PREDICTING STUDENT PERFORMANCE USING COGNITIVE AND NON-COGNITIVE INFORMATION

By

Sara Sultana

NUST201463931MSEECS60014F

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Information Technology

School of Electrical Engineering and Computer Science (SEECS)

National University of Sciences and Technology (NUST),

Islamabad, Pakistan.

January 2017

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Mr/Ms __ Sara Sultana __, (Registration No __ NUST201463931MSEECS60014F __), of School of Electrical Engineering and Computer Science (SEECS) has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members and foreign/local evaluators of the scholar have also been incorporated in the said thesis.

Signature: _____

Name of Supervisor: ____Dr. Sharifullah Khan____

Date: _____

Signature (HOD): _____

Date: _____

Signature (Dean/Principal): _____

Date: _____

NUST School of Electrical Engineering and Computer Sciences

A center of excellence for quality education and research

CERTIFICATE

Certified that the contents of thesis document titled *Predicting Student Performance Using Cognitive And Non-cognitive Information* submitted by Mr./Miss Sara Sultana have been found satisfactory for the requirement of degree.

Advisor: Dr. Sharifullah Khan

Committee Member 1: Dr. Asad Anwar Butt

Committee Member 2: Ms. Farzana Ahmad

Committee Member 3: Mr. Fahad Satti

CERTIFICATE OF ORIGINALITY

I here by declare that the research paper titled ***Predicting Student Performance Using Cognitive And Non-cognitive Information*** is my own work and to the best of my knowledge. It contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at SEECS or any other education institute, except where due acknowledgment, is made in the thesis. Any contribution made to the research by others, with whom I have worked at SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the projects design and conception or in style, presentation and linguistic is acknowledged. I also verified the originality of contents through plagiarism software.

Author Name: Sara Sultana

Signature: _____

DEDICATION

To All Those

Who focus on the light at the end of the tunnel, and Not on the length of the tunnel

To those

Who create stories that all of us aspire to have

But only a few have the courage, determination and will to create

Dedicated to three such souls

Saeed ur Rehman Roomi, Jawad Khan and Sidra Sultana

ACKNOWLEDGEMENT

It gives me immense pleasure to thank my supervisor Dr. Sharifullah Khan for his continuous support and excellent guidance which enabled me to conduct this research. It is he who taught me that a research is meaningless if not meant to contribute towards society by solving real world problems. I attribute the quality of my research and completion of my thesis to his efforts and encouragement, and of course to his apparently tough, but in reality, disciplined nature which helped me to avoid procrastination while doing research.

I offer my gratitude to Dr. Muhammad Azeem Abbss for his valuable input in this research regarding methodology of research and analysis. I would like to thank Dr. Asad Anwar Butt for being a helpful committee member and for identifying improvements wherever possible. I sincerely thank Sir Fahad Satti for his prompt responses whenever I needed guidance. It was his critical review of my research that strengthened my belief in the productivity of critique and value of justified appreciation.

I would like to thank Madam Farzana Ahmad for her motivation and for helping me in this research. Her involvement helped me a lot to explore the educational and behavioral aspect of this investigation. I would also like to thank Sidra Sultana who, besides my supervisor and committee members, helped me in data collection for this research.

Finally, I am indebted to my family, specially my husband Jawad Khan, whose support, prayers and belief in me helped me in completing this research.

ABSTRACT

Higher education is a privilege in developing countries like Pakistan where citizens are fighting even for getting basic education. In the past two decades, more and more students have started to enroll in IT and engineering related programs in Pakistan but a significant number of these students dropout before completing their degrees which results into loss of time, money and seats which could be offered to other deserving students. This problem demands university administrations and educators to device mechanisms through which student drop out rate can be controlled, if not totally eliminated. Besides financial support, one such mechanisms which can help in controlling drop outs is accurate prediction of student performance so that the students on the verge of failing could be identified and alarmed. This will help them in realising the efforts needed to show good academic performance. At present, the prediction methods use academic or cognitive records of students to predict their future performance. Although the non-cognitive and behavioral aspects are critical in improving student performance, their role in prediction is yet to be explored. In this research, an effort is made to improve student performance prediction by predicting performance through combined use of cognitive and non-cognitive features. The result analysis of two different data sets has shown that by adding non-cognitive variables in prediction, prediction accuracies increase using decision tree algorithm; however the addition does not play significant role in other techniques. The research also highlighted those individual cognitive features which might help students and educators to cater for drop outs.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

At the time Pakistan got independence, there was one public sector university and no private sector universities or degree awarding institutions in the country [1]. The number has increased to 179 universities and degree awarding institutions in 2016 producing 445,000 university graduates annually [2]. Developing countries like Pakistan are subsidising their higher education and are making financial and administrative arrangements in order to increase higher education literacy rate [3] since higher education has become one of key indicators of growth and development in the knowledge based economies. In Pakistan, the rate of higher education is on rise particularly after the inception of the Higher Education Commission (HEC) yet there are many hurdles facing the growth and penetration of higher education in the country [4]. Lack of student awareness, financial limitations, poor performance, job opportunity, parental pressure to earn, early child marriages and student dropout rate are few reasons among them [5]. Student dropout rate is a critical problem both for the universities and the students as both the parties invest significant time, energy and resources which are wasted when a student drops out of college or a university without accomplishing the degree enrolled. Besides that, the student seat gets wasted which could have been offered to another deserving pupil. The economic costs of student dropout include reduced literacy rate in the country, creation of non-innovative environment and loss of USD 200 billion in lost earnings and unrealized tax revenue [5].

It is observed that the focus of existing higher education plans in Pakistan is on increasing the number of student enrollment in colleges and universities and little is being done to control the dropout rate. As an example, the 11th five year development plan of Pakistan, which caters to the upcoming needs during year 2013-2018, calls the attention of the policy makers towards offering more scholarships to the university students, developing more campuses, hiring more faculty and establishing increased number of universities [3]. However, the student dropout in later semesters, particularly in the field of engineering and IT, is a big challenge which has to be faced and resolved so that the students not only enroll in universities, but also complete

degrees and become effective contributors in the knowledge based economy of Pakistan.

Fortunately, the issue of student dropout is not uncontrollable. It can be controlled by using accurate prediction methods to predict student grades. Anticipated poor grades can be used to warn the student regarding a potential dropout and then encouraging the student to put increased efforts in academic activities and spend more time in studying effectively. Effective performance prediction can also help the teacher in reviewing her own teaching methodology and in finding how can she engage the students better.

Almost all the higher education institutions have information systems containing student data where records of the student's academic performance are kept. Using a relatively newer interdisciplinary field of educational data mining (EDM) the universities can find how educational data can help in producing information which facilitates in making effective decisions for improving educational procedure [6]. EDM can help in solving dropout problem by predicting student performance. In this regard, EDM methods like regression analysis, decision tree algorithm, naive bayes theorem and neural networks can be very helpful. This research compares these four techniques and finds that decision tree algorithm offers more accurate student performance prediction as compared to the other techniques. Previously prediction was carried out only on the basis of cognitive information because it is easily available. However, as universities in developed countries are using behavioral and non-cognitive features for understanding and enhancing student performance, these features like self-image, leadership, time management are used here to predict student performance.

1.1 Motivation

In the poor as well as developing countries, higher education is a dream of most of the young population and only a few can access it because even primary and secondary literacy rates are low in these countries. Being a developing country struggling with its economic constraints, Pakistan also has relatively low literacy rate where 58 percent population holds primary education only. This is because although 79 percent male students aged between 5 to 9 years, and 65 percent female students enroll in primary schools, many of which drop-out before education completion [7]. The higher education literacy rate is far less than that although an increasing

number of students are getting admissions in the universities every year, as mentioned, but unfortunately not all of them are able to graduate. On one hand, these students waste their own and the universitys time, money and energy, and on the other hand, some deserving students are deprived of the seats that are wasted because of student dropout. Students drop out of institutions because of many reasons but, the most common among these are poor academic performance and financial limitations. Therefore, the motivation behind this research is to find ways in which the scarce educational resources of Pakistan can be effectively utilized by accurately predicting poor performers and avoiding drop outs.

1.2 Problem Statement

It is a common observation that sometimes the students with bright past record start facing difficulties in getting good grades or even passing marks because they are facing issues related to family, financial support, an increased responsibility to support family, distraction due to unnecessary socialization or hobbies. On the other hand, the once-poor performers can become bright students if they start allocating more time to studies and exam preparations rather than on activities that are not productive. In short, often the important non-cognitive and behavioral factors are ignored in student performance prediction and only past grades are relied upon which explain students abilities but cannot give complete picture of what student is capable of. Thus, while the cognitive factors like attendance, previous grades and marks in assignments and projects are important in student performance prediction, the role of is non-cognitive factors like time management, self-concept, and self-esteem are also important in improving student performance.

1.3 Research Questions

The study will focus on finding answers to the following research questions:

Q1. How does the addition of non-cognitive factors in student performance prediction impact accuracy of prediction?

Q2. Which machine learning prediction tools offer more accurate results comparatively?

Q3. Which non-cognitive variables are strong predictors of performance?

1.4 Hypothesis

Following hypothesis will be tested in this research:

H1. The inclusion of non-cognitive variables in the prediction process helps in making prediction results more sound and accurate.

1.5 Significance of Study

At present, the performance prediction methods use only the cognitive and demographic variables for prediction purposes but, the important non-cognitive factors are ignored. This study tends to find the role of non-cognitive variables in student performance prediction, and how these factors will allow the faculty members to identify which students are weak in particular non-cognitive areas. In future, such results can help in improving the role of university counselling centers so that they can help students in, for example, time management or team work, so that the overall academic performance of the students increase.

1.6 Organization of Thesis

This thesis is organized into six chapters. Chapter 1 is about Introduction of topic, its significance, the hypothesis, research questions and the motivation behind the study. Next is chapter 2 on Background which briefly introduces the concept of Educational Data Mining, steps in data mining and the main methods of prediction. Chapter 3 covers the Literature Review where several previous researches have been investigated and critically analyzed to find what is performance prediction, how it is carried out, what are common and uncommon variables of performance prediction and what are the findings of these performance prediction studies so far. This chapter also identifies gap in existing work and suggests what should be done next. Chapter 4 is about the Methodology to cover the existing gap in student performance prediction by including non-cognitive variables in research. This chapter also explains data collection method, data preprocessing and data transformation steps. Chapter 5 briefs about the Results of the study. It also discusses and critically analyzes the results. Chapter 6 presents Conclusion and Future Work related to the study.

