International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 04 | Apr-2018 www.irjet.net p-ISSN: 2395-0072

Student Result Analysis System and Predicting Difficulty Level of Subject

Dhekane Rushikesh A[1], Dhumal Megha P[2], Jadhav Rutuja Yuvraj[3]

[1,2,3] *Department of Information Technology*
SVPM's College of Engineering, Malegaon, Tal-Baramati, Dist-Pune, M.S., India

---

ABSTRACT- *SPPU introduces Credit based system as a result to analyse the performance of student which was introduced by SPPU from academic year 2015-16. During engineering studies there are subject dependency, means to study advance subject knowledge of some basic subject is requires. So, We also propose the functionality through which student can predict the difficult subject in the next year based on the current understanding level of basic subject by using machine learning algorithm (C4.5 Decision Tree Algorithm).For predicting difficulty level of subject different subject dependencies are used which is used to measure performance of student during that year. Its result is based on SGPA and CGPA & Credit earn by student. Displayed result is in the form of distinction, first class, second class and higher second class, pass class and fail. The CGPA 7.75 or more than7.75 indicates Distinction, CGPA is 6.75 or less than 7.75 indicates First Class, CGPA is 6.25 or less than 6.75 indicates Higher Second Class, CGPA is 5.5 or less than 6.25 indicates pass class. The proposed system extract PDF file and convert into text file using java application which store text file into database. This record is analysing to find first three toppers of department, subject toppers, failed students, ATKT students, difficulty level of subject for student to improve their performance.*

Index terms **- CGPA, SGPA, ID3, C4.5 algorithm, decision tree, data mining, result analysis, prediction.**

1. INTRODUCTION

SPPU has recently introduce a credit based system to analysis the performance of student which was introduced in academic year 2015-2016. Result is based on SGPA and CGPA and Credit earned by student. Predicting difficulty level of subject for student is also provided by proposed system.

The proposed system makes result analysis of university annul examination and the semester examination which is based on Credit Grade Based System. Proposed system is mainly based on database technology, PDF data extraction and data mining for furcating system.

The examination result PDF is declared by the university and give to the individual colleges. This result file is in a single format and very hard to distributes in various departments and analysis toppers and fail students so this proposed system show the result which is calculated and create the separate form in distinction, first class, higher second class, second class, pass class, department wise toppers, subject wise toppers and also provide the functionality to predict the difficult subject in the next year based on current understanding level of the student. The system is developed for analysis of student result and generate the department result.

This system uses java for developing the application. Machine learning techniques are used for prediction of subject difficulty. System uses classification decision tree algorithm (C4.5 algorithm).The main aim of proposed system is to predict difficulty level and improve student's achievement and success more effectively in educational system. Describing how the analysis-forecasting result can be used to find out the factors which can affect students marks, so some negative learning habits or behaviors of students can be revealed and corrected in time

This proposed system is useful for storing student information and annual result in database and shows the student performance. The main goal of proposed system is to sort result list and find out how many students are first class toppers list, individual subject toppers, overall result and predicting the difficult subject in next year.

The proposed system is to reduce manual work to provide efficiently way of holding result data the result module the work of staff of making entries of each department result manually. It also proposed to improve students' performance in next semester.

2. LITERATURE SURVEY:

For educational measurement processes test and result analysis of annual examination based on the university exam of student. This system is mainly based on the database technology and the Choice Base Credit Based system. This system is developed for analysis of the MSc (Computer science) result [1]. This paper analyses and predicts students performance using data mining techniques for two data sets of 1000 students each one for Mathematics, and the other for System Analysis, and Design. This study can help the education community to understand learning behaviour of students as far as courses of varying difficulty are concerned. J48 supplemented by AdaBoost performs excellent for System Analysis, and Design but perform worst for mathematics and M5P generates best results for early prediction of students' marks in the major test [2]. Using decision tree algorithm C4.5 to establish a classification rule and an analysis forecasting model for students marks. The effectiveness and correctness of analysis and forecasting model and classification for students marks based on

decision tree algorithm C4.5 has been examined by an example [3].

A multiclass classification refers to the classification of the instance into more than two classes. Multiclass classification and prediction is suitable for hand written digit recognition, hand written character recognition, speech recognition and body parts recognition etc. This paper compares five classification algorithms namely Decision Tree, Nave Bayes, Nave Bayes Tree, K-Nearest Neighbour and Bayesian Network algorithms for predicting students grade particularly for engineering students. the classifiers then Bootstrap method is used to improve the accuracy of the each classifier. Bootstrap method is a resample function available in WEKA tool kit[4]. In this paper, a model is proposed to predict the performance of students in an academic organization. The algorithm employed is a machine learning technique called Neural Networks[5]. This paper presents the analysis of student performance on the basis of academic performance, research and innovation, self-development and extracurricular activities. Performance Analyzer, Score Card, Student Development, Student Performance, Student, Classification, Association rules [6]. . This paper discusses use of decision trees in educational data mining Decision tree algorithms are applied on engineering students past performance data to generate the model and this model can be used to predict the students' performance [7].

