EDUCATIONAL DATA MINING

Minor Project 1

Submitted by:

RUPANSH VIJ (9915103098)

ARPIT GUPTA (9915103113)

ASHWYN SINGH (9915103114)

DISHA NAGPAL (9915103117)

Under the supervision of

Mrs. Akanksha Bhardwaj

Department of CSE/IT

Jaypee Institute of Information Technology University, Noida

NOVEMBER 2017

ACKNOWLEDGEMENT

We would like to place on record our deep sense of gratitude to **Dr. Hariom Gupta**, Director, Jaypee Institute of Information Technology, India for his generous guidance, help and useful suggestions.

We express our sincere gratitude to **Mrs. Akanksha Bhardwaj**, Department of CSE, JIIT, for her invigorating guidance, continual encouragement and supervision throughout the course of present work.

We also wish to extend our thanks to our batch mates for their insightful comments and constructive suggestions to improve the quality of this project work.

Signature(s) of Students

RupanshVij (9915103098)

Arpit Gupta (9915103113)

Ashwyn Singh (9915103114)

Disha Nagpal (9915103117)

ABSTRACT

Educational Data Mining refers to techniques, tools, and research designed for automatically extracting meaning from large repositories of data generated by or related to people' learning activities in educational settings. This project highlights the importance of using student data to drive improvement in education planning. The need for prediction of a student's performance is to enable the university to intervene and provide assistance to low achievers as early as possible Working on the same principle we are analysing the result (grades) of students from a particular university by first collecting the data via Google form. Pre-processing the collected data, selecting the attributes and then the dimensionality of data set is reduced by a technique known as PCA. Moreover Decision Tree is applied on test set to predict their grades at the end of the first year.

Keywords—Data Mining; Education; Students; Performance

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVATIONS

1. DM: Data Mining
2. RDM: Relational Data Mining
3. KDD: Knowledge Discovery in Database
4. EDM: Educational Data Mining
5. PCA: Principle Component Analysis

Chapter 1

INTRODUCTION

1.1 EDUCATION DATA MINING

Educational Data Mining (EDM) is a new trend in the data mining and Knowledge Discovery in Databases (KDD) field which focuses in mining useful patterns and discovering useful knowledge from the educational information systems, such as, admissions systems, registration systems, course management systems and any other systems dealing with students at different levels of education, from schools, to colleges and universities [6].

1.2 PURPOSE

The main purpose is to analyze the students' data and information to classify students, or to create association rules, to make better decisions or to enhance student's performance is an interesting field of research, which mainly focuses on analyzing and understanding student's educational data that indicates their educational performance, and predictions to help students in their future educational performance. Educational Data Mining (EDM) is used to analyze collected student's information through a survey, and provide classifications based on the collected data to predict students' performance in their upcoming semester.

1.3 APPROACH

In our project we propose an approach based on Data Mining techniques that are able to predict the grades of students in first year based on dimensionality reduction using Principal Component Analysis that draws inferences from datasets consisting of input data and Decision Tree(J48) algorithm based on a supervised learning algorithm where you need a target variable for prediction.

CHAPTER 2

BACKGROUND STUDY

Education is a high priority of world society, which claims to enhance the scope, quality, efficiency, and achievements of educational systems. According to the fast evolution of computers, communications, internet, and heterogeneous platforms that facilitate the interaction of man–machine anywhere and anytime, the educational settings and data are growing exponentially. Educational data mining is a field for solving education-related problems [5]. It seeks solutions to improve academic performance for the learners.

Through EDM, we try to understand how different individuals engage with the educational system. These approaches can be applied into modeling students' individual differences and respond to those differences by providing a way, which help improve students' performance.

2.1 DIMENSIONALITY REDUCTION

Principal component analysis is a method of extracting important variables (in form of components) from a large set of variables available in a data set. It extracts low dimensional set of features from a high dimensional data set with a motive to capture as much information as possible. With fewer variables, visualization also becomes much more meaningful. PCA is more useful when dealing with 3 or higher dimensional data. It is always performed on a symmetric correlation or covariance matrix. This means the matrix should be numeric and have standardized data.

