SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876 Research Paper © 2012 IJFANS. All Rights Reserved, Journal Volume 11, Iss 11, 2022

A Pattern Mining Approach for Prediction of School Student Performance using Classification Algorithms

Khushbu Agrawal Research Scholar MATS School of Information Technology MATS University, Raipur (C.G.), India Email: khushi9r@gmail.com

Abstract—Evaluating student performance is very difficult task for academic area. The educational performance play a vital role in classifies the student in higher education. The student performance effect various factor like learning process, personal and social. This paper demonstrates the impact of student positive or negative performance of student success. Here the most commonly used prediction algorithms were LWL, random forest and bagging. After apply this three algorithm we present a novel model combination of this three algorithm name as ensemble. With the help of data mining algorithm we predict the student dropout rate, which is helpful for academic progress. For this we have to collect big and authentic data which can be done through uniform district information system for education DISE).

Keywords— student performance, prediction, classification algorithm, educational data mining (EDM), academic performance

I. INTRODUCTION

Educational data mining is a scientific research area, it use the multiple algorithm to improve academic result and procedure for further decision making. Predicting student performance in academic data is an important issue in elearning environments. Student performance is based on various factors such as personal, social, psychological and other issues. Data mining techniques is a promising tool to attain these objectives, data mining techniques are use to bring hidden information, patterns and relationship among the large dataset, which help us in categorization of data into knowledgeable facts. To identify the prediction of risk students with a large no. of student data set, it is very difficult and time consuming to using traditional data mining research methods such as questionnaires. Using traditional method in data mining has some limitations like it cannot properly handle the missing values, requires detailed information about the data, and cannot deal with uncertainty or vagueness in any information domain. Various tools and techniques required for achieving the best result from data ining like data cleansing, AI, association rule mining, clustering, regression, machine learning and classification. So the classification is one of the most useful predictive data mining techniques to solve this problem, and customized traditional method by applying various classification techniques. The prediction of student performance with high accuracy is beneficial for identifying the students with low

Dr. Bhavana Narain Professor MATS School of Information Technology MATS University, Raipur (C.G.), India Email: <u>narainbhawna@gmail.com</u>.

academic achievements. It is required that the identified students can be assisted more by the teacher so that their performance is improved in future.

I.I Contribution and organization of paper

Several empirical studies have been conducted the data mining in academic data I have study 12 research paper; most of the research work is done in small scale and structured dataset that are discussed below:

Romero et al., [1] have studied about data mining techniques and present early prediction of student performance. They search 133 research papers and 82 papers were used to solve their five research question.

Viswanathan and Vengatesh Kumar [2] have found that student performance prediction and define methodology for implementing student prediction. They have proposed a novel model of ensemble support vector machine (esvm) as a tool for data mining algorithm.

Namdeo et al., [3] presented a comparative study on the effectiveness from prediction of student's performance through educational data mining techniques. They describe a various research paper related to prediction of student performance and which use different classification technique.

Mangat & Singh Saini [4] has found that predict student performance in higher educational institutions by using machine learning algorithm. Here Naïve Bayesian algorithm provides more accurate result the study was done on seven hundred student's data set.

Zhang et al., [5] have studied about OLP (online teaching learner) learner scores for data mining technique. They have proposed course score analysis model and EM clustering was adopted for score features and salient features were obtained through PCA.

Ahmad et al., [6] have studied about EDM is the process of transforming raw data obtained from educational systems into useful data that can be used to make data-driven decisions.

Hossain et al. [7] have a proposed a new K-means Cluster Algorithm for data mining. Earlier method of Kmeans cluster techniques (original) have problem that if the number of clusters is to be chosen small then there is a higher probability of adding dissimilar items into the same group.



