Modeling NUST MS Admission Policy / Process using Machine Learning Methods

By

Tariq Jamil

(Spring19 - MS CS&E - 00000281815)

Supervisor

Dr. Zamir Hussain

Department of Computational Science and Engineering

Research Centre for Modelling and Simulation (RCMS)

National University of Sciences and Technology (NUST)

Islamabad, Pakistan

December 2021

Modeling NUST MS Admission Policy / Process using Machine Learning Methods

By

Tariq Jamil

(Spring19 - MS CS&E - 00000281815)

Supervisor

Dr. Zamir Hussain

A thesis submitted in conformity with the requirements for

the degree of *Master of Science* in

Computational Science and Engineering

Department of Computational Science and Engineering

Research Centre for Modelling and Simulation (RCMS)

National University of Sciences and Technology (NUST)

Islamabad, Pakistan

December 2021

DEDICATION

I dedicate this dissertation to my Parents.

For their endless love, support and encouragement

Declaration

I, *Tariq Jamill*, declare that this thesis titled “Modelling NUST MS Admission Policy / Process using Machine Learning Methods ” and the work presented in it are my own and has been generated by me as a result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a Master of Science degree at NUST
2. Where any part of this thesis has previously been submitted for a degree or another qualification at NUST or any other institution, this has been clearly stated
3. Where I have consulted the published work of others, this is always clearly at-tributed
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work
5. I have acknowledged all main sources of help
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself

Tariq Jamil,
00000281815

Copyright Notice

Acknowledgments

I would first like to thank my thesis supervisor Dr. Zamir Hussain and Co-supervisor Dr, Adnan Maqsood of the Research Centre for Modelling and Simulation, NUST. The door to his office was always open whenever I was stuck in a problem related to development or thesis writing. He consistently allowed this thesis to be my own work, but always guided me to right direction whenever he thought I needed it. I would also like to pay special thanks to Dr. Mian Ilyas Ahmed, Principal RCMS and Dr. Rizwan Riaz, Pro Rector, Research Innovation and commercialization (RIC) , NUST. He remained a source of motivation for this project with his timely support. I would also like to thank my other GEC members including Dr. Muhammad Tariq Saeed and Dr. Rizwan Riaz for their support and help throughout the research phase.

I would also like to thank Management of RCMS, NUST was very supportive and was there to assist for arranging presentations and demos of this research work. During the research phase, my fellow researchers of Data Analyst Group (DAG) also helped me during research phase.

Contents

List of Abbreviations and Symbols

GAT	Graduate assessment test
INT	Interview
ACAD	Academics
ANN	Artificial Neural Network
DT's	Decision Trees
MLT	Multi-layer Perceptron
RBF	Radial Based Function
NUST	National University of Science & Technology
PCA	Principal Component Analysis
ICT	Information and Communication technology
SPSS	Statistical Package for the Social Sciences
TPR	True Positive Rate
TNR	True Negative Rate
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

List of Tables

List of Figures

Abstract

Universities aims to identify and admit the applicants who will perform best. Therefore, they usually base their admission decisions/processes on a combination of various characteristics or measures of the applicants. Development of consensus between different universities/institutions on a common set of measures for the process of selection is not easy. However, homogeneity can be achieved by introducing dynamic features supported by empirical analysis of indigenous data. The effects of uniformity in the admission processes are manifold. It can bring harmony, respect, and better coordination between universities. Secondly, it will give confidence to the applicants with better focus to develop their attitude towards evident features of assessment. Admission policy is often seen as a quality response to upcoming generations and, consequently, higher education institutions have become a wide source of producing new talented scholars and doctors in the last decade. This research study analyzed and investigated the relationship, effectiveness and weightage distribution of different variables / parameters used in the admission process of National University of Sciences and Technology (NUST) at postgraduate level. NUST admission policy / process consists of three variables: i) interview (INT) conducted at the time of admission by a concern school/center/institution of NUST, ii) academic record of the student (ACAD), and iii) graduate assessment test (general) (GAT) (a test conducted by National Testing Service of Pakistan for higher education commission of Pakistan) or graduate record examination (general). The current weightage of INT in current admission policy is 25%, ACAD is 25%

and GAT is 50%. Since this research study is an empirical analysis, therefore, an archival contains the data of 13094 applicant is used. Span of the data is seven years, provided by the ICT directorate of NUST. Evidence collected from the literature review, the range and size of the data used for analysis are adequately sufficient to derive significance results and conclusions regarding the effectiveness of the process. Comparative analysis has been used to analyze the relationship of these variables between the admitted and not admitted student. Interview and academics have a statistically significant linear relationship with r =0.203, and $p < 0.01$. Furthermore, Cohan's d analysis has been used to analyze the practical significance of these variables between the admitted and not admitted student. These result shows that the interview variables have a marginal difference from GAT and ACAD. Results shows that Interview (INT) is practically significant, and effect of the association is large with $d = 0.704$. Furthermore, principal component analysis has been used for dimension reduction and suggestion of new weightages to these variables. By using the coefficient of linear relationship between these three variables, the new suggested weightages are 36.15% for interview, 31.35% for GAT and 32.50% for ACAD.

Moreover, different machine learning models are developed to check the predictive ability of these variables. Machine learning model's includes radial based function (RBF), decision trees (DT), multi-Layer perceptron (MLP) and binary logistic regression (BLR). Results reveals that the average accuracy of these models is 63.1%. Results concluded from the predictive modeling shows that these variables are not balanced in terms of subjective weightage assigned to them. Furthermore, predictive models also concludes that these variables are not complete because of lack in predictive ability. Therefore, there is a need to review the weightages and includes other variables in the analysis like program popularity, financial stability of the applicant, location and place of residence, facility of hostel etc.,

These all factors will provide useful insight for the suggestion of new admission policy. Further research for analyzing data of different universities can be a step towards uniform national admission policy at postgraduate level.

Keywords: *Admission, Admission Policy, Graduate assessment test, Education Quality, Academics, Modern Education, Admission Process*

Chapter 1

Introduction

In this chapter, background and introduction of the research project are discussed. In this research, admission policy for postgraduate students by using machine learning method is presented. A statistical and machine learning based approach is used to analyze the existing policy for postgraduate students. This research is also interrogated whether the weightages of existing policy for post graduate student is properly distributed or not. In the subsequent sections, these admission policies are discussed in context with the research work. Afterwards, different types of techniques to suggest new weightages for the stated variables. This study aims to analyze three important features of MS admission policy followed at National University of Sciences and Technology (NUST). The idea is to analyze the appropriateness, completeness and distribution of the variables being assessed. The current admission policy of NUST for postgraduate students depends on three characteristics /variables:

i. Previous academic record of an applicant (ACAD),
ii. Graduate record examination (general) (GRE) or graduate assessment test (general) (GAT) (A test conducted by national testing service of Pakistan for HEC),
iii. Interview (INT) (conducted in the concerned school/department).

Further details of each variable are as follows:

a) ACAD: The score of the variable ACAD, out of 25, has been allotted for each applicant. Cumulative grade point average (CGPA) or percentage obtained in terminal degree/transcript of the applicant has been used to obtain the score of ACAD. The weightage of this variable in NUST existing admission policy is 25%. The details of the formula or ranges to allocate marks of ACAD are provided in Appendix-I.

b) GAT: It is a test conducted by National Testing Service (NTS) of Pakistan, which is mandatory for admission in MS/M.Phil. and PhD programs in National University of Sciences and Technology (NUST). This test is consist of 100 multiple choice questions and having sections of analytical reasoning, verbal reasoning, quantitative reasoning with weightages varying for various disciplines. The total score of this test is 100 and a candidate having at least 50 marks is considered as qualified / passed. The weightage of this variable in NUST existing admission policy is 50%.

c) INT: To evaluate the suitability of a candidate, an interview is conducted in the concerned/specific institution/school/center of NUST. The procedure used to obtain the marks of INT for each candidate is based on Performa including various attributes assessed by the interview committee. This Performa is available in Appendix II. The weightage of this variable in NUST existing admission policy is 25%.

Another aim of this study is to investigates appropriateness of the current weightages of the three variables used for the calculation of merit of an applicant. The purpose is to propose an adequate model using machine learning methods based on the three stated variables as guidelines for the development of uniform admission process/system of universities for the student’s seeking admission in post-graduation (MS/MPhil/PhD) programs. Literature reveals that the standardization and modernization of admission policy has positively influence on the quality of education

1.1 Admission Policy

There is proceedings debate about how to improve the quality of education, reduce socio-economic and different segregation in universities. To this end, many countries have affirmative action programs, intended to increase college admission rates for targeted populations.

We are a conscious learning focused nation where each student is motivated to accomplish their potential and to turn this inspiration into an ethically confident, and internationally minded citizen of tomorrow. The NUST community is joined by a shared obligation to this Mission. It drives all parts of university life, including admission in university. Admission to NUST is very competitive, however open to all students who will benefits from a different programs and extracurricular activities, within internationally accepted framework. Successful candidates show high level of motivation and steadiness.

In general, existing admission policy of NUST focus on different process to check the applicant eligibility. The main characteristics which are involved in the process of evaluation is graduate assessment test (GAT), Interview (INT) and academics (ACAD). On the basis of these three variables a merit is calculated through which applicant can get admission in different universities.

NUST seek dynamic, open-minded, and responsible students who will contribute to our diverse school community, and whose desire to learn is matched by their enthusiastic engagement with peers and school faculty and staff. Our curriculum embraces an intentionally international perspective, and we encourage students to take full advantage of the opportunities available to them in academics, athletics, community service and extracurricular activities, with other students from around the world.

1.2 Impact of Admission Policy on Education

Admission policy plays a vital role in the enrollment of a student and introduce new opportunities to brings students in different career fields. In current situation, the ability to get admission in higher education is a main challenge. The ability of getting admission in higher education and who does not, is consequently a significant issue in shaping a dynamic and progressive societies. While in other aspect admission systems have the task of selecting those who have the potential to succeed in higher education. Therefore, admission policy/process have a huge impact to provide an efficient and effective route to study successfully. Post graduate admission is not a simple process which occurs at the end of bachelor’s education. It is a process which may begin from the moment a student is streamed in bachelor’s education.

1.3 Existing Admission Policy of NUST

Current admission policy of NUST depends on three main characteristics

1.3.1 Graduate Assessment Test

It is a test conducted by National Testing Service (NTS) of Pakistan, which is mandatory for admission in MS/M.Phil. and PhD programs in National University of Sciences and Technology (NUST). This test is consist of 100 multiple choice questions and having sections of analytical reasoning, verbal reasoning, quantitative reasoning with weightages varying for various disciplines. The total score of this test is 100 and a candidate having at least 50 marks is considered as qualified / passed. The weightage of this variable in NUST existing admission policy is 50%.

1.3.2 Academics

The score of the variable ACAD, out of 25, has been allotted for each applicant. Cumulative grade point average (CGPA) or percentage obtained in terminal degree/transcript of the applicant has been used to obtain the score of ACAD. The weightage of this variable in NUST existing admission policy is 25%. The details of the formula or ranges to allocate marks of ACAD are provided in Appendix-I.

1.3.3 Interview

To evaluate the suitability of a candidate, an interview is conducted in the concerned/specific institution/school/centre of NUST. The procedure used to obtain the marks of INT for each candidate is based on performa including various attributes assessed by the interview committee. This performa is available in Appendix II. The weightage of this variable in NUST existing admission policy is 25%.

1.4 Research Questions

Every year hundreds of graduate students enrolled in graduate schools for higher education and this trend will continue in future as well. Therefore, there is a need of standardized and modernize the graduate education admission process to sustain the quality of education and career of upcoming generations. The main research questions are

i. Is the current admission process of NUST for post Graduate student is appropriate?

ii. Are the weightages given to different variables is balanced, in terms of characteristics being assessed?

1.5 Objectives

This study provides an in-depth analysis of three important characteristics/variables of admission process of applicants at postgraduate level considered by NUST. These variables are GAT, ACAD, and INT as stated above. The analyses will be based on secondary information/data of the stated factors of more than thirteen thousand applicants intended for admission in various disciplines/degree programs for a span of 2008 to 2014. The results of the study would be a step towards uniform admission procedures/systems/policies for postgraduate students in the universities. Based on the stated questions in the research question section, the main objectives of this study are:

i. To analyze trends and tendencies of variables with respect to the performance of admitted and not admitted students using comparative analysis including correlation analysis and testing practical significance through effect size calculations, etc. Secondly, investigating various myths related to GRE and ACAD or ACAD vs INT or GRE VS ACAD or GRE VS INT, etc. For instance, does interview marks effected by performance in GRE and/or ACAD?

ii. Development of a predictive model for estimating chances of admission of an applicant by using machine learning methods like decision trees multi-level perceptron etc.

iii. Performance evaluation of developed models in terms of precision and accuracy.

iv. Assess the adequacy of the current weightages of variables and propose new weightages, if required.