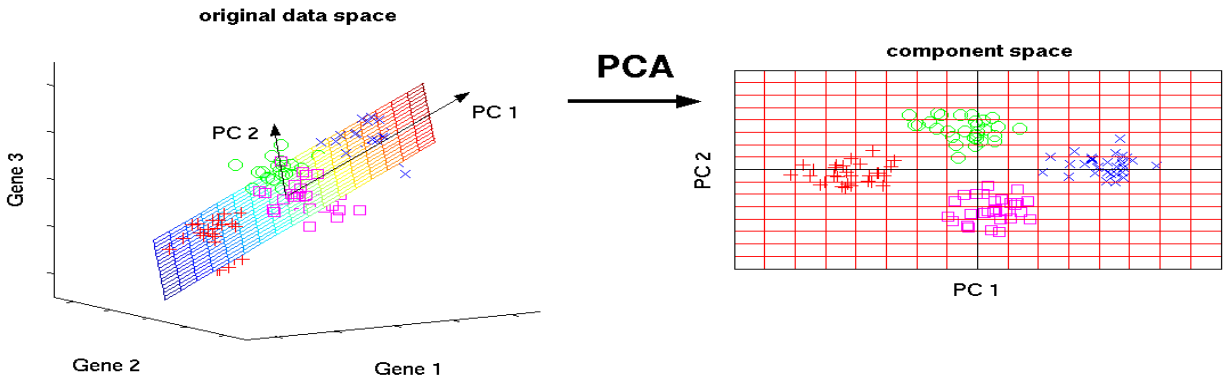

Figure 1.Transformation of a high dimensional data to low dimensional data using PCA (Source:https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/)

2.2 DECISION TREE

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is based on the principle of Divide and Conquer.

Decision tree builds classification 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.

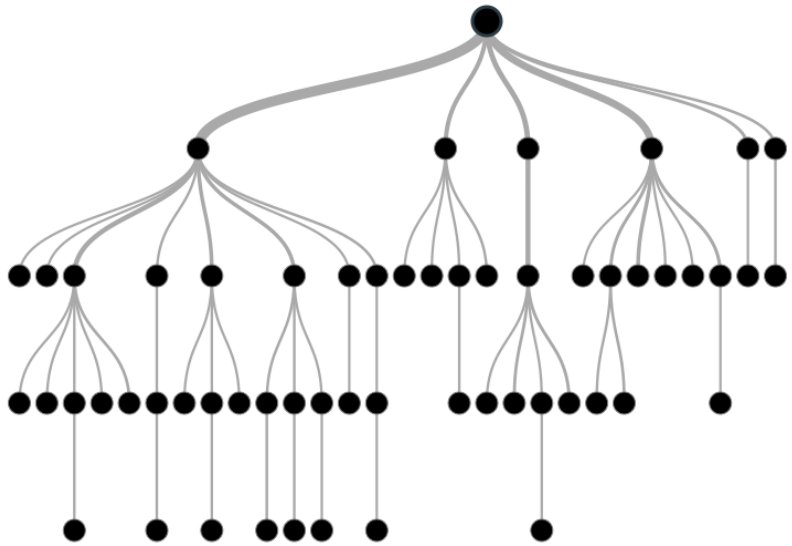

Figure 2. A decision tree layout
(Source:https://blog.bigml.com/2012/01/23/beautiful-decisions-inside-bigmls-decision-trees/)

A decision tree helps us make conclusions about an item's target value using its observations. It does so by making a predictive model. The J48 algorithm for constructing decision trees works top-down, by choosing a variable at each step that best splits the set of items

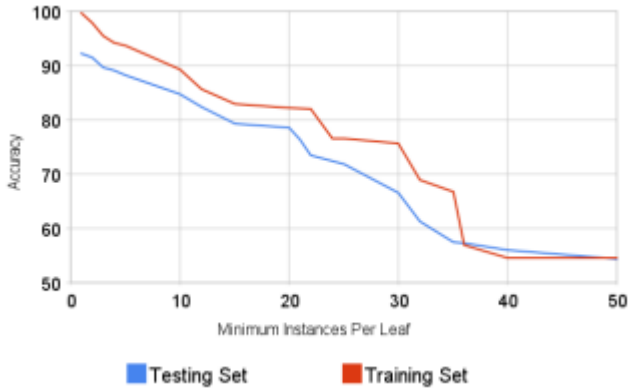

Figure 3. Classification accuracy after pruning a test domain

Proper utilization of pruning methods and techniques has shown to increase classification accuracy given a decision tree. It was found that increasing the minimum instance requirement reduced the accuracy of the classifier [3].

