IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 22, Issue 4, Ser. III (Jul. – Aug. 2020), PP 01-07
www.iosrjournals.org

A Predictive Analysis on Enrollment and Grading System of Government School in Jorhat District using Data Mining Algorithm

Rinkumani Kalita[1] , Joydip Kumar Sarmah[2], Dr. Siddhartha Baruah[3]
[1] *(Management Information System (MIS), SSA Jorhat Assam, India)*
[2]*(Department of Computer Applications, Jorhat Engineering College, Assam, India*
[3]*(Department of Computer Applications, Jorhat Engineering College, Assam, India*

Abstract:
The enrollment of the primary government school always possess less compare to the private inspite of the facility and availabilty of the school. The paper studies all the possible factor which affects the enrollment of the student in the government school. The paper put forward the maximum factor in the form of attributes to analysis the underlying condition of the primary school in jorhat district. The analysis carried the data mining tool and techniques used in WEKA for algorithmic accuracy and study. The classification techniques are accepted for analysis of the attributes and conditions. Cross fold condition and slip value condition is also checked for data consistency and accuracy.

Key Word*: Primary Government School; Enrollment; Data Mining Techniques; WEKA; Algorithms.*

Date of Submission: 11-08-2020 Date of Acceptance: 27-08-2020

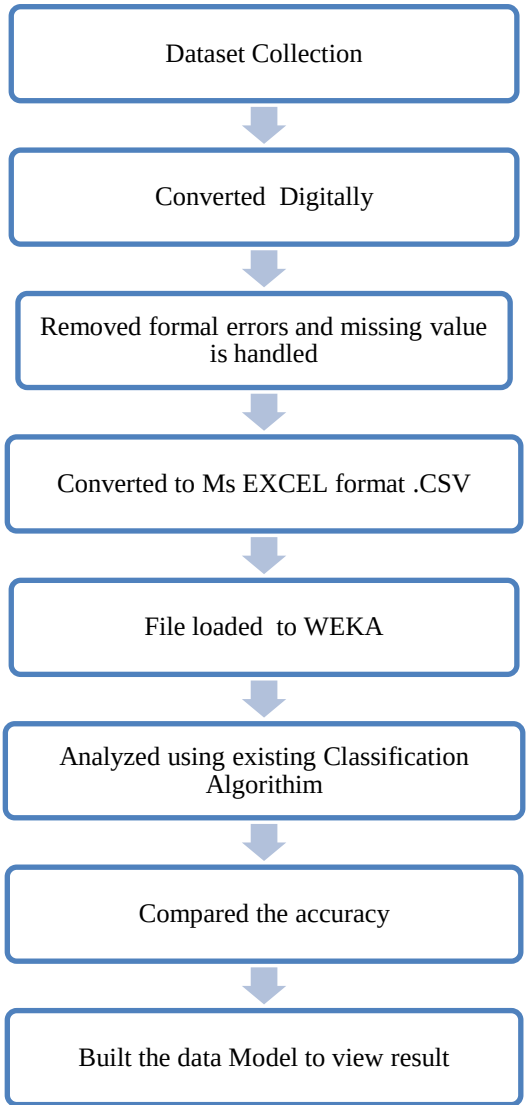
I. Introduction

The elementary education of India refers to from class 1 to class 8.the most important and crucial step for developing future generation. Primary education system is provided free and compulsory education from the government of India. The education system is provided in every government school in every district of India. In Assam school education is parted into primary, middle, high and higher secondary school. Free education system is accessible to up to 14 years of age in Assam. Theschools are either under state government to private or venture. The status of government school is little low compared to other state. There are many reasons but this paper actually studies the impact of school infrastructure part respective to grade of school which is constructed by government constructor. The infrastructure system is very poor in many districts of Assam. The study is conducted in the Jorhat district of Assam. This paper took a detail study of much school to understand the infrastructure system. It was observed that the infrastructure is very much dependent on the grade and enrolment of students. This paper reviewed all the government school and constructed attributed that will help to understand the present status of government school. The paper takes Data Mining techniques to predict the enrollment and grade of the school. Data mining is rich with its algorithm and with the help of algorithm the proper training data is judged and processed to generate accuracy. The tool Weka has been used to generate and test the data with the algorithm.