1.6 Scope of the study:

The proposed study will proceed with the following limitations:

i. There may exist more factor in the process of admission in various universities, i.e. no: of seats as per department, specific program popularity, proposal of the research, etc. The results of this study will be limited to the provided data and stated factors.

ii. The analysis related to finding results based on the secondary data already collected by the principal investigators of the study. Furthermore, analysis will depend upon the appropriateness of data/information and availability of data including the factors which are being missed.

iii. In this study we only focus on the practical significance of existing admission policy and weightage distribution of the variables used in this policy.

Chapter 2

Literature Review

For the last many years, the process of admission in academics institutions were investigated in several studies using various types of data sources and methods. Literature shows that there exist a variety of characteristics assessed for screening of an applicant for admission especially at post graduate level. Some of them includes undergraduate grade point average (GPA) of their terminal degree, recommendation letters from supervisors and faculty members, Graduate Record Examination (GRE) or Graduate Management Admission Test (GMAT) and Interview (INT) etc.

In view of the question of this research project: " Is the admission process of NUST for post graduate student is complete & appropriate? Are the weightages of different variables are balanced? ". The supreme focus of the research is to analyze the completeness of postgraduate student at NUST. If we study the admission process, we will see different variables used for the evaluation of an applicant. There are different research studies available around the world however none of them focus on completeness and objectivity of the admission process. There is no empirical and scientific study available in Pakistan which focus on admission process.

2.1 Admission Process Rational

An Image of filtering process arise in our mind when we think about the admission process in desired universities or colleges. The principal behind this selection is the competition between the different applicants as well availability of less slot and more applicants. Moreover, the main focus and purpose of the admission process is to choose the best applicant on the basis of their academic achievements [13]. Thus, to solve the issue of selecting best applicant, there is a need of uniform and efficient admission process [14].

2.2 Before 2000 Historical Literature:

GRE/GMAT/GAT and previous cumulative grade point average (CGPA) are used to select highest scoring individuals (Goldenberg & Alliger, 1992). In perspective of applicants, the academic characteristics of the students should be properly assessed. The admission process needs to introduce a standard evaluation system.

The study of Thornel and McCoy’s (1985) suggested a strong relation between GRE and GGPA. Their results were base on (n=582 graduate students) and (coefficient of correlation = r = 0.43 having p-value < 0.05).

Harvanicl & Gordon (1986) collected data from 619 master’s degree graduate students to determine, the relationship between graduate grade point average (GGPA) and GRE scores. Their results were also statistically significant with (r = 0.48, p < 0.05).

Milner, King and McNeil (1984) reveal the results from a group of 145 full time graduate students and found statistically insignificant relationship between GRE and GGPA (r = 0.238). Robertson and Nielsen (1961) also found a weak correlation (r = 0.29) between the GRE and the faculty rating students potential to complete Ph.D. program.

William & Robert (1977) stated that the admission process for graduate school of medium size and state supported university is analyzed by using a combination of bayesian and cost/benefit decision analysis technique. First technique require that the researcher identify the decision-making group headed by the Dean of the graduate school. Second technique identify the selection criteria upon which to base the selection process i.e., academic performance, test scores and establishing cutoff score for each criteria. Method which is used for organize information required in decision analysis is to build a decision matrix (William L & Lynch, 1977).

2.3 Post 2000 Literature:

In 2014, Nienke R Schripsema, Anke M van Trigt, Jan C C Borleffs conducted a study to analyze the performance of students admitted to a school based on pre university grades, multifaceted selection process and lottery. They collected a data (n = 1055) from university of Groningen Netherland with a span of three years (2009, 2010, 2011). The sample data is divided into two categories (i) Student who had not participated in multifaceted selection process, (ii) Students who had been rejected in multifaceted selection process. They used logistic regression, anova modelling and Bonferroni post hoc multiple comparison tests to analyze the collected data. The study reported that the pre-university grade group achieved higher knowledge test score and highest professionalism score more often than the lottery admitted group that had not participated in the multifaced selection process.

In 2018, Heena Sabnani, Mayur More, Prashant Kudale have used the techniques of data mining and machine learning to analyze the scenario of admission by predicting the enrolment behavior of students. They have used the Apriori technique to analyze the behavior of students who are seeking admission to a particular college. They have also used the Naïve Bayes algorithm which will help students to choose the course and help them in the admission procedure. In their project, they were conducting a test for students who were seeking admissions and then based on their performance, they were suggesting students a course branch using Naïve Bayes Algorithm. But human intervention was required to make the final decision on the status.

In 2020, Amal AlGhamdi, Amal Barsheed, Hanadi AlMshjary, Hanan AlGhamdi used the techniques of the machine learning and statistics to automatically predict the best suitable university for post graduate admissions. They used logistic regression model, linear regression and decision trees. This paper evaluates these models to select the best model in term of accuracy and least error. Logistic regression model shows the most accurate prediction in this paper. Table 1 contains the variables which is discussed and analysis in different research studies.

Table 1: Admission Variables Summary

Year of Publication & Author's	Variables Names
Moruzi and Norman (2002)	i. GPA ii. Autobiographical submission iii. Simulated Tutorial iv. Personal Interview v. Licensing Examination
Nienke R. S, Anke M. T, J.C. Borleff and , Janke C.S (2014)	i. GPA ii. Course Credit (First Year) iii. Test Scores
Thomas, Barbara, and Razack (2017)	i. GPA ii. Personal Statements iii. MMI iv. Reference Letters
Abdulmohsin and Abdulaziz (2020)	i. SAAT ii. High school grade iii. GAT

The concept of admission policy and relationship between variables came in 1961 with several advantages. Previous literature found the relation between different variables by using different methods which is descriptive statistics, Bayesian and cost/benefit decision. The main aim of this research is to analyze the admission process of NUST keeping in view the good practices suggested in various published studies.

Following are the details of different heterogeneous variables used in the different research study. Shown in figure 2.1. Therefore, uniform admission process in all universe is not easy.

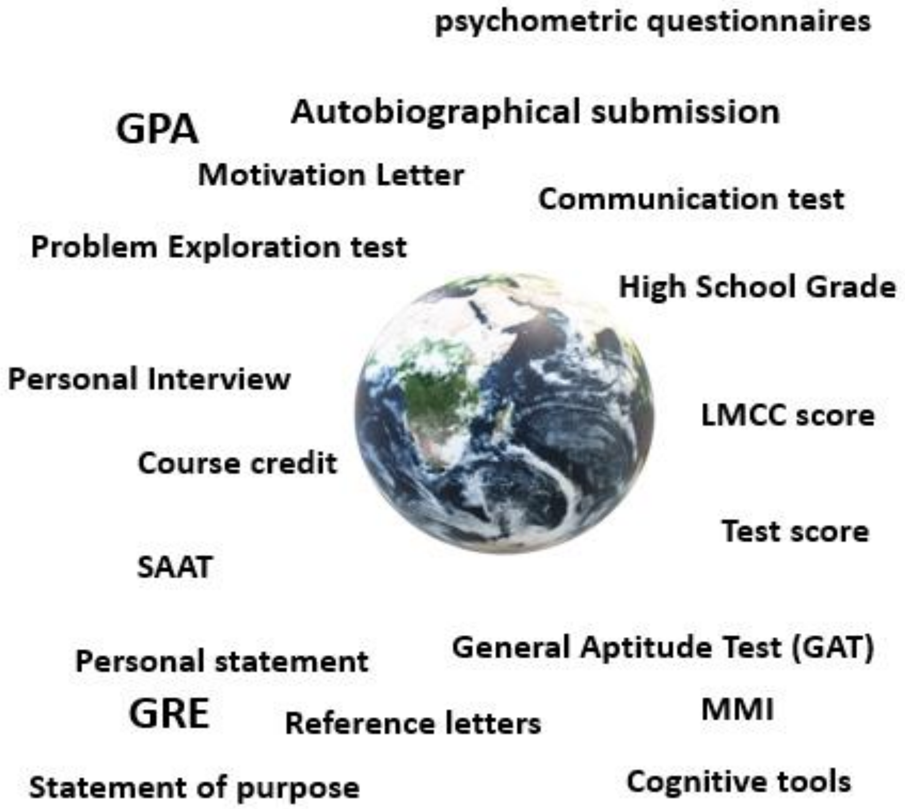


Figure 1 : Heterogeneous Variables around the globe

2.4 Admission Criteria of NUST for PG Students

Being the leading top university of Pakistan, NUST followed the same pattern mentioned in earlier literature. NUST followed the pattern of admission variables which is Graduate assessment test (GAT), Academics (ACAD) and Interview (INT), detail of these variables are available in the 1st chapter. These three variables are used to choose the best candidate. This research study is primarily focus on the completeness and objectivity of the current admission policy of NUST for post graduate students.

To quantify these variables university used the merit formula for an applicant. In current merit formula the weightage of GAT is 50%, ACAD is 25% and INT is 25%. The fundamental question concerning the process in practice is : Are the properly or improperly differentiate between admitted students and not-admitted students? This research will be based on the secondary data provided by the information and communication department of NUST. The span of the data is more then five years. The dataset contains thirteen thousand applicants' data. This research will focus on the relationship between these variables, Practical significance and the suggestion of new weightages to aforementioned variables. The result would be a step towards uniform national admission process for post graduate students in the universities of Pakistan.

Chapter 3

Methodology

In this section, we are going to discuss the methodology used in this research to achieve the objectives. As the main objective of this research is to find the relationship between different variables by using correlation analysis techniques. In this research study, a dataset of 13094 applicant's is analyzed and used the method of principal component analysis to reduce the dimension and proposed new weightages. Use of different machine learning methods for the development of predictive models. In the upcoming subsections, the methodology of these research study is discussed in detail.

3.1 Data Collection

This study depends on recorded method, which includes describing information that existed before the start of this research study. Data of all applicants are collected from different schools, colleges, and research center of NUST located at H-12 sector of Islamabad for the time span of 2008 to 2014 has been utilized for the current investigation. Before 2008, this data was not available in electronic form. An applicant was characterized as an individual who had applied at any school/college/research centre of NUST in postgraduate programs. The complete information of the applicants is provided by the information and communication technologies (ICT) directorate of the NUST. The branch of ICT, NUST has collected this information by gathering it from every schools/colleges/research centre of NUST for universities record. Data consist of 13094 number of applicants with four characteristics which is ACAD, GAT, INT, and STATUS of the applicant in which 5458 applicants' value is admitted and 7636 not admitted. The sample size used for investigation of this research is very large to conclude the evidence regarding the stated variables. The results will provide the valuable guidelines to NUST for the regulation of admission process.

3.2 Correlation Analysis

Correlation analysis is the technique in which we measure the strength of relationship between different variables. In this research study correlation analysis is used to break the different myths related to GAT vs INT, GAT vs ACAD and INT vs ACAD. For determining the relationship between these independent variables, we use the techniques of correlation. The coefficient of correlation will be used to calculate the intensity and direction of linear relationship between the continues independent variables. Coefficient of Correlation is denoted by " r ". The range of correlation " r " is between 0 and 1.

- If the value of r is close to 0, then there is a weak correlation.
- If the value of r is close to 1 then there is a strong correlation.
- A positive sign with the value of r means a positive correlation, which means both variables is directly proportional to each other.
- A negative sign with the value of r means a negative correlation, which means both variables is inversely proportional to each other.
- If the value of r is +1, then there is perfect positive correlation.

- If the value of r is -1, then there is perfect negative correlation.

3.3 Practical Significance

Usually there are two terms which is used in statistics. Statistical significance in which we use to check the significance of a hypothesis by using t-test and practical significance is the technique in which we use Cohan's *d* method to check the significance. Cohan's D is the most common technique which is used for calculating effect size or practical significance. The general rule of thumb which Cohan's said used be cautiously. The measure of Cohen (1988, 1992) can be used as a guideline when measuring the correlation coefficient. Table:4.2 indicates the measurement scale for the correlation coefficient to check the practical significance.

