Students Academic Performance Prediction in Education Sector Using Classification

A. Priyadharsini[1], Mrs. B. Azhagusundari[2]
[1, 2] Department of Computer Science
[1, 2] N.G.M, College, Pollachi, India

Abstract- *Educational Data mining is based on collecting knowledge from educational databases or data warehouses and the information collected that had never been known before, it is valid and operational. In recent years, the biggest challenges that educational institutions are dealing with is the explosive growth of educational data and to use this data to improve the quality of managerial decisions. The challenge is to improve the quality of the educational processes so as to enhance student's performance in academic and placement. Thus, it is extremely important decision to set new strategies and plans for a better management of the current processes. Educational data mining model helps to declare student's future learning outcomes using data sets of ongoing students. Prediction of student's academic exam results in educational environments is important as well. In this paper, Naïve Baye's classification algorithm is used to predict the semester exam result for the first year students.*

Keywords- Educational Data Mining, Decision Tree, Classification, Student Academic Result analysis, Naïve Baye's

I. INTRODUCTION

Data mining is considered as the process of extracting important patterns from a given database and it always act as a valuable tool for converting data into usable information. Data mining has a wide range of applications in different areas that include marketing field, banking sector, educational research, surveillance, telecommunications fraud detection, and scientific discovery (Han & Kamber, 2008). More specifically, data mining can discover hidden information to support decision-making in various domains. The education data mining is one of these domains in which the primary concern is the evaluation and, in turn, enhancement of organizations based on education domain.

Data mining techniques are used to discover hidden information patterns and relationships of Educational data, which is helpful in decision making. Data mining can be applied to wide variety of applications in the educational sector for the purpose of improving the performance of students as well as the status of the educational institutions. Educational data mining is rapidly developing as a key technique in the analysis of data generated in the educational domain A single data contains valuable information. The type of information produced by the data and it decides the processing method of data. A collection of data that can produce valuable information, in education sector contains that information needed for mining, which helps the education sector to capture and compile low cost information. For this type of information and communication, technology is used. Now-a-days usage of educational database is increased rapidly because of the large amount of data that can be stored and analyzed.

Classification and prediction are two forms of data analysis. Categorical class labels are predicted using classification whereas prediction models are to predict continuous valued function. Classification process includes two steps, first step is building a classifier or model and second step is using classifier for classification. In predictive model, a model predictor is constructed to predict the continuous valued function or ordered value.

Educational Data Mining (EDM) is an emerging field exploring data in educational context by applying different Data Mining (DM) techniques/tools.

Purpose of Study: This study has aims to implement several prediction techniques in data mining to assist educational institutions with predicting their student's semester exam results. If students are predicted to have low academic performance or less chance to pass in semester exams, then extra efforts can be made to improve their academic performance activity.

II. LITERATURE SURVEY

A number of reviews pertaining to not only the diverse factors like personal, socio-economic, psychological and other environmental variables that influence the performance of students but also the models that have been used for the performance prediction are available in the literature and a few specific studies are listed below for reference.

Brijesh Kumar Baradwaj et al., describes the main objective of higher education institutions is to provide quality education to its students. One way to achieve highest level of quality in higher education system is by discovering knowledge for prediction regarding enrolment of students in a particular course, detection of abnormal values in the result sheets of the students, prediction about students' performance and so on, the classification task is used to evaluate student's performance and as there are many approaches that are used for data classification, the decision tree method is used here.

Mohammed M. Abu Tair and Alaa M. El-Halees (2012) applied the educational data mining concerns with developing methods for discovering knowledge from data that come from educational domain. Used educational data mining to improve graduate students' performance, and overcome the problem of low grades of graduate students and try to extract useful knowledge from graduate students data collected from the college of Science and Technology.

Abeer Badr El Din Ahmed, Ibrahim Sayed Elaraby (2014), currently the amount huge of data stored in educational database these database contain the useful information for predict of students performance. The most useful data mining techniques in educational database is classification. In this paper, the classification task is used to predict the final grade of students and as there are many approaches that are used for data, classification method is used here.

