"University Classification and Prediction Using Data Mining"

Project Report

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Engineering

by

Mohammed Shadir Fakir Mohammed Fatima (12CO44)

Ashiq Ahamed Abdul Subhan RehamanBee (12CO23)

Baig Mannan Rehman Sabira (12CO24)

Shaikh Abdul Alim JamilurRehman Sufiyana (12CO57)

Supervisor

(Prof.Javed Sheikh)

Co-Supervisor

(Prof.Muhammed Salman Shamsi)

Department of Computer Engineering,
School of Engineering and Technology
Anjuman-I-Islam's Kalsekar Technical Campus
Plot No. 2 3, Sector -16, Near Thana Naka, Khanda Gaon,
New Panvel, Navi Mumbai. 410206
Academic Year : 2015-2016

CERTIFICATE

Department of Computer Engineering,

School of Engineering and Technology,
Anjuman-I-Islam's Kalsekar Technical Campus
Khanda Gaon,New Panvel, Navi Mumbai. 410206

This is to certify that the project entitled ***University Classification and Prediction using Data Mining*** is a bonafide work of **Mohammed Shadir Fakir Mohammed Fatima**(12CO44),**Ashiq Ahamed Abdul Subhan Rehaman Bee**(12CO23),**Baig Mannan Rehman Sabira** (12CO24),**Shaikh Abdul Alim Jamilur Rehman Sufiyana**(12CO57) submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **Bachelor of Engineering** in **Department of Computer Engineering**.

Prof.Javed Sheikh — Supervisor/Guide

Prof.Muhammed Salman Shamsi — Co-Supervisor/Guide

Prof. Tabrez Khan — Head of Department

Dr. Abdul Razak Honnutagi — Director

Project Approval for Bachelor of Engineering

This project entitled ***University Classification and Prediction using Data Mining*** by ***Mohammmed Shadir Fakir Mohammed , Ashiq Ahamed Abdul Subhan, Baig Mannan, Shaikh Abdul Alim JamailurRehman*** is approved for the degree of ***Bachelor of Engineering in Department of Computer Engineering.***

Examiners

1. .....
2. .....

Supervisors

1. .....
2. .....

Chairman

.....

Declaration

We declare that this written submission represents our ideas in our own words and where others ideas or words have been included, We have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Mohammed Shadir Fakir Mohammed Fatima (12CO44)

Ashiq Ahamed Abdul Subhan RehamanBee(12CO23)

Baig Mannan Rehman Sabira(12CO24)

Shaikh Abdul Alim JamilurRehman Sufiyana(12CO57)

Abstract

Tittle: University Classification and Prediction using Data Mining

From the medieval age university quality has been widely known by writings of students, greats names of professors and academic reputation of universities. But since the last quarter of the twentieth century, university ranking systems have widely observed international universities and updated world universities ranking annually.

During recent years, university rankings have gained a considerable importance not only among the academia but also amongst students, parents, industry and businesses. Common stakeholders, the students and their parents, may not be aware of the intricacies of ranking processes and elements / criteria of rankings but they are definitely keen to know the position of the University of their Interest in the ranking lists. This paper will review the trend and existing approaches of the most common and popular university ranking systems and evaluations and describe various Quantitative / Qualitative criteria used to determine the rankings.

The process involves various surveys besides using statistics and rankings are conducted on national, regional and global levels for institutions,departments, schools or specific academic programs. It is opined that although university rankings are considered inherently controversial for Not being absolutely objective and definitive, they are still used as reference to assist in making certain crucial decisions

Mohammed Shadir Fakir Mohammed Fatima(12CO44)

Ashiq Ahamed Abdul Subhan RehamanBee(12CO23)

Baig Mannan Rehman Sabira(12CO24)

Shaikh Abdul Alim JamilurRehman Sufiyana(12CO57)

B.E. (Computer Engineering)
University of Mumbai.

Contents

List of Figures

Keywords And Glossary

Keywords :

University Ranking System, Data Analytics, Statistics Bi-dimensional Ranking, Perfomance Indicators Naive Baye's Classifier, Webometrics, Probability Factor(pf).

Glossary :

Accredited :Official recognition that a college or university meets the standards of a regional or national association. Although international students are not required to attend an accredited college or university in the United States, employers, other schools, and governments worldwide often only recognize degrees from accredited schools.

Assistantship: A financial aid award granted to a graduate student to help pay for tuition that is offered in return for certain services, such as serving as a teaching assistant or research assistant.

Course: A regularly scheduled class on a particular subject. Each college or university offers degree programs that consist of a specific number of required and elective courses.

Academic adviser:A member of a school's faculty who provides advice and guidance to students on academic matters, such as course selections.

Academic year:Annual period during which a student attends and receives formal instruction at a college or university, typically from August or September to May or June. The academic year may be divided into semesters, trimesters, quarters, or other calendars

Bachelor's:An undergraduate degree awarded by a college or university upon successful completion of a program of study, typically requiring at least four years (or the equivalent) of full-time study. Common degree types include bachelor of arts (B.A. or A.B.), which refers to the liberal arts, and bachelor of science (B.S.). A bachelor's is required before starting graduate studies.

Campus: The grounds and buildings where a college or university is located

College: A postsecondary institution that typically provides only an undergraduate education, but in some cases, also graduate degrees. "College" is often used interchangeably with "university" and "school." Separately, "college" can refer to an academic division of a university, such as College of Business. (See U.S. News's rankings of Best Colleges.

Curriculum:A program of study made up of a set of courses offered by a school.

Enroll:To register or enter a school or course as a participant.

Fees:An amount of money charged by colleges and universities, in addition to their tuition, to cover costs of services such as libraries and computer technology.

Graduate job:The term âœgraduate jobâ may refer to any job that requires a degree, or more specifically to a job that is part of a companyâ^{TM}s graduate scheme â" a formal training scheme, typically lasting one or two years.

GRE (Graduate Record Examination):A standardized graduate school entrance exam administered by the nonprofit Educational Testing Service (ETS), which measures verbal, quantitative, and analytical writing skills. The exam is generally required by graduate schools, which use it to assess applicants of master's and Ph.D. programs. Some business schools accept either the GMAT or GRE; law schools generally require the LSAT; and medical schools typically require the MCAT. Effective August 2011, the GRE will incorporate key changes in the content, length, and style of the exam. (See the U.S. News GRE guide for more information

Letter of recommendation:A letter written by a student's teacher, counselor, coach, or mentor that assesses his or her qualifications and skills. Colleges, universities, and graduate schools generally require recommendation letters as part of the application process

Ranking and sorting::When we search a database, our results are displayed in some order. This is called ranking or sorting. The most common orders we see are most recent first and relevance.

Scholarship:A type of financial aid that consists of an amount of free money given to a student by a school, individual, organization, company, charity, or federal or state government. "Scholarship" is often used interchangeably with "grant."

University:A postsecondary institution that typically offers both undergraduate and graduate degree programs. "University" is often used interchangeably with "college" and "school."

Work-study:A financial aid program funded by the U.S. federal government that allows undergraduate or graduate students to work part time on campus or with approved off-campus employers. To participate in work-study, students must complete the FAFSA. In general, international students are not eligible for work-study positions.

Chapter 1

Project Overview

1.1 Statement of Project

A qualitative evaluation shall be utilized for this research project leveraging subjective methods such as interviews and observations to collect substantive and relevant data. These interviews shall be conducted with practicing diplomats from any University as well as visiting diplomats to another universities conferences.Such a qualitative approach is valuable here due to the varying experiences of the diplomats in any country.

As per our own research efforts, We will have the opportunity to implement by intercultural education expertise and develop a ground-breaking and full pledged university ranking system.our software will rank the universities according to the prescribed qualification of any student for choosing the best university for him/her. Since in today generation, its not only important to decide what to study..what degree to pursue but also where to study. A minor degree from a reputed outstanding university can be more rewarding than a major degree from an average university. hence choosing where to pursue your education also plays an important role in your future career as it exposes you to the competitive world to expand your horizons.

1.1.1 Motivation

The main motivation behind creating this project is to aid student regarding their decision to choose best possible university for themselves.important motivation factors of international students that play a important role in decision making process to choose foreign universities to study are highlighted as study destination reasonable cost of study and living,reputation and quality of education and tourist attraction and also university's characteristics such as offered program,quality of courses and reputation.

The main reason for ranking universities is the vast growth in higher education across the world in recent decades The number of students has grown (College can cost a six-figure sum, so you need to know you are going to the right one). The courses and programs offered by universities

are increasingly diverse. Support to increase Quality in higher Education. Goal of Ranking System is not to label foreign universities as Best or worst the main Intention is to provide valuable information to Student.

The proposed architecture, by making use of data mining, offers a solution to those problems. That is, for student there is no need to visits many websites or consultancy service for getting information about admission process as well as recommendation of university based on his/her own budgets. These technology acts as an agent on the behalf of many consultancy service center, a user can get prediction and classified instances like yes or no without visiting many website as well as without wasting of time in visiting of website. university prediction and classification using Data Mining provides the a stage to student where the student can get various of variety of a particular university information hands on The user only needs to use this technology like java to be install in system .Hence this application minimizing the use of resources

1.1.2 Advantages Over Current System

1. Our Application will be bi-dimensional
   Our System will not only focus on research, surveys but also on teaching and learning quality of knowledge transfer and also other factors which have an impact on University Ranking.

2. User - friendly Interface with tips to ensure every user becomes capable to use the software without any complexity.
   To increase the compatibility and user friendliness of the system to the user, we will be providing much simple,smoother, well formatted UI with Shortcuts and dropdowns which will make even the novice users to use it efficiently

3. Regular Check ups of Processing Units.
   All the processing units, may it be functional modules or may be the application server, it will be checked regularly for their efficient usage and rectification if required.

4. Reminder to Back up data daily.
   To avoid data loss, we will ensure to back up data everyday with our cloud storage.

5. For Overcoming vulnerabilities of Cloud Storage, and also Dos attacks and similar malware attacks, we will take the help of Amazon Cloud Services which promises top-notch protection

1.2 Proposed System Architecture

Our System Architecture is divided into three main functional Categories :

1) Data Acquisition : Administrator of our system will first gather all the relevant data from real time structured database, Excel Sheets, User Inputs (Comments and research surveys) and Text files which will be uploaded in our main database warehouse which will be directly connected to cloud technology.

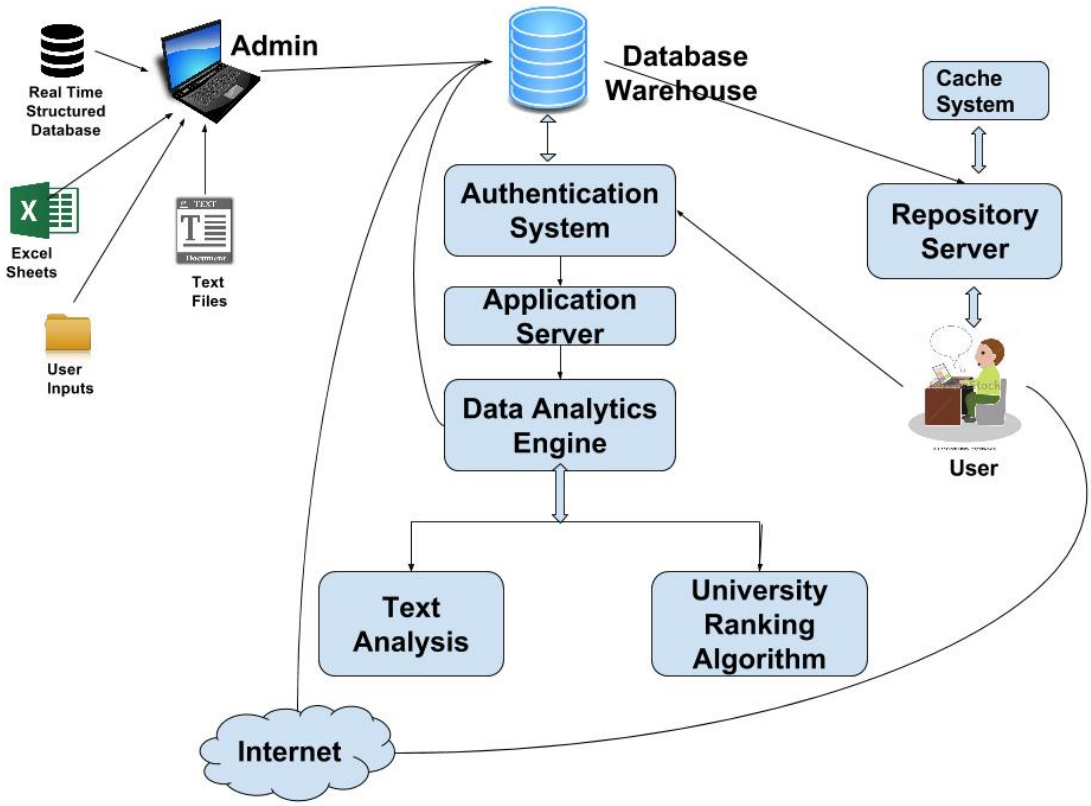

Figure 1.1: System Architecture

2) Cloud Data Storage : For implementing this Integrated DBMS Software we will be using Cloud technology to keep the system Online,acessible to all potential customers. Our Database Warehouse will be connected to every module of the system using Cloud data Storage,databases will be accessed Online making the system dynamic as well as ready to serve anytime, anywhere and at any place.

3) Presentation / GUI : Our application will be equipped with Simplistic and Well-designed User Interface to ensure high user interaction using graphical icons and visual indicators such as secondary notation, as opposed to text-based interfaces, typed command labels or text navigation. Well-designed graphical user interfaces can free the user from learning complex command languages.