2. BACKGROUND

This research uses different techniques of Educational Data Mining to predict student performance. The process and steps involved in educational data mining are discussed here.

2.1 Educational Data Mining

Universities, like any other organization, are generating huge amount of data on daily basis. Educational Data Mining is an emerging discipline which guides educators, administrators and policy makers to use this educational data to enhance their understanding of students, educational process and to make educational institutions more productive. In Educational Data Mining or EDM, a multidisciplinary approach is used where tools and applications of data mining, machine learning and statistics are used to device intelligent systems and processes for improving educational process and tutoring [6].

Educational Data Mining or knowledge discovery in the field of education varies a little from the other methods of data mining because it often deals with hierarchal data where the students, staff, faculty and administration all contribute data in the system and then it is used for information generation which benefits all the stakeholder in the educational arena. The hierarchal data in educational data mining can be explained through an example i.e. the interest of a student in a particular computer software can be explained using hours invested on the software, correct answers given in a test related to software, number of sessions attended, number of classes attended, or number of diplomas or courses attended in developing proficiency in certain software. Thus, the data scientists as well as teachers are part of educational data mining process to help in identifying right level of data hierarchy that suits certain research.

2.2 Steps in Educational Data Mining

The Educational Data Mining field translates data into meaningful information, but there are some steps that should be followed to get these meaningful results. These steps include follow-

ing:

2.2.1 Step 1:Data Generation or Collection

In the first step, if the data is available in local or online database then it is accessed from there, else the data is generated through interviews, questionnaires or surveys.

2.2.2 Step 2:Data Pre-Processing and Transformation

Raw data is of no or little use for the analysis purpose. Therefore data is pre-processed and transformed into usable forms by adopting a format that is more understandable. In the same step, the incomplete, inconsistent and erroneous data is handled so that data becomes complete, consistent and error free. For example, the missing data is handled through various methods like ignoring instances with missing fields, deleting instances with missing fields, replacing missing values with means or weighted means etc.

2.2.3 Step 3:Data Mining

Data mining or DM is basically an interdisciplinary field in the IT and computer science. DM is computational process where patterns are discovered in data sets of large sizes. As a step in EDM, different patterns in the data are identified and relationships among different variables are established using DM techniques. Outliers are also identified in data mining to find which behaviors adhere to common patterns and which stand out.

2.2.4 Step 4:Evaluation and Interpretation

In this step, meanings are extracted out of patterns so that knowledgeable information is outlined. This knowledge is then used by decision makers to make important decisions in the field of education.

2.3 Prediction Methodologies

Student dropout rate, discussed earlier, can be controlled using different methods of educational data mining particularly the prediction tools. In EDM, the prediction methodologies include classification and regression. Classification helps in predicting categorical labels assigned to each class, thus the students results can be categorized into pass or fail classes, teachers can be

categorized as full-time or part-time faculty members etc. Prediction is also possible using regression methods, but unlike classification where discrete class values are predicted, regression analysis predicts continuous values. Under classification, there are many tools like the Decision Trees, Neural Networks, Bayesian Networks, and Support Vector Machines. Regression based prediction can be carried out using Simple Linear Regression or Linear Regression.

3. LITERATURE REVIEW

Significant amount of research has been conducted on prediction of student performance. This chapter discusses common techniques and variables used in student performance prediction. Performance prediction is an application of educational data mining which starts from data collection, involves pre-processing and transforming of data into right format, application of data mining methods and results evaluation [8]. Student Performance Prediction is normally concluded in the form of IF-THEN format e.g. IF the student submits assignment =on-time THEN result = Pass. Most commonly the decision tree prediction method is used. For instance, in [8], both regression and decision tree methods are used and compared for prediction, but decision trees method was found to be more accurate. The student performance prediction can be carried out to predict grades like in [9] or only to predict the PASS/FAIL status.

3.1 Common Factors and EDM Techniques of Performance Prediction

Research shows that past academic records are widely used to predict student performance. [9] used student's past academic data like marks, grades in assignment and lab works to predict their grades as Poor, Average or Good. In [10], the academic as well as the demographic data of students was used to predict performance. This research used student's profile, pre test data and post test data and compared which prediction tool, decision tree or regression methods predicts results better, and decision trees offered more accurate predicted results. [11] used decision trees to predict grades in major courses based on performance in programming courses, rather than predicting overall grade during semester. This research found that J48 algorithm is helpful in making better performance predictions.

Besides using past grades and academic records for prediction purposes, [12] used student's financial status, gender and motivation to study as the variables for predicting performance. Another study used variables like gender, birth year, birth place, living place and country, type of previous education and the institute to predict grades [13]. This research comprised of data of 347 undergraduate students and offered 66 percent prediction accuracy. The grades

attained during first two years of university were used in [14] to predict student performance during final year and the results were predicted with significant accuracy. [15] claimed that the result prediction of 98 students improved by using previous mathematics scores, but the same research found that gender and grades at A-Level and O-Level are poor predictors of student performance and explain only 5 percent of performance during university life.

Prediction method of Multiple Instance Learning or MIL was used in [16] to predict student's final grades. The research ended in successfully predicting majority of the grades on the basis of marks in quizzes, assignments and activities. [17] on the other hand conducted same task of performance prediction using J48 decision tree algorithm to classify 1500 students. This research accurately predicted results of 1268 students thus achieving prediction accuracy of 84.53 percent. The research claimed that higher prediction accuracy can be achieved when predicting results in narrow range e.g. predicting PASS/FAIL offers more accuracy as compared to predicting three classes. However, it is interesting to note that [15] found that A-level grades were poor predictors of student performance while [17] found this variable helpful in predicting results.

Performance prediction helps students in the longer run as well as in the short run by identifying their weak performance areas. In [18] decision tree and regression techniques were used to predict and monitor student performance, and it was found that decision trees were more helpful in prediction and monitoring which helped student in improving performance during the second attempt. On the contrary, [9], [8] and [19] found that the linear regression and logistic regression techniques offer more reliable predictions. [20] compared prediction results of Decision Tree and Bayesian Network on Vietnamese and Thailand based students and achieved 73 percent prediction accuracy for former and 71 percent for later in four-class (fail, fair, good, very good) analysis and achieved 94 percent accuracy for former group and 93 percent accuracy for the later group in two-class (fail, pass) analysis. The results offered better accuracies using Decision Tree as compared to the Bayesian Method. [21] used different methods of performance prediction like decision trees and used variables like gender, age, family backgrounds, past GPA and lab work not only to predict grades but also to predict which students should be offered admission in the university and which should be denied admissions.

Using Regression analysis, [22] found that the gender was not helpful in predicting first year grades and that previous results in Liberal Arts were poor predictors of students' success in B.Sc programs while performance in Mathematics was a strong predictor of success particularly in

engineering. In another research, similar findings were made where analysis on 1000 students revealed that Regression analysis offers better result prediction using grades in Natural Sciences and Mathematics as compared to grades in Liberal Arts and Linguistics [23].

3.2 Non-Cognitive Variables and Student Performance

In the developed countries like the USA, the examination and testing services evaluate students on the basis of their academic and behavioral performance. Committee responsible for designing GRE tests found that the behavioral aspects of students help in determining how they will perform during their university lives [24]. The non-cognitive variables that Kyllonen (2005) says are important and are being covered in GRE and SAT tests include affective competencies (like creativity, emotional intelligence and confidence), performance factors (like leadership, team work and discipline), attitudinal constructs (like self concept, values and ethics), learning skills (like time management and stress coping) and basic personality factors (like extroversion and agreeableness). [25] emphasized that student self perception helps students in performing better and also helps teachers in imparting education in a better manner. In [26], researchers found that student behavior and attitude towards assignments affect their performance and determines how much time they will allocate to a subject at home. In his thesis, Flynt (2008) stated that student hostility, extroversion, self-image and self-esteem and leadership abilities help in determining therm performance [27]. In another research, it was proved that student's procrastination was a performance deteriorate and effective time management helps in attaining higher grades [28]. The same research also concluded that self-esteem, self-discipline and self-management are key to student's success in academia.

3.3 Critical Analysis

Despite availability of reasonable data regarding importance of student behavior in determining his performance, these behavioral factors are not included in the performance prediction most of the times. Although the previous performance in studies and past grades help in determining the future grades to much extent, the complete picture is not drawn without taking into account the behavior and attitude of the student. It is therefore recommended that these behavioral and non-cognitive attributes should be included in the performance prediction process.

4. PROPOSED METHODOLOGY

The review of the literature suggests that student performance prediction is mainly carried out on the basis of demographic and cognitive factors, and critical non-cognitive performance factors are ignored while prediction. In order to fill this gap, this research conducts student performance prediction using common machine learning tools and predicts results on the basis of demographic, cognitive and non-cognitive variables to test a hypothesis that whether the inclusion of non-cognitive variables makes results more sound or not. This section of proposed methodology briefs about the research design, sampling technique, selected sample, type of data collected, tool for data collection and data analysis methods. Figure 4.1 exhibits research design adopted to perform performance prediction and to compare the results of prediction with non-cognitive variables and without non-cognitive variables.

The overall research is divided into three stages. The first stage of research deals with sampling and data collection. In the second stage, data analysis is performed and prediction evaluation takes place in the third stage.

4.1 Sampling and Data Collection

In sampling and data collection, the researcher first defines which sampling technique to adopt. After that, a sample of population is identified and then the data collection tool is composed and shared with the sample population. Once the target population provides responses, the grades of students are collected, after semester completion, from the exam branch of the university. A second data set is downloaded from the internet in order to test the hypothesis and verify the soundness of the prediction model.