2.3 LITERATURE SURVEY

2.3.1 Data mining in education (2013)

Author: Cristobal Romero and Sebastian Ventura

URL: https://pdfs.semanticscholar.org/c73b/0424e1a4ab2574cfce2e41c505f71f46940e.pdf

According to this paper, applying data mining in education is an emerging interdisciplinary research field also known as educational data mining. It is concerned with developing methods for exploring the unique types of data that come from educational environments. Its goal is to better understand how students learn and identify the settings in which they learn to improve educational outcomes and to gain insights into and explain educational phenomena.

2.3.2 Dimensionality Reduction: A Comparative Review (2009)

Author: Laurens van der Maaten, Eric PostmaJaap and van den Herik

URL:https://www.tilburguniversity.edu/upload/59afb3b8-21a5-4c78-8eb3-6510597382db_TR2009005.pdf

According to this paper, in recent years, a variety of dimensionality reduction techniques have been proposed that aim to address the limitations of traditional techniques such as PCA. The paper presents a review and systematic comparison of these techniques. The performances of

the nonlinear techniques are investigated on artificial and natural tasks. The results of the experiments reveal that PCA performs well on selected artificial tasks for dimensionality reduction, but that this strong performance may not reflect in real world tasks.

2.3.3 Predicting Student Performance: A Statistical and Data Mining Approach(2013)

Author: V.Ramesh, P.Parkavi and K.Ramar

URL: http://research.ijcaonline.org/volume63/number8/pxc3885242.pdf

This is a research paper in which a dataset was collected from high school students in Tamil Nadu schools by using a questionnaire. Several socio-economic and extracurricular related questions were asked in it. After this, the dataset was pre-processed and analyzed. The test dataset was taken from the internet. Furthermore Decision Tree (J48) and other classification algorithms were applied on the dataset for classification. It was found that J48 gave one of the best prediction accuracy among all the algorithms applied.

CHAPTER 3

REQUIREMENT ANALYSIS

3.1 SOFTWARE REQUIREMENTS

- Operating System : Window 7 or above
- Software : WEKA

3.2 HARDWARE REQUIREMENTS

- Processor : Any Processor above 500 MHz
- Ram : 1 GB.
- Hard Disk : 10 GB.
- Input Device : Standard Keyboard and Mouse.
- Output Device : VGA and High Resolution Monitor.

3.3 FUNCTIONAL REQUIREMENTS

- Google forms
- Train Set and Test Set

3.4 NON FUNCTIONAL REQUIREMENTS

- Storage
- Accessibility
- Flexibility
- Configuration
- Security

3.5 USER REQUIREMENTS

- False negative rate should be less
- Easy to operate
- Providing simple user interface
- Quick response

