrade!=(B)|(D)|(A)
|  |  |  |  MatricGrade =(B)|(A)|(65.0)
|  |  |  |  |  ITCTeach=(Y)|(C)|(Z)|(X)|(E)|(F)|(D)|(B): Dropout(4.0/0.0)
|  |  |  |  |  ITCTeach!=(Y)|(C)|(Z)|(X)|(E)|(F)|(D)|(B): Retention(3.0/0.0)
|  |  |  |  MatricGrade !=(B)|(A)|(65.0): Retention(5.0/0.0)
|  |  CalTeacher !=(G)|(E)|(SR)|(B)|(C)|(D): Retention(29.0/16.0)
|  EngTeach!=(X): Retention(193.0/69.0)
CPTeach!=(A)|(C)|(X)|(Z)|(Y): Retention(128.0/26.0)
```

**Figure 13: BFD Decision Tree**

The Best First Decision tree is shown in the figure (13). There are multiple paths from start to root node. Best First Decision tree shows also interesting patterns on student data. For example, CP/ ITC teacher impact on attrition of students. The accuracy achieved by Best First Decision tree is 70%.

## 4.2 Statistical Approaches

To check the impact of teacher methodology on attrition of the students we applied some statistical approaches on give data set. The input attributes were:SSG, HSSG, ITC teacher, CP teacher, English teacher, and Calculus teacher. The selected tool was Microsoft excel. The below graph shows the impact of teacher methodology on attrition of the students in introduction to computer (ITC) course;

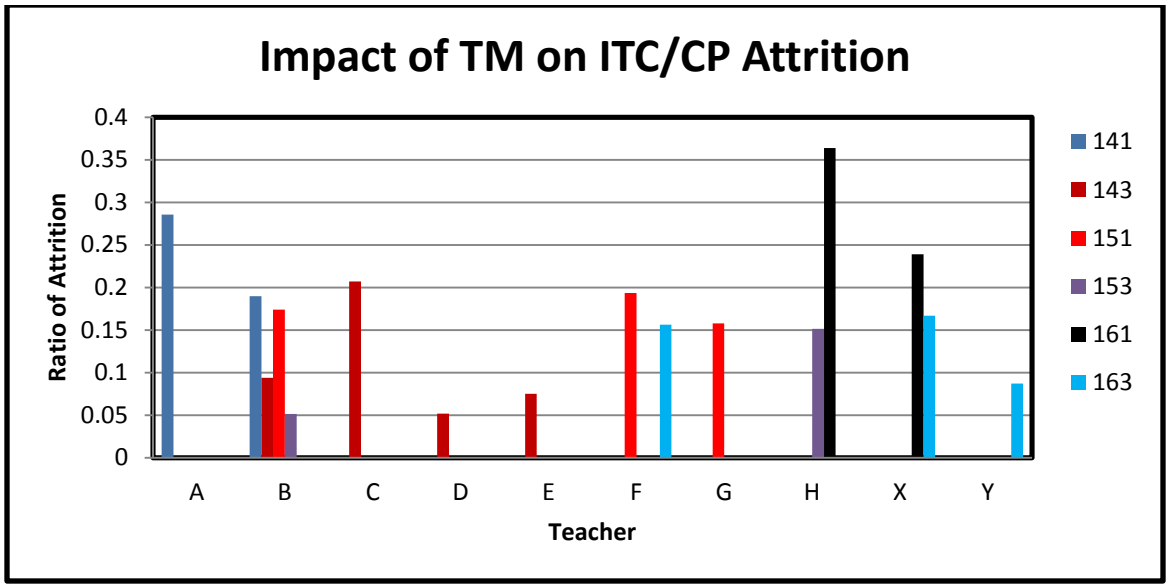


**Figure 14: Impact of ITC/CP Teacher**

We collect data of ITC from term 141 to 163 of each section. In figure (4-13) x-axis shows teacher name and in Y-axis shows blue bar shows total number of students that a teacher taught and red bar shows number of students drop out against each particular teacher. The result shows that there is no significant impact of teacher methodology on attrition of students in ITC course. Furthermore, we have observed that teacher 'Y' and teacher 'D' have relatively weak retention rate and teacher X and F have relatively good retention rate. We have forwarded our findings to the department for appropriate decisions.

Next to check the impact of teacher on attrition of students in Computer programming we collect data of CP from term 141 to 163. The collected data is CP (Computer Programming) their grade, and teacher name against each section. In figure (4-14) x-axis shows teacher name and in Y-axis shows blue bar shows total number of students that a teacher taught and red bar shows number of students drop out against each particular teacher. The result shows that there is no significant impact of teacher methodology on attrition of students in ITC course. Furthermore,

we have observed that teacher ‘X’ and teacher ‘Y’ have relatively low retention rate and teacher ‘D’ and “B” have relatively good retention rate.

