

Predicting Student's Performance using CART approach in Data Science

M.E Dissertation

Submitted in partial fulfillment of the requirements

For the degree of

Master of Engineering

in

Information Technology

by

Mr. Madhav S. Vyas

Roll No. 15IF1004

(University PG Registration No. RAIT/589/23-09-2015)

Supervisor

Ms.Reshma Gulwani



Department of Information Technology

Dr. D. Y. Patil Group's

Ramrao Adik Institute Of Technology

Dr. D. Y. Patil Vidyanagar, Sector 7, Nerul, Navy Mumbai 400706.

(Affiliated to University of Mumbai)

2017



Ramrao Adik Institute of Technology

(Affiliated to the University of Mumbai)

Dr. D. Y. Patil Vidyanagar, Sector 7, Nerul, Navi Mumbai 400706.

CERTIFICATE

This is to certify that, M.E Dissertation entitled

“ Predicting Student’s Performance using CART approach in Data Science ”

is a bonafide work done by

Mr. Madhav S. Vyas

and is submitted in the partial fulfillment of the requirement for the degree of

Master of Engineering

in

Information Technology

to the

University of Mumbai

Supervisor

Ms.Reshma Gulwani

PG Co-ordinator
(Mrs.Nilima M. Dongre)

Head of Department
(Dr. Ashish Jadhav)

Principal
(Dr. Ramesh Vasappanavara)

Certificate of Approval by Examiners

This M.E Dissertation report entitled “*Predicting Student’s Performance using CART approach in Data Science* ” is a bonafide work done by *Mr. Madhav S. Vyas* under the supervision of *Ms.Reshma Gulwani* approved for the award of *Master’s Degree in Information Technology, University of Mumbai*.

Examiners :

1.

2.

Supervisors :

1.

2.

Principal :

.....

Date :

Place :

Declaration

I declare that this written submission represents my ideas in my own words and where other's ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I 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.

.....

(Signature)

(Mr. Madhav S. Vyas)

(Roll No. 15IF1004)

(University PG Registration No.

RAIT/589/23-09-2015)

Date :

Contents

Listoffigures	vi
1 Introduction	1
1.1 Predictive Modelling	2
1.2 Classification based Prediction	3
2 Literature Review and Related Work	4
3 Comparative Analysis	7
3.1 Advantages of using Decision Tree Approach for Student Prediction	8
4 Problem Statement	9
5 Proposed System	10
5.1 Classification and Regression Tree Approach	10
5.2 Applying CART in Preparing Predictive Models for Students	11
6 Implementation Detail	15
6.1 System Information	15
6.2 Performance Analysis	22
7 Conclusion	24
8 Future Scope	25
References	26

List of Figures

1.1	Prediction Process	2
5.1	Student Performance Prediction Steps	12
6.1	CSV File Upload	16
6.2	Update Database	16
6.3	Run DT and MLP Algorithm for Result Prediction Page	17
6.4	First Output Overview Page	18
6.5	Result Prediction Table: Training Data	18
6.6	Result Prediction Table: Test Data	18
6.7	Decision Tree Structure Based on Prediction Result: Scenario 1	19
6.8	Prediction Accuracy by Using Precision Score and Recall Score: Scenario 1	20
6.9	Decision Tree Structure Based on Prediction Result: Scenario 2	21
6.10	Prediction Accuracy by Using Precision Score and Recall Score: Scenario 2	22

List of Tables

6.1	Comparing results of CART and MLP on Accuracy	23
-----	---	----

Abstract

Since a few years, student's data stored in educational organisations is used to analyse their performance. Faculties are working on active teaching methodologies. It requires them to be aware of each student's current performance and identify the students who need special attention. An application to make use of the collected data is needed which will make use of the educational data mining approach to predict student's performance. Based on student's current performance and some measurable past attributes the performance can be predicted and judged to classify them among good or bad performers. A student model can be created by the previous years records and they can be helpful in telling us the factors contributing to a student's success or failure in first year IT subject. Usually in professional IT courses like engineering, it is very important to keep students interested in its core subjects like programming since the beginning. It is necessary to keep track of students performance during the first year of the course especially. Students shying away from the core subjects like programming is common phenomena in engineering colleges nowadays. It gives a scope for us to apply data science techniques to analyse and predict a student's performance in this area. A student's interest in programming or other core subjects can be monitored and encouraged as needed. It will thus enable a faculty to pay attention to the weak students and plan sessions for them accordingly.

Chapter 1

Introduction

Educational data mining is useful in studying the students records for concluding their performance pattern and make predictions at different level. A high quality education means that knowledge is being imparted to all students in a course successfully. Knowledge cannot be tested on some absolute scale but examination is one of the parameters that can be used for it. We need to understand that apart from class tests, there are some other attributes as well which contribute to a student's good or bad performance in a semester exam. The various combination of these attributes help us in finding the how they are related to a student's performance. Also it tells the most important factors which causes a student to belong to different categories of performers. Data collection is the first and very important task as it is the base on which further analysis is possible. After analysing the data the next step is to prepare predictive models. It is then used to predict student's end semester performance. The predictive models will make faculty aware of the student's who are likely to struggle during the final exams. Extracting the useful information, depends on the attributes taken into consideration and our aim of prediction. From an IT engineering course perspective, the first exposure to programming logic is quite crucial to students. It helps in building a good foundation for further subjects. Typically an engineering student learns C language as the first IT course. The reason to study student's performance in this subject is because it is a first step this will help students to learn other subjects related to IT. If he/she doesn't find this as interesting then probably he/she may not take interest in any of the IT subjects to be taught in further semesters. If a student finds something difficult to work with or learn then it naturally reflects in his performance. It is thus important to study a student's college level performance so that their semester performance can be predicted in advance. It will thus help faculty to identify and help such students who are finding difficult to learn the programming logic.