II. Data Pre-Processing Step

The data pre processing step is the overall layout of the methodology used in this paper for study. It contains several step and procedure which is shown by the diagram representation. The below steps are used to process and find the conclusion of the result. The steps are shown below. Pre Processing steps explain the work layout towards conclusion. All the steps that are conducted for research and study has been illustrated purely with explanation. The collection of data from different primary government schools has been initiated. The interaction with school bodies has done thoroughly for quality data collection process.

Table no 1: Shows preprocessing step of collecting information.



1. Dataset Collection: - The dataset is collected by interacting with various head and official of districts as well institute basically the lower, upper primary and higher secondary schools in Jorhat district. Initially one block of the Jorhat district is selected from the five blocks for data collection process. The attributes also include the internal and external types of the value used in government Gunostav assessment. The problem is being understood and took forward in paper layout for further study. Some attributes which affects the grade and enrollment of the school has been picked and data value has been collected.
2. Converted Digitally: -The paper study has been converted digitally with the help of simple database analysis tool MS Access. Initially the MS Access application has been selected to create a database model for storing and studying data for further analysis.
3. Removal of unwanted missing value: - Missing value with redundancy has been handled with the help of MS Access and various attribute relationship has been selected.
4. Converted to MS Excel format: - The file is again uploaded to MS Excel for format variation and converting to .CSV format.
5. File loaded to WEKA: - WEKA tool supports the .CSV format file which can only converted by WEKA in .arff format.
6. Analyzed using existing Classification Algorithms: - The data after uploading has been tested in WEKA using various constrain like cross fold analysis, slip analysis, testing and training slip analysis for understanding the effect nature of school in respect to grade and enrollment. Three best three algorithm has

been considered for the study – J48, Random Tree and Naïve Bayes algorithm. Among the three J48 gives average accuracy support with less fluctuate value. Random tree has shown the analysis on one condition the cent percent which effect the value if constrain is changed.

7. Compare the accuracy: - Accuracy of all three algorithms has been studied with best analysis.
8. Data Model Building: - J48 random tree shows the tree structure for analysis of class system and values which are affected the grading system and enrollment of the school.

III. Dataset Information

Various attribute has been considered for the analysis of the data which as been illustrated in below table

Table 2: List of attributed along with their value and descriptions

Sl. No.	Attributes Name	Description
1.	blkname	Block of the Jorhat district
2.	cluname	Cluster of division of the blocks
3.	schcd	UDISE code
4.	schname	School Name
5.	schmgt_desc	School Management
6.	new_schmgt_desc	Category of the school
7.	viltype	Area or address of the school
8.	tchposn	Total enrolled teacher
9.	tchMale	No. of male teacher
10.	tchFemale	No. of female teacher
11.	tchRegular	No. of regular teacher
12.	tchContract	No. of teacher in contact (TET)
13.	tchContract_SP	No. of State Pool Teacher (TET)
14.	tchOthers	Other teacher (if any)
15.	bldcond_desc	School building condition
16.	totclroom	Total classroom
17.	clgood_p	Total classroom in good condition
18.	cl_nd_rprng	Total classroom that require repairing
19.	toilet_b	Toilet facility for boys
20.	toilet_g	Toilet facility for girls
21.	toiletb_func	Boys Toilet functional condition
22.	toilet_func	Girls Toilet functional condition
23.	Toiletwater_b	Water Facility in boy's toilet
24.	Toiletwater_g	Water facility in girl's toilet
25.	Urinals_b	No of urinal for boys
26.	Urinals_g	No. of urinals for girls
27.	Uniwater_b	Water Facility in boy's urinals
28.	uniwater_g	Water Facility in girl's urinals
29.	handwash_desc	Handwash Facility
30.	water_desc	Water Facility
31.	electric_desc	Electric Facility
32.	bndrywall_desc	School Boundary wall Type
33.	pground_desc	Playground Facility
34.	library_desc	Library Facility
35.	bookinlib	No. of books in library
36.	bookbank_desc	Bookbank Facility

