CONTRIBUTION OF STUDENTS' ACADEMIC EFFORTS ON PREDICTION OF STUDENTS' PERFORMANCE USING WEKA

Saurabh Bhagvatula[1], K.J.Shreyas[2], Saurav Gupta[3], Mamta Singh[4]
Student, B.E.(7th Semester), Department of Computer Science & Engineering,
Bhilai Institute of Technology, Durg, Chhattisgarh, India[1,2,3],
Asst. Director, SaiMahavidyalaya,Sector – 6,Bhilai,Chattisgarh,India[4]
Email:Saurabhbh21@gmail.com[1], shreyaskj10@gmail.com[2], sourav77gupta@gmail.com[3],
mamtas_singh@yahoo.com[4]

Abstract— Educational Data Mining, a technique to explore the insight of academia with the machine learning provides various hidden trend and features which provide Educational Institutes a design policy of their educational techniques. Moving further in this field of research, our focus is on finding the attribute set which contributes more towards the students' performance. In this direction, we carried out comparative analysis of students' performance on three different attribute set having attributes which are highly coupled, medium coupled and loosely coupled with current state of the student over three different machine learning algorithms named as SVM, Naïve Bayes and J48. It was observed that the attributes which are highly coupled and more likely to change are contributing more towards student's performance. Prediction of result(Pass/Fail/Supplementary) before the end semester examination assists the students lagging behind in academics to discover the factor for the same and helps them work over it. Hence resulting in improved performance of students as well as help the research scholars to select the dominating attributes as per precedence order.

Index Terms— Attributes, Educational Data Mining, J48, Machine Learning, Naïve Bayes, Support Vector Machine.

I. INTRODUCTION

Educational data mining methods to develop predictive models that can help monitoring and anticipating student performance and take actions in issues related to student teaching and learning processes. Educational organizations foresee themselves from a business-perspective view, catering excellent quality of teaching-learning environments in order to satisfy their students.EDM should be applied to predict the learners' and educators' academic appraisals instead of constraining their objectives to admissions and enrollment procedures only.With the rapidly increasing levels of educational databases, the data miners are eager to explore the meaningful information at managerial levels. It was concluded that demographic characteristics are not significant predictors of student aggregate scores or success rates.

The aim of this study is to contribute to the prediction of students' academic performance in on-going courses. The prediction model is useful in identifying weak students who are likely to perform poor in their studies.

This paper proposes a framework for identify the most contributed attributes towards academia, for the performance of second year students of computer science and application course. An appropriate supervised machine learning model is applied upon our set of inherent attributes. The some features are non-changeable and so do not contribute in upgraded academic performances of the students. As they do not reveal any added academic effort. In this study, authors decided to work upon only external features of students by assigning weights that reflect their residual efforts put in for those features. Thus, the model is able to extract the fitness procedure sequences of external effort put up by each student who is predicted in 'at-risk' category in on-going course. This precedence relation can be used to identify and resolve the most unfit governing factor for enhancing students' academic performance.

II. EDM REVIEW IN WEKA

Educational Data Mining has emerged as one of the powerful technique which can be used in educational field to enhance our understanding of learning process and to focus on determining,extracting and evaluating attributes related to the learning process of students as described by Alaa el-Halees [5].

- Pandey and Pal studied the Investigation regarding the new comer students who applied to seek admission in PGDCA (one-year 'Post-Graduate Diploma Course in Computer Applications' course run nation-wide for graduate students) will come under the category of performers or non-performers, It is termed as dichotomous classification model [1].
- Pandey and Pal used the association rule mining methodology to correlate the class attendance of actual students' with the above the 'Students' interestingness using the Language medium of Class-room teaching attribute[2].
- Yadav and Pal carried out the experiment, along with classification modeling performed to predict End Semester result of students, the enrollment management system for granting admission in MCA course was also resolved using DT classifiers.The same experimental setup was used to explore similar kind of modeling based upon identifying, "Student's Retention for the pursued course" by using the decision-tree classifiers:ID#, C4.5 and ADT [3] [4].