3.6 UML Diagrams

3.6.1 Use Case Diagram

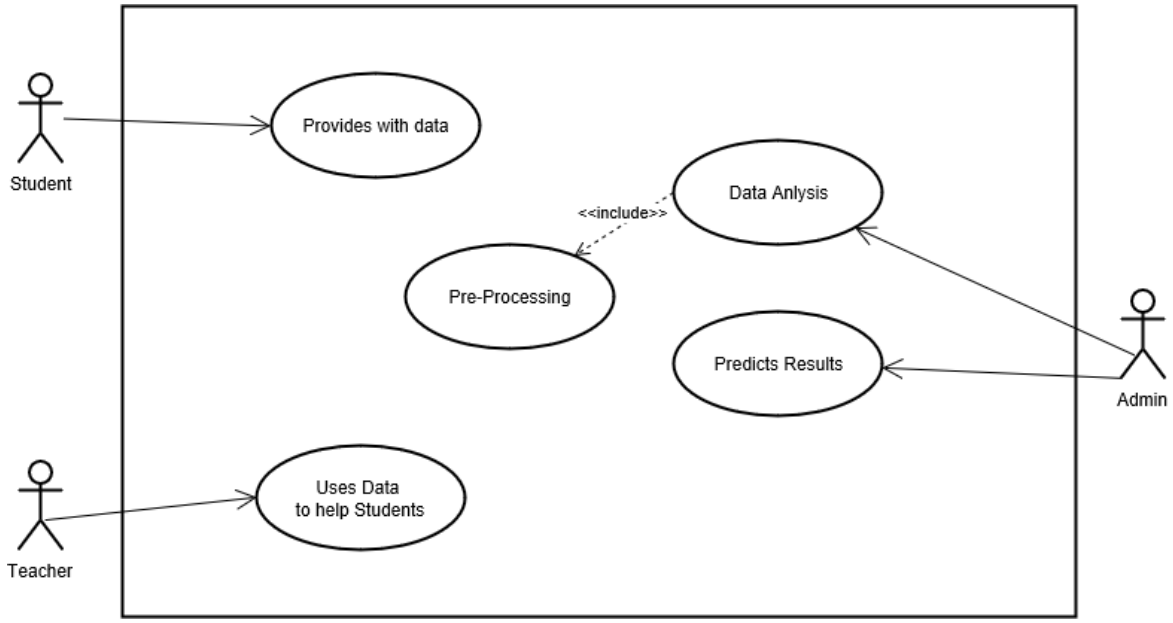

Figure 4. UML Diagram

CHAPTER 4

DETAILED DESIGN

4.1 PROCESS MODEL

A survey is conducted to collect the raw data. The categorical data is converted to numerical data using in preprocessing. Principal Component Analysis is then used to reduce the dimensionality of the dataset. J48 Decision Tree algorithm is then used to classify the dataset on basis of grades on both the reduced dataset and the preprocessed dataset. The classified model is then applied on the test set and the results are compared.

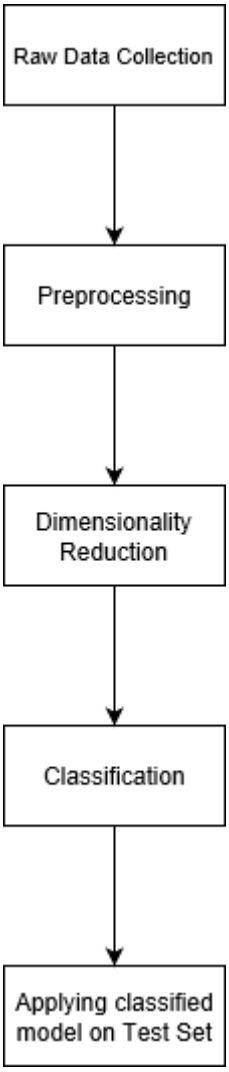

Figure 5. Flow chart

CHAPTER 5

IMPLEMENTATION

1. COLLECTION OF DATA

First of all student's data is collected with the help of Google form. The form was filled by the students of JIIT of both the branches CSE and ECE. The sample of the form is given below:

Attribute	**Description**	**Possible Values**
Branch	Student's branch	CSE, ECE
Gender	Student's Gender	Male/Female
X percentage	Percentage in X class	0-60,60-70,70-80,80-90,90-100
X board	Board opted in X	CBSE, ICSE, Other
XII Percentage	Percentage in XII class	0-60,60-70,70-80,80-90,90-100
XII Board	Board opted in XII	CBSE, ICSE, Other
Year gap	Year gap after high school	0,1, greater than 1
Region	Region you belong to	North India, South India, West India, East India
Entry	Entry in college	JIIT, Lateral
Accommodation	Where do you now live during the college year	Campus Hostel Residence (house, apartment, etc.) within walking distance of the institute. Residence (house, apartment, etc.) within driving distance.

Companion	With whom do you live during the college year	no one, I live alone, one or more other students, my parents, other relatives, friends who are not students at this institute
Computer Access	access to a computer or laptop where you live, that you can use for the college work	Yes, No
Father's Qualification	Qualification of father	Secondary School, High School, Graduate, Post Graduate
Mother's Qualification	Qualification of Mother	Secondary School, High School, Graduate, Post Graduate
Family Business	Family business	Yes, No
Liking of college	Your liking for college	I like the college, I am more or less neutral about it, I don't like it.
Co-curricular activities	Co-curricular activities	Yes, No
Programming Experience	Studied C, C++, Java, Python in school.	Yes, No

Means of travelling	Means of travelling	College Bus, Metro, Cab, Personal Vehicle, Walking.
Financial Crisis	Financial crisis in family	Yes, No
Sibling	Any sibling you have	Yes, No
Time to reach college	Time to reach college	15 min 15-30 min 30 min 1 hour >1 hour
Post-Graduation	Willing to pursue Post graduation	Yes, no
Health Status	Current health status	Very Bad, Bad,Good, Very Good, Excellent.

