SAVITRIBAI PHULE PUNE UNIVERSITY

A PROJECT REPORT ON

StudentCareerMap: A Learning based Student Performance and career Analyzer

SUBMITTED TOWARDS THE
PARTIAL FULFILLMENT OF THE REQUIREMENTS OF

BACHELOR OF ENGINEERING (Computer Engineering)

BY

Shivani Bhosale	B120334230
Pranjal Nimse	B120334334
Siddhi Wadgaonkar	B120334403
Aishwarya Yeole	B120334411

Under The Guidance of

Dr. Pravin Futane

DEPARTMENT OF COMPUTER ENGINEERING
Pimpri Chinchwad College Of Engineering
Sector-26, Pradhikaran, Nigdi, Near Akurdi Railway Station, Pune,
Maharashtra 411044

Pimpri Chinchwad College Of Engineering
DEPARTMENT OF COMPUTER ENGINEERING

CERTIFICATE

This is to certify that the Project Entitled

StudentCareerMap: A Learning based Student Performance and career Analyzer

Submitted by

Shivani Bhosale	B120334230
Pranjal Nimse	B120334334
Siddhi Wadgaonkar	B120334403
Aishwarya Yeole	B120334411

is a bonafide work carried out by Students under the supervision of Dr. Pravin Futane and it is submitted towards the partial fulfillment of the requirement of Bachelor of Engineering (Computer Engineering).

Dr. Pravin Futane
Internal Guide
Dept. of Computer Engg.

Dr. K. Rajeswari
H.O.D
Dept. of Computer Engg.

Dr. A. M. Fulambarkar

Principal

Pimpri Chinchwad College Of Engineering

PROJECT APPROVAL SHEET

A Project Title

StudentCareerMap: A Learning based Student Performance and career Analyzer

Is successfully completed by

Shivani Bhosale	B120334230
Pranjal Nimse	B120334334
Siddhi Wadgaonkar	B120334403
Aishwarya Yeole	B120334411

at

DEPARTMENT OF COMPUTER ENGINEERING

(Pimpri Chinchwad College Of Engineering)

SAVITRIBAI PHULE PUNE UNIVERSITY,PUNE

ACADEMIC YEAR 2016-2017

Dr. Pravin Futane	Dr. R. K. Rajeshwari
Internal Guide	H.O.D
Dept. of Computer Engg.	Dept. of Computer Engg.

Abstract

Considering the competition in today's world, getting a job is a very crucial aspect of every engineering student. Every institute wants to improve their placement records every year. Placement of a student in a particular type of company is dependent on many aspects such as academic performance , extra-curricular activities, projects etc. Also the institute takes certain actions that help students in the placement activities. The system is aimed at analysing the different attributes on which a student's placement depends. The target audience of this project is final year students and faculty of Computer Dept, Pimpri Chinchwad College of Engineering. Random Forest approach is used for predicting the type of company the student would most probably get placed in. Depending on the this, the faculties and TPO can take certain actions , especially for those aspects that are strongly corelated with placement. More efforts in these fields will help in better placement results. This system has taken input in form of categories for different attributes. It then predicts the class label (Company type) for all unlabelled tuples. Also dynamic class label prediction is possible by entering the tuple details. Visualization in form of reports and graphs is available as a part of output.

Acknowledgments

It gives us great pleasure in presenting the project report on ***'StudentCareerMap: A Learning based Student Performance and career Analyzer'****.*

We would like to take this opportunity to thank our internal guide ***Dr. Pravin Futane*** *for giving us all the help and guidance we needed. we really grateful to them for their kind support. Their valuable suggestions were very helpful.*

we are also grateful to ***Dr. K. Rajeswari****, Head of Computer Engineering Department, Pimpri Chinchwad College of Engineering for her indispensable support, suggestions.*

In the end our special thanks to ***non-teaching staffs*** *for providing various resources such as laboratory with all needed software platforms, continuous Internet connection, for Our Project.*

Shivani Bhosale
Pranjal Nimse
Siddhi Wadgaonkar
Aishwarya Yeole
(B.E. Computer Engg.)

INDEX

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION TO STUDENT CAREER PREDICTION SYSTEM

1.1 WHAT ARE PREDICTION SYSTEMS ?

A prediction is a statement about an uncertain event. It is often, but not always, based upon experience or knowledge. There is no universal agreement about the exact difference between the two terms; different authors and disciplines ascribe different connotations. A prediction system predicts the future occurance of an event. Prediction of an event requires vague, imperfect and uncertain knowledge.Complexity in a prediction system is its intrinsic characteristic. Predicting a system is usually done by learning from the past for which historical data is obtained and analyzed to study the resulting pattern in the market. Predicting any event requires knowledge about past performance. Data from the past is used mainly to learn the patterns that existed. Historical data provides information on the specific pattern of learning the data. Learning from the past provides knowledge about future to some extent.

1.2 WHY USE PREDICTION SYSTEMS ?

Earlier, prediction systems were built with rules formed manually. Rules became complicated with increase in the number of inputs and predicting an event grew tedious. Engineering helps build a prediction system that could adapt to the increasing number of inputs and frame rules accordingly. The accuracy and speed obtained is superior to manual prediction schemes.

1.3 DIFFERENT PREDICTION ALGORITHMS :

The following displays different algorithms used in prediction systems alongwith the percentage accuracy obtained by testing it on data-set.

- Naive bayes - 68.5
- Regression trees - 71.3
- Random forest algorithm - 77.6
- Support vector machines - 73.7

1.4 PROBLEM DEFINITION

To implement an application that analyses students' overall data for placement prediction and gives analysis output in form of reports and graphs so that relevant actions can be taken.

1.5 JUSTIFICATION OF PROBLEM

The predicted system takes into consideration the overall data for prediction of placement company category and uses the most efficient algorithm for prediction, hence almost accurate results are predicted.

1.6 NEED OF PROBLEM

Existing career prediction system consists of prediction depending on few parameters such as overall performance, psychometric tests etc. The proposed system takes into consideration various parameters such as academic performance, technical skills and courses, volunteering activities, coding skills. Considering all these parameters the accuracy of the system can be increased. Hence the need to build such a system.

1.7 PURPOSE OF YOUR SYSTEM

StudentCareerMap - Our system aims at supporting the Department Faculties and the Training and Placement Officer in making relevant decisions with respect to placement activities and preparation. It will analyse the previous data of students (Academic Performance , Extra-curricular activities , Volunteering , Technical Courses attended etc.) and derive relationship between these attributes and Placement (class Label). Various reports, graphs etc. will make it easy to visualize the results of analysis. The system can also predict the Type of Company (class label) for a user-entered tuple.

1.8 LITERATURE SURVEY

- **”Performance Analysis of Undergraduate Students Placement Selection using Decision Tree Algorithms by T. Jeevalatha, N. Ananthi, D. Saravana Kumar”**

  Data mining is a new approach for education. The main objectives of higher education institutions are to provide quality education to its students for their better placement opportunity. We could use Decision tree algorithms to predict student selection in placement. It helps us to identify the dropouts of the student who need special attention and allow the teacher to provide appropriate placement training. This paper describes how the different Decision tree algorithms used to predict student performance in placement. In the first step we have gathered the last two years passed out students details from placement cell in Dr.N.G.P Arts and Science College. In the second step preprocessing was done on those data and attributes were selected for prediction and in the third step Decision tree algorithms such as ID3, CHAID, and C4.5 were implemented by using Rapid Miner tool. Validation is checked for the three algorithms and accuracy is found for them. The best algorithm based on the collected placement data is ID3 with an accuracy of 95.33

- **”Use of ID3 Decision Tree Algorithm for Placement Prediction. Hitarthi Bhatt 1 , Shraddha Mehta 2 , Lynette R. D’mello 3 Dept. of Computer Engineering, D.J. Sanghvi College of Engineering, Mumbai University Mumbai, India”**

  Every year corporate companies come to colleges in order to recruit students. Recruitment is one of the most essential processes for any organization as they look for skilled and qualified professionals to fill up the positions in their organization. Many companies hire students through campus recruitment process. Campus recruitment is an efficient way to get the right resources at the right time with minimal cost and within minimum time frame. While the industry hires candidates from different institutes, students too get a chance to start their career with some of the best companies in the corporate sector. The main aim of this paper is to identify relevant attributes based on quantitative and qualitative aspects of a student's profile such as CGPA, academic performance, technical and communication skills and design a model which can predict the placement of a student. For this purpose ID3 classification technique based on decision tree has been used. The result of this analysis will assist the academic planners to design a strategy to improve the performance of students that will help them in getting placed at the earliest.

- **"Data mining for building an informed decision making model for career prediction, by S. Sathyavathi, N.Niraimathi, K.Priyadarshini"**

  The impact of the credits on the career choice is determined by using mining tools classification (ID3, CHAID). The steps performed for mining the data is classification. The tool used is Rapid miner. The algorithm used for comparing the datasets using classification are ID3, CHAID, Decision tree are used for classification. The performance of the algorithm are compared and it was found that Decision tree provide the maximum accuracy for classifying the fittings prediction based on the skills.

CHAPTER 2

ANALYSIS OF STUDENT CAREER PREDICTION SYSTEM

2.1 PROJECT PLAN

2.1.1 Project Estimates

2.1.1.1 Reconciled Estimates

- Time Estimates Accurate time estimation is a skill essential for good project management. It is important to get time estimates right for two main reasons:
  - Time estimates drive the setting of deadlines for delivery and planning of projects, and hence will impact on other people assessment of your reliability and competence as a project manager.
  - Time estimates often determine the pricing of contracts and hence the Dritability of the contract/project in commercial terms.

Table 2.1: Time estimation

Sr. No.	Period	Estimated Time
1	Definition	1 month
2	Concept	2 month
3	Design	4 months
4	Technical Implementation	2.5 months
5	Testing	1.5 months

2.1.2 Project Resources

- Project Team Members : 4
- Hardware Resources: : Laptop or PC with suitable RAM and processor on which the software is deployed .
- Software Resources :
  Operating system: Windows or Ubuntu
  IDE: Netbeans IDE

2.2 MATHEMATICAL MODELLING

2.2.1 Goals and Objectives

The main objective of this project is to support faculties and TPO in taking certain decisions related to placement activities and preparation. The class label of prediction is the type of company that the student is most likely to get placed in .The system should predict the class labels with maximum accuracy possible. This will help the faculties to take some extra efforts on the poor students or give special courses for students who have more potential to get placed in good companies. Also, it will help the TPO to decide what kind of companies to invite based on students caliber and also take certain actions to increase the potential of students.

2.2.2 Statement of scope

The scope of this project is limited to the students of Computer Department , Pimpri Chinchwad College of Engineering (2017 batch). We have collected the students' data for all 4 years of engineering tenure. Since, other institutes have different placement scenarios and policies, this system will be accurate mostly for PCCOE.

2.2.3 Major Constraints

- Redundant data.
- Incomplete data i.e missing attribute values should be ignored.
- Heterogeneous data.
- Real time and consistent data to be used as database.
- Ensure that accurate data is provided as training data-set.
- User friendly GUI should be created.

2.2.4 Methodologies of Problem solving and efficiency issues

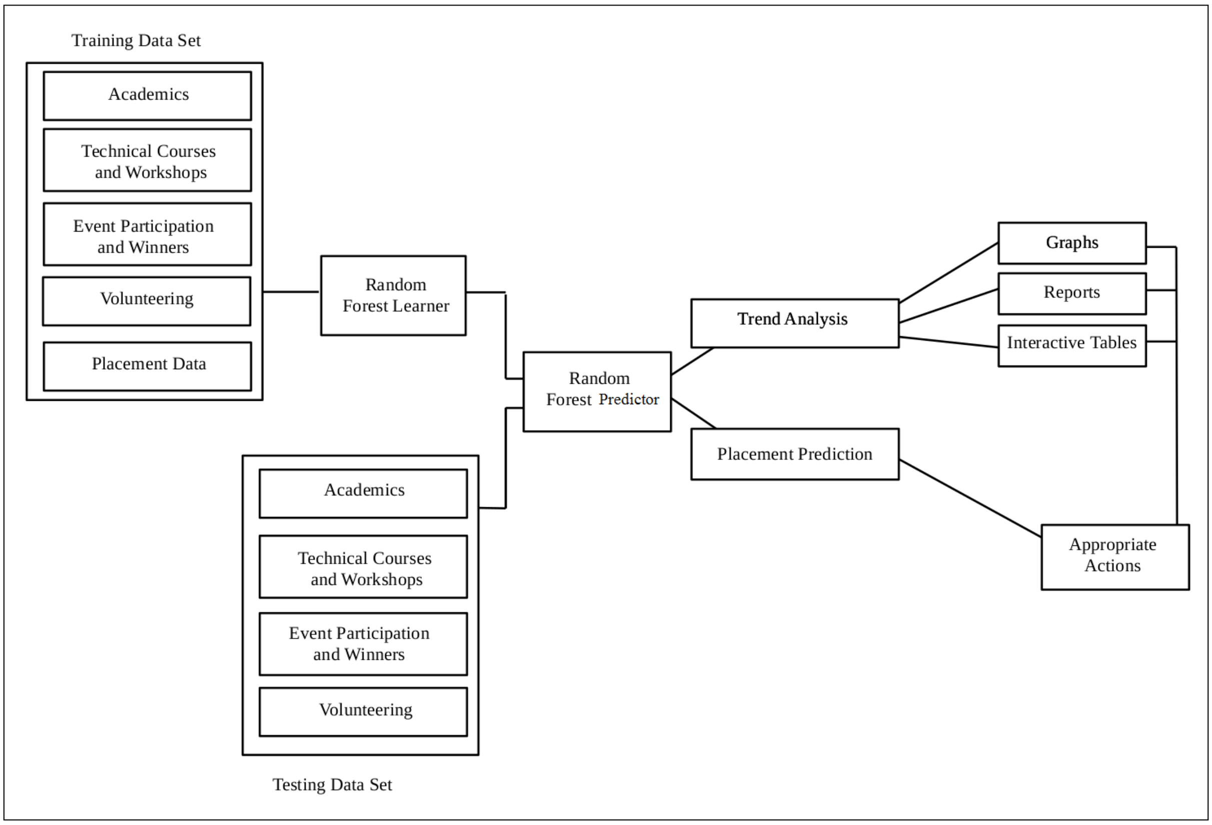


Figure 2.1: System Diagram

Figure 2.1 depicts the basic system architecture. The training data set has the following attributes : Academics , Technical Courses and Workshops , Event Participation and Winners , Volunteering details , Placement Data (class label). This training data set is given as input to the Random Forest Learner using which the system learns and trains itself. Then, the testing data set is given as input to the Random Forest Predictor which actually predicts the class label for testing data. Also, Trend analysis is one aspect of the output which will represent the results in form of graphs , reports etc. Finally, the output will help the faculties and the TPO to take appropriate actions in order to improvise the Placement.

Efficiency : As the main subject (domain) of our project are Students and their performance , it varies from every student to student. So, it is very difficult to get very high accuracy like 90% . Also, there are a few other factors that also affect the class label (Placement) that cannot be considered as it cannot be quantified. Currently, there is 78% accuracy. As the system learns more with increased number of tuples, it will be able to classify more accurately.

2.2.5 Outcome

Outcome of this system is prediction of type of company the student would most probably get placed in.

2.2.6 Applications

This project aims at analysing the different attributes on which a student's placement depends. It aims at predicting the type of company the student would most probably get placed in. Depending on the this, the faculties and TPO can take certain actions , especially for those aspects that matter the placement the most.

2.3 REQUIREMENT ANALYSIS

2.4 ADVANCES/ADDITIONS/UPDATING THE PREVIOUS SYSTEM

Existing career prediction system consists of prediction depending on few parameters such as overall performance, psychometric tests etc. The proposed system takes into consideration various parameters such as academic performance, technical skills and courses, volunteering activities, coding skills. Considering all these parameters the accuracy of the system can be increased.

2.5 TEAM ORGANIZATION

2.5.1 Team Structure

As all of us are immature to project management software engineering. Mutual discussion and coordination were essential aspects of democratic team structure were best suited in order give equal opportunities and bring best things out.

Table 2.2: Roles Description

Role	**Expected output**
Coders	Number of functions coded.
Designers	Designing and modeling of system.
Analyzers	Requirement gathering analysis.
Testers	Testing system functionality.

Table 2.3: Roles played by team members

Team Member	**Role Played**
Shivani Bhosale	Designer, Tester, Coder and Analyzer
Pranjal Nimse	Designer, Tester, Coder and Analyzer
Siddhi Wadgaonkar	Designer, Coder and Analyzer
Aishwarya Yeole	Designer, Coder and Analyzer

CHAPTER 3

DESIGN OF STUDENT CAREER PREDICTION SYSTEM

3.1 SOFTWARE REQUIREMENT SPECIFICATION

3.1.1 Problem Statement

StudentCareerMap: A Learning based Student Performance and career Analyzer

3.1.2 Problem definition

To implement a system This system is designed to support faculties and TPO in taking certain decisions related to placement activities and preparation. The main aim of the project is to identify the relationship between certain performance attributes of students with their placement. The system needs a complete dataset for training

3.1.3 External interface requirement

User Interface: This will be a desktop application and user will communicate with it through GUI for predicting the career path of student by entering its details. Graphs, Tables and other visualizing figures will also be used in analysis part for communication with user.

- Hardware Interface: This application requires 64- bit Operating System of Processor speed: 1.5 GHz.

- Software Interface: The application will be a desktop application and does not require internet.

