

# Study of factors effecting to getting internship to Second/Third year students: A data mining Approach

Ashish Pant<sup>1</sup>, DevendraKushwaha<sup>2</sup>, JitendraMaurya<sup>3</sup>

Computer Science and information technology, I.E.T.M.J.P.R.U<sup>1,2,3</sup>

**Abstract:** The main concerns of any higher educational system is evaluating and enhancing the educational organization so as to improve the quality of their services and satisfy their customer's needs. This is an attempt to find suitable predication techniques using data-mining tool called WEKA. As the second and third year is very important for the engineering student cause after this a student go out for internship and get experience in his field. So we are trying to evaluate the factors for getting internship after 2nd year or 3<sup>rd</sup> year and if he is not been a better chance to get internship so he can work in the field in the third year. So the education institute can check his student performance and give student the notification that he has to do such work and have to work hard to increase his performance. The process of finding a suitable prediction algorithm was also described.

**Keywords:** Classification, Prediction, Student performance, data mining.

## I. INTRODUCTION

Data Mining is the vast field containing research in various amount of data. Educational data mining is one of its field. Educational Data Mining (EDM) describes a research field concerned with the application of data mining, machine learning algorithm and mining various statistics that are collected from various field. This field concern with improving the performance of students or evaluating factors that effect the students in order to discover new insights about how people learn in the context of such settings. Educational data mining is contributed to the theories of learning investigated by researcher in an field of education. The field is closely tied to that of learning analytics, and the two have been compared and contrasted. With the continuous development of database technology and the extensive applications of database management system, the data volume stored in database increases rapidly and in the large amounts of data much important information is hidden. So data mining tools are very important in this field. These tools are very important for evaluating the factor that are effecting the students' performance, so that one could improve the performance. So in this paper we are trying to evaluate the factors that are important for student for getting internship to 2<sup>nd</sup> and 3<sup>rd</sup> year students in getting internship.

## II. REVIEW LITERATURE

[1] This study measure the quantity of student behaviour, teacher attributes, and course characteristics on class attendance and performance. Several notable factors that influence attendance and grades are motivation, prior grade point average (GPA), self-financing by students, hours worked on jobs, quality of teaching, and nature of class lectures. It also provide the student positive effect on class attendance on performance of students. [2] Social work educators concur that writing and critical thinking are basic components of effective practice, yet students are often deficient in these skills. Although there is agreement

among educators about the need to enhance students' writing skills, there is little understanding of the nature of students' problems—a necessary step in the development of effective educational approaches. This article reports on a qualitative study that sought to understand which factors contributed to the weak writing skills of a group of social work undergraduates. Based on its findings and analysis of previous literature, the authors suggest a set of interrelated components that influence students' writing. The study's findings may contribute to the development of potential solutions to this persistent problem

[3] On the basis of educational persistence and motivational theory models, the PSFs (physiological and study skill factors) were categorized into 9 broad constructs: achievement motivation, academic goals, institutional commitment, perceived social support, social involvement, academic self-efficacy, general self-concept, academic-related skills, and contextual influences. Two college outcomes were targeted: performance (cumulative grade point average; GPA) and persistence (retention). Meta-analyses indicate moderate relationships between retention and academic goals, academic self-efficacy, and academic related skills. The best predictors for GPA were academic self-efficacy and achievement motivation. Supplementary regression analyses confirmed the incremental contributions of the PSF over and above those of socioeconomic status, standardized achievement, and high school GPA in predicting college outcomes.

[4] J.F. Superby conducted a study to investigate to determine the factors to be taken into account we will use a model adapted from that of Philippe Parmentier (1994). In other words the idea is to determine if it is possible to predict a decision variable using the explanatory variables which we retained in the model.

[5] Bray in his study has made conclusion that the number of students taking private tutoring in India is much higher

than in other countries. It also shows that the impact of private tutoring on the performance of the students and intensity of private tutoring depends on the collective factor namely socio-economic conditions.

### III. METHODOLOGY

Through extensive search of the literature, discussion and by experience of experts on student performance, a number of socio-economic, environmental, academic, and other related factors that are considered to have influence on the performance of university student were identified. These factors were carefully studied and harmonized into a manageable number suitable for computer coding within the context of the familiar algorithms. These influencing factors were categorized as input variables. The output on the other hand represents some possible levels of performance of a candidate in terms of the present college grading system.