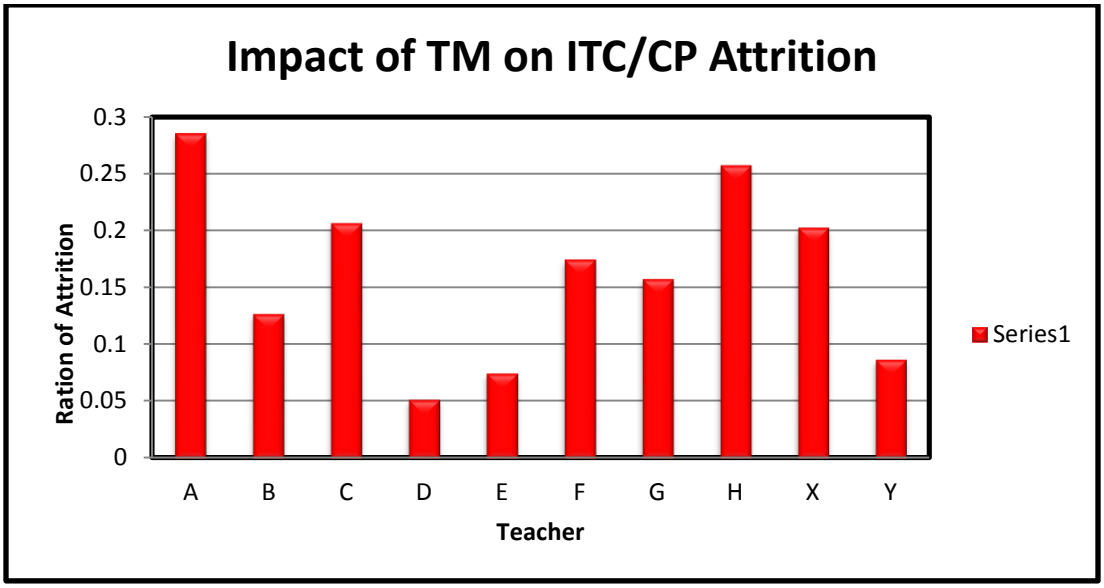


**Figure 15: Average Impact of Teacher on Attrition**

Further to check the impact of teacher on attrition of students in Calculus we collected data of Cal-I from term 141 to 163. The collected data is Calculus (Cal-I) grade, and teacher name against each section. In figure (4-15) x-axis shows teacher name and in Y-axis shows blue bar shows total number of students that a teacher taught and red bar shows number of students drop out against each particular teacher. The result shows that teacher ‘A’ and teacher ‘Y’ have low attrition and high retention rate and teacher ‘Z’ and “C” have relatively weak retention rate and high attrition rate. We have forwarded our findings to the department for appropriate decisions.

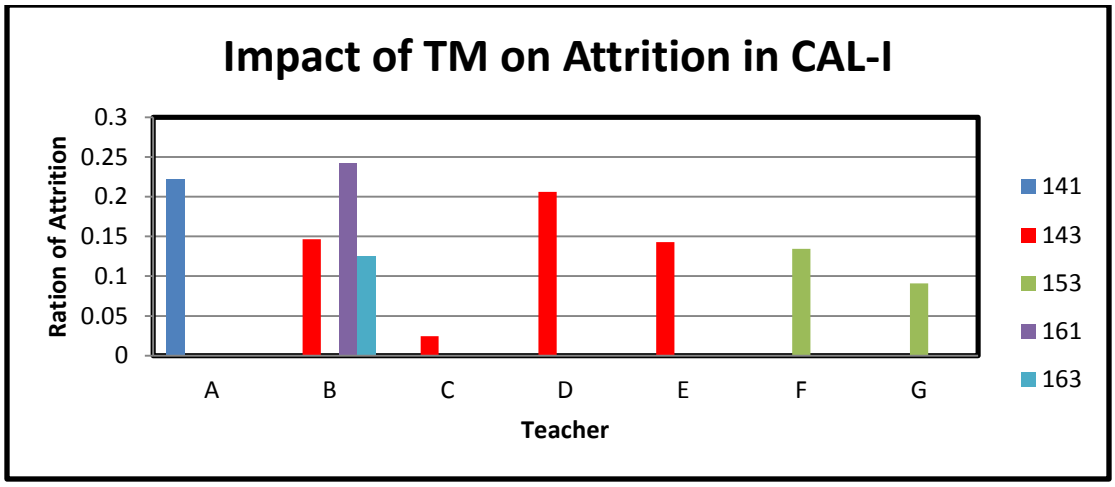


**Figure 16: Impact of Cal-I teacher on Attrition**

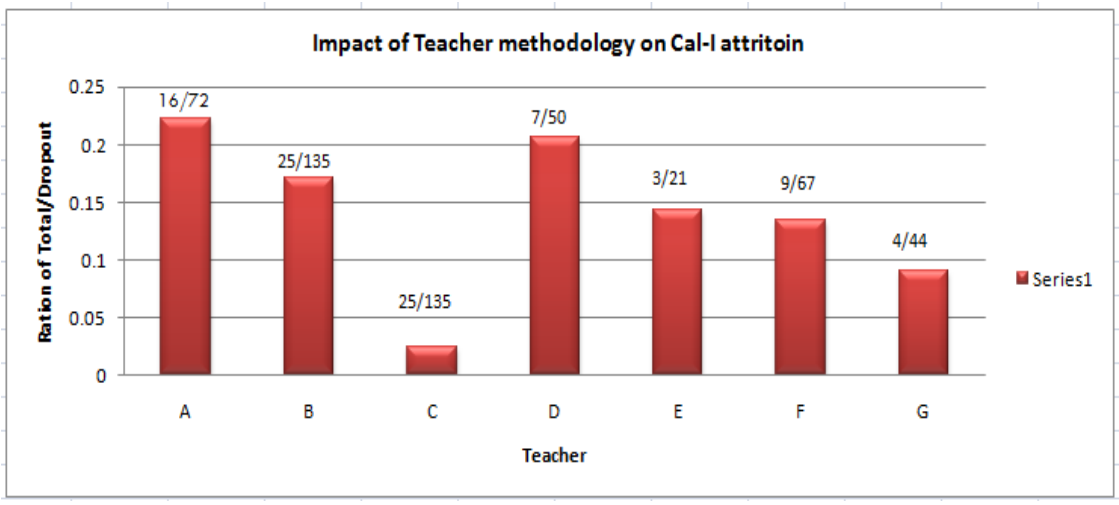


**Figure 17: Impact of Cal-I teacher on Attrition**

**RQ.4: To check the impact of introducing tutorials in first semester on student performance**

Department started offering tutorials of Calculus-I and Computer Programming courses from Fall 2016 semester. In this research question, we wanted to evaluate impact of tutorial over attrition. We took the pre-qualification data (SSSC and HSSC grades), tutorial offered and class label (dropout or retained). The students from Spring 2014 till Spring 2016 (5 batches) were not

offered tutorial, whereas the batch of Spring 163 were offered tutorial, The output of J48 and REP has been shown in figure (18) and figure (19).