4.2 Sampling Technique

This research utilizes convenience sampling technique whereby the researcher collected data from students that could later be conveniently contacted to know about their final grades, or

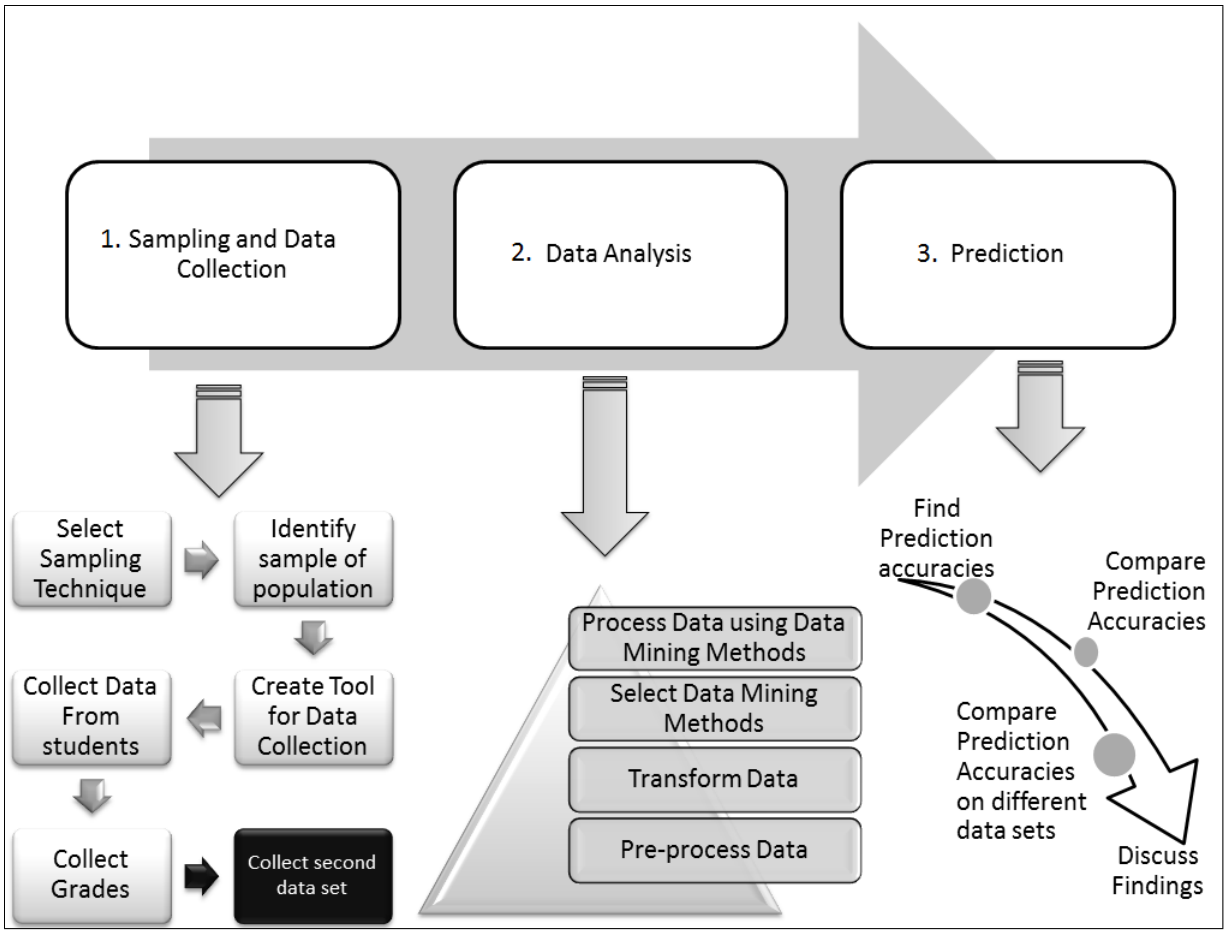


Fig. 4.1: Proposed Research Design for Student Performance Prediction

at least their grades could be traced with the help of exam branch of respective university. SEECS, NUST and Abasyn University Islamabad were chosen on the basis of their location, administrative suitability and ease of data collection.

4.3 Data Analysis

The data collected from primary source (survey questionnaire attached in Appendix) and secondary source (online [29]) are then pre-processed and transformed into usable formats. Pre-processing handles the issue of missing values and values of non-cognitive features which is also explained in detail in section 4.5.1. After that, data reduction takes place where-by the dimensions of data are reduced and data is transformed e.g. by converting five scale responses (e.g. ranges of GPA i) Above 3.5; ii) 3.00 3.49; iii) 2.5 2.99; iv) 2.00 2.49; and v) below 2) into three result categories of Pass, Fail and Probation/Warning. Section 4.5.1 summarizes the dimension reduction of data set 1, and section 4.5.2 summarizes dimension reduction of data

set 2.

Once the data is pre-processed, transformed and reduced, data mining methods are selected. These methods include Decision Trees J48 algorithm, Logistic Regression, Nave Bayes and the Neural Networks. The main reason behind the selection of these methods is frequent usage of these techniques in other related studies like [12] and [14]. The pre-processed data in excel sheets is converted to comma-separated values or .CSV format and then processed using each of the four mentioned techniques in Weka which is a freely available data mining and machine learning software [30]. Weka is a Java based collection of algorithms that assists in data pre-processing, classification, regression and visualization.

4.4 Prediction

Processing data in Weka gives output showing how the classification of data took place, what are the given values of dependent variable (i.e. Results), predicted values of Results, prediction accuracies and confusion matrices. From these outputs, the prediction accuracies are selected for four methods. For each method, there are two prediction accuracies separately calculated; first for the combination of demographic and cognitive variables and second time for the prediction based on demographic, cognitive and non-cognitive variables. The prediction accuracies of four methods are then compared to find whether prediction accuracy increased, decreased or remained unchanged after the inclusion of non-cognitive variables.

4.5 Case Studies

The above mentioned proposed methodology is applied to two data sets in this research. Following two case studies discuss application of proposed methodology on each data set.

4.5.1 Case Study 1

The first case study covers student performance prediction using data set collected through survey questionnaire. The data collection, data analysis and prediction on data set 1 is discussed below.

Primary Data Collection

In this research convenience sampling technique was adopted. It is a type of non-probability sampling technique where each individual in the population does not have equal chance of being selected in the sample. Rather, the participants in the sample are those individuals that are easily, or conveniently, approachable [31]. Convenience sampling technique expedites the data collection process, the data collection is made easier and cost effective through this technique therefore, it was used.

The convenient sample comprised of students from two easily approachable universities in Islamabad. First was the National University of Sciences and Technology (NUST) and the second one was Abasyn University Islamabad. Students of under graduation and post-graduation programs in the discipline of Electrical Engineering participated in the research. In order to collect data from these students, a survey tool was designed. The designed questionnaire was based on the Study Habit Questionnaire available at IEEE.org [32]. The Study Habit Questionnaire was adopted form of Virginia Gordons University Survey: A Guidebook and Readings for New Students, and the questionnaire was further adopted under this research in order to include the critical non-cognitive factors identified in the related work section. The questions related to community and social support were adopted from Social Support Questionnaire [33], leadership questions were extracted from LeadershipSkillsQuestionnaire [34], Student Self-Assessment Questionnaire [35] provided questions on the aspects of realistic self-appraisal and self-concept. The compiled survey tool comprised of 42 questions covering demographic, cognitive and non-cognitive aspects of the individuals profile. Some of these questions like students name, identification number and university name were not used in prediction and helped only in retrieving students final results from the exam branch of respective universities. The details about number of questions on each variable are discussed in Table 4.1.

The answers of these questions helped in defining whether a student was doing great, OK or needed help in the particular non-cognitive area. After the compilation of questionnaire, it was shared among the students using Google documents which helped in online data collection. Total 128 students responded to the survey questionnaire out of which 113 instances were complete and usable. After the students responded to the questions, the students grades were collected from the exam departments of respective universities based on the student IDs.

Tab. 4.1: Variables and the number of questions related to each variable

Type of Variable	Variable	Number of Questions
Demographic Variables	Gender	1
	Status	1
	Mother Education	1
Cognitive Variables	Previous Results	1
	Sessional	1
	Quizzes	1
	Assignments	1
	Projects	1
Non-Cognitive Variables	Time Management	8
	Self Concept	2
	Realistic Self Appraisal	5
	Leadership	9
	Community Support	7

Data Analysis

The data collected from survey questionnaire was pre-processed and transformed into usable format. During pre-processing, issue of missing values was handled. Since all other questions were mandatory to answer except name and roll number, the instances with both missing name and roll number were excluded from data set, since their final grades could not be accessed from exam departments. The next step was data dimension reduction where dimensions were reduced as shown in Table 4.2.

4.5.2 Case Study 2

In the second case study, an online data set was used for prediction purpose. The data collection, data analysis and prediction process of this data set are discussed below.

Data Collection

The data set 2 was a larger data set with 650 instances available in an online repository. It was a student profile data set used by Cortez and Silva (2008) [36]. The online data set covered 32

Tab. 4.2: Data Dimension Reduction in Case Study 1

	Variable	Original Dimensions	Reduced Dimensions
Demographic Variables	Mother's Education	Neo-Literate Primary Level Secondary Level Graduation Level Level Post-Graduation Level No Formal Education None	Basic Advance None
Cognitive Variables	Previous Results	Above 3.5 3.00 3.49 2.5 2.99 2.00 2.49 Below 2	Pass Probation Warning
Non-Cognitive Variable	Time Management	9 Questions (Count the number of Agrees)	6-8 Agree Great 3-5 Agree OK 0-2 Agree Need Help
Non-Cognitive Variable	Self-Concept	2 Questions (Count the number of Agrees)	2 Agree Great 1 Agree OK 0 Agree Need Help
Non-Cognitive Variable	Realistic Self Appraisal	5 Questions (Count the number of Agrees)	4-5 Agree Great 2-3 Agree OK 0-1 Agree Need Help
Non-Cognitive Variable	Leadership	9 Questions (Count the number of Agrees)	6-9 Agree Great 3-5 Agree OK 0-2 Agree Need Help
Non-Cognitive Variable	Community Support	7 Questions (Count the number of Agrees)	5-7 Agree Great 3-4 Agree OK 0-2 Agree Need Help

variables, of which 15 independent and 1 dependent variable were selected for this study, based on general acceptance in research and relevance of these variables with the culture of Pakistan.

Tab. 4.3: Variables in Data Set 2

Type of Variable	Variable
Demographic Variables	Gender
	Parent's Cohabitation Status
	Guardian
Cognitive Variables	Absenteeism
	First Sessional
	First Sessional
Non-Cognitive Variables	Study Performance
	Study Time
	Free Time
	Independence Level
	Proximity to College
	Health
	Go Out
	School Support
	Plan for Future Studies

The variables covered in the two data sets are presented in Table 4.3.

Data Analysis

The second data set needed little pre-processing. First of all, there were no missing values therefore, all the instances were usable. Few variables in data set 2 needed data dimension reduction which was performed as mentioned in Table 4.4.