SCIENCES ISSN PRINT 2319 1775 Online 2320 7876

Research Paper © 2012 IJFANS. All Rights Reserved, Journal Volume 11, Iss 11, 2022

Ali et al. [8] investigate k-nearest neighbor classification (k-NN) performance on heterogeneous data sets which include the combination of both numerical and binary data. Earlier traditional k-NN works on numerical data only.

Sankari et al. [9] studied about the educational data mining model and learning analytics. The paper also investigate on how educational data mining (EDM) correlate between the student learning performance with other different learning models.

Fairos et al. [10] data mining algorithm was applied to predict student performance either excellent or non-excellent. The student performance was conducted on selected university in Malaysia.

Majeed and Naaz [11] has reviewed 39 research paper for understand the scope and the importance of academic data mining. The Expectation Maximization Algorithm, C4.5 and Page Ranker.

Qiao and Hong [12] have found that didactic of analyzing process data in log file by data mining methods. They proposed four supervised learning methods- Classification and Regression Tree (CART), gradient boosting, random forest, and SVM are explored to develop classifiers and unsupervised methods self-organizing (SO) and k-means methods are use by students to examine same and different score.

II. Material and methods

In the study the main problem is to organise the data set to get relevant information about the student performance in the educational industry. This problem is being sorting out with the help of data mining technique using uniform district information system for education (UDISE).



Fig.1 Prediction of student performance using UDISE

A. Data collection and Description

Table 1.Student database using UDISE

2019-20	P(I-V)Total	P-Dropout	UP(VI-VIII)Total	UP-Dropout	S(IX-X)Total	S-Dropout
Location						
BALOD	64851	0.01	39257	0.83	27857	15.68
BALODABAZAR	148282	0.95	85115	4.15	55936	19.73
BALRAMPUR	91517	2.98	44849	8.16	25516	18.44
BASTER	84203	4.9	42286	5.55	26443	19.45
BEMETARA	89485	0.41	53571	3.32	34811	24.77
BIJAPUR	33402	11.21	11661	12.27	6300	16.57
BILASPUR	226018	0	125801	4.46	83763	19.09
DANTEWADA	30943	9.29	13166	3.76	7624	19.96
DHAMTARI	69774	0.71	41245	1.47	30043	19.46
DURG	146365	0	86985	1.52	57084	14.57
GARIABAND	59143	1.9	33011	6.76	19457	17.22
JANJGIR - CHAMPA	173628	0.47	100845	1.91	67052	17.53

For our work we have presented sample datasets of year 2019-2020 from UDISE. Sample of this data set present here total number of student enrol_ in school has been taken. Students are divided in 3 categories they are primary, upper primary and secondary level. In this section total no. of enrol_ student and dropout percentage of students are considered. Here, location and total number of students are independent variables and dropout percentage of students is dependent variable. By using out location were dropout percentage of student is minimum.

B. Feature extraction and feature selection

Table 2 .Show year 2019-20 statistical representation

	P(1-	P-		
Statistic	V)Total	Dropout	Missing 1(3%)	
Minimum	17412	0		
			P(1-	P-
maximum	2651484	11.21	V)Total	Dropout
			Distinct	Distinct
Mean	189391.7	2.292	28	22
			Unique	Unique
stdDev	485407.1	3.429	28(97%)	20(69%)

	UP(VI-	UP-	UP-Total
Statistic	VIII))Total	Dropout	,dropout
			Missing
Minimum	7817	0.83	1(3%)
Maximum	1481381	12.27	Distinct 28
Mean	105812.9	4.592	Unique



ISSN PRINT 2319 1775 Online 2320 7876 Research Paper © 2012 IJFANS. All Rights Reserved, Journal Volume 11, Iss 11, 2022

			28(97%)
StdDev	271452.2	2.513	

	S(IX-	S-	S-total,
Statistic	X)Total	Dropout	dropout
			Missing
Minimum	4548	12.94	1(3%)
Maximum	943808	24.77	Distinct 28
			Unique
Mean	67414.86	18.058	28(97%)
StdDev	17303037	2.333	