3.2 SYSTEM FEATURES

3.2.1 System diagram

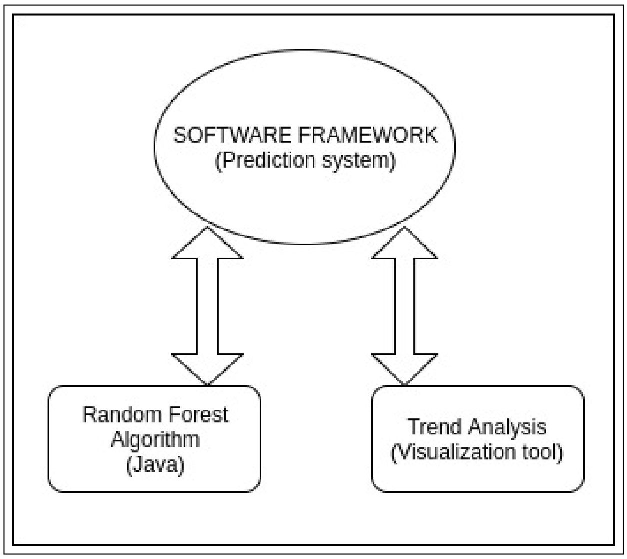


Figure 3.1: System Diagram

3.2.2 The main prediction system

- Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
- Random decision forests correct for decision trees' habit of over-fitting to their training set.
- Features of Random forests :
  - It is unexcelled in accuracy among current algorithms.
  - It runs efficiently on large data bases.
  - It can handle thousands of input variables without variable deletion.
  - It gives estimates of what variables are important in the classification.

- It generates an internal unbiased estimate of the generalization error as the forest building progresses.
- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.
- It has methods for balancing error in class population unbalanced data sets.
- Generated forests can be saved for future use on other data.
- Prototypes are computed that give information about the relation between the variables and the classification.
- It computes proximities between pairs of cases that can be used in clustering, locating outliers, or (by scaling) give interesting views of the data.
- The capabilities of the above can be extended to unlabeled data, leading to unsupervised clustering, data views and outlier detection.
- It offers an experimental method for detecting variable interactions.

- Random forests are among the most popular machine learning methods thanks to their relatively good accuracy, robustness and ease of use.

- They also provide two straightforward methods for feature selection: mean decrease impurity and mean decrease accuracy.

- Mean decrease impurity :

  - Random forest consists of a number of decision trees.
  - Every node in the decision trees is a condition on a single feature, designed to split the data-set into two so that similar response values end up in the same set.
  - The measure based on which the (locally) optimal condition is chosen is called impurity.
  - For classification, it is typically either Gini impurity or information gain/entropy and for regression trees it is variance. Thus when training a tree, it can

be computed how much each feature decreases the weighted impurity in a tree.

  - For a forest, the impurity decrease from each feature can be averaged and the features are ranked according to this measure.

3.2.3 Trend analysis on data

- Trend analysis is the rampant practice of collecting information and attempting to spot a pattern.

- A trend analysis is an aspect of technical analysis that tries to predict the future movement of a stock based on past data.

- Trend analysis is based on the idea that what has happened in the past gives traders an idea of what will happen in the future. There are three main types of trends: short-, intermediate- and long-term.

- A trend is the general direction the market is taking during a specified period of time.

- Trends can be both upward and downward, relating to bullish and bearish markets, respectively.

- While there is no specified minimum amount of time required for a direction to be considered a trend, the longer the direction is maintained, the more notable the trend.

- Trend analysis is the process of trying to look at current trends in order to predict future ones and is considered a form of comparative analysis.

- Although trend analysis is often used to predict future events, it could be used to estimate uncertain events in the past.

- In project management, trend analysis is a mathematical technique that uses historical results to predict future outcome. This is achieved by tracking variances in cost and schedule performance.

- Trend analysis can be done using graphs or charts. A graph can be thought of as a ”picture” of data. Graphs are useful because they can reveal patterns or trends and can be used to make predictions. You can analyze and evaluate data found in graphs to make predictions.

3.3 OTHER NON- FUNCTIONAL REQUIREMENTS

3.3.1 Performance requirements

- System should train the algorithm using an accurate real time data-set for proper results.
- Time and space complexity should be less.
- Facility to provide data manually for any random values with exception handling.
- User friendly GUI.
- Predicted results should be relevant.

3.3.2 Safety requirements

- The application is designed in modules where errors can be detected and fixed easily and can be reduced over the time as dataset increases.

3.3.3 Software Quality attributes

Our software has many quality attribute that are given below:-

- Adaptability: This software is adaptable by all users as per their use and requirement.
- Availability: This software is easily available to all its audience.
- Maintainability: After the deployment of the project if any error occurs then it can be easily maintained by the software developer.

- Reliability: The performance of the software is better as the system will learn which will increase the reliability of the Software.
- User Friendliness: Since, the software is a GUI application; the output generated is much user friendly in its behavior.
- Integrity: Integrity refers to the extent to which access to software or data by unauthorized persons can be controlled.
- Security: Authentication is done for users to use the system as well as for the developers so that the integrity of the system and importantly dataset can be maintained.
- Testability: The software will be tested considering all the aspects. Time and space complexity should be less.

3.4 RISK ASSESSMENT

3.4.1 Risk Management w.r.t. NP Hard Analysis

This section discusses project risks and the approach to manage them.

3.4.2 Risk Identification

For risks identification, review of scope document, requirements specifications and schedule is done. Answers to questionnaire revealed some risks.

1. Have top software and customer managers formally committed to support the project?
   -Yes
2. Are end-users enthusiastically committed to the project and the system/product to be built?
   -Yes
3. Are requirements fully understood by the software engineering team and its customers?
   -Yes

4. Have customers been involved fully in the definition of requirements?

   -No

5. Do end-users have realistic expectations?

   -Yes

6. Does the software engineering team have the right mix of skills?

   -Moderate

7. Are project requirements stable?

   -Yes

8. Is the number of people on the project team adequate to do the job?

   -Yes

9. Do all customer/user constituencies agree on the importance of the project and on the requirements for the system/product to be built?

   -Yes

3.4.3 Risk Analysis

The risks for the Project can be analyzed within the constraints of time and quality.

Table 3.1: Risk Table

ID	**Risk Description**	**Probability**	**Impact**		
			Schedule	**Quality**	**Overall**
1	Obtaining Quality product from the project.	Medium	Medium	Medium	High
2	Accurate prediction of one's career depending on different attributes	Medium	Medium	Medium	Medium
3	Accurate data used for training model	High	Medium	High	High

Table 3.2: Risk probability description

Probability	**Value**	**Description**
High	Probability of occurrence is	$> 75\%$
Medium	Probability of occurrence is	$26-75\%$
Low	Probability of occurrence is	$< 25\%$

Table 3.3: Risk Impact definitions

Impact	**Value**	**Description**
Very high	$> 10\%$	Schedule impact or Unacceptable quality
High	$5-10\%$	Schedule impact or Some parts of the project have low quality
Medium	$< 5\%$	Schedule impact or Barely noticeable degradation in quality Low Impact on schedule or Quality can be incorporated

3.4.4 Overview of Risk Mitigation, Monitoring, Management

1. Obtaining Quality product from project

2. Accurate prediction of one's career depending on different attributes Our project considers the different attributes like academic scores, performance in extra-curricular activities, courses undertaken, certifications received etc. to predict the career of a student. It uses a training dataset of students who are already placed in different companies.

Following are the details for each risk.

Table 3.4: Risk ID 1

Risk ID	1
Risk Description	Obtaining Quality product from project
Category	Development Environment
Source	Software requirement Specification document.
Probability	Moderate
Impact	High
Response	Mitigate
Strategy	Successful implementation and publicity of project to make it a quality product
Risk Status	Identified

Table 3.5: Risk ID 2

Risk ID	2
Risk Description	Accurate prediction of one's career depending on different attributes
Category	Development Environment
Source	Software Design Specification documentation review.
Probability	Moderate
Impact	Medium
Response	Mitigate
Strategy	Use of a prediction algorithm with high accuracy
Risk Status	Identified

Table 3.6: Risk ID 3

Risk ID	3
Risk Description	Accurate data used for training model.
Category	Requirements
Source	Software Design Specification documentation review.
Probability	High
Impact	High
Response	Mitigate
Strategy	Collection of data from relevant sources.
Risk Status	Identified

CHAPTER 4

MODELLING OF STUDENT CAREER PREDICTION SYSTEM

4.1 UML DIAGRAM

4.1.1 Use-cases

The below table describes the necessary use cases:

Table 4.1: Use Case Description

Sr No.	Use Case	Description	Actors	Assumptions
1	Use Case 1	Add Data and Pre-process	Lab Assistant	Data in required format
2	Use Case 2	Predict for a real time query	Dept Faculties	Correctly entered fields
3	Use Case 3	View Analysis	Dept Faculties	
4	Use Case 4	View results for unlabelled tuples	TPO	Training data is correct
5	Use Case 5	Suggest or take actions	Dept Faculties & TPO	Correct Results

4.1.2 Use Case View

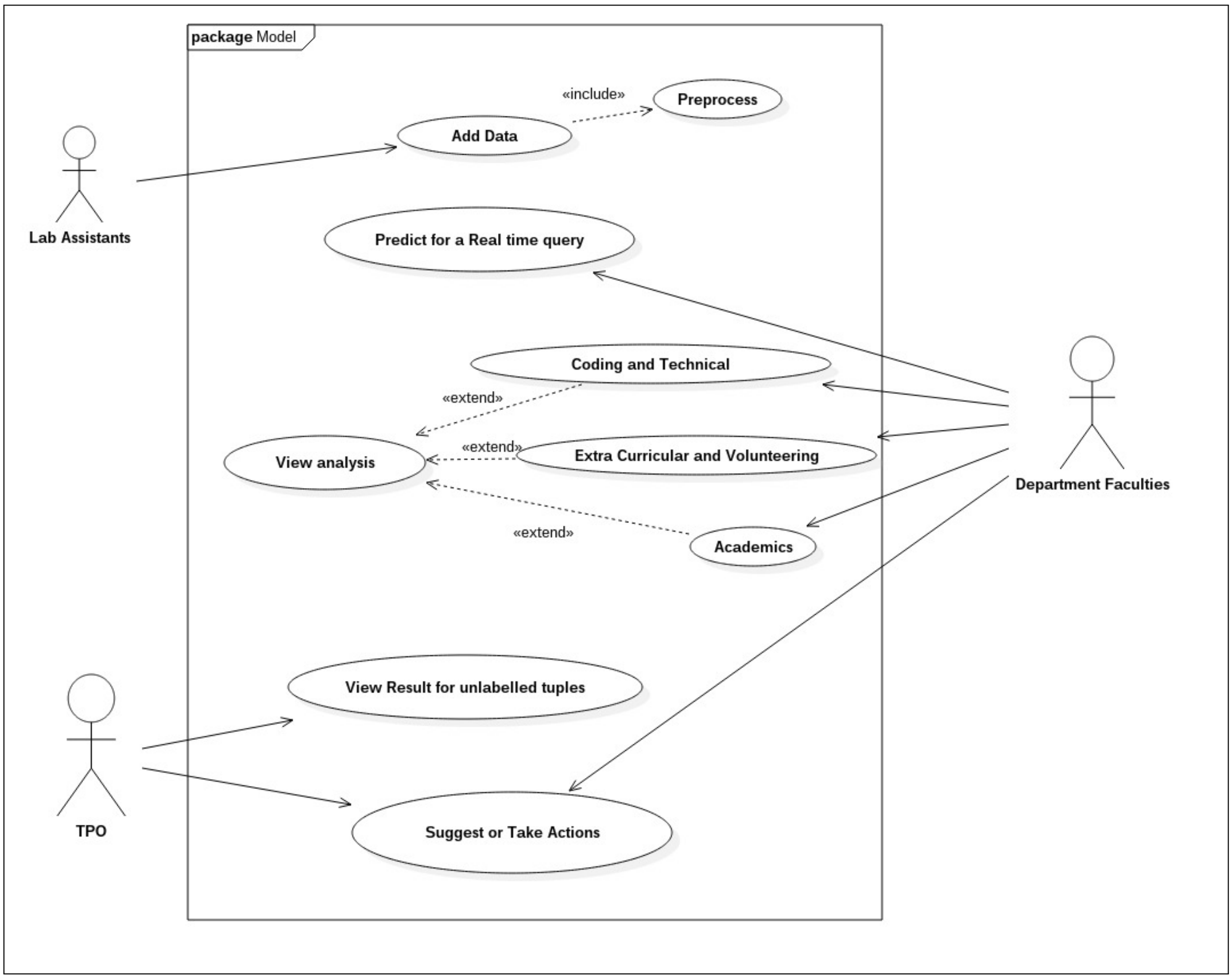


Figure 4.1: Use Case Diagram

Figure 4.1 shows the main actors of the system along with the use cases. There are three main actors that will interact with the system. Lab assistants will be responsible for the timely updation of the database. Department Faculties will interact with the system to see the overall results of the analysis, view result for a particular student or view some patterns etc. The TPO interacts with the system to check what factors are most important for the placement and will take necessary actions.

4.1.3 Class Diagram

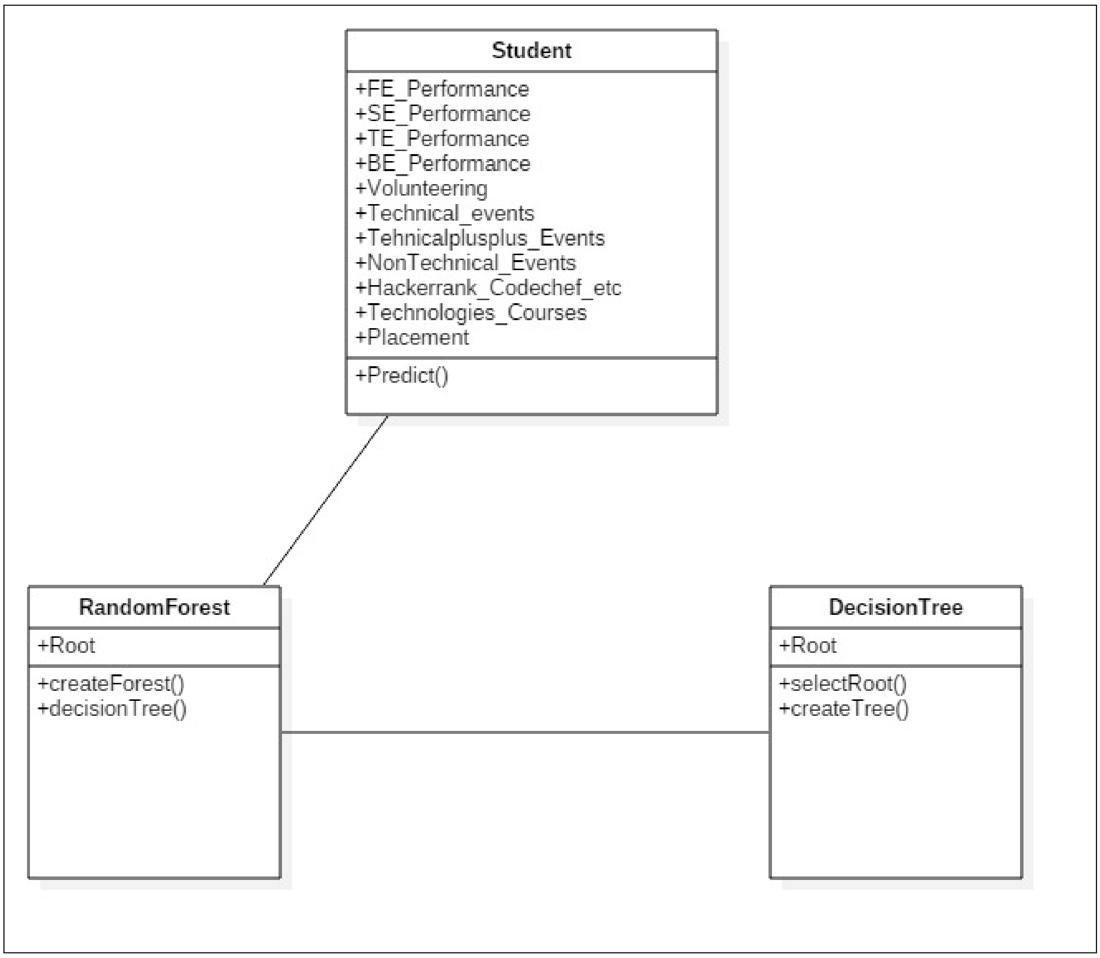


Figure 4.2: Class Diagram

Various classes of the system are depicted in Figure 4.2 . The three major classes are : Student class that contains all the performance attributes of every student . The Random Forest class has all the information related to RandomForest algorithm which in turn calls the DecisionTree algorithm several times. DecisionTree has actual implementation of of the DecisionTree algorithm.