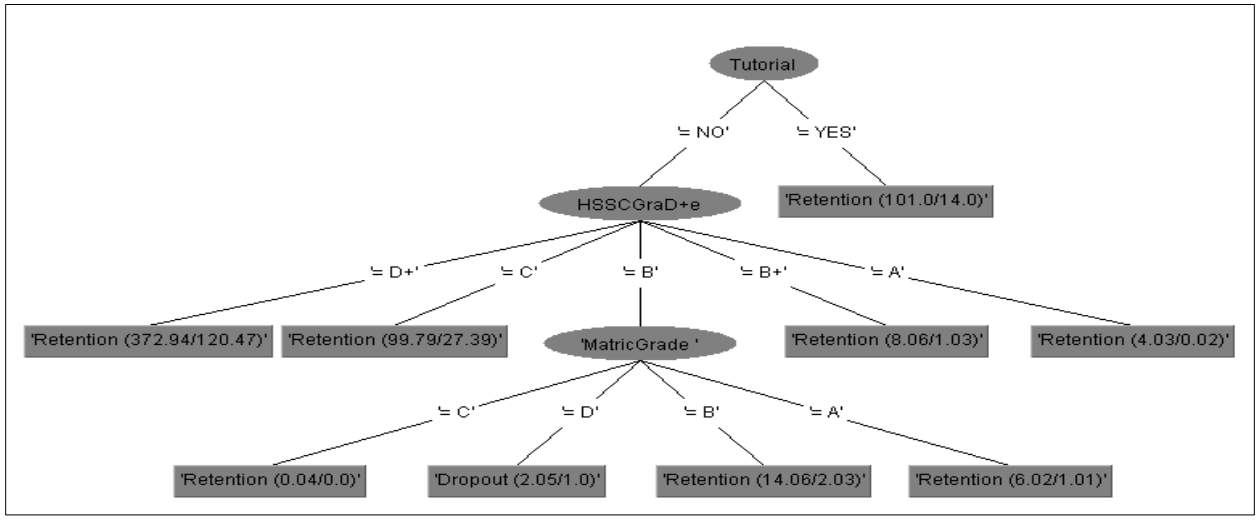


**Figure 18: J48 Decision Tree**

The tree generated by J48 algorithm is shown in the figure (18). The tree started from tutorial attributes it means this attribute play very effective. The rule which is established from the tree is that students in which tutorial is offered they have greater chances of retention. The next important rule which is established from the tree is that the term in which tutorial is not offered they depend upon HSSC grade and SSC grade. The accuracy achieved by J48 classifier is 70%.

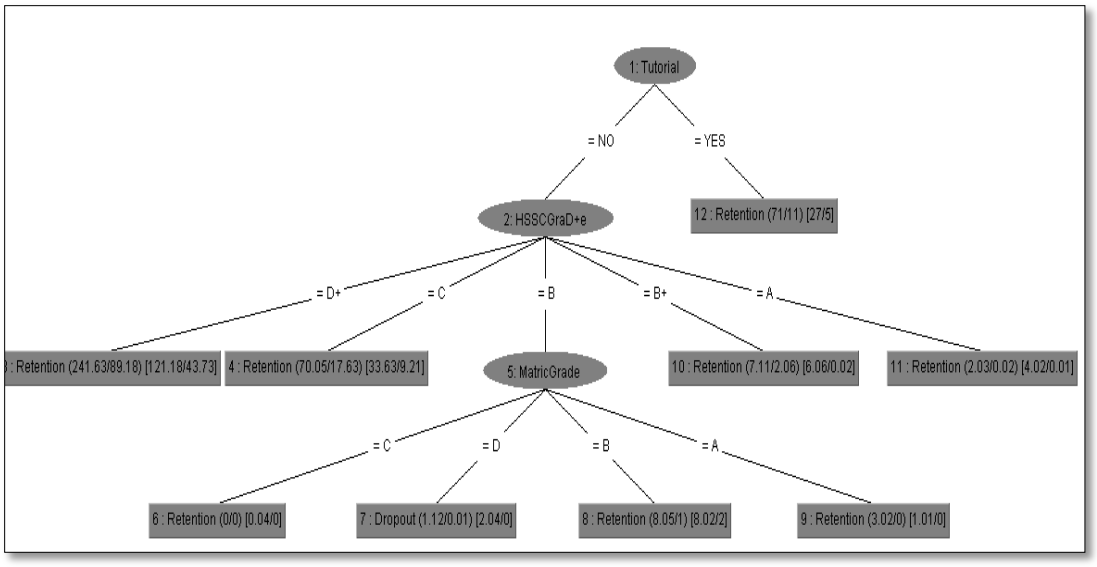


**Figure 19: REP Decision Tree**

The tree generated by REP algorithm is shown in figure (19). The tree started from tutorial attribute which shows it is very effective attribute. There are some numeric values on the leaf node. The first value in the brackets () is the amount of correctly classified instances from the training set under that leaf while the second value is the amount of instances which were under the leaf but had a different classification value, the second value in the [] brackets shows the amount of correct classification from the pruning set and the second number is the wrong classifications. The rule which is established from the tree is that students in which tutorial is offered they have greater chances of retention. The next important rule which is established from the tree is that the term in which tutorial is not offered they depend upon HSSC grade and SSC grade. The accuracy achieved by REP classifier is 70%.