2. PREPROCESSING

The collected data (Train and Test) set is in nominal form and hence it is converted in numeric form with the help of method known as Frequency Encoding that works on the principle of frequency count.

3. DIMENSIONALITY REDUCTION

The training dataset is then preprocessed and taken under analysis by using PCA (Principle Component Analysis) along with Ranker method which reduces the dimensionality of our dataset resulting in the attributes that have contribution more than 70% i.e. from 23 attributes to 13 attributes respectively.

4. CONVERSION OF TEST SET

The test set is then reduced according to our train set that includes 13 attributes using the technique known as Batch Filtering which works on the command line. The code of which is given below.

** java WEKA.filters.supervised.attribute.AttributeSelection -b -i 3rdyear.csv -o train_pca.arff -r 1styear.csv -s test_pca_output.arff -E "WEKA.attributeSelection.PrincipalComponents -R 0.7 -A 5" -S "WEKA.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1" **

5. PREDICTION

Now we are done with our train and test set and now our major part i.e. prediction comes into play. For prediction various techniques can be used but we are using Decision Trees (J48 algorithm) as they are easy to understand, reliable and suits best for our dataset. We have done two predictions one without applying PCA and one after applying PCA.

The tree formed after predictionsare shown below:

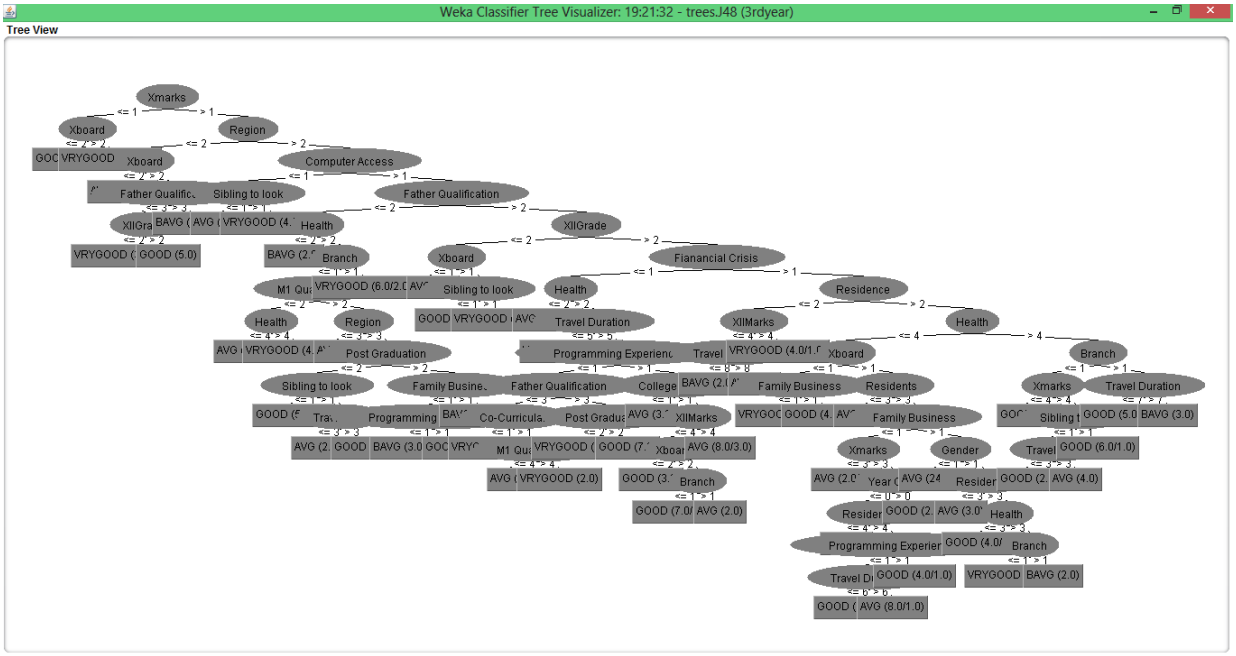

Figure 6. Complete decision tree after applying PCA.