Table 2 – Correlation measurement scale (Cohan 1998, 1992)

Correlation Coefficient	Effect	Explanation
d = 0.20	Small	The effect explains 1% of the total variance
d = 0.50	Medium	The effect explains 9% of the total variance
d = 0.80	Large	The effect explains 25% of the total variance

The formula which is used for Cohan's *d* effect size is :

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\text{n}}$$

Where

- *$\bar{x}$: mean*
- *n : sample size*
- *s : standard Deviation*

Where *s* can be calculated using this formula for unequal group size:

$$\mathbf{n} = \frac{\sqrt{(\mathbf{n_1} - \mathbf{1})\mathbf{s}_1^2 + (\mathbf{n_2} - \mathbf{1})\mathbf{s}_2^2}}{\mathbf{n_1} - \mathbf{n_2}}$$

Where *s* can be calculated using this formula for equal group size:

$$\mathrm{n} = \sqrt{s_1^2 + s_2^2}/2$$

Pooled variance is used when there are two different sample sizes. The large sample mean will affect more the total mean of both samples. To overcome this difference pooled variance is used.

3.4 Dimension Reduction

3.4.1 Principal Component Analysis

Large dataset is commonly used in research, and it is difficult to interpret with large data set. Many techniques are being developed for this purpose, but Principal component analysis (PCA) is one of the oldest and widely used technique. Principal component analysis (PCA) is a technique for compressing a lot of data into something that captures the essence of original data. Idea of PCA is to reduce the number of variable while preserving as much information as well. PCA searches to find the linear combination with largest variances, and then divide them into Principal Components (PC). In PCA the largest variance is captured by the highest component in order to extract the fruitful information.

Steps of Principal Component analysis:

Step 1: **Standardization**

The main goal of this step is to normalize the range of all continues variables so that every single one of them contributes equally to analysis. The step of standardization is more important prior to PCA. If there are a large difference between the ranges of different variables, the higher ranges will dominate over the small one's which will prompt one sided result.

Mathematically, this step can be done by subtracting the mean of each variable from the value of each variable and dividing by standard deviation.

$$Z = \frac{x - \bar{x}}{Std.Dev}$$

Step 2: **Covariance Matrix Computation**

The main goal of this step is to understand, how the value of input dataset is different from the mean of each value respectively or to see the relationship between them. Sometimes the input data contain more redundant data which is related to each other. To identify this relationship, we compute the covariance matrix.

The covariance matrix is a p × p symmetric matrix (Number of dimensions = p) that has as entries the covariances associated with all possible pairs of the initial variables. For a 3 × 3 matrix the covariance matrix will be of this form:

$$\begin{bmatrix} Cov\,(x,x) & Cov\,(x,y) & Cov\,(x,z) \\ Cov\,(y,x) & Cov\,(y,y) & Cov\,(y,z) \\ Cov\,(z,x) & Cov\,(z,y) & Cov\,(z,z) \end{bmatrix}$$

Since the covariance matrix is the variance of itself. As well the covariance matrix is the commutative (Cov(a,b) = Cov(b,a)), the entries of the covariance matrix are symmetric with respect to the main diagonal, which means that the upper and the lower triangular portions are equal.

Step 3: To identify the principal components compute the eigen values and eigen vectors

This both eigenvectors and eigenvalues are the linear algebra concepts which will be compute from covariance matrix to identify the principal components of the data as well the variance of these principal components. The eigenvalues and eigenvectors are always come in pairs, every eigenvector has an eigenvalue. And the number of the eigenvalues and eigenvector is equal to the numbers of the dimensions of the data. Before going to start the explanation of these concepts, let’s first understand the principal components.

Principal Component: Principal components are the new variables that are constructed, which is the linear combination of all variables. Preserving much information by computing a new variable - principal component which is uncorrelated and consider for the total variations of data. The Eigen value greater than 1 should be selected for further analysis according to Kaiser criterion.

So, the main idea behind principal components is 3-dimentional data will gives you 3 principal components, but principal component analysis tries to put maximum possible information in the first component, then in the second component and so on, until we got something like shown in the diagram scree plot below.

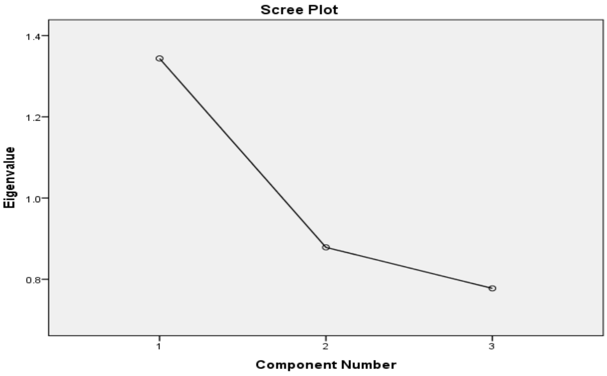


Figure 2: Scree Plot diagram of Eigen Value and principal Components

In view of the above diagram, principal components show the direction of the data that explain the amount of variance. The relationship between information and variance here is directly proportional, larger the variance by the line, larger the dispersion of the data will be. And larger the dispersion along a line, the more the information it has.

Computing Principal Components:

PC = Coefficient of (v1) * Standardized value (v1) + Coefficient of (v2) * Standardized value (v2) + Coefficient of (v3) * Standardized value (v3)

3.4.2 Weightage distributions

One important objective of this thesis is to analyze the weightage of different variables used in admission policy. The variables are Graduate assessment test (GAT), academics (ACAD) and interview (INT). The old weightages of the variables are defined as follows:

Table 3 – Existing Weightage Distribution of the Variables

Variable	GAT	ACAD	INT
Weightages	50%	25%	25%

In view of the objective, the principal component analysis will be used to define the new weightages. The eigen value of principal component one is more than 1 and the variation of principal component one is more then all. So, the coefficient of principal component one will be used in weightages distribution formula.

For weightages distribution we will use the following formula's:

i) Weightage Distribution (INT):

$$= \frac{Coefficient\ of\ \ (INT)}{Coefficient\ of\ \ (INT) + Coefficient\ of\ \ (ACAD) + Coefficient\ of\ (GAT)} * 100$$

ii) Weightage Distribution (ACAD):

$$= \frac{Coefficient\ of\ \ (ACAD)}{Coefficient\ of\ \ (INT) + Coefficient\ of\ (ACAD) + Coefficient\ of\ \ (GAT)} * 100$$

iii) Weightage Distribution (GAT):

$$= \frac{Coefficient\ of\ \ (GAT)}{Coefficient\ of\ \ (INT) + Coefficient\ of\ (ACAD) + Coefficient\ of\ (GAT)} * 100$$

3.5 Predictive Modeling using Machine Learning Methods

Machine learning has much importance and popularity in current trends because of modern technologies and artificial intelligence. Machine learning is mostly used to learn different patterns by training a model from different dataset and apply them on dataset for prediction. Nowadays, machine learning gained much supremacy over the old methods of analysis and classification because of fast prediction and high accuracy. However, many issues occurred during the training phase like overfitting, multi-collinearity which may make models suspicious, but these issues can be handled by the selection of suitable related techniques. In this project, noisy data and duplication of data were removed in the data cleaning step to avoid the different issues related to model training. Multiple machine learning models have been trained and built by using artificial neural network (ANN), and decision trees for classification by using SPSS (statistical package for social sciences) software in order to differentiate between admitted and not admitted students. SPSS (statistical package for social sciences) is a tool which is used for analysis of complex statistical data. It contains a collection of machine learning algorithms for predicative modeling and data analysis, it also has visualization function through which easily transform the data to attractive visual.

3.5.1 Artificial Neural Network

Artificial neural network (ANN) is being used by analyst and researchers of almost all fields for classifications, predictions, pattern recognition, voice recognition, and relatively competitive to conventional machine learning models and regression. It contains different algorithms which behaves like nervous system of human brain and process the information for computation by using interconnected nodes. A feedforward neural network (FFNN) contains many layers and nodes, and each node of a layer is connected to all other nodes in the next layers. Connections of on layer and nodes to another layer and nodes are not same because each connection can have a different strength. The output of all connections is generated when all the weight and strength pass through an activation function [162]. Besides the weights and strength nodes also contribute to final decision as if the summation of weight is positive, then output will be 1, if the summation of weight is negative, then output will be 0. The information passes from one layer to another layer until it reaches to output nodes. The flow of information is single direction which is called "feedforward" [163].

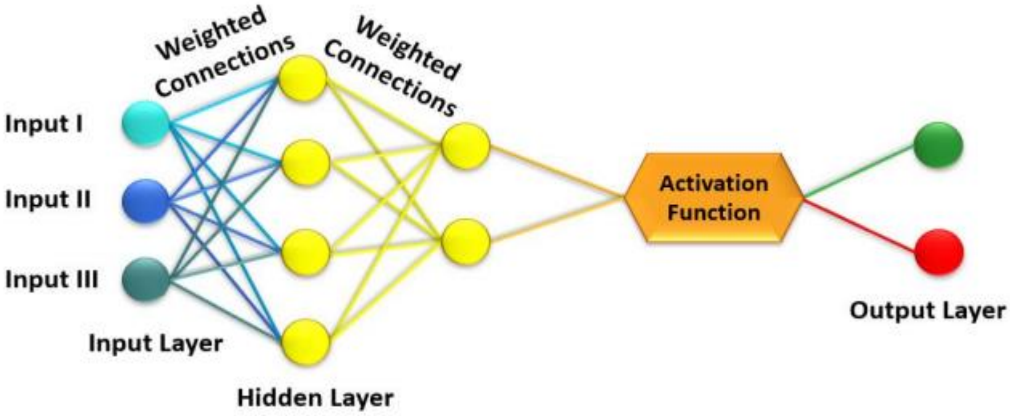


Figure 3: General Neural network with fundamental elements consists of input layer, hidden layer, and output layer

3.5.1.1 Multi-layer Perceptron

Multi-Layer perceptron (MLP) is a perception with multi layers. It consists of three layers or nodes: an input layer, a hidden layer and output layer. Except of the input layer, every layer is a neuron that uses a non-linear activation function. The input and output layer could be 1 or 0, but the hidden layer is anyone of them. Multi-layer perceptron uses a back propagation technique for training which is the part of supervised machine learning.

Input layer: This introduces input values to the network. No activation function needed

Hidden layer: This layer Perform classification of feature. Most probably two hidden layers is sufficient to solve any problem. The feature with lass layer will produce a batter result.

Output layer: hidden layer passes the function to the output layer to show the results.

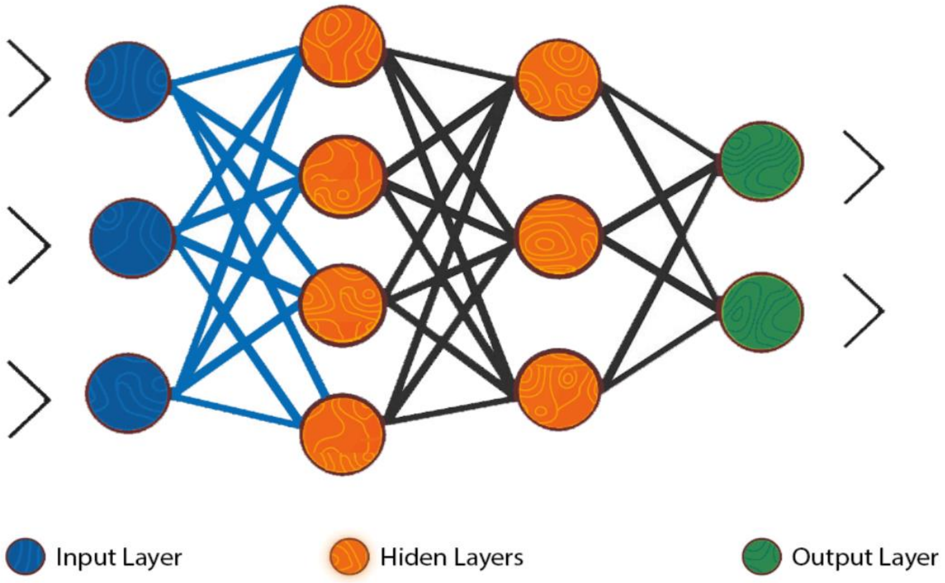


Figure 4: General flow of multi-layer perceptron consists of input layer, hidden layer, and output layer

3.5.1.2 Radial basis Function

There are three layers in radial basis function (RBF) embedded in a three-layer artificial neural network (ANN). RBF consists of an input layer, hidden layer which perform a radial activated function, and output layer which contains linear or non-linear neurons. In this artificial neural network, a single neuron is connected with all neurons of next layer but there is no connection exist within the same layer (Livingstone, 2019). In this neural network the input is non-linear, but the output is in linear form. Because of non-linear input RBF model, a complex mapping in hidden layer as compared to MLP which works with multiple hidden layers (Wilusz, 1995) (Bors, 2001). In hidden layer RBF used a non-linear Gaussian function while in the output layer RBF used a non-linear sigmoid function (Nguyen and Keip, 2018).