1.1 Predictive Modelling

Prediction based modeling applies statistical concepts to predict future events or classify them with respect to models. Majority of the time the event one wants to predict is in the future, but predictive modelling can also be applied to any type of unknown event, irrespective of when it happened. At times predictive models are often used to detect crimes and identify suspects, after the crime has taken place. In many cases the model is chosen on the basis of detection theory to try to guess the probability of an outcome given a set amount of input data, for example given an email determining how likely that it is spam.

Models can be made using one or multiple classifiers in an attempt to find the belonging of a value to a set. Also it can be used to find the probability of a set of data belonging to another set. Depending on requirement, predictive modelling is synonymous with, or largely combination of the area of machine learning, as it is more commonly termed to in educational or research and development contexts. When it is made use of commercially, predictive modelling is oftenly known as predictive analytics.

Predicting student's performance helps in identifying the students at risk of failing. In Predictive Mining there are 3 basic categories - 1) Classification. 2) Regression. 3) Density Estimation. With respect to predicting student's performance the application of Classification technique and Regression technique are more popular. The steps involved in a predictive mining approach are:

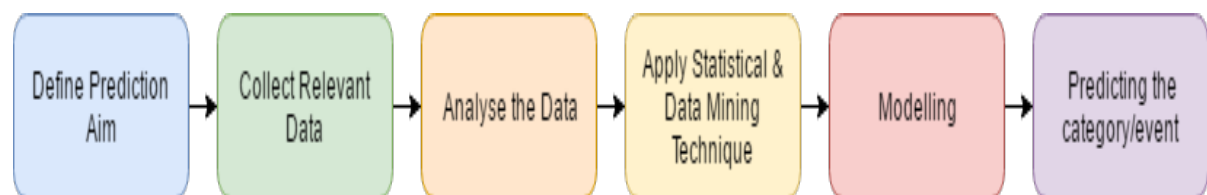


Figure 1.1: Prediction Process

1.2 Classification based Prediction

Classification is also called Supervised Learning. The training data (observations, measurements, etc) are used to learn a classifier. The training data are labeled data and the new data (unlabeled) are classified using the training data. Classification based Prediction constructs models (functions) based on some training examples. It then describes and distinguishes classes or concepts for future prediction. Finally it predicts some unknown class labels.

There are two main steps in classification

- **Step1: Model Construction** (learning step, or training step). Construct a classification model based on training data. Training data consists of:
 - A set of tuples: Each tuple is assumed to belong to a predefined class
 - Labeled data (ground truth).

A classification model can be represented by one of the following forms:

- Classification rules
 - Decision trees
 - Mathematical formulae
- **Step2: Model Usage** Before using the model, we first need to test its accuracy. To measure the accuracy of a model we need test data. Test data is similar in its structure to training data (labeled data). The known label of test sample is compared with the classified result from the model. Accuracy rate is the percentage of test set samples that are correctly classified by the model.

Evaluating Classification Methods

- **classifier accuracy:** The ability of a classifier to predict class labels
- **predictor accuracy:** How close is the predicted value from true value.
- **training time:** Time to construct the model.
- **time to use the model:** Classification/Prediction time
- **Robustness:** Handling noise and missing values.
- **Interpretability:** Level of understanding and insight provided by the model.

Chapter 2

Literature Review and Related Work

Benilda Eleoner V.Comendador, Lorena W. Rabago and Bartolome T. Tanguilig[1] studied educational data mining. Educational Data Mining process is divided into following sections-1) Data Pre-Processing; 2)Data Mining; 3)Pattern Analysis. In the first part, the user's log data is collected and preserved for analysis.Then feature extraction is carried out to understand the factors contributing to a student's performance. In the second stage, Data mining algorithm is applied based on the type of target data set used in comparison. In the final stage the information is extracted to make a knowledgeable conclusion.

Keno C. Piad, Melvin A. Ballera, Menchita Dumlao and Shaneth C. Ambat[2] stated that classification is the most popular technique used in educational data mining. It is because of it accuracy and efficiency with which it classifies data for information extraction. The authors used it for predicting IT employability . Weka is a popular tool to asses the data input given, which they also used in predicting IT employability from student's data.

Midhun Mohan, Siju Augustin and Dr Kumar Roshani V S[3] discussed about a new field in data mining known as Learning Analytics. It is a combination machine learning, artificial intelligence, information retrieval, statistics and visualization. They have used HSC CBSE students data across India for dropout prediction.

Amir Mohammad Amiri and Giuliano Armano[4] proposed an automatic method to segment heart sounds, by applying classification and regression trees approach.They implemented a heart sounds diagnostic system, to be used to help physicians in the auscultation of patients, with the goal of reducing the number of unnecessary echocardiograms and of preventing the release of newborns that are in fact affected by a heart disease.

Sonia Singh and Priyanka[5] Gupta reviewed the three most commonly used Decision Tree algorithms namely ID3, C4.5 and CART(Classification and Regression Tree). They highlighted the advantages and scenario of using each algorithm along with their limitations as well. It provided an overview of which type of decision tree algorithm used in which scenario.

Classification Technique is the most popular technique used by many researchers to study student's performance. Benilda Eleoner V.Comendador, Lorena W. Rabago and Bartolome T. Tanguilig[1] identified the most crucial factors for learning outcomes of learners.They used J48 decision tree algorithm and Multiple linear Regression to determine how likely is a student to pass in distance learning course.To study the course effectiveness they considered dropout rate as the most important factor.

In professional courses like engineering, it is a usual scenario where students do not complete their course in stipulated time frame. Also it is observed that some dropout of the course midway. Tismy Devasia, Vinushree T P and Vinayak Hegde[6] proposed a web based application which used Naive Bayesian technique to propose a student's performance in the semester examination. It predicted the students who may dropout from a course. It followed the following steps-a)Generation of data source of predictive variables. b) Identification of various features or factors which affects the performance of student's learning during academic career.c) Construction of a prediction model with the help of classification data mining techniques on the basis of predictive variables which is readily identified. d) Validation of the model which is developed for universities with students performance. The attributes they considered are gender, category, medium, food habits, other habits, location, accomodation, family type, annual income, ssc score, father qualification.