4.1.4 Activity Diagram

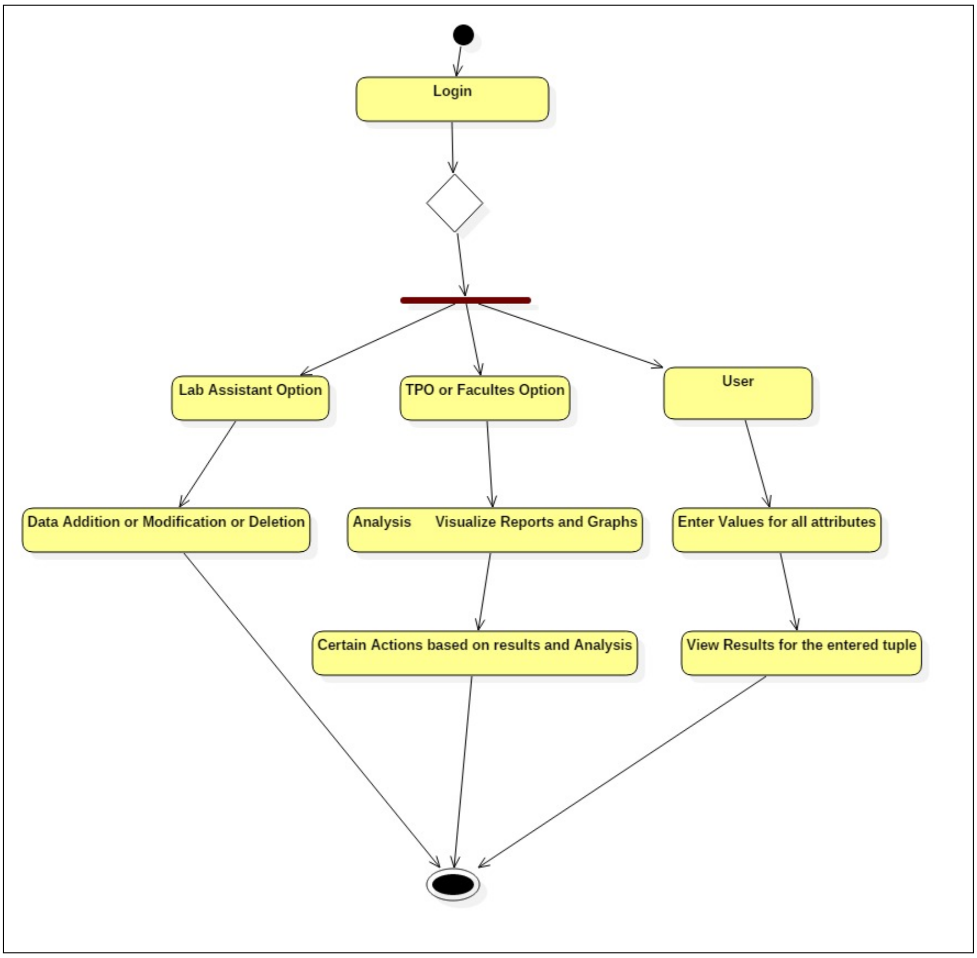

Login
Lab Assistant Option
TPO or Facultes Option
User
Data Addition or Modification or Deletion
Analysis Visualize Reports and Graphs
Enter Values for all attributes
Certain Actions based on results and Analysis
View Results for the entered tuple

Figure 4.3: Activity Diagram

Figure 4.3 shows the Activity diagram for the system. It shows the basic flow from one activity to the other. Here, the system can take either of the three paths depending upon the user who is using the system. Accordingly, the features and actions will vary.

4.1.5 Sequence Diagram

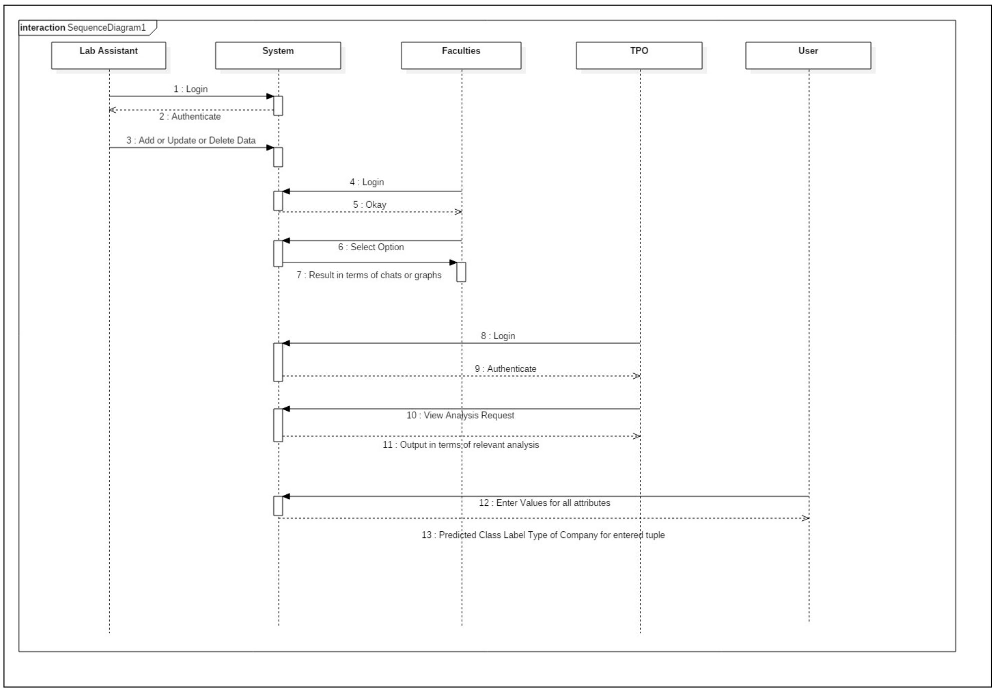

interaction SequenceDiagram1
Lab Assistant
System
Faculties
TPO
User
1 : Login
2 : Authenticate
3 : Add or Update or Delete Data
4 : Login
5 : Okay
6 : Select Option
7 : Result in terms of chats or graphs
8 : Login
9 : Authenticate
10 : View Analysis Request
11 : Output in terms of relevant analysis
12 : Enter Values for all attributes
13 : Predicted Class Label Type of Company for entered tuple

Figure 4.4: Sequence Diagram

Figure 4.4 depicts the basic interactions with the system in the form of Sequence diagram. The main lifelines are System , Faculties , TPO , Lab Assistants.

4.1.6 State Diagram

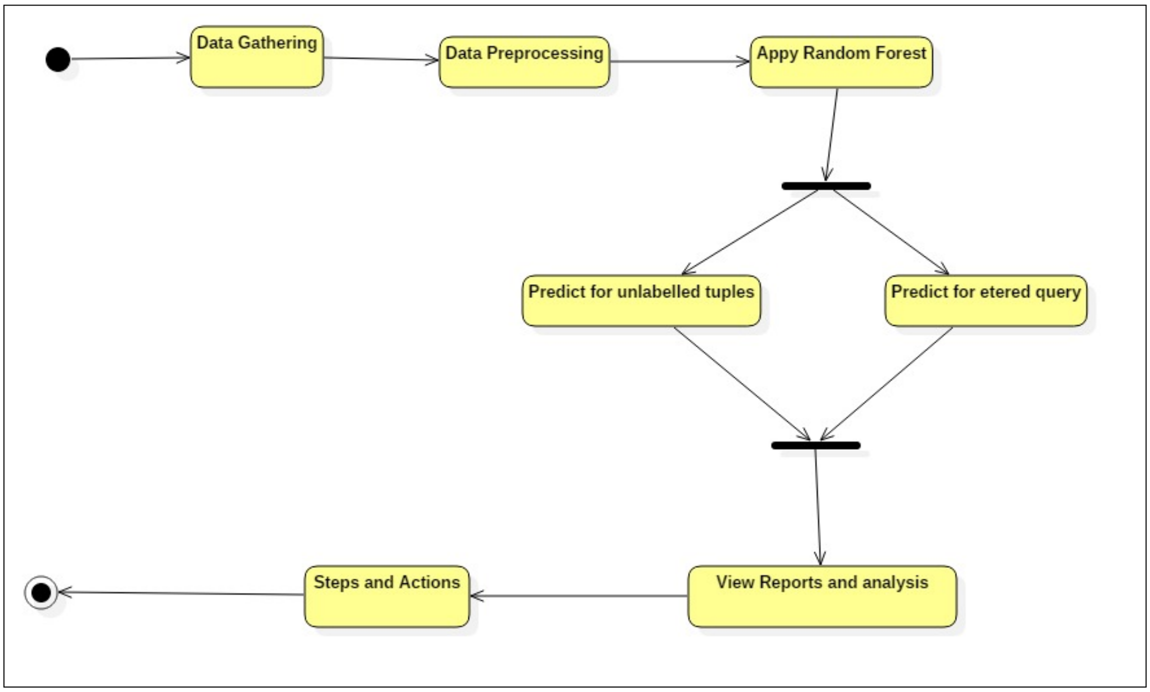


Figure 4.5: State Diagram

Figure 4.5 illustrates the different states ,the system can take, in form of State Chart diagram. Once data is ready , the system can be used for predicting the class labels for either already existing unlabelled tuples or user-entered queries. Then it will go in the Visualization state where graphical representation of results is shown.

4.2 ENTITY RELATIONSHIP DIAGRAM

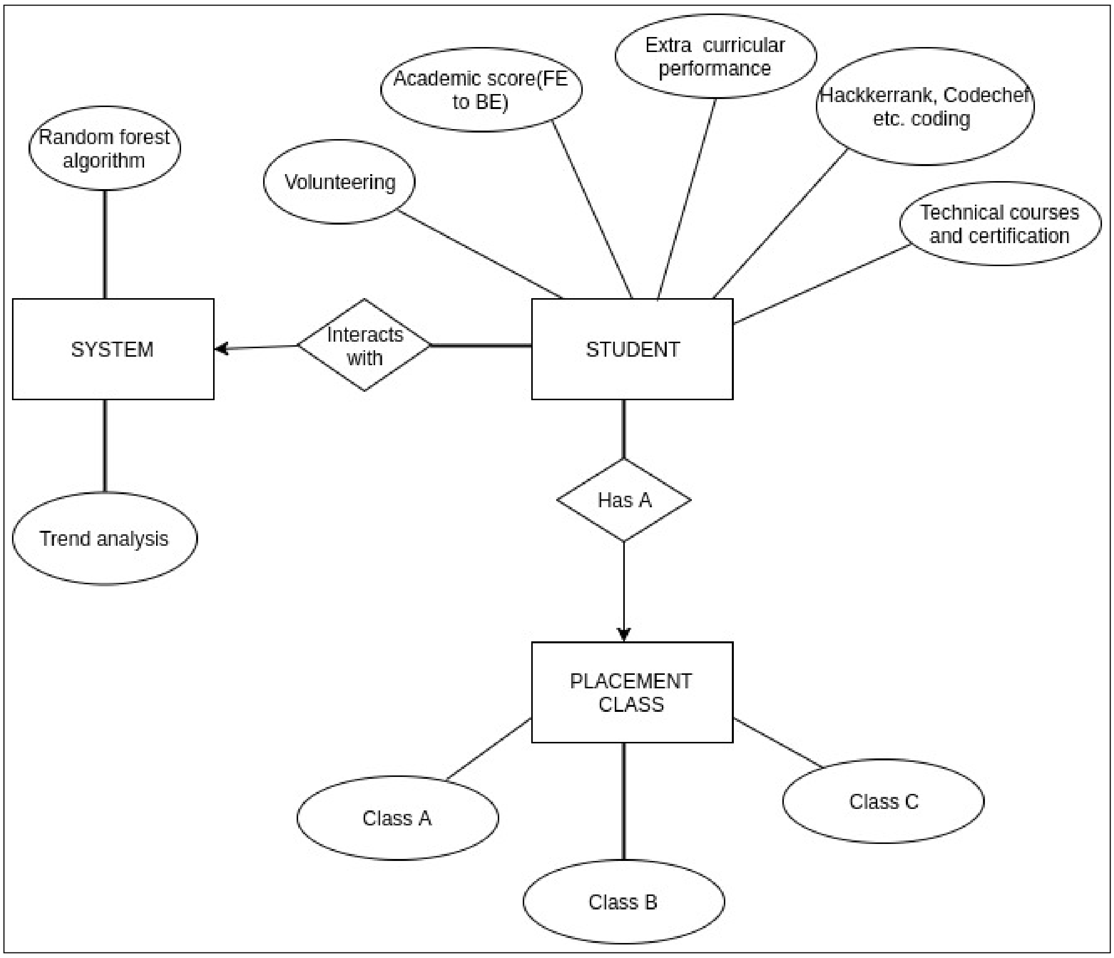


Figure 4.6: ER Diagram

Figure 4.6 illustrates the Entity-Relationship diagram of the system.

CHAPTER 5

CODING

5.1 ALGORITHMS / FLOWCHARTS

5.1.1 Random Forest Algorithm

Random forest is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees. Each decision tree is constructed using the following algorithm:

- Let the number of training cases be N, and the number of variables in the classifier be M.
- We are told the number m of input variables to be used to determine the decision at a node of the tree; m should be much less than M.
- Choose a training set for this tree by choosing n times with replacement from all N available training cases (i.e. take a bootstrap sample). Use the rest of the cases to estimate the error of the tree, by predicting their classes.
- For each node of the tree, randomly choose m variables on which to base the decision at that node. Calculate the best split based on these m variables in the training set.
- Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).
- For prediction a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in.
- This procedure is iterated over all trees in the ensemble, and the average vote of all trees is reported as random forest prediction.

5.2 SOFTWARE USED

NetBeans IDE : NetBeans IDE lets us to quickly and easily develop Java desktop, mobile, and web applications, as well as HTML5 applications with HTML, JavaScript, and CSS. This IDE also provides a great set of tools for PHP and C/C++ developers.
The following features are provided in the current stable version:

- Fast and Smart Code Editing
- Rapid User Interface Development.
- Easy and Efficient Project Management.
- Write Bug Free Code.
- Support for Multiple Languages.
- Cross Platform Support.
- Rich Set of Community Provided Plugins.

5.3 HARDWARE SPECIFICATION

Laptop or PC with suitable RAM and processor on which the software is deployed

Table 5.1: Hardware Requirements

Sr. No.	Parameter	Minimum Requirement	Justification
1	CPU Speed	1 GHz	For fast execution
2	RAM	512 MB	Minimum RAM required is 512 MB

5.4 PROGRAMMING LANGUAGES

5.4.1 Java

Programming language used student career prediction system is Java.
Java is used because it has following features:

- Simple.
- Object-Oriented.
- Portable.
- Platform independent.
- Secured.
- Robust.
- Architecture neutral.
- Dynamic.

5.5 PLATFORM

- Windows
- Linux

5.6 COMPONENTS

- Training Component:It includes the student data-set which has all class labels
- Testing Component:It includes the student data-set which has class labels excluding the one which needs to be predicted.

5.7 TOOLS

Tableau Tool: Tableau is one of the fastest evolving Business Intelligence (BI) and data visualization tool. It is very fast to deploy, very easy to learn and very intuitive to use for a customer. The following features are provided in the current stable version:

- The greatest strength of Tableau is its speed with which it analyzes hundreds of millions of rows and gives the required answers with in seconds.
- Another strength of Tableau is that it is very easy to use. Its basic drag and drop. One can start using Tableau even with no prior programming experience.
- The Dashboard of Tableau is very interactive and gives dynamic results.
- Tableau allows the users to directly connect to databases, cubes, and data warehouses.

5.8 CODING STYLE FORMAT

- A program should be readable. All the other rules on this page can be overridden by this primary rule. Your programs should display good taste, which you can generally acquire only by paying attention to the practices of other, expert programmers.
- Every class (except for anonymous inner classes) should have a Javadoc comment that specifies the purpose of the class and describes its public interface in general terms. Except for nested classes, the Javadoc comment should include an @author tag that lists your name.
- Use indentation to display the structure of your program. The body of a class definition should be indented. The body of a method definition should be indented. When a statement is nested inside another statement, it should be indented one additional level.
- Use meaningful names for your variables, methods, and classes.
- A method should have a clear, single, identifiable task.

- A class should represent a clear, single, identifiable concept.

CHAPTER 6

RESULT SET

6.1 DATA SET PREPARATION

The scope of the system is limited to the students of Computer Department , PCCOE. Hence, the data is collected from the Computer Department previous records.

Details of the data collected :

1. Academic Details

   - University Results (FE , SE , TE , BE)

   - Unit Test Marks (FE , SE , TE , BE)

   - Term Work Marks (FE , SE , TE , BE)

2. Attendance (FE , SE , TE , BE)

3. Events Participation / Winners

   - Technical Events : Core Technical Events that require participants to showcase excellent technical skills with strong basic knowledge. These events identify Students who are technically more strong . Eg. Coding Competitions , Technical-Quiz , Hackathon coding rounds , Debugging Competitions etc.

   - Non-Technical Events - Fun events which shows the enthusiasm of students to participate and indulge in activities outside the classroom. There can be a wide range of events under this category such as : Photography related events like Snaphunt , Pictionary etc . Also events like Extempore , Debate Competition etc come under this category.

   - Technical ++ - This category of events is reserved for events that requires students to be strong in Technical skills as well as Soft Skills. Winners in this category will have strong technical base and are also able to present their ideas and work with confidence which means that they have excellent Communication Skills.

4. Volunteering Details

   Depending on how active is the student in volunteering , he/she will be categorized into one of the three categories :

A - Very Active

B - Moderately Active

C - Not Active

We collected different data about Training and Placement Cell volunteers , Volunteers of different conferences held in the institute and different events held in 4 years.

5. Technologies and Courses attended

   Apart from syllabus , students also attend various workshops in the domain that interests them. Attending workshops not only increases the knowledge of the students, but in turn makes them more competent. Hence, we collected the following data of every student with the mentioned scores :-

   1. C / C++ : 0 / 1

   2. Core Java / Advanced Java

   Workshop 1 / 2

   Course (SEED , NIIT etc) 2 / 3

   Oracle Certification 5 / 6

   3. Python : 0 / 1

   4. .NET (C# , ASP etc) : Workshop : 1

   Course : 2

   Microsoft Certification : 5

   5. Networking Workshops : 1 / 5

   6. Databases : 1 / 5

   7. Big Data / Hadoop : 0 / 2

   8. Web Technologies : 0 / 1

   9. LISP . Perl , Ruby , Scala etc. : 0 / 1

6. Coding - Hackerrank , Codechef etc.

As the project aims at Placement Prediction of Computer Engineering Department students , Coding is one of the major aspects. To check the Coding competency of students , we asked them how frequently do they solve challenges on such coding platforms.