Admin Responsibilities :
1)Collect and Upload Relevant Data
2)Updation and integration of Database using Cloud technology
3)Regular Checking of Processing Units.

Server Side :
1)University Ranking along with predicting pF with the Help of Naive Bayes Classifier.
2)Text Analysis of research and text files.
3)Repository Server for efficient Data usage using cache System

1.2.1 Formulation of Problem With using Technology

It is really a tedious and time-consuming process that students go through when choosing the best suited foreign university for their further education. The process might start with a search for a particular interested university from which several links to different universities are returned. The student typically visits each website to check different factors such as Programs Offered, Cost of Living, Tution fees, Quality of Educational Degree etc. This could involve considering alternate universities from an online catalog, program and courses availability, location options etc. After all relevant information is gathered, the student will then enroll into university for getting himself admitted to it using a educational qualification as a gateway. It should be noted whenever the student want to get admission into any university through online from any source he/she does visit the many Universities websites for getting the desirable interest fields and information about the university which will help the student in his decision making process whether to opt for that university or not. Like this student surf lots of time in visiting of universities websites for getting admission into desired university or affiliated program. The student not only surfs lots of time in visiting of university sites, and affiliated programs, but also he/she suffers from limited option to choose the programs offered.

The University Classification and Prediction is applicable in many application fields from which some areas are specially recognized as University Ranking, Educational Consultancy and Colleges. These areas need high transaction of money. Now the user can directly enter his educational qualification and preferred University of choice to analyse the probability of getting admission or not and also can predict for himself the best suited university among the varied list of options with respect to his educational qualification and area of interest. This is one reason which instigates much research effort in University Classification using Data Mining which has its own benefits when used in these areas.

The proposed architecture, by making use of Naive Bayes Classifier, offers a solution to those problems. That is, for students there is no need to visits many university's websites for gathering information about the desired university's criteria of admission. The student just needs to enter his data to the application which will automatically classify the probability factor whether the designated student will be able to get admission or not in among many different universities stored in our database. The application will also help the student by predicting the most suitable university the student will be able to get admission to by comparing the student sample data to the trained model which will be classifying and predicting based on a Naive Bayes Classification Model. This model will also be regressively tested on testing data for improving its accuracy

and predict out the best suited university for the student. hence this application minimizing the use of resources.

1.3 Organization of the Project

The remaining part of the project is organized as follows.

Chapter 2 Presents a review of related work.

Chapter 3 Introduces the Software and Hardware Requirement of the project.

Chapter 4 Proposes the Project Design of the Project . It represent the architectural design, front end design and database design of the project.

Chapter 5 Introduces the system model and some basic assumptions and Dependencies of our work.

Chapter 6 Presents the Results and Test cases related work.

Chapter 7 Described the time management and time utilization during the Project implementation.

Chapter 8 Described the Workload distribution.

Chapter 9 Provides some concluding remarks and directions of our future work.

Chapter 2

Review Of Literature

2.1 Webometric Ranking of World Universities: Introduction, Methodology, and Future Developments

2.1.1 Description

Today the worldwide web (web) is one of the main sources of information and the main showcase for everyone (institutions, business enterprises, individuals, etc.) who wants to be recognized on in the "real world". At the academic level, universities have a very important role as a means to communicate scientific and cultural achievements. Web publication by scholars is not only a tool for scholarly communication but it is also a means to reach larger audiences and in general a reflection of the performance of the institutions. There have been several efforts to develop web indicators that can ultimately lead to build a university's rankings. This paper presents the Webometric Ranking of World Universities which is built using a combined indicator called WR that takes into account the number of published web pages (S) (twenty five percent), the number of rich files, those in pdf, ps, doc and ppt format (R) (12.5 percent), the number of articles gathered from the Google Scholar Database (Sc) (12.5 percent,) and the total number of external in links (V) (fifty percent). The results show that there is a larger than expected academic digital divide between higher education institutions in the United States and those in the European Union.This kind of rankings using web indicators should be used to measure universities' performance in conjunction with more traditional academic indicators.

2.1.2 Pros

- In this System, Webometric Ranking of World Universities is built using a combined indicator called WR based on number of published web pages (S) (twenty five percent), the number of rich files, those in pdf, ps, doc and ppt format (R) (12.5 percent), the number of articles gathered from the Google Scholar Database (Sc) (12.5 percent,) and the total number of external inlinks (V) (fifty percent).

2.1.3 Cons

- Ranking System currently available on the World wide web judge the performance of any university which respect to certain predefined performance indicators namely Academic Reputation of the university ,Employer Reputation, Faculty Students and Campus Placements.This is because they are focused primarily on research and do not fully cover teaching and learning quality of knowledge-transfer

- It may not be updated regularly enough therefore there may be chances of getting wrong result.

2.1.4 How we overcome Those problem in Project

- Our Application will be bi-dimensional Our System will not only focus on research, surveys but also on teaching and learning quality of knowledge transfer and also other factors which have an impact on University Ranking.

- We are using a periodic trigger in our project which will regularly update the product records stored in database.

2.2 The Use of University Rankings in the United Kingdom

2.2.1 Description

University league tables and rankings are produced annually in the United Kingdom by the leading "quality" newspapers using statistical data on universities and colleges published each year. The newspapers claim that the up-to-date independent measures which they provide annually on the changing quality of universities are necessary to guide students in choosing universities in which to enroll. Many, however, are skeptical about the accuracy (as opposed to the precision) of the data, where year-to-year movements in the rankings of institutions are as much artifacts of the data and their manipulation as real changes. Prospective students seem not to be strongly influenced by the annual changes in the league table position of a given institution but are influenced more strongly by competition for places and the quality of the learning environment offered by the university.

2.2.2 Pros

- In this System, year-to-year movements in the rankings of institutions are evaluated and their manipulation is also calculated which acts as necessary to guide students in choosing universities in which to enroll.

2.2.3 Cons

- An individual who posses the knowledge about how the university ranking process works, and also have a descriptive knowledge about the current universities worldwide can only be able to operate the system.

- Setting up Domains, Servers and Also other stuffs will cost initially high.

2.2.4 How we overcome Those problem in Project

- User - friendly Interface with tips to ensure every user becomes capable to use the software without any complexity. 5 To increase the compatibility and user friendliness of the system to the user, we will be providing much simple,smoother, well formatted UI with Shortcuts and dropdowns which will make even the novice users to use it efficiently.

2.3 Academic quality, league tables, and public policy: A cross-national analysis of university ranking systems

2.3.1 Description

The global expansion of access to higher education has increased demand for information on academic quality and has led to the development of university ranking systems or league tables in many countries of the world. A recent UNESCO/CEPES conference on higher education indicators concluded that cross-national research on these ranking systems could make an important contribution to improving the international market for higher education. The comparison and analysis of national university ranking systems can help address a number of important policy questions. First, is there an emerging international consensus on the measurement of academic quality as reflected in these ranking systems? Second, what impact are the different ranking systems having on university and academic behavior in their respective countries? Finally, are there important public interests that are thus far not reflected in these rankings? If so, is there a needed and appropriate role for public policy in the development and distribution of university ranking systems and what might that role be? This paper explores these questions through a comparative analysis of university rankings in Australia, Canada, the UK, and the US

2.3.2 Pros

- In this System, Global Ranking is done through a comparative analysis of university rankings in Australia, Canada, the UK, and the US and also makes an important contribution to improving the international market for higher education

2.3.3 Cons

- There should be annual changes made to the system with respect to the changes in the selection procedure of the university as well as their merits and demerits Maintenance and Repairing Cost will be High if something crashes. If the system crashes it might take a toll to bring it back to normal again.

- Many Security issues are to be found in the present system, many loop holes which may provide unauthorized access

2.3.4 How we overcome Those problem in Project

- To avoid data loss, we will ensure to back up data everyday with our cloud storage.

- For Overcoming vulnerabilities of Cloud Storage, and also Dos attacks and similar malware attacks, we will take the help of Amazon Cloud Services which promises top-notch protection

2.4 Technological Review

2.4.1 Java

In implementing the project, we are making use of java language as the main language.The Desktop GUI,file converter and classification and prediction algorithm's are developed in java. Since java is platform independent and has rich functionality such as sophisticated multi-threading facility,Inheritance,Applet, java Swing therefore java has been roped in to develop the Graphical user interface of the application.

2.4.2 R

In our project For Graphical analysis of student data we make use of R which show statistical analysis as well as various graphs for understanding the overview of the data R is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

2.4.3 WEKA

In our project for Data mining algorithm like classification and prediction we make use of weka(Waikato Environment for Knowledge Analysis) which a open source data mining tool used for predictive modeling and data analysis. Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection.

2.4.4 MySQL

For backend of our project we used MySQL as a database server it is an open-source relational database management system (RDBMS) as it provide some good feature like Multiple storage engines, allowing one to choose the one that is most effective for each table in the application.Native storage engines InnoDB, MyISAM, Merge, Memory (heap), Federated, Archive, CSV and Cross platform support with graphical user interaction

Chapter 3

Requirement Analysis

3.1 Platform Requirement :

3.1.1 Supportive Operating Systems :

The supported Operating Systems for client include:

- Windows xp onwards
- Linux any flavour.
- Any other Operating System Which has Java.

Windows and Linux are two of the operating systems that will support comparative Application. Since Linux is an open source operating system, This system which is will use in this project is developed on the Linux platform but is made compatible with windows or Any Other Operating System as it is developed in java which is platform Independent.The comparative application will be tested on both windows and mac.

The supported Operating Systems for server include: The supported Operating Systems For server include Linux. Linux is used as server operating system. For web server we are using apache 2.0

3.2 Software Requirement :

The Software Requirements in this project include:

- Java
- R
- Weka
- MySQL
- NetBeans
- Internet Explorer, Mozila FireFox, Google Chrome etc

In this project, the use of weka and Mysql are used for creating the backbone structure of the comparative are java Netbeans(Java Applets) that are used for developing desktop modules such as login system, data Management,Predictive model etc. Weka is a specialized tool developed by Waikato written in java programming language for Data mining Algorithm on test Data such as Classification,clustering,Assoication etc weka libraries are imported in the project for developing our main classification algorithm. Through Weka libraries, classifier can perform efficient classification and thus increasing the performance of our analysis instruction.Mysql is used for deleting and updating user data. Naive Bayes is the classification algorithm which is more efficient classifier compared to other algorithm's which makes it extremely important in the entire application.

Java language is the fundamental language being used in the development of the project. MySQL is a sql which is used as a database for storing all student records MySQL reduces the complexity of maintaining relationship.With the help of MySQL, Map reduce techniques can be performed on the data Thus facilitating operations of data mining on the database.

3.3 Hardware Requirement :

3.3.1 Hardware Required For Project Development:

- Dual Core 2.4 GHz processors or higher with 8 GB RAM Minimum Or Intel I3 processor or equivalent CPU with 2GB RAM Recommended: Intel I5 processor or equivalent CPU with 4GB+ RAM
- 50+ gigabytes hard drive space, SATA 7.2k RPM, 16MB cache

- LAN 100Mbps hardware connectivity
- SVGA or better display resolution

Chapter 4

Project Design

4.1 Design Approach

Design is the first step in the development phase for any techniques and principles for the purpose of defining a device, a process or system in sufficient detail to permit its physical realization. Once the software requirements have been analyzed and specified the software design involves three technical activities design, coding, implementation and testing that are required to build and verify the software. The design activities are of main importance in this phase, because in this activity, decisions ultimately affecting the success of the software implementation and its ease of maintenance are made. These decisions have the final bearing upon reliability and maintainability of the system. Design is the only way to accurately translate the customer requirements into finished software or a system. Design is the place where quality is fostered in development. Software design is a process through which requirements are translated into a representation of software. Software design is conducted in two steps. Preliminary design is concerned with the transformation of requirements into data.

4.2 Software Architectural Designs

Our system is follow the three tier architecture . First tier consist of GUI, Comparision tier and the Database.

1. GUI: The GUI(Graphical User Interface) in our project deals with the interface for the user where the user enters the marks, gre score, Budget Area of interest along with basic details he/she wants to predict.The GUI provides a platform for the user to communicate with the database.

2. Comparision block: The comparision block is the block where the actual processing of our project is done. This block connects the GUI to the database i.e. It acts as a connector as well as communicator which connects the database and helps in transfer of data between the

GUI and the database. It's main function is to transfer our database data into fileformat like csv arff and do the classification prediction on the basis of the user requirement and other aspects.

3. Database: Database tier is the tier used for the storage of data. This tier contains all the data that is need for the processing of the whole project. The data in this tier is related to the product details such as the BE agg,Gre score, Budget,Area of Interest, along with other detail require for analysis.