3. PROPSED SYSTEM:

A] Result Analysis System

The proposed system is to provide to the college level system to generate result analysis report from the PDF file, this pdf file is given by the university. This pdf file analysis and distribution is done by admin section. It reduces the work load of staff in college and manpower. The generated result analysis report is in simple form and easy to understand.

Modules

1] Admin Module

I]Registration:

User's (Teacher Staff's) registration is done by authorised person from department (Admin).In registration module some simple attributes are used like first name, last name, user id/name, password, email and department etc.

II] Login :

After successful registration process admin provides the user id and password to user by using this user can login to the system. i.e.Users are allowed to enter and use the system.

III] Upload PDF File:

Only Admin is allowed to upload the pdf file to the system. PDF file browse from the system storage. When attach the pdf file it will show the successful pop up message on screen "PDF File Successfully Uploaded".

2] Teacher Module

After successful upload od PDF file is extracted itext java library is used for extraction of PDF file. This library read the pdf file line by line and store the result in buffer memory to generate the result report in required format.

3] Result Generator Module

Result analysis report is display to the user in proper format. It shows the First Class, Distinction, Second Class, Pass Class, Fail, Subject wise toppers, Report is save in system.

B] Predicting Difficulty Level

By using Machine Learning Techniques is proposed to improve students' performance. The main aim of proposed system is to predict difficulty level and improve student's achievement and success more effectively in educational system.

Modules

1] **Registration and Login Module**

In registration module user(student) has to register first. After registration process is complete, authorized user can login to the system.After successful login user is allow to enter the system.

2] Upload CSV

Only admin can upload the csv file i.e. the training data set or previous data to the system. CSV is comma separated value file format. Each line is a record. Each record consists of one or more fields separated by comma.

3] Input from User Module

In this module user has to choose subject dependencies after that user need to enter the marks of subjects.

4] Prediction Module

In this module difficulty level of subject for student is predicted, according to input from students as marks of subjects. System uses Weka for predicting difficulty level.

4. ARCHITECTURE:

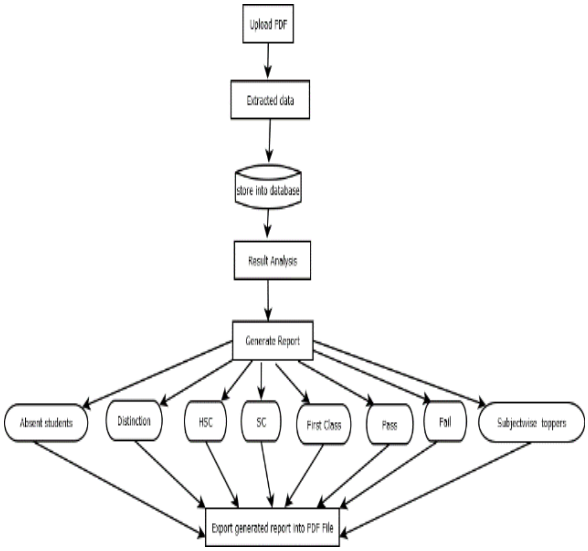

Fig. 1. result analysis system

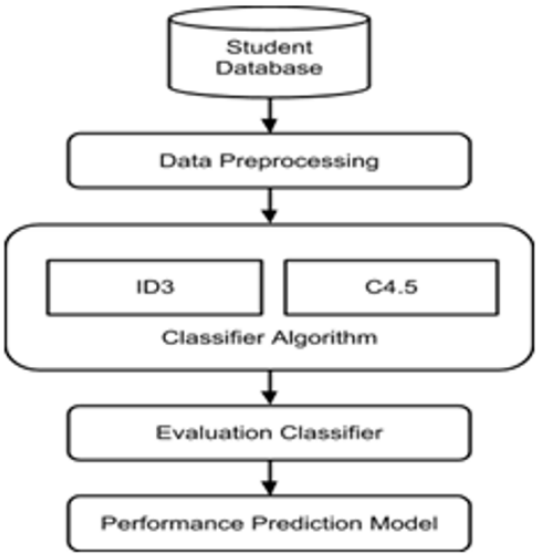

Fig. 2. prediction system

5. ALGORITHM C4.5:

I] Algorithm introduction

The C4.5 algorithm was proposed in 1992 by Ross Quinlan to overcome the limitation of the ID3 algorithm. C4.5 uses a divide and conquer approach to growing decision tree. The default splitting criteria used by C4.5 is gain ratio.