A - Very Frequently

B - Sometimes

C - Never

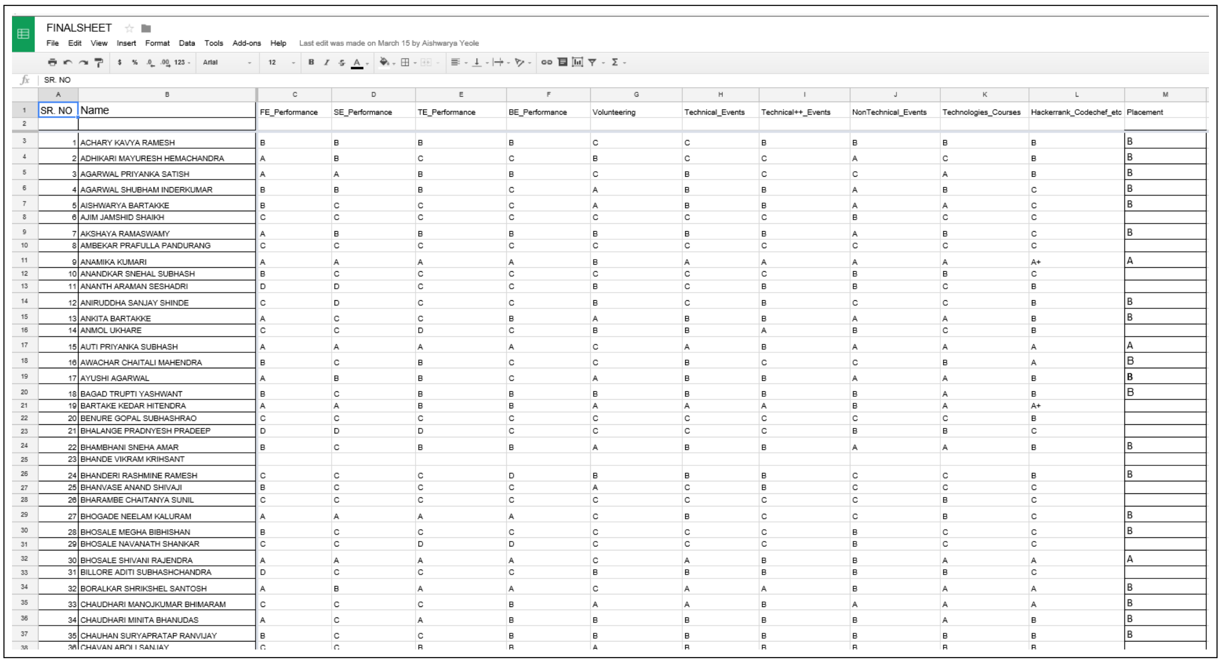
7. Placement

In the Training dataset , the class label Placement is filled with either of the three categories A , B , C. This categorization is based on the companies and the job profiles they offer in PCCOE.

A - These Companies offer a higher level post with a higher package than other average companies.

B - These are the average offers with moderate packages and profiles .

C - Communication and Support Jobs



SR. NO	Name	FE_Performance	SE_Performance	TE_Performance	BE_Performance	Volunteering	Technical_Events	Technical++_Events	NonTechnical_Events	Technologies_Courses	Hackerrank_Codechef_etc	Placement
1	ACHARY KAVYA RAMESH	B	B	B	B	C	C	B	B	B	B	B
2	ADHIKARI MAYURESH HEMACHANDRA	A	B	C	C	B	C	C	A	C	B	B
3	AGARWAL PRIYANKA SATISH	A	A	B	B	C	B	C	C	A	B	B
4	AGARWAL SHUBHAM INDERKUMAR	B	B	B	C	A	B	B	A	B	C	B
5	AISHWARYA BARTAKKE	B	C	C	C	A	B	B	A	A	C	B
6	AJIM JAMSHID SHAIKH	C	C	C	C	C	C	C	B	C	C	
7	AKSHAYA RAMASWAMY	A	B	B	B	B	B	B	A	B	C	B
8	AMBEKAR PRAFULLA PANDURANG	C	C	C	C	C	C	C	C	C	C	
9	ANAMIKA KUMARI	A	A	A	A	B	A	A	A	A	A+	A
10	ANANDKAR SNEHAL SUBHASH	B	C	C	C	C	C	C	B	B	C	
11	ANANTH ARAMAN SESHADRI	D	D	C	C	B	C	B	B	C	B	
12	ANIRUDDHA SANJAY SHINDE	C	D	C	C	B	C	B	C	C	B	B
13	ANKITA BARTAKKE	A	C	C	B	A	B	B	A	A	B	B
14	ANMOL UKHARE	C	C	D	C	B	B	A	B	C	B	
15	AUTI PRIYANKA SUBHASH	A	A	A	A	C	A	B	A	A	A	A
16	AWACHAR CHAITALI MAHENDRA	B	C	B	C	C	B	C	C	B	A	B
17	AYUSHI AGARWAL	A	B	B	C	A	B	B	A	A	B	B
18	BAGAD TRUPTI YASHWANT	B	C	B	B	B	B	B	B	A	B	B
19	BARTAKE KEDAR HITENDRA	A	A	B	B	A	A	A	B	A	A+	
20	BENURE GOPAL SUBHASHRAO	C	C	C	C	C	C	C	C	C	B	
21	BHALANGE PRADNYESH PRADEEP	D	D	D	C	C	C	C	B	B	C	
22	BHAMBHANI SNEHA AMAR	B	C	B	B	A	B	B	A	A	B	B
23	BHANDE VIKRAM KRIHSANT											
24	BHANDERI RASHMINE RAMESH	C	C	C	D	B	B	B	C	C	B	B
25	BHANVASE ANAND SHIVAJI	B	C	C	C	A	C	B	C	C	C	
26	BHARAMBE CHAITANYA SUNIL	C	C	C	C	C	C	C	C	B	C	
27	BHOGADE NEELAM KALURAM	A	A	A	A	C	B	C	C	B	C	B
28	BHOSALE MEGHA BIBHISHAN	B	C	C	C	C	C	C	B	C	C	B
29	BHOSALE NAVANATH SHANKAR	C	C	D	D	C	C	C	B	C	C	
30	BHOSALE SHIVANI RAJENDRA	A	A	A	A	C	A	B	B	A	A	A
31	BILLORE ADITI SUBHASHCHANDRA	D	C	C	C	B	B	B	B	B	C	
32	BORALKAR SHRIKSHEL SANTOSH	A	B	A	A	C	A	A	B	A	A	B
33	CHAUDHARI MANOJKUMAR BHIMARAM	C	C	C	B	A	A	B	A	A	A	B
34	CHAUDHARI MINITA BHANUDAS	A	C	A	B	B	B	B	B	A	B	B
35	CHAUHAN SURYAPRATAP RANVIJAY	B	C	C	B	B	B	B	B	B	B	B
36	CHAVAN ABOLI SANJAY	C	C	B	B	A	B	B	B	B	B	

Figure 6.1: Final Sheet

6.2 DISCUSSION

Various algorithms such as Random Forest algorithm,Naive Bayes algorithm , Regression Tree algorithm,SVM algorithm,etc were applied on the student database.
A comparison study of working of each algorithm with respect to student database was done.
Accuracy was calculated depending upon the training data using k-folds.
Accuracy of each algorithm was as follow:
Random Forest Algorithm:77.6
Naive Bayes Algorithm:68.5
Regression Tree Algorithm:71.3
SVM Algorithm:73.7
Comparison study concluded that Random Forest algorithm has the maximum accuracy. So,Random Forest Classifier is used to predict student placement class.
The student career prediction system predicts the type of company that student would most probably get placed in.Depending on the this, the faculties and TPO can take certain actions in-order to increase the placement.
This system aims in reducing the manual work of analysing the student data and helps in managing the placement work. TPO can focus on inviting the companies according to caliber of students.

6.3 SCREEN SHOTS

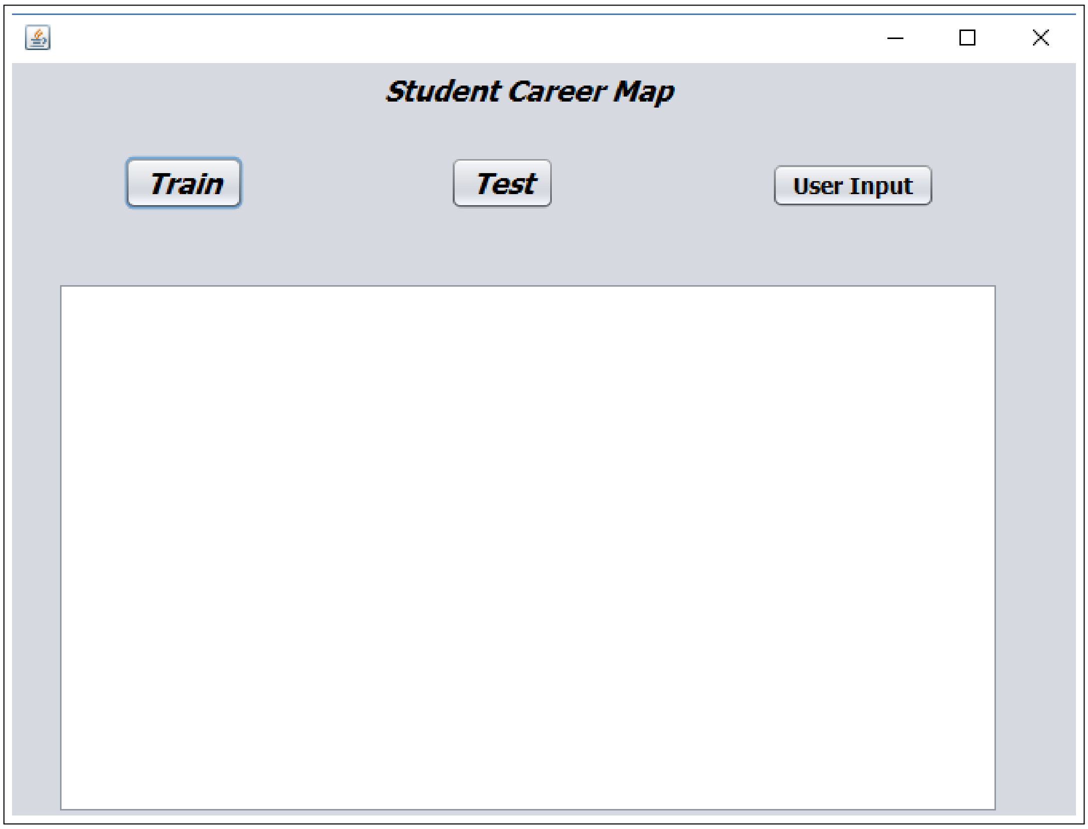
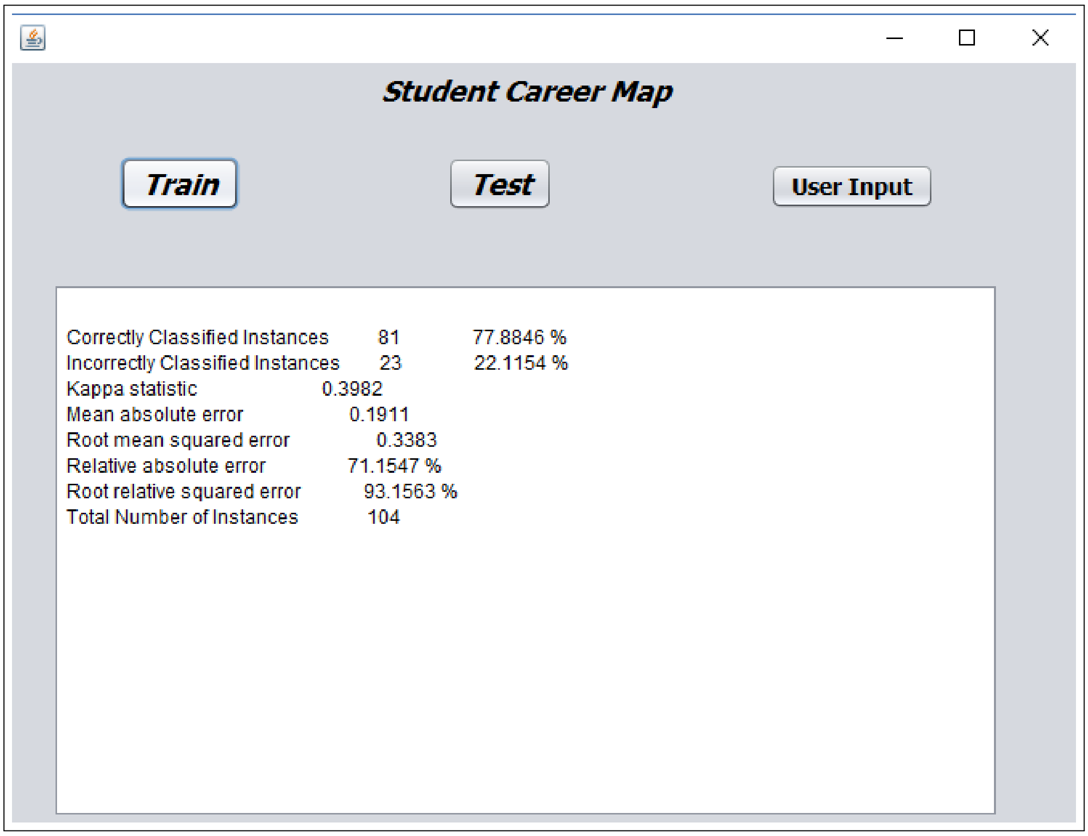


Figure 6.2: GUI

Fig 6.2 describes the GUI of the system.

6.4 OUTPUTS

STEP 1: Train the data-set

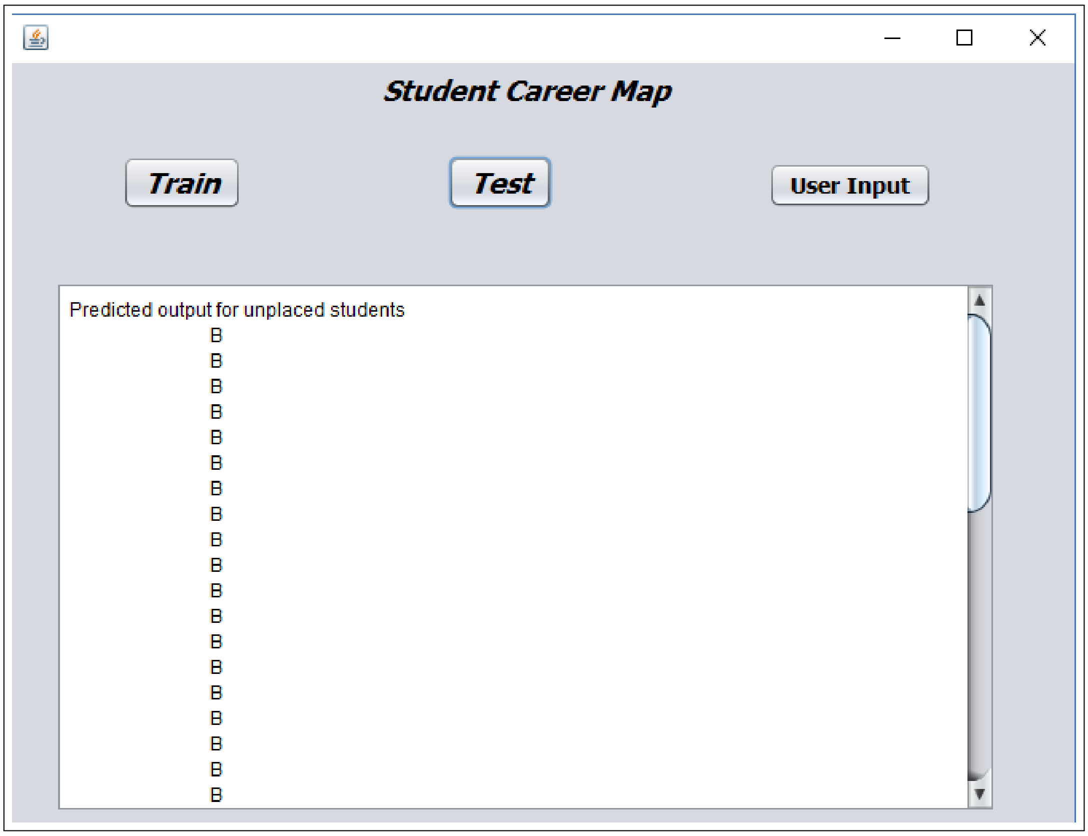


```
Correctly Classified Instances       81          77.8846 %
Incorrectly Classified Instances     23          22.1154 %
Kappa statistic                  0.3982
Mean absolute error                0.1911
Root mean squared error              0.3383
Relative absolute error            71.1547 %
Root relative squared error          93.1563 %
Total Number of Instances            104
```

Figure 6.3: Train the data set

Fig 6.3 describes the result of training the data.It displays the accuracy,correctly and incorrectly classified instances.