37.	**bookbank_books**	**No of books in Book Bank**
38.	**readcorner_desc**	**Reading Corner Facility**
39.	**readcorner_books**	**No. of books in reading corner**
40.	**newspaper_desc**	**Newspaper Facility**
41.	**calLAB_desc**	**Computer aided Lab Facility**
42.	**medchk_desc**	**Medical Check up**
43.	**total_enrolment**	**Total enrollment of student**
44.	**GRADE_2018**	**Grade provided by Gunostov2019**

The dataset has two class attributes i.e. Enrollment and Grade. Grade is being collected from the Gunostov available grade marking. Enrollment basically shows the no of student participation in government schools. The school infrastructure has been considered for the enrollment of the student and various factor which affects the enrollment has all being considered for the result. The total attributes value is 44 which has been specifically selected for analysis. This attribute value has been collected from various school available in different block of the Jorhat district.

IV. Data mining Algorithms

Data mining techniques and algorithms are implemented in various predictive analyses for decision support. There are many algorithms available in data mining tools for understanding the pattern of data. Similarly, in educational data mining task many algorithms are being used for decision parameters. In this paper following data mining algorithm is being used for its advantages to analysis educational data.

1. J48 algorithm
    a. Advantages: -
        i. Easy to implement the structure in tress specification
        ii. J48 is used to built analysis by forming the tree structure which is less.
        iii. Normalization of data is not required
        iv. Missing value data does not affect the analysis of data.
    b. Disadvantages: -
        i. Data complexity is less handled by J48 algorithm
        ii. Can not apply functionally of regression technique for implementation.

2. Random tree
    a. Advantages: -
        i. Ensemble technique is implemented in Random Tree algorithm.
        ii. More than one model can be preferred for data analysis.
        iii. Can handle large data set.
        iv. Complex data can be handles by this algorithm.
        v. Missing value is handled very efficiently by this algorithm.

    b. Disadvantages: -
        i. It cannot provide nature prediction.
        ii. Time consuming in data analysis process.
3. Naïve Bayes
    a. Advantages: -
        i. Required small amount of training data to estimate test data.
        ii. Easy to implement.
        iii. It holds independent prediction compare to another algorithm.
    b. Disadvantages: -
        i. It assumes that all attributes are mutually independent.
        ii. There may be of change in data accuracy dependent upon the data attributes.

V. WEKA Implementation

WEKA stands for Waikato Environment for Knowledge Analysis. It is data mining software develop by the University of Waikato in New Zealand. WEKA is a collection of Machine learning algorithms .The data file normally used by WEKA is in ARFF file for-mat, which consists of special tags to indicate different things

in the data file foremost: attribute names, attribute types, and attribute values and the data. The GUI allows us to try out different data preparation, transformation and modeling algorithms on data set. It allows running different algorithms in batch and compares the result. The comparative study using WEKA is very liable and understandable. The three algorithms J48, Random Tree and Naïve Bayes is tested and compared using various statistical measures. WEKA has pre defined structural form of algorithmic data for testing procedure. The data set is divided into test and training data. Training data is implemented and test data is predicted for grade system. It is seen that all test procure of the algorithm J48 gives average and good result compare to other ones. J48 consistently provide the average accuracy level and don't fluctuate like other two techniques.

VI. Data Analysis

The three algorithmic techniques are compared using certain statistical measure like

1. Kappa Statistics: -It helps to measure the inter-rater reliability for qualitative data variable.
2. Mean Absolute Error: - It is used to measure the difference between two continuous variables.
3. Root Mean Squared Error: -It shows the regression percentage of certain continuous variables.
4. Relative Absolute Error: -The relative absolute error is very similar to the relative squared error in the sense that it is also relative to a simple predictor, which is just the average of the actual values.
5. Root Relative Squared Root: -The root relative squared error is relative to what it would have been if a simple predictor had been used. More specifically, this simple predictor is just the average of the actual values.