Further to check the impact of tutorial on performance of students we applied some statistical approaches in Microsoft excel on collected data. We collect Calculus final grade from term 143 to 163.In figure (4-18) x-axis shows grade earned students from 143 to 163 and y- axis

shows percentage of students. The graph shows that performance of students in 163 term is slightly better as compared to other terms.

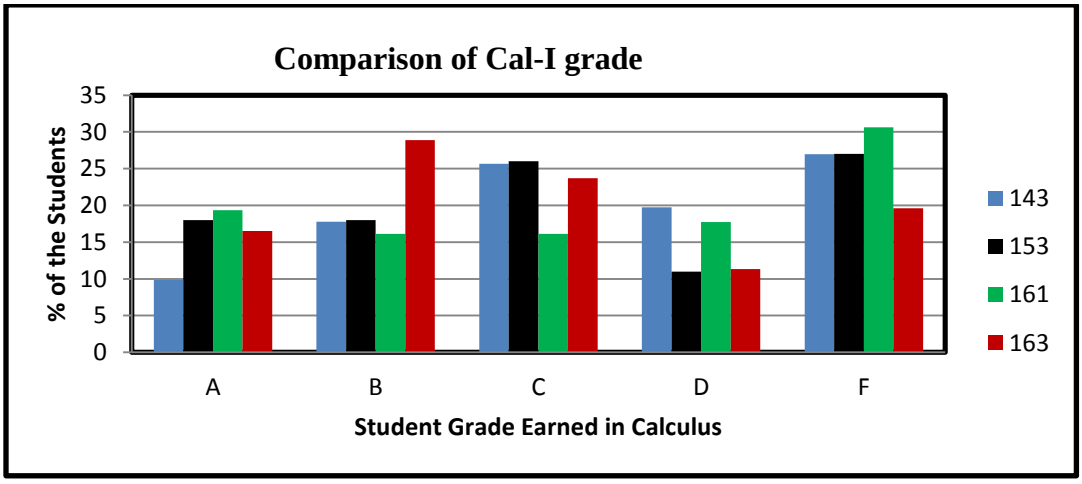


**Figure 20: Cal-I Grade Comparison**

Next we calculate average grade of each term and compared the performance of terms. This graph more clearly describes the impact of tutorial on term 163. In this figure (20) x-axis shows the number of terms and Y-axis shows the average percentage of each terms. From the graph we can observe that the average grade of term 163 is better than others. From these findings we can established that tutorial has produced impact on performance of students.

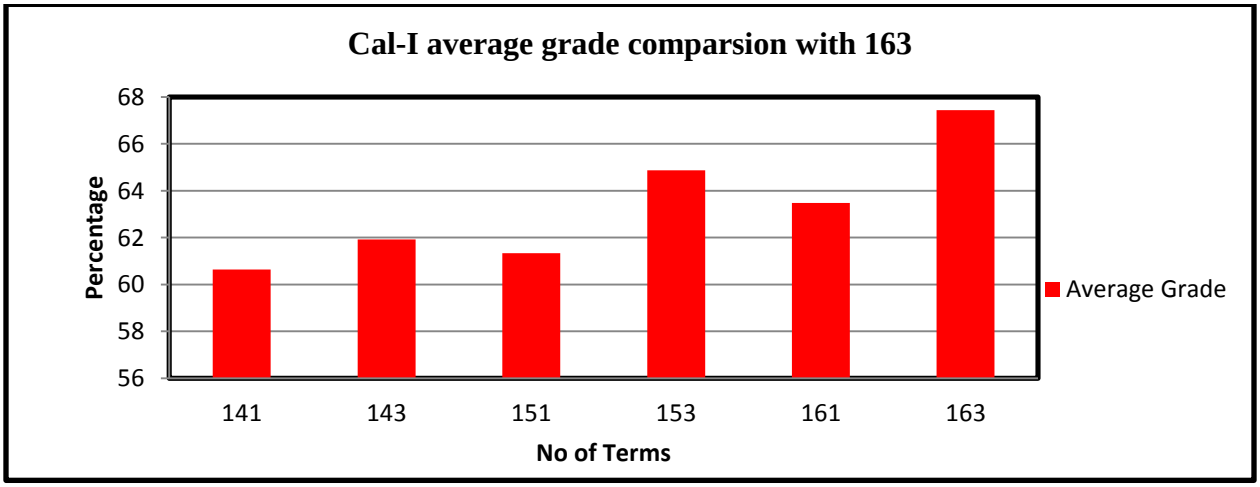


**Figure 21: Cal-I average grade comparison**

Next we collect Computer programming final grade from term 143 to 163.In figure (21) x-axis shows number of terms and y- axis shows percentage of students grade earned. The graph shows that performance of students in 163 term is slightly better as compared to other terms. The first rule which is established from the graph is that percentage of students having "F" grade is less in term 163 as compared to other terms. The second rule which is established from the graph is that the percentage of students having "A" grade in term 163 is slightly better than remaining terms.