Step 2: Test the output for unplaced students

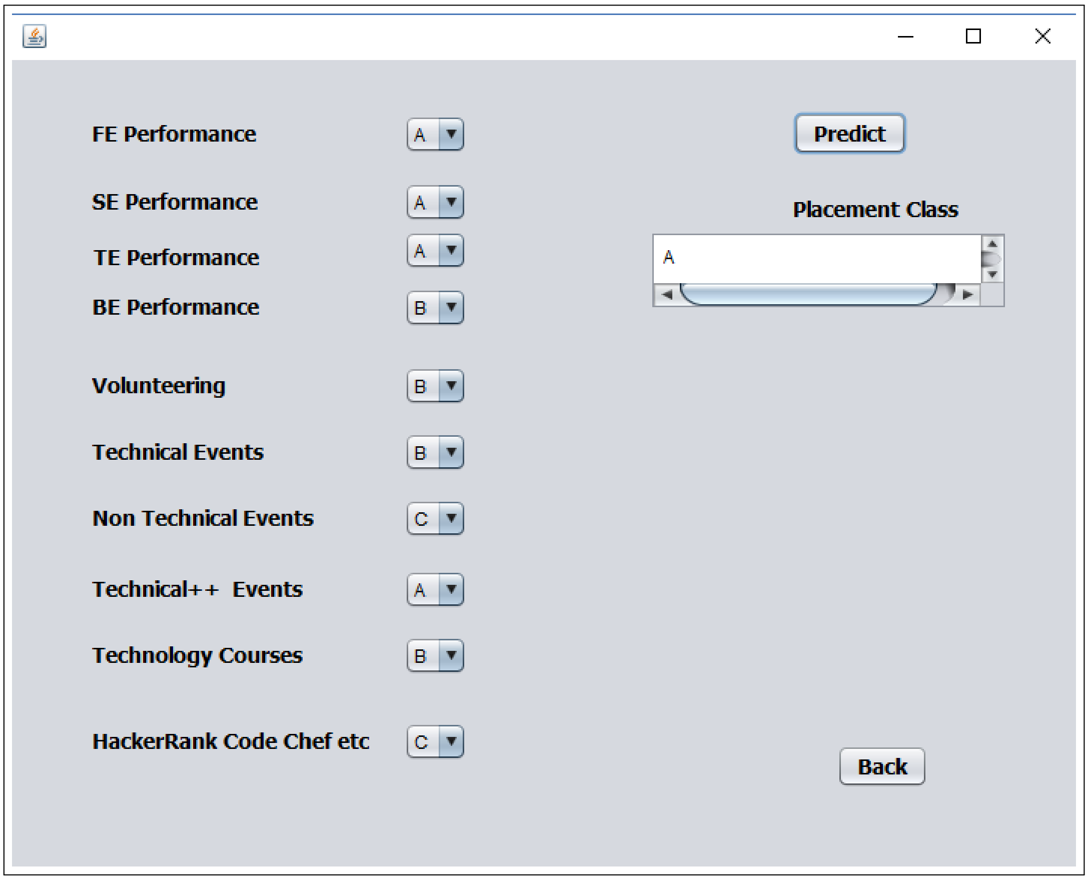


```
Predicted output for unplaced students
                    B
                    B
                    B
                    B
                    B
                    B
                    B
                    B
                    B
                    B
                    B
                    B
                    B
                    B
                    B
                    B
                    B
                    B
                    B
```

Figure 6.4: Test

Fig 6.4 describes the result of testing the data the data of unplaced students.it displays their placement class .

Step 3: Check for user input



FE Performance	A
SE Performance	A
TE Performance	A
BE Performance	B
Volunteering	B
Technical Events	B
Non Technical Events	C
Technical++ Events	A
Technology Courses	B
HackerRank Code Chef etc	C

Placement Class

A

Figure 6.5: User Input

Fig 6.5 describes the result by considering the user input.it displays their placement class .

CHAPTER 7

TESTING

7.1 TESTING

In general, testing is finding out how well something works. In terms of human beings, testing tells what level of knowledge or skill has been acquired. In computer hardware and software development, testing is used at key checkpoints in the overall process to determine whether objectives are being met. For this system manual testing has been done at different level. Test data-set has been used to check the accuracy of algorithm used in system (Random Forest). White Box and Block Box testing is also done.

7.2 FORMAL TECHNICAL REVIEWS

The term formal technical review is used to mean a software inspection. A 'Technical Review' may also refer to an acquisition life-cycle event. Formal technical review is conducted to analyse the following questions:

- Have you done alpha testing ?
- Have you done beta testing ?
- Have you validated the requirements, design and code as per standards ?
- Have you performed GUI testing of Project ?
- Have you tested the code using standard data-set available in your area of project ?
- After integration of all components, whether total performance of system is checked or not ?
- Have you done black box testing, white box testing, unit testing, integration testing ?

7.3 TEST PLAN

7.3.1 Type of Testing used:

7.3.1.1 Black Box Testing:

Under this testing effort all the input to the software were tested one by one. The error found in this testing were immediately examined and fixed. In this testing only the specification of the systems are checked over here not how the code was written. following testings are applied on our system under Black Box testing:

- All the buttons are checked for their functionality by giving input. For eg: Train Button, Test Button.
- System Output for particular unlabeled student is checked over actual output by giving its attribute as input who got placed later.
- System Output is checked by giving invalid inputs. In this system, Giving 'D' input for coding rank which is invalid printed proper error message.

7.3.1.2 White box testing:

Under this testing all the modules of the software were tested thoroughly with the different test cases. Here mainly the code of the software is checked and not the specification of the software. following testings are applied on our system under White Box testing:

- Random Forest code is checked for its correct functions.
- Statement coverage is tested for student class where each line is checked step by step.
- Branch coverage is tested for true or False input. For output of higher studies module (if else module for whether student will go for higher studies or not) is checked step by step.

7.3.2 Testing Tool

JUNIT TOOL

JUnit is a unit testing framework for the Java programming language. JUnit has been important in the development of test-driven development, and is one of a family of unit testing frameworks which is collectively known as xUnit that originated with SUnit.

Characteristics of JUnit Testing:

Following are the characteristics of the JUnit testing:

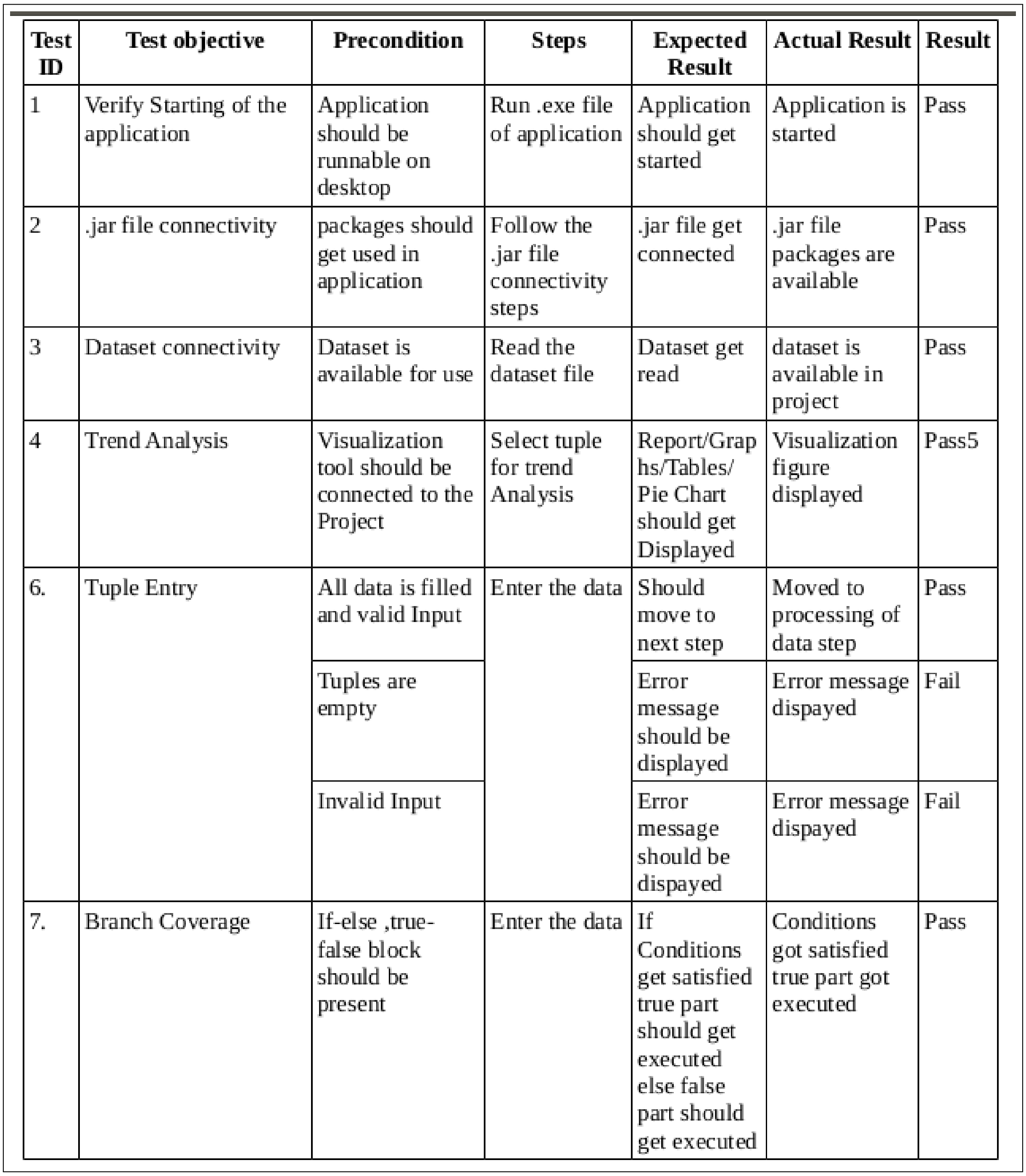
- Provides Annotation to identify the test methods.
- Provides Assertions for testing expected results.
- Provides Test runners for running tests.
- JUnit tests allow you to write code faster which increasing quality
- JUnit is elegantly simple. It is less complex and takes less time.
- JUnit tests can be run automatically and they check their own results and provide immediate feedback. There's no need to manually comb through a report of test results.
- JUnit tests can be organized into test suites containing test cases and even other test suites.
- Junit shows test progress in a bar that is green if test is going fine and it turns red when a test fails.

Advantages of JUnit Testing:

JUnit can be used separately or integrated with build tools like Maven and Ant and third party extensions, such as dbUnit cor database testing operations. Provides Assertions for testing expected results. xmlUnit which simplifies XML testing (comparing and evaluating).

Testing Type: Black Box and White Box Testing

7.4 TEST CASES AND TEST RESULTS



Test ID	Test objective	Precondition	Steps	Expected Result	Actual Result	Result
1	Verify Starting of the application	Application should be runnable on desktop	Run .exe file of application	Application should get started	Application is started	Pass
2	.jar file connectivity	packages should get used in application	Follow the .jar file connectivity steps	.jar file get connected	.jar file packages are available	Pass
3	Dataset connectivity	Dataset is available for use	Read the dataset file	Dataset get read	dataset is available in project	Pass
4	Trend Analysis	Visualization tool should be connected to the Project	Select tuple for trend Analysis	Report/Graphs/Tables/ Pie Chart should get Displayed	Visualization figure displayed	Pass5
6.	Tuple Entry	All data is filled and valid Input	Enter the data	Should move to next step	Moved to processing of data step	Pass
		Tuples are empty		Error message should be displayed	Error message dispayed	Fail
		Invalid Input		Error message should be dispayed	Error message dispayed	Fail
7.	Branch Coverage	If-else ,true-false block should be present	Enter the data	If Conditions get satisfied true part should get executed else false part should get executed	Conditions got satisfied true part got executed	Pass

Figure 7.1: Test Case

CHAPTER 8

CONFIGURATION MANAGEMENT PLAN

8.1 INSTALLATION AND UN-INSTALLATION

1. Installation

   - Copy the exe file.
   - Install exe file.

2. Un-Installation

   - Un-install exe file.

8.2 USER HELP

1. Install the software.
2. If you want to train the data set then click on Train.
3. If you want to test for unplaced students data set then click on Test.
4. If you want to dynamically take user input then click on User input.
5. Then select the values of attributes and click Predict.

CHAPTER 9

CONCLUSION AND FUTURE SCOPE

9.1 CONCLUSION

In this project placement class of students of computer department is predicted using random forest algorithm. The prediction is done depending on students' academic performance,participation in volunteering of events,technical and non technical courses and extra curricular activities.This will support the faculties and TPO in taking certain decisions related to placement activities and preparation.The system predicts the class labels with maximum accuracy possible using the most efficient algorithm.This will help the faculties to take some extra efforts on the poor students or give special courses for students who have more potential to get placed in good companies. Also, it will help the TPO to decide what kind of companies to invite based on students caliber and also take certain actions to increase the potential of students.

9.2 FUTURE SCOPE

This project can be further extended for different departments for prediction of placement of students. BE final performance can also be predicted depending on previous performance of student.

CHAPTER 10

REFERENCES

1. T. Jeevalatha N. Ananthi D. Saravana Kumar ”Performance Analysis of Undergraduate Students Placement Selection using Decision Tree Algorithms” International Journal of Computer Applications (0975 8887) Volume 108 No 15, December 2014

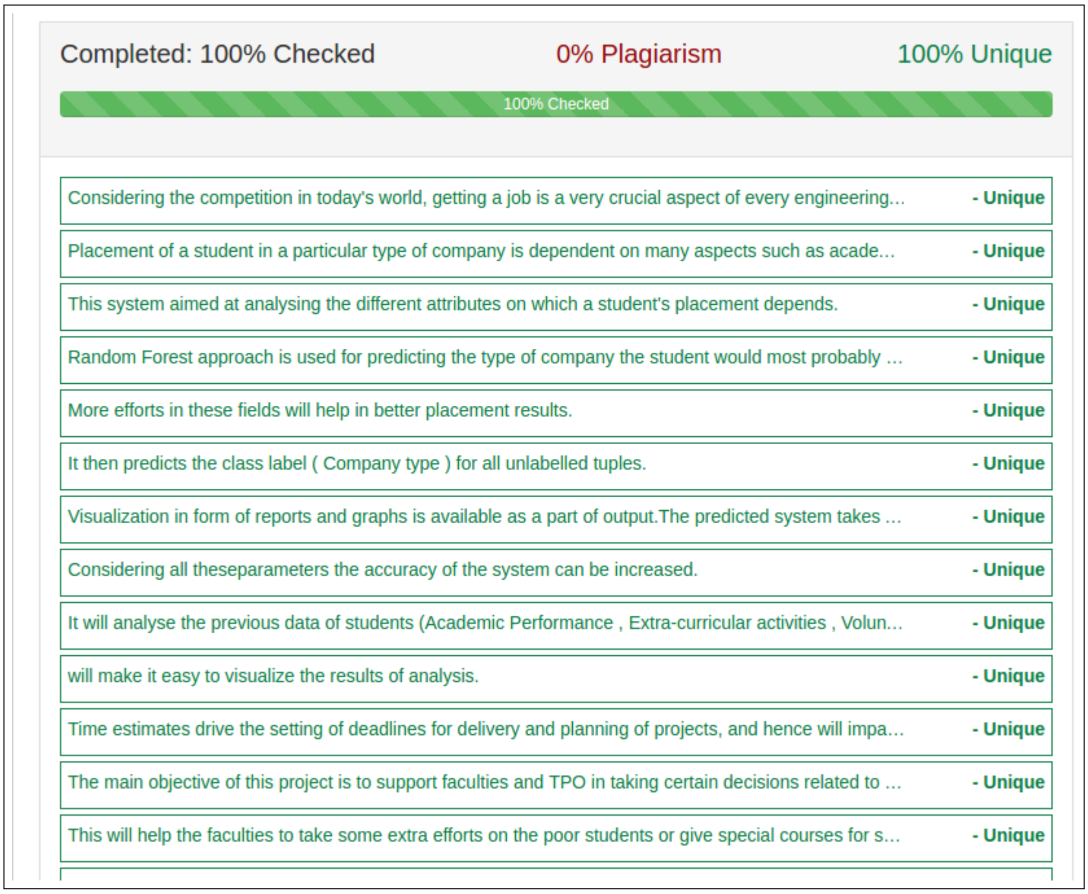
2. Hitarthi Bhatt , Shraddha Mehta , Lynette R. D’mello ”Use of ID3 Decision Tree Algorithm for Placement Prediction” Hitarthi Bhatt et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (5) , 2015, 4785-4789

3. S. Sathyavathi, N.Niraimathi, K.Priyadarshini ”DATA MINING FOR BUILDING AN INFORMED DECISION MAKING MODEL FOR CAREER PREDICTION ” IJRCS - International Journal of Research in Computer Science ISSN: 2349-3828

4. Rakesh Kumar Arora, Dr. Dharmendra Badal ”Placement Prediction through Data Mining ” Volume 4, Issue 7, July 2014 ISSN: 2277 128X

CHAPTER 11

PLAGIARISM REPORT



Completed: 100% Checked

0% Plagiarism

100% Unique

Considering the competition in today's world, getting a job is a very crucial aspect of every engineering... - Unique

Placement of a student in a particular type of company is dependent on many aspects such as acade... - Unique

This system aimed at analysing the different attributes on which a student's placement depends. - Unique

Random Forest approach is used for predicting the type of company the student would most probably ... - Unique

More efforts in these fields will help in better placement results. - Unique

It then predicts the class label (Company type) for all unlabelled tuples. - Unique

Visualization in form of reports and graphs is available as a part of output.The predicted system takes ... - Unique

Considering all theseparameters the accuracy of the system can be increased. - Unique

It will analyse the previous data of students (Academic Performance , Extra-curricular activities , Volun... - Unique

will make it easy to visualize the results of analysis. - Unique

Time estimates drive the setting of deadlines for delivery and planning of projects, and hence will impa... - Unique

The main objective of this project is to support faculties and TPO in taking certain decisions related to ... - Unique

This will help the faculties to take some extra efforts on the poor students or give special courses for s... - Unique

Figure 11.1: Plagiarism Report

CHAPTER 12

TERM-II PROJECT LABORATORY ASSIGNMENTS

Assignment No. 1

AIM: Project workstation selection, installations and setup.