Tab. 4.4: Data Dimension Reduction in Case Study 2

	Variable	Original Dimensions	Reduced Dimensions
Cognitive Variables	Absences	Any number between 0-100	0-1 Few
			11-20 Need Help
			21-50 Warning
			Above 51 Alarming
	First Sessional	Any number between 0-20	0-5 Poor
			6-10 Fair
			11-15 Good
			16-20 Excellent
	Second Sessional	Any number between 0-20	0-5 Poor
			6-10 Fair
			11-15 Good
			16-20 Excellent
Dependent Variable	Results	Any number between 0-20	0-5 Poor
			6-10 Fair
			11-15 Good
			16-20 Excellent

5. RESULTS AND ANALYSIS

5.1 Correlation among Independent and Dependent Variables in Case Study 1

The correlation coefficient between two variables tells how strong or weak is a relationship between them. There are many correlation measures that can be used and this research relies upon one of the most common methods i.e. Pearson's R correlation value. If the value of Pearson's R for two variables is close to 1, this means that there is a strong relationship between the variables, and if the value is close to 0 then this means that the variables are weakly correlated. A positive Pearson's R value shows that increase in value of one variable will increase the value of other variable too and the decrease will cause decrease in value of other variable. On the other hand, negative Pearson's R value shows that the variables have inverse relationship and increase in value of one variable decreases the other. If the correlation value is significant, this means that the relationship between the two variables does not exist by accident, on the other hand, insignificant correlation value means that there is a chance that the relationship occurs by accident. In this research, the correlation is significant at the 0.01 level (2-tailed). Statistically, the Sig (2-Tailed) value which is less than or equal to .05 shows that a statistically significant correlation exists between the two variables.

Table 5.1 lists different independent variables and their correlation with the dependent variable 'Results' calculated in SPSS. As it can be seen, according to the values, only Gender and Previous Results have significantly positive relationship with the Results, while other variables have correlation that is not significant and can be attributed to chance. However, the correlation coefficient calculated on smaller dataset can sometimes be misleading, hence further analysis will show whether these independent variables affect the student results or not.

Some interesting observations were made during the correlation analysis regarding variables having negative correlation with the results. These correlations are briefly discussed in following sections:

Tab. 5.1: Correlation of Variables with Results

Variable	Pearson Correlation	Sig. (2-tailed)
Gender	.274**	.003
Status	0.26	0.787
Mother's Education	-.060	.527
Previous Results	.613**	.000
Sessional	.102	.283
Quizzes	.033	.732
Assignments	.119	.208
Projects	-.017	.860
Time Management	.083	.383
Self Concept	-.167	.077
Realistic Self Appraisal	-.088	.354
Leadership	-.125	.188
Community Support	.077	.419

**. Correlation is significant at the 0.01 level (2-tailed).

5.1.1 Mother's Education

Mother's education is a demographic variable and correlation analysis showed that this variable was negatively correlated with Results, having Pearson's R value of -.060. This weak correlation means that it is likely that the increase in mother's education will not help the student in getting good results at the university level. The negative correlation can be explained in following ways:

- Since the data set comprises of students from Computer Science, IT, and Software Engineering backgrounds, and these fields have penetrated in the Pakistan's education sector only a decade or two ago, the parents of the current young generation are less likely to have education in the same field. The mother's education, therefore, might not be relevant to the education of university students thus, it does not significantly help them in getting good grades.

- Mother's education seems to be more helpful to the children during their early education particularly during their school years when they are studying general courses. The

Tab. 5.2: Interpretation of Correlation Values

Correlation Scale	Interpretation
.00 - .19	Very Weak
.20 - .39	Weak
.40 - .59	Moderate
.60 - .79	Strong
.80 1.0	Very Strong

mother's education in general subjects (like English, Islamiat, Urdu, Education) can help students when they are in schools, but after that it does not help them.

5.1.2 Projects

The data set gave negative correlation between project grades and the Results and produced Pearson's R value of -.017. This means, getting good marks in projects might result into poor final grades. The reasons might include:

- The projects might not be well designed and therefore, they might not truly gauge the abilities of a student. Asking a student to do an easier or unrelated project might overestimate his abilities, consequently the student is not able to show similar output in exams which are tougher than the projects.
- It is likely that the students are not doing their own projects themselves and getting projects done by someone else. In this way, they can cheat the teacher in project (by turning in good projects) but they fail to keep up the same performance in final results where monitoring is stricter.

5.1.3 Self Concept

The students' self concept i.e. the idea that one holds about himself, is weakly negatively correlated, i.e. -.167 value, with student results in the collected data set. This is opposite to the common expectation that a person with higher self-image should be a better performer. This inverse finding can be explained in the following manner:

- It is possible that the students are mistaking over-confidence in themselves as self-concept. They think that getting a B grade is not a big deal in university and that the grading sys-

tem is unfair when they are offered poor grades. Such underestimating of toughness of university educational system can cause student to perform poorly in exams and get poor grades.

5.1.4 Realistic Self Appraisal

The ability to evaluate one's own strengths and weaknesses is called realistic self appraisal, which is negatively correlated with results in the data collected under this study. The correlation value is -.088. This shows that:

- The students need to improve their abilities to evaluate their strengths and weaknesses, as so far they are unable to identify and overcome their weakness i.e. poor performance in exams.

- A more comprehensive approach towards student appraisal might be adopted the next time, i.e. the peer and teacher side appraisal will offer a better evaluation of strengths and weaknesses of the student. It seems that the student alone is not able to clearly see his strengths and weaknesses and needs the help of people around him. The 360 degree appraisal is adopted now-a-days for employee appraisal where colleagues, bosses and subordinates all evaluate the performance of a person to exclude bias factor. Similar approach can be used for student appraisal.

5.1.5 Leadership

Leadership is normally considered to be a helpful quality, but the data analysis shows that students in computer science and engineering disciplines might get poorer results if they are good at leadership and vice versa. The weak negative correlation between leadership and results i.e. -.125 can be interpreted as:

- Leadership quality is not as relevant to good performance of computer science and engineering students as it might be for the students of social sciences, marketing and business administration, because there are no grades associated with leadership in computer science and engineering.

- Computer science demands more technical and analytical skills from the students than the social and interpersonal skills. Thus, leadership might divert them from their actual

performance goals. Therefore, being more leader-like might mean being less productive in computer science.

5.2 Prediction on Data Set 1

Prediction accuracy of a method tells how many instances a particular method can correctly classify. Higher the number of correctly classified instances, higher is the accuracy of prediction. Accuracy is the number of true positives and true negatives out of all the true positives and negatives plus false positives and negatives.

$$Accuracy = (TP + TN)/(TP + FP + FP + FN) \tag{5.1}$$

The research tends to find out whether the addition of non-cognitive variables in the list of commonly used predictors enhances the accuracy of prediction or not. Several data mining prediction techniques are used below to test the hypothesis.

5.2.1 Decision Trees Algorithm

Decision Tree is one of the most commonly used prediction tools. This research processes data using J48 algorithm of Decision Trees to predict grades and find the accuracy of prediction.

Without Non-Cognitive Variables

In this section, Results are predicted using eight independent variables including the gender, employment status of the student, mother's education, previous result, marks in sessional exams, marks in quizzes, assignments and projects. Based on these eight attributes, the J48 algorithm offered 61 percent accuracy of prediction. The decision Tree produced as a result of this prediction scheme is expressed in IF-THEN format in Figure 5.1.

With Non-Cognitive Variables

The accuracy of the result prediction based on eight independent variables is compared with the prediction based on thirteen variables, where all the five new variables belong to the category of non-cognitive aspect and these include Time Management, Self Concept, Realistic Self Appraisal, Leadership and Community Support. These thirteen variables offered 65 percent accurately predicted results i.e. showed the improvement of about 4 percent. The comparison

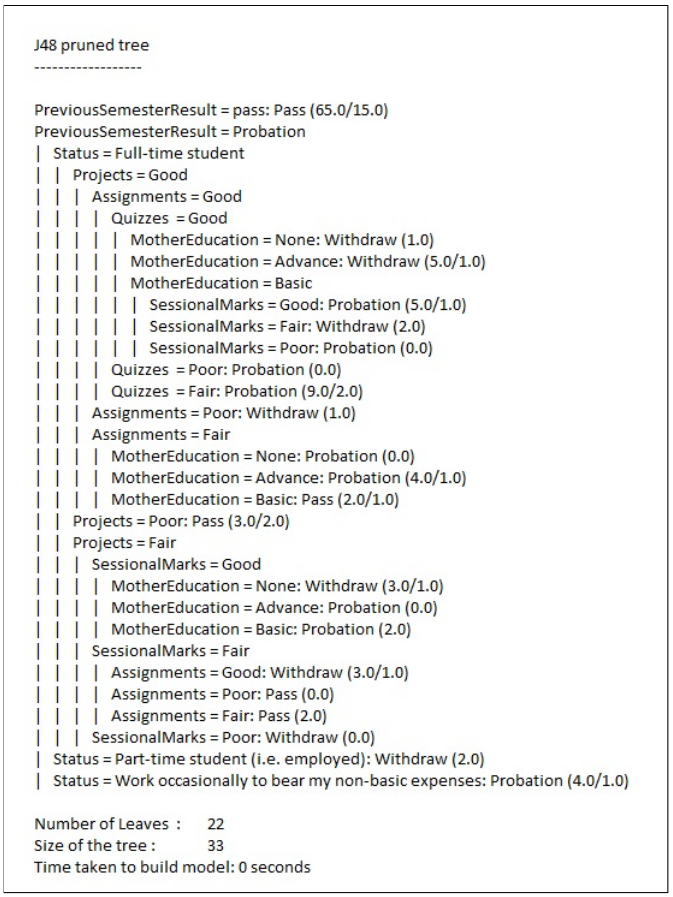


Fig. 5.1: Decision Tree Result based on Demographic and Cognitive Variables

of the two prediction accuracies i.e. without non-cognitive variables and with non-cognitive variables is shown in 5.2

5.2.2 Logistic Regression

Logistic Regression helps in analyzing a dataset where multiple categorical variables are used to predict an outcome. Data analysis using logistic regression shows that the prediction accuracy based on eight independent variables i.e. demographic and cognitive variables is 54 percent. On the other hand, when the prediction was carried out on the basis of thirteen independent variables where five variables were non-cognitive, the prediction accuracy of 51 percent was observed. 5.3 shows the comparison of prediction accuracies with and without non-cognitive variables using logistic regression function.