II] Algorithm Steps

1) Collection of metadata of student's marks.

2) Calculation of information gain and information gain ratio for each attribute.

Calculation of information expectation for each classification of training set

$$I(r_1,r_2,r_3,\dots r_m) = -\sum_{i=0}^{n} P_i \times \log_2 P_i \quad (1)$$

Where ,

$$P_i = \frac{r_i}{|T|}$$

Calculation of information entropy of attribute

$$E(A) = \sum_{i=0}^{n} \left({}^{T_{1j},T_{2j},T_{3j},\dots} I(T_{1j},T_{2j},\dots T_{mj}) \right) (2)$$

Where,

$$I(T_{1j},T_{2j},\dots T_{mj}) = -\sum_{i=0}^{n} P_{ij} \times \log_2 P_{ij}$$

Calculation of information gain

$$G(A) = I(r_1,r_2,r_3,\dots r_m) - E(A)$$

Calculation of information gain ratio

$$GR(A) = \frac{G(A)}{E(A)}$$

Construction of decision tree

6. EXAMPLE:

1. collection of metadata

The student marks of different subjects with some subject dependencies are collected. This data is used as training dataset for analysing and prediction purpose.The marks dataset contains M 1(maths), M 2(maths) and M3(maths) marks. M1, M2 and M3 subjects have dependencies between them.

Table 1

Sr.No	M1	M2	M3
1	80	75	70
2	74	67	74
3	78	45	72
4	35	66	63
5	66	55	65
6	23	42	30
7	67	54	60
8	78	77	66
9	90	87	80
10	42	23	35
11	77	60	65
12	33	45	50
...	...	...	...
270	86	77	71

2. Preparation of data

Arrange the available data in simplified manner. If marks less than 40, it will be written as fail in result. If marks more than 40, it will be written as pass in result. All marks are get sampled from training dataset.

TABLE 2

Sr.no	M1	M2	Result(M3)
1	42	23	Fail
2	86	77	Pass
3	77	60	Pass
...	...	...	...
270	33	45	Pass

TABLE 3

Sr.no	M1	M2	M3
Pass	172	188	159
Fail	98	82	111

TABLE 4 Combine of two subjects

M1	M3	Students No.
1	1	125
1	0	47
0	1	52
0	0	46
M2	M3	Students No.
1	1	145
1	0	43
0	1	38
0	0	44

TABLE 5 combine of 3 subjects

M1	M2	M3	Student No
1	1	1	63
1	1	0	51
1	0	1	62
1	0	0	18
0	1	1	29
0	1	0	2
0	0	1	12
0	0	0	23

3. Decision tree construction by using C4.5 algorithm

TABLE 3 shows that classifications are defined on subjects (attribute of data) from the sample training set, and two subsets (pass and fail) of students' number are contained in each classification.

The subjectM3 is selected as the tag attribute. To find out the dependencies between subject M3 and the other subjects (Decision attribute) is the purpose of constructing the decision tree. The training data set contains 270 tuples, the number of tuple in the subset corresponding to classification M3 is: pass number of student r1=159; fail number r2=111. For calculating the information gain of each decision attribute (the subjects other than M3), the information expectation of subject M3 (Tag attribute) must be calculated firstly, as per (1)

$I(r1,r2) = I(159,111)$

$= -\frac{159}{270}\log_2\frac{159}{270} - \frac{111}{270}\log_2\frac{111}{270} = 0.9770$

Entropy calculation of M1

$E(M1) = -\frac{172}{270} * I(125,47) + \frac{98}{270} * I(52,46)$

Where,

$I(172,98) = \frac{125}{172}\log_2\frac{125}{172} - \frac{47}{172}\log_2\frac{47}{172}$

$=0.8460$

$I(52,46) = \frac{52}{98}\log_2\frac{52}{98} - \frac{46}{98}\log_2\frac{46}{98}$

$=0.9972$

Finally calculated entropy is,

E (M1) = 0.6370*0.8460+0.3629*0.9972

=0.9007

Information gain of M1 is, as per (4)

G (M1) = I(r1,r2)-E(M1)

=0.9770-0.9007=0.0847

Information gain ratio of M1 is, as per (5)

$GR(M1) = \frac{G(M1)}{E(M1)}$

$= \frac{0.0763}{0.9007} = 0.0847$

The calculation method for all subjects is same for all subjects.