Table 3 .Show year 2018-19 statistical representation(2018-19)

	P(1-	P-		
Statistic	V)Total	Dropout	Missing 1(3%)	
Minimum	17412	0		
			P(1-	P-
maximum	2651484	11.21	V)Total	Dropout
			Distinct	Distinct
Mean	189391.7	2.292	28	27
			Unique	Unique
StdDev	485407.1	3.429	28(97%)	2(90%)

	UP(VI-	UP-	
Statistic	VIII)Total	Dropout	
Minimum	7805	1.19	Missing 1(3%)
Maximum	1517322	19.22	Distinct 28
			Unique
Mean	108380.1	0.104	28(97%)
StdDev	278006	3.816	

	S(IX-	S-	
Statistic	X)Total	Dropout	
			Missing
Minimum	4289	14.72	1(3%)
Maximum	946427	31.73	Distinct 28
			Unique
Mean	67601.93	20.695	28(97%)
StdDev	173494.9	3.78	

Table 4 .Show year 2017-18 statistical representation(2017-2018)

	P(1-	P-	P(1-	P(1-
Statistic	V)Total	Dropout	V)Total	V)Dropout
Minimum	19543	0.005	Missi	ng 1(3%)
			Distinct	Distinct
maximum	2671372	4.28	26	28
			Unique	Unique
Mean	190812.3	1.831	24(83%)	28(97%)
stdDev	488913.6	1.116		

	UP(VI-		UP-total,
Statistic	VIII))Total	UP-Dropout	dropout
			missing1
Minimum	8267	0.09	(3%)
Maximum	1631532	8.35	Distinct 26
Mean	116538	4.819	Unique
StdDev	298862	2.037	24(83%)

Statistic	S(IX-X)Total	S-Dropout	S-total, dropout
Minimum	4317	13.44	Missing1(3%)
Maximum	947230	28.87	Distinct 28
Mean	67659.29	19.966	Unique
StdDev	173589	3.566	28(97%)

Table 5.Show year 2016-17 statistical representation

(2016-2017)

Statistic	P(1-V)Total	P-Dropout		
Minimum	18761	0	Missin	g 1(3%)
			P(1-	P-
maximum	2710696	12.88	V)Total	Dropout
			Distinct	Distinct
Mean	193621.143	1.855	28	18
			Unique	Unique
stdDev	496260.866	3.73	28(97%)	17(59%)



SCIENCES ISSN PRINT 2319 1775 Online 2320 7876 Research Paper © 2012 IJFANS. All Rights Reserved, Journal Volume 11, Iss 11, 2022

	S(IX-	S-	S-
Statistic	X)Total	Dropout	total,dropout
			Missing
Minimum	4078	14.78	1(3%)
Maximum	943703	30.24	Distinct 28
			Unique
Mean	67407.351	21.066	28(97%)
StdDev	172971.952	3.824	

Here table 2,3,4 and 5 show statistical representation of primary (I-V) total, dropout and upper primary (VI-VIII) total, dropout and secondary (IX-X) total, dropout values of 3 years student dataset, and show minimum, maximum, mean and stdDev values on each class. And also given the value of missing value distinct value or given unique value. This value is accessed by running data on weka plateform.

C. Proposed methodology

Research has its special significance in solving various operational problems. Research methodology is the way to systematically solve the research problem.

Hypothesis of Research Work

Hypothesis₁: Is there no relation between transition rate, promotion rate and dropout rate?

Hypothesis₂: Are data gaps deteriorating data quality?

Hypothesis₃: Did the model successfully classify the U-DISE data or not, in terms of location wise enrolment and learning performance?

D. Algorithms used for Classification

During the intense study of around few contributions, various architecture of machine learning model has been Studied. In most of the contribution authors have suggested different models of machine learning suitable for Student performance prediction. The major contributions are as follows.