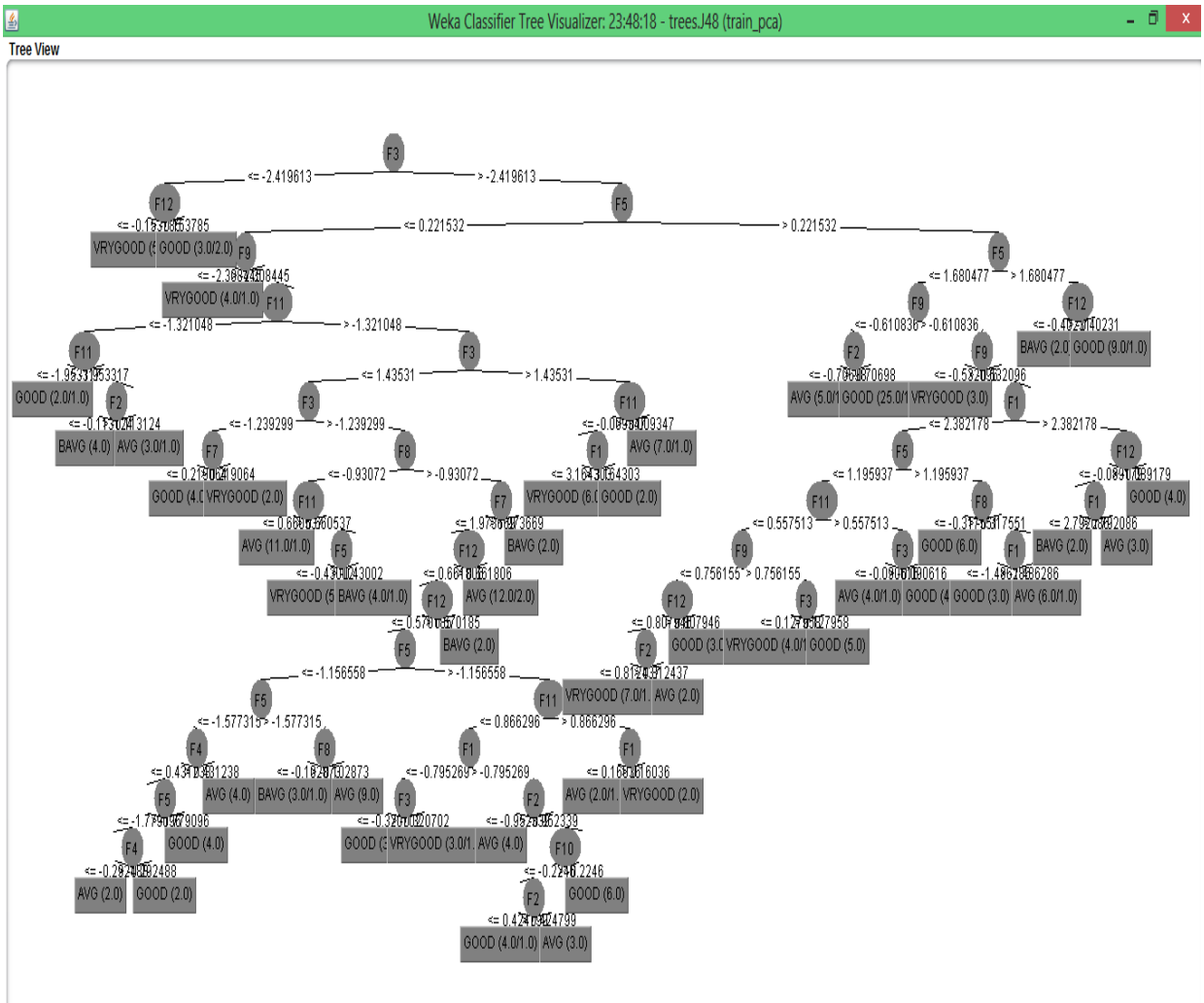

Figure 7. Complete decision tree before applying PCA.

6. RESULT

1. On applying the test set on our train set using the supplied test method we get an accuracy of 85.1986 % which means 236 instances are correctly classified out of 277 instances.