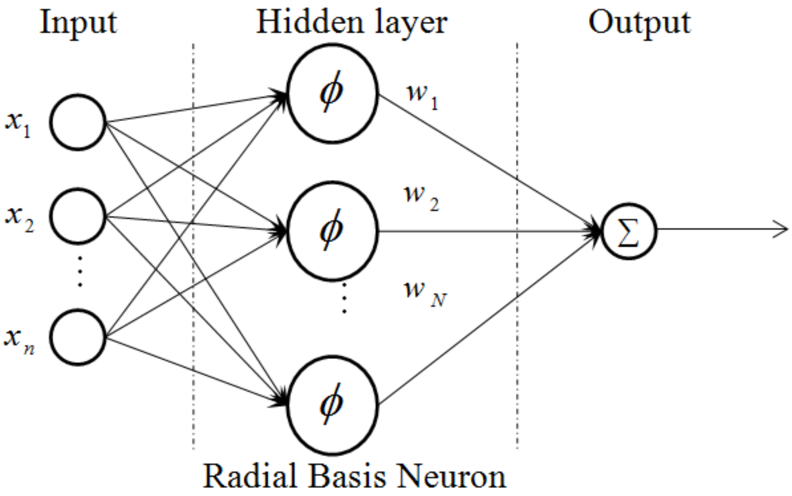


Figure 5: General flow of Radial Basis Function consists of input layer, hidden layer, and output layer

3.5.2 Decision Tree

Decision trees are in the category of supervised machine learning methods used for regression and classification tasks. Decision trees are closely related to human reasoning and very easy to understand. Decision trees sequential models consist of many nodes such as root node, internal node, and leaf node. The best corresponded predictor is the root node or the top most node. Root nodes will further be divided into two or more internal nodes which represents a “test”. While the internal node further divided into leaf nodes which represents the class label [158]. Our training set contains actives (labelled as 1) and in-actives (labelled as 0) based on the activity values (admitted =1, not-admitted = 0). Tree size may generate complexity which includes total number of attributes, total number of nodes, and total number of leaves affects model accuracy hence, the size should be controlled. For this purpose, pruning technique is used which not only reduces the complexity of the final classifier but also improves predictive accuracy by reduction of overfitting.

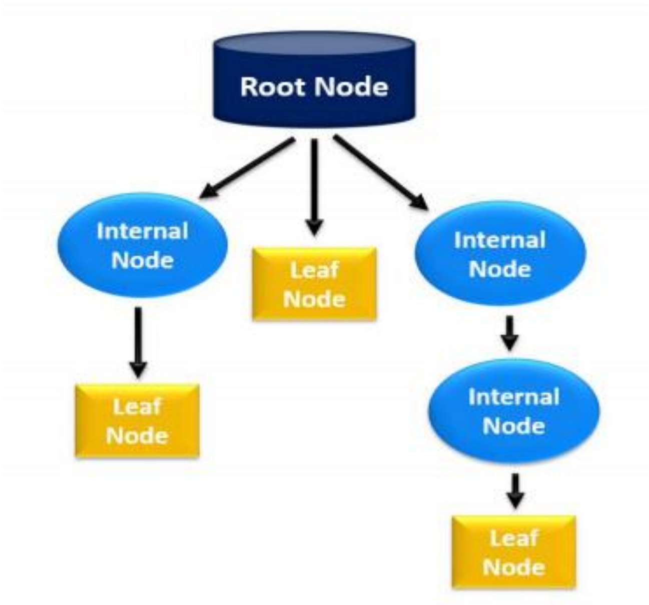


Figure 6: Basic scheme of Decision Tree. The most important node is root node and the leaf take final decision

3.5.2.1 Binary Logistic Regression

Predictive modeling is play's very important role in current analytical techniques. The main aim behind this is to check the significance of variables and different trends, pattern in the data. Already used different techniques for predictive modeling but the accuracy of these all analysis is not improved. The predictive modeling technique, binary logistics regression is commonly used, which is a well-known supervised learning technique. The variables used in logistic regression are carried out from principal component analysis. There are two independent principal component variables which is continues and one dependent variable, status of the student which is categorical. The justification of using binary logistic regression is the nature of our binary response variables instead of using linear regression technique. In the end binary logistic regression tells us about the probability of student being admitted. The tool used for this technique is SPSS and there are different method within the binary logistic regression which is forward likelihood ratio, forward Wald and backward Wald. I used the enter method to perform this analysis.

3.6 Assessment of Machine Learning Models

In this research study, we have independent variables in a form of student scores (GAT, ACAD and INT) while the dependent variable is student status (admitted / not admitted). Different machine learning models built which is selected based on different parameters used for analysis. It is very important to evaluate these models based on different parameters include true positive rate (TPR) or sensitivity, true negative rate (TNR) or specificity, and accuracy. Sensitivity tells us the rate of active that are correctly predicted as admitted and specificity explain rate of not admitted which are correctly predicted as not admitted. The formulas of these parameters are given below.

3.6.1 Sensitivity:

Sensitivity/Recall is also called true positive rate (TPR) is the ability to measure proportion of a students with positive index amongst all the students that are correctly identified. i.e. correctly measure the proportion of students who are admitted (Goutte and Gaussier, 2005).

$$\text{Sensitivity} = \frac{TP}{TP+FN} \qquad \textbf{3.6.1}$$

3.6.2 Specificity

Specificity which is also called true negative rate (TNR) is the ability of a test to measure the proportion of a students with negative index amongst the total actual negative instances. i.e., correctly measure the proportion of student who are not admitted (Goutte and Gaussier, 2005).

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{3.6.2}$$

3.6.3 Accuracy

Accuracy is the defined as the level of the right forecasts made by the model and is the representation of connection between the sensitivity and specificity (Goutte and Gaussier, 2005).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP} \tag{3.6.3}$$

3.6.4 Precision

It is additionally considered as positive predictive value that defines the occurrence of a positive class among the overall predicted positive observations. (Goutte and Gaussier, 2005).

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3.6.4}$$

Table 4: Confusion Matrix

Test Outcome		**Predicted**	
		Positive (1)	**Negative (0)**
Observed	**Positive (1)**	TP	FP
	Negative (0)	FN	TN

Descriptive details of assessment analysis are provided below:

i. True positive (TP): It represents admitted students that are detected admitted by the model.

ii. True negative (TN): It represents is the probability of not admitted students that are detected not admitted by the model.

iii. False positive (FP): It the accuracy of the model to detect admitted students as not admitted students.

iv. False negative (FN): It represents the probability of not admitted students that are detected as admitted students. The confusion matrix is shown in Table no. 3.6.

Chapter 4

Results and Discussion

In this chapter, we will discuss the results of the research presented in this thesis. The process of testing and validation of the software system is also thoroughly presented in this section. A conclusion is also presented at the end of this chapter along with the future work. It is important to test the performance of the application through the methods adopted in previous research techniques. The application was tested through two different perspectives. Firstly, the performance was measured in terms of framerate on edge devices. Secondly, the performance of the application was measured in terms of accuracy. In this regard, it was required to test the performance of the application on the datasets of drivers.

4.1 Data Collection

Data is collected from the information and communication technologies (ICT) directorate of the NUST. The aim of this research study is to check the practical significance of current admission process of NUST at postgraduate level. Data has been collected by information and communication technology directorate of NUST to maintain the record of university. The span of the data is from 2008 to 2014 which is overall seven years. Data consists of 13094 students with five columns including YEAR, GAT, ACAD, INT and STATUS. Before 2008 the data is not available in electronic form.

4.2 Correlation Analysis

To measure the relationship between these three variables a bivariate correlation analysis is used. The analyzed variables are Academics represented as ACAD, graduate assessment test represented as GAT, and Interview represented as INT. The correlation coefficient range is from -1 to +1. The perfect positive relationship value is +1 and the perfect negative relationship value is -1. If the coefficient of correlation is 0, then there is no relationship between two variables. Table 4.2.1 shows the results of correlation analysis.

Table 5 – Correlation Analysis Measurement Scale

Variables Comparison	Correlation coefficient	Significance level	Effect
ACAD Vs GAT	0.125	0.01	Small
ACAD Vs INT	0.203	0.01	Small
INT Vs GAT	0.189	0.00	Small

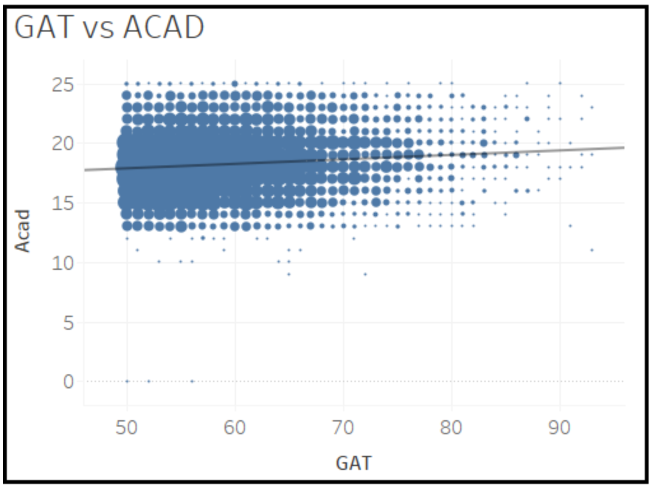


Figure 7: Correlation Analysis GAT vs ACAD

In the view of the above correlation analysis the relationship between academics (ACAD) and graduate assessment test (GAT) are given below:

- ACAD and GAT have a statistically significant linear relationship (r = 0.125, p < 0.01).
- The direction of the relationship is positive (i.e., ACAD and GAT are positively correlated), meaning that these variables tend to increase together
- High Score in ACAD is associated with high score in GAT).
- The magnitude, or strength, of the association is approximately small (0.1 < | r | < 0.3).

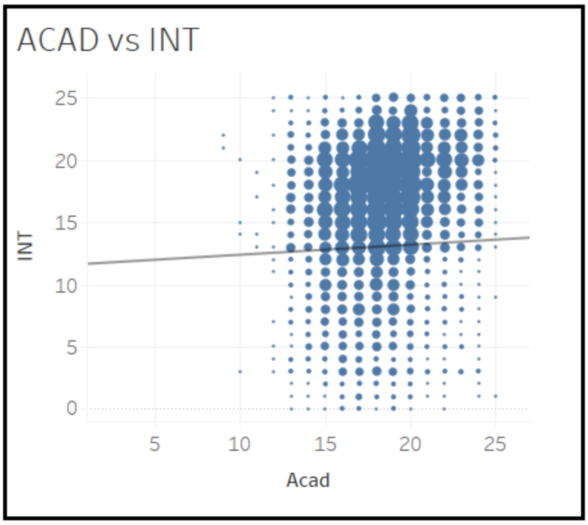


Figure 8: Correlation Analysis ACAD vs INT

Second correlation analysis reveals that the relationship between interview (INT) and academics (ACAD) are as follows:

- INT and ACAD have a statistically significant linear relationship (r =0.203, p < 0.01).
- The direction of the relationship is positive (i.e., ACAD and INT are positively correlated), meaning that these variables tend to increase together
- High Score in ACAD is associated with high score in INT).
- The magnitude, or strength, of the association is small (0.1 < | r | < 0.3).

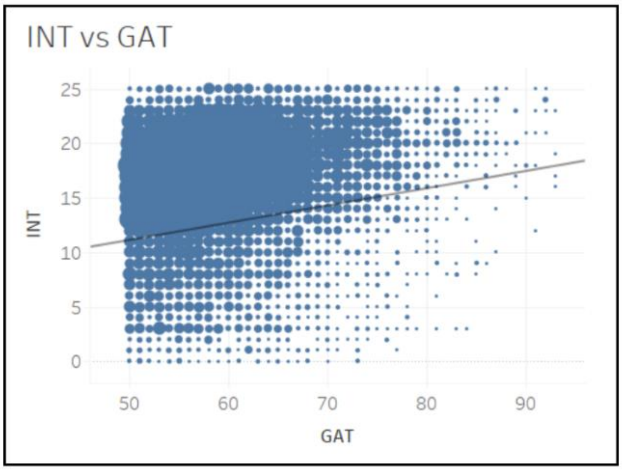


Figure 9: Correlation Analysis INT vs GAT

Third correlation analysis figure no: 4.2.3 results reveals that the relationship between interview (INT) and graduate assessment test (GAT) are as follows:

- INT and GAT have a statistically significant linear relationship (r =0.189, p < 0.01).
- The direction of the relationship is positive (i.e., GAT and INT are positively correlated), meaning that these variables tend to increase together.
- High Score in GAT is associated with high score in INT.
- The magnitude, or strength, of the association is approximately small (0.1 < | r | < 0 .3).