• LWL

Local weight learning is approximation technique. It is find the underlying relationship between input and output. When we use dataset or Training data were each input is associated with one output and its use to create model that predicts values which come and close to the correct/true function.LWL use local functions and create a local model.

Random Forest

Random forest is a supervised machine learning algorithm. It can be use for both regression and classification problem solving schemes used in machine learning. It follows the



	UP(VI-	UP-		
Statistic	VIII)Total	Dropout	Missin	g 1(3%)
			UP(1-	
Minimum	8127	0	V)Total	Dropout
			Distinct	Distinct
maximum	1639555	17.52	28	26
			Unique	Unique
Mean	117111.071	5.793	28(97%)	25(86%)
stdDev	300346.181	3.8		

concept of ensemble learning algorithm which is the combination of multiple classifiers and solves the difficult problem with a great accuracy and also improves the model performance. It is use in Banking, Medicine, Land use or Marketing. Random forest contains a number of decision tress on various subsets of given dataset and predicting the majority of higher voting. It works with two phase first it create the random forest by combining N decision tree and second phase is to make predictions for each tree created in the phase.



Fig.2 The above diagram explains the working of Random forest.

• Bagging

Bagging is also known as bootstrap aggregation, is the ensemble learning technique, which is generally use to improve the stability and accuracy of machine learning algorithms and reduces variance within a noisy dataset. It is help to avoid over fitting and it can be use in different type of method like regression or classification specially decision tree method.

SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

Research Paper © 2012 IJFANS. All Rights Reserved, Journal Volume 11, Iss 11, 2022



Fig. 3 This diagram presents the Bagging process.

• Ensemble

Ensemble is a machine learning algorithm that combines more than two or more models and makes one optimal predictive model. There is Bagging, Staking and Boosting are three main classes of ensemble technique. It produces more accurate solution compare to a single model would.

E. Performance Evaluative method

In our work we have used Weka tool for evaluating the dataset.

(TP) and true negative (TN), respectively; the false positive (FP) and false-negative (FN) denotes the misclassification of normal and infected images, respectively; P = TP + FN and N = TN + FP.

(TP) and true negative (TN), respectively; the false positive (FP) and false-negative (FN) denotes the misclassification of normal and infected images, respectively; P = TP + FN and N = TN + FP.

Accuracy (ACC) = $TP + TN/P + N \times 100$

Specificity = $TN/N \times 100$

Precision = TP/ TP + FP \times 100

Recall = TP/ $P \times 100$

 $F1 - Measure = (2 \times precision \times recall / precision + recall) \times 100$

Area under Curve (AUC) = 1/2 (TP/ P + TN/ N)

Matthews Correlation Coefficient (MCC) = $TP \times TN - FP \times FN/\sqrt{(TP + FP) \times P \times N \times (TN + FN)}$

Finally, the obtained result is statistically validated using ztest and Friedman average ranking and Holm (Holm, 1979) and Shaffer (Shaffer, 1986) post-hoc multiple comparison methods.

Table The Prototype of the proposed automatic sequential Model ((Abbreviations: LWL, Random Forest, Bagging, Ensemble)