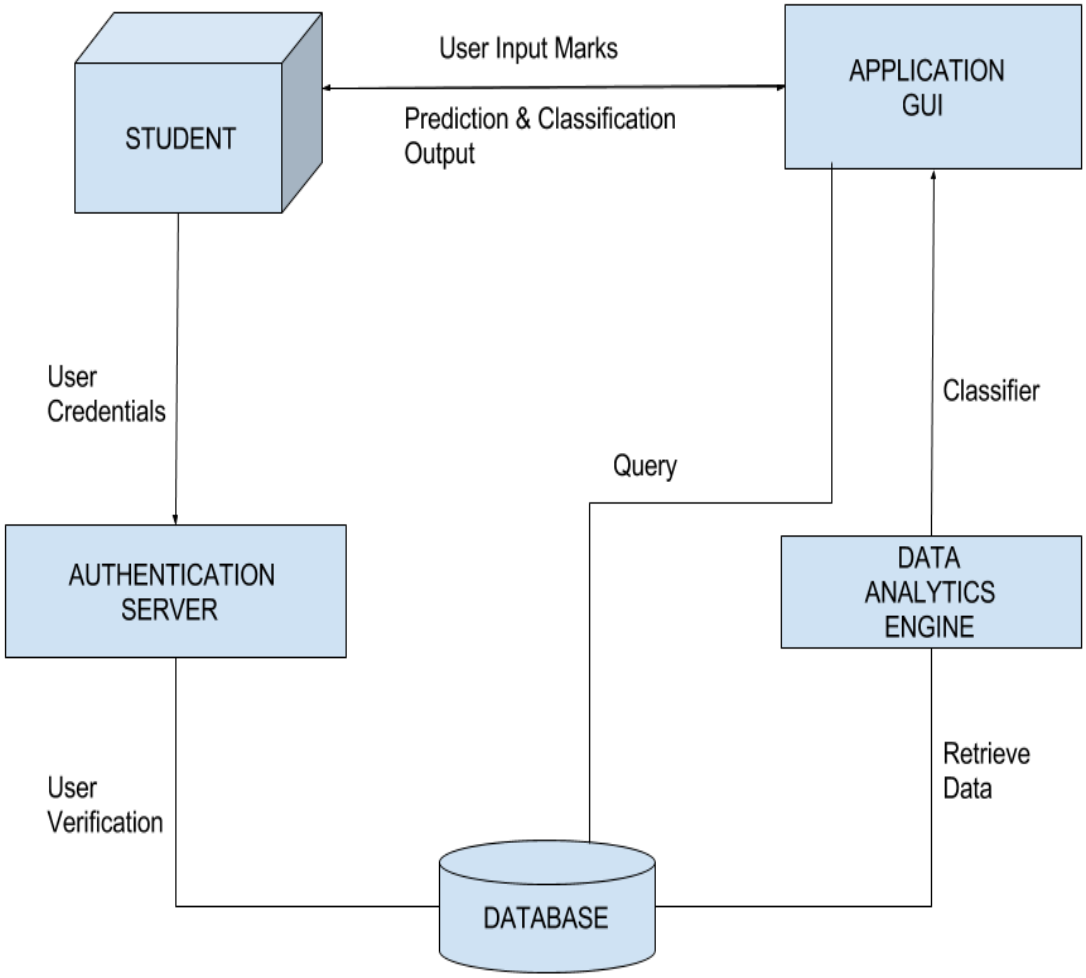

Figure 4.1: Software architecture Design

4.2.1 Front End Designs

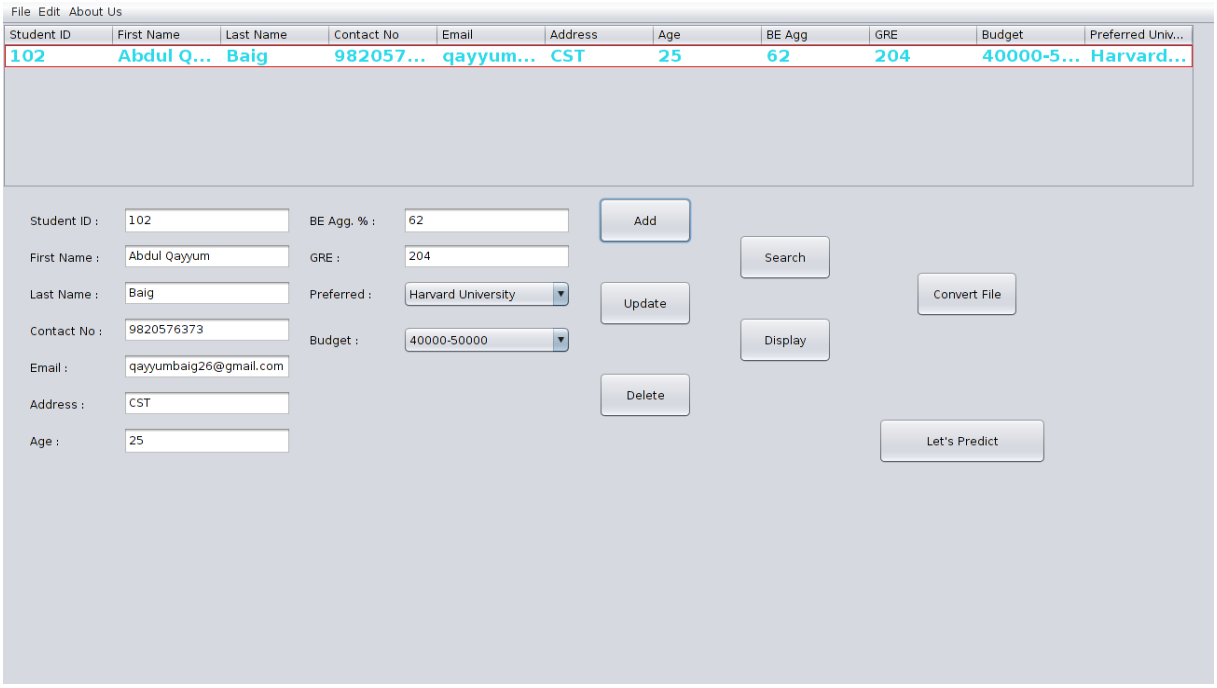

Figure 4.2: Front End Design

4.2.2 Component Diagram

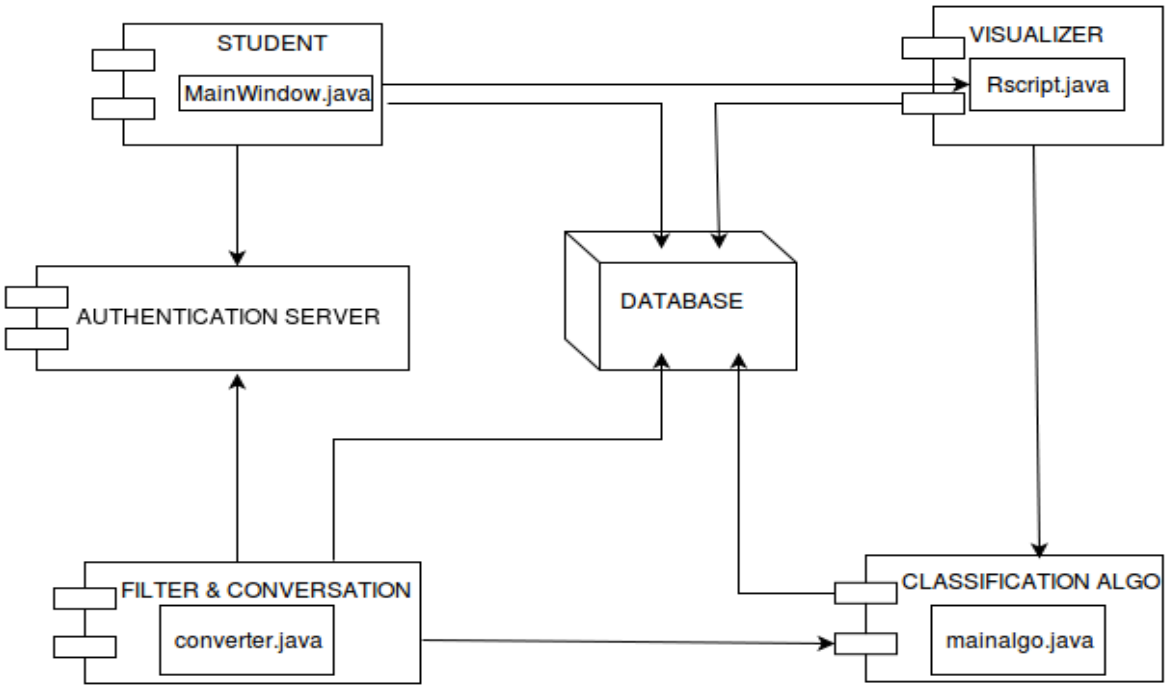

Figure 4.3: Component Diagram of University Classification Prediction System

4.2.3 Deployment Diagram

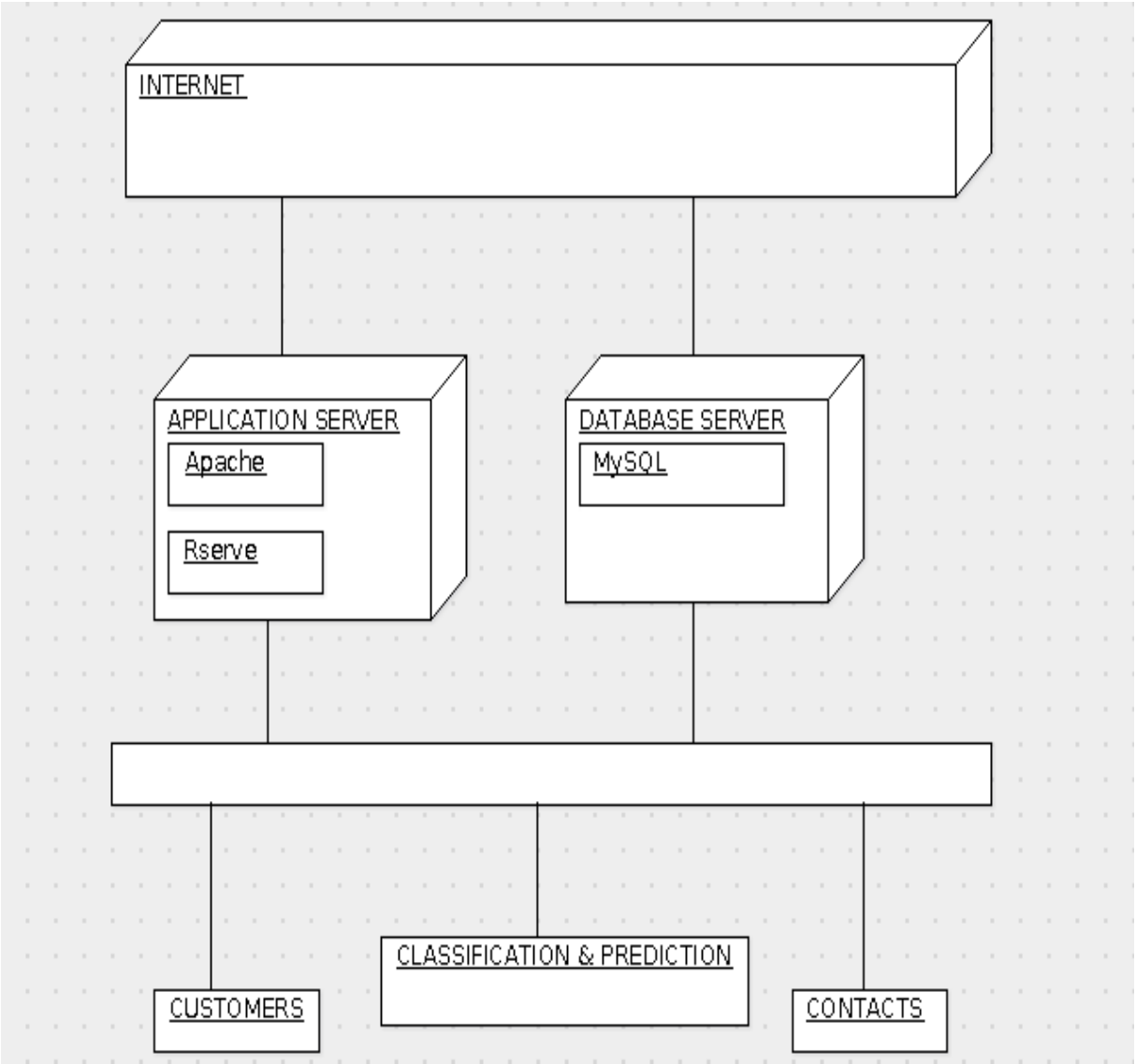

Figure 4.4: Deployment Diagram of University Classification Prediction System

4.3 Database Design

4.3.1 E-R Diagram

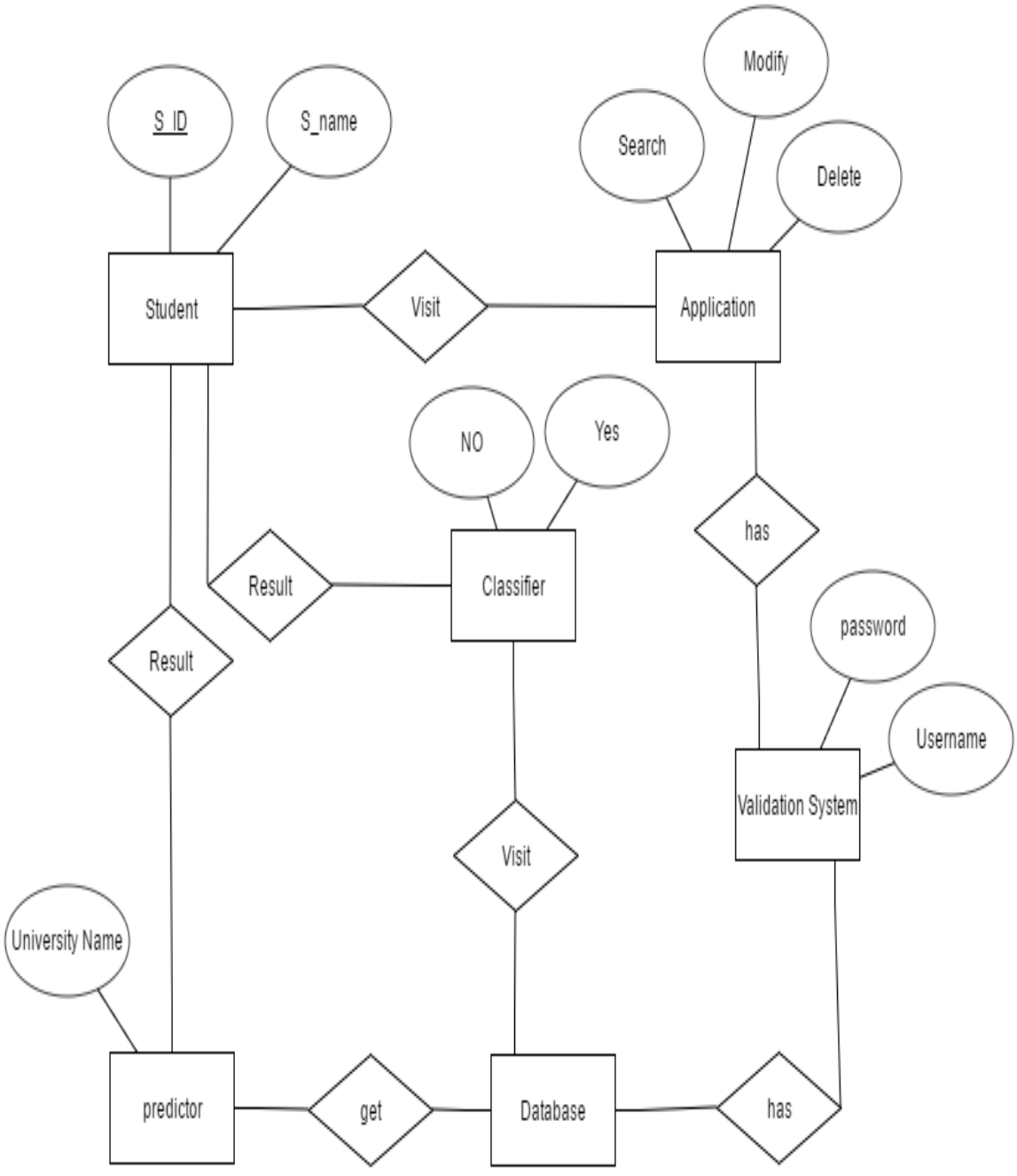

Figure 4.5: E-R Diagram of University Classification Prediction System

4.4 Work-flow Design

4.4.1 Flow Diagram

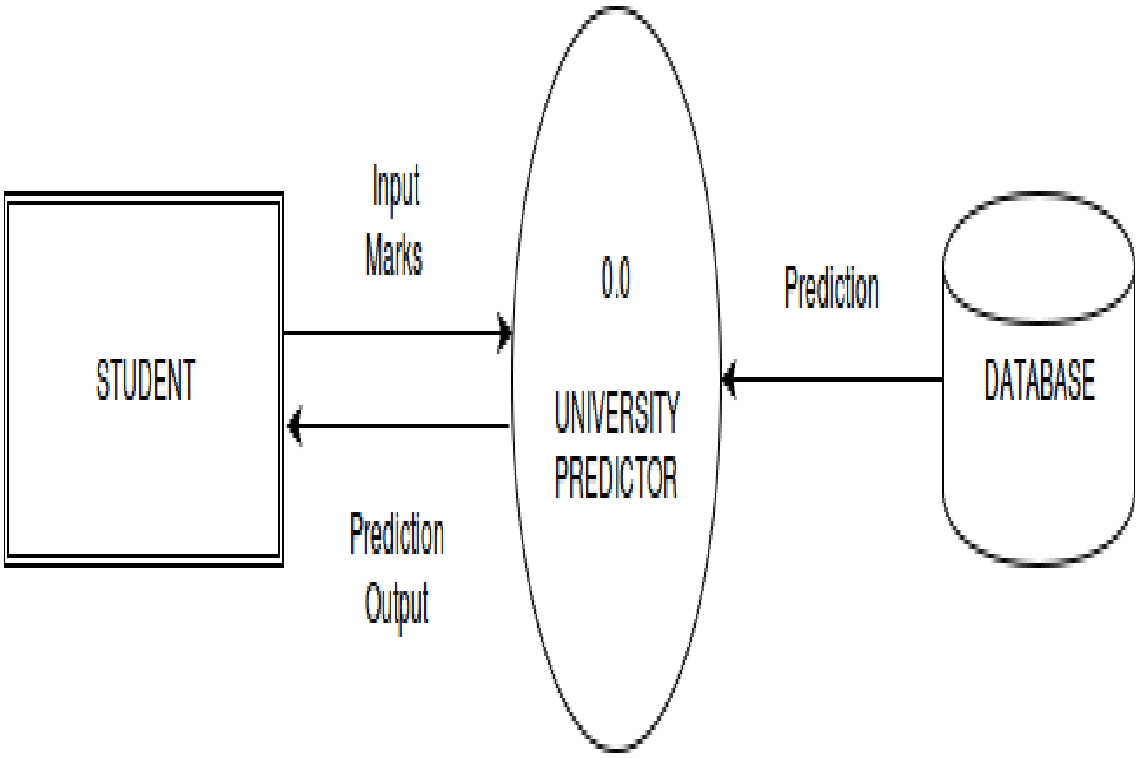

Figure 4.6: Level 0 DFD of university classification and predictions system

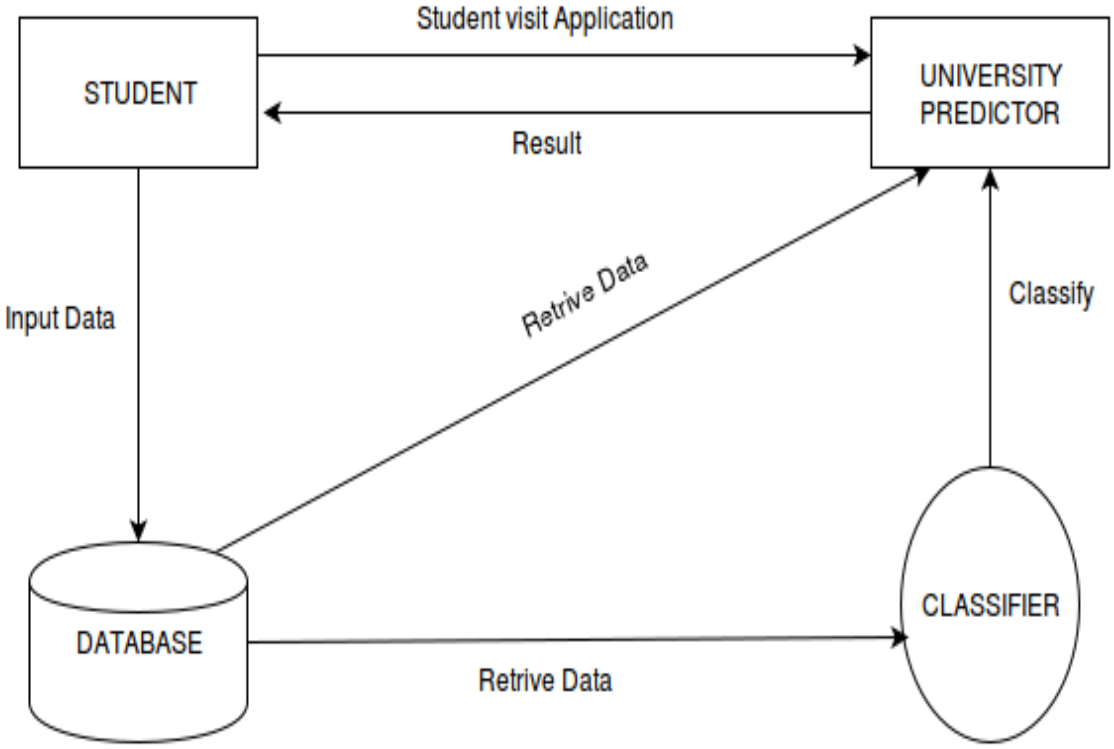

Figure 4.7: Level 1 DFD of university classification and prediction system

Chapter 5

Implementation Details

5.1 Assumptions And Dependencies

5.1.1 Assumptions

The following Assumption was taken into consideration:

- Weka NaiveBayes algorithm is Assume that it fetches the data from student enter into database and stores in the local database. The algorithm fetches the data from different tables in database and updates the local database if any updates are made by the Student. Therefore it is assume that the algorithm has to be very effective and fast in fetching the database. The local database is assumed to be scalable and robust so that it can store huge amount of data with time and maintain consistency of the data.

- The classification algorithm has to very optimize in performing the prediction As soon as the user click predict the database the search algorithm is assumed to bring the required and accurate results. The user interface should be simple and clean that allows soothing effect to the user. The comparative algorithm is assumed to be very effective that allows comparison amongst different products.

5.1.2 Dependencies

The dependencies are as follows:

- For backend processing,Weka classification and R is being used. The backbone structure of the system is developed by making use of java Swing,java AWt and Netbeans IDE software. In creation of application,java and MySQL is used. Weka is a specialized tool developed by WAIKTO University which is written in java language For performing data

mining tasks like clustering,classification and association etc. The extracted dataset are stored in MySQL database.

- R is the visualization server tool which is put into use for handling server side request and response. Real Time student data analysis using graphs generation it is basically serves as a mediator between java application and Database.It depends on data extracted from different student and university.

5.2 Implementation Methodologies

Different Modules created are Classifier, search, update, predictor, inventory module, fileformat and testing. In classification module, the classifier is created using weka libraries, java and MySQL. The classifier module contains code for Fetching data from database and filtering data for making train .The filter data are then stored in the database. The classifier module is then invoked to store the Data in inventory module

5.2.1 Modular Description of Project

5.3 Detailed Analysis and Description of Project

Inventory Module: In this module, the admin can add, delete and update student information. Store script module contains all the queries for the creation of Database in the MySQL and the required queries to perform operations on the database.

File viewer module: In Fileviewer the java coding is done and the front end is developed.In this module, the student can select the file and put sample data it into database.When the student clicks on the prediction, then it is redirected to the actually algorithm which will do the main operation