Many different data mining approaches have been used to analyse students drop out rate. They stress on how predictive analytics can hint in identifying major attributes in a student record set. Amirah Mohammad Shahiri, Wahidah Hussain and Nur'aini Abdul Rashid[7] have tried different combination of attributes such as CGPA, Internal Assessment,External Assessment and Demographics to check the accuracy of these mining algorithms. They concluded that Decision Tree and Neural Networks both performed with high accuracy. Similarly, Jiawei han Michelin Kamber[8] compared four data mining algorithms such as Decision Trees, Random Forest, Neural Networks and Support Vector Machine.It used Weka Tool to test data sets of student records and applied on various classification algorithm. It reported that Decision Tree had the highest accuracy followed by Neural Networks. Felix Castro, Alfred Vellido, Angela Nebot and Francisco Mugica[9] reviewed various data mining techniques applied to E-learning problems. For the student performance prediction category it has listed Neural Networks, Bayesian Network, Linear regression Decision Tree as the most preferred algorithms.

There has been a long time argument that data mining methods are better than statistical methods for

processing large sized data. Hina Gulati[10] took this statement as a base to identify a prediction model based on the key attributes which are relevant to predicting student's performance. Author used classification algorithm in two steps- 1)Collecting all identified attributes and apply algorithms on them. 2)Applied attribute selection algorithm and ranked all attributes on the basis of their occurrence. Here student dropout rate studied using Rule based classification using Jrip in Weka,then decision tree algorithm used to represent information of classification using J48.

With the growing digital student data, it becomes important to preserve each student's data. At the same time when so much of data is to be processed for gaining information, we need an integration of best identified algorithm on Big Data framework. Banica Logica, Radulescu Magdalena[11] presented the integration of Big Data with E-Learning System. It highlighted the influence of using 3 step architecture for a consortium of universities. They stated that such an integration will be useful analyse,organise and access huge data sets in the cloud environment. They have concentrated on dealing with unstructured data using the graphical Gephi tool.

Student Dropout prediction is an important aspect of application of educational data mining. Mohammad Nurul Mustafa, Linkon Chowdhury and Md. Sarwar kamal[12] proposed a dynamic dropout prediction model for universities, institutes and colleges. They used chi square test to separate factors such as gender, financial condition and dropping year to classify the successful from unsuccessful students. Degree of freedom is used to P-value (probability value) for best predictors of dependent variable. After being separation of factors we have had examined by using data mining techniques Classification and Regression Tree (CART) and CHAID tree.Based on results from feature selection the CHAID and the CART trees presentation it was found that the most important factors that help separate successful from unsuccessful students are financial support , age group and gender.

Various data analysis techniques used were studied by Glyn Hughes and Chelsea Dobbins[13].As such, the area of machine learning is a popular area of research that can be applied to heterogeneous sets of data to find patterns for predictive modeling, i.e., training data is used to predict the behavior of the previously unseen test data. This type of learning can either be supervised (classification), where the data is labeled to determine how powerful the algorithm is at learning the solution to the problem, or unsupervised (clustering), where the data is unlabeled and the system forms natural groupings (clusters) of patterns automatically.

Chapter 3

Comparative Analysis

After reviewing all related papers we came to know about the the different data mining algorithms which are used in predictions. Some studies highlighted that If there is a high non-linearity and complex relationship between dependent & independent variables, a decision tree model will outperform a classical regression method in predicting student performance. Also if we need to build a model which is easy to explain to people, a decision tree model will always do better than a linear model. Decision tree models are even simpler to interpret than linear regression. A decision tree is able to process both numerical and categorical data. Also Large amounts of data can be analysed using standard computing resources in lesser time.

The implementation of Data Science in various domains due to increasing data have also proposed a need to use the combination of data mining and statistical approach to predict student's performance. Student's prediction has been general in nature. Course specific requirements are different which was not a part of previous studies. Also the combination of personal life factors are more prominent in school level. At college the personal life background does not contribute much to student's performance.

3.1 Advantages of using Decision Tree Approach for Student Prediction

- 1. Easy to Understand:** Decision tree output is very easy to understand even for people from non-analytical background. It does not require any statistical knowledge to read and interpret them. Its graphical representation is very intuitive and users can easily relate their hypothesis.
- 2. Useful in Data exploration:** Decision tree is one of the fastest way to identify most significant variables and relation between two or more variables. With the help of decision trees, we can create new variables / features that has better power to predict target variable.
- 3. Less data cleaning required:** It requires less data cleaning compared to some other modeling techniques. It is not influenced by outliers and missing values to a fair degree.
- 4. Data type is not a constraint:** It can handle both numerical and categorical variables.
- 5. Non Parametric Method:** Decision tree is considered to be a non-parametric method. This means that decision trees have no assumptions about the space distribution and the classifier structure.

Chapter 4

Problem Statement

IT undergraduate students at times find their first encounter difficult with programming. It is important to analyse students from the semester beginning and predict their performance in end semester examination. A system is required which will help the faculty in classifying students and focus on the students which the system shows probable poor performers. For this the right set of attributes have to be considered. Some authors in the past have used attributes which in practical life do not contribute much to a student's performance. A work specifically to address this programming challenge in undergraduate courses have not been worked upon so far. As per a survey by times 91.28% candidates lack programming skills[14]. A majority of the students describe their first experience with programming as escaping ones, where they did not learn enough to be good at it. Infact majority students did not clear the subject in their first attempt during course.