Table 6: Heat map of Pearson correlation Matrix

Correlations Matrix			
	INT	GAT	ACAD
INT	1	.189	.203
GAT	.189	1	.122
ACAD	.203	.122	1

4.3 Practical Significance (Cohen's D Analysis)

4.3.1 Effect Size Analysis

To check the practical significancy of these three variables an effect size analysis (Cohan's d Method) is used. The analyzed variables are Academics represented as ACAD, graduate assessment test represented as GAT, and Interview represented as INT. The measure of Cohen (1988, 1992) can be used as a guideline when measuring the correlation coefficient. Table: 4.3.1 contain the results of Cohan's d analysis.

Table 7: Results of Cohan's d analysis

	Admitted Student			Not admitted Student		
Variable Name	**Sample size**	**Mean**	**Standard Deviation**	**Sample size**	**Mean**	**Standard Deviation**
INT	5458	17.693	3.944	7636	14.185	5.834
GAT		60.798	7.873		59.187	7.33
ACAD		18.341	2.736		17.819	2.589

Admitted vs Not admitted	Cohan's *d* Value
Interview	0.704
Graduate assessment test	0.211
Academics	0.195

Cohan's *d* analysis used to find the comparison between two groups. Group of admitted students and group of not-admitted students. IBM - SPSS tool is used to find the results of mean, standard deviation of these two groups and to find the results of Cohan's *d* analysis used the online Cohan's calculator [12]. Cohan's d analysis (Table no 4.2) reveals the practical significance results which are discussed as follows.

Interview:

- The value of mean for interview is 17.69 and standard deviation is 3.94 for admitted students' group.
- The value of mean for interview is 14.18 and standard deviation is 5.83 for not-admitted students' group.

- The results of Interview (INT) between admitted students and not-admitted students' group have a practically significant result and effect of the association is large with (*d* = 0.704).

Graduate Assessment Test:

- The value of mean for graduate assessment test is 60.79 and standard deviation is 7.87 for admitted students' group.
- The value of mean for graduate assessment test is 59.18 and standard deviation is 7.33 for not-admitted students' group.
- The results of Graduate assessment test (GAT) between admitted students and not-admitted students' group have a practically significant result but effect of the association is small with (*d* = 0.211).

Academics:

- The value of mean for academics is 18.34 and standard deviation is 2.73 for admitted students' group.
- The value of mean for academics is 17.81 and standard deviation is 2.58 for not-admitted students' group.
- The results of Academics (ACAD) between admitted students and not-admitted students' group have a practically significant result but effect of the association is small with (*d* = 0.195).

4.4 Dimension Reduction

3.4.1 Principal Component Analysis

In PCA, 13094 applicants with three independent variables (INT, GAT and ACAD) and one dependent variable (Status of the student) were analyzed by using a statistical tool SPSS. The principal component that are practically significant are determined by their eigen values. The eigenvalues and eigenvectors of the correlation matrix were calculated. The eigenvectors are onstructed from linear combinations of original attributes in the data set. The results of principal component's eigen values are given below table no 4.4.1.

Table 8: Results of PCA eigen values and variance of eigen values

Total Variance Explained			
Principal Component	**Initial Eigenvalues**		
	Eigen Value	**% of Variance**	**Cumulative %**
Principal Component 1	1.344	44.792	44.792
Principal Component 2	.879	29.286	74.078
Principal Component 3	.778	25.922	100.000
Extraction Method: Principal Component Analysis.			

The eigen value of principal component one is 1.344 , Eigen value of second principal component is 0.879 and the eigen value of third principal component is 0.778. The eigen value of first principal component is greater the 1. According to Kaiser Criterion the eigen value greater than 1 should be placed for further analysis. The variance of the first principal component is 44.79% and the variance of second principal component is 29.28% and the variance of third principal component is 25.92%. The cumulative variance of first two component is 74%, which shows that more information can be extracted from these two principal components.

The coefficients of each original attribute give the index of agreement or disagreement in the original attributes towards the new dimension (principal component). The calculated principal components are given below in table no 4.4.2:

Table 9: Extracted principal component values

Variable	Principal Component 1	Principal Component 2	Principal Component 3
Interview_marks	0.625	-0.056	0.779
Gat_marks	0.542	0.749	-0.381
Acad_marks	0.562	-0.661	-0.498

The number of principal components that have practical significance are determined using eigenvalues. The principal components are constructed from linear combinations of original variables in the data set. The coefficients of each original variable give the index of agreement or disagreement towards the new dimension (principal component). Finally, a loading plot where, selected eigenvalues (PCs) and corresponding eigenvectors are plotted, is used to visualize the variation of original attributes on the selected PCs and a score plot where, selected PCs and samples that were transformed into these PCs are plotted to identify possible grouping and outliers in the sample set. The scree plot diagram is given below:

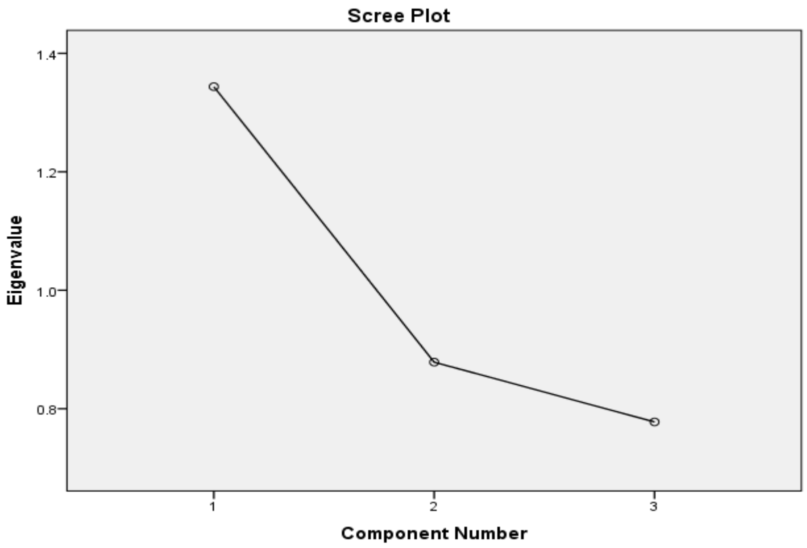


Figure 10: Scree Plot diagram of Eigen Value and principal Components

In view of the above scree plot diagram the eigen values are on y-axis and the principal component which is called eigen vector on x-axis. These results shows that the first two eigen values have the highest cumulative variance which will be used for further analysis.

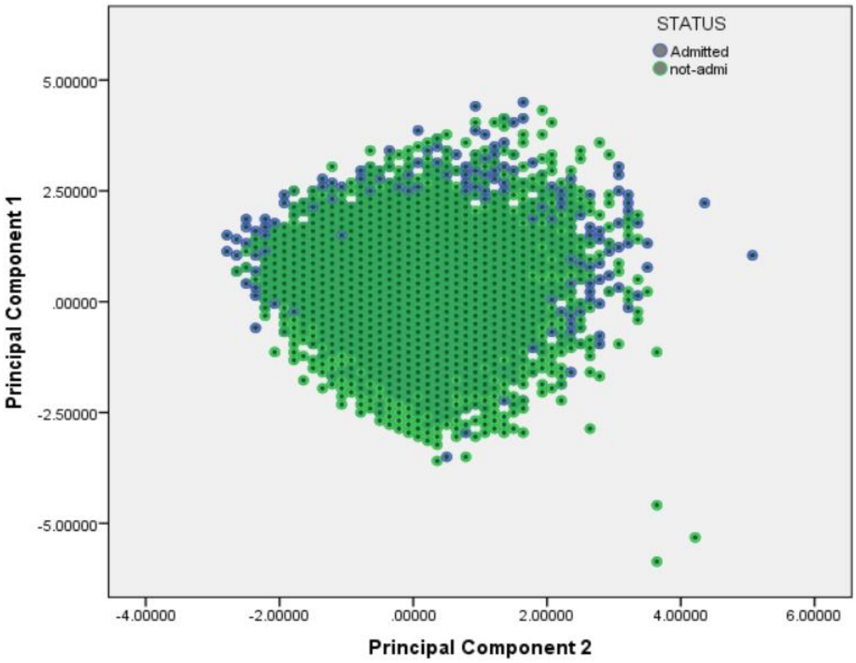


Figure 11: Scatter Plot between Principal Component 1 and Principal component 2

4.4.2 Weightages Distribution

As mentioned in the methodology chapter, the formulas for the weightage distribution and the calculated value of these formulas are defined below:

1. Weightage Distribution (INT):

$$= \frac{Principal\ Component\ (INT)}{Principal\ Component\ (INT) + Principal\ Component\ (ACAD) + Principal\ Component\ (GAT)} * 100$$

$$= \frac{0.625}{0.625 + 0.542 + 0.562} * 100$$

INT = 36.15 %

2. Weightage Distribution (GAT):

$$= \frac{Principal\ Component\ (GAT)}{Principal\ Component\ (INT) + Principal\ Component\ (ACAD) + Principal\ Component\ (GAT)} * 100$$

$$= \frac{0.542}{0.625\ +0.542+0.562} * 100$$

$$\text{GAT} \ = 31.35\ \%$$

3. Weightage Distribution (ACAD):

$$= \frac{Principal\ Component\ (ACAD)}{Principal\ Component\ (INT) + Principal\ Component\ (ACAD) + Principal\ Component\ (GAT)} * 100$$

$$= \frac{0.562}{0.625\ +0.542+0.562} * 100$$

$$\text{ACAD} \ = \ 32.50\ \%$$

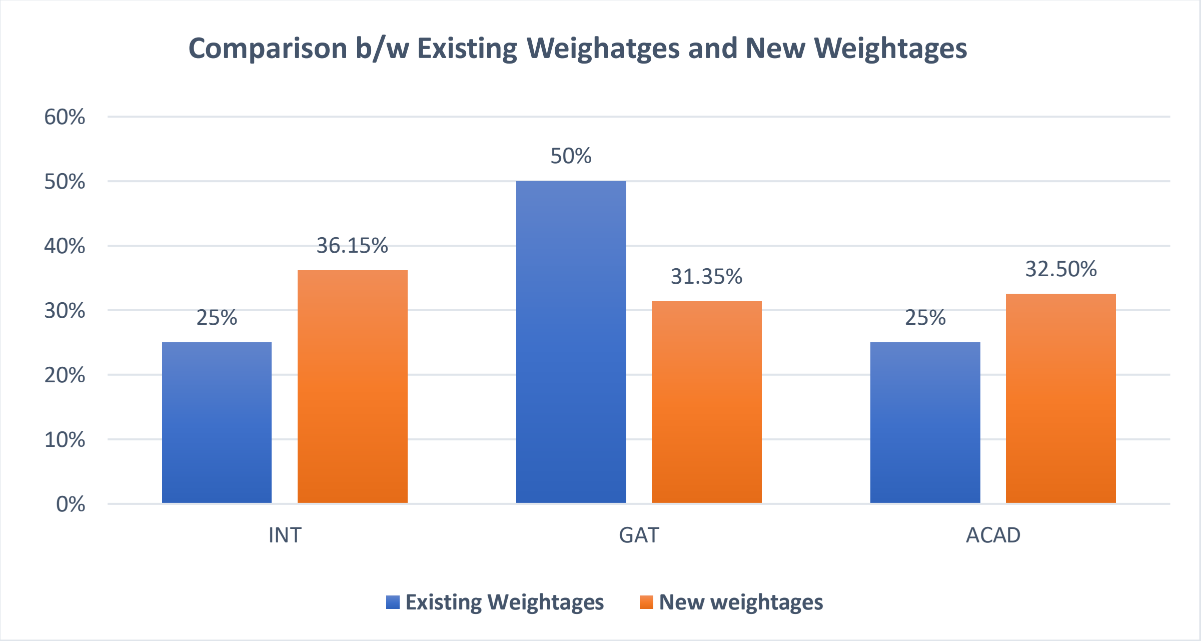


Figure 12: Comparison of existing weightage and new weightages

As per the results comparison of the existing and new weightages of different variables used in admission policy. The new suggested weightage for interview is 26.15% while the existing weightage of interview in existing admission policy is 25%. Same as the new weightage of graduate assessment test is 31.35% which is suggested by using principal component analysis, while the existing weightage of graduate assessment test in the existing admission policy is 50%. The new suggested weightage for academics is 32.50%, while the existing weightage of academic variable in the existing admission policy is 25%.

4.5 Predictive Modeling using Machine Learning Methods

In this research a most popular machine learning technique is used for machine learning model which is artificial neural networks. In artificial neural networks, the main wo techniques, radial based Function, and multi-layer perceptron are used for predictive modeling. Beside this Decision trees is also used for predictive modeling to check the comparison of these all-machine learning models.

