INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 5)
Available online at: www.ijariit.com

An automated prediction system for academic performance

Shahina K. M.
shahinakm2@gmail.com
University of Calicut, Malappuram, Kerala

ABSTRACT

Data analytics has a key role in the educational system and it helps to predict the student's performance based on academic and non-academic details. Any institution spends huge amount of work to analyze student performance. Accuracy of manual prediction depends on different factors like relation between the student and teacher, the time spends for the analysis, etc. So prediction with high accuracy requires huge amount of time and mental work. This paper presents a system for student performance analysis and grade prediction. This system greatly helps both the academicians and students. Grade prediction system helps the students to improve their performance and reduce dropout. And academicians can draw a clear picture of each student in the institution by this prediction. This scenario uses a data set consist of 33 features of 382 students. Python Anaconda distribution is the tool used for analysis and prediction. Four classification methods, linear regression, random forest, gradient boosting and logistic regression were used. The result shows that random forest outperforms all other methods inaccuracy. The system also finds the high influential factors and calculates the correlation between the final mark with different features like study time, father occupation and time spent on internet access, etc.

Keywords—

1.INTRODUCTION

Educational institutions are the center of Educational Data Mining (EDM). Educational data mining is a key research area because each institution handles thousands of student-related information in every academic year. Handling this huge amount of information using databases is a tedious task. Especially data analysis and prediction of student's performance requires huge amount of time and money. Prediction of student performance based on academic and non-academic attributes using python anaconda platform greatly helps to improve student performance. In current situation student dropout and year back, rates are increasing day by day from each university in an academic year, especially in the case of professional courses. This may cause lots of time and money wastage. So here comes the importance of data analytics systems. This system predicts student performance and end semester grade based on student data. And by using this system students can understand their academic range and identify the area they have to give more importance. Also, the educational institutions will get a clear picture of their students range at any time without any expense. So, they can identify the students having low academic performance and can arrange the required coaching for them to avoid year back and dropout.

2. LITERATURE REVIEW

A literature survey is done to review similar existing systems used to perform student performance prediction. Four existing systems are chosen because these systems are similar to the proposed system.

Snehal Kekane, Dipika Khairnar, and Prof. N. Gawande applied classification algorithms to analyze and monitor students' academic performance automatically. Different types of association rules are created to evaluate the performance and generate a scorecard. Which in turn assist the teachers to identify the weak areas of their student [1].

Chew Li Sa, Dayang Hanani bt. Abang Ibrahim, Emmy Dahliana Hossain, Mohammad bin Hossin generated a Student Performance Analysis System (SPAS) to keep track of students' results in the Faculty of Computer Science and Information Technology (FCSIT) in University Malaysia Sarawak. The system uses different data mining classification algorithms such as J48, Simple CART, BFTree, Random Tree and J48graft to predict students end semester result and which assist the lectures of the department to identify students with bad performance. But the system shows a maximum of 61 percent accuracy [2].

Ms.Tismy Devasi, Ms.Vinushree T P and Mr.Vinayak Hegde conducted a study for investigating students' future performance over 700 students' with 19 attributes in Amrita Vishwa Vidyapeetham, Mysuru. They used the Naive Bayesian mining technique for comparison and prediction. The system uses students' admission details, course details, mark details, course details, etc, to predict the end semester performance [3].

Sadiq Hussain, Neama Abdulaziz Dahan, and Fadl Mutaher Ba-Alwi targeted on the techniques and strategies of Educational Data Mining and Analysis of Students' Academic Performance Using WEKA. They considered 24 factors of 300

© 2019, www.IJARIIT.com All Rights Reserved

students from 3 different colleges from Assam, India. Four classification algorithms including J48, PART, BayesNet, and Random Forest are applied on the dataset using the data mining tool WEKA [4].

3. PROPOSED SYSTEM

The proposed system builds on anaconda distribution using python language. The system uses the student data set that consists the details of 382 students from two Portugal schools. Each data have thirty-three attributes such as student age, mother occupation, travel time and internal marks, etc. Then the student data is analyzed using anaconda python distribution where sci-kit learn is used for prediction. The system predicts the end semester grade of each student by considering the academic and non-academic details. The prediction includes several stages. They are listed below

1. Collection of student data from different classes of educational institutions.
2. Pre-process the data using anaconda python distribution
3. Plotting correlation matrix for the given data to identify the factors that heavily determines the final grade.
4. Splitting the data set into two for training and testing
5. Applying regression algorithms to predict the end of the semester grade.