Classifier: The classifier fetches the data from train dataset In this project, a Naivebayes Classifier is used. The job of the NaiveBayes is to periodically fetch data from sample dataset. Since application database hold huge amount of data, the Classifier has to be very efficient in extracting student Required data. For this purpose, a specialized tool called weka is being used. The weka is a specialized tool developed by WAIKTO which holds specialization in dataset extraction and collection. Then we perform filtration of the dataset in order to get the right and useful data for prediction and removes the unuseful data.

Visualization module: In this module the data which we got from student are pass to R script program throught mysql database which will create bar graphs,pie charts,histogram of student data so that by just seeing graphs admin can check student interest section like area of interest,budgets and number of student got higher marks/lower marks as well as a overview of data entities.

User Search: Whenever admin searches for the particular student in the search bar, the database is queried in order to retrieve the accurate results. Now the admin is presented with a particular student information with different attributes likes student marks,gre score,student budgets.This allows the admin to compare student based on say marks. In this project, the search is required to be very effective and efficient. For this purpose various factors must be taken into consideration such as time, address of the student etc.We used MySQL query for making search more effective as MySQL provide best searching tasks .Also to track user behavior and search patterns, algorithms are used that keeps track on users and their behaviors and their searching patterns in order to provide user with better search experience.

Prediction Module: Prediction module techniques are performed on the data present in the database. In this Module prediction algorithm is performed which is give feedback to student about which university is most suitable college/university base upon his/her own sets of marks and also provide suggestion on which are the best university in preferred location. The database is queried in order to get the results which the user searches.

File Converter: This module is important because our algorithm works on arff file format for that we need to convert the student data which is recorded from student through form.converter will first fetch data and convert into CSV file then CSV is converted into arff file format with the help of weka arff converter functionality libraries.

5.3.1 Usecase Report

Title:	University Classification and Prediction Using Data Mining
Decription:	University Classifier and Prediction Using Data Mining,provide to student wheteher he/she will get a admission or not and also predict the best suitable university according to his/her academic details .
Primary Actor:	Student
Preconditions:	Student to visit the appliction
Post conditions:	Student Fill Details Properly
Main Success Scenario:	1. Student fill the Required Inputs like score,gre marks into the database through Gui application,Application will fetch the data and send to classification & prediction algorithm. 2. Algorithm will make analysis on the student data and respond back to student in terms of (yes or no) and University name. 3. Student will get to know the whether he/she will be getting admission in preferred or any other university.
Frequency of Use:	Student can use at any number time.
System Requirement:	Normal
Pre-Processing:	A weka function,which will fetch the data from database and convert it into csv and then arff file format which is used to supply as train dataset to algorithm.