4.5.1 Artificial Neural Network (ANN)

The two main popular techniques of artificial neural network, Radial based function, and multi-layer perceptron with on hidden layer have been used. For hidden layer activation multi-layer perceptron used the hyperbolic Tangent and for the activation of output layer MLP used the SoftMax. However, for Radial Based Function neural network, SoftMax is used for hidden layer and identity is used for the output layer. Details and results of these two modeling techniques are given below:

Table 10: Artificial Neural Network (ANN) results

	Training			
Methods	**Sensitivity**	**Specificity**	**Precision**	**Accuracy**
Multi-Layer Perceptron	54.77%	70.08%	63.29%	62.65%
Radial Based Function	54.46%	68.61%	59.51%	62.12%

	Testing			
Methods	**Sensitivity**	**Specificity**	**Precision**	**Accuracy**
Multi-Layer Perceptron	55.28%	71.27%	62.35%	63.87%
Radial Based Function	55.45%	70.83%	60.93%	63.90%

The MLP and RBF neural networks for binary categorical variables are shown in Table No 4.5.1. In this classification of applicant's data, we denoted the applicant which are not admitted as 0 and admitted as 1. The important step of modeling to perform this analysis the data into training and testing sets. As mentioned in data collection we have 13094 samples of applicants to deal with, therefore a ratio of 30% and 70% has been used for training and testing of the model. In view of the evaluation matrix in percentage based on assessment measure described in section no 3.4 of chapter no 3 and the performance of both techniques are approximately same.

If we see the results of both techniques, the results of RBF yield overall accuracy of 62.12% for training and 63.90% for testing which is approximately same with the results of MLP which is 62.12% for training and 63.87% for testing. For training dataset, the results of RBF yield specificity of 70.08% while the results of MLP is 68.61% for training i.e. the RBF measure correctly the proportion of applicants who are not admitted for training dataset. While for the testing dataset, MLP perform better then RBF with specificity of RBF is 70.83% and MLP is 71.27%. Now, if we see the Sensitivity of both techniques, for training dataset RBF has the same performance with value 54.46% while the MLP have 54.77%. As well for testing dataset they both have approximately same performance with RBF value 55.45% and MLP value 55.28%. The precision value of RBF for training dataset is 59.51% and MLP has the value 63.29%. While for testing dataset the value of RBF is 60.93% and MLP have the value 62.35%.

The independent variable importance with quantified significance in the development of prediction model is given in the table 4.5.2 and 4.5.3. The result shows that RBF independent variables importance value is more than the MLP values. The important point in both techniques is that interview has been taken as the most important variable then academics and graduate assessment test.

Table 11: Independent Variable Importance RBF

Independent Variable Importance RBF

	Importance	Normalized Importance
INT	.472	100.0%
GAT	.173	36.6%
ACAD	.355	75.2%

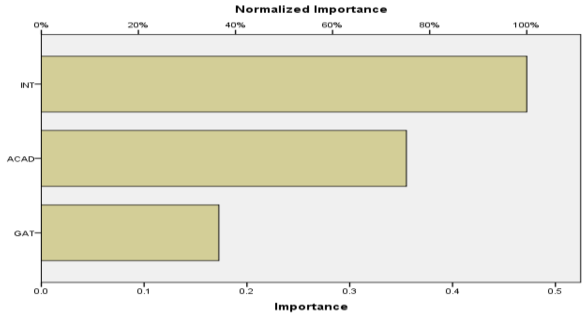
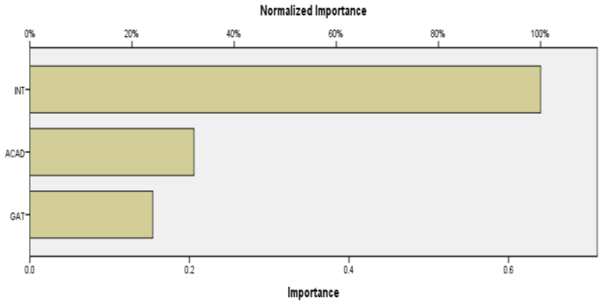


Table 12: Independent Variable Importance MLP

Independent Variable Importance MLP

	Importance	Normalized Importance
INT	.640	100%
GAT	.154	24.1%
ACAD	.206	32.1%



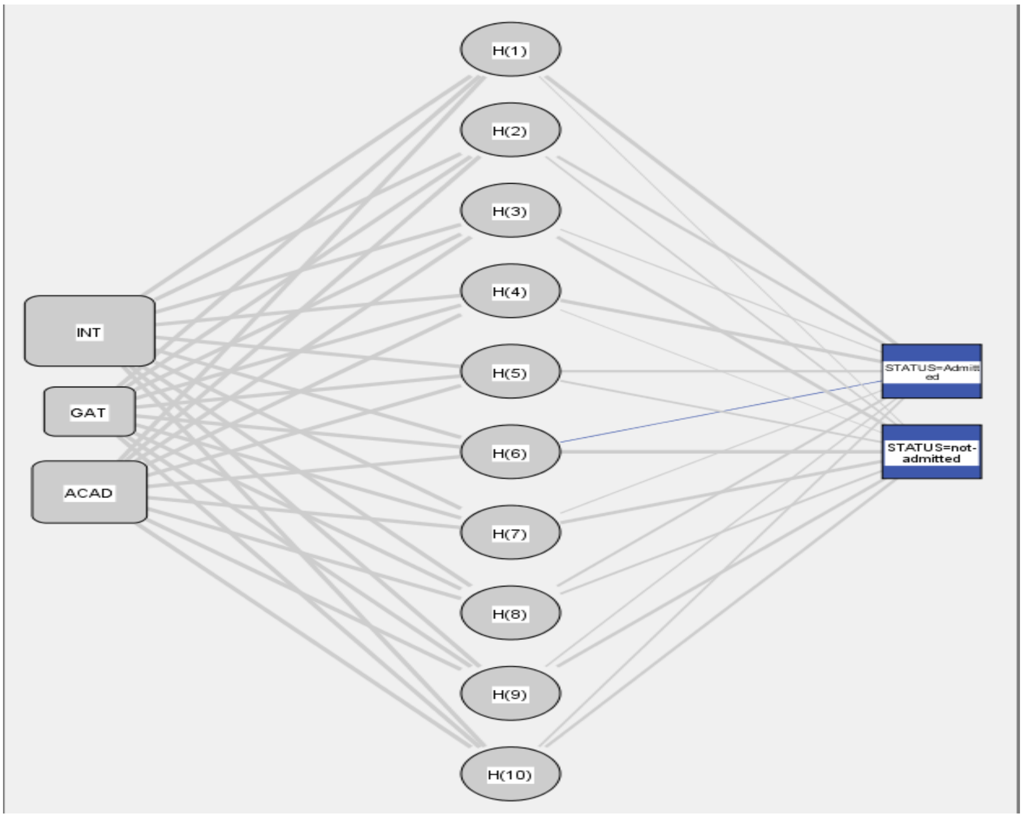


Figure 13: Radial Based Function. Hidden layer activation is Hyperbolic SoftMax, and output layer activation function is Identity

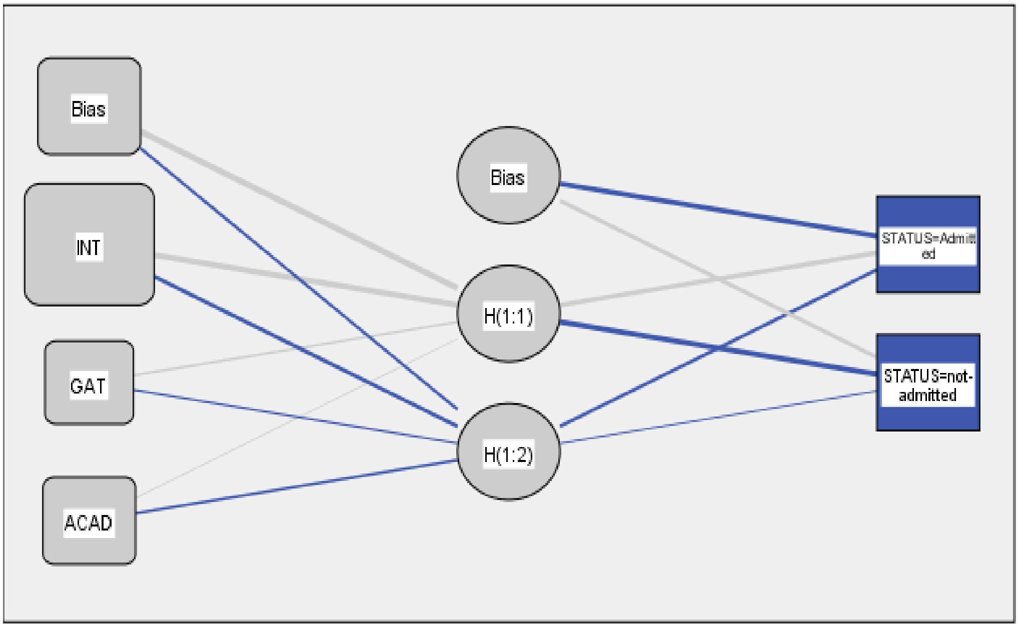


Figure 14: Multi-Layer Perceptron. Hidden layer activation is Hyperbolic tangent, and output layer activation function is SoftMax

4.5.2 Decision Tree (DT)

The decision tree classification model is built by using decision tree algorithm in special package for social sciences (SPSS). The growing method used for modeling decision tree is classification and regression trees (CRT) and Chi-Squared Automatic Interaction Detection (CHAID). CRT basically splits the data into classification and regression which are as homogenous as possible with respect to the dependent variables. There is also homogeneous pure node which have same value of for dependent variable in terminal nodes. While the CHAID is unsupervised technique which use entire model to build the tree. The results of total node and terminal node are given in Figure No: 4.5.2.1 and 4.5.2.2, which shows that the depth value of the decision tree in CRT technique are 5 while the depth of decision tree in CHAID technique are 3. The total nodes of the decision tree in CRT technique are 23 while the number of nodes of decision tree in CHAID technique are 28. In CRT technique the terminal node are 12 while in CHAID technique the terminal nodes are 18.

Table 13: Results of Decision Tree

Results of Decision tress				
Methods	**Sensitivity**	**Specificity**	**Precision**	**Accuracy**
Decision Trees (CRT)	54.37%	74.52%	73.16%	63.22%
Decision Trees (CHAID)	54.41%	72.88%	69.66%	63.02%

The results of decision tree are shown in Table No: 12. In the classification of applicant's data we denoted the applicant which are not admitted as 0 and admitted as 1. The performance of both techniques are results of CHAID which is 63.02%. The results of CRT yield specificity of 74.52% while the results of CHAID is 72.88% which means the CRT perform better then

CHAID in the results of specificity, i.e. the CRT technique measure correctly the proportion of applicants who are not admitted. While in results of sensitivity, the performance of both techniques is approximately same. The specificity of CRT is 54.37% and CHAID is 54.41%. The precision value of CRT for is 73.16% and CHAID has the value of 69.66%.