Chapter 5

Proposed System

Tree based learning algorithms seem to be one of the best and most useful predictive mining algorithms. Tree based methods empower student predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they will be able to map non-linear relationships among students attributes quite well. They are adaptable at solving any kind of problem at hand (classification or regression). Data Science application for predicting future outcomes can be applied to predict a student's performance as well. In order to prepare predictive models the Classification and Regression Trees approach can be used. It is used to be able to classify students as per the different predictive models. To this tree structure model an Ensemble technique can be applied to improve its efficiency. Data Science field is growing with availability of statistical analyses and visualization packages. It thus makes it easier to develop predictive models based on the prime attributes responsible for performance outcome.

5.1 Classification and Regression Tree Approach

CART stands for Classification and Regression Trees. It is characterized by the fact that it constructs binary trees, namely each internal node has exactly two outgoing edges. When provided, CART can consider misclassification costs in the tree induction. It also enables users to provide prior probability distribution. An important feature of CART is its ability to generate regression trees. Regression trees are trees where their leaves predict a real number and not a class. In case of regression, CART looks for splits that minimize the prediction squared error (the least-squared deviation). The prediction in each leaf is based on the weighted mean for node. It has the following advantages-

- CART can easily handle both numerical and categorical variables.
- CART algorithm will itself identify the most significant variables and eliminate non-significant ones.
- CART can easily handle outliers.

The reason for using CART for student performance prediction is that it can handle outlier values. In student performance prediction it can handle the not so relevant values and ignore such factors for splitting the tree structure.

5.2 Applying CART in Preparing Predictive Models for Students

Here to predict a student's performance, attributes to be considered are -

- Existing knowledge value(based on whether student had undertaken vocational course)
- Attendance
- Class test marks(College Term tests average)
- Lab activity outcomes(program output shown/algorithm verified)

CART algorithm will be applied to prepare the following predictive models

- Likely to Pass students
- Likely to Fail students

The system will classify students as per the predictive models they will belong to. It will be based on the attributes listed above. It can be explained with the following steps:

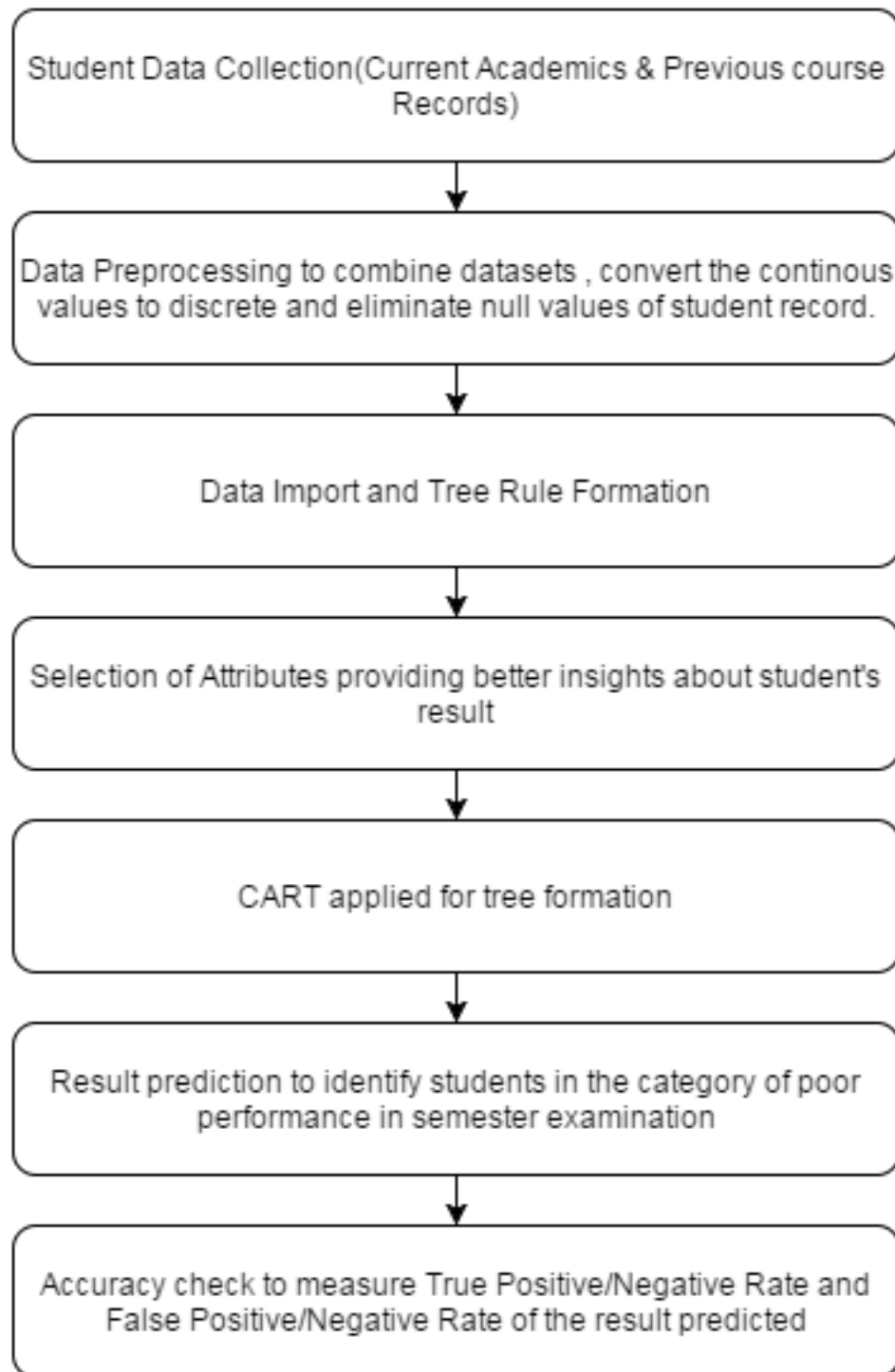


Figure 5.1: Student Performance Prediction Steps

Step I: Student Data Collection. Here the aim is to collect relevant background information of student like his name, gender, hsc percentage, vocational programming course done in high school (yes/no). Also his/her current data is collected in different phases like attendance, term test marks, lab activity marks (based on code executed and explained).