To enrich the EDM modeling parameters, efforts were made to develop on-line educational assessment.

III. METH0DOLOGY

The broad steps of the procedure are shown in fig.1 and each of which are elaborated below:

A. *GENERATING EXPERIMENT SETS*

Three experiment sets were designed based on their close connection with the final effect on student's overall performance in a given academic session. Attributes are classified on the basis of the degree of coupling with the status of performance. They are classified as tightly coupled, moderately coupled and loosely coupled. First experiment set contains only tightly coupled attributes (4 attributes). Second Experiment set contains tightly coupled as well as moderate coupled attributes(6 attributes).Third Experiment set all the attributes relevant to this experiment(9 attributes).Sets are designed to study the degree of performance generated by each attribute type on overall results and their effect on each other's performance(synergistic or antagonistic).Final Stage.

The diagrammatic representation of this step is in fig.2.

B. *COLLECTION OF DATA*

At this step, the raw data was collected which contains the numerous information relevant to a student. The graphical representation of this step is in fig.3.

C. *PRE-PROCESSING STEP*

In this phase the data collected in the previous step which is available in the raw form are processed and converted into 4-attribute, 6-attribute and 9-attribute training and testing dataset. The fig.5 represents the processing of this step. Also, fig.6 and fig.7 represents the training and testing dataset for 4-attribute experiment.

D. *CLASSIFICATION USING MACHINE LEARNING ALGORITHMS IN WEKA*

In this step, the three experiment sets are performed i.e. trained and tested over three different machine learning classification algorithms (Support Vector Machine, Naïve Bayes and J48) and the accuracy was recorded. The reasons for selecting these algorithms are:

- To determine sets of attributes most responsible for output of student's performance, every algorithm will give diff. results.
- One algorithm would not be enough for arriving at right conclusion.
- Each experiment set is run through all mentioned algorithm to arrive at some general trend. This trend can be analysed to arrive at the right conclusion.

E. ACCURACY ANALYSIS

At this stage, after performing the three experiments over the three different machine learning algorithm, the accuracy was recorded are analyzed for the behavior each attribute show towards students' performance.

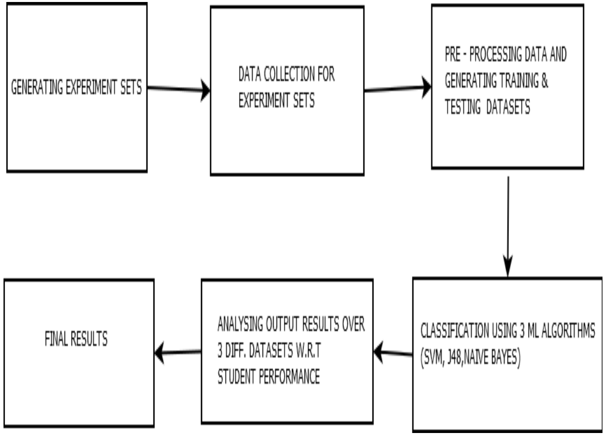

GENERATING EXPERIMENT SETS
DATA COLLECTION FOR EXPERIMENT SETS
PRE - PROCESSING DATA AND GENERATING TRAINING & TESTING DATASETS
CLASSIFICATION USING 3 ML ALGORITHMS (SVM, J48,NAIVE BAYES)
ANALYSING OUTPUT RESULTS OVER 3 DIFF. DATASETS W.R.T STUDENT PERFORMANCE
FINAL RESULTS