III. EDUCATIONAL DATA MINING USING CLASSIFICATION

Educational Data mining refers to extracting or "mining" knowledge from large amounts of educational data. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. Currently, the data are stored in educational database, these database contain the useful information to predict students performance. The most useful data mining techniques in educational database is classification. In this paper, the classification task is used to predict the final grades of students and as there are many approaches that are used for data classification.

We have Figure: 3. 1 that represents working methodology based on the framework. It is important to have a working methodology to govern our work before applying data mining techniques. The work methodology begins with problem definition, data collection and data preprocessing that includes data selection for testing and training. It precedes with data mining classification techniques with pruning which leads to discovering knowledge that is benefit to us.

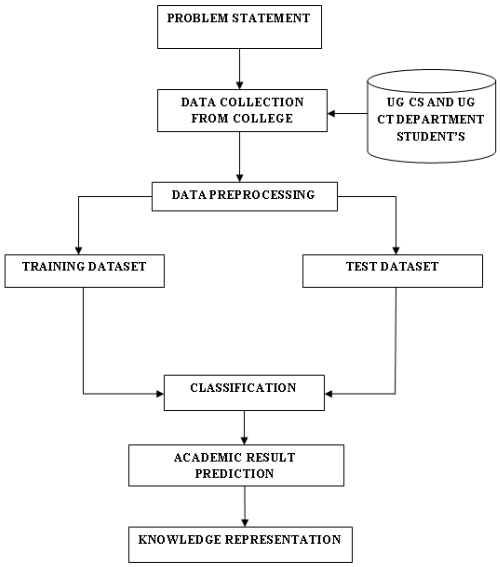

Figure: 3.1. Data mining work methodology

3.1. PROBLEM STATEMENT

The data set used in this study is obtained from NGM College of Arts and Science, Pollachi. The placement and academic marks of previous students of UG- CS (Aided) and UG-CT (self finance) is collected from the college database. The personal, academic details of the first, second and third year student's details directly collected from the students through questionnaire. All those details are stored in database and it is used to predict the performance analysis of students.

3.2. DATA COLLECTION

We have collected student's data of first year, pre final year and final year students of UG- CS (Aided) and UG-CT (self finance) departments through questionnaire that includes academic details.

3.3. PREPROCESS

In present day's educational system, a student's academic performance is determined by the Admission Type, marks scored in SSLC, HSC, Board and medium of study, activities done in the class like Assignments, Seminars, Paper Presentations, Attendance, etc., UG marks are based on the internal assessment and external exam mark. The internal assessment is carried out by the teacher based upon student's performance in educational activities such as class test, seminar, assignments, general proficiency, attendance and lab work. The end semester examination is one that is scored by

the student in semester examination. Each student has to get minimum marks to pass a semester in internal as well as end semester examination.

3.4. TRAINING DATA

In this step only those fields were selected which were required for data mining. A few derived variables were selected. While some of the information for the variables was extracted from the database. All the predictor and response variables which were derived from the database are given in Table 3.4.1 for reference.

Table 3.4.1

Variables	Description	Possible Values
Gender	Gender	{male,female }
Dept	Department	{CT,CS}
A Type	Admission Type	{Aided, Self Finance}
Board10	Board of Studies in 10^{th}	{CBSE, STATE}
Category	Communal Turn	{FC, BC, OBC, SC, ST, UNRESERVED}
SSLCMarks	Percentage of Marks in SSLC	{35% to 100%}
Board12	Board of Studies in 12^{th}	{CBSE, STATE}
HSC Marks	Percentage of Marks in HSC	{35% to 100%}
Assignment	Assignment Submitted	{Yes, No}
Seminar	Seminars Taken	{Yes, No}
PaperPresent	Papers Presented	{Yes, No}
Attend	Class Attendance	{Good, Average, Poor}
Medium	Medium of Study in School	{English, Regional}
Locality	Living Locality	{Village, Town, Taluk, District}
Bag log	Arrears in college	{Yes, No}
	academic	
UGPerc	Percentage of marks in UG	{35% to 100%}
HigherStudy	Interest in higher studies	{Yes, No}
Job	Job after UG	{Yes, No}

3.5. TEST DATA

Totally 101 instances with 20 attributes such as Name, Gender, Department, Admission Type, Board10, Category, SSLCMarks, Board12, HSCMarks, Assignment, Seminars taken, Paper Presented, Attendance, Medium of study, Locality, Baglog, UG Percentage, Higher Study after UG and Job after UG are passed to Naïve Baye's classification with training and test data sets. Attributes and instance of the first year student's data placed in the test set. Final year and second year students data are placed in the training set.