III. Experimental result and discussion



Algorithm		1	(2019-2)	0)P-Total			(2019-2	0jup-total		(2019-20)s-total				
		use tranning set	supplied test set	cross validation(10 fold)	persentage split(66%)	use tranning set	supplied test set	cross validation(10 fold)	persentage split(66%)	use tranning set	supplied test set	cross validation(10 fold)	persentage split(66%)	
	Time taken to test model	0.03 second	0 second	0 second	0 second	0.03	0.03	(0	0.03	6		
	Correlation coefficient	0.998	0.998	0.4021	0.6245	0.9978	0.9978	0.4094	0.6536	0.9978	0.9978	0.4174	0.69	
LWL	Mean absolute error	21911.0065	21911.0069	114298.5835	28360.3592	12269.3743	12269.3743	64026.5426	15612.4633	7565.054	7565.054	40264.6235	9397.875	
	Root mean squared error	30373.4036	30373.4036	469441.298	42647.546	1779.7592	1779.7592	261698.9759	25034.894	11229.8394	11229.8394	166443.2789	15459.461	
	Relative absolute error	12.09%	12.09%	61.29%	21.02%	12.08%	12.08%	61.29%	21.19%	12%	12%	60.42%	20.355	
	Root relative squared error	6.37%	6.37%	95.04%	29.34%	7%	7%	94.79%	31.01%	6.61%	6.61%	94.60%	30.495	
	Total Number of Instances	22	28	28	10	28%	28%	25	10	28	28	25	1	
	Ignored Class Unknown Instances	1	1	3	null	1%	1%	i 3	null	1		3	null	
	Time taken		0	0.03		0.03	0.01	0.05	0	0	0	0.06		
	Correlation coefficient	0.9995	0.9995	0.0165	0.6256	0.9997	0.9997	0.0243	0.271	0.9997	0.9997	0.0888	0.9	
	Mean absolute error	45109.5493	46109.5493	149243.5269	94743.6603	24344.5772	24344.5772	83388.6934	55020.7297	15952.015	15952.015	52005.1127	27742.889	
	Root mean squared error	139442.726	139442.7263	478473.021	104031.124	77663.35%	77663.35%	267531.723	64113.2619	49582.105	49582.105	169964.8179	30549.260	
tandom rores	Relative absolute error	25.44%	25.44%	80.02%	70.23%	23.97%	23.97%	79.83%	74.67%	24.66%	24.66%	78.04%	60.079	
	Root relative squared error	29.25%	29.25%	96.87%	71.57%	29.14%	29.14%	96.90%	79.40%	29.18%	29.18%	96.60%	60.245	
	Total Number of Instances	25	28	28	10	28	28	25	10	28	28	28	1	
	ignored Class Unknown Instances	1	1	1	null	1	1	1	null	1	1	1	null	
	Time taken		0		0			0 0	0	0	0		1	
	Correlation coefficient	0.9938	0.9938	-0.5194	0	0.9935	0.9935	-0.5164	0	0.9935	0.9935	-0.5134		
	Mean absolute error	73637.9376	73637.9376	197641.2858	163967.213	41122.5181	41122.5181	110803.5644	90091.9046	26140.5316	26140.5316	70537.5493	56551.161	
	Root mean squared error	243610.95	243610.9897	496268.179	172663,465	136127.245	136127.245	277409.4748	95949.3983	86736.6317	\$6736.6317	176754.8264	60307.356	
Bagging	Relative absolute error	40.63%	40.63%	105.97%	121.54%	40,49%	40.49%	106.07%	122.26%	40.34%	40.34%	105.85%	122.455	
	Root relative squared error	51.11%	51.11%	100.47%	118.79%	51.07%	51.07%	100.48%	118.83%	51.05%	51.05%	100.46%	118.925	
	Total Number of Instances	21	28	28	10.00%	28	28	28	10	28	28	28	1	
	Ignored Class Unknown Instances		1	1	null		1	1	cull.	1	1		Inull	

Table 6.Represented data in the year 2019-20 total and dropout value by running weka platform.