Fig.1: Broad View of processing steps

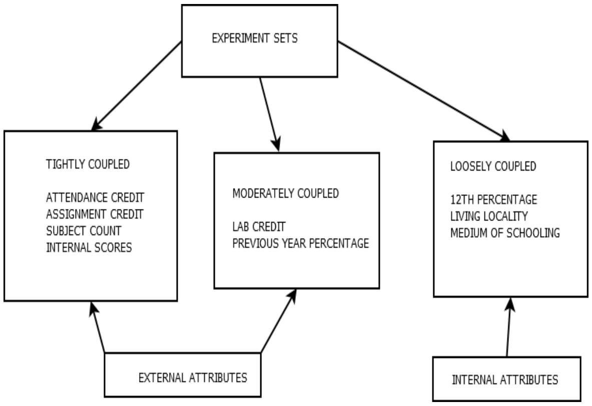


Fig.2: Deep View of Experiment set Generation step

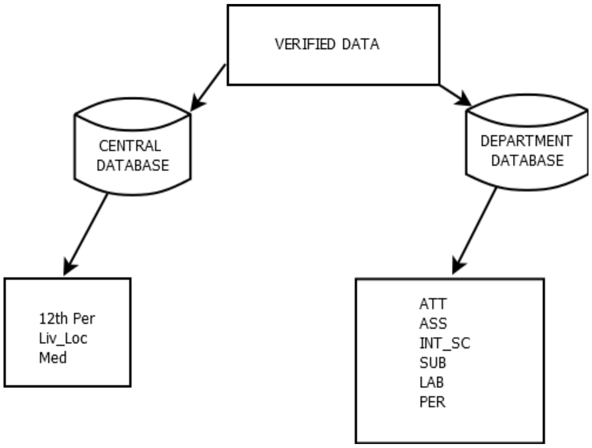
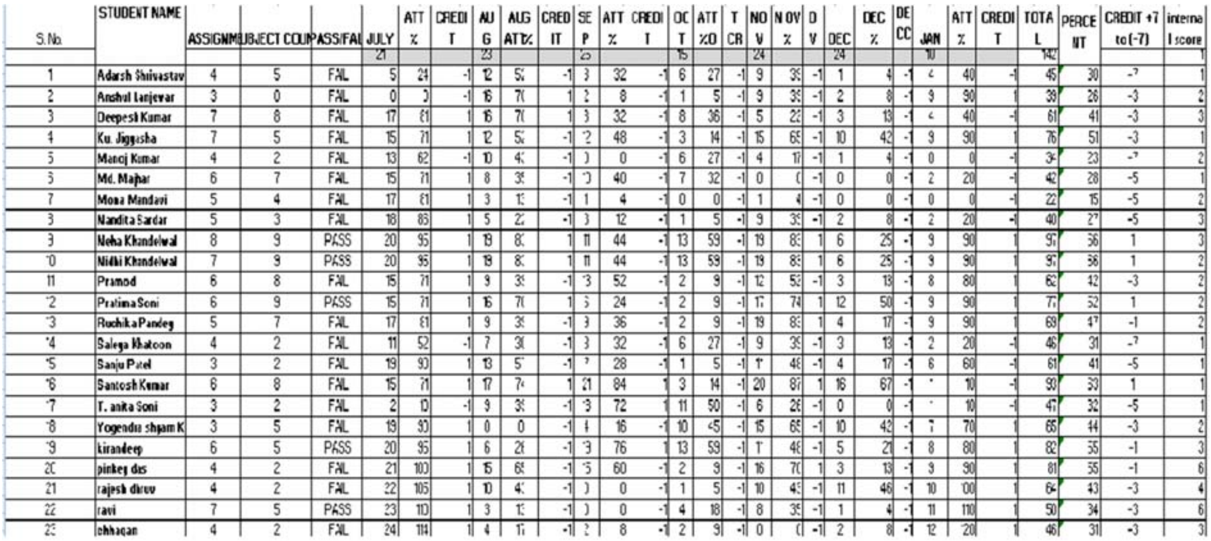