3.6. CLASSIFICATION

Classification is supervised learning method. It consists of two steps: 1. Model is built by analyzing the data tuples from training data. 2. Test data is used to check accuracy. There are various classification techniques such as Decision Tree algorithm that include ID3, C4.5, Bayesian Network, Neural Network and Genetic algorithm etc can be used.J48 is a java implementation of c4.5 in tool. These techniques can be used to build the classification model

3.6.1. NAÏVE BAYE'S ALGORITHM

The performance of the student is predicted using data mining technique called as classification rules. The Naïve Baye's classification algorithm is used by the administrator to predict the performance of the First and Second year students in the upcoming semester based on their SSLC and HSC Marks, Attendance, Seminars, Paper presentation, board of study, Admission type, bag log and previous semester result. This classification also used to predict the placement status for the Final year students based on the academic performance and behavior of the password students with UG Percentage, SSLC and HSC Marks.

3.6.1.1 ADVANTAGES OF NAÏVE BAYE'S ALGORITHM

- Naïve Baye's classifier requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the

variances of the variables for each class need to be determined and not the entire covariance matrix.

- It improves the classification performance by removing the irrelevant features.
- High Performance.
- It is short computational time

3.6.1.2 STEPS OF NAÏVE BAYE'S ALGORITHM

Step 1: Scan the dataset (storage servers)
Step 2: Calculate the probability of each attribute value. [n, n_c, m, p]
Step 3: Apply the formulae P(attribute value (ai)/subject value vj)=(n_c + mp)/(n+m) Where:

- n = the number of training examples for which v = vj
- n_c = number of examples for which v = vj and a = ai
- p = 1/number of subject values
- m = the equivalent sample size [number of attributes]

Step 4: Multiply the probabilities by p
Step 5: Compare the values and classify the attribute values to one of the predefined set of class.

IV. EXPERIMENTAL RESULT

4.1 WEKA TOOL

WEKA is a data mining tool, open source Java based tool issued under General Public License, for implementation of machine learning algorithm for data mining task. The WEKA tool provides a number of options associated with data mining algorithm suited for new machine learning schemes. In case of potential over fitting pruning can be used as a tool for précising. In other words algorithms can possibly used directly on the dataset. This algorithm it generates the rules from which particular identity of that data is generated. The objective is to apply classification algorithm Naïve Baye's for training set and obtain a classifier to classify test data until it gains equilibrium of flexibility and accuracy.