	- Southern			cross-19			and becaused as	Levis 191	aropool	lannet	-	10	- 10- 19		Lanner	
			use tranning	supplied test set	validation(persentage split(66%)	use tranning set	supplied test set	cross validation(persenta split/66%	ge use) tranni	ng test s	lied o	validation(1	split	tage
		Time taken to test model	0.01	0.00	au runaj	0		0.03	(restal	0	0 1	0.00	0.02	ronaj	-	0
		Correlation coefficient	0.9448	0.9286	0.9203	0.9041	0.9063	0.9053	0.37	1 0.2	129 0.5	458 0	8458	-0.020	1 0	7717
LATY	IWI	Mana shool da arror	1.0025	5/2220	3 3133	2 1032	1.7741	503341	0.57	10 2.41	50 1.4	0.05 1	1005	2.16		1023
LALI		Reat many second error	1.2002	5 0110	2.6526	2.1002	1 500.0	1 500.0	3.67	10 2.00	06 20	010 1	1004	4.54		65.00
		Relative abrolute error	41 02%	70 6250	45 £91/	£2,4534 53.40k	49.000	49.00%	103.95	19 3.3. K 103.9	50 2.0	070 2.	sel/	100.100	0 2/	2166
	-	Renarrie advorace error	12.052	67 6050	37.102	45 600	43.300	43.305	02.55	10 AND 7	CE/ E.4	100 64	110/	116 200	00	655/
		Root relative squared error	34.03/8	07/00/0	37.2276	40.0076	42.33/1	42.03/	25.20	10 2021	10	15/8	12.0	110.75	00	100/1
	- 1	Total number of Instances	28	10	26	10		1	-	1	10	20	280		8	10
		ignored class unknown instances	- 1		1				-	1	-	-1	- 1		-	_
	-	Time takes			0.06			0.01		-	0		-	0.0		0
		Correlation conflictent	0.057.0	0.002.0	0.00	0.03334	0.0070	0.03	0.00	10 O.C.	150 0	0		0.0		0
THEF		Mana absolute arrest	1,1393	1,1383	0.9538	0.7534	0.9976	0.3570	0.00	0.0	20 0.	983 0	422	3.461	2 U.	003
THEE	1.00 00 1	mean ausorate error	1.120/	1.120/	2.3704	2/4011	0.0421	0.0423	LiLP	0 2.1	01 0	322 4	1.922	2.001	1	.003
	- Random forest	Root mean squared error	1.5141	1.5141	4.1428	2.92/9	1.2135	1.2155	3.20	4 2.3	45 1.3	1 252	53559	3.735	2	5801
	-	Relative absolute error	24,34%	24,5479	60.79%	62,4875	31.729	31.72%	81.23	82.3	576 33.1	90% 33	1000	91.58	12	10976
	-	Root relative squared error	22.55%	22.55%	38.12%	54.5775	32.437	52.437	81.99	8 84.7	576 50.4	4/76: 52	A/76	90.197	84	.5/7t
	-	Total Number of Instances	28	28	28	10	25	28	-	8	10	28	28	2	8	10
		Ignored Class Unknown Instances	1	1	1		3	1 1		1	-	1	1	1	-	_
									-	-	-	-	-		-	-
		Time taken		0	0	0		0.01	0.	14	0	0	0	0.0	9	0
MEIA	1 3	Correlation coefficient	0.95555	0.95553	-0.5935	0	0.9948	0.9648	-0.55	1/	0 0.9	660 0.	9990	-0.571	1	0
	I	Mean absolute error	1.6284	1.6284	4.8397	3.5722	0.9935	0.9935	2.7	3 2.5	.06 1.0	065 1.	0565	2.925	1 2	4175
	Bagging	Root mean squared error	2.5321	2.5321	7.1055	5.2435	1.5098	1.5098	3.90	2.9	05 1.6	577 1.	6577	3.915	7 3.	0791
	_	Relative absolute error	35.12%	35.12%	99.69%	90.69%	37.439	37,439	99.04	\$ 95.7	5% 38.1	30% 38	.80%	100.999	100	.68%
	-	Root relative squared error	37.35%	37.35%	99.69%	98.44%	40.30%	40.30%	100	§ 97.7	0% 44.1	56% 44	.66%	100.82	E 100	.69%