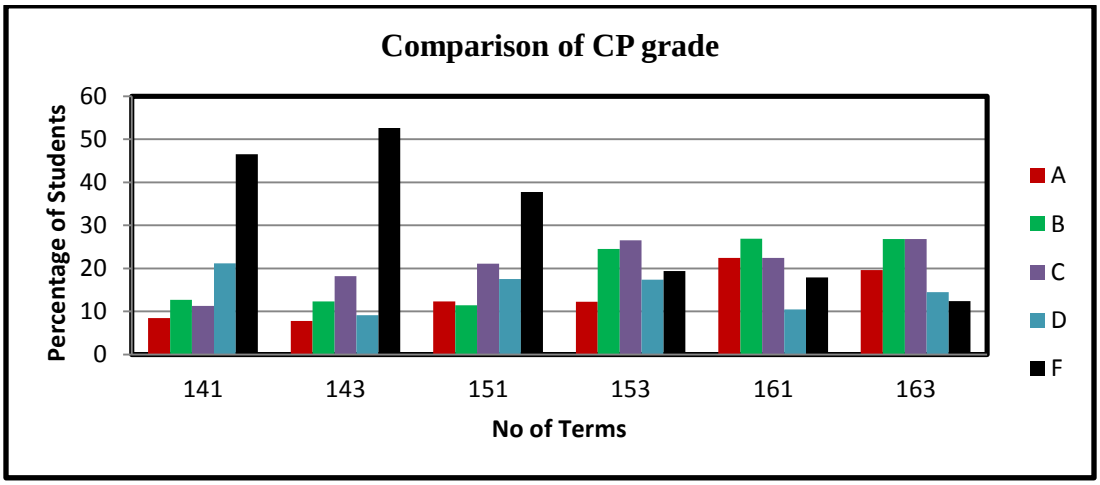


**Figure 22: CP/ITC Grade Comparison**

Next we calculate average grade of each term and compared the performance of terms. This graph more clearly describes the impact of tutorial on term 163. In this figure (4-21) x-axis shows the number of terms and Y-axis shows the average percentage of each terms. From the graph we can observe that the average grade of term 163 is better than others. From these findings we can established that tutorial has produced impact on performance of students.

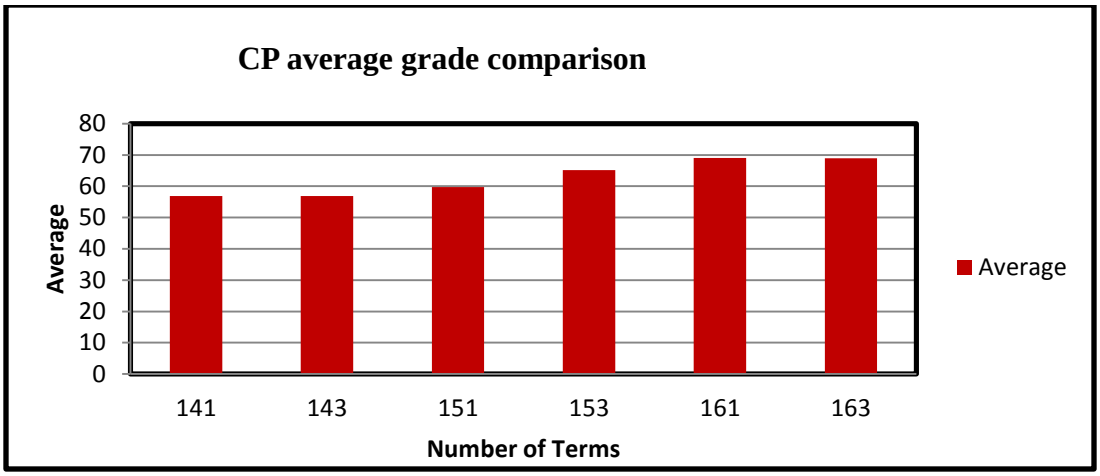


**Figure 23: CP/ITC average grade comparison**

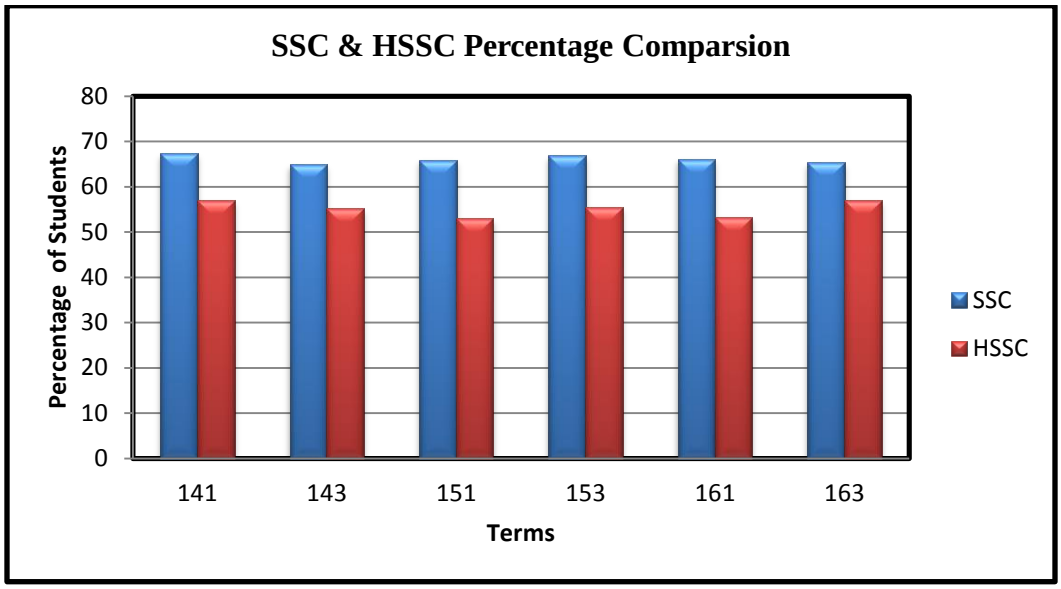


**Figure 24: SSC & HSSC Percentage comparison with Tutorial**

In figure (24) the x-axis shows the number of terms and Y-axis shows the number of students. The graph shows that the overall performance of students in SSC and HSSC is almost same from all terms. So this graph reveals that the attrition rate of 163 is reduced due to tutorial impact not by SSC and HSSC percentages.