TABLE 6

	M1	M2
Gain	0.1345	0.0763
Gain Ratio	0.0847	0.1596

Table 6 shows the calculations of information gain and information gain ratio. From above table we know that decision attribute M2 has maximum gain ratio, so M2 will be chosen as root node in decision tree. M2 has only two values as pass or fail, therefore root node has two branches. From table 4 we see where the marks of M2 and M3 all are of 1 (pass), the student number is 145, so the ratio to all the pass number of M2 188 is:145/188=0.7712 It means that accuracy on branch 1 is 77.12% and the criterion 75% we set is fulfilled, therefore branch 1 can be stopped splitting.In branch 0 the fail number of student in M2 is 82, the fail number of M3 is 44 (Table 4), the splitting accuracy can't be fulfilled; we have to do the further splitting.

For determining the next splitting node, we calculated the information ratio of the other subjects, exclude the root node, following the above method, and the result shows that subject M1 has the maximum gain ratio, so it is chosen as the next splitting node for branch 0 of the root node.

Subject M1 has also 2 values of 1 and 0, so two Branches can be split for it. We know from table 5 that among the fail students of M2 and M1there are 23 students whose mark of M3 is 0 (fail), and 12 is 1 (Pass), so in branch 0 of M1 the fail assessment accuracy of M3 is 23/35=65.71% In addition, table 5 also shows that among the fail students of M2 and the pass students of M1 there are 62 students whose mark of M3 is 1 (pass), and 18 is 0 (fail), so in branch 1 of M1 the pass assessment accuracy of M3 is 62/80=77.50% The pre-set criterion is fulfilled, branch 1 can also be stopped splitting.

GENERATED DECISION TREE

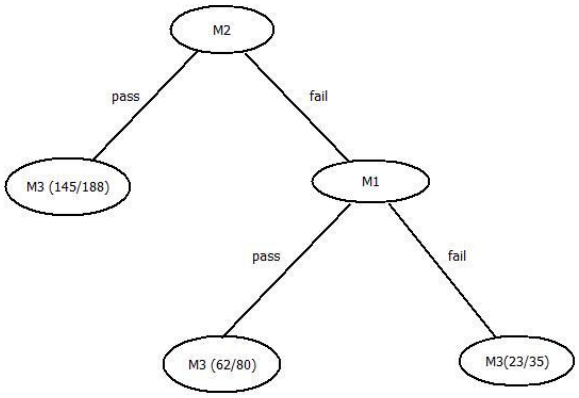

Fig. 2. prediction system

CLASSIFICATION RULE

Here only the classification rule of attribute M3 is extracted, describe as follows:

- If the mark of M2 is pass, then the mark of M3 is also pass usually, the accuracy is 77.12%, and the covering rate of the number of students is 188/270=69.62%.
- If the mark of M2 is fail, and the mark of M1 is also fail, then the mark of M1 is also fail usually, the accuracy is 53.65%, and the covering rate of the number of students is 35/270=12.96%.
- If the mark of M2 is fail, but the mark of M1 is pass, then the mark of M3 can also be pass , the accuracy is 81.25%, and the covering rate of the number of students is 80/270=29.62%.

CONCLUSION:

The goal of system is successfully achieved by generating result analysis report and predicting difficulty level of subject for student. Required manpower and time consuming problems are solved by system. The system helps students to achieve success in educational system. the teaching effect of the teacher can be checked, the teaching management can also be assisted. It will enable to identify the students in advance who are likely to fail and allow the teacher to provide appropriate inputs.

This project can be easily used by college for generating result analysis report. This system is user friendly and generates reports very fast.

ACKNOWLEDGEMENT

We thank Prof. Chatse R.V. Without his guidance, this paper could never have been accomplished.

REFERENCES:

1. Shubhangi Shankar Shinde1, Dr. Bhatambrekar S.S.2 Dipali Meher'"Result Analysis of Choice Base Credit System"2016.
2. Kamaljit Kaur* and KuljitKaur," Analyzing the Effect of Difficulty Level of a Course on Students Performance Prediction using Data Mining."2015.
3. ZhiwuLiu,Xiuzhi Zhang," Prediction and Analysis for Students Marks Based on Decision Tree Algorithm."2010
4. S.Taruna, Mrinal Pandey," An Empirical Analysis of Classification Techniques for Predicting Academic Performance."2014
5. Havan Agrawal, HarshilMavani," Student Performance Prediction using Machine Learning."2015
6. SnehalKekane, DipikaKhairnar, Rohini Patil, Prof. S.R.Vispute,Prof. N. Gawande, " Automatic Student Performance Analysis and Monitoring. "2016
7. R. R.Kabra,R. S. Bichkar, " Performance Prediction of Engineering Students using Decision Trees"2011