Step II: Data gets preprocessed. Changes are done to the values of some missing or incorrect format values. For e.g. if educational background value is not showing whether the student did a vocational course or not then we have to set such null-values to a numeric value. Then, the continuous variables are converted into discrete variables, which provide a much more comprehensible view of the data. For e.g. gender value can be converted from male/female to 0's and 1's respectively.

Step III: Data is imported and decision Tree formation begins, as per the decision rule of the system and splitting criteria. Here since we are going to use CART algorithm, so the splitting criteria will be of Gini Index. For e.g. there will be different student models as per the rule defined in the tree. So if students belong to 3 different categories of performers equally then Gini Index will be highest. In ideal case it is 0.5. If the students model shows that majority belong to the same model then Gini Index is minimum.

Step IV: Attributes are chosen in varied forms like Correlation Based Feature Selection (CFS). The attributes that are chosen using the Feature Selection Techniques are called, Correlation Based Feature Selection. It selects the set of features that are strongly related with class, in addition to those features those are less related. For e.g. If the student data set reveals that majority of the good performers in the programming subject are from some programming vocational background or the people with good performance in lab and with good attendance are the good performers. Then such attributes are taken mainly to predict the student's performance.

Step V: Data obtained will be classified by implementing classification and regression tree. Here the students can be classified into different category of performers. The Gini Index is taken to choose the test attribute at every node of the decision tree.

Step VI: Student's classification based prediction result is generated. They are classified as likely to fail or pass with values 0's and 1's respectively. A Classification and Regression Tree based result and Multilayer Perceptron based result table is generated along with the actual result column.

Step VII: At last the accuracy of the result predicted is measured by checking accuracy related terms like True positive rate, False positive rate, True negative rate and False negative rate. Here Precision Score and Recall Score is used to check the accuracy of the result generated by Classification and Regression Tree Approach and Multilayer Perceptron (MLP) approach, as MLP is one of the most used technique after Decision Tree for research done in prediction related work.

Chapter 6

Implementation Detail

6.1 System Information

The core project concept of using Data Science was implemented using Python Language. The analysis and prediction code was first tested on IPython (Jupyter) Notebook to understand the prediction process. Then for the system development the server side scripting language used is PHP, the Database connectivity is done using MySQL. Python is used along with PHP during the data processing and prediction role for the system to be able to develop predictive models. My system specification for this project are Windows 8(64 bit), with Core i3 Processor. The main process of this system is to upload student data, divide it into training and test data. Then the system learns from training data and tries to make prediction based on it. CART technique produces a tree structure to make the understanding of splitting criteria and factors contribution importance clear. Then at last we have measured accuracy of Classification and Regression Tree algorithm with the help of Precision Recall Algorithm. Also the accuracy check is done for the results obtained by Multilayer Perceptron Classifier.

The System working is as follows:

- Admin Updates CSV File containing student data

Upload Files

Year: Select

Upload CSV: Choose File No file chosen

Submit

Sr.No.	Class Year	Updated To DB	Uploaded On	Action
1	2015	No	24 Apr 2017 01:06 PM	Delete

Figure 6.1: CSV File Upload

- Student Data file is to Updated to the System's Database

Add Data to DB

Sr.No.	Class Year	Updated To DB	Uploaded On	Action
1	2015	No	24 Apr 2017 01:06 PM	Please run DB script

Figure 6.2: Update Database

- The Classification and Regression Tree Algorithm along with Multilayer Perceptron Algorithm are run to process the data and generate Output Page Of Prediction

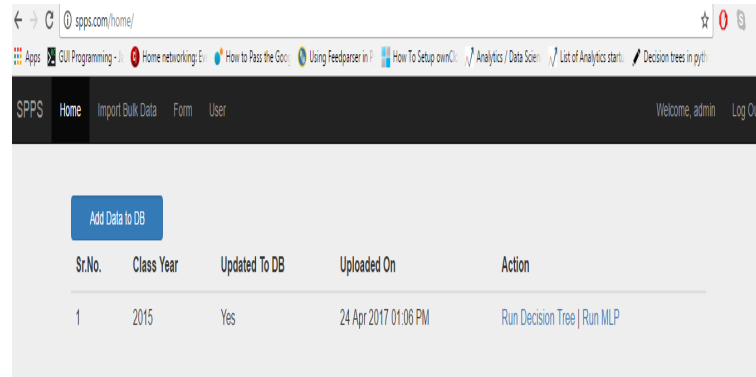


Figure 6.3: Run DT and MLP Algorithm for Result Prediction Page

- The Admin lands on Output Page where the following elements of the system are visible:

Prediction of Class Result

Tree Structure

Comparison between CART and MLP Performance

Precision Score and Recall Score

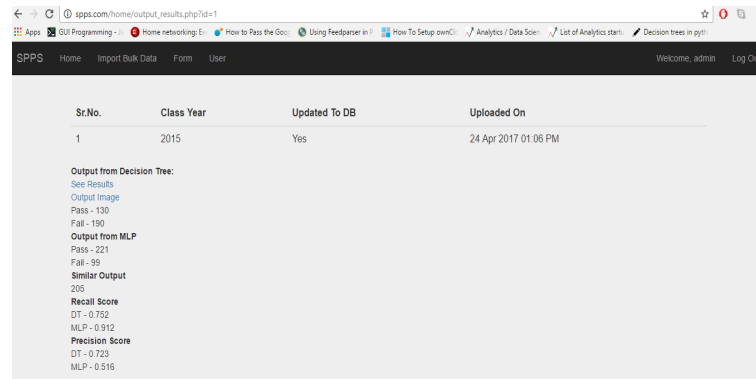


Figure 6.4: First Output Overview Page

- A Tabular Data is generated which shows the actual data, predicted data from CART and predicted data from MLP. It consists of Training Data and Test Data