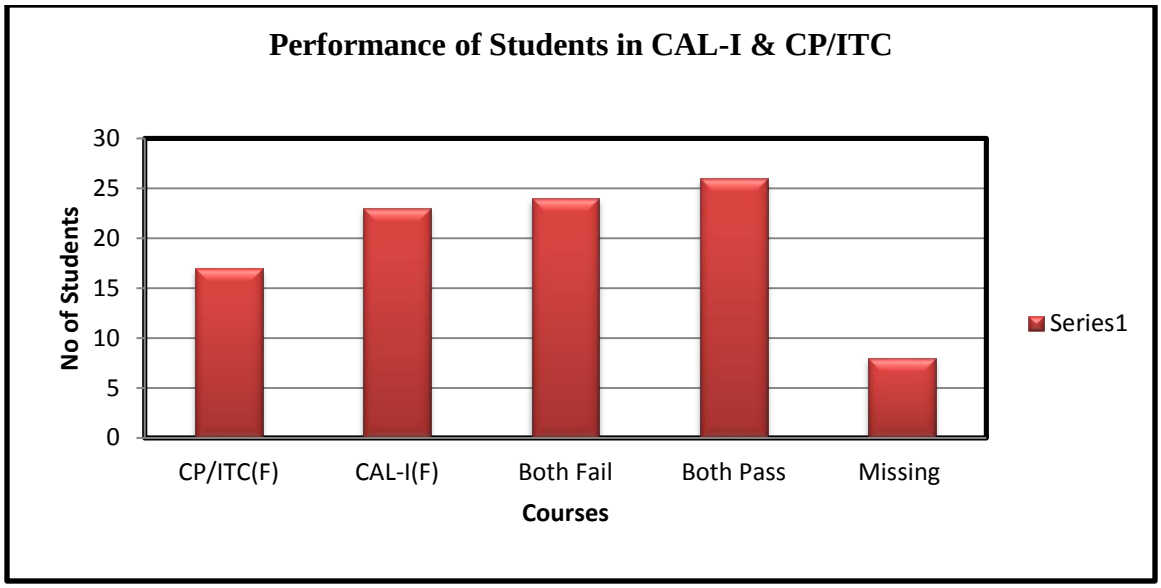


**Figure 25: Attrition Comparison**

The above graph shows the attrition rate of students in CP/ITC and CAL-I. The x-axis shows the students which fails only Cal-I and CP/ITC and which fails/ pass in both courses. The results shows that overall percentage of students which pass in both subjects is high but we take the average of CP, Cal-I and both their failure rate is high from both pass.

## 4.3 General Findings

This section describes the general findings of our research work. We have applied some statistical approaches on whole data set in Microsoft excel. We collect data of students from 141 to 161term. The total instances were 507 and total dropout students were 154. Then we collect CGPA data of all students and find that who much CGPA or grade impact on performance of students.

In below figure (24) the x-axis shows the grade earned by students and Y-axis shows the percentage of student's dropout and retention. The total instances were 507 and including 154 dropouts. The graph shows that students having highest grade like "A" and "B" they have good

retention percentage and students having poor grade like “D” and “F” they have poor retention percentage and their attrition rate is high. This shows that CGPA is clearly impact on attrition of the students.

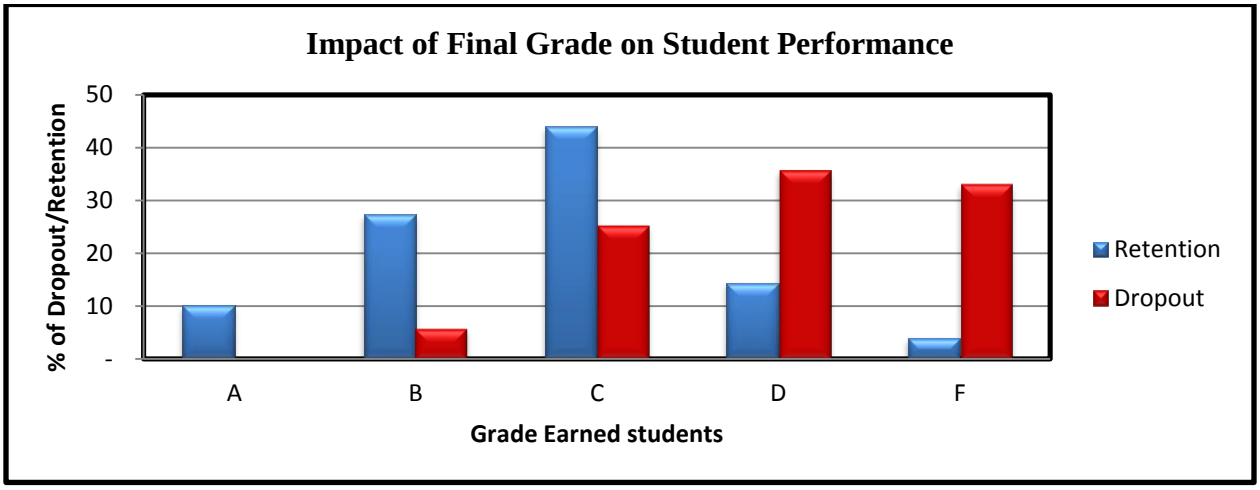


**Figure 26: Student Performance against Grade**

The below graph showed the performance of students in CP against their GPA. In figure (25) the x-axis shows the number of terms and Y-axis shows the no of dropout students. The rule which is established from the graph that the students which have GPA (0.00 to 1.00) in computer programming course they have highest dropout rate**.**

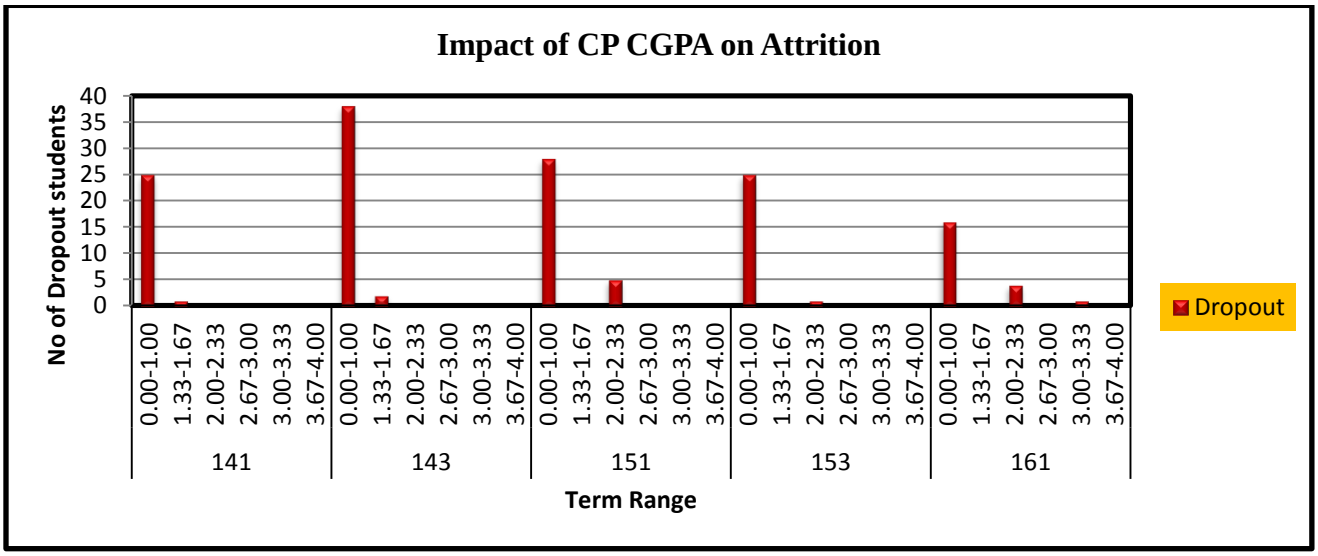


**Figure 27: CP Dropout Rate against GP**

The below figure show the impact of pre-college grade on student performance. In figure (25) the x-axis shows the HSSC percentage and Y-axis shows the number of dropout and retention students. The rule which is established from the graphs is that the students which have low percentage in HSSC they have greater chances of dropout. The dropout ratio is high at low percentage. The dropout ratio is decreasing as percentage in HSSC is increased.