4.2 DATASET

Second and third year student details collected from students through questionnaire are used as training set to evaluate and create the classifier model for Nearly 20 scaling points are used while collecting data from the respondent. Here our respondents are students from first, second and final year under graduation. We have collected sample data from two departments as a real time data .We have considered only those fields needed for classifying is considered. Figure 4.2.1 specify the dataset set used.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Name	Gender	Age	Category	SSLC	SSLC Mark	HSC	HSC Mark	Branch	Admission T	Assignme	Seminar	Paper Pre	Attendan	Mediu
2	YOGESHWARI.A	Female	20	OBC	State	89	State	88	CS	Aided	Yes	Yes	No	Good	Regior
3	V.SANGEETHA	Female	20	OBC	State	87	State	89	CS	Aided	Yes	Yes	No	Good	Regior
4	V.MAHESHWARI	Female	20	OBC	State	90	State	90	CS	Aided	Yes	Yes	No	Good	Regior
5	UMAMAHESWARI.I	Female	19	OBC	State	82	State	79.35	CS	Aided	No	No	No	Good	Regior
6	T.KALPANA	Female	20	OBC	State	84	State	81	CS	Aided	No	Yes	No	Good	English
7	SOUNDARYA.C	Female	20	OBC	State	81	State	86	CS	Aided	Yes	Yes	No	Good	Regior
8	SHANMUGAPRIYA.I	Female	19	OBC	State	68	State	72.3	CS	Aided	No	No	No	Good	Regior
9	S.SAJITHA	Female	19	OBC	State	88	State	81	CS	Aided	No	Yes	No	Good	Regior
10	S.PAVITHRA	Female	20	OBC	State	92	State	91	CS	Aided	No	Yes	No	Good	English
11	S.KALAIVANI	Female	19	OBC	State	72	State	75	CS	Aided	No	No	No	Good	Regior
12	S.AISHWARYA	Female	19	OBC	State	87	State	89	CS	Aided	No	No	No	Good	Regior
13	REVATHI.S	Female	20	OBC	State	84	State	95	CS	Aided	No	Yes	No	Good	Regior
14	RAGAVI.A	Female	18	UNRESER\	State	86.4	State	86	CS	Aided	Yes	Yes	No	Good	Regior
15	R.VIGNESHKUMAR	Male	20	OBC	State	74	State	76	CS	Aided	No	No	No	Poor	Regior
16	PUSHPALATHA.M	Female	19	OBC	State	89	State	86	CS	Aided	Yes	Yes	No	Good	Regior
17	PRIYADHARSINI.K	Female	19	OBC	State	88.5	State	80	CS	Aided	No	Yes	No	Good	Regior
18	PAVITHRA.R	Female	19	OBC	State	75	State	91.2	CS	Aided	Yes	Yes	No	Good	English
19	PAVITHRA.D	Female	20	OBC	State	89	State	92.66	CS	Aided	No	Yes	No	Good	Regior
20	NIVETHITHA.S	Female	20	OBC	State	92	State	85	CS	Aided	Yes	Yes	No	Good	English
21	N.SARANYA	Female	20	OBC	State	89	State	91	CS	Aided	No	Yes	No	Good	English
22	N.KRISHNAVENI	Female	21	OBC	State	90.1	State	90	CS	Aided	Yes	Yes	No	Good	English
23	MANJOTHILAKSHN	Female	20	OBC	State	92.5	State	82	CS	Aided	No	Yes	No	Good	English
24	Karthikeyan	Male	20	OBC	State	73.6	State	76.6	CT	Self Finance	Yes	Yes	No	Good	Regior
25	KARTHIKADEVI.G	Female	20	OBC	State	93	State	93	CS	Aided	Yes	Yes	No	Good	English