Table 14: Model Summary Results for CRT

RESULTS OF MODEL SUMMARY CRT		
Results	Independent Variables Included	INT, ACAD, GAT
	Number of Nodes	23
	Number of Terminal Nodes	12
	Depth	5

Table 15: Model Summary Results for CHAID

RESULTS OF MODEL SUMMARY CHAID		
Results	Independent Variables Included	INT, GAT, ACAD
	Number of Nodes	28
	Number of Terminal Nodes	18
	Depth	3

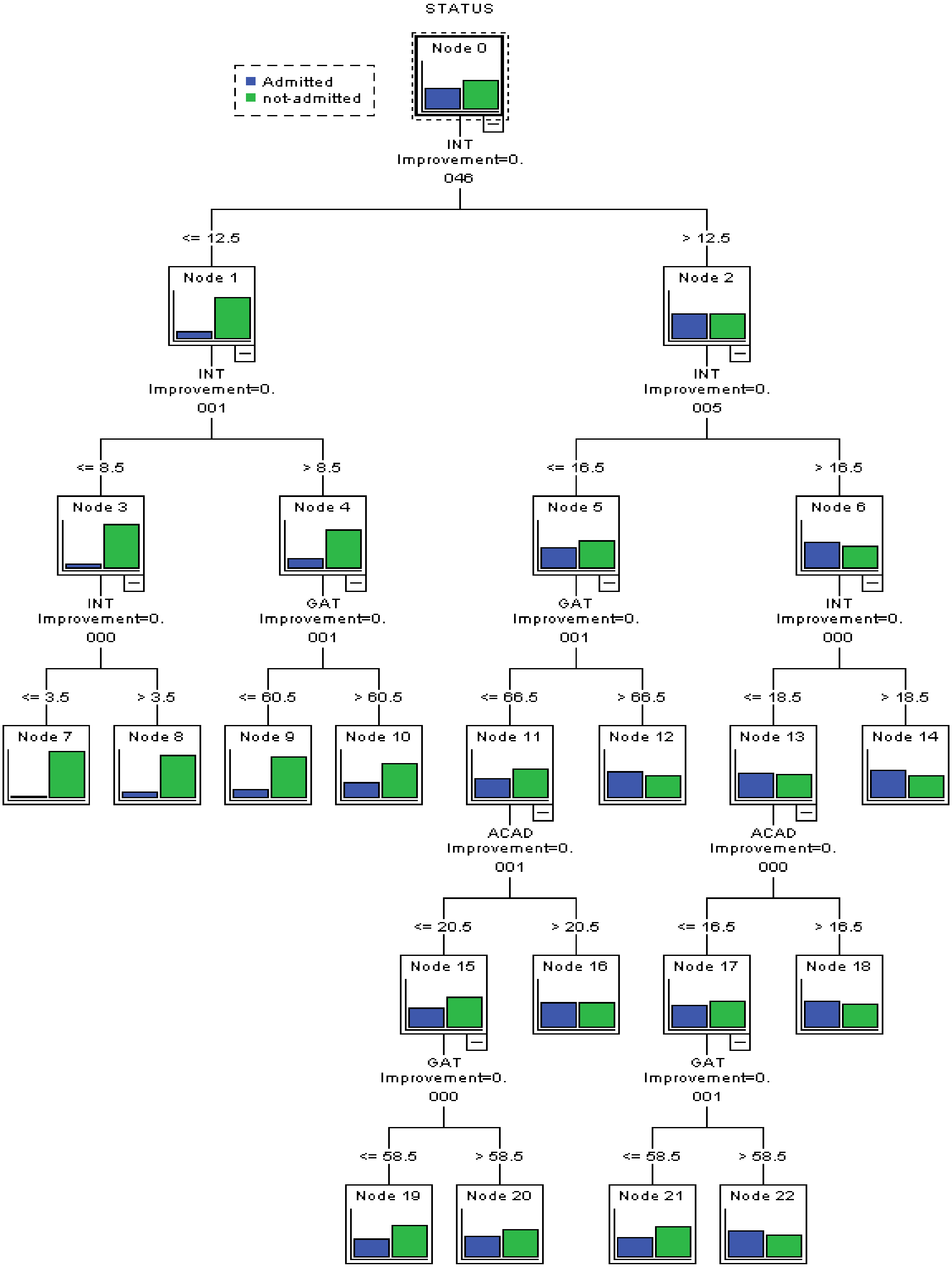


Figure 15: Diagram of Decision Tree for Classification and regression of tree (CRT)

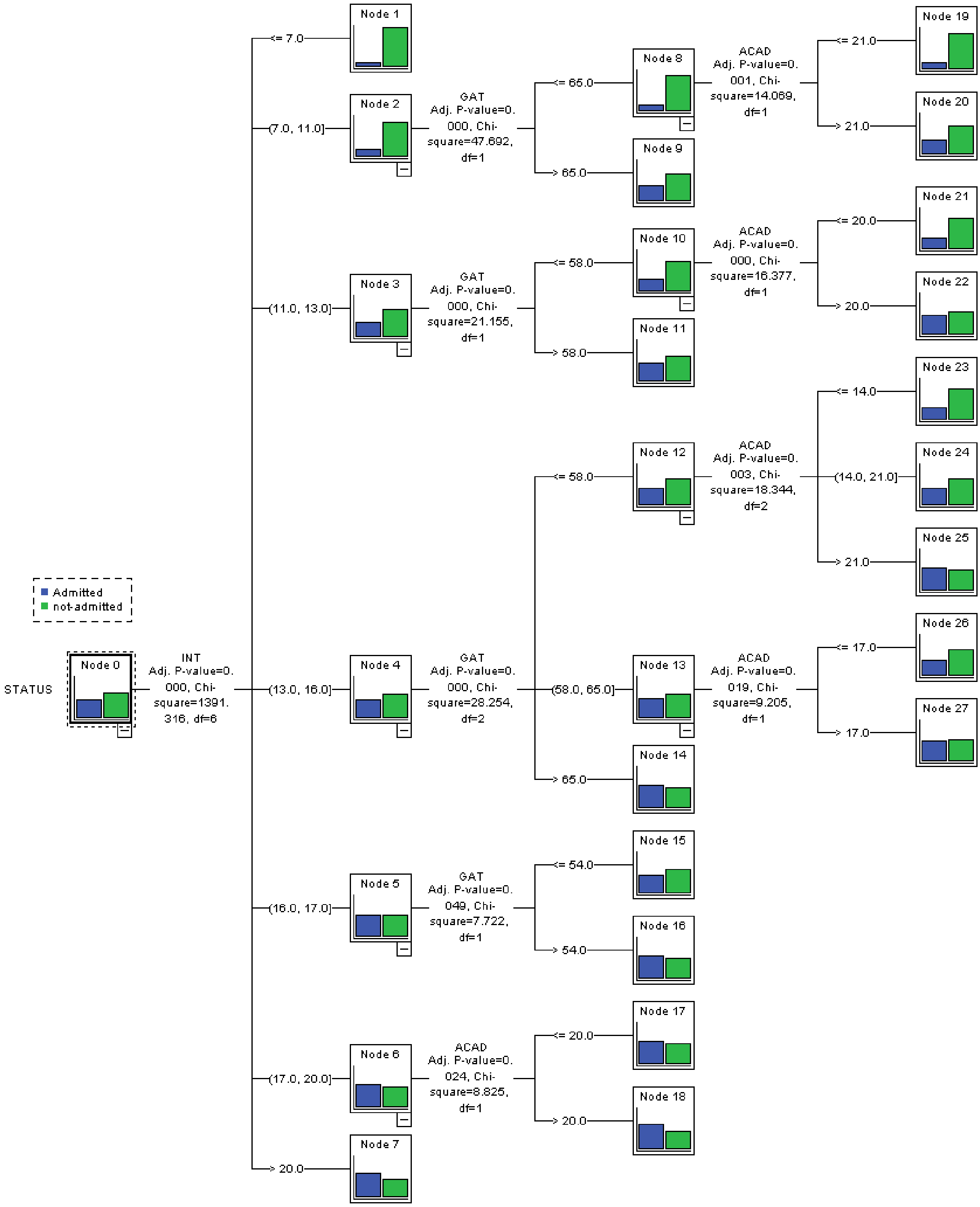

Admitted
not-admitted
STATUS
Node 0
INT Adj. P-value=0.000, Chi-square=1391.316, df=6
<= 7.0
Node 1
(7.0, 11.0]
Node 2
GAT Adj. P-value=0.000, Chi-square=47.692, df=1
<= 65.0
Node 8
ACAD Adj. P-value=0.001, Chi-square=14.069, df=1
<= 21.0
Node 19
> 21.0
Node 20
> 65.0
Node 9
(11.0, 13.0]
Node 3
GAT Adj. P-value=0.000, Chi-square=21.155, df=1
<= 58.0
Node 10
ACAD Adj. P-value=0.000, Chi-square=16.377, df=1
<= 20.0
Node 21
> 20.0
Node 22
> 58.0
Node 11
(13.0, 16.0]
Node 4
GAT Adj. P-value=0.000, Chi-square=28.254, df=2
<= 58.0
Node 12
ACAD Adj. P-value=0.003, Chi-square=18.344, df=2
<= 14.0
Node 23
(14.0, 21.0]
Node 24
> 21.0
Node 25
(58.0, 65.0]
Node 13
ACAD Adj. P-value=0.019, Chi-square=9.205, df=1
<= 17.0
Node 26
> 17.0
Node 27
> 65.0
Node 14
(16.0, 17.0]
Node 5
GAT Adj. P-value=0.049, Chi-square=7.722, df=1
<= 54.0
Node 15
> 54.0
Node 16
(17.0, 20.0]
Node 6
ACAD Adj. P-value=0.024, Chi-square=8.825, df=1
<= 20.0
Node 17
> 20.0
Node 18
> 20.0
Node 7

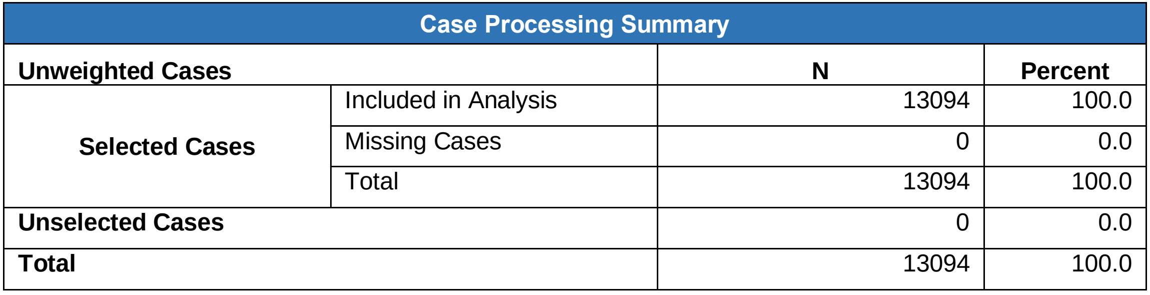
Figure 16: Decision Tree for Chi-Squared Automatic Interaction Detection (CHAID)

4.5.3 Logistic Regression

There are two independent continues variables and one dependent categorical variable. The two independent variables have been derived by using principal component analysis technique. The first two principal components are used in logistics regression because the cumulative variance of these two components is 74 %. This principal component is denoted by PC1 and PC2. The third variable is categorical which is denoted by status. Basically, there are two categories in status variable, admitted and not admitted students. The admitted student is represented as 0 and not admitted as 1 by default in SPSS. Before diving into binary logistic regression there is a need to check all the pre-requisites, then move forward with SPSS(version 20) for the analysis and development of binary logistic regression. The model is developed to predict the status of the student based on provided variables from principal component.

These all independent and dependent variable were added respectively. It can be seen from Table no: 4.5.3.1, that there are no missing values, total number of values used in binary logistic regression is 13094.

Table 16: Case Processing Summary Results for BLR



Case Processing Summary			
Unweighted Cases		**N**	**Percent**
Selected Cases	Included in Analysis	13094	100.0
	Missing Cases	0	0.0
	Total	13094	100.0
Unselected Cases		0	0.0
Total		13094	100.0

The SPSS automatically encode the status of student 0 for admitted and 1 for not admitted. The null model output has no explanatory variables which is Block 0. and then the model with all the explanatory variables (Block 1). Table 4.5.3.2 shows us all the explanatory variables (Block 1), which actually tells us about the model prediction. The overall percentage accuracy of block 0 model is 58.3%. However, the Table 4.5.3.3 shows us the null model. The constant (B) term in null model is significant with p-value = 0.000 < 0.01, because the size of the data is large.

Table 17: Classification Table. Results of BLR for Step 0

Classification Table					
	Observed		Predicted		
			STATUS		Percentage Correct
			Admitted	not-admi	
Step 0	STATUS	Admitted	0	5458	0
		not-admi	0	7636	100
	Overall Percentage				58.3
a. Constant is included in the model.					
b. The cut value is .500					

Table 18: Variables in the Equation results for Step 0

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	0.336	0.018	358.894	1	0	1.399

After analyzing the null model, step forward towards the main model which includes all the explanatory variables. Comparison between these models on the basis of significance are shown in the table no: 4.5.3.4. In view of these results the model is significant because of the chi-square value and p-value (chi-square = 685.34, p-value < 0.000). The value of "step" and "block" are same because I added all the variables at once instead of using stepwise approach. The variance of this model in the outcome is 6.9% shown in table no: 4.5.3.5.

Table 19: Omnibus Tests of Model Coefficients

Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	685.34	2	.000
	Block	685.34	2	.000
	Model	685.34	2	.000

Table 20: Model Summary Results of BLR

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	17102.829[a]	0.051	0.069
a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.			

This model behaves accurately then the baseline model. The accuracy of this model in percentage is 60.6 % shown in table no: 4.5.3.5, there is increase of 2.30% from the null model's correct percentage which is shown in table no: 4.5.3.2.

Table 21: Classification Table. *Results of BLR for Step 1*

Classification Table					
	Observed		Predicted		
			STATUS		Percentage Correct
			Admitted	not-admi	
Step 1	STATUS	Admitted	1654	3804	30.3
		not-admi	1351	6285	82.3
	Overall Percentage				60.6
a. The cut value is 0.500					

After looking at the significancy of all models it is essential to look at each explanatory variables significancy. From table no: 4.5.3.7, principal component 1 (PC1) has high significant variable with value of (Wald = 557.657 and p-value < 0.000) and Principal component 2 (PC2) is the least significant with value of (Wald = 96.13, p-value < 0.00). Coefficients (B) for principal component 1 is significant and negative which indicates that 1 unit change in the principal component 1 will decrease the odds for falling into the not-admitted group. The odd ratio of the corresponding principal component variables are shown in Exp (B) column. With the increase of 1 mark in principal component 1 there are 38% less chance to be not admitted. Principal component 2 has their odds ration greater than 1 which show a negligible association between independent and dependent variable.