### IV. DATA COLLECTION

For the evaluation of real world data's are collected from first year engineering students. For collection we choose M.J.P.R.U campus and then evaluate the performance. A sample of 360 students was taken from college 2<sup>nd</sup> and 3<sup>rd</sup> year students. Students were sit down on their hostels and we told them briefly about the survey so they are true in their data given. The primary data was collected using a questionnaire. Which include questions (i.e. with predefined options) related to several personal, socio-economic, psychological and school and college related variables that were expected to affect student performance. The questionnaire was reviewed by us and tested on a small set of 50 students in order to get a feedback i.e. how true these students are. The final version contained 23 questions in a single A4 sheet and it was answered by more than 300 students. Latter we selected a sample of 250 from the whole. All questionnaires were filled with the response rate of 100% out of which 100 were females and 150 were males.

The secondary data such as semester mark details, attendance percentage, and class test performance were collected by checking their marks and their attendance is asked by the students and confirmed by the class representative. All the predictor and response variables which were derived from the questionnaire.

	Variable Name	Description	Domain
1	Sex	Student sex	{male, female}
2.	Com	Student community	{general, obc, sc, st}
3	Living	Living area	{
4	Q1	Rank of department	{1,2,3,4,5,6}
5	Po	Parent occupation	daily wages, farmer, weaver, ex-serviceman, govt. ,

			Business, private)
6	Mother_0	Mother occupation	{ house wife, daily wages, farmer, weaver, private, government }
7	E_o	Occupation of elder sibling	{ daily wages, farmer, weaver, ex-serviceman, govt. , Business, private }
8	Elder	How many Elder sibling	{1,2,3}
9	Mos	Student's medium of study	{hindi, English }
10	H_s_p	High-school percentage	{ O - 90% - 100%, A - 80% - 89%, B - 70% - 79%, C - 60% - 69%, D - 50% - 59%, E - 40% - 49%, F - < 40% }
11	i_p	Intermediate percentage	{ O - 90% - 100%, A - 80% - 89%, B - 70% - 79%, C - 60% - 69%, D - 50% - 59%, E - 40% - 49%, F - < 40% }
12	Ad_method	Admission method	(management, quota)
13	Stay_place	Student staying place	{day scholar, hostel }
14	Grade first	Marks in 1 <sup>st</sup> year	{ O - 90% - 100%, A - 80% - 89%, B - 70% - 79%, C - 60% - 69%, D - 50% - 59%, E - 40% - 49%, F - < 40% }
15	Grade_second	Grade in 2 <sup>nd</sup> year	{ O - 90% -

			100%, A - 80% - 89%, B - 70% - 79%, C - 60% - 69%, D - 50% - 59%, E - 40% - 49%, F - < 40%
16	Project	Project u have made	{1,2,3}
17	Paper_present	Paper presented	{1,2,3}
18	Patent	Any patent	{Yes,no}
19	Technical	Member of any technical  cultural  sports community	{1,2,3,4}
20	Tech_attendance	Student class attendance	{<50, 50, 60, 70, 80, 90}
21	Area_of_interest	Any area of interest (mention number)	{1,2,3}
22	Pqual	Parent qualification	{ illiterate, schooling, degree, diploma}
23	Co_cir	Good in Co-circular activity	{yes,no}

Table 1.Factors for evaluating studentsperformace

### V.Algorithm And Implementation

Although many classification models exist, only some have been selected within the scope of this study. The selected algorithms are Naïve Bayesian algorithm, MLP, SMO, J48 and Decision table are used.

First, data cleaning was applied on the datasets. According to the missing data analysis, missing data have been removed from the datasets. Other than missing data analysis, datasets were also cleaned to remove noisy data. Unnecessary space characters or other spelling mistakes were also cleaned in the datasets. Another usual step in data pre-processing is data discretization. Although some algorithms are said to perform better when the numerical input variables are discretised, in this study numerical variables have not been put into binned intervals in order to maintain the same conditions for all algorithms.

WEKA is open source software issued under the GNU General Public License. WEKA has been utilized as the tool to run different classification algorithms. 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

### V1. Result And Discussion

This paper is an attempt to use classification techniques to analyse and evaluate student academic data and to enhance the quality of the higher educational system. Findings from the factors influencing academic performance were very significant. When pointing at the performance results of the classifier, its classification accuracy is actually measured. Accuracy is calculated by determining the percentage of instances correctly classified. Costs for wrong assignment can also be applied in classification problems; however, misclassification costs are not within the scope of this study.

	<b>Naïve Bayes</b>	<b>MLP</b>	<b>SMO</b>	<b>J48</b>	<b>Decision tree</b>
<b>Accur acy</b>	<b>10.8</b>	<b>69.5</b>	<b>58.8</b>	<b>64.8</b>	<b>51.7</b>

Table 1. Comparison of classification accuracy  
The accuracy values of the multiple dataset implementations according to each classifier can be seen in Table 4.1 (in percentage). The different classification algorithm predicts the grade more accurately. From this study Multi-Layer Perceptron predict the all the grades more accurately. The results of accuracy of different classifier are given above.