Workstation Used: NetBeans

- NETBEANS INSTALLATION STEPS:

  1. Step 0: Install JDK

     To use NetBeans for Java programming, you need to first install Java Development Kit (JDK). See ”JDK - How to Install”.

  2. Step 1: Download

     Download ”NetBeans IDE” installer from http://netbeans.org/downloads/index.html. There are many ”bundles” available. For beginners, choose the 1st entry ”Java SE” (e.g., ”netbeans-8.2-javase-windows.exe” 95MB).

  3. Step 2: Run the Installer

     Run the downloaded installer.

- NetBeans Java PACKAGES: Packages are used in Java in order to prevent naming conflicts, to control access, to make searching/locating and usage of classes, interfaces, enumerations and annotations easier.

  1. Step 1: Open Netbeans and expand your project in the Projects window panel on the left. Right click on Libraries and select the Add JAR/Folder option.

  2. Step 2: In the resulting dialog window that appears, browse to the location of the JAR file on your computer and then click Open. The JAR file should now be automatically added to the Class Path.

INSTALLATION STEPS:

EXECUTION STEPS:

1. Open the NetBeans application in windows.

2. Run the project in NetBeans.

3. Project will run on your desktop

Assignment No. 2

Aim: Programming and GUI.

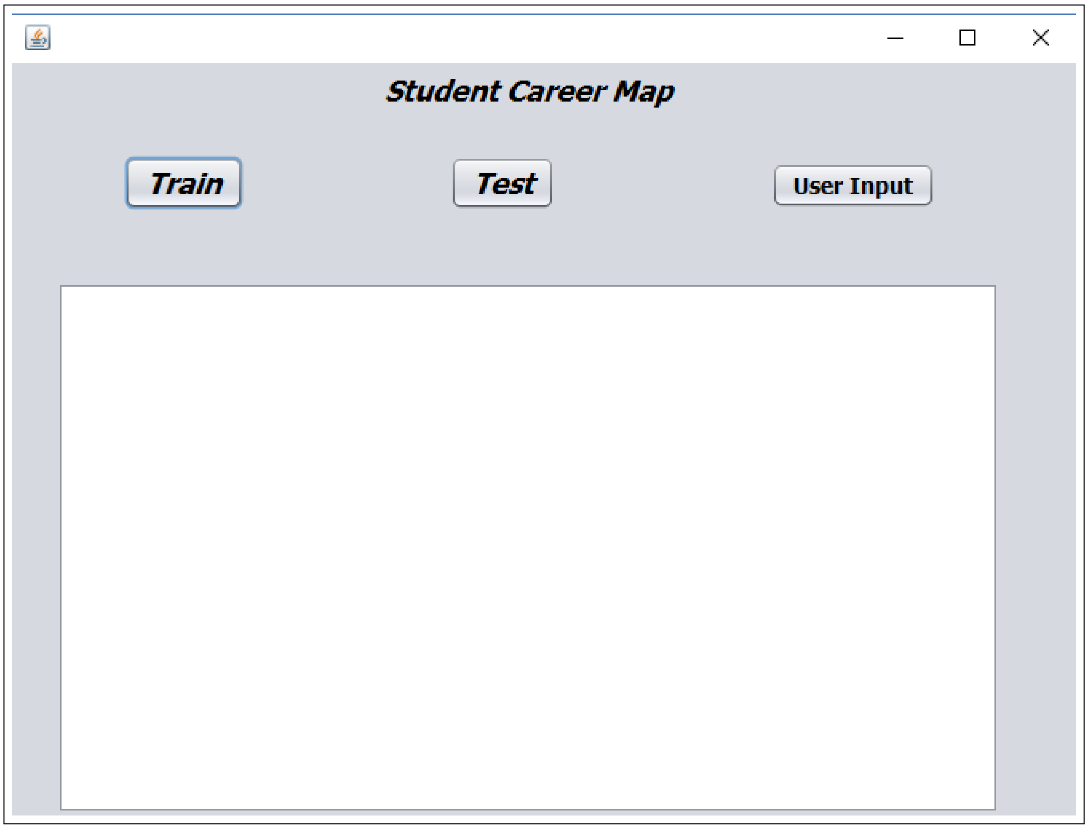
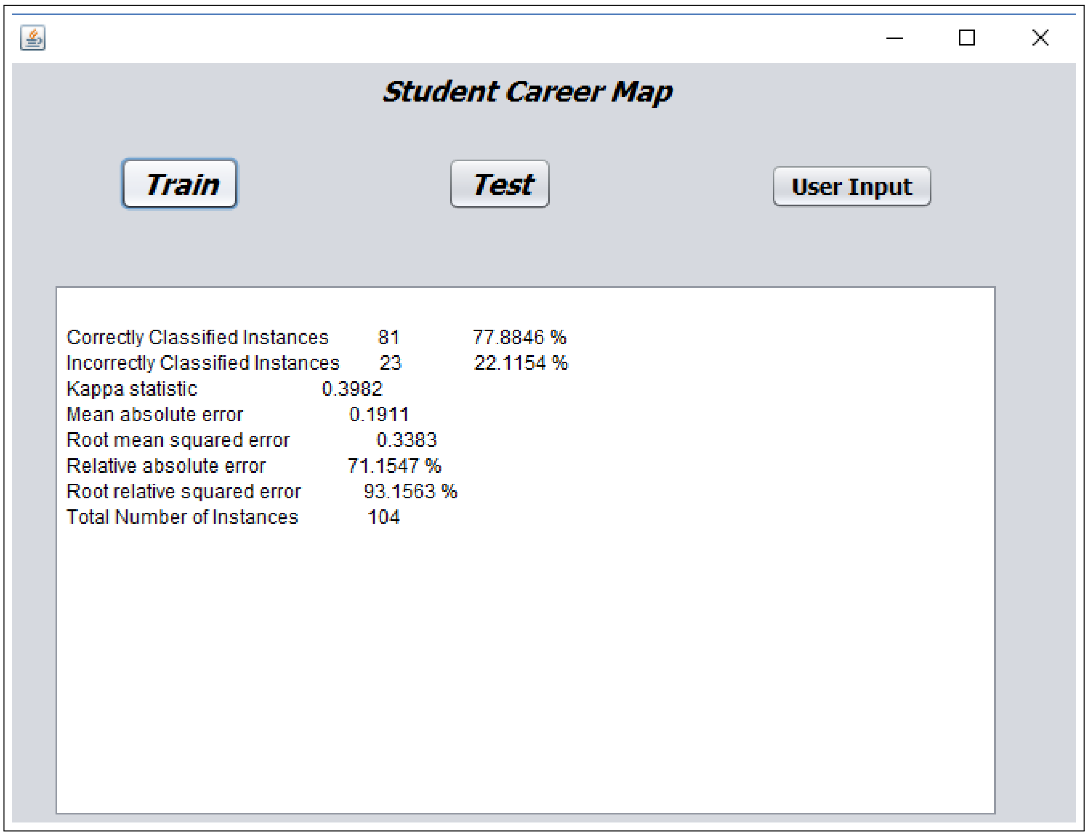


Figure 12.1: GUI

Fig 12.1 describes the GUI of the system.

STEP 1: Train the data-set

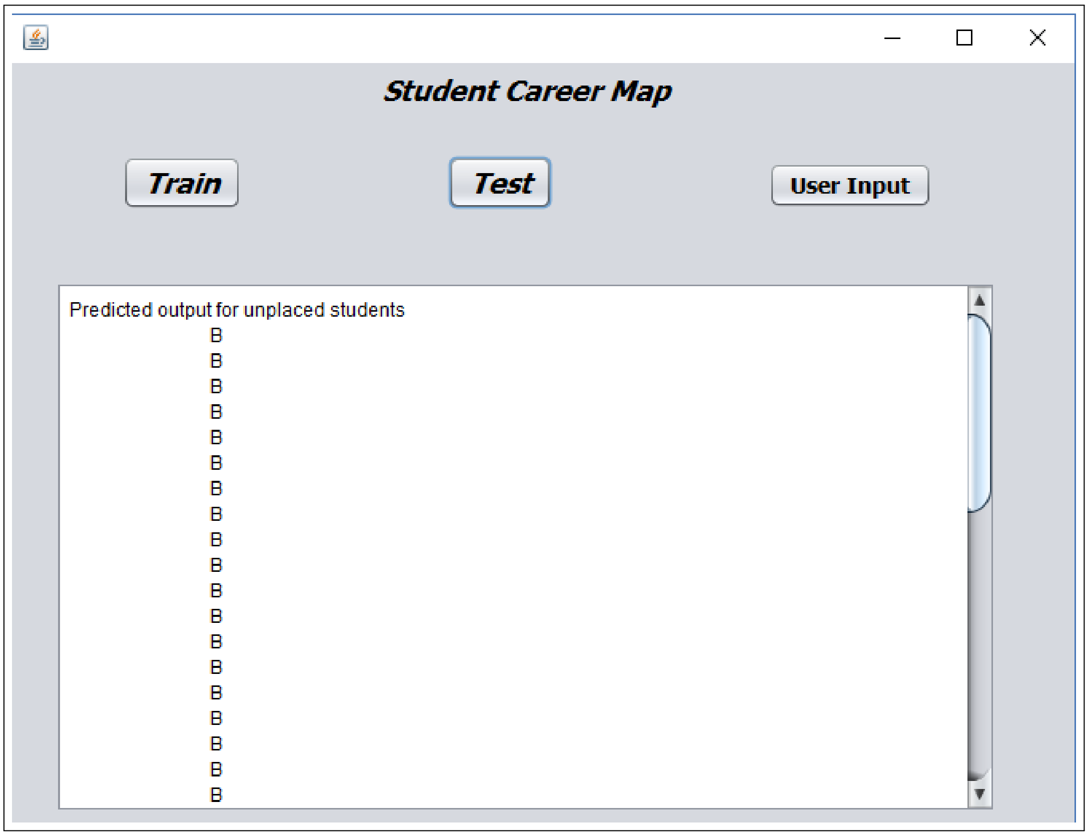


```
Correctly Classified Instances       81          77.8846 %
Incorrectly Classified Instances     23          22.1154 %
Kappa statistic                  0.3982
Mean absolute error                  0.1911
Root mean squared error                  0.3383
Relative absolute error              71.1547 %
Root relative squared error          93.1563 %
Total Number of Instances            104
```

Figure 12.2: Train the data set

Fig 12.2 describes the result of training the data.It displays the accuracy,correctly and incorrectly classified instances.

Step 2: Test the output for unplaced students

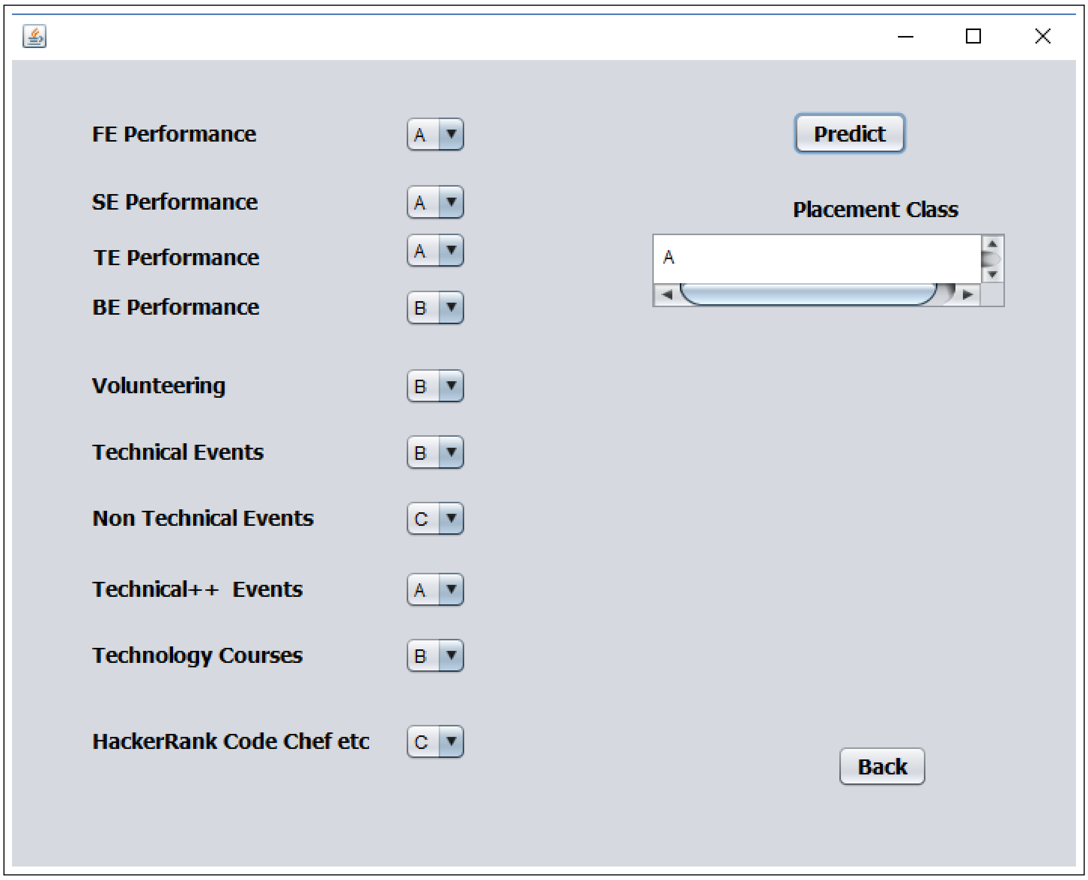


```
Predicted output for unplaced students
                     B
                     B
                     B
                     B
                     B
                     B
                     B
                     B
                     B
                     B
                     B
                     B
                     B
                     B
                     B
                     B
                     B
                     B
                     B
```

Figure 12.3: Test

Fig 12.3 describes the result of testing the data the data of unplaced students.it displays their placement class .

Step 3: Check for user input



FE Performance A

SE Performance A

TE Performance A

BE Performance B

Volunteering B

Technical Events B

Non Technical Events C

Technical++ Events A

Technology Courses B

HackerRank Code Chef etc C

Placement Class

A

Figure 12.4: User Input

Fig 12.4 describes the result by considering the user input.it displays their placement class .

Assignment No.3

Aim: Programming of the project functions, interfaces and GUI (if any) as per 1 st Term term-work submission using corrective actions recommended in Term-I assessment of Term-work.

Problem Statement : StudentCareerMap: A Learning based Student Performance and Career Analyzer

Main Project functions are: Classification and Visualization

1. Data Gathering and Preprocessing :

   Collecting the data of students for all 4 years of Engineering from various sources. All data of a particular student is needed i.e. Academics , Extra curricular activities , Volunteering , Achievements etc.

2. Decision Tree Algorithm : Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches . Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

   The core algorithm for building decision trees called ID3 by J. R. Quinlan which employs a top-down, greedy search through the space of possible branches with no backtracking. ID3 uses Entropy and Information Gain to construct a decision tree.

   Entropy A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous). ID3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one.

   Information Gain The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain (i.e., the most homogeneous branches).

Step 1: Calculate entropy of the target.
Step 2: The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain, or decrease in entropy.
Step 3: Choose attribute with the largest information gain as the decision node, divide the dataset by its branches and repeat the same process on every branch.
Step 4a: A branch with entropy of 0 is a leaf node.
Step 4b: A branch with entropy more than 0 needs further splitting.
Step 5: The ID3 algorithm is run recursively on the non-leaf branches, until all data is classified.

3. RandomForest :

Random forest is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees. Each decision tree is constructed using the following algorithm:

- Let the number of training cases be N, and the number of variables in the classifier be M.
- We are told the number m of input variables to be used to determine the decision at a node of the tree; m should be much less than M.
- Choose a training set for this tree by choosing n times with replacement from all N available training cases (i.e. take a bootstrap sample). Use the rest of the cases to estimate the error of the tree, by predicting their classes.
- For each node of the tree, randomly choose m variables on which to base the decision at that node. Calculate the best split based on these m variables in the training set.

- Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).
- For prediction a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in.
- This procedure is iterated over all trees in the ensemble, and the average vote of all trees is reported as random forest prediction.

Assignment No.4

Aim: Test tool selection and testing of various test cases for the project performed and generate various testing result charts, graphs etc. including reliability testing.

Problem Statement :

StudentCareerMap: A Learning based Student Performance and career Analyzer

Testing Tool: JUNIT TOOL

JUnit is a unit testing framework for the Java programming language. JUnit has been important in the development of test-driven development, and is one of a family of unit testing frameworks which is collectively known as xUnit that originated with SUnit.