		Total Number of Instances	28.00%	28.00%	28	10	26	26	4 3	18	10	28	28	2	8	10
		Ignored Class Unknown Instances	1	1	1		1	ĵ		1		1	1			
ssificatio	n Algorithm	Ignored Class Unknown Instances	1	{201	1 1-19)P-TOT/	4		(20:	18-19Jup-TO	1 TAL		1	1	(2018-19)5-	TOTAL	
ssificatio	n Algorithm	Ignored Class Unknown Instances	1	201 supplie	1 - 19]P-TOT/ d cross	L.	ge use trans	(20:	18-19jup-TO lied	TAL pe	rsentage	1	supt	(2018-19)s plied cros	TOTAL	persent
ssificatio	n (Algorithm	lignored Class Unknown Instances	use tranni	201 supplie test se	1 I-19)P-TOT/ d cross validati 10 fold)	d. persentaj sn[split[66%	ge use trane .) set	(20: ling supp test :	18-19]up-TO lied cross valid 0 fold	TAL pe sp ution[1 0	rsentage it(66%)	1 use tranning set	1 supp test	(2018-19)s plied cros set valid 0 fol	TOTAL lation(1 d)	persent split
ssificatio	n (Algorithm	Ignored Class Unknown Instances	use transi set	(201 supplie test se re 04 (0	1 - 19)P-TOT/ d cross validati 10 fold) .02	L persentaj on(split(66%	ge use trane) set	(20: ing supp test :	18-19]up-TO lied iet cross valid 0 foic 0.02	TAL pe sp tion(1 0	rsentage it(66%)	1 use tranning set	1 supp test	(2018-19)s- plied cros set valic 0 fol 0.03	total lation(1 d)	persent split
ssificatio	n Algorithm	Ignored Class Unknown Instances	use transi set 0	(201 supplie test se 04 (0 67 0.9	1 1 Sip-TOT/ d cross validati 20 fold) 102 107 0.5	L persentaj on(split(66% 0 223 0.3	ge use trane) set	(20) ling supp test : 0.01 .9964 0	18-19jup-TO lied iet cross valid 0 fold 0.02	1 TAL pe sp ition(1) 0.5344	it(66%) 0.7574	use tranning set 0.845	1 supp test	(2018-19)s- plied cros set valic 0 fol 0.03 0.8454	totAL lation(1 d) -0.0707	persent split
sificatio	n Algorithm	Ignored Class Unknown Instances	use tranni set 0 0.99 28332.91	(201 supple test se 04 (0 67 0.9 76 28832.	1 19)P-TOT/ d cross validati 10 fold) 02 167 0.5 18 467692	4. persentaj on(split(66% 0 223 0.3 189 35187	ge use trane) set 0 7232 0 1.046 17081	(20) ling supp test : 0.01 .9964 (.4308 1708	18-19]up-TO lied cross valid 0 fold 0.02 0.9964 11.431 703	1 TAL pe sp ition(1) 0.5344 71.6748 2:	rsentage it[6656] 0 0.7574 196.6809	1 use tranning set 0.845 1.488	1 supp test 0 8 0 5 1	(2018-19)s plied cros set valid 0 fol 0.03 0.8454 L4885	total lation(1 d) -0.0707 3.163	persent split 0.7 2.
sificatio	n Algorithm	Ignored Class Unknown Instances	1 use transi set 0.91 28832.91 39248.30	201 suppli- test se 04 (0 67 (0.9 76 (28832) 41 (39248)	1 1-19)P-TOT/ d cross validati 10 fold) 02 107 0.5 118 467692 104 65.5	4. persenta split(66% 213 0.1 189 35187 195 47185.5	ge use trane) set 0 7232 0 0.046 17081 5493 23475	(20: ing supp test : 0.01 9964 (0 .4308 1708 .3569 2343	18-19)up-TO lied iet cross valid 0 fold 0.02 1.1431 703 75.357 2686	TAL per sp tion(1 0 0.5344 71.6745 22 55.7103 20	o 0.7574 196.6809	1 use tranning set 0.845 1.488 2.009	1 supp test 0 5 1 6 2	(2018-19)s- plied cross valic 0 fol 0.03 0.8454 L4885 2.0096	totAL (ation(1 d) -0.0707 3.163 4.536	persent split 0.7 2.6