Table no 2: Shows comparison depending upon various statistical data

Algorithm	Kappa statistics	MAE(Mean absolute error)	RMSE (Root mean squared error)	RAE(Relative absolute error)	RRSE(Root relative squared error)
Naïve Bayes	0.2511	0.2354	0.4377	85.3006 %	118.1256 %
Random tree	1	0	0	0	0
J48	0.7497	0.1016	0.2254	36.818 %	60.8354 %

Table no 3: Shows the accuracy measurement of three algorithms

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	No of Correctly Classified Instances	No of Incorrectly Classified Instances
Naïve Bayes	38.9262 %	61.0738 %	58	91
Random tree	100 %	0 %	149	0
J48	83.2215 %	16.7785 %	124	25

Table no 4: Shows the statistical analysis TEST MODE:10 fold CROSS VALIDATION

Algorithm	Kappa statistics	MAE(Mean absolute error)	RMSE (Root mean squared error)	RAE(Relative absolute error)	RRSE(Root relative squared error)
Naïve Bayes	0.0507	0.2985	0.5058	108.0034 %	136.397 %
Random tree	0.1618	0.2482	0.4272	89.8014 %	115.1897 %
J48	0.2628	0.2121	0.4015	76.7416 %	108.2745 %

Table no5: Shows Accuracy (Test mode: 10-fold cross-validation)

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	No of Correctly Classified Instances	No of Incorrectly Classified Instances
Naïve Bayes	24.8322 %	75.1678 %	37	112
Random tree	44.9664 %	55.0336 %	67	82

J48	51.0067 %	48.9933 %	76	73

Table no6: Shows Statistical Analysis (Test mode: split 66.0% train, remainder test) (Total Number of Instances: 51)

Algorithm	**Kappa statistics**	**MAE(Mean absolute error)**	**RMSE (Root mean squared error)**	**RAE(Relative absolute error)**	**RRSE(Root relative squared error)**
Naïve Bayes	0.0652	0.2738	0.4873	98.394 %	130.592 %
Random tree	-0.0268	0.2931	0.4885	105.3108 %	130.9261 %
J48	0.2216	0.2277	0.4226	81.8188 %	113.2684 %

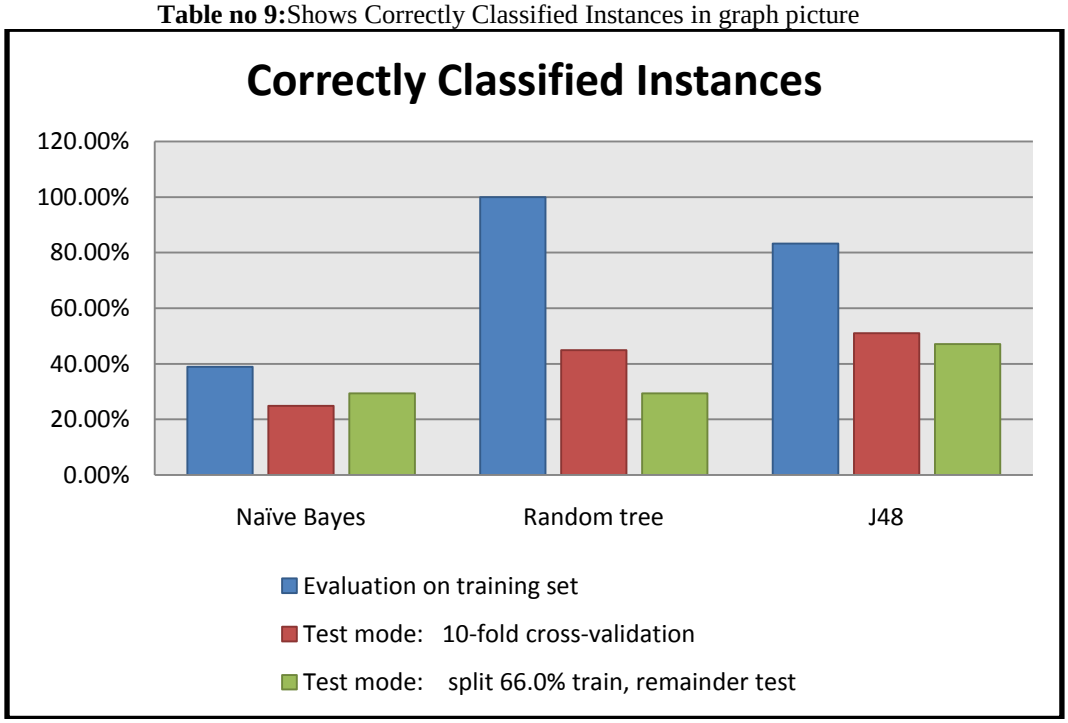
Table no 7: Accuracy (Test mode: split 66.0% train, remainder test)(Total Number of Instances :51)