Figure 5.1: Usecase Report

5.4 Class Diagram

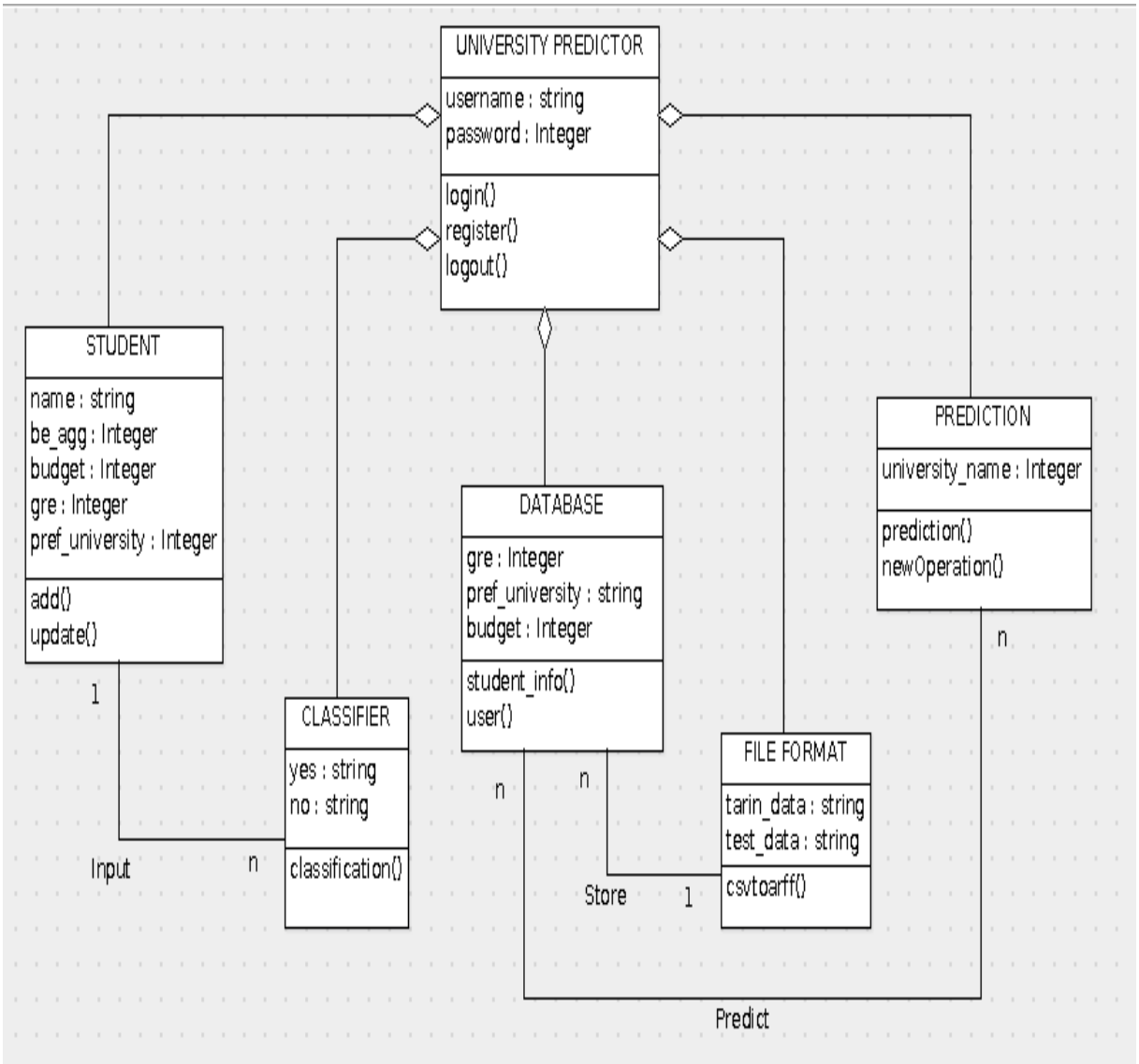

Figure 5.2: Class Diagram

5.4.1 Class Diagram Report

Title:	University Classification and Prediction Using Data Mining
Description:	University Classifier and Prediction Using Data Mining,provide to student wheteher he/she will get a admission or not and also predict the best suitable university according to his/her academic details .
Primary Actor:	Student
Preconditions:	Student to visit the appliction
Post conditions:	Student Fill Details Properly
Prediction:	This Entity Perform the Prediction of University name by using naive bayes algorithm.
Classifier:	Algorithm will make analysis on the student data and respond back to student in terms of (yes or no).
File Format:	Train & Test data format are prepared from database to csv,then csv to arff.
Databases:	Database is used for storing the information of student like gre,budget,pref_university etc.

Figure 5.3: Class Diagram Report

Chapter 6

Results and Discussion

6.1 Test cases and Result

We have tested our application by considering following test cases:

6.1.1 Unit Testing

In this testing we are making test on different working units of our projects Example , Below picture shows the validation of different end-user of the application as well as student searching unit testing are also done.

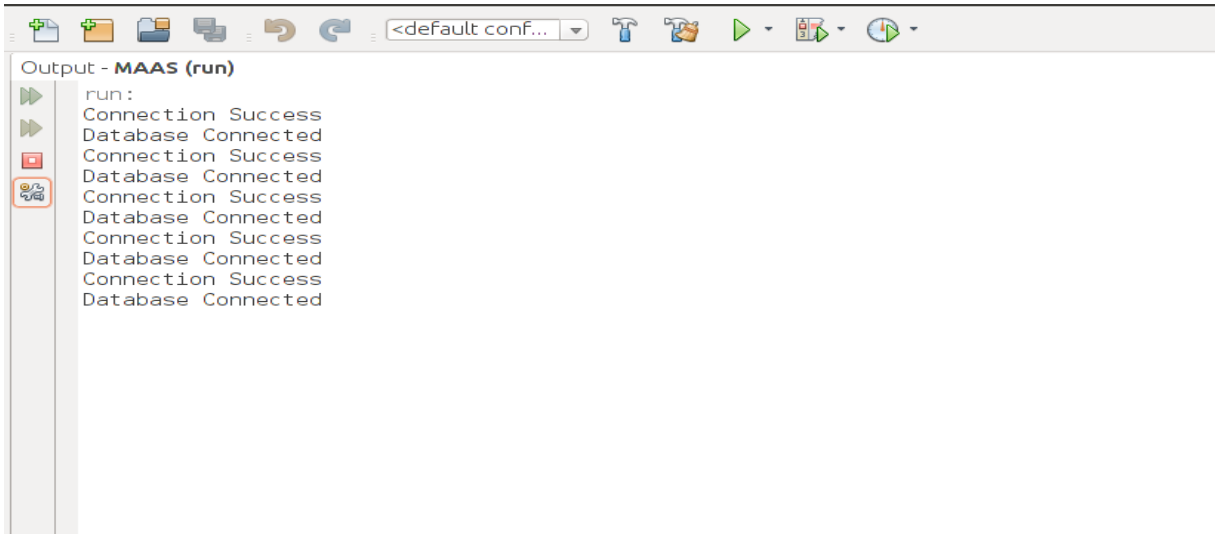

Figure 6.1: Database Connectivity

Figure 6.2: Validation of User

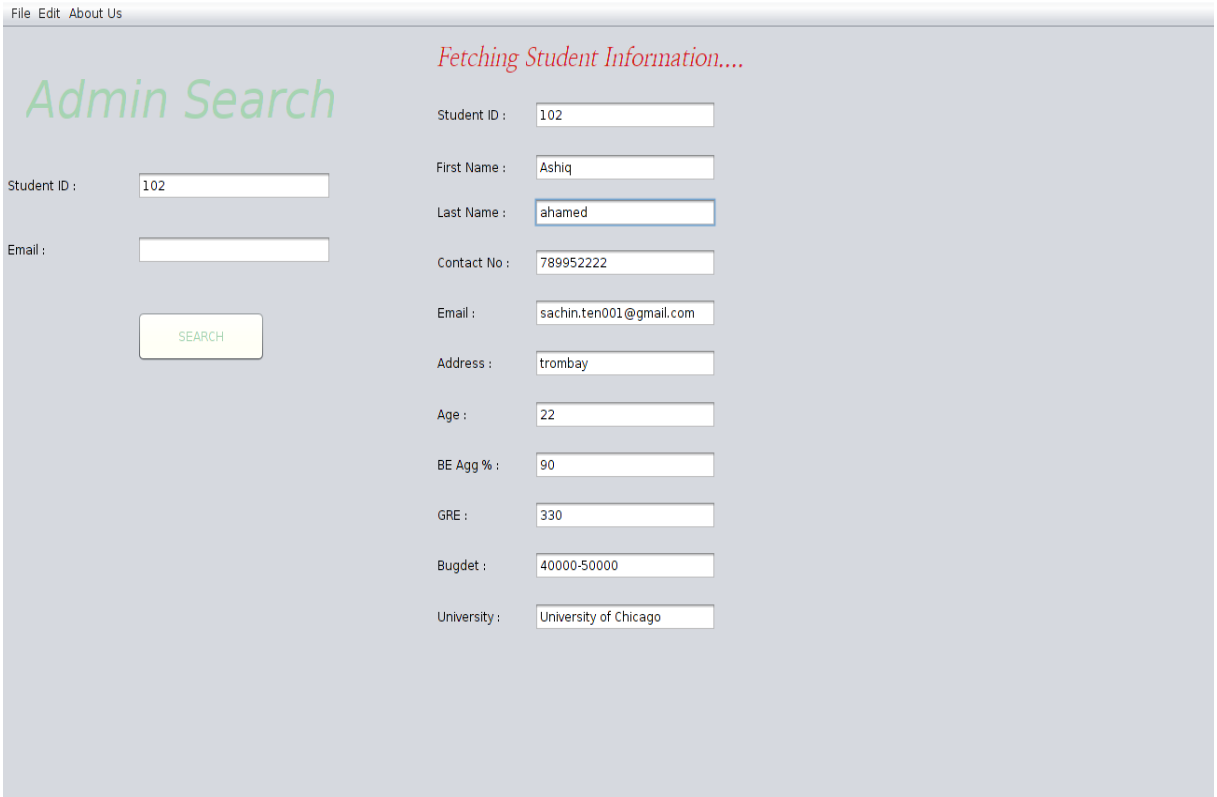

Figure 6.3: Student Searching

```
mysql> select * from Student;
+------+---------+---------+-----------+-----------------------+----------+-------+-------+-----+------------+-----------------------
------------+
| S_id | S_Fname  | S_Lname | S_contact  | email                 | S_add     | S_age | BEagg | GRE | Budget      | prefUniv
           |
+------+---------+---------+-----------+-----------------------+----------+-------+-------+-----+------------+-----------------------
------------+
|    1 | Mohammed | Shadir  | 789456123  | shadhir99@gmail.com      | andheri   |    21 |    65 | 190 | 40000-50000 | Harvard university
           |
|   12 | Alim     | Shaikh  | 9820576373 | shaikhalim20@gmail.com   | mumbra    |    22 |    61 | 170 | 30000-45000 | MIT University
           |
|   13 | Rehan    | Khan    | 8798546587 | rehan12@gmail.com        | Vashi     |    25 |    58 | 265 | 40000-50000 | Stanford University
           |
|   14 | Sahil    | Khan    | 9845786598 | sahilkh12@gmail.com      | Thane     |    23 |    61 | 250 | 30000-35000 | University of Chicago
           |
|   15 | Inayat   | Ansari  | 7893154526 | inayatam12@gmail.com     | Kalwa     |    24 |    55 | 230 | 50000-60000 | Massachusetts Institute o
f Technology |
|   16 | Amit     | Patil   | 9878451265 | amitpatil12@gmail.com    | Panvel    |    22 |    59 | 200 | 50000-60000 | Yale University
           |
|   17 | Ayjaz    | Khan    | 7802351548 | ayjazkhan14@gmail.com    | Bandra    |    23 |    59 | 210 | 40000-50000 | Princeton University
           |
|   18 | Mustafa  | Khan    | 9765031455 | mustafakhan125@gmail.com | Govandi   |    23 |    60 | 190 | 40000-50000 | University of Chicago
           |
|   19 | Arif     | Shaikh  | 7965840233 | shaikharif122@gmail.com  | Ghatkopar |    22 |    65 | 250 | 35000-40000 | Massachusetts Institute o
f Technology |
|   20 | Mohsin   | Khan    | 8833654587 | mohsinkhan145@gmail.com  | Mumbra    |    22 |    58 | 180 | 40000-50000 | Yale University
           |
|   21 | Intekhab | Khan    | 8844512217 | intekhab145@gmail.com    | Andheri   |    23 |    55 | 180 | 30000-35000 | Harvard University
           |
|   24 | Mannan   | Baig    | 9768827443 | mannanbaig38@gmail.com   | Chembur   |    22 |    60 | 220 | 35000-50000 | Stanford University
           |
|  102 | Ashiq    | ahamed  | 789952222  | sachin.ten001@gmail.com  | trombay   |    22 |    90 | 330 | 40000-50000 | University of Chicago
           |
+------+---------+---------+-----------+-----------------------+----------+-------+-------+-----+------------+-----------------------
------------+
13 rows in set (0.00 sec)
```

Figure 6.4: Data Stored

6.1.2 Functional Testing

We have tested our web application on the server by integrating all the units of our project. In this testing we focus on the output is as per the requirement or not. which is as follow:

Testing accuracy of the model by providing test data to algorithm(if accuracy is acceptable then model is evaluated on sample data) :-