Fig.3: Deep view of data collection step



S.No.	STUDENT NAME	ASSIGNM	UBJECT COU	PASS/FAI	JULY	ATT %	CREDIT	AUG	AUG ATT%	CREDIT	SEP	ATT %	CREDIT	OCT	ATT %O	T CR	NOV	NOV %	DV	DEC	DEC %	DECC	JAN	ATT %	CREDIT	TOTAL	PERCENT	CREDIT +7 to(-7)	internal score
					21			23			25			15			24			24			10			142			1
1	Adarsh Shivastav	4	5	FAIL	5	24	-1	12	5	-1	8	32	-1	6	27	-1	9	3	-1	1	4	-1	4	40	-1	45	30	-7	1
2	Anshul Lanjewar	3	0	FAIL	0	0	-1	16	7	1	2	8	-1	1	5	-1	9	3	-1	2	8	-1	9	90	1	39	26	-3	2
3	Deepesh Kumar	7	8	FAIL	17	81	1	16	7	1	8	32	-1	8	36	-1	5	2	-1	3	13	-1	4	40	-1	61	41	-3	3
4	Ku. Jiggasha	7	5	FAIL	15	71	1	12	5	-1	12	48	-1	3	14	-1	15	6	-1	10	42	-1	9	90	1	76	51	-3	1
5	Manoj Kumar	4	2	FAIL	13	62	-1	10	4	-1	0	0	-1	6	27	-1	4	17	-1	1	4	-1	0	0	-1	34	23	-7	2
6	Md. Majhar	6	7	FAIL	15	71	1	8	3	-1	10	40	-1	7	32	-1	0	0	-1	0	0	-1	2	20	-1	42	28	-5	1
7	Mona Mandavi	5	4	FAIL	17	81	1	3	1	-1	1	4	-1	0	0	-1	1	4	-1	0	0	-1	0	0	-1	22	15	-5	2
8	Nandita Sardar	5	3	FAIL	18	85	1	5	2	-1	3	12	-1	1	5	-1	9	3	-1	2	8	-1	2	20	-1	40	27	-5	3
9	Neha Khandelwal	8	9	PASS	20	95	1	19	8	1	11	44	-1	13	59	-1	19	8	1	6	25	-1	9	90	1	9	36	1	3
10	Nidhi Khandelwal	7	9	PASS	20	95	1	19	8	1	11	44	-1	13	59	-1	19	8	1	6	25	-1	9	90	1	9	36	1	2
11	Pramod	6	8	FAIL	15	71	1	9	3	-1	13	52	-1	2	9	-1	12	5	-1	3	13	-1	8	80	1	62	42	-3	2
12	Pratima Soni	6	9	PASS	15	71	1	16	7	1	6	24	-1	2	9	-1	17	74	1	12	50	-1	9	90	1	7	52	1	2
13	Ruchika Pandey	5	7	FAIL	17	81	1	9	3	-1	9	36	-1	2	9	-1	19	8	1	4	17	-1	9	90	1	69	47	-1	2
14	Saleya Khatoon	4	2	FAIL	11	52	-1	7	3	-1	8	32	-1	6	27	-1	9	3	-1	3	13	-1	2	20	-1	46	31	-7	1
15	Sanju Patel	3	2	FAIL	19	90	1	13	5	-1	7	28	-1	1	5	-1	11	4	-1	4	17	-1	6	60	-1	61	41	-5	1
16	Santosh Kumar	6	8	FAIL	15	71	1	17	7	1	21	84	1	3	14	-1	20	8	1	16	67	-1	1	10	-1	93	33	1	1
17	T. anika Soni	3	2	FAIL	2	10	-1	9	3	-1	18	72	1	11	50	-1	6	2	-1	0	0	-1	1	10	-1	47	32	-5	1
18	Yogendu shyam K	3	5	FAIL	19	90	1	0	0	-1	4	16	-1	10	45	-1	15	6	-1	10	42	-1	7	70	1	65	44	-3	2
19	kirandeep	6	5	PASS	20	95	1	6	2	-1	19	76	1	13	59	-1	11	4	-1	5	21	-1	8	80	1	82	55	-1	3
20	pinkey das	4	2	FAIL	21	100	1	15	6	-1	15	60	-1	2	9	-1	16	7	1	3	13	-1	9	90	1	81	55	-1	6
21	rajesh dhruv	4	2	FAIL	22	105	1	10	4	-1	0	0	-1	1	5	-1	10	4	-1	11	46	-1	10	100	1	64	43	-3	4
22	ravi	7	5	PASS	23	110	1	3	1	-1	0	0	-1	4	18	-1	8	3	-1	1	4	-1	11	110	1	50	34	-3	6
23	chhagan	4	2	FAIL	24	114	1	4	1	-1	2	8	-1	2	9	-1	0	0	-1	2	8	-1	12	120	1	46	31	-3	3