Algorithm	**Correctly Classified Instances**	**Incorrectly Classified Instances**	**No of Correctly Classified Instances**	**No of Incorrectly Classified Instances**
Naïve Bayes	29.4118 %	70.5882 %	15	36
Random tree	29.4118 %	70.5882 %	15	36
J48	47.0588 %	52.9412 %	24	27

Table no 8: Shows Correctly Classified Instances

Algorithm	**Evaluation on training set**	**Test mode: 10-fold cross-validation**	**Test mode: split 66.0% train, remainder test**
Naïve Bayes	38.93%	24.83%	29.41%
Random tree	100%	44.97%	29.41%
J48	83.22%	51.01%	47.06%

Table no 9: Shows Correctly Classified Instances in graph picture



VII. Result

The following paper has observed some important aspects which are stated as below:-

1. Algorithm Random forest is best suitable for the observed data set.
2. The accuracy of the Random Forest algorithm is better compare to J48.
3. The Enrollment of the student is highly depend on the infrastructure and quality of teacher.
4. The role of teacher is influenced by the grade of the student.
5. The infrastructure of the school makes grade and enrollment effected, the good is the infrastructure the better is the performance.

VIII. Conclusion

The paper tried to analysis the government school infrastructure which is needed to be improved in many places in Jorhat for effective and influential grade and enrollment of the student. The proposed study points out some factor which will try to help the grading system of the student. Best Enrollment needs proper structure of teacher appointment. The future research will try to figure out the background structure of the school and their need to improve the infrastructure system. This paper also figures out the attributes which correlate with the enrollment and grading system.

References

[1]. B. K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance "*International Journal of Advanced Computer Science and Applications*, vol. 2, 2011.

[2]. A. Nandeshwar and S. Chaudhari. (2009). Enrollment Prediction Models Using Data Mining. [Online]. Available: http://nandeshwar.info/wp-content/uploads/2008/11/DMWVU_Project.pdf

[3]. D. Kabakchieva, "Predicting student performance by using datamining methods for classification," *Cybernetics and InformationTechnologies*, vol. 13, 2013.

[4]. Z. J. Kovačić, "Early prediction of student success: Mining students enrolment data," presented at the Informing Science & IT Education Conference, 2010.

[5]. J. Han and M. Kamber, "Data mining: Concepts and Techniques," (Morgan-Kaufman Series of Data Management Systems). San Diego: Academic Press, 2001.

[6]. A. Chaudhuri, K. De, and D. Chatterjee, "A Comparative Study of Kernels for the Multi-class Support Vector Machine," *icnc, FourthInternational Conference on Natural Computation*, vol. 2, pp. 3-7, 2008.

[7]. I. H. Witten and E. Frank, and D. Mining, "Practical Machine Learning Tools and Techniques with Java Implementations," Academic Press, 2000.

[8]. Harwatia, ArditaPermataAlfiania, FebrianaAyuWulandaria, "Mapping Student‟s Performance Based on Data Mining Approach (A Case Study)" ScienceDirect,Agriculture and Agricultural Science Procedia 3 (2015) 173 – 177.

[9]. C. Ricketts, S. J. Wilks, (2002),"Improving Student Performance Through Computer-based Assessment:insights from recent research "Assessment & Evaluation in Higher Education, Volume 27, Number 5, 1, pp. 475-479

[10]. "WEKA Data Mining Book" (n.d.)http://www.cs.waikato.ac.nz/~ml/weka/book.html.

Rinkumani Kalita, et. al. "A Predictive Analysis on Enrollment and Grading System of Government School in Jorhat District using Data Mining Algorithm." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 22(4), 2020, pp. 01-07.