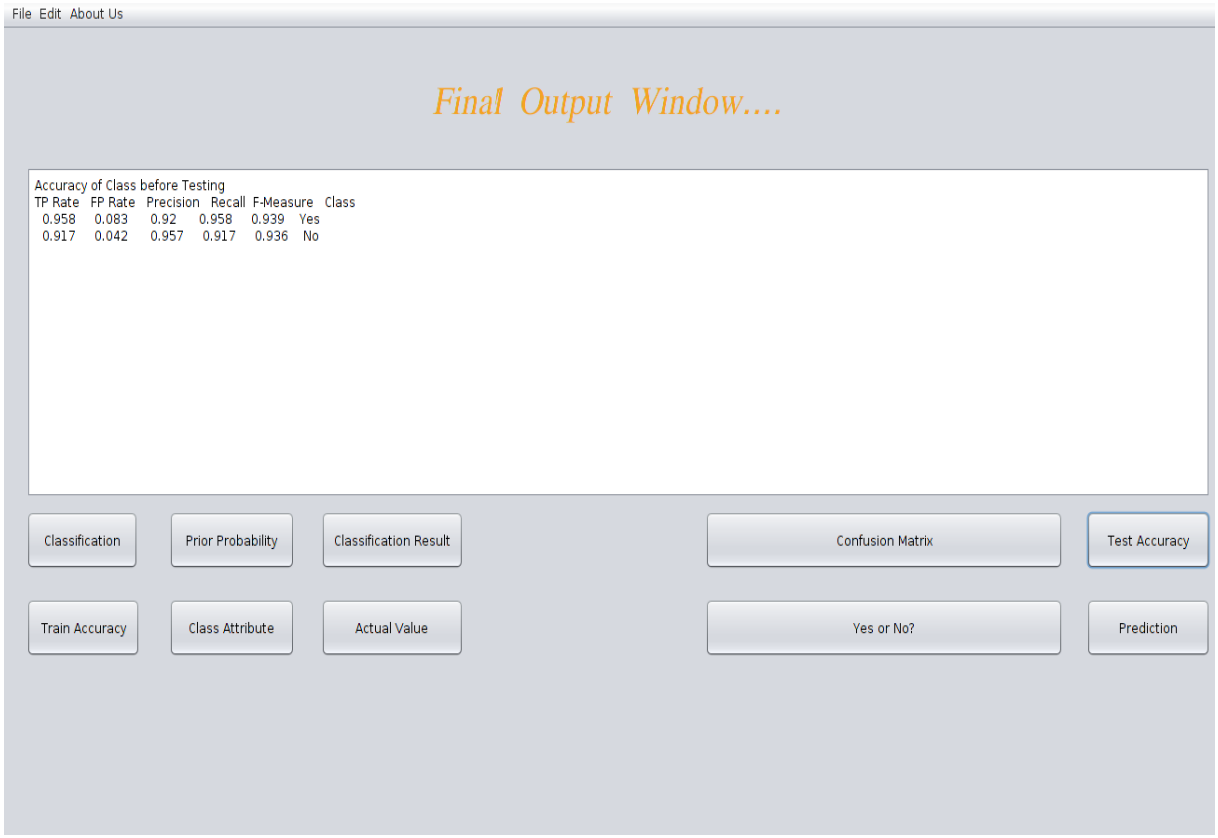

Figure 6.5: Result of Testing on Train Data

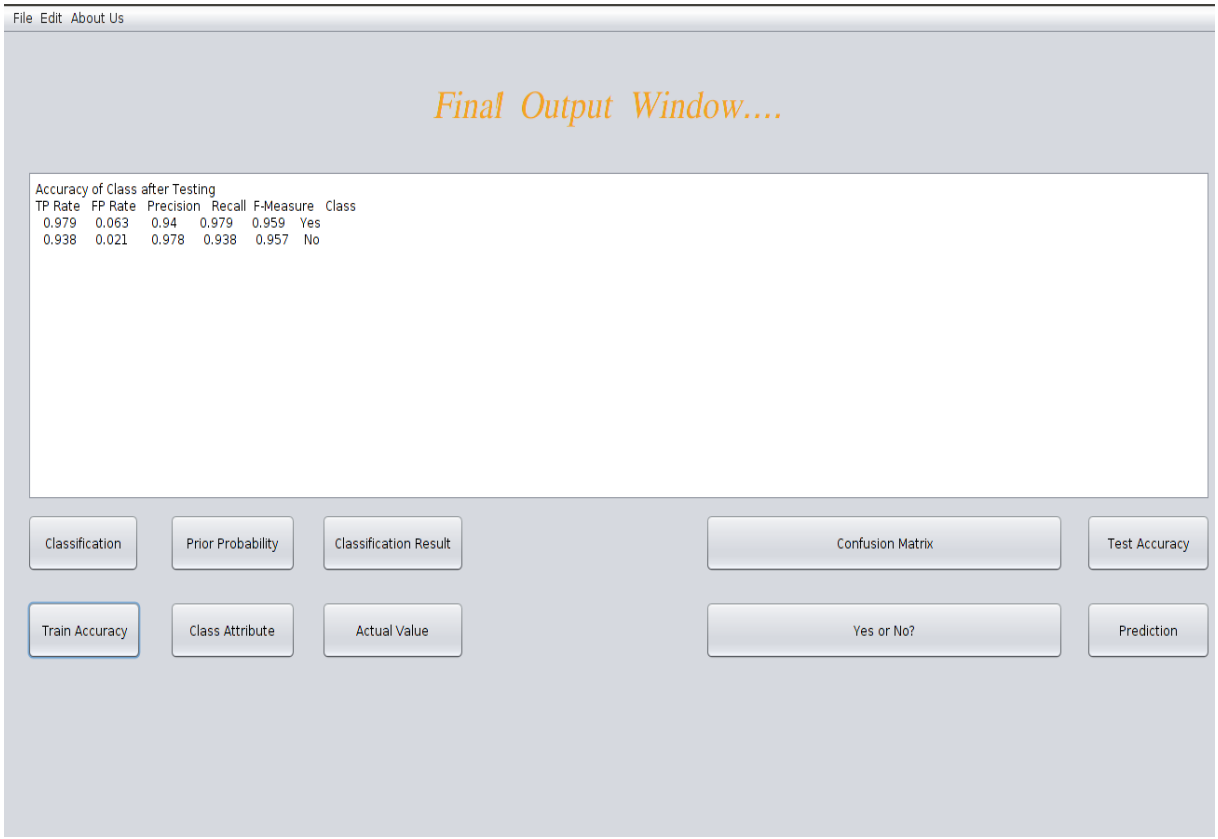

Figure 6.6: Result of Testing after giving Test Data

6.1.3 Dataset

Relation: train

No.	be_agg Numeric	gre Numeric	area of interest Nominal	budget Nominal	pref_university Nominal	admitted Nominal
1	75.0	320.0	MS Computer ...	5000...	Massachuse...	Yes
2	58.0	323.0	MS computer ...	5000...	Massachuse...	Yes
3	55.0	311.0	MS Computer ...	4000...	Harvard Univ...	Yes
4	80.0	310.0	MS computer ...	4000...	Harvard Univ...	Yes
5	52.0	308.0	MS Embedded...	5000...	Massachuse...	No
6	50.0	305.0	MBA	5000...	Massachuse...	No
7	50.0	320.0	MS Embedded...	4000...	Harvard Univ...	No
8	53.0	300.0	MBA	4000...	Harvard Univ...	No
9	60.0	330.0	MS Computer ...	3500...	Stanford Uni...	Yes
10	70.0	309.0	MS computer ...	3500...	Stanford Uni...	Yes
11	59.0	325.0	MS Computer ...	3000...	Princeton Un...	Yes
12	69.0	313.0	MS computer ...	3000...	Princeton Un...	Yes
13	49.0	270.0	MS Embedded...	3500...	Stanford Uni...	No
14	45.0	290.0	MBA	3500...	Stanford Uni...	No
15	45.0	315.0	MS Embedded...	3000...	Princeton Un...	No
16	53.0	310.0	MBA	3000...	Princeton Un...	No
17	73.0	322.0	MS computer ...	5000...	University of ...	Yes
18	58.0	314.0	MS computer ...	5000...	University of ...	Yes
19	55.0	330.0	MS Computer ...	4000...	Yale University	Yes
20	65.0	315.0	MS Computer ...	4000...	Yale University	Yes
21	51.0	265.0	MS Embedded...	5000...	University of ...	No
22	45.0	320.0	MBA	5000...	University of ...	No
23	40.0	325.0	MS Embedded...	4000...	Yale University	No
24	48.0	300.0	MBA	4000...	Yale University	No
25	50.0	320.0	MS Computer ...	5000...	Massachuse...	No
26	48.0	300.0	MS computer ...	5000...	Massachuse...	No
27	56.0	320.0	MS Embedded...	5000...	Massachuse...	Yes
28	60.0	335.0	MBA	5000...	Massachuse...	Yes
29	52.0	250.0	MS Computer ...	4000...	Harvard Univ...	No
30	52.0	311.0	MS computer ...	4000...	Harvard Univ...	No
31	65.0	314.0	MS Embedded...	4000...	Harvard Univ...	Yes
32	60.0	330.0	MBA	4000...	Harvard Univ...	Yes
33	50.0	320.0	MS Computer ...	3500...	Stanford Uni...	No
34	51.0	308.0	MS computer ...	3500...	Stanford Uni...	No
35	65.0	316.0	MS Embedded...	3500...	Stanford Uni...	Yes
36	60.0	323.0	MBA	3500...	Stanford Uni...	Yes
37	49.0	320.0	MS Computer ...	3000...	Princeton Un...	No
38	52.0	300.0	MS computer ...	3000...	Princeton Un...	No
39	57.0	316.0	MS Embedded...	3000...	Princeton Un...	Yes
40	63.0	325.0	MBA	3000...	Princeton Un...	Yes
41	54.0	317.0	MS Computer ...	5000...	University of ...	No

Figure 6.7: Train Dataset

Relation: test

No.	be_agg Numeric	gre Numeric	area of interest Nominal	budget Nominal	pref_university Nominal	admitted Nominal
1	50.0	320.0	MS Computer ...	5000...	Massachuse...	No
2	48.0	300.0	MS computer ...	5000...	Massachuse...	No
3	56.0	320.0	MS Embedded...	5000...	Massachuse...	Yes
4	60.0	335.0	MBA	5000...	Massachuse...	Yes
5	52.0	250.0	MS Computer ...	4000...	Harvard Univ...	No
6	52.0	311.0	MS computer ...	4000...	Harvard Univ...	No
7	65.0	314.0	MS Embedded...	4000...	Harvard Univ...	Yes
8	60.0	330.0	MBA	4000...	Harvard Univ...	Yes
9	50.0	320.0	MS Computer ...	3500...	Stanford Uni...	No
10	51.0	308.0	MS computer ...	3500...	Stanford Uni...	No
11	65.0	316.0	MS Embedded...	3500...	Stanford Uni...	Yes
12	60.0	323.0	MBA	3500...	Stanford Uni...	Yes
13	49.0	320.0	MS Computer ...	3000...	Princeton Un...	No
14	52.0	300.0	MS computer ...	3000...	Princeton Un...	No
15	57.0	316.0	MS Embedded...	3000...	Princeton Un...	Yes
16	63.0	325.0	MBA	3000...	Princeton Un...	Yes
17	54.0	317.0	MS Computer ...	5000...	University of ...	No
18	53.0	275.0	MS computer ...	5000...	University of ...	No
19	65.0	310.0	MS Embedded...	5000...	University of ...	Yes
20	65.0	326.0	MBA	5000...	University of ...	Yes
21	47.0	318.0	MS Computer ...	4000...	Yale University	No
22	50.0	302.0	MS computer ...	4000...	Yale University	No
23	62.0	314.0	MS Embedded...	4000...	Yale University	Yes
24	60.0	330.0	MBA	4000...	Yale University	Yes
25	75.0	320.0	MS Computer ...	5000...	Massachuse...	Yes
26	58.0	323.0	MS computer ...	5000...	Massachuse...	Yes
27	55.0	311.0	MS Computer ...	4000...	Harvard Univ...	Yes
28	80.0	310.0	MS computer ...	4000...	Harvard Univ...	Yes
29	52.0	308.0	MS Embedded...	5000...	Massachuse...	No
30	50.0	305.0	MBA	5000...	Massachuse...	No
31	50.0	320.0	MS Embedded...	4000...	Harvard Univ...	No
32	53.0	300.0	MBA	4000...	Harvard Univ...	No
33	60.0	330.0	MS Computer ...	3500...	Stanford Uni...	Yes
34	70.0	309.0	MS computer ...	3500...	Stanford Uni...	Yes
35	69.0	325.0	MS Computer ...	3000...	Princeton Un...	Yes
36	69.0	313.0	MS computer ...	3000...	Princeton Un...	Yes
37	49.0	270.0	MS Embedded...	3500...	Stanford Uni...	No
38	45.0	290.0	MBA	3500...	Stanford Uni...	No
39	45.0	315.0	MS Embedded...	3000...	Princeton Un...	No
40	53.0	310.0	MBA	3000...	Princeton Un...	No
41	73.0	322.0	MS computer ...	5000...	University of ...	Yes