Figure: 4.2.1 Second and Third Year Student's Data Set – Training File

Data collected from first year under graduate students through questionnaire are used for test data evaluation. Some of the fields used for scaling are based on the previous educational qualification of students. Figure 4.2.2 show the dataset used for testing.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	Name	Gender	Age	Category	SSLC	SSLC Mark	HSC	HSC Mark	Branch	Admissior	Assignme	Seminar	Paper Pre	Attendan	Medi
2	Gowthami	Female	18	OBC	State	90.6	State	92.33333	CS	Aided	Yes	Yes	Yes	Average	Englis
3	Harisham	Female	17	OBC	State	98	State	87	CS	Aided	No	No	No	Good	Englis
4	Suganya.M	Female	18	OBC	State	82	State	80	CS	Aided	Yes	Yes	Yes	Good	Englis
5	V.KOKILA	Female	18	OBC	State	81.5	State	85	CS	Aided	Yes	Yes	Yes	Good	Regio
6	M.RUBA	Female	17	SC	State	94.5	State	85.2	CS	Aided	Yes	Yes	Yes	Good	Regio
7	HARSHINI.D	Female	18	OBC	State	96	State	85.51	CS	Aided	Yes	Yes	No	Good	Englis
8	A.NANDHINI	Female	17	UNRESER\	State	96.2	State	85.5	CS	Aided	Yes	Yes	No	Good	Regio
9	S.HARITHA	Female	17	OBC	State	96	State	92	CS	Aided	Yes	No	No	Good	Englis
10	K.SUGANYA	Female	17	SC	State	71	State	73.5	CS	Aided	Yes	Yes	No	Good	Regio
11	MAHASAKTHI.T	Female	17	OBC	State	93.4	State	86.5	CS	Aided	Yes	Yes	Yes	Good	Regio
12	VIJAYARAGAVI.S	Female	17	OBC	State	91.5	State	87	CS	Aided	No	No	No	Good	Regio
13	BALADIVYA	Female	17	OBC	State	87	State	88	CS	Aided	Yes	Yes	No	Good	Regio
14	NATHIYA.V	Female	17	OBC	State	91.8	State	88.6	CS	Aided	Yes	Yes	No	Good	Regio
15	MALATHI	Female	18	FC	State	95	State	91	CS	Aided	Yes	Yes	Yes	Good	Regio
16	GURUSAMY.P	Male	17	FC	CBSE	95.6	CBSE	96.2	CS	Aided	Yes	Yes	No	Good	Regio
17	MADHANKUMAR.S	Male	18	OBC	State	94	State	72	CS	Aided	Yes	Yes	No	Good	Regio
18	HARISHANAND.D	Male	18	OBC	State		State	CS		Aided	Yes	Yes	No	Good	Englis
19	AKILAN.A	Male	17	UNRESER\	State	84	State	74	CS	Aided	Yes	Yes	No	Good	Englis
20	BALAGHAN.M	Male	17	OBC	State	80	State	73	CS	Aided	Yes	Yes	No	Good	Regio
21	PRABU.M	Male	18	ST	State	66	State	56.4	CS	Aided	Yes	Yes	No	Poor	Englis
22	KEERTHANA.S	Female	18	OBC	State	87	State	74	CS	Aided	Yes	Yes	No	Good	Regio
23	NASREENBANU.A	Female	17	OBC	State	93	State	84	CS	Aided	Yes	Yes	No	Good	Regio
24	JAYAPRIYA.K	Female	18	OBC	State	92.4	State	79.6	CS	Aided	Yes	Yes	No	Good	Regio
25	MEENA.M	Female	17	SC	State	91	State	80	CS	Aided	Yes	No	No	Good	Regio

Figure: 4.2.2 First years test file

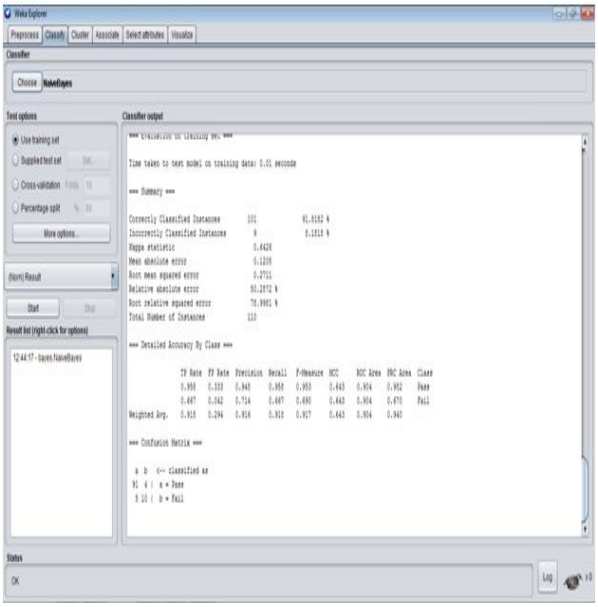

Figure: 4.2.3 Naïve Baye's Classification – training file

On evaluating the data set under Naïve Baye's classifier the result generated with the correctly classified instance by 91.89 %, F-Measure value 0.917 and Recall value 0.918.