Characteristics of JUnit Testing:

Following are the characteristics of the JUnit testing:

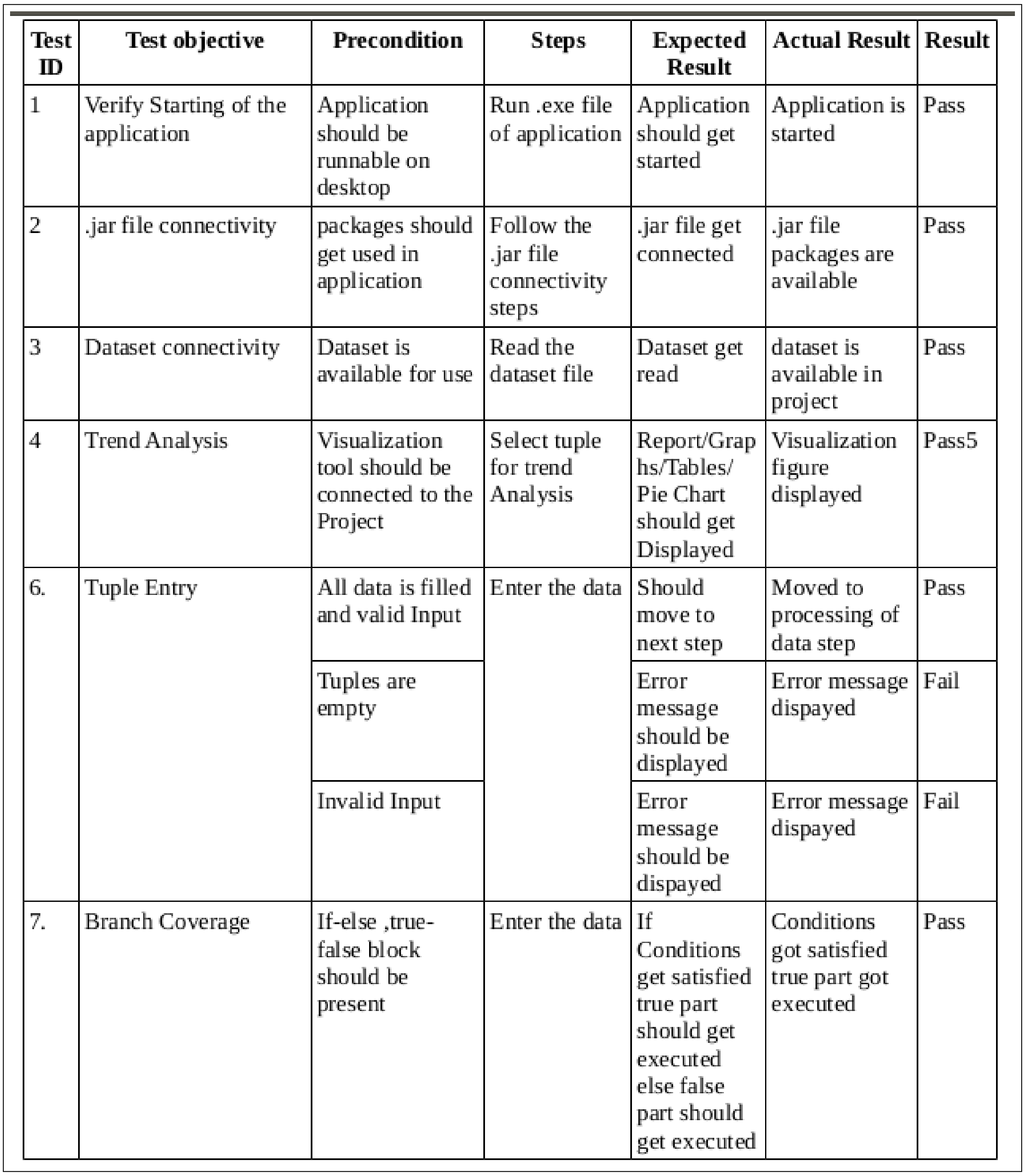
- Provides Annotation to identify the test methods.
- Provides Assertions for testing expected results.
- Provides Test runners for running tests.
- JUnit tests allow you to write code faster which increasing quality
- JUnit is elegantly simple. It is less complex takes less time.
- JUnit tests can be run automatically and they check their own results and provide immediate feedback. There's no need to manually comb through a report of test results.
- JUnit tests can be organized into test suites containing test cases and even other test suites.
- Junit shows test progress in a bar that is green if test is going fine and it turns red when a test fails.

Advantages of JUnit Testing:

JUnit can be used separately or integrated with build tools like Maven and Ant and third party extensions, such as dbUnit cor database testing operations. Provides Assertions for testing expected results. xmlUnit which simplifies XML testing (comparing and evaluating).

Testing Type: Black Box and White Box Testing

Test Cases:

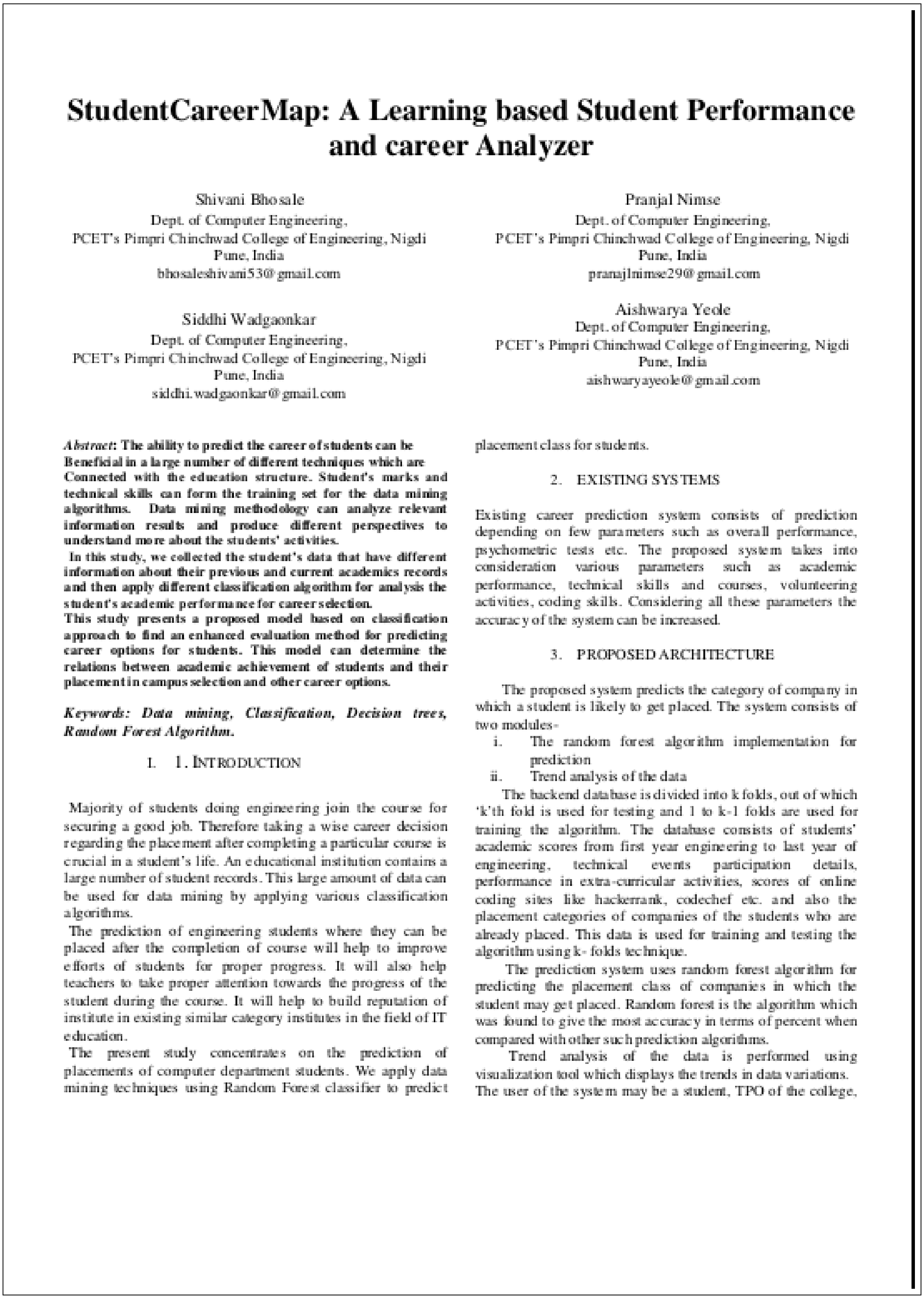


Test ID	Test objective	Precondition	Steps	Expected Result	Actual Result	Result
1	Verify Starting of the application	Application should be runnable on desktop	Run .exe file of application	Application should get started	Application is started	Pass
2	.jar file connectivity	packages should get used in application	Follow the .jar file connectivity steps	.jar file get connected	.jar file packages are available	Pass
3	Dataset connectivity	Dataset is available for use	Read the dataset file	Dataset get read	dataset is available in project	Pass
4	Trend Analysis	Visualization tool should be connected to the Project	Select tuple for trend Analysis	Report/Graphs/Tables/ Pie Chart should get Displayed	Visualization figure displayed	Pass5
6.	Tuple Entry	All data is filled and valid Input	Enter the data	Should move to next step	Moved to processing of data step	Pass
		Tuples are empty		Error message should be displayed	Error message dispayed	Fail
		Invalid Input		Error message should be dispayed	Error message dispayed	Fail
7.	Branch Coverage	If-else ,true-false block should be present	Enter the data	If Conditions get satisfied true part should get executed else false part should get executed	Conditions got satisfied true part got executed	Pass

Figure 12.5: Test Case

CHAPTER 13

PAPER PUBLICATION/PATENT/COPYRIGHT



StudentCareerMap: A Learning based Student Performance and career Analyzer

Shivani Bhosale
Dept. of Computer Engineering,
PCET's Pimpri Chinchwad College of Engineering, Nigdi
Pune, India
bhosaleshivani53@gmail.com

Pranjal Nimse
Dept. of Computer Engineering,
PCET's Pimpri Chinchwad College of Engineering, Nigdi
Pune, India
pranajlnimse29@gmail.com

Siddhi Wadgaonkar
Dept. of Computer Engineering,
PCET's Pimpri Chinchwad College of Engineering, Nigdi
Pune, India
siddhi.wadgaonkar@gmail.com

Aishwarya Yeole
Dept. of Computer Engineering,
PCET's Pimpri Chinchwad College of Engineering, Nigdi
Pune, India
aishwaryayeole@gmail.com

***Abstract*: The ability to predict the career of students can be Beneficial in a large number of different techniques which are Connected with the education structure. Student's marks and technical skills can form the training set for the data mining algorithms. Data mining methodology can analyze relevant information results and produce different perspectives to understand more about the students' activities.**

In this study, we collected the student's data that have different information about their previous and current academics records and then apply different classification algorithm for analysis the student's academic performance for career selection.

This study presents a proposed model based on classification approach to find an enhanced evaluation method for predicting career options for students. This model can determine the relations between academic achievement of students and their placement in campus selection and other career options.

Keywords: Data mining, Classification, Decision trees, Random Forest Algorithm.

I. 1. INTRODUCTION

Majority of students doing engineering join the course for securing a good job. Therefore taking a wise career decision regarding the placement after completing a particular course is crucial in a student's life. An educational institution contains a large number of student records. This large amount of data can be used for data mining by applying various classification algorithms.

The prediction of engineering students where they can be placed after the completion of course will help to improve efforts of students for proper progress. It will also help teachers to take proper attention towards the progress of the student during the course. It will help to build reputation of institute in existing similar category institutes in the field of IT education.

The present study concentrates on the prediction of placements of computer department students. We apply data mining techniques using Random Forest classifier to predict placement class for students.

2. EXISTING SYSTEMS

Existing career prediction system consists of prediction depending on few parameters such as overall performance, psychometric tests etc. The proposed system takes into consideration various parameters such as academic performance, technical skills and courses, volunteering activities, coding skills. Considering all these parameters the accuracy of the system can be increased.

3. PROPOSED ARCHITECTURE

The proposed system predicts the category of company in which a student is likely to get placed. The system consists of two modules-

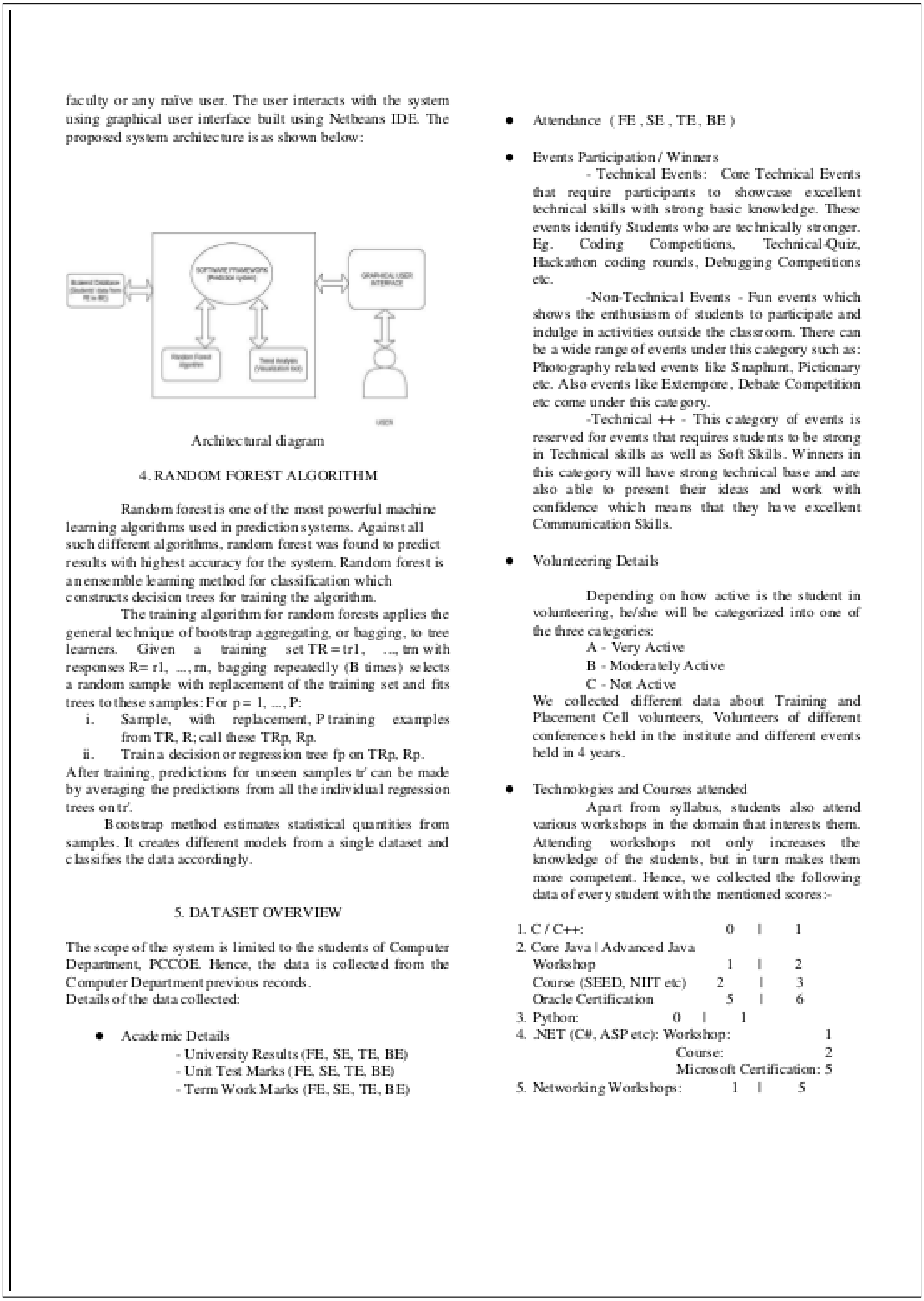
i. The random forest algorithm implementation for prediction
ii. Trend analysis of the data

The backend database is divided into k folds, out of which 'k'th fold is used for testing and 1 to k-1 folds are used for training the algorithm. The database consists of students' academic scores from first year engineering to last year of engineering, technical events participation details, performance in extra-curricular activities, scores of online coding sites like hackerrank, codechef etc. and also the placement categories of companies of the students who are already placed. This data is used for training and testing the algorithm using k- folds technique.

The prediction system uses random forest algorithm for predicting the placement class of companies in which the student may get placed. Random forest is the algorithm which was found to give the most accuracy in terms of percent when compared with other such prediction algorithms.

Trend analysis of the data is performed using visualization tool which displays the trends in data variations. The user of the system may be a student, TPO of the college,

Figure 13.1: Test Case


Student Database (Students' data from FE to BE)
SOFTWARE FRAMEWORK (Prediction system)
GRAPHICAL USER INTERFACE
Random Forest Algorithm
Trend Analysis (Visualization tool)
USER

faculty or any naïve user. The user interacts with the system using graphical user interface built using Netbeans IDE. The proposed system architecture is as shown below:

Architectural diagram

4. RANDOM FOREST ALGORITHM

Random forest is one of the most powerful machine learning algorithms used in prediction systems. Against all such different algorithms, random forest was found to predict results with highest accuracy for the system. Random forest is an ensemble learning method for classification which constructs decision trees for training the algorithm.

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set TR = tr1, ..., trn with responses R= r1, ..., rn, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples: For p = 1, ..., P:

i. Sample, with replacement, P training examples from TR, R; call these TRp, Rp.
ii. Train a decision or regression tree fp on TRp, Rp.

After training, predictions for unseen samples tr' can be made by averaging the predictions from all the individual regression trees on tr'.

Bootstrap method estimates statistical quantities from samples. It creates different models from a single dataset and classifies the data accordingly.

5. DATASET OVERVIEW

The scope of the system is limited to the students of Computer Department, PCCOE. Hence, the data is collected from the Computer Department previous records.
Details of the data collected:

- Academic Details
  - University Results (FE, SE, TE, BE)
  - Unit Test Marks (FE, SE, TE, BE)
  - Term Work Marks (FE, SE, TE, BE)

- Attendance (FE , SE , TE , BE)

- Events Participation / Winners
  - Technical Events: Core Technical Events that require participants to showcase excellent technical skills with strong basic knowledge. These events identify Students who are technically stronger. Eg. Coding Competitions, Technical-Quiz, Hackathon coding rounds, Debugging Competitions etc.
  -Non-Technical Events - Fun events which shows the enthusiasm of students to participate and indulge in activities outside the classroom. There can be a wide range of events under this category such as: Photography related events like Snaphunt, Pictionary etc. Also events like Extempore, Debate Competition etc come under this category.
  -Technical ++ - This category of events is reserved for events that requires students to be strong in Technical skills as well as Soft Skills. Winners in this category will have strong technical base and are also able to present their ideas and work with confidence which means that they have excellent Communication Skills.