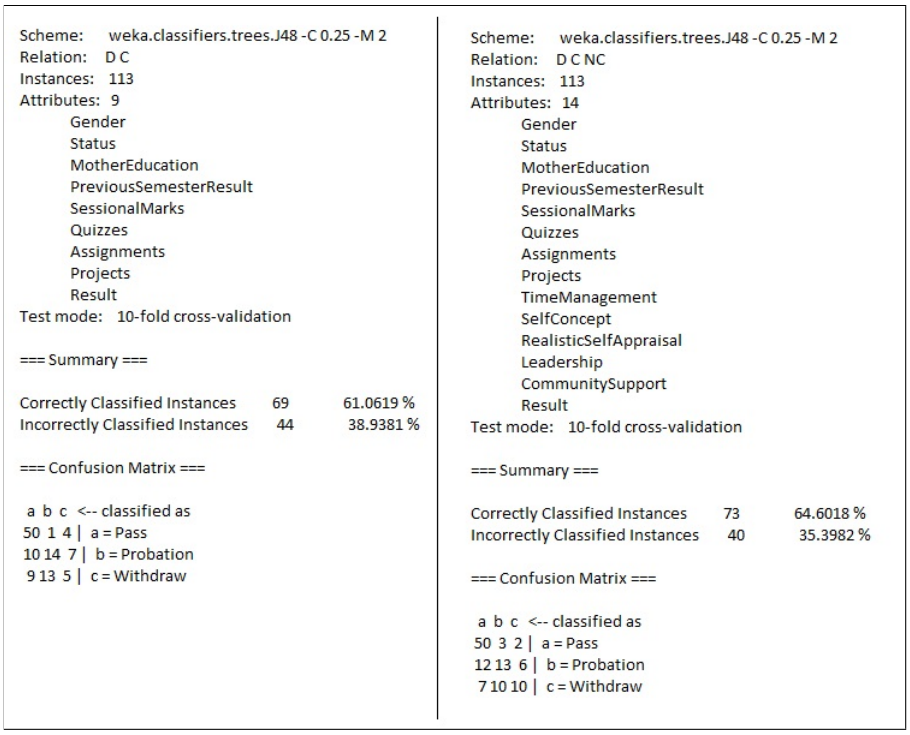


Fig. 5.2: Comparison of Prediction Accuracies with and without Non-Cognitive Variables using Decision Tree

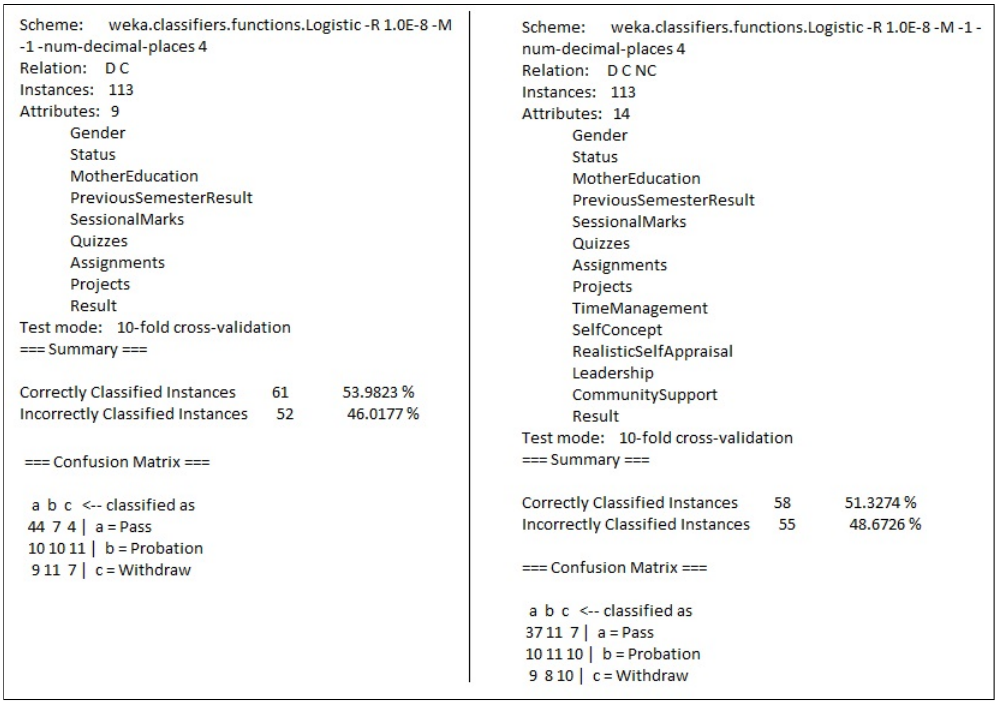


Fig. 5.3: Comparison of Prediction Accuracies with and without Non-Cognitive Variables using Logistic Regression

5.2.3 Naive Bayes Function

The Naive Bayes classifier helps in probabilistically finding relationship between variables. Weka based Naive Bayes analysis of demographic and cognitive variables to predict results

shows that 61 percent accuracy of prediction can be achieved without the use of non-cognitive variables. On the other hand, when the same tool and technique were used to predict results with the addition of five non-cognitive variables, the same accuracy of prediction was observed i.e. 61 percent. 5.4 exhibits that prediction accuracy remained same using Naive Bayes classifier whether or not the non-cognitive variables were used.

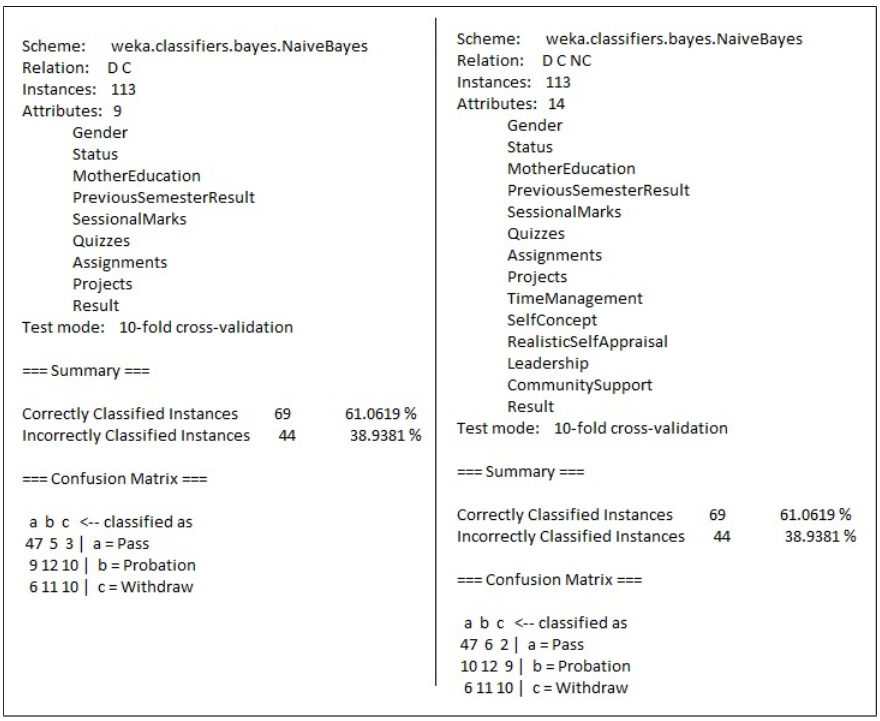


Fig. 5.4: Comparison of Prediction Accuracies with and without Non-Cognitive Variables using Naive Bayes Classifier

5.2.4 Neural Networks

Neural Networks are often used for pattern recognition and prediction purposes. In this research, the Neural Networks function was used to predict student results and to find whether a student will pass, go on probation or will be asked to withdraw. The prediction based on demographic and cognitive variables offered 54 percent accuracy while inclusion of non-cognitive variables offered prediction accuracy of 52 percent. This showed that using neural networks, prediction accuracy is better when the non-cognitive variables are not included.

In summary, the prediction accuracies on the data set 1 using four different techniques with and without the non-cognitive variables are shown in Table 5.3 representing an increasing trend in accuracies using decision trees, no change using Naive Bayes function and a decreasing trend that was experienced using logistic regression and neural networks.

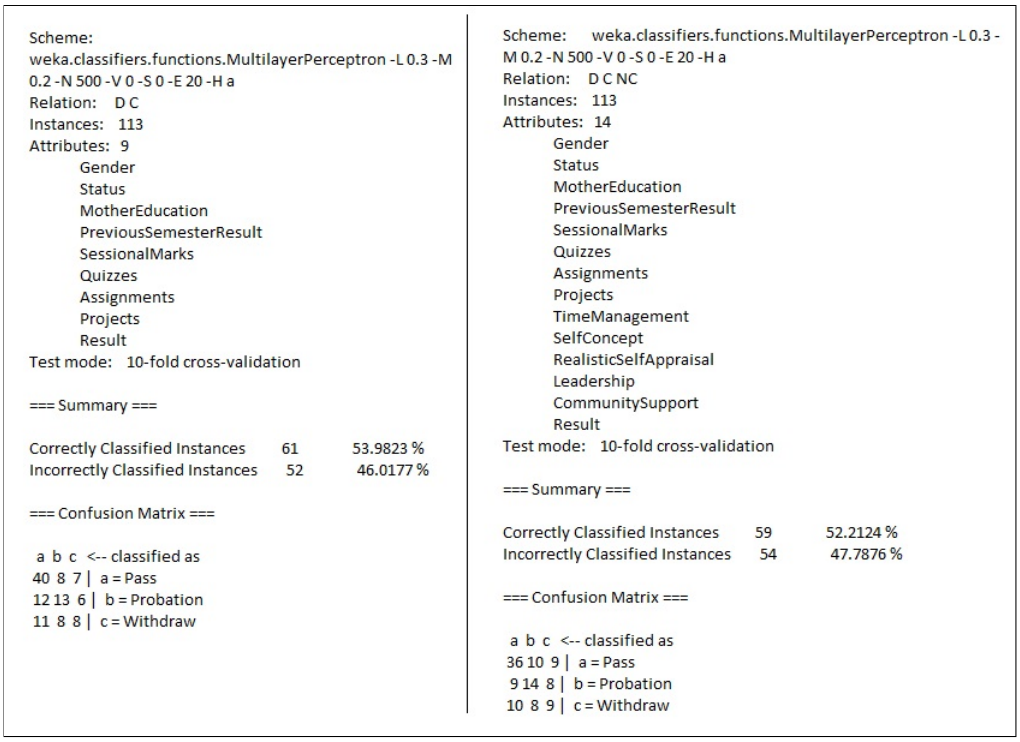


Fig. 5.5: Comparison of Prediction Accuracies with and without Non-Cognitive Variables using Neural Network

5.3 Prediction on Data Set 2

The online data set had fifteen independent variables including demographic variables (gender, parent's cohabitation status, guardian), cognitive variables (absenteeism, first sessional grades, second sessional grades) and the non-cognitive variables (study preference, study time, free time, independence level, proximity to college, health, go out with friends, school support and plan for future studies). The data was analyzed using four selected prediction techniques and the results of prediction techniques are discussed below.