Roll No.	First Name	Last Name	Actual Result	DT Result	MLP Result
101	madhav	vyas	1
102	aditya	vyas	1
103	poopa	sharma	1
104	snetha	chopra	1
105	prathic	vaidya	1
106	vishakha	Narote	1
107	Ritesh	Pradham	1
108	Aishwarya	Shetty	1
109	Rishabh	Prasadgiri	1
110	aniket	singh	1
111	Mayuresh	Shinde	0
112	Shikhar	Rattan	1
113	Siddhesh	Shinde	1

Figure 6.5: Result Prediction Table: Training Data

Roll No.	First Name	Last Name	Actual Result	DT Result	MLP Result
181	SHRAVYA	SHETTY	0	0	0
182	ADITYA	SINGH	0	0	0
183	ASHA	WAGH	0	0	0
184	SNEHA	SAGAR	1	1	1
185	SHREYA	MOOLYA	0	0	0
186	ROHIT	AMBRE	0	0	0
187	TEJAL	KAMBLE	0	0	0
188	SHARVARI	BARGE	0	0	0
189	SUPRIYA	SURYAWANSHI	0	0	0
190	GITANJALEE	JAGDALE	1	0	1
191	ANKITA	SANGLE	1	1	1
192	SHUBHADA	AHINAVE	0	0	0

Figure 6.6: Result Prediction Table: Test Data

- On clicking the 'Output Image' link the Decision Tree structure is created highlighting the factors crucial for result prediction viz Fail(0) or Pass(1) based on the Splitting Criteria. It is helpful in identifying the factors based

on which result is dependent most

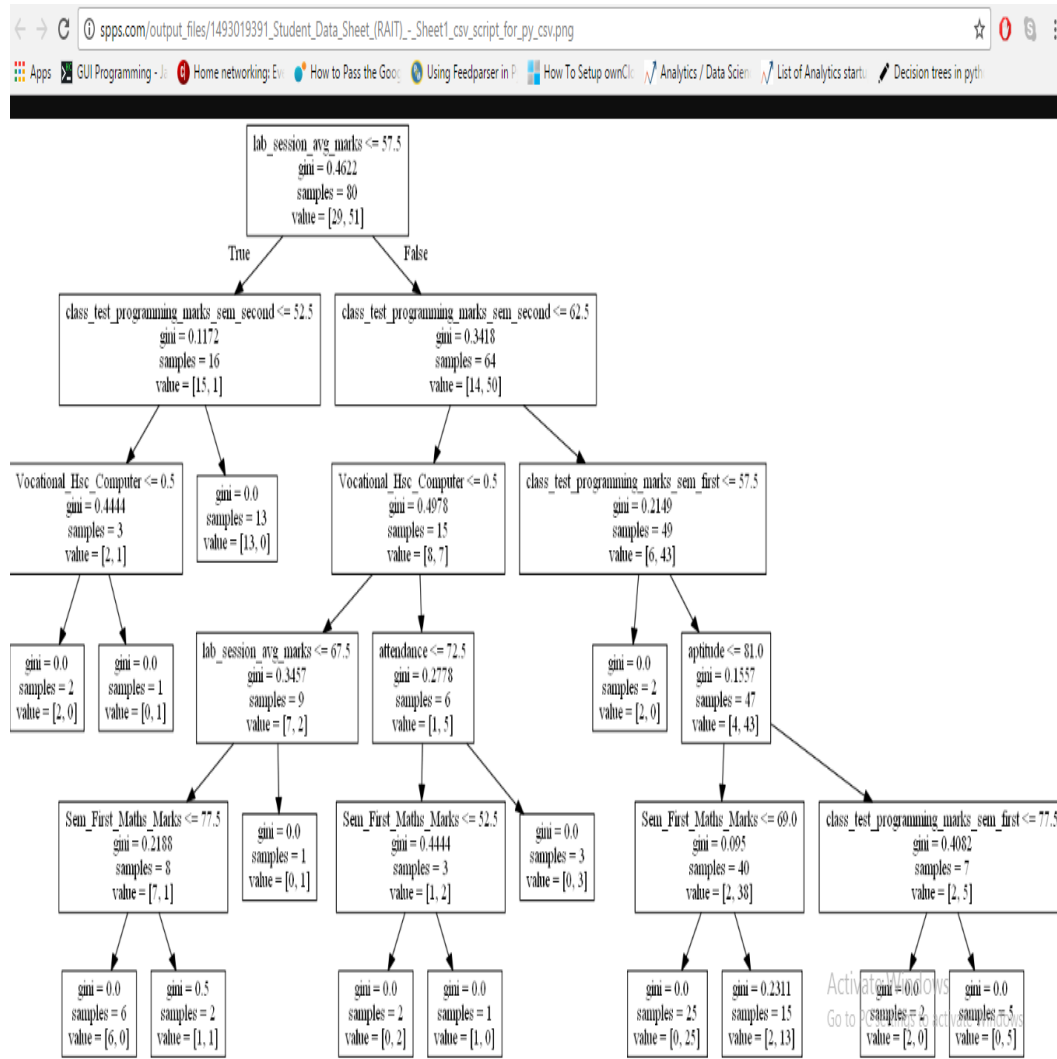


Figure 6.7: Decision Tree Structure Based on Prediction Result:Scenario 1

Output from Decision Tree:
[See Results](#)
[Output Image](#)
Pass - 130
Fail - 190
Output from MLP
Pass - 221
Fail - 99
Similar Output
205
Recall Score
DT - 0.752
MLP - 0.912
Precision Score
DT - 0.723
MLP - 0.516

Figure 6.8: Prediction Accuracy by Using Precision Score and Recall Score:Scenario 1