- Volunteering Details

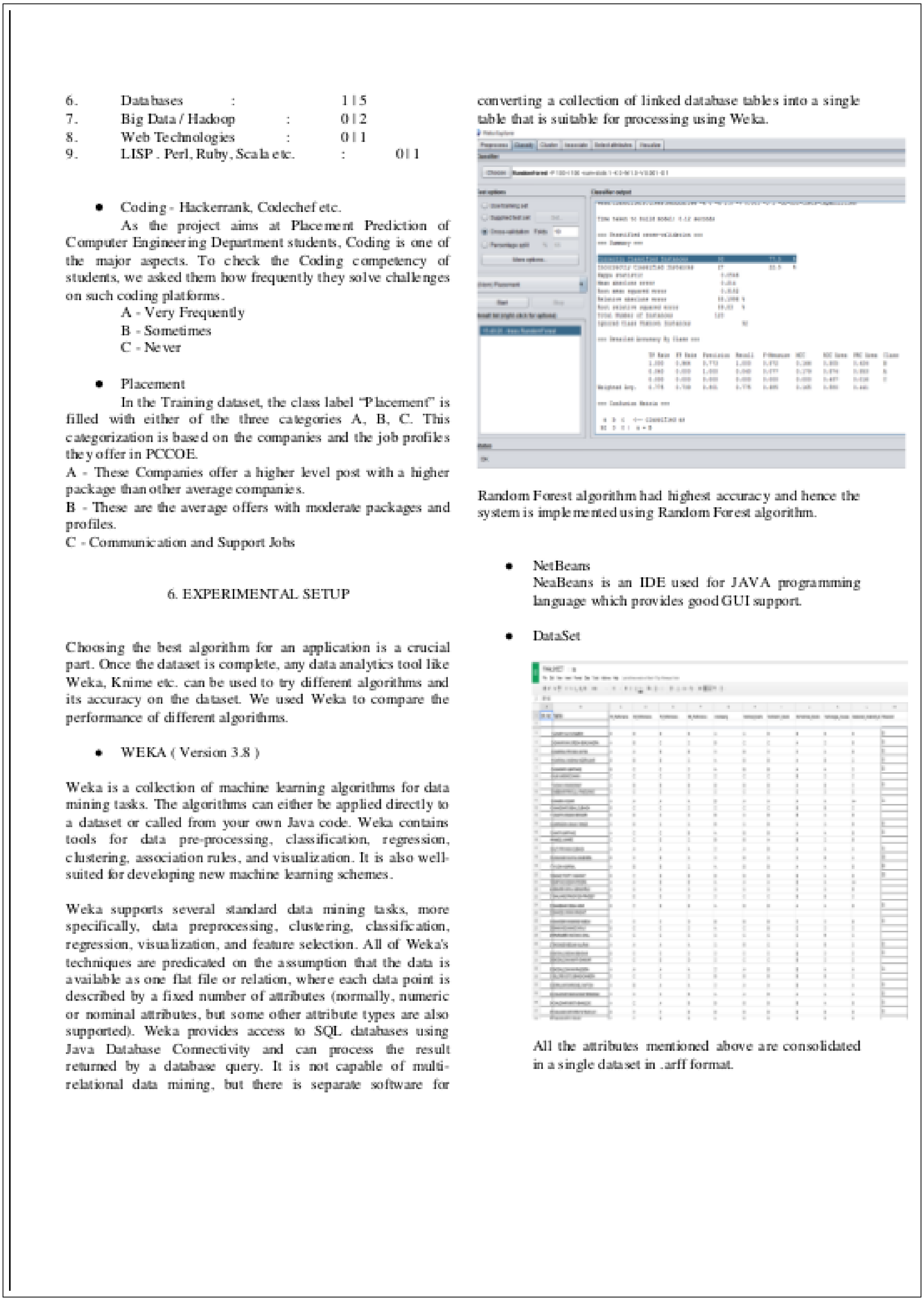
Depending on how active is the student in volunteering, he/she will be categorized into one of the three categories:
A - Very Active
B - Moderately Active
C - Not Active
We collected different data about Training and Placement Cell volunteers, Volunteers of different conferences held in the institute and different events held in 4 years.

- Technologies and Courses attended

Apart from syllabus, students also attend various workshops in the domain that interests them. Attending workshops not only increases the knowledge of the students, but in turn makes them more competent. Hence, we collected the following data of every student with the mentioned scores:-

1. C / C++: 0 | 1
2. Core Java | Advanced Java
   Workshop 1 | 2
   Course (SEED, NIIT etc) 2 | 3
   Oracle Certification 5 | 6
3. Python: 0 | 1
4. .NET (C#, ASP etc): Workshop: 1
   Course: 2
   Microsoft Certification: 5
5. Networking Workshops: 1 | 5

Figure 13.2: Test Case



6.	Databases	:	1 | 5
7.	Big Data / Hadoop	:	0 | 2
8.	Web Technologies	:	0 | 1
9.	LISP . Perl, Ruby, Scala etc.	:	0 | 1

- Coding - Hackerrank, Codechef etc.

As the project aims at Placement Prediction of Computer Engineering Department students, Coding is one of the major aspects. To check the Coding competency of students, we asked them how frequently they solve challenges on such coding platforms.

A - Very Frequently
B - Sometimes
C - Never

- Placement

In the Training dataset, the class label "Placement" is filled with either of the three categories A, B, C. This categorization is based on the companies and the job profiles they offer in PCCOE.

A - These Companies offer a higher level post with a higher package than other average companies.

B - These are the average offers with moderate packages and profiles.

C - Communication and Support Jobs

6. EXPERIMENTAL SETUP

Choosing the best algorithm for an application is a crucial part. Once the dataset is complete, any data analytics tool like Weka, Knime etc. can be used to try different algorithms and its accuracy on the dataset. We used Weka to compare the performance of different algorithms.

- WEKA (Version 3.8)

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka.

Random Forest algorithm had highest accuracy and hence the system is implemented using Random Forest algorithm.

- NetBeans

NeaBeans is an IDE used for JAVA programming language which provides good GUI support.

- DataSet

All the attributes mentioned above are consolidated in a single dataset in .arff format.

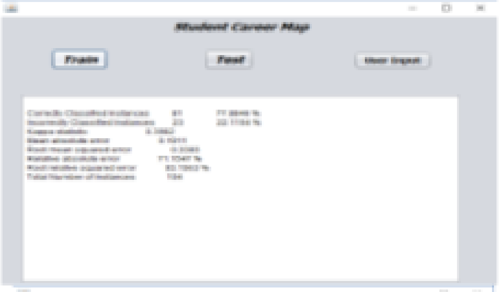
Figure 13.3: Test Case

7. RESULTS

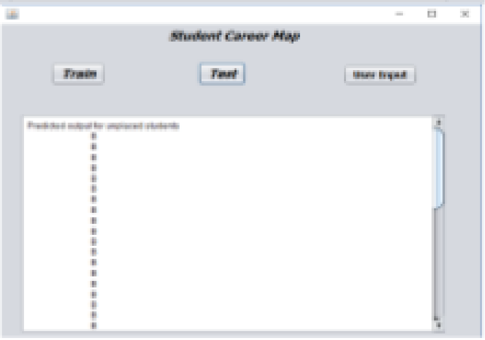
Various algorithms such as Random Forest algorithm, Naive Bayes algorithm, Regression Tree algorithm, SVM algorithm, etc were applied on the student database.
A comparison study of working of each algorithm with respect to student database was done.
Accuracy was calculated depending upon the training data using k-folds.
Accuracy of each algorithm was as follow:

- Naive Bayes Algorithm:68.5
- Regression Tree Algorithm:71.3
- SVM Algorithm:73.7
- Random Forest Algorithm:77.5

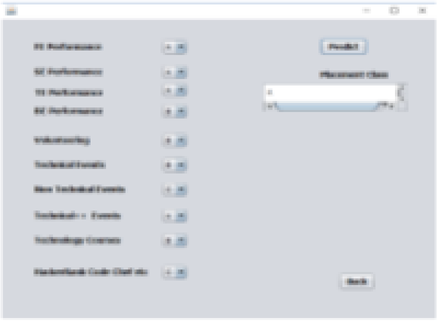
Comparison study concluded that Random Forest algorithm has the maximum accuracy.
So, Random Forest Classifier is used to predict student placement class.
The student career prediction system predicts the type of company that student would most probably get placed in. Depending on this, the faculties and TPO can take certain actions in-order to increase the placement.
This system aims in reducing the manual work of analyzing the student data and helps in managing the placement work. TPO can focus on inviting the companies according to caliber of students.
Screenshots of the system:



```
Correctly Classified Instances        81          77.8846 %
Incorrectly Classified Instances      23          22.1154 %
Kappa statistic                    [illegible]
Mean absolute error                   [illegible]
Root mean squared error                 [illegible]
Relative absolute error               [illegible] %
Root relative squared error            [illegible] %
Total Number of Instances             104
```



Predicted output for unplaced students



8. CONCLUSION

In this paper we present a system which predicts the placement class of students of Computer Department, Pimpri Chinchwad College of Engineering using Random Forest algorithm. The prediction is done by taking into consideration the student's performance of four academic years, participation in volunteering of different events, technical and non-technical courses and extra-curricular activities.

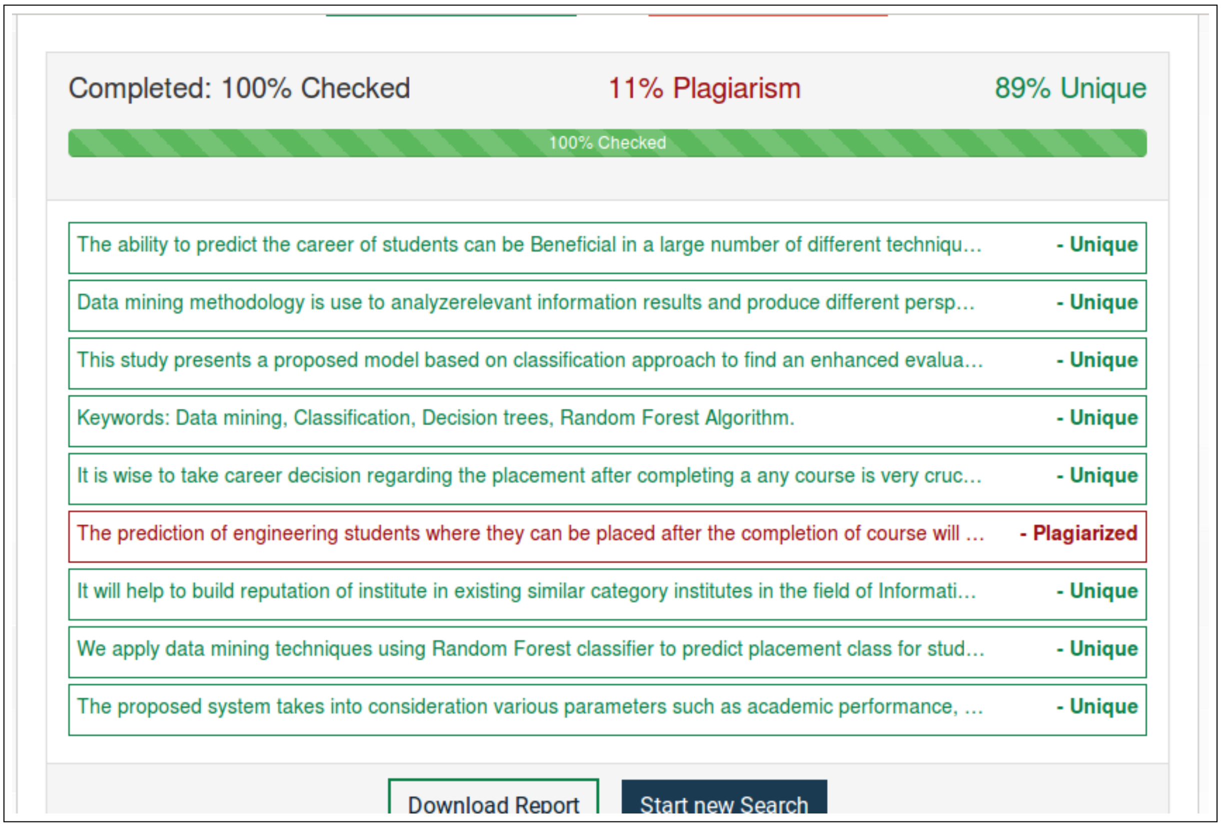
9. ACKNOWLEDGEMENT

We express our sincere thanks to project guide Dr. P. R. Futane for contributing his thoughts and ideas. We also thank Persistent Sytems for the project sponsorship and Mr. Snehkumar Shahani, our mentor from Persistent Systems, for his guidance.

10. REFERENCES

1. T. Jeevalatha N. Ananthi D. Saravana Kumar "Performance Analysis of Undergraduate Students Placement Selection using Decision Tree Algorithms" International Journal of Computer Applications (0975 – 8887) Volume 108 – No 15, December 2014
2. Hitarthi Bhatt , Shraddha Mehta , Lynette R. D'mello "Use of ID3 Decision Tree Algorithm for Placement Prediction" Hitarthi Bhatt et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (5) , 2015, 4785-4789
3. S. Sathyavathi, N.Niraimathi, K.Priyadarshini "DATA MINING FOR BUILDING AN INFORMED DECISION MAKING MODEL FOR CAREER PREDICTION " IJRCS - International Journal of Research in Computer Science ISSN: 2349-3828

Figure 13.4: Plagarism Report of Paper



Completed: 100% Checked

11% Plagiarism

89% Unique

The ability to predict the career of students can be Beneficial in a large number of different techniqu… - Unique

Data mining methodology is use to analyzerelevant information results and produce different persp… - Unique

This study presents a proposed model based on classification approach to find an enhanced evalua… - Unique

Keywords: Data mining, Classification, Decision trees, Random Forest Algorithm. - Unique

It is wise to take career decision regarding the placement after completing a any course is very cruc… - Unique

The prediction of engineering students where they can be placed after the completion of course will … - Plagiarized

It will help to build reputation of institute in existing similar category institutes in the field of Informati… - Unique

We apply data mining techniques using Random Forest classifier to predict placement class for stud… - Unique

The proposed system takes into consideration various parameters such as academic performance, … - Unique

Figure 13.5: Test Case

CHAPTER 14

INFORMATION OF PROJECT GROUP MEMBERS

1. Name: Shivani Bhosale
2. Date of Birth:10/11/1995
3. Gender:Female
4. Permanent Address:Chinchwad
5. E-Mail: bhosaleshivani53@gmail.com
6. Mobile/Contact No:9689704889
7. Placement Details: KPIT,Persistent

1. Name: Pranjal Nimse
2. Date of Birth: 29/07/1995
3. Gender: Female
4. Permanent Address:Nigdi
5. E-Mail: pranjalnimse29@gmail.com
6. Mobile/Contact No: 9860218475
7. Placement Details: KPIT,Persistent

1. Name: Siddhi Wadgaonkar

2. Date of Birth: 28/03/1996

3. Gender: Female

4. Permanent Address: Pradhikaran

5. E-Mail: siddhi.wadgaonkar@gmail.com

6. Mobile/Contact No: 8805016801

7. Placement Details: KPIT,Bombardier

1. Name: Aishwarya Yeole
2. Date of Birth: 08/09/1995
3. Gender: Female
4. Permanent Address: Pimpri
5. E-Mail: aishwaryayeole@gmail.com
6. Mobile/Contact No: 7020372684
7. Placement Details: KPIT,Xoriant

CHAPTER 15

REPORT DOCUMENTATION

Report Documentation

Report Code: CS-BE-Project 2016-2017 **Report Number:GB-17**

Report Title: StudentCareerMap: A Learning based Student Performance and career Analyzer

Address (Details):

Pimpri Chinchwad College of Engineering

Sector-26 Pradhikaran, Nigdi,Pune-411044

Author	Shivani Bhosale	Pranjal Nimse	Siddhi Wadgaonkar	Aishwarya Yeole
Address	Near Gurudwara, Chinchwad 411033	Krishnanagar, Nigdi 411019	Indiranagar Chinchwad 411033	Vallabh Nagar Pimpri 411018
E-mail	bhosaleshivani53@gmail.com	pranjalnimse29@gmail.com	Siddhi.wadgaonkar@gmail.com	aishwaryayeole@gmail.com
Roll No.	BECOA115	BECOB223	BECOB266	BECOA271
Cell No.	9689704889	9860218475	8805016801	9028536495

Year: 2016-17

Branch: Computer Engineering

Key Words: *Databases, Data mining , Random Forest classifier ,Decision Tree , Prediction*

Type of Report	Report Checked By:	Report Checked Date:	Guides Complete Name:	Total Copies
FINAL	Dr.Pravin Futane		Dr.Pravin Futane	6

Abstract:

Considering the competition in today's world, getting a job is a very crucial aspect of every engineering student. Every institute wants to improve their placement records every year. Placement of a student in a particular type of company is dependent on many aspects, for eg. his academic performance , other activities , projects etc. Also the institute takes certain actions that help students in the placement activities. This project aims at analysing the different attributes on which a student's placement depends. It aims at predicting the type of company the student would most probably get placed in. Depending on the this, the faculties and TPO can take certain actions , especially for those aspects that matter the placement the most. More efforts in these fields will help in better placement results. Our system takes input in form of categories for different attributes. It then predicts the class label (Company type) for all unlabelled tuples. Also dynamic class label prediction is possible by entering the tuple details. Visualization in form of reports and graphs is available as a part of output.