A. Data Preparation

Student related data were collected from two schools in Portugal. And the data set consists of 382 students with 33 attributes.

```
In [9]: import pandas as pd
        import numpy as np
        import seaborn as sns
        import matplotlib.pyplot as plt
        %matplotlib inline
        df_por = pd.read_csv("student-por.csv",sep=';')
        df_por.head()
```

Out[9]:

school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	4	0	11	11
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	2	9	11	11
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	6	12	13	12
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	0	14	14	14
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	0	11	13	13

5 rows × 33 columns

Fig 1: Student data set

B. Data Selection and Transformation

The correlation between different factors and end semester mark is printed. Based on the correlation matrix and box plot some attributes that contribute very less for predicting the end semester grade is removed from the feature set.

```
G3            1.000000
G2            0.918548
G1            0.826387
studytime     0.249789
Medu          0.240151
Fedu          0.211800
famrel        0.063361
goout        -0.087641
absences     -0.091379
health       -0.098851
age          -0.106505
freetime     -0.122705
traveltime   -0.127173
Walc         -0.176619
Dalc         -0.204719
failures     -0.393316
```

Fig 2: Correlation matrix

From the correlation matrix, it is clear that the internal mark and parent's education is directly proportional to the end semester grade and failure is inversely proportional. A corresponding box plot is also given below. Where y-axis represents the end semester grade and the x-axis represents the level of mother education.

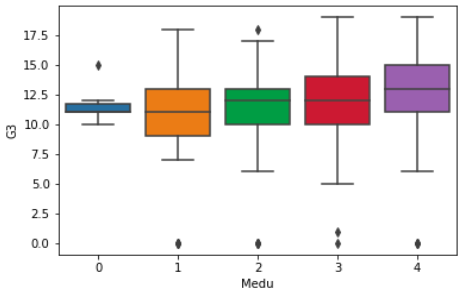

Fig 3: Box plot of mother education

C. Experimental Result

The end semester mark is classified into four grades and the random forest model predicted the result with 100 % accuracy. The corresponding scatters diagram is shown below.

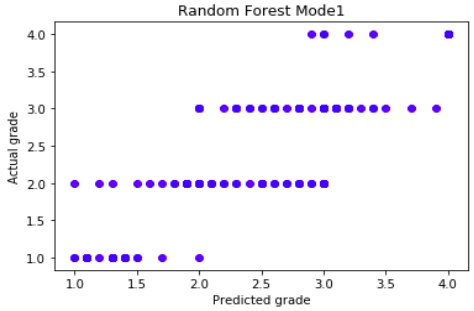

Fig 4: Scatter diagram

4. CONCLUSION

In this paper, 33 factors of 382 students from two schools are analyzed using anaconda python distribution. The correlation matrix shows that the major factors that constitute the output are internal marks, study time and mother education. Where 70% of the data is used for training and 30% is used for testing. The end semester mark is predicted in the form of grades and the result shows that random forest regression outperforms in accuracy and precision over the other three methods, Linear regression, Gradient boosting and logistic regression.

5. REFERENCES

[1] Snehal Kekane, Dipika Khairnar, and Prof. N. Gawande, "Automatic Student Performance Analysis and Monitoring", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 1, January 2016.

[2] Ms. Tismy Devasia, Ms. Vinushree T P and Mr. Vinayak Hegde, "Prediction of Students Performance using Educational Data Mining", International Journal of Innovative Research in Science, Engineering and Technology, vol 3,Special issue 3, March 2017...

[3] Chew Li Sa, Dayang Hanani BT. Abang Ibrahim, Emmy Dahliana Hossain, Mohammad bin Hossin, "Student Performance Analysis System (SPAS)".

[4] Sadiq Hussain, Neama Abdulaziz Dahan, and Fadl Mutaher Ba-Alwi, "Educational Data Mining and Analysis of Students' Academic Performance Using WEKA", Indonesian Journal of Electrical Engineering and Computer Science, Vol. 9, No. 2, February 2018, pp. 447~459.

© 2019, www.IJARIIT.com All Rights Reserved