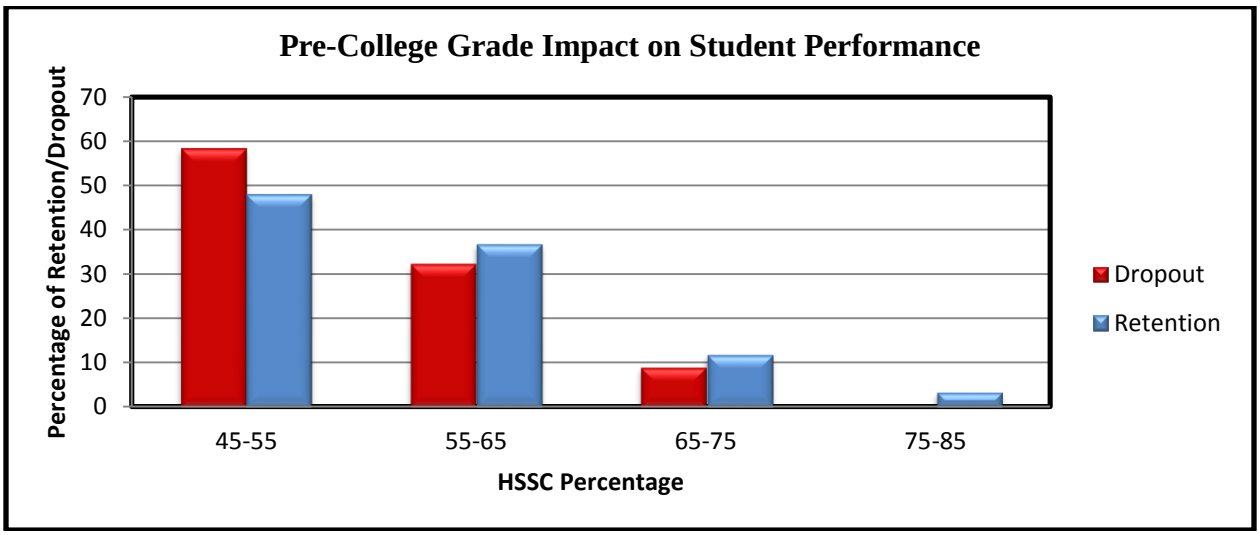


**Figure 28: Pre-College Grade Impact on Student Performance**

**Summary:** This chapter presents the results obtained from proposed methodology utilizing the data collected from registrar office. The most significance factors that affecting student retention was: The result suggested that persister students had better CGPA as compared dropout students. In addition the dropout students had poor performance in CP/ITC and Cal-I courses. However, there is clear difference between attrition rate associated with different teachers' classes that MAY be due to teachers' methodology that need to be further investigated. Furthermore it is investigated that performance of students has improved due to tutorial period.

CHAPTER 5

# 5 CONCLUSION & FUTURE WORK

## 5.1 Conclusion

Student retention is a serious issue in the higher education institutions since many years. As a result, the scientific community has been focused towards solving this problem by proposing data mining tools and techniques. From university perspective it is very costly and time consuming to bring new students in the system. This issue also impact personally on student's life.

In this thesis a comprehensive literature related to topic is critically reviewed and their strengths and weakness of these state of the art approaches are discussed and their results and limitations are presented. Initially this issue was solved through traditional statistical approaches but from last couple of year's data mining approaches have been proposed to solve this problem. Their result reveals that data mining approaches performs well. A student enters in the institutions with some family background and pre college characteristics. Researchers used student demographic, pre-college and academic data to conduct their experiments. The study shows that majority of researchers has used only one category of data to conduct experiments. The research study shows that GPA, ACT score, SSG, HSSG, and p-occupation attribute play critical role behind the attrition of the attrition. I have seen that majority of research has been done in foreign institutes and they have used limited number of attributes to conduct their experiments.