Table 22: Variables in the Equation results of BLR for Step 1

Variables in the Equation									
		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	PC1	-0.386	0.016	557.657	1	.000	0.68	0.659	0.702
	PC2	0.191	0.019	96.013	1	.000	1.21	1.165	1.257
	Constant	0.354	0.018	376.466	1	.000	1.425		
a. Variable(s) entered on step 1: PC1, PC2.									

For evaluation of binary logistic regression model to see the error trend and magnitude. I used SPSS and confusion matrix calculator to find the value of specificity , sensitivity, accuracy, and precision. The overall accuracy of BLR model is 60.63 % shown in table no: 4.5.3.8. The value of specificity 62.39% which means this model 62.39% accurately identified the proportion of the student which is not admitted. The value of Sensitivity id 55.04% which means this model 55.04% accurately identified the proportion of student which are admitted. The precision value for this model is 30.30% shown in table no 4.5.3.8.

Table 23: Performance Measure of evaluations for BLR

Results of Logistics Regression				
Methods	**Sensitivity**	**Specificity**	**Precision**	**Accuracy**
Logistics Regression (Principal Component)	55.04%	62.30%	30.30%	60.63%

Chapter 5

Summary

In this chapter we will conclude the results of previous chapters. In the 1st chapter we discuss the introduction of this thesis and introduce the existing admission policy of NUST for post graduate students. In 2nd chapter a literature review related to this topic is discussed which conclude that there are different variables which is under discussion related to admission policy and enrolment process. In 3rd chapter a stepwise methodology is used to analyze the existing post graduate admission policy of NUST. 4th chapter describes the different results, predictive ability and weightage distribution of different variables used in admission process. This chapter presents the summary, conclusions, and recommendations for future research.

5.1 Summary

Different case studies and research is conducted on admission policy or enrollment process, comparison of accurate prediction between different machine learning techniques. Most of them are interested in behavioral and psychological assessment of different applicants score and the relationship[p between these variable's scores. Most of the literature related to this study focus on different factors used in the enrollment process or on the predictive ability / power of those factors. are interested in comparison of different variables. In review of different literature, I didn't find the research which is focused on subjective vs objective. The uniqueness of this research study is that we focused on the analysis of employed policy at postgraduate level in NUST, Pakistan.

5.2 Conclusion and Findings

The objectives of this study were:

I. To analyze trends and tendencies of variables with respect to the performance of admitted and not admitted students using comparative analysis including correlation analysis and testing practical significance through effect size calculations, etc. Secondly, investigating various myths related to GRE and ACAD or ACAD and INT or GRE VS ACAD or GRE VS INT, etc. For instance, does interview marks effected by performance in GRE and/or ACAD?

II. Development of a predictive model for estimating chances of admission of an applicant by using machine learning methods like decision trees multi-level perceptron etc.

III. Performance evaluation of developed models in terms of precision and accuracy.

IV. Assess the adequacy of the current weightages of variables and propose new weightages, if required.

5.2.1 Correlation and practical significance analysis:

In the step of correlation analysis and practical significance analysis, all the variables were analyzed to investigate the different variables and the relationship between them. Correlation analysis is used to investigate the myths related to GAT, ACAD and INT. The correlation between ACAD and GAT is relatively small and positive. Which means high score in ACAD has a very small impact on the score of GAT. Same correlation value between the result of INT and GAT, which means there is also a very small impact. The correlation value between ACAD and INT is medium, which means high score in ACAD has a medium impact on the score of interview.

Effect size analysis is used to check the practical significance between the admitted students and not admitted students. The effect size analysis value of the interview (INT) variable is large then ACAD and GAT. Which means that interview is highly significant in the analysis of practical significance.

5.2.2 Principal Component analysis - Weightage distribution:

In this step we reduce the dimension of the variables and then we use these dimensioned reduced variables in the method of logistic regression to improve the accuracy. Secondly, we compute the eigenvalues and eigenvectors by using the SPSS tool. Further we use these eigenvalues and eigenvectors for the weightage distributions.

In the step of accuracy improvement there is only 1 % increase in accuracy and slightly better results than other machine learning methods. In the second step of weightage distribution there is more difference between the new and old weightages. The suggested new weightage for interview (INT) is 35%, Academics (ACAD) is 31.5% and Graduate assessment test (GAT) is

31.5%. While the old weightage for interview (INT) is 25%, Academics (ACAD) is 25% and Graduate assessment test (GAT) is 50%.

5.2.3 Predictive power of these variables:

Two main machine learning method were used to analyze the predictive power of these variables. Artificial neural network and Decision trees are the focused techniques in this case study. In artificial neural network, multi-layer perceptron and radial based function techniques were used. In which both techniques were not able to predict with more accuracy. Because the variables are not extensively defined. The accuracy of multi-layer perceptron is 62.65% While the accuracy of radial based function is 62.12%. In this phase, the interview (INT) has more importance in the prediction of applicant status. While the weightage of interview is less then graduate assessment test (GAT). The academics (ACAD) has the second most importance in the prediction of applicant status, while the weightage of ACAD is also less then graduate assessment test (GAT).

The second focused technique is decision tress. In this technique we also achieved the accuracy of 63.02% which is not suitable to predict the status of the applicant.

Currently we achieved 63% maximum possible accuracy. In view of these results, we achieved that these three variables, ACAD, GAT and INT are incomplete for predicting the merit of an applicant. These variables are not sufficient to provide the complete information of an applicant.

The conclusion of this case study is among all variables in the admission policy, interview is most important and influenced variable in term of creating difference concerning applicant status. While the weightage of interview is not consistent with subjective weightage assigned to all these three variables. Thus, the question arose in mind that why GAT has the highest weightage percentage then other two variables.

The conclusion of this case study was to answer the objectivity of NUST postgraduate admission policy. This research comes up with evidence collected form the applied methodologies. After analyzing the obtained results, its concluded that the subjective process didn’t support the objective analysis. Thus, the implemented admission policy is not complete in term of objectivity.

5.3 Future Work and Recommendations

I. To improve admission policy and accuracy improvement there should be more data related to an applicant, e.g., hostel facility, student address (residence of an applicant), financial stability status, degree, and program popularity, etc.

II. Formulation of an entry test system for the postgraduate students at NUST same as the entry test system for undergraduate students.

References

[1] Abedi, J. (1991). Predicting Graduate Success from Undergraduate Academic Performance: A Canonical Correlation Study. Educational and Psychological Measurement, 51, 151-160.

[2] Basturk, R. (1999, October 13-16). THE Relationship of Graduate Record Examination Aptitude Test Scores and Graduate School Performance of International Student at the United States Universities. Annual Conference of the Mid-Western Educational Research Association Chicago, Illinois, 15.

[3] Delgado-Marquez, B. L., Escudero-Torres, M. A., & Hurtado-Torres, N. (2013). Being highly internationalised strengthens your reputation: an empirical investigation of top higher education institutions. Springer Science+Business Media Dordrecht. doi:10.1007/s10734-013-9626-8

[4] Fisher, J. &. (1990). Standardized Testing and Graduate Business School Admission: A review of Issues and an Analysis of a Baruch College MBA Cohort. College and University, 55, 137-148.

[5] Flier, H. v., b, G. D., & Zaaiman, H. (2003). Selecting students for a South African mathematics and science foundation programme: the effectiveness and fairness of school-leaving examinations and aptitude tests. International Journal of Educational Development, 23, 399-409.

[6] Goldenberg, E. L., & Alliger, G. (1992). Assessing the Validity of the GRE for Students in Psychology: A Validity Generalization Approach. Educational and Psychological Measurement, 52, 1019-1027.

[7] Graham, L. (1991). Predicting Academic Success of Students in a Master of Business Administration Program. Educational and Psychological Measurement, 51, 721-727.

[8] HAMILTON, D. J. (1990). Multiple Regression Analysis and Predicting of GPA upon Degree. College Student Journal, 24, 91-96.

[9] Hurtado-Torres, N. E., Delgado-Marquez, B. L., & Escudero-Torres, M. A. (2013). Being highly internationalised strengthens your reputation: An empirical investigation of top higher education institutions. Higher Education, 16.

[10] William L, D., & Lynch, R. M. (1977). Graduate Admission Policy: A Bayesian Analysis. The Journal of Experimental Education, 45, 8-41.

[11] WILSON, K. M. (1986). The Relationship of GRE General Test Scores to First Year Grades for Foreign Graduate Students. Educational testing Service, (ETS). Princeton, NJ, 20.

[12] Cohan’s Calculator : https://www.socscistatistics.com/effectsize/default3.aspx

[13] Chadi, A., & de Pinto, M. (2017). Selecting successful students? Undergraduate grades as an admission criterion. Applied Economics, 50(28), 3089–3105. doi:10.1080/00036846.2017.1418072

[14] Zimmermann, J., von Davier, A., & Heinimann, H. R. (2017). Adaptive admissions process for effective and fair graduate admission. International Journal of Educational Management, 31(4), 540–558. doi:10.1108/ijem-06-2015-0080

Appendices

Appendix I

Scheme of allocation of Marks for the variable ACAD (out of 25)

CGPA (in terminal degree/transcript)	Or	Percentage (in terminal degree/transcript)	Marks Allotted
4.00		98.00 - 100.00	25.00
3.90 - 3.99		95.00 - 97.99	24.00
3.80 - 3.89		92.00 - 94.99	23.00
3.70 - 3.79		89.00 - 91.99	22.00
3.60 - 3.69		86.00 - 88.99	21.00
3.50 - 3.59		83.00 - 85.99	20.00
3.40 - 3.49		80.00 - 82.99	19.50
3.30 - 3.39		79.00 - 79.99	19.00
3.20 - 3.29		77.00 - 78.99	18.50
3.10 - 3.19		75.00 - 76.99	18.00
3.00 - 3.09		73.00 - 74.99	17.50
2.90 - 2.99		71.00 - 72.99	17.00
2.80 - 2.89		69.00 - 70.99	16.50
2.70 - 2.79		67.00 - 68.99	16.00
2.60 - 2.69		65.00 - 66.99	15.50
2.50 - 2.59		63.00 - 64.99	15.00
2.40 - 2.49		60.00 - 62.99	14.50
2.30 - 2.39		58.00 - 59.99	14.00
2.20 - 2.29		57.00 - 57.99	13.50
2.10 - 2.19		56.00 - 56.99	13.00
2.00 - 2.09		55.00 - 55.99	12.50
- - -		54.00 - 54.99	12.00
- - -		53.00 - 53.99	11.50
- - -		52.00 - 52.99	11.00
- - -		51.00 - 51.99	10.50
- - -		50.00 - 50.99	10.00
- - -		Less than 50	9.00

Note: Percentage will only be valid if cumulative grade point average (CGPA) is not mentioned in terminal degree/transcript.

Appendix II

MS Individual Interview Proforma

Roll No.____________________Name:_______________________________Degree
____

No.	Parameter	Total Marks Allocated	Total Marks Earned
1	**Personality**		
	a. Appearance	1	
	b. Mannerism	2	
	c. Emotional stability / Maturity	2	
	Sub-Total	**5**	
2	**Communication Skills**		
	a. Writing Skill (must ask candidate to write statement of purpose)	3	
	b. Fluency in expression	2	
	Sub-Total	**5**	
3	**Motivation / Zeal / Commitment**		
	a. Commitment to complete MS	2	
	b. Employbility after graduation or already employed	1	
	c. Potential for success	1	
	d. Leadership Qualities/ Teamworker	1	
	Sub-Total	**5**	
4	**Knowledge of Applied Discipline**		
	a. Degree of knowledge & expression of interest in the applied program	1	
	b. No of core / elective courses taken in UG, relevant to applied program.	1	
	c. Situational Awareness/ General knowledge	1	
	d. Ability to apply existing knowledge	1	
	e. Ability to apply new concepts	1	
	Sub-Total	**5**	
5	**Research Aptitude**		
	a. Research experience / publication	2	
	b. Rapidity in thinking & reasoning	1	
	c. Commitment to intense learning	1	
	d. UG project/ MS thesis	1	
	Sub-Total	**5**	
TOTAL SCORE		**25**	

Remarks (if any):
____
___ **Signature:_____________ Name:___________**
Date:________________