Fig.4 : Collected data to be used for each experiment set

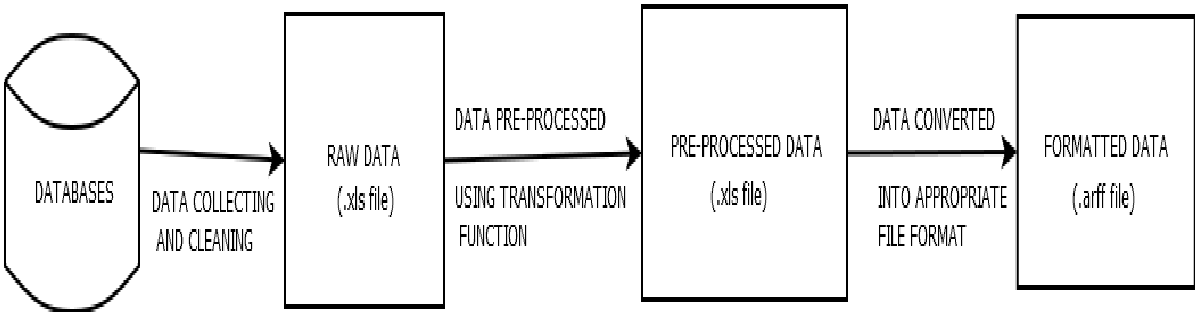
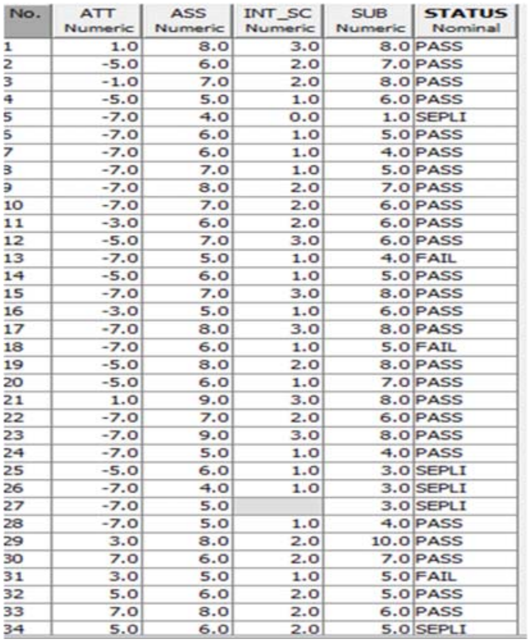
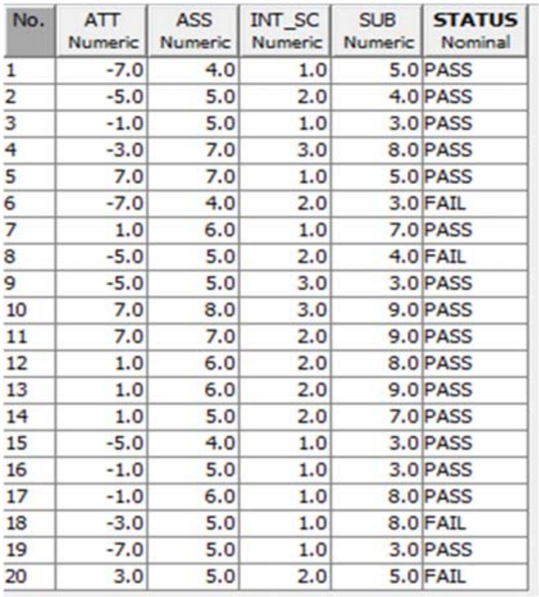