The motivation behind this thesis is to use attributes from three domains demographic, pre-college and academic to perform experiments. Further we have checked whether the overseas identified factors behind the attrition are valid in our context. Therefore in this thesis we collect data from demographic, pre-college and institutional domain.

To conduct experiments, a comprehensive data set of BSCS students of Capital university of Science and Technology has been used. The dataset has been collected from registration department. The dataset consist of academic, demographic and pre-college attributes. Some attributes are used from previous research because they play important role in attrition of the students and some of them were new which we want to check whether they play part behind the attrition of the students.

The Weka data analysis tool was used for the experiments. According to the literature review, all variations of the Tree based classifier were used for the classification task. For the evaluation, the 10-k fold cross validation method was used. The results were exhaustively compared and explained for each experiment. The findings of this research are discussed below;

The objective of this research work was to check whether the overseas identified factors GPA, ACT score, SSG, HSSG, and p-occupation are valid in our local context. For achieving this objective first we collect the required attributes CGPA, HSSC grade, SSC grade, Gender and city. Before exploring the objective of the research work the collected data set was verified and then pre-processed in required form. The result shows that CGPA and HSSC grade are very effective for predicting the attrition of the students. Further, to check the most influential subjects behind the attrition of students we prepared another experiments data file which consists of Computer programming course grade, Introduction to computer grade, Calculus grade and

English grade of first semester. The proposed techniques applied on input data the results reveals that computer programming (CP), Calculus (Cal-I) and introduction to computer (ITC) are effective attributes for predicting the attrition of the students. Then to check the impact of teacher methodology on attrition of students we prepared another file for experiments which include teacher name of all above mentioned courses from 141 to 163 terms. Then applied data mining and some statistical approaches on it he result reveals that computer programming and calculus teacher methodology impact slightly on attrition of students. Then to check the impact of tutorial on student performance we collected data of 163 term in which tutorial was offered and applied data mining and some statistical techniques on it the results reveals the performance of students term 163 in which tutorial was offered is slightly better than other students. These findings are important for researchers and decision maker. The decision makers of Capital University of science and technology can use these finding to improve the retention rate of institution.

## 5.2 Future work

This research was conducted in private institute of Pakistan. This can be extended in other public large residential universities in Pakistan. Secondly, this study was conducted on undergraduate computer science students further it can be extended for MS and PHD students and conducted on large datasets. Thirdly, this research was conducted on computer science department's data set in future this research can be extended in other departments like management and engineering.

# CHAPTER 6

# 6 REFERENCES


Jayanthi, M. A., Kumar, R. L., Surendran, A., & Prathap, K. (2016, October). Research contemplate on educational data mining. In *Advances in Computer Applications (ICACA), IEEE International Conference on* (pp. 110-114). IEEE..

Shah, R. (2014). The Approach, Methods and Impact of a Non-Governmental Organization in Education of Child Workers in India: A Case Study..

Nandeshwar, A., Menzies, T., & Nelson, A. (2011). Learning patterns of university student retention. *Expert Systems with Applications*, *38*(12), 14984-14996. [2.879]

Pal, S. (2012). Mining educational data to reduce dropout rates of engineering students. *International Journal of Information Engineering and Electronic Business*, *4*(2), 1..

Jia, J. W., & Mareboyana, M. (2013). Machine Learning Algorithms and Predictive Models for Undergraduate Student Retention. In *Proceedings of the World Congress on Engineering and Computer Science* (Vol. 1).

Djulovic, A., & Li, D. (2013). Towards freshman retention prediction: a comparative study. *International Journal of Information and Education Technology*, *3*(5), 494.

Sherrill, B., Eberle, W., & Talbert, D. (2011). Analysis of Student Data for Retention Using Data Mining Techniques.

Ngemu, J. M., Elisha, O. O., William, O. O., & Bernard, M. M. Student Retention Prediction in Higher Learning Institutions: The Machakos University College Case. *International Journal of Computer and Information Technology (ISSN: 2279–0764) Volume*. [**0.876**]

Dagley, M., Georgiopoulos, M., Reece, A., & Young, C. (2016). Increasing retention and graduation rates through a STEM learning community. *Journal of College Student Retention: Research, Theory & Practice*, *18*(2), 167-182. [0.192]

Alkhasawneh, R., & Hobson, R. (2011, April). Modeling student retention in science and engineering disciplines using neural networks. In *Global Engineering Education Conference (EDUCON), 2011 IEEE* (pp. 660-663). IEEE..

Jia, J. W., & Mareboyana, M. (2014). Predictive models for undergraduate student retention using machine learning algorithms. In *Transactions on Engineering Technologies* (pp. 315-329). Springer, Dordrecht.

Grier-Reed, T., Arcinue, F., & Inman, E. (2016). The African American student network: An intervention for retention. *Journal of College Student Retention: Research, Theory & Practice, 18*(2), 183-193. [0.192]]

Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Mining Education data to predict student's retention: a comparative study. *arXiv preprint arXiv:1203.2987*.

Fike, D. S., & Fike, R. (2008). Predictors of first-year student retention in the community college. *Community college review*, *36*(2), 68-88.

Mamiseishvili, K., & Deggs, D. M. (2013). Factors affecting persistence and transfer of low-income students at public two-year institutions. *Journal of College Student Retention: Research, Theory & Practice, 15*(3), 409-432.

Pal, S. (2012). Mining educational data using classification to decrease dropout rate of students. *arXiv preprint arXiv:1206.3078*.

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