To verify relationship among the attributes, hypothesis is formed and tested.Result of these test are given below.

<b>Attribute</b>	<b>Hypothesis</b>
Student quota through they joined, Grade obtained at semester examination	Ho rejected
Student area they are belonging, Grade obtained at semester examination	Ho rejected
Student staying place, Grade obtained at semester examination	Ho rejected
Parent's qualification, Grade obtained at semester examination	Ho rejected
Student department, Grade obtained at semester examination	Ho rejected
Medium of schooling, Grade obtained at semester examination	Ho rejected
Student sex, Grade obtained at semester examination	Ho rejected

Table2. Testing Hypothesis

Rules that satisfy both a minimum support threshold and a minimum confidence threshold are called strong. In our case, we get several strong rules in our association since they satisfy these requirements. However, not all of them are useful for us. Because of that, we have to choose what the rules we need are and we should apply to. By selecting those useful rules, we can use them to do the prediction

like we did before. Of course, even some attributes and rules are not our interests now; we might still need them later. When we want to predict other different attributes, we just need to repeat the same processes. In this analysis following are the strong association rules.

1. Department=Mechanical → Sex=M
2. Area=urban Quota=management → Staying place=day scholar
3. 10th/12th medium=English Staying place=hostel → Quota=counselling
4. Department=Parent qualification=Schooling → Sex=M
5. Department=Mechanical 10th/12th medium=Hindi → Sex=M
6. 10th/12th medium=English 10th/12th percentage= 70+ Department → E.E.
7. 10th/12<sup>th</sup> percentage =65-75 medium=Hindi department=mechanical → 7+ gpa /ii year
8. Elder sibling occupation= Engineer → internship=approved
9. Elder Brother occupation=scientist → Paper presentation=yes
10. 10th/12th medium=Hindi 10th/12th percentage=60-65 Staying place=hostel → member of department cricket team
11. Patent=yes → internship=approved
12. Area of interest=coding → internship=approved

## V. CONCLUSION

In this paper we have analysed various factors influencing the academic performance of the students at engineering college level and predict the grade of the student if these factors are given as input. Performance of different classification algorithms are compared for classifying students using a Weka mining tool. We have shown that some algorithms improve their classification performance. We have also indicated that a good classifier model has to be both accurate and comprehensible for instructors. These included students' staying place, whether they stayed in hostel or day scholar in their first year of study and the area him/her belonging. This study will give a timely and an appropriate warning to students at risk. This work may improve student performance; reduce failing ratio by taking appropriate steps at right time to improve the quality of education. Therefore, it seems to us that data mining has a lot of potential for education.

## ACKNOWLEDGMENT

We would like to thanks to my parents which help me in making dreams come true. I am indebted to faculty members who help me in motivating to work in this field. I am also thankful to **Mr. Prashant, Miss. Urooj, Mr. Arun Yadav** who help us in reviewing the work and help me in interacting the students of other branches. I would like to thanks other persons also who help me in collecting data. Special thanks to the **Mr. Anil Yadav, Mr. Atiq Ahmad** for helping us while collecting data of questionnaire. Special thanks to Guest Lect. Miss Smriti whose extreme support and help made our work possible.

## REFERENCES

- [1] Evaluation of Factors Influencing Student Class Attendance and Performance by Stephen Deva doss and John Foltz.
- [2] Beyond "Writing to learn": Factors influencing students writing outcomes by Jay shree S. Jani & Marcela Sarmiento Mellinger.
- [3] Do Psychosocial and Study Skill Factors Predict College Outcomes? A Meta-Analysis. Robbins, Steven B.; Lauver, Kristy; Le, Huy; Davis, Daniel; Langley, Ronelle; Carlstrom, Aaron.
- [4] J.F. Superby J.-P. Vandamme & N. Meskens of Catholic University of Mons "Determination of factors influencing the achievement of the first-year university students using data mining methods".
- [5] Rosemary Win and Paul W. Miller: The Effects of Individual and School Factors on University Students' Academic Performance.

## BIOGRAPHIES



**Ashish Pantis** pursuing his final year of B.S. from I.E.T.M.J.P.R.U, Bareilly in Computer Science and Information Technology.



**Devendra Kushwaha** is in final year of graduate. Area of interest is management. Final year project is "Management system".



**Jitendra Murya** is in final year of graduation. His are of interest is A.I. He is being working on text to speech conversion in his final year project.