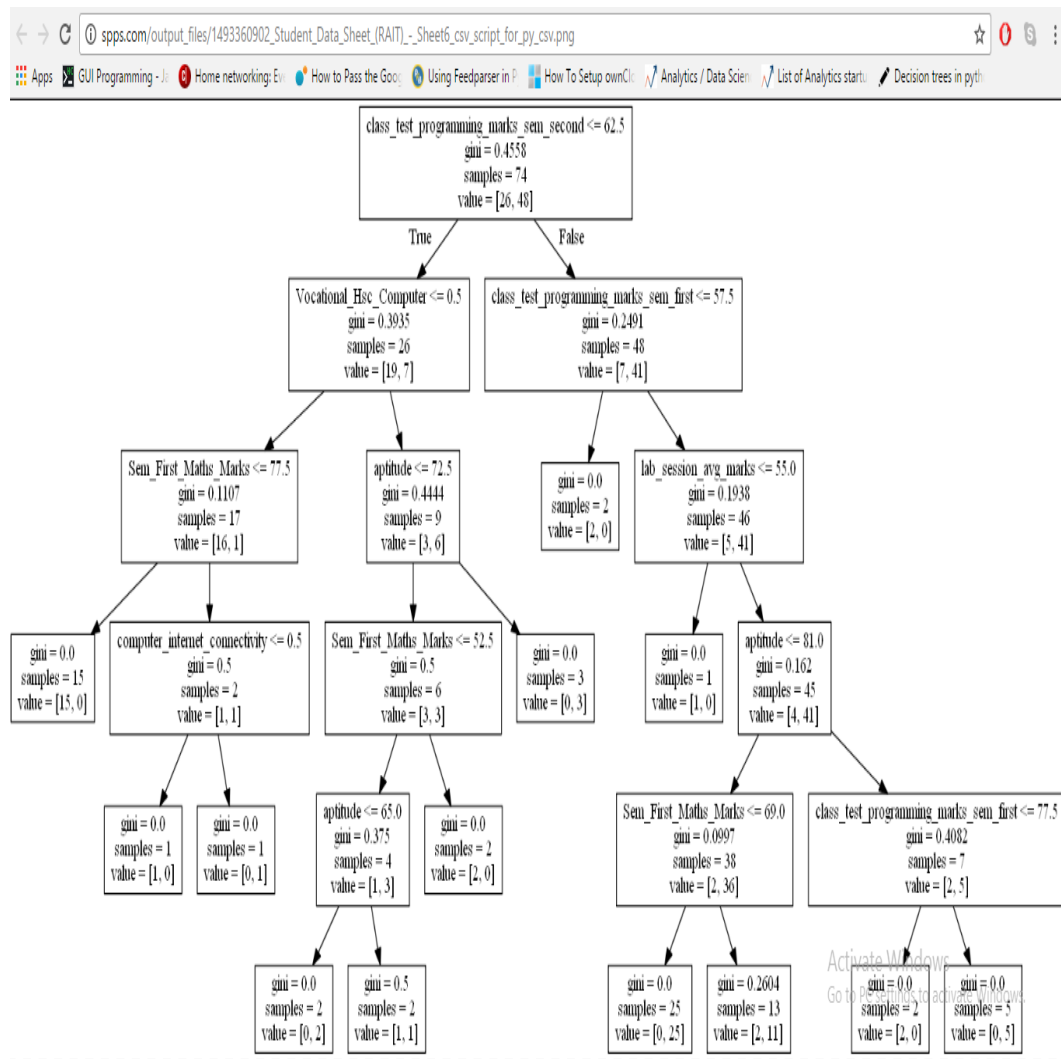
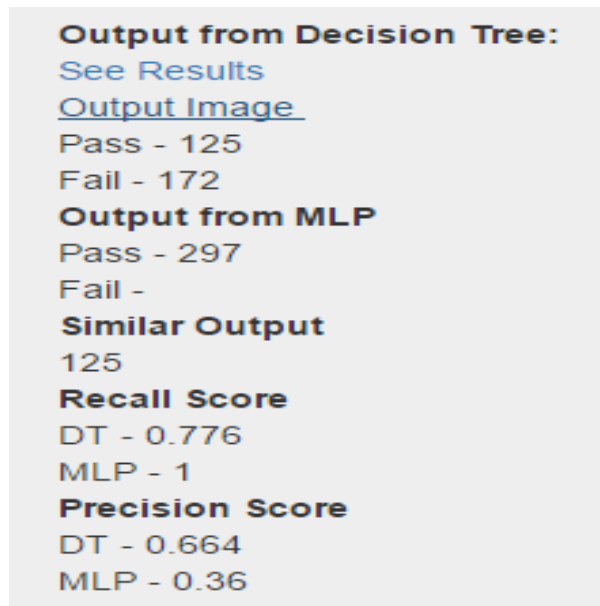


Figure 6.9: Decision Tree Structure Based on Prediction Result:Scenario 2

6.2 Performance Analysis

A comparative indicator for the accuracy of result is shown in the output page. Precision Score and Recall Score is generated to understand the prediction accuracy. It compares the Decision Tree's accuracy and MLP's accuracy with respect to correct result prediction.



The screenshot displays the output of a web application. It features a light gray background with black text. The content is organized into several sections, each starting with a bold header. The first section, 'Output from Decision Tree:', includes a blue link 'See Results' and a blue link 'Output Image'. Below these are the counts 'Pass - 125' and 'Fail - 172'. The second section, 'Output from MLP', shows 'Pass - 297' and 'Fail -'. The third section, 'Similar Output', lists the number '125'. The fourth section, 'Recall Score', provides 'DT - 0.776' and 'MLP - 1'. The final section, 'Precision Score', shows 'DT - 0.664' and 'MLP - 0.36'.

```
Output from Decision Tree:
See Results
Output Image
Pass - 125
Fail - 172
Output from MLP
Pass - 297
Fail -
Similar Output
125
Recall Score
DT - 0.776
MLP - 1
Precision Score
DT - 0.664
MLP - 0.36
```

Figure 6.10: Prediction Accuracy by Using Precision Score and Recall Score:Scenario 2

Technique	Precision	Recall	Conclusion
Classification and Regression Tree	0.723 & 0.664	0.752 & 0.776	Higher accuracy detected when compared with prediction result table.
Multilayer Perceptron	0.516 & 0.36	0.912 & 1	Accuracy Lesser detected when compared with prediction result table.