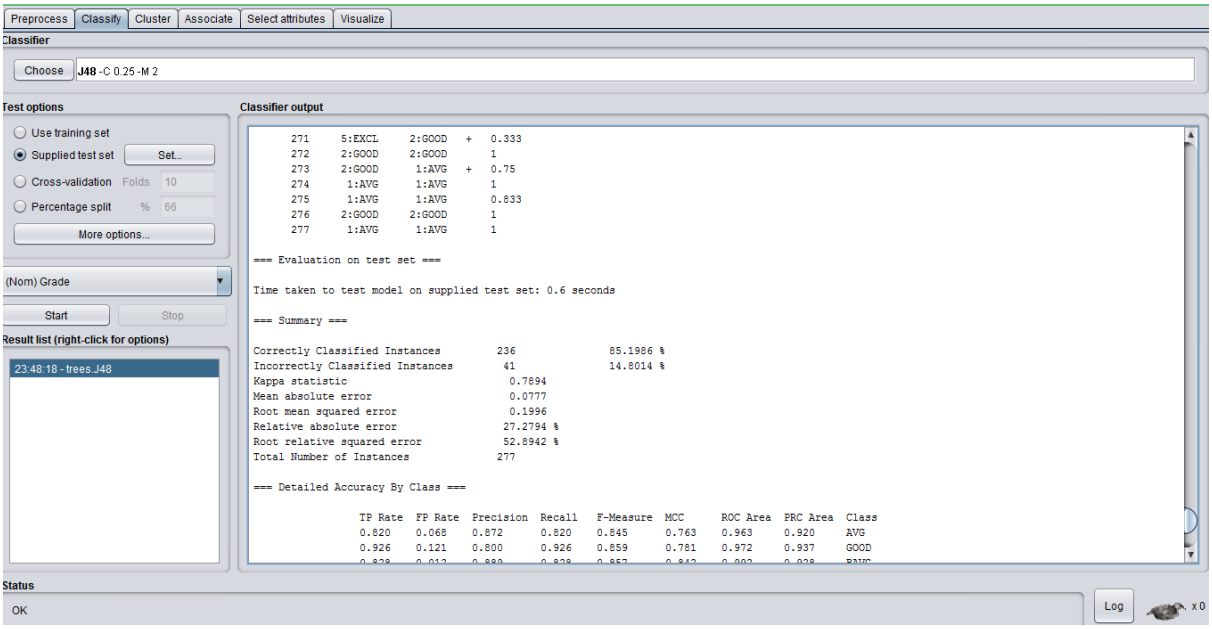

Figure 8. Accuracy of test set applied on train set after PCA.

2. On applying the test set on our train set using the supplied test method we get an accuracy of 83.3935 % which means 231 instances are correctly classified out of 277 instances. This shows that we have done prediction in correct manner.