	NAME	Gender	Branch	Admissior	SSLC Mark	HSC Mark	UG	'prediction margin'	'predicted Placement'
2	YOGESHWARI.A	Female	CS	Aided	89	88	68	1	No
3	V.SANGEETHA	Female	CS	Aided	87	89	67	0.935484	No
4	V.MAHESHWARI	Female	CS	Aided	90	90	68	1	No
5	UMAMAHESWARI.N	Female	CS	Aided	82	79.35	70	1	No
6	T.KALPANA	Female	CS	Aided	84	81	63	0.935484	No
7	SOUNDARYA.C	Female	CS	Aided	81	86	69	1	No
8	SHANMUGAPRIYA.M	Female	CS	Aided	68	72.3	55	0.935484	No
9	S.SAJITHA	Female	CS	Aided	88	81	61	0.935484	No
10	S.PAVITHRA	Female	CS	Aided	92	91	77	-0.294118	Yes
11	S.KALAIVANI	Female	CS	Aided	72	75	66	0.935484	No
12	S.AISHWARYA	Female	CS	Aided	87	89	68	1	No
13	REVATHI.S	Female	CS	Aided	84	95	68	1	No
14	R.VIGNESHKUMAR	Male	CS	Aided	74	76	61	0.935484	No
15	PUSHPALATHA.M	Female	CS	Aided	89	86	62	0.935484	No
16	PRIYADHARSINI.K	Female	CS	Aided	88.5	80	60	0.935484	No
17	PAVITHRA.R	Female	CS	Aided	75	91.2	65	0.935484	No
18	PAVITHRA.D	Female	CS	Aided	89	92.66	69	1	No
19	NIVETHITHA.S	Female	CS	Aided	92	85	68	-0.294118	Yes
20	N.SARANYA	Female	CS	Aided	89	91	70	1	No
21	N.KRISHNAVENI	Female	CS	Aided	90.1	90	61	-0.294118	Yes
22	MANJUTHILAKSHMI.K	Female	CS	Aided	92.5	82	80	-0.294118	Yes
23	KARTHIKADEVI.G	Female	CS	Aided	93	93	72	-0.294118	Yes
24	KALLESWARI.S	Female	CS	Aided	90	92	72	1	No

Figure: 4.2.4. Predicted academic result file

V. CONCLUSION

In this paper, the Baye's classification is used predicts the academic performance of the first year students based on the performance of the second year and final year students.

As there exist many approaches that are used for data classification, the Baye's Classification algorithm is used here. Information's like Board of studies in SSLC and HSC, Marks obtained in SSLC and HSC, Attendance, Seminars, Assignments, Paper presentation, bag logs, category, living locality, percentage were collected from the current students.

This study will help to the students and the professors to improve the academic performance of those who are at the risk of less chance to pass in semester exams. This study will also work to identify those students who needed special attention to placement interviews and taking appropriate action for the academic related activity.

REFERENCES

[1] Han, J. Kamber. M., "Data Mining: concepts and techniques. 2nd Edition, Morgan Kaufmann publishers (2008).

[2] Brijesh Kumar Baradwaj, Saurabh Pal, "Data mining: machine learning, statistics, and databases", 1996.

[3] Mohammed M. Abu Tair, Alaa M. El-Halees, "Mining Educational Data to Improve Students' Performance: A Case Study", 2012.

[4] Abeer Badr El Din Ahmed, Ibrahim Sayed Elaraby, "Data Mining: A prediction for Student's Performance Using Classification Method", World Journal of Computer Application and Technology 2(2): 43-47, 2014.

[5] Udeni Jayasinghe, Anuja Dharmaratne, Ajantha Atukorale, "Students' Performance Evaluation in Online Education System Vs Traditional Education System", IEEE 2015 12th International Conference on Remote Engineering and Virtual Instrumentation (REV).

[6] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques, 2nd edition", 2006.

[7] P. Ajith, M.S.S.Sai, B. Tejaswi, "Evaluation of Student Performance: An Outlier Detection Perspective", 2013.

[8] Varun Kumar, Anupama Chadha, "An Empirical Study of the Applications of Data Mining Techniques in Higher Education", 2011.

[9] Hongjie Sun, "Research on Student Learning Result System based on Data Mining", 2010.