DATABASES
DATA COLLECTING AND CLEANING
RAW DATA (.xls file)
DATA PRE-PROCESSED USING TRANSFORMATION FUNCTION
PRE-PROCESSED DATA (.xls file)
DATA CONVERTED INTO APPROPRIATE FILE FORMAT
FORMATTED DATA (.arff file)

Fig.5: Deep view of data-pre processing step



No.	ATT Numeric	ASS Numeric	INT_SC Numeric	SUB Numeric	STATUS Nominal
1	1.0	8.0	3.0	8.0	PASS
2	-5.0	6.0	2.0	7.0	PASS
3	-1.0	7.0	2.0	8.0	PASS
4	-5.0	5.0	1.0	6.0	PASS
5	-7.0	4.0	0.0	1.0	SEPLI
6	-7.0	6.0	1.0	5.0	PASS
7	-7.0	6.0	1.0	4.0	PASS
8	-7.0	7.0	1.0	5.0	PASS
9	-7.0	8.0	2.0	7.0	PASS
10	-7.0	7.0	2.0	6.0	PASS
11	-3.0	6.0	2.0	6.0	PASS
12	-5.0	7.0	3.0	6.0	PASS
13	-7.0	5.0	1.0	4.0	FAIL
14	-5.0	6.0	1.0	5.0	PASS
15	-7.0	7.0	3.0	8.0	PASS
16	-3.0	5.0	1.0	6.0	PASS
17	-7.0	8.0	3.0	8.0	PASS
18	-7.0	6.0	1.0	5.0	FAIL
19	-5.0	8.0	2.0	8.0	PASS
20	-5.0	6.0	1.0	7.0	PASS
21	1.0	9.0	3.0	8.0	PASS
22	-7.0	7.0	2.0	6.0	PASS
23	-7.0	9.0	3.0	8.0	PASS
24	-7.0	5.0	1.0	4.0	PASS
25	-5.0	6.0	1.0	3.0	SEPLI
26	-7.0	4.0	1.0	3.0	SEPLI
27	-7.0	5.0		3.0	SEPLI
28	-7.0	5.0	1.0	4.0	PASS
29	3.0	8.0	2.0	10.0	PASS
30	7.0	6.0	2.0	7.0	PASS
31	3.0	5.0	1.0	5.0	FAIL
32	5.0	6.0	2.0	5.0	PASS
33	7.0	8.0	2.0	6.0	PASS
34	5.0	6.0	2.0	5.0	SEPLI

Fig.6: 4-attr. Training dataset



No.	ATT Numeric	ASS Numeric	INT_SC Numeric	SUB Numeric	STATUS Nominal
1	-7.0	4.0	1.0	5.0	PASS
2	-5.0	5.0	2.0	4.0	PASS
3	-1.0	5.0	1.0	3.0	PASS
4	-3.0	7.0	3.0	8.0	PASS
5	7.0	7.0	1.0	5.0	PASS
6	-7.0	4.0	2.0	3.0	FAIL
7	1.0	6.0	1.0	7.0	PASS
8	-5.0	5.0	2.0	4.0	FAIL
9	-5.0	5.0	3.0	3.0	PASS
10	7.0	8.0	3.0	9.0	PASS
11	7.0	7.0	2.0	9.0	PASS
12	1.0	6.0	2.0	8.0	PASS
13	1.0	6.0	2.0	9.0	PASS
14	1.0	5.0	2.0	7.0	PASS
15	-5.0	4.0	1.0	3.0	PASS
16	-1.0	5.0	1.0	3.0	PASS
17	-1.0	6.0	1.0	8.0	PASS
18	-3.0	5.0	1.0	8.0	FAIL
19	-7.0	5.0	1.0	3.0	PASS
20	3.0	5.0	2.0	5.0	FAIL

Fig.7: 4-attr Testing dataset

IV. DATA PROCESSING

During the initial stage Data is available in the form of excel sheets. Raw data containing various inconsistency require Data Cleaning, followed by the Transformation step to provide desired shape to the required attributes using variegated mathematical functions (in some of the cases). Attributes along with parameters are described below

Internal Attributes (remain constant) consist of given attributes : Living Locality, 12th score, medium.

- **12th Score** is scaled on the basis of student's performance in 12th class.
- **Schooling Medium** stores value on the basis of medium of study during school.
- **Living Locality** denotes the status of living.

External Attributes (vary as per performance) are listed below : Attendance Credit(ATT), Assignment Credit (ASS), Internal Score(INT SC), Subject Count (SC), Lab Credit (LAB) and Previous Year % (Per).

- **Assignment Credit** denotes the score given to each student on the basis of their performance related to assignments allotted to them.
- **Internal Score** denotes the score based on internal assessment.

- **Subject Count** denotes the student's performance on class tests.
- **Lab Credit** denotes the average of student's performance in laboratory.
- **Previous Year %** is scaled on the basis of student's previous year's overall performance.
- **Status** is class attribute (whose value will be predicted by ML algorithms).

V. EXPERIMENTAL SETUP

This project requires use of WEKA describes as follows :-

Weka (**Waikato Environment for Knowledge Analysis**) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License.

Three sets of experiments has been performed, which are as follows :-

- **4 – attribute Exp.** – Att. Credit , Ass. Credit, Int. Score, Sub. Count
- **6 – attribute Exp.** – All attribute from 4 attribute Exp. along with Lab credit and Previous Year %
- **9 – attribute Exp.** – All attribute from 6 attribute Exp. along with 10th score, 12th score and medium.

Each experiment involves two datasets :-

- **87** tuples dataset used for training purpose.
- **20** tuples dataset used for testing purpose.

All experiments are performed in WEKA software. Classification is performed using following three algorithms :-

1) **Naïve Bayes** :- works on the principle of Bayes' theorem for determining probability and establishing strong correlation between attributes. It works on probability model i.e. it predicts the value of class on the basis of probability given the value of attributes.
   The formula for calculating the probability for given class value $\mathbf{C_k}$ is:
   $\mathbf{p(C_k \mid x_1 , x_2 , \ldots.., x_n)}$ where $\mathbf{x_1 , x_2 , \ldots..,}$ $\mathbf{x_n}$ are values of dependent attributes.
2) **Support Vector Machine** :- It is a functional classifier i.e. it utilises sets of mathematical functions to plot the values into multi-dimensional space and predict the outcome on the basis of its position in space. SVM uses hyperplane to classify tuples. It first clusterise tuples on the basis of their proximity with other tuples. Then it draws hyperplane such that any point plotted on either side of plane will be automatically classified.
3) **J48** :- This classifier is decision tree based i.e. it creates decision tree model on the basis of training dataset and use it to classify test dataset. It uses two properties to calculate decision tree, namely Information Entropy which is amount of randomness present in the dataset.
   Formula to calculate Entropy of a given set S is as follows :-
   $$\mathbf{Entropy(S) = \Sigma\ \{ – (S_c /S) \log_2(S_c /S)\}}$$
   where S is total length of concerned set
   S_c is length of set belonging to particular class label.
   The above property is used to calculate Information Gain of each attribute which is as follows:-
   $$\mathbf{Gain(S,A) = Entropy(S) - \Sigma\ \{ (|S_v| /|S|) * Entropy(S_v)\}}$$
   $\mathbf{S_v}$ is subset of S for given attr. A having class label v
   $|\ \mathbf{S_v}\ |$ is length of subset $\mathbf{S_v}$.

   $|\ \mathbf{S_v}|$ is length of subset **S.**

VI. EVALUATION AND COMPARISON

Appropriate training set is selected and loaded in Weka for building a machine-learning model. A machine-learning model is basically a mathematical model generated by machine learning algorithms to be used for data prediction. Now we load the test data into the generated model, algorithms predict the result on the basis of training data and final result is provided on the basis of available values.

Results are obtained on the basis of parameters such as Accuracy, Precision, F – measure etc. These measures are very useful in comparison of machine learning algorithm's performance.

The parameters influencing the performance of different algorithms are explained below :-

- **Accuracy** :- This is primary unit of measurement for efficiency of algorithm on given dataset. It is calculated as ratio of correctly classified instances to the total instances.
- **Precision / +ve predictive value :-** It is the fraction of retrieved instances that are relevant,
- **Recall / Sensitivity:-** The fraction of relevant instances that are retrieved.
- **F-measure** :- IT is harmonic mean of precision and recall.
- **TP-rate :-** The ratio of successfully classified relevant instances to total no. of available relevant instances.
- **FP-rate :-** The ratio of instances mistakenly classified as relevant instances to that of total no. of available irrelevant instances.

Below are the parameters calculated for each individual value of class attribute :-

- **True Positives(TN)** :-Number of tuples that are predicted as given class and belong to it.
- **False Positives (FP)** :- Number of tuples that are predicted as given class but not belong to it.
- **False Negative (FN)** :- Number of tuples that are not predicted as given class but belong to it.
- **True Negatives (TN)** :- Number of tuples that are not as predicted as given class and not belong to it.

Table1: Performance Parameters

Parameter	**Formulae**
Accuracy	**Σ** TP / Total number of Instances
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)
F- measure	2 . TP / (2.TP + FP + FN)
TP – rate	TP / (TP + FN)
FP – rate	TP / (TP + FP)

Table2: Accuracy Comparison

Alg. / Exp.	**4-att.**	**6-att.**	**9-att.**
Naïve Bayes	50	55	45
J48	60	55	55
SVM	80	70	75

VII. CONCLUSION

In this paper, we have performed prediction for students' performance with the help of three machine algorithms on three different pairs of datasets and compared their results to evaluate the optimized algorithm for this purpose. From the given results, we can infer that Support Vector Machine resulted accuracy in the range of 70-80 %. Also it has been observed that optimum result negates with more attributes, which implies that highly coupled attributes provide maximum accuracy.

VIII. ACKNOWLEDGEMENT

The authors sincerely thank the Management, Sai College, Sector-6,Bhilai for supporting the current on-going work in Educational Data Mining and Analytics and also to the Department of Computer Science for providing students' data of on-going batch of BCA second year, one of the approved courses at Sai College, affiliated to Pt. Ravi Shankar Shukla University, Chhattisgarh, India. Also, the authors are grateful towards Dr. A.Rawal, Professor and Dr. M.V.Padmavati, Head, Department of Computer Science & Engineering, Bhilai Institute of Technology, Durg, affiliated to Chhattisgarh Swami Vivekanand Technical University, Bhilai.

REFERENCES

[1] U. K. Pandey and S. Pal, "Data Mining: A prediction of performer or underperformer using classification", Int. J. Computer Science and Information Technologies, vol. 2, pp. 686-689, 2011.

[2] U. K. Pandey and S. Pal S, "Data Mining View on Classroom Teaching Language",Int. J. Computer Science Issues, vol. 8, No. 2, , pp. 277-282, 2011.

[3] Yadav S.K. and Pal S. (March 2012), "Data Mining Application in Enrollment Management: A Case Study", Int. J. Computer Applications, vol. 41, No. 5, pp. 1-6, March 2012.

[4] Alaa el-Halees, "Mining Students Data to Analyze e-Learning Behavior: A Case Study", 2009.

[5] Witten, I. H., Frank, E., Hall, M. A., ―Data Mining: Practical Machine Learning Tools and Techniques‖, 3rd Ed. Morgan Kaufmann, 2011.

[6] Kember, D., ―Open Learning Courses for Adults: A model of student progress‖. Englewood Cliffs, NJ.: Educational Technology Publications, 1995.

[7] Brijesh Kumar Bhardwaj, Saurabh Pal," Data Mining: A prediction for performance improvement using classification", *(IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4, April 2011*