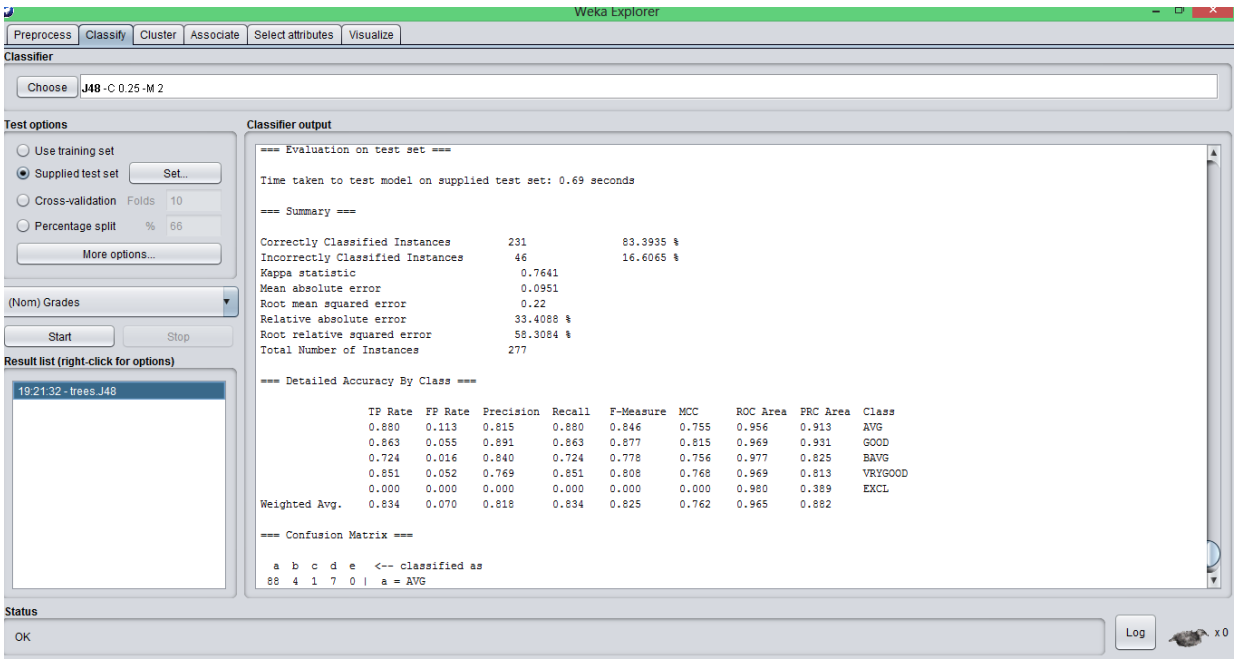

Figure 9. Accuracy of test set applied on train set before PCA.

Table 2. Table describing the result after each step

Serial No.	Task	Result
1.	Collection of student's data	We collected the dataset of 277 B.tech. students.
2.	Preprocessing	The collected dataset is converted into numerical form for further tasks.
3.	Dimensionality Reduction	The preprocessed data is reduced with the help of PCA.
4.	Conversion of Test set	Test set is converted into the suitable set for prediction using Batch Filtering.
5.	Prediction	1.When prediction is done using decision tree(J48), we get an accuracy of 83.3935% without applying PCA. 2.When prediction is done using the same algorithm on reduced dataset, we get an accuracy of 85.1986%.

Chapter 6

TESTING REPORT

Table 3. Test Table

Test Id	Modules	Type Of Testing
1.	Frequency Encoding	Black Box
2.	Batch Filtering	Black Box
3.	Threshold	Black Box

Table 4. Testing Report

Test case id	Input	Expected Output	Status
1.1	Collected Train Set	Numeric Train Set	Pass
1.2	Pre –Processed Train Set	Numeric Test Set	Pass
2.1	Test Set	Reduced Dimensions in Test Set	Pass
3.1	Threshold : 0.95	23 Attributes	Pass
3.2	Threshold : 0.7	13 Attributes	Pass

CHAPTER 7

CONCLUSION AND FUTURE SCOPE

In the current study, it was slightly found that the student's performance is not totally dependent on their academic efforts, instead, there are many other factors that have equal or greater influences as well. In conclusion, this study can motivate and help universities to perform data mining tasks on their students' data regularly to find out interesting results and patterns which can help both the university as well as the students in many ways.

Moreover, the students' data that was collected in this research included a classic sampling process which was a time consuming task, it could be better if the data was collected as part of the admission process of the university, that way, it would be easier to collect the data, as well as, the dataset would have been much bigger, and the university could run these data mining tasks regularly on their students to find out interesting patterns and maybe improve their performance.

CHAPTER 8

REFERENCES

[1] Baradwaj, Brijesh Kumar, and Saurabh Pal. "Mining educational data to analyze students' performance." arXiv preprint arXiv:1201.3417 (2012).

[2] Nisbet, John, and Jennifer Welsh. "Predicting student performance." Higher Education Quarterly 20.4 (1966): 468-480.

[3] Drazin, Sam, and Matt Montag. "Decision tree analysis using weka." Machine Learning-Project II, University of Miami(2012): 1-3.

[4] http://romisatriawahono.net/lecture/rm/survey/machine%20learning/Ayala%20-%20Educational%20Data%20Mining%20-%2020

[5] Cheng, Jiechao. "Data-Mining Research in Education." arXiv preprint arXiv:1703.10117 (2017).

[6] Voznika, Fabricio, and Leonardo Viana. "Data Mining Classification." (2007).

[7] Zuur, Alain F., Elena N. Ieno, and Chris S. Elphick. "A protocol for data exploration to avoid common statistical problems." Methods in Ecology and Evolution 1.1 (2010): 3-14.