5.3.1 Decision Trees Algorithm

The prediction accuracy on the data set 2 was 82 percent without the use of the non-cognitive variables and when the non-cognitive variables were included in prediction, the accuracy improved to 84 percent. The Figure 5.6 compares the prediction with and without non-cognitive variables.

Tab. 5.3: Comparison of Prediction Accuracies using Four Prediction Techniques

Factors/Tech	Decision Trees	Logistic Regression	Naive Bayes	Neural Networks
Demographic and Cognitive Variables	61	54	61	54
Demographic, Cognitive and Non-Cognitive Variables	65	51	61	52

```
Scheme:     weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:   demo cog
Instances:  649
Attributes:  7
        gender
        Pstatus
        guardian
        absences
        FirstSessional
        SecondSessional
        Result
Test mode:  10-fold cross-validation

=== Summary ===

Correctly Classified Instances         535               82.4345 %
Incorrectly Classified Instances       114               17.5655 %

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 326   0  42   2   0   0 |   a = good
   0   0   0   2   0   0 |   b = excellent
  29   0 149   0   2   0 |   c = fair
  23   0   1  54   0   0 |   d = excellent
   1   0  10   0   6   0 |   e = poor
   1   0   0   1   0   0 |   f = vpoor
```

```
Scheme:     weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:   Demo Cog Non Cog
Instances:  649
Attributes:  16
        gender
        Pstatus
        guardian
        absences
        FirstSessional
        SecondSessional
        preferstudies
        studytime
        freetime
        IndependenceLevel
        Proximitytocollege
        health
        goout
        schoolsup
        PlanforFutureStudies
        Result
Test mode:  10-fold cross-validation
=== Summary ===

Correctly Classified Instances         541               83.59 %
Incorrectly Classified Instances       108               16.41 %


=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 326   0  42   2   0   0 |   a = good
   0   0   0   2   0   0 |   b = excellent
  23   0 155   0   2   0 |   c = fair
  23   0   1  54   0   0 |   d = excellent
   1   0  10   0   6   0 |   e = poor
   1   0   0   1   0   0 |   f = vpoor
```

Fig. 5.6: Comparison of prediction accuracies using Decision Trees on data set 2

5.3.2 Logistic Regression

With the help of logistic regression, the prediction accuracy of 84 percent was achieved when the non-cognitive variables were not included in the prediction process. However, the inclusion of these factors gave the prediction accuracy of 82 percent as shown in the Figure 5.7.

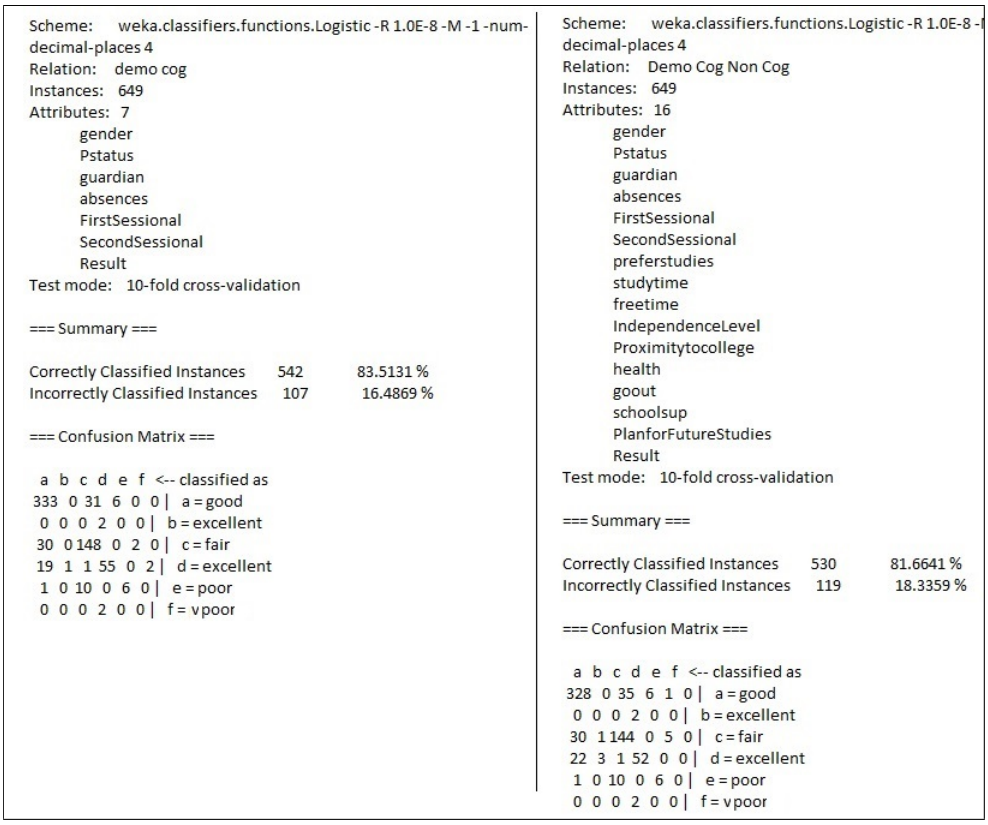


Fig. 5.7: Comparison of prediction accuracies using Logistic Regression on data set 2

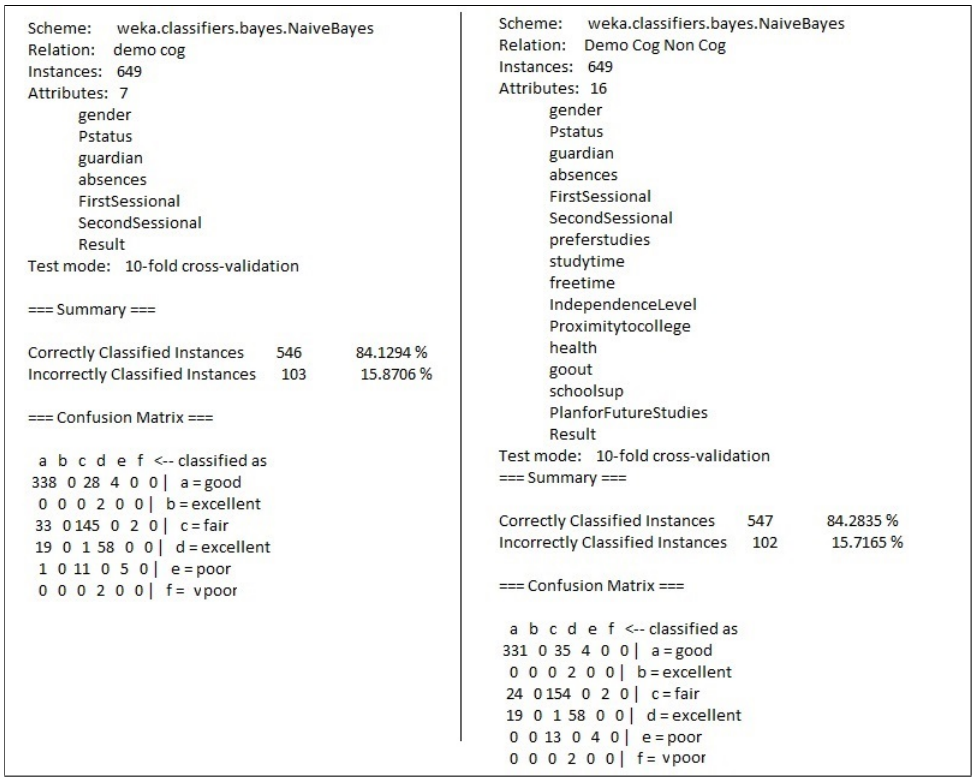
5.3.3 Naive Bayes Function

The second data set showed same result prediction accuracies before and after the inclusion of the non-cognitive variable. The accuracy was 84 percent when only the demographic and cognitive variables were used for prediction and it was again 84 percent when these variables were added in the prediction process as mentioned in Figure 5.8.

5.3.4 Neural Networks

Neural Networks based prediction on data set 2 offered prediction accuracy of 82 percent when the analysis was carried out without the non-cognitive variables. However, when the non-cognitive variables were added, the prediction accuracy reduced to 76 percent which is presented in Figure 5.9.

The observation made in the data analysis of first data set is also repeated in the analysis of results of data set 2. The prediction accuracy increases using decision trees, decreases using logistic regression and neural networks and remains unchanged when the Naive Bayes function is used.



```
Scheme:     weka.classifiers.bayes.NaiveBayes
Relation:   demo cog
Instances:  649
Attributes:  7
        gender
        Pstatus
        guardian
        absences
        FirstSessional
        SecondSessional
        Result
Test mode:  10-fold cross-validation

=== Summary ===

Correctly Classified Instances         546               84.1294 %
Incorrectly Classified Instances       103               15.8706 %

=== Confusion Matrix ===

  a  b  c  d  e  f  <-- classified as
338  0 28  4  0  0 |  a = good
  0  0  0  2  0  0 |  b = excellent
 33  0 145  0  2  0 |  c = fair
 19  0  1 58  0  0 |  d = excellent
  1  0 11  0  5  0 |  e = poor
  0  0  0  2  0  0 |  f = vpoor
```