ssificatio LAZY	twi	Ignored Class Unknown Instances	1 use transi set 0 28832.91 39248.30 16.0	201 supplie test se 04 (0 67 (0.9 76 (28832, 41 (39248, 25) (16.0	1 1-19)P-TOT/ d cross validati 20 fold) .02 .02 .03 .02 .04 .05, .05 .05 .05 .05 .05 .05 .05 .05	4. persenta on(split(66% 223 0.3 189 35187 195 47185.1 195 47185.2	ge use trans) set 7232 0 0.046 17081 4195 16	(20) ing supp test 1 0.01 9964 (4308 1708 3569 2343 446% 1	18-19jup-TO lied et cross valid 0 fold 0.954 11.431 703 5.357 2686 6.46%	1 TAL pe sp tion(1 0 0.5344 71.6748 2: 55.7103 2: 65.935	0 0.7574 196.6809 27.95%	1 use tranning set 0.845 1.488 2.009 551	1 supp test 0 5 1 6 2 6	(2018-19)s- plied cros set valic 0 fol 0.03 0.0456 L4855 2.0096 5355	total ation(1 d) -0.0707 3.163 4.536 109.19%	persent split 0.7 2. 2.6 86.3
ssification LAZY	tWL	Ignored Class Unknown Instances	1 use transi set 0.99 28832.91 39248.30 16.0 8.2	2011 supplie test se 04 (0 07 0.9 76 28832 41 39248. 25 16.0 26 8.2	1 -19)P-TOT/ d cross validati 10 fold) .02 .03 .04 .05 .05 .05 .05 .05 .05 .05 .05	4. persentaj split(66% 0 213 0.7 189 35187 195 47185.1 195 28.6 206 32.	0 17232 0 0.046 17081 5493 23475 4155 8	(20: ing supp test 1 0.01 .9964 (0 .4308 1708 .3569 2342 .460% 1 .60%	18-19]up-TO lied uet cross valid 0.9964 11.431 703 75.357 2086 6.40%	TAL per sp 1 0 0 0 3 3 4 5 5 5 5 5 5 5 5 5 5 5 5 5	0.7574 196.6809 27.95% 34.10%	1 use tranning set 0.845 1.488 2.009 551 54,135	1 supp test 5 1 6 2 6 5	(2018-19)s- plied cross set valid 0 fol 0.03 0.0454 1.4885 1.4885 5.555 5.555	total ation(1 d) -0.0707 3.163 4.536 109.19%	persent split 0.7 2.6 86.1 86.1
ssificatio	I Algorithm	Ignored Class Unknown Instances	1 use transi set 0 0.91 28832.91 39248.30 16.0 8.2	201 supplie test se ng 04 (0 07 0.9 76 28832 41 39248, 25 186,0 25 18,0 28	1 1-19)P-TOT// d cross validati 10 fold) 002 1067 0.5 118 467092 106 65.3 25 95.3 05 28-4 28	4 persenta split(66% 0 233 0.7 189 35187 195 47185.9 109 32.1 1	1 1 1 1 1 1 1 1 1 1 1 1 1 1	(20: 100 test 1 0.01 9964 (0 4308 1708 3569 2347 446% 1 5.60% 28	L8-19jup-TO lied et cross valid 0 fold 0.02 11.431 703 75.357 2686 6.40% 8.60% 28	TAL per sp 1 0 0 0 5 5 7 1 6 5 5 5 7 1 6 5 5 5 5 5 5 5 5 5 5 5 5 5	0 0.7574 196.6809 133.4089 27.95% 34.10%	1 use tranning set 0.845 1.488 2.009 551 54,135 2	1 supp test 0 5 1 6 2 6 5 5 5 5 5 5	(2018-19)s- plied cross set valid 0 fol 0.03