Figure 6.8: Test Dataset

6.1.4 Graphical Analysis

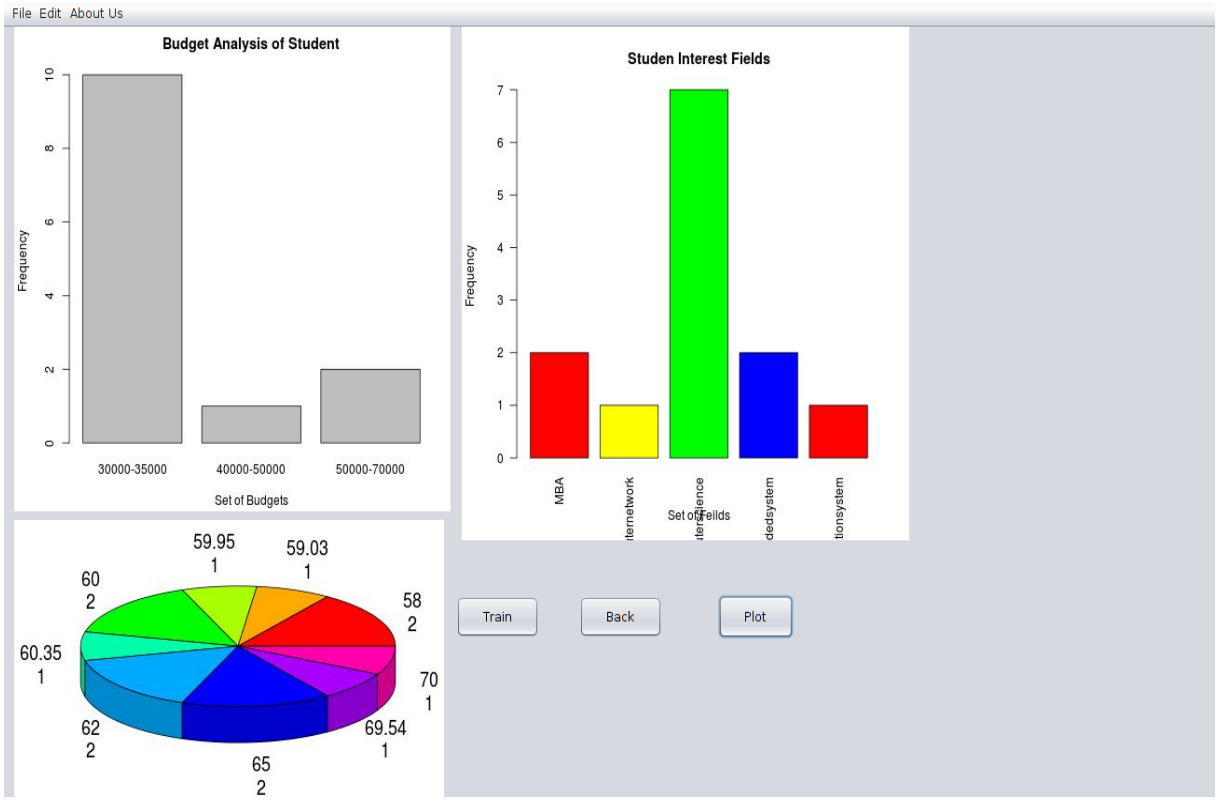

Figure 6.9: Train Graphs

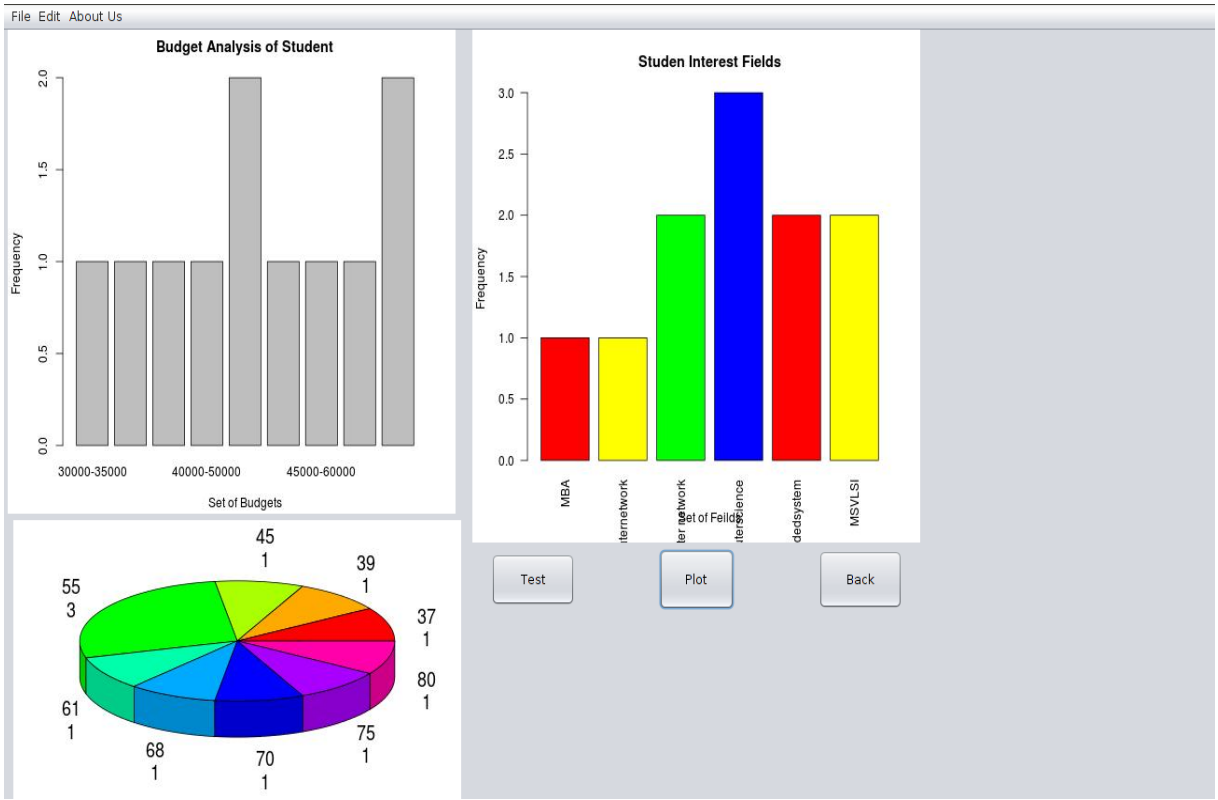

Figure 6.10: Test Graphs

6.1.5 Final Result

Sample Data Given input to Algorithm :-

Relation: test-weka.filters.unsupervised.instance.RemovePercentage-P50.0

No.	be_agg Numeric	gre Numeric	area of interest Nominal	budget Nominal	pref_university Nominal	**admitted** Nominal
1	90.0	310.0	MS Computer ...	4000...	Princeton Un...	

Figure 6.11: Sample Data for Classification

Relation: train2-weka.filters.unsupervised.instance.RemovePercentage-P50.0

No.	be_agg Numeric	gre Numeric	area of interest Nominal	budget Nominal	**pref_university** Nominal
1	90.0	310.0	MS Computer ...	4000...	

Figure 6.12: Sample Data for Prediction

Final Output Given to Student :-

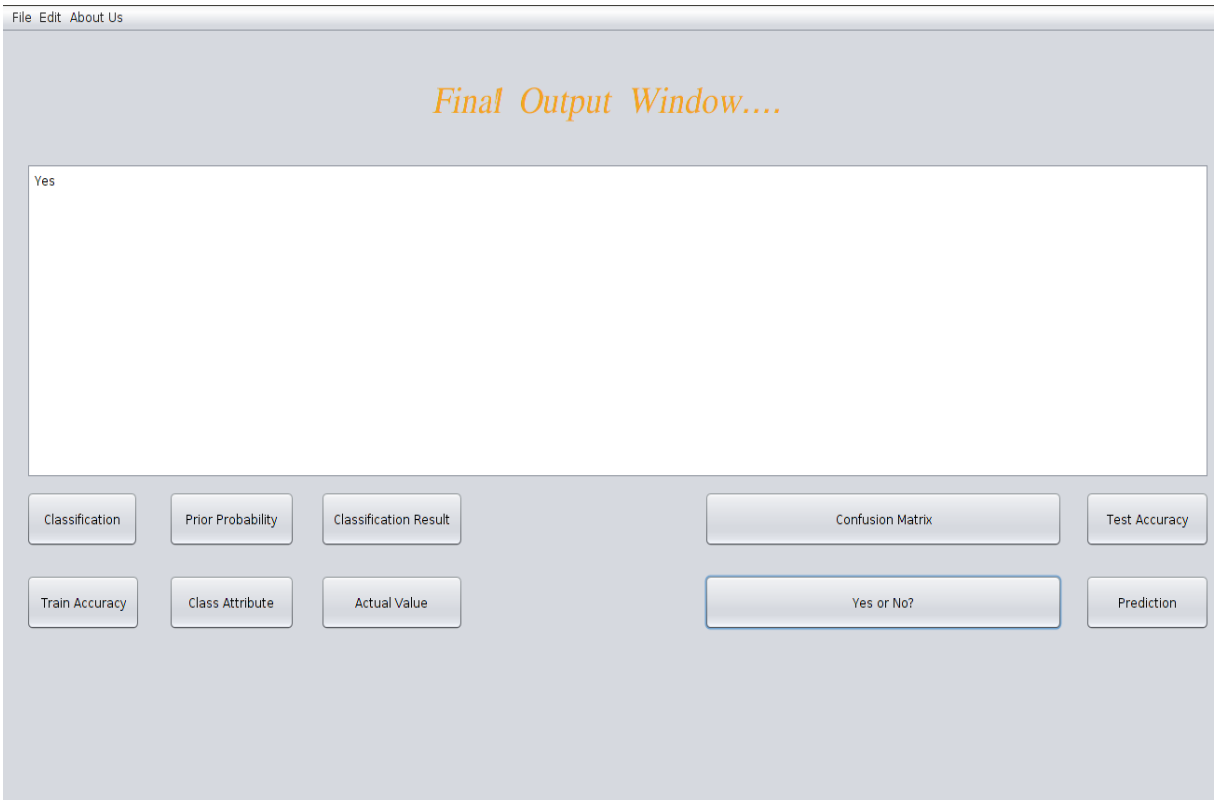

Figure 6.13: Result of Classification

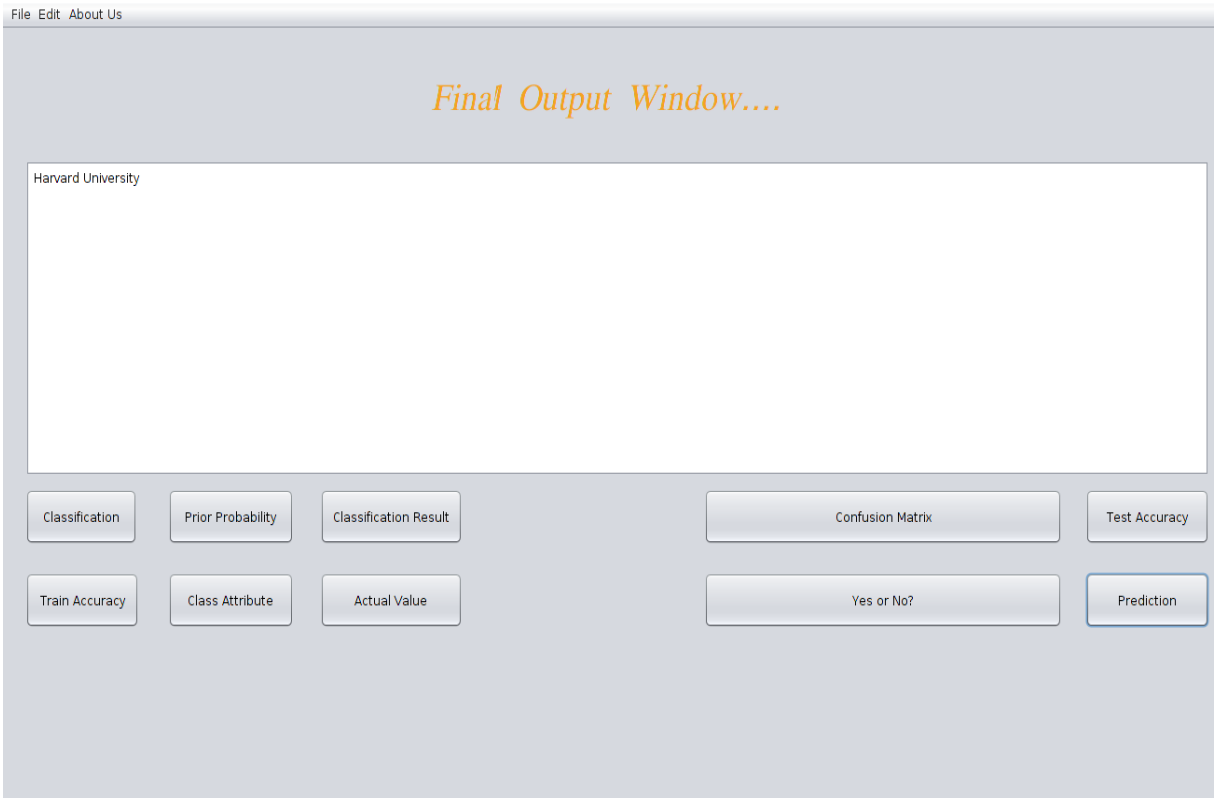

Figure 6.14: Result of Prediction

Chapter 7

Project Time Line

7.1 Project Time Line Matrix

	Name	Duration	Start	Finish
	1(a)Requirement Gathering	10 days	1/1/16 10:00 AM	1/15/16 10:00 AM
	1(b)Confirm Requirement	2 days	1/16/16 11:00 AM	1/19/16 5:00 PM
	2(a) Frontend-User Interface	8 days	1/20/16 10:00 AM	2/1/16 10:00 AM
	2(b)Backend Database Design	7 days	2/2/16 11:00 AM	2/11/16 11:00 AM
	3(a)Frontend Coding	10 days	2/12/16 10:00 AM	2/26/16 10:00 AM
	3(b) Database Creation	8 days	2/27/16 11:00 AM	3/9/16 5:00 PM
	3(c) Coding for Algorithm	10 days	3/10/16 11:00 AM	3/24/16 11:00 AM
	3(d) Creation of test case	5 days	3/25/16 10:00 AM	4/1/16 10:00 AM
	4(a)Unit Test	4 days	4/2/16 11:00 AM	4/7/16 5:00 PM
	4(b)System Test	5 days	4/8/16 11:00 AM	4/15/16 11:00 AM
	4(c)Alpha and Beta Test	6 days	4/16/16 10:00 AM	4/25/16 5:00 PM
	5(a)Deployment	1 day	4/26/16 10:00 AM	4/27/16 10:00 AM