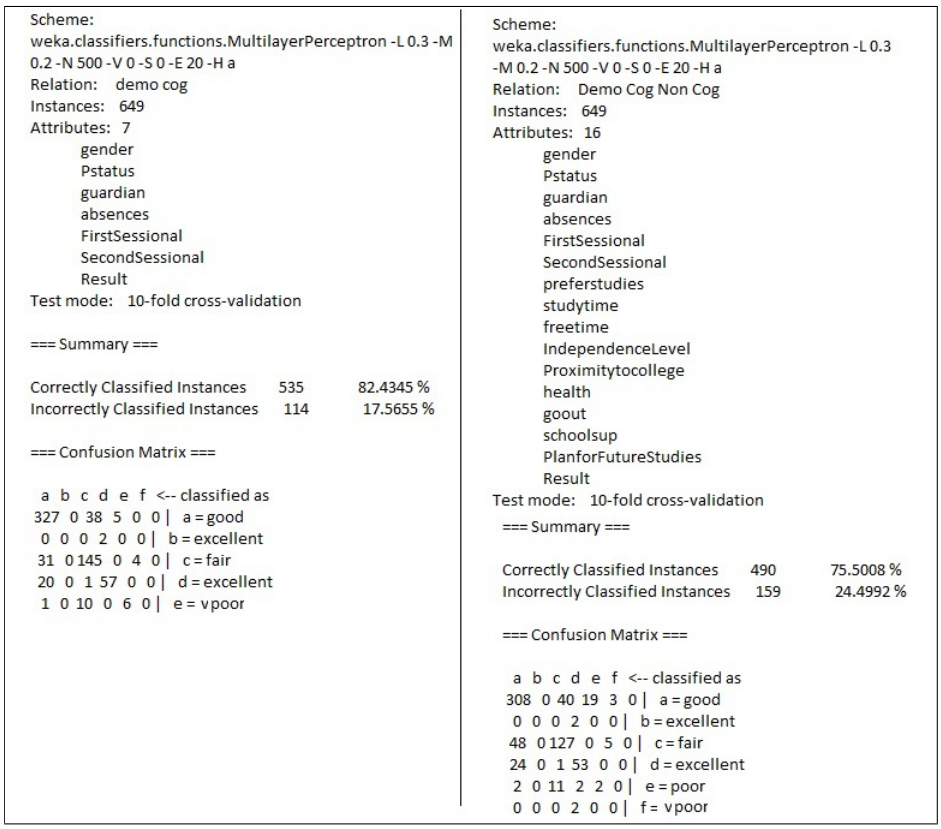
```
Scheme:     weka.classifiers.bayes.NaiveBayes
Relation:   Demo Cog Non Cog
Instances:  649
Attributes:  16
        gender
        Pstatus
        guardian
        absences
        FirstSessional
        SecondSessional
        preferstudies
        studytime
        freetime
        IndependenceLevel
        Proximitytocollege
        health
        goout
        schoolsup
        PlanforFutureStudies
        Result
Test mode:  10-fold cross-validation
=== Summary ===

Correctly Classified Instances         547               84.2835 %
Incorrectly Classified Instances       102               15.7165 %

=== Confusion Matrix ===

  a  b  c  d  e  f  <-- classified as
331  0 35  4  0  0 |  a = good
  0  0  0  2  0  0 |  b = excellent
 24  0 154  0  2  0 |  c = fair
 19  0  1 58  0  0 |  d = excellent
  0  0 13  0  4  0 |  e = poor
  0  0  0  2  0  0 |  f = vpoor
```

Fig. 5.8: Comparison of prediction accuracies using Naive Bayes on data set 2



```
Scheme:
weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M
0.2 -N 500 -V 0 -S 0 -E 20 -H a
Relation:   demo cog
Instances:  649
Attributes:  7
        gender
        Pstatus
        guardian
        absences
        FirstSessional
        SecondSessional
        Result
Test mode:  10-fold cross-validation

=== Summary ===

Correctly Classified Instances         535               82.4345 %
Incorrectly Classified Instances       114               17.5655 %

=== Confusion Matrix ===

  a  b  c  d  e  f  <-- classified as
327  0 38  5  0  0 |  a = good
  0  0  0  2  0  0 |  b = excellent
 31  0 145  0  4  0 |  c = fair
 20  0  1 57  0  0 |  d = excellent
  1  0 10  0  6  0 |  e = vpoor
```

```
Scheme:
weka.classifiers.functions.MultilayerPerceptron -L 0.3
-M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
Relation:   Demo Cog Non Cog
Instances:  649
Attributes:  16
        gender
        Pstatus
        guardian
        absences
        FirstSessional
        SecondSessional
        preferstudies
        studytime
        freetime
        IndependenceLevel
        Proximitytocollege
        health
        goout
        schoolsup
        PlanforFutureStudies
        Result
Test mode:  10-fold cross-validation
  === Summary ===

  Correctly Classified Instances         490               75.5008 %
  Incorrectly Classified Instances       159               24.4992 %

  === Confusion Matrix ===

    a  b  c  d  e  f  <-- classified as
  308  0 40 19  3  0 |  a = good
    0  0  0  2  0  0 |  b = excellent
   48  0 127  0  5  0 |  c = fair
   24  0  1 53  0  0 |  d = excellent
    2  0 11  2  2  0 |  e = poor
    0  0  0  2  0  0 |  f = vpoor
```

Fig. 5.9: Comparison of prediction accuracies using Neural Networks on data set 2

Tab. 5.4: Comparison of Prediction Accuracies on Data Set 2

Factors/Tech	Decision Trees	Logistic Regression	Naive Bayes	Neural Networks
Demographic and Cognitive Variables	82	84	84	82
Demographic, Cognitive and Non-Cognitive Variables	84	82	84	76

The two case studies had different number of instances and variables and these offered different prediction accuracies before and after addition of non-cognitive variables as shown in Table 5 and 10. However, each of the methods used in the two case studies showed same trends in prediction accuracies i.e. the prediction accuracy increased after addition of non-cognitive variables using decision trees in both data sets. In both case studies, the prediction accuracies did not change after the inclusion of non-cognitive variables when Nave Bayes Method was used, however, a downward trend was observed when the non-cognitive variables were added in prediction using Logistic Regression and Neural Networks. The trends are, therefore, same in the two case studies as shown in Table 5.10

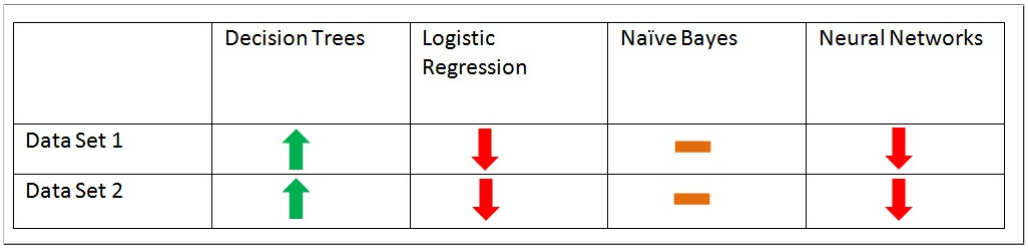


Fig. 5.10: Comparison of prediction accuracies on two data sets

It is observed that the addition of non-cognitive variables results into increase in prediction accuracy, but the impact is not as simple as is the impact of addition of cognitive variables. Not all the non-cognitive variables help in increasing prediction accuracy and some of these help in increasing the accuracy more than the others. Leadership, for example either does not help in increasing prediction accuracy or does so by a very little percentage. Community support helps

in increasing accuracy more often, both individually as well as in combination with other non-cognitive variables. Therefore, the non-cognitive variables increase the prediction accuracy but the researchers critical job is to find which non-cognitive variables to include in prediction and which to ignore. Leadership might be a more relevant performance predictor for students of social sciences, but the students of electrical engineering might not benefit a lot from it.

6. CONCLUSION AND FUTURE WORK

6.1 Conclusion

The primary objective of this research was to use educational data mining tools in such a way that students can be alarmed about a potential poor performance so that they can improve their performance and avoid drop-out from university. So far, the student performance prediction is carried out using demographic and cognitive variables. This research used demographic, cognitive and non-cognitive variables for prediction purpose after realizing the importance of behavioral factors in student performance. It was hypothesized that addition of non-cognitive variables in performance prediction process will offer more sound predictions. The data analysis and results prove that certain non-cognitive variables help in improving the prediction accuracy of results.

6.2 Future Work

This research has opened new doors for research in performance prediction. First of all, the university counselling centers might learn the importance of keeping non-cognitive data about students for identifying student strengths and weaknesses which can be used for improving student performance. In future, a more careful selection of non-cognitive variables can be made because some non-cognitive variables are more helpful in predicting student performance than the other variables. Another thing which can help researchers in future is the collection of large size data sets so that the models are trained more effectively.

Appendices

Thesis Data Collection

Alsalam o alaikum Dear Fellow Students
This is Sara Sultana conducting a research titled "Predicting student performance using cognitive and non-cognitive information". Please fill this questionnaire to the best of your knowledge and support me in collecting data for research.
JazakAllah

*Required

Demographic Data

1. **Q1: University Name**
   Mark only one oval.
   - ◯ Abasyn University Islamabad
   - ◯ NUST

2. **Q2: Name of Student (Optional)**

3. **Q3: University Registration ID** *

4. **Q4: Gender** *
   Mark only one oval.
   - ◯ Male
   - ◯ Female

5. **Q5: I study at University as** *
   Mark only one oval.
   - ◯ Full-time student
   - ◯ Part-time student (i.e. employed)
   - ◯ Work occasionally to bear my non-basic expenses

6. **Q6: My Mother's Education is** *
   Mark only one oval.
   - ◯ Primary level
   - ◯ Secondary level
   - ◯ Graduation level
   - ◯ Post-Graduation level
   - ◯ She could not get formal education
   - ◯ Neo-literate (can only read/write)
   - ◯ None
   - ◯ Other: .....

Cognitive Data

7. **Q7: My previous semester GPA is** *
   Mark only one oval.
   - 2.00 - 2.49
   - 2.5 - 2.99
   - 3.00 - 3.49
   - Above 3.5

8. **Q8: My marks in sessionals during this semester can be described as** *
   Mark only one oval.
   - Good (Above 60 percent)
   - Fair (40 to 50 percent)
   - Poor (Below 40 percent)

9. **Q9: My marks in quizzes during this semester are.... ** *
   Mark only one oval.
   - Good (Above 60 percent)
   - Fair (40 to 50 percent)
   - Poor (Below 40 percent)

10. **Q10: I can say my assignments during the semester are... ** *
    Mark only one oval.
    - Good (Above 60 percent)
    - Fair (40 to 50 percent)
    - Poor (Below 40 percent)

11. **Q11: My projects/presentations marks so far in this semester are... ** *
    Mark only one oval.
    - Good (Above 60 percent)
    - Fair (40 to 50 percent)
    - Poor (Below 40 percent)

Non-Cognitive Factors (1/5)

Student Time Management

12. **Q12: Do you make a Time Table for each semester?** *
    Mark only one oval.
    - Yes
    - No

13. **Q13: How often do you update your Time Table?** *
Mark only one oval.

- Once a week
- Once fortnight
- Once a month
- Once every two months
- Never
- Other (Please specify)

14. **Q14: Do you stick to your Time Table?** *
Mark only one oval.

- Yes
- No
- Sometimes

15. **Q15: On weekly basis, do you manage to get time for exercise and socializing with friends?** *
Mark only one oval.

- Yes
- No
- Sometimes

16. **Q16: Do you get at least 6 hours of sleep each night?** *
Mark only one oval.

- Yes
- No
- Sometimes

17. **Q17: Do you study at least two hours every week for each course at home?** *
Mark only one oval.

- Yes
- No
- Sometimes

18. **Q18: Do you get your assignments done on time?** *
Mark only one oval.