Table 6.1: Comparing results of CART and MLP on Accuracy

The Results for different scenarios suggests that the trees structure highlighting the factor responsible for classifying the students may vary as per the student's class records. In the first scenario based on Decision Tree, Lab session is more responsible as it is on the root node. In the second scenario the Class test marks determine the classification. Also in the second scenario the Lab session performance and vocational background is among the high value splitting criteria. So we can say that the Lab session and Vocational background are the major factors for a student to be classified. The comparison shows that accuracy based on trade off between precision score and recall score is better for Classification and Regression Tree than Multilayer Perceptron when compared with the actual results.

Chapter 7

Conclusion

Student's enrolling in the first year of a graduation are crucial to be monitored for their performance. It is a crucial learning phase of their career and so it is necessary to be able to analyse their current performance in IT related subjects. Based on their existing knowledge profile and other performance during the semester, their final exam performance has to be predicted. So faculties can pay more attention to students likely to fail based on the prediction report. Once equipped with a good base of programming and/or any IT foundation course, they will develop more interest in their field of study. The outcome will be that every student will be able to learn the subject in the coming semesters in a better way.

Chapter 8

Future Scope

The Prediction Results obtained by using CART Approach in Data Science can also be taken ahead from other perspective. It can be used to suggest students about different working profiles that exists in IT industry. As per their programming or other skill set score related to IT, a recommendation system can be developed. After predicting their technical skills a career path can be recommended to students.

References

- [1] B. E. V. Comendador, L. W. Rabago and B. T. Tanguilig, "An educational model based on Knowledge Discovery in Databases (KDD) to predict learner's behavior using classification techniques," 2016 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Hong Kong, 2016
- [2] K. C. Piad, M. Dumlao, M. A. Ballera and S. C. Ambat, "Predicting IT employability using data mining techniques," 2016 Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC).
- [3] M. G. M. Mohan, S. K. Augustin and V. S. K. Roshni, "A BigData approach for classification and prediction of student result using MapReduce," 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS).
- [4] A. M. Amiri and G. Armano, "Early diagnosis of heart disease using classification and regression trees," The 2013 International Joint Conference on Neural Networks (IJCNN).

- [5] Sonia Singh and Priyanka Gupta, "Comparative Study ID3, CART AND C4.5 Decision Tree Algorithm: A Survey" 2014 International Journal of Advanced Information Science and Technology (IJAIST).
- [6] T. Devasia, Vinushree T P and V. Hegde, "Prediction of students performance using Educational Data Mining," 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE).
- [7] Amirah Mohamed Shahiri, , Wahidah Husain, Nur'aini Abdul Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques", 2015 The Third Information Systems International Conference.
- [8] Jiawei han Michelin Kamber, "Data mining: concepts and techniques" 2000 The Morgan Kaufmann Series in Data Management Systems.
- [9] Felix Castro, Alfred Vellido, Angela Nebot and Francisco Mugica, "Applying Data Mining Techniques to e-Learning Problems" 2006 Proceeding WBE'06 Proceedings of the 5th IASTED international conference on Web-based education.
- [10] Hina Gulati, "Predictive analytics using data mining technique," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom).
- [11] Banica Logica, Radulescu Magdalena, "Using Big Data in the Academic Environment" 2015 7th International Conference, The Economies of Balkan and Eastern Europe Countries in the changed world.
- [12] M. N. Mustafa, L. Chowdhury and M. S. Kamal, "Students dropout prediction for intelligent system from tertiary level in developing coun-

try,” 2012 International Conference on Informatics, Electronics & Vision (ICIEV).

- [13] Glyn Hughes and Chelsea Dobbins, ”The utilization of data analysis techniques in predicting student performance in massive open online courses (MOOCs)” 2015 Research and Practice in Technology Enhanced Learning, Springer Open.
- [14] Parneet Kaur, Manpreet Singh and Gurpreet Singh Josan ”Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector” 2015 Procedia Computer Science, Elsevier, Volume 57.
- [15] <http://timesofindia.indiatimes.com/city/mumbai/Only-18-engineering-grads-are-employable-says-survey/articleshow/38438996.cms>
- [16] <http://scikit-learn.org/>
- [17] <https://www.analyticsvidhya.com/>
- [18] <http://ipython.org/documentation.html>
- [19] <https://www.edx.org/course/introduction-python-data-science-microsoft-dat208x-5>
- [20] <http://pandas.pydata.org/pandas-docs/stable/>

Papers Presented in Conference

- [1] "Predicting Students Performance using CART approach in Data Science", IEEE International conference on Electronics, Communication and Aerospace Technology (ICECA 2017), RVS Technical Campus, Coimbatore on 20-22 April 2017.
- [2] "Predictive Analytics for E-Learning System", International Conference on Inventive Systems and Control (ICISC 2017), JCT College Of Engineering and Technology, Coimbatore on 19-20 January 2017.

Acknowledgement

I take this opportunity to express my profound gratitude and deep regards to my guide **Ms. Reshma Gulwani** for her exemplary guidance, monitoring and constant encouragement throughout the completion of this report. I am truly grateful to her efforts to improve my technical writing skills. I am also thankful to **Dr. Ashish Jadhav**, Head of Department of Information Technology and PG Co-ordinator **Mrs. Nilima M. Dongre**. I take this privilege to express my sincere thanks to **Dr. Ramesh Vasappanavara**, Principal, RAIT for providing the much necessary facilities . Last but not the least I would also like to thank all those who have directly or indirectly helped me in completion of this report, Department of Information Technology, RAIT, Nerul, Navi Mumbai.

Mr. Madhav S. Vyas

15IF1004