Figure 7.1: Time Line Matrix

7.2 Project Time Line Chart

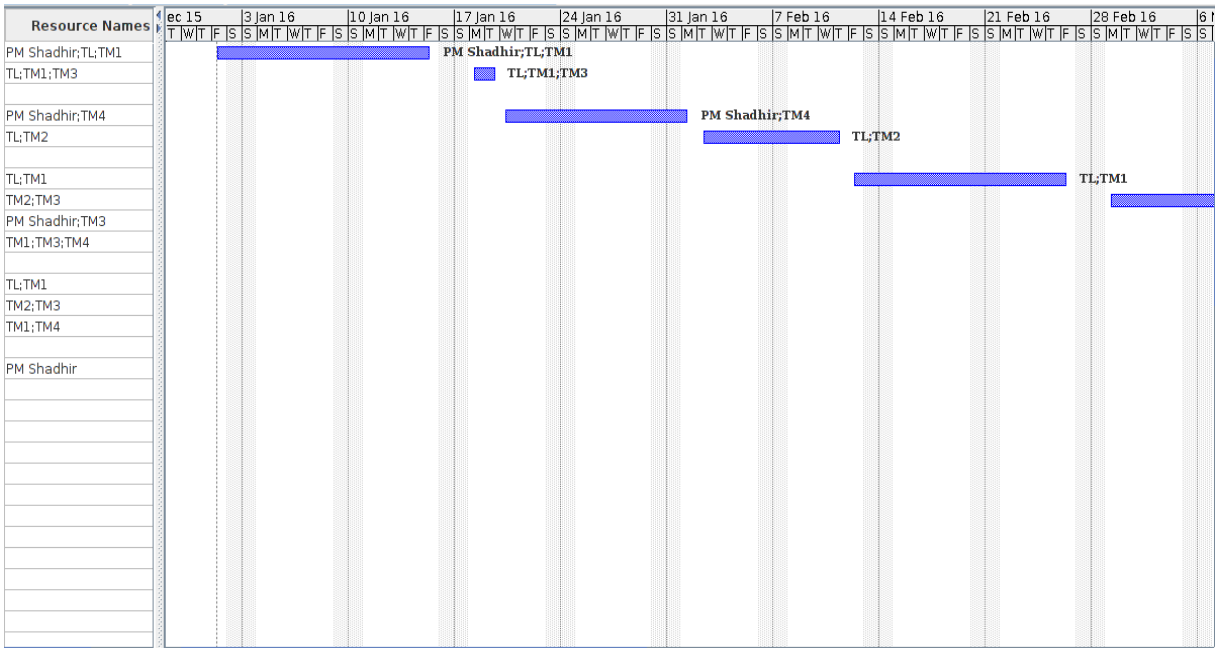

Figure 7.2: Time Line Chart

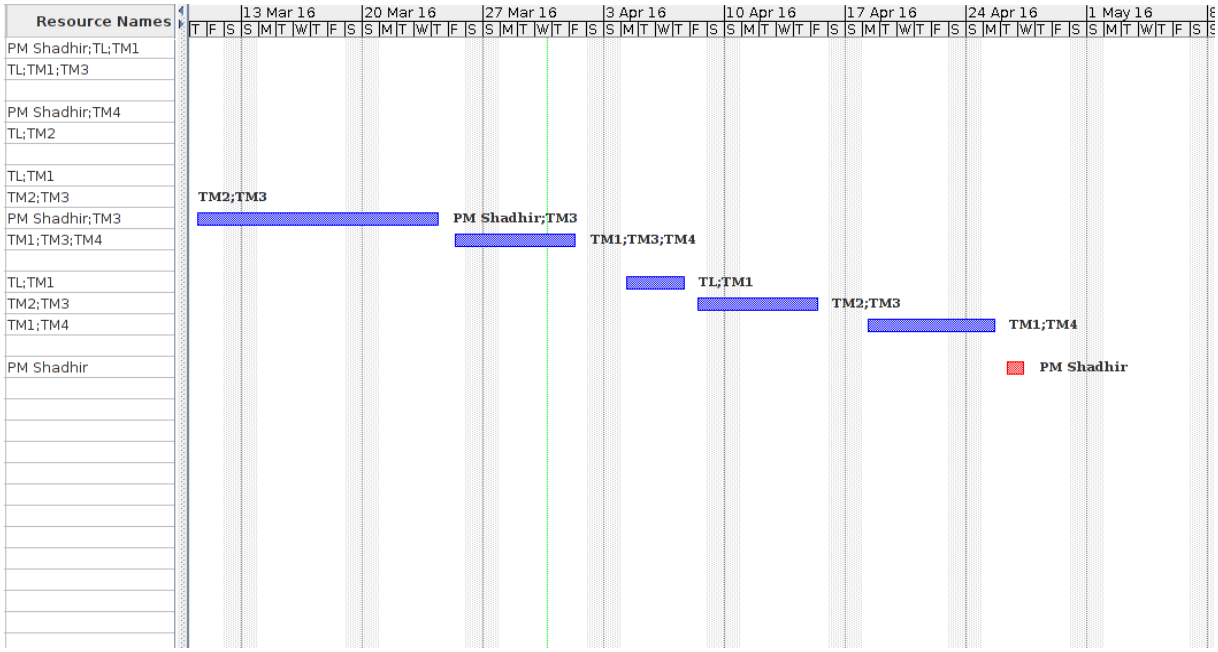

Figure 7.3: Time Line Chart

Chapter 8

Task Distribution

8.1 Distribution of Workload

8.1.1 Scheduled Working Activities

Activity	Time Period	Comment
Requirement Gathering	10 Days	Requirement gathering has took placed through searching on internet and taking the ideas, sharing the views among group members.
Planning	04 Days	Planing has done by Reviewing of literature of IEEE papers.
Design	04 Days	Designing has done by creating UML diagram, By creating Charts.
Implementation	90 Days	Implementation has done First creating the backend and then front end module by module.
Testing	10 Days	Testing has done by perfoming unit testing, alpha & Beta testing, integrated testing and system testing.
Deployment	05 Days	Deployment has done by installing project on the server.

Figure 8.1: Scheduled Working Activitiess of University Classification Prediction System

8.1.2 Members actvities or task

Member	Activity	Time Period	Start Date	End Date	Comment
M1, M2, M3, M4	Requirement Gathering	04 Days	01/01/16	05/01/16	M1 and M2 has perfomed the searching for project requirement on the internet by reviewing the related literature and by anlaysing the related project which is already available in the market. Regularly inform to the other member of team.
M1, M2, M3, M4	Analysis of the requirement	03 Days	06/01/16	09/01/16	M1, M2, M3, M4 done the requirement analysing of project by sharing the ideas, and by discussing on related information which is gather by the M1, And M2. M3 and M4 has created the list of requirement after every meeting.
M1, M2, M3, M4	Finalysing the requirement	01 Days	10/01/16	11/01/16	Whole team finalize the requirement. M1 and M4 has created a list of finalise requirement.
M1, M2, M3, M4	Planning	04 Days	12/01/16	16/01/16	Planning has done by walkthrough and by analysing the available product. M2 and M3 creates a list of function which will be implement in the project. Each and every module were discuss in every group meeting and M1 and M2 creates a blue print for project .

M1, M2	Front End design	04 Days	17/01/16	21/01/16	M1 and M2 creates the UML diagram for front end of the system and data flow diagrams and informed to the whole team regularly.
M3, M4	Back End design	04 Days	17/01/16	21/01/16	M3 and M4 creates the UML diagram for back end of the system and data flow diagrams and informed to the whole team regularly.
M3, M4	Installation of tools and technology for front end	01 Days	21/01/16	22/01/16	M3 and M4 installed the all the require tools which is use for front end design.
M1, M2	Installation of tools and technology for back end	02 Days	21/01/16	22/01/16	M3 and M4 installed all the require tools which is use for back end design.
M3, M4	Implementation of GUI	04 Days	23/01/16	27/01/16	M3 and M4 creates the GUI of the project using netbeans and informed to other member.
M1	Implementation of Naive Bayes Algorithm	20 Days	28/01/16	17/02/16	M1 implemented the Algorithm using WEKA and discuss on it with other team member

M1,M2	Navie Bayes Algorithm Connectivity with GUI	12 Days	18/02/16	02/03/16	M1 and M2 makes the connection among Algorithm and GUI. M1 and M2 Explain the codes to the other member of team.
M1,M2	Database connectivity	04 Days	02/03/16	06/03/16	M1 and M2 created the database connectivity with the WEKA.
M3,M4	GUI connectivity	04 Days	07/03/16	11/03/16	M3 and M4 created the connectivity GUI with database.
M3, M4	Data gathering into database	2 Days	12/03/16	14/03/16	M3 and M4 gather the data of university from different university ranking website.
M1, M2, M3, M4	Integration of all modules	10 Days	15/03/16	25/03/16	M1, M2,M3 and M4 integrated all the module. Implemented whole system properly.
M1,M2	Unit testing	2 Days	26/03/16	28/03/16	M1 and M2 performed the unit testing and noted down results and discuss with other member of team.
M3, M4	Functional testing	5 Days	18/03/16	28/03/16	M3 and M4 performed the functional testing and noted down results and discuss the result of testing with other member.
M1, M2, M3, M4	Deployment	05 Days	29/03/16	04/04/16	

Figure 8.2: Member Activities and Task

Chapter 9

Conclusion and Future Scope

9.1 Conclusion

In this project, we are trying to solve every students problem who wish to pursue higher education in foreign universities but donâ^{TM}t have any idea regarding which options he/she have with respect to his/her overall profile and also havenâ^{TM}t quite exposed to the procedure how students get admitted at foreign universities. Unlike Indian Universities, Marks isnâ^{TM}t the only criteria on which basis admission takes place, its all about what have you done in your college life except for attending lectures and sitting exams i.e extra curricular activities and Internships etc. we are mainly focusing on,providing the masses a software application which gives a total comprehensive solution for selecting the best universities according to their preferences their own parameters for their further education.

9.2 Future Scope

- Upon collecting the qualitative data derived from said interviews, careful analysis shall be done (both manually and utilizing our software) to prepare a SWOT (strengths, weaknesses, opportunities, and other aspects) of any university, then we will rank them using a combination of perfomance indicators.

- Since there is no such application build until now, we can embed many things like we can develop a mobile application which will integrate itself with the application server and provide the system in userâ^{TM}s mobile.

- In addition to foreign universities, we can also rank our Indian universities with respect to their own criteria.

- Get our software as a government authorised platform for favouring the students by providing our services.

- Include Voice recognition and Voice search for better user interaction with support for multiple languages.

- Build this application for every other current Operating systems including all flavours of linux and Mac ...

References

- http://www.shanghairanking.com/ARWU2013.
- http://www.webofknowledge.com
- http://www.highlycited.com
- http://www.topuniversities.com
- http://www.timeshighereducation.co.uk/world-university-rankings
- http://www.cwur.org
- http://admin-apps.webofknowledge.com/JCR/JCR

Bibliography

[1] U Dill, D. D., Soo, M. (2005).Academic quality, league tables, and public policy:*A cross-national analysis of university ranking systems* Higher Education, 49(4), 495-533. (Pubitemid 40965811)

[2] U Isidro F. Aguillo; Jose Luis Ortega; Mario Fernandez *"Webometric Ranking of World Universities: Introduction, Methodology, and Future Developments"*, Higher Education in Europe, Vol. 33, No. 2/3, July-October 2008

[3] A. M. Adegbore *"University faculty use of electronic resources: a review of the recent literature"*, PNLA Quarterly (The Officially Publications of the Pasific Northwest Library Association), vol. 75, no. 4, 2011

[4] H. Deng *"Emerging patterns and trends in utilizing electronic resources in a higher education environment: An empirical analysis"* New Library World, vol. 111, no. 3/4, pp.87 -103 2010

Chapter 10

Appendix I

10.1 What is NaiveBayes?

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes has been studied extensively since the 1950s. It was introduced under a different name into the text retrieval community in the early 1960s and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate preprocessing, it is competitive in this domain with more advanced methods including support vector machines. It also finds application in automatic medical diagnosis.

10.1.1 Why NaiveBayes is more Efficient than other classification algorithms?

Super simple, you're just doing a bunch of counts. If the NB conditional independence assumption actually holds, a Naive Bayes classifier will converge quicker than discriminative models like logistic regression, so you need less training data. And even if the NB assumption doesn't hold, a NB classifier still often performs surprisingly well in practice. A good bet if you want to do some kind of semi-supervised learning, or want something embarrassingly simple that performs pretty well.

- Naive Bayes is used a lot in robotics and computer vision, and does quite well with those tasks.
- You should train a large training set to use NB well. This would allow you to better assume that samples are "independent" of each other.
- generally no requirements, good for few categories variables, compute the multiplication of

independent distributions, suffer multicollinearity.

10.2 WEKA

Waikato Environment for Knowledge Analysis (Weka) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License.

Weka (pronounced to rhyme with Mecca) is a workbench that contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. The original non-Java version of Weka was a Tcl/Tk front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and a Makefile-based system for running machine learning experiments.

10.2.1 Features of WEKA

- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- Ease of use due to its graphical user interfaces.
- A comprehensive collection of data preprocessing and modeling techniques.
- It supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection
- Java Library can be imported and used it's functionality in other java program
- provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query.
- Free availability under the GNU General Public License.

ACKNOWLEDGMENT

We would like to take the opportunity to express our sincere thanks to our guide **Prof. Javed Sheikh**, Assistant Professor, Department of Computer Engineering, AIKTC, School of Engineering, Panvel for his invaluable support and guidance throughout our project research work. Without his kind guidance & support this was not possible.

We are grateful to him for his timely feedback which helped us track and schedule the process effectively. His time, ideas and encouragement that he gave his help us to complete our project efficiently.

We would also like to thank **Dr. Abdul Razak Honnutagi**, AIKTC, Panvel, for his encouragement and for providing an outstanding academic environment, also for providing the adequate facilities.

We are thankful to **Prof. Tabrez Khan**, HOD, Department of Computer Engineering, AIKTC, School of Engineering, Panvel and all my B.E. teachers for providing advice and valuable guidance.

We also extend our sincere thanks to all the faculty members and the non-teaching staff and friends for their cooperation.

Last but not the least, We are thankful to all our family members whose constant support and encouragement in every aspect helped us to complete our project.

Mohammed Shadir Fakir Mohammed Fatima (12CO44)

Ashiq Ahamed Abdul Subhan RehamanBee (12CO23)

Baig Mannan Rehman Sabira (12CO24)

Chapter 10. Appendix I

Shaikh Abdul Alim JamilurRehman Sufiyanas (12CO57)

(Department of Computer Engineering)

University of Mumbai.