- Yes
- No
- Sometimes

19. **Q19: Do you regularly attend your classes?** *
Mark only one oval.

- Yes
- No
- Sometimes

Non-Cognitive Factors (2-3/5)

Self Concept

20. **Q20: Getting a B grade (3.0 GPA) is not very hard?** *
Mark only one oval.
- Agree
- Neutral
- Disagree

21. **Q21: My high school (College/HSSC) grades don't really reflect what I can do (I am capable of much more)** *
Mark only one oval.
- Agree
- Neutral
- Disagree

Realistic Self Appraisal

22. **Q22: I have experienced, or am currently experiencing, difficulties in my personal life which can affect my grades** *
Mark only one oval.
- Agree
- Neutral
- Disagree

23. **Q23: I have experienced, or am currently experiencing, difficulty adjusting to university.** *
Mark only one oval.
- Agree
- Neutral
- Disagree

24. **Q24: I need, or have needed, to develop organizational skills.** *
Mark only one oval.
- Agree
- Neutral
- Disagree

25. **Q25: I tended or tend to panic when I writing exams or making presentations.** *
Mark only one oval.
- Agree
- Neutral
- Disagree

26. **Q26: I had a hard time getting through the assigned readings.** *
Mark only one oval.

- Agree
- Neutral
- Disagree

Non-Cognitive Factors (4/5)

Leadership

27. **Q27: I am a member of a community, society at my university and participate actively. (Leadership)** *
Mark only one oval.

- Agree
- Neutral
- Disagree

28. **Q28: During discussions or brainstorming sessions in class, university society, people most often seek my opinion/advice. (Leadership)** *
Mark only one oval.

- Agree
- Neutral
- Disagree

29. **Q29: Seeing the big picture comes easily for me. (Leadership)** *
Mark only one oval.

- Agree
- Neutral
- Disagree

30. **Q30: I like making strategic plans for my life/family/society etc. (Leadership)** *
Mark only one oval.

- Agree
- Neutral
- Disagree

31. **Q31: I am effective at problem solving (Leadership)** *
Mark only one oval.

- Agree
- Neutral
- Disagree

32. **Q32: I can clearly see a pathway for the implementation of a vision, including not only the process but also the people and resources needed. (Leadership)** *
Mark only one oval.

- Agree
- Neutral
- Disagree

33. **Q33: Rather than being annoyed when team members have issues preventing them from doing their jobs effectively, I see the issues as an opportunity to serve and help those people. (Leadership)** *
Mark only one oval.

- Agree
- Neutral
- Disagree

34. **Q34: I find great personal satisfaction in helping other people/fellow students/siblings become more successful. (Leadership)** *
Mark only one oval.

- Agree
- Neutral
- Disagree

35. **Q35: I use my emotional energy to motivate others. (Leadership)** *
Mark only one oval.

- Agree
- Neutral
- Disagree

Non-Cognitive Factors (5/5)

Community Support

36. **Q36: There is someone I can turn to for advice about handling problems with my family. (Community Support)** *
Mark only one oval.

- Agree
- Neutral
- Disagree

37. **Q37: There is someone in university that I can turn to for advice when I face academic problems. (Community Support)** *
Mark only one oval.

- Agree
- Neutral
- Disagree

38. **Q38: If I were sick, I could easily find someone to help me with my daily chores. (Community Support)** *
Mark only one oval.

- Agree
- Neutral
- Disagree

39. **Q39: I feel that there is no one I can share my most private worries and fears with. (Community Support)** *
Mark only one oval.

- Agree
- Neutral
- Disagree

40. **Q40: I often get invited to do things with others. (Community Support)** *
Mark only one oval.

- Agree
- Neutral
- Disagree

41. **Q41: If I needed some help in moving to a new house or apartment, I would have a hard time finding someone to help me. (Community Support)** *
Mark only one oval.

- Agree
- Neutral
- Disagree

42. **Q42: If I decide one afternoon that I would like to go to a movie that evening, I could easily find someone to go with me. (Community Support)** *
Mark only one oval.

- Agree
- Neutral
- Disagree

BIBLIOGRAPHY

[1] P. Hoodbhoy, "Pakistans higher education system what went wrong and how to fix it," *The Pakistan Development Review*, vol. 48, no. 4, p. 581594, 2009. [Online]. Available: http://www.pide.org.pk/pdf/PDR/2009/Volume4/581-594.pdf

[2] HEC, "Hec recognized universities and degree awarding institutions," 2016. [Online]. Available: http://www.hec.gov.pk/english/universities/Pages/DAIs/HEC-Recognized-Universities.aspx

[3] "11th five year plan," Government of Pakistan, Tech. Rep., 2011. [Online]. Available: http://www.pc.gov.pk/wp-content/uploads/2015/11/Ch1-Background-and-Vision1.pdf

[4] T. Nation.com, "Higher education in pakistan," 2016. [Online]. Available: http://nation.com.pk/letters/13-Aug-2016/higher-education-in-pakistan

[5] A. Latif, A. Choudhary, and A. Hammayun, "Economic effects of student dropouts: A comparative study," *Journal of Global Economics*, vol. 3, no. 137, 2015. [Online]. Available: http://www.esciencecentral.org/journals/economic-effects-of-student-dropouts-a-comparative-study-2375-4389-1000137.php?aid=57059

[6] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting student dropout in higher education," *ArXiv e-prints*, 2016.

[7] Unicef, "Statistics:pakistan," Unicef, Tech. Rep., 2013.

[8] R. B. Sachin and M. S. Vijay, "A survey and future vision of data mining in educational field," *ACCT '12 Proceedings of the 2012 Second International Conference on Advanced Computing & Communication Technologies*, pp. 96–100, 2012.

[9] B. K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance," *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 6, pp. 63–69, 2011.

[10] S. B. Kotsiantis, "Use of machine learning techniques for educational proposes: a decision support system for forecasting students grades," *Artificial Intelligence Review*, vol. 37, no. 4, p. 33144, 2011.

[11] M. A. Al-Barrak and M. Al-Razgan, "Predicting students final gpa using decision trees: A case study," *International Journal of Information and Education Technology*, vol. 6, no. 7, p. 528533, 2016.

[12] K. D. Kolo, S. A. Adepoju, and J. K. Alhassan, "Decision tree approach for predicting students academic performance," *IJEME International Journal of Education and Management Engineering*, vol. 5, no. 5, p. 1219, 2015.

[13] D. Kabakchieva, "Predicting student performance by using data mining methods for classification," *Cybern Inf Technol*, vol. 3, no. 1, p. 6172, 2013.

[14] R. Asif, A. Merceron, and M. K. Pathan, "Predicting student academic performance at degree level: A case study," *International Journal of Intelligent Systems and Applications*, vol. 7, no. 1, p. 4961, 2014.

[15] P. Golding and S. Mcnamarah, "Predicting academic performance in the school of computing & information technology," *Proceedings Frontiers in Education 35th Annual Conference*, pp. 16–20, 2005.

[16] A. Zafra, C. Romero, and S. Ventura, "Predicting academic achievement using multiple instance genetic programming," *2009 Ninth International Conference on Intelligent Systems Design and Applications*, p. 307314, 2009.

[17] B. Khan, M. S. H. Khiyal, and M. D. Khattak, "Final grade prediction of secondary school student using decision tree," *International Journal of Computer Applications*, vol. 115, no. 21, p. 3236, 2015.

[18] W. Xing, R. Guo, E. Petakovic, and S. Goggins, "Participation-based student final performance prediction model through interpretable genetic programming: Integrating learning analytics, educational data mining and theory," *Computers in Human Behavior*, vol. 47, p. 168181, 2015.

[19] N. Thai-Nghe, L. Drumond, A. Krohn-Grimberghe, and L. Schmidt-Thieme, "Recommender system for predicting student performance," *Procedia Computer Science*, vol. 1, no. 2, p. 28112819, 2010.

[20] N. Thai-Nghe, P. Janecek, and P. Haddawy, "A comparative analysis of techniques for predicting academic performance," *2007 37th annual frontiers in education conference - global engineering: knowledge without borders, opportunities without passports*, vol. 37, 2007.

[21] A. M. Shahiri, W. Husain, and N. A. A. Rashid, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, p. 414422, 2015.

[22] D. E. Winter and D. Dodou, "Predicting academic performance in engineering using high school exam scores," *International Journal of Electrical Engineering Education*, vol. 27, no. 6, pp. 1343–1351, 2011.

[23] S. A. Naser, I. Zaqout, M. A. Ghosh, R. Atallah, and E. Alajrami, "Predicting student performance using artificial neural network: in the faculty of engineering and information technology," *International Journal of Hybrid Information Technology*, vol. 8, no. 2, p. 221228, 2015.

[24] P. C. Kyllonen, Ed., *The Case for Noncognitive Assessments*. R&D Connections, 2005.

[25] A. Wigfield and R. D. Harold, "Teacher benefits and childrens achievement selfperceptions: A developmental perspective," *Student Perceptions in the Classroom*, pp. 95–121, 1992.

[26] S. Janssen and M. O'Brien, "Disentangling the effects of student attitudes and behaviors on academic performance," *International Journal for the Scholarship of Teaching and Learning*, vol. 8, no. 2, 2014.

[27] C. J. Flynt, "Predicting academic achievement from classroom behaviours," Ph.D. dissertation, Virginia Polytechnic Institute and State University, August 2008. [Online]. Available: https://theses.lib.vt.edu/theses/available/etd-09162008-100711/unrestricted/Dissertation4.pdf

[28] F. A. Adebayo, "Time management and students academic performance in higher institutions, nigeria a case study of ekiti state," *International Research in Education*, pp. 1–12, 2015.

[29] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," *Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC)*, 2008.

[30] "Weka 3: Data mining software in java." Tech. Rep., 2016. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/

[31] I. Etikan, S. Musa, and R. S. Alkassim, "Comparison of convenience sampling and purposive sampling," *American Journal of Theoretical and Applied Statistics*, vol. 5, no. 1, pp. 1–4, 2016.

[32] "Study habit questionnaire. adopted from virginia gordons university survey: A guidebook and readings for new students."

[33] I. G. Sarason, H. Levine, and R. Basham, "Assessing social support: The social support questionnaire," *Journal of Personality and Social Psychology*, vol. 44, pp. 127–139, 1983.

[34] "Leadershipskillsquestionnaire." [Online]. Available: http://www.csueu.org/DesktopModules/Bring2mind/DMX/Download.aspx

[35] "Student self-assessment questionnaire. ryerson university." [Online]. Available: http://www.ryerson.ca

[36] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," *Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)*, pp. 5–12, 2008.