Evaluation of Five-Class Student Model based on Hybrid Feature Subsets

Rahul Misra[1], Dr. Ramkrishan Sahay[2]

[1]PhD Scholar, Department of CSE, Mahamaya Technical University, Noida, India

[2]Professor, Department of CSE, Mahamaya Technical University, Noida, India

misra.rrahul@gmail.com

Abstract- **The academic achievement of higher secondary school education in India is a turning point in the life of any student, as it serves as a very important link between the higher and higher secondary education of students. But, there are determinants like demographic, academic and socio-economic factors of students that restrict the students' performance. In this paper present the evaluation of five-class student model based on hybrid feature subsets.**

Keywords- **Education, Student, Performance, Predication Models, Feature Selection.**

I. INTRODUCTION

Education is a process of imparting or acquiring knowledge and habits through instruction or study and this process results in desirable changes in the behavior of human beings. It provides the skills to individuals to become self-confident, self-reliant and self- sustained and inculcates buoyancy to face challenges in all walks of life. It enhances the ability of individuals to manage health problems, improve nutrition and childcare, and prepare for the future. It sustains the human values which contribute to individual and collective well-being. It is the key which allows people to move up in the world, seek better jobs, and ultimately succeed in their lives. It is essential for eradicating poverty and it allows people to be more productive playing greater roles in economic life and earn a better living. It is worth mentioning that education forms the basis for lifelong learning in the context of human development and it is one of the fundamental requirements of democracy. It makes the people to aware of opportunities and rights that in turn result in more responsible and informed citizens. These citizens can have a voice in politics and society, which is essential for sustaining democracy and so education, is the only tool which takes the country to greater heights.

As education provides multifaceted developments of human beings, it is imperative to conduct researches in education for its effective implementation for the benefits of end users. One of the major goals of educational research is to investigate behavioral patterns in pupils, students, teachers and other participants in schools and other educational institutions. In fact, educational researchers like other social science researchers use a variety of techniques which can be broadly summarized as well as categorized in to two forms of methods viz. qualitative and quantitative research methods. Both these research methods are used in the fields of natural science, social science and technology; though these methods are differ to each other in all aspects

II. FEATURE SELECTION PROCESS

Feature selection is a process commonly used in machine learning, wherein a subset of the features available from the data is selected for application of a learning algorithm. The best subset contains the least number of dimensions that mostly contribute to accuracy; we can discard the remaining, unimportant dimensions. This is an important stage of preprocessing and is one of two ways of avoiding the curse of dimensionality. Reducing the number of irrelevant/redundant features drastically reduces the running time of a learning algorithm and yields more general concept. This helps in getting better insight into the underlying concept of a real-world classification problem. Feature selection methods try to pick a subset of features that are relevant to the target concept.

According to Dash and Liu (2007) feature selection attempts to select the minimally sized subset of features with the following criteria.

- The classification accuracy does not significantly decrease, and
- The resulting class distribution, given only the values for the selected features, is as close to the original class distribution as possible, given all features

Fig 1 show that how an optimal feature subset can be generated from the original data set through the sequence of step by step feature subset selection procedure.

There are four basic steps (Dash and Liu, 2007) in a typical feature selection method and they have been mentioned in Fig 1:

- A generation procedure used to generate the next candidate subset,
- An evaluation function used to evaluate the subset under examination,
- A stopping criterion to decide when to stop, and
- A validation procedure to check whether the subset is valid

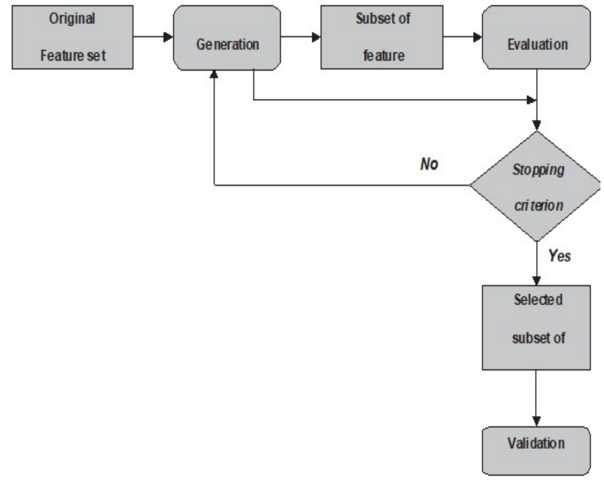

Fig. 1: Feature selection process

III. EVALUATION OF FIVE-CLASS STUDENT MODEL BASED ON HYBRID FEATURE SUBSETS

By repeating the evaluation of hybrid-based methods with top 20 features ranked according to their merits by CHI, CFS and ING methods, the predictive accuracy has been shown in Table 4.24. The results of the study exposed that the two classifiers BayesNet and NaiveBayes performed well against ranked based feature subsets and there was no effective improvement on the other three classifiers – J48, DT and MLP.

Table 1: Performance Evaluation Results of Hybrid-Based Classifiers for Five-Class Student data

Modles/FSS	FFS	F1M-CHI-13	F1M-CFS-19	F1M-ING-13	ROC-CHI-5	ROC-CFS-12	ROC-ING-5
J48	71.2806	65.0155	67.9841	65.0155	51.2254	59.2164	51.2254
DT	52.8133	51.5361	52.0366	51.5361	50.6904	50.932	50.6904
BayesNet	42.7511	47.4629	47.4629	47.4629	49.1025	47.4629	49.1025
NaiveBayes	39.5927	41.8019	42.613	41.8019	45.2192	44.3907	45.2192

The poor performances of the classifiers against these hybrid-based features were due to fact that smaller number of features were chosen based on F1-value and ROC-value (Fig 2). Another possibility for getting poor predictive performance was that the maximum number of instances was on particular class "***good***". In other words, ROC might not be an ideal measurefor multi-class problem.

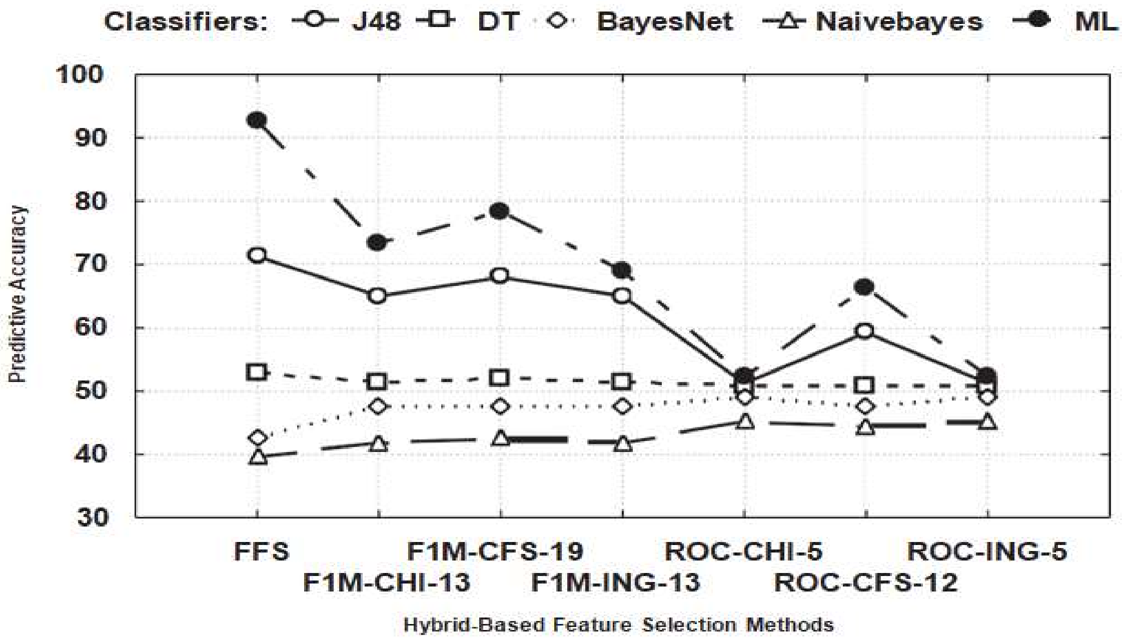

Fig. 2: 2D-line plot showing comparative performance of five classifiers-J48, DT, BayesNet, NaiveBayes and MLP with Hybrid-based feature selection methods for Five-Class student data set

Alternatively, the performance of these five classifiers was assessed through misclassification cost measure. The relative ranking for five-class problem was fixed as shown in Table 1 and its associated cost matrix for three-class has been given in Table 2. Heavy penalty was fixed for misclassification of “excellent “class into “fail” class.

Table 2: Relative Result Ranking for Five-Class

Results	**excellent** (90% and above)	**very-good** (75**%** and above)	**good** (60% and above)	**fair** (40% and above)	**fail** less than 40%of mark
Ranking	**0.0**	**0.1**	**0.2**	**0.3**	**0.9**

Table 3: Matrix representing Degree of Misclassification for Five-Class

			Predicted Results				
R			**excellent**	**very-good**	**good**	**fair**	**fail**
			0.0	0.1	0.2	0.3	0.9
e s	**excellent**	0.0	0	-0.1	-0.2	-0.3	-0.9
	very-good	0.1	0.1	0	-0.1	-0.2	-0.8
	good	0.2	0.2	0.1	0	-0.1	0
	fair	0.3	0.3	0.2	0.1	0	-0.6
	fail	0.4	0.9	0.8	0.7	0.6	0

The final cost matrix for five-class problem was obtained from the degree of misclassification using equation (4.3), with m = 0.9 and S = 100.and it has been shown in Table 3.

Table 4: Cost Matrix for Five-Class

		Predicted Results				
		O	A	B	C	F
True Results	O	0	2	4	6	18
	A	3.333333	0	2	4	16
	B	6.666667	3.333333	0	2	0
	C	10	6.666667	3.333333	0	12
	D	30	26.66667	23.33333	20	0

Table 5 shows the performance results of five classifiers against filtered subsets obtained by CFS, CSS, CHI, GAR and ING evaluation methods. The performance results of these classifiers showed that the rank value of both cost measure and predictive measures in filter-based approach were quit similar for MLP and J48 classifiers.

Table 5: Performance Evaluation Results of Filter-Based Five-Class Classifiers

Classifiers	Based on Misclassification cost Measure		Based on Accuracy Measure	
	Cost	Ranking	Accuracy	Ranking
Bayes-CFS	25.54592	18	49.1025	17
Bayes-CHI	27.0665	21	47.4629	19
Bayes-CSS	27.59583	22	49.0162	18
Bayes-FSS	24.51467	15	42.7511	21
Bayes-GAR	29.30358	24	47.4629	19
Bayes-ING	29.30358	24	47.4629	19
DT-CFS	27.87417	23	49.4477	16
DT-CHI	24.51467	15	51.6741	14
DT-CSS	25.60515	19	49.7929	15
DT-FSS	24.43254	13	52.8133	12
DT-GAR	24.05142	11	51.9676	13
DT-ING	24.51467	15	51.6741	14
J48-CFS	24.06144	12	54.591	11
J48-CHI	15.66173	9	68.4674	9
J48-CSS	15.43349	7	70.8146	6
J48-FSS	15.13625	5	71.2806	5
J48-GAR	15.33592	6	68.5537	7
J48-ING	15.65809	8	68.4846	8
Naive-CFS	26.83961	20	44.6151	20
Naive-CHI	24.69793	17	40.3003	24
Naive-CSS	25.23449	18	41.8882	22
Naive-FSS	24.55009	16	39.5927	25
Naive-GAR	24.49796	14	41.0079	23
Naive-ING	24.69793	17	40.3003	24
MLP-CFS	21.82812	10	59.7169	10
MLP-CHI	11.84857	4	81.6362	4
MLP-CSS	9.863847	2	85.951	2
MLP-FSS	4.338674	1	92.7166	1
MLP-GAR	10.03112	3	82.6717	3
MLP-ING	10.03112	3	82.6717	3

On considering the performance of Wrapper-based classifiers, MLP and J48 turned out as top ranked classifiers in terms of both cost measure and accuracy measure for Full Feature Set (FFS).

Table 6: Performance Evaluation Results of Wrapper-Based Five-Class Classifiers

Classifiers	Based on Misclassification cost Measure		ed on AccuracyMeasure	
	Cost	Ranking	Accuracy	Ranking
BayesNet-FFS	24.57467	6	42.7511	9
BayesNet-NB-BF	26.51579	9	48.2396	8
DT-FFS	24.43254	5	52.8133	5
DT-NB-BF	27.13642	10	49.4822	6
J48-FFS	15.13625	2	71.2806	3
J48-NB-BF	18.2123	4	62.9272	4
NaiveBayes-FFS	25.33629	7	39.5927	10
NB-NB-BF	25.86213	8	48.2741	7
MLP-FFS	4.338674	1	92.7166	1
MLP-NB-BF	18.17748	3	72.2817	2

As regards the performance of the Hybrid-based five-class classifiers (Table 6) are concerned, the classifier MLP had top ranked for FSS and F1M-CFS-19 feature subsets. The classifier J48 had also performed well for both FSS and F1M-CFS-19 feature subsets following the MLP classifier. The other feature subsets did not influence the predictive measure of the five classifiers.

Table 7: Performance Evaluation Results of Hybrid-Based Five-Class Classifiers

Classifiers	Based on Misclassification cost Measure		Based on AccuracyMeasure	
	Cost	Ranking	Accuracy	Ranking
BayesNet-F1M-CFS-19	29.30358	24	47.4629	18
BayesNet-F1M-CHI-13	29.56385	25	47.4629	18
BayesNet-F1M-ING-13	29.56385	25	47.4629	18
BayesNet-FFS	24.51467	13	42.7511	21
BayesNet-ROC-CFS-12	29.30358	24	47.4629	18
BayesNet-ROC-CHI-5	27.59583	21	49.1025	17
BayesNet-ROC-ING-5	27.59583	21	49.1025	17
DT-F1M-CFS-19	24.23805	10	52.0366	12
DT-F1M-CHI-13	25.75095	16	51.5361	13
DT-F1M-ING-13	25.75095	16	51.5361	13
DT-FFS	24.43254	12	52.8133	10
DT-ROC-CFS-12	26.80697	19	50.932	15
DT-ROC-CHI-5	28.9028	23	50.6904	16
DT-ROC-ING-5	28.9028	23	50.6904	16
J48-F1M-CFS-19	15.68065	4	67.9841	6
J48-F1M-CHI-13	17.05488	6	65.0155	8
J48-F1M-ING-13	17.05488	6	65.0155	8
J48-FFS	15.13625	3	71.2806	4

J48-ROC-CFS-12	20.88832	9	59.2164	9
J48-ROC-CHI-5	28.78956	22	51.2254	14
J48-ROC-ING-5	28.78956	22	51.2254	14
NB-ROC-ING-5	24.24043	11	42.613	22
NB-F1M-CFS-19	26.35597	17	41.8019	23
NB-F1M-CHI-13	26.35597	17	41.8019	23
NB-F1M-ING-13	24.55009	14	39.5927	24
NB-FFS	25.33629	15	44.3907	20
NB-ROC-CFS-12	27.2351	20	45.2192	19
NB-ROC-CHI-5	27.2351	20	45.2192	19
MLP-F1M-CFS-19	13.73761	2	78.426	2
MLP-F1M-CHI-13	15.81593	5	73.4553	3
MLP-F1M-ING-13	17.15441	7	69.0024	5
MLP-FFS	4.338674	1	92.7166	1
MLP-ROC-CFS-12	17.42424	8	66.3445	7
MLP-ROC-CHI-5	26.64032	18	52.261	11
MLP-ROC-ING-5	26.64032	18	52.261	11

IV. CONCLUSION

The academic achievement of higher secondary school education in India is a turning point in the life of any student, as it serves as a very important link between the higher and higher secondary education of students. But, there are determinants like demographic, academic and socio-economic factors of students that restrict the students' performance. This necessitates the need for some forecasting systems to predict the academic performance of students at plus two examinations. This is an attempt made first time in this aspect, which is mainly devoted to design and develop a prediction model by taking into account variables pertaining to the Indian society, for Indian educational system. Wide literature review on academic performance of students and its prediction by using performance models was carried out. But, it was noticed that limited research investigations have been executed not only on the factors that are influencing the academic performance of the students at high school/ higher secondary level but also on the prediction of the academic performance of the students using different classification algorithm in data mining. In this paper present and analysis of the evaluation of five-class student model based on hybrid feature subsets.

V. REFERENCES

[1] V.Ramesh, P.Parkavi and K.Ramar, "Predicting Student Performance: A Statistical and Data Mining Approach", International Journal of Computer Applications, Vol.-63, No.8, pp. 35-39, February 2013.

[2] Jagannath Mohanty, “Modern Trends in Indian Education , Second Revised & Enlarged Edition”, Deep &Deep Publication Pvt. Ltd., New Delhi, 2004.

[3] U. bin Mat, N. Buniyamin, P. M. Arsad and R. Kassim, "An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention," 2013 IEEE 5th Conference on Engineering Education (ICEED), pp. 126-130, 2013.

[4] Z. Ibrahim, D. Rusli, Predicting students academic performance: comparing artificial neural network, decision tree and linear regression, in: 21st Annual SAS Malaysia Forum, 5th September, 2007.

[5] M. Ramaswami and R.Bhaskaran, “A Child Based Performance Prediction Model in Educational Data Mining”, International Journal of Computer

Science Issues Vol. 7, Issue 1, No. 1, January 2010.

[6] Nguyen Thai-Nghe, Andre Busche, and Lars SchmidtThieme, “Improving Academic Performance Prediction by Dealing with Class Imbalance”, 2009 Ninth International Conference on Intelligent Systems Design and Applications.

[7] L.Arockiam, S.Charles, I.Carol, P.Bastin Thiyagaraj, S. Yosuva, V. Arulkumar, “Deriving Association between Urban and Rural Students Programming Skills”, International Journal on Computer Science and Engineering Vol. 02, No. 03, pp. 687-690, 2010.

[8] P. Cortez, and A. Silva, “Using Data Mining To Predict Secondary School Student Performance”, In EUROSIS, A. Brito and J. Teixeira (Eds.), pp.5-12, 2008.

[9] D. Kabakchieva, “Student performance prediction by using data mining classification algorithms”, International Journal of Computer Science and Management Research, vol.1, 2012.

[10] V. Ramesh, “Predicting student performance: A statistical and data mining approach”, International Journal of Computer Applications, vol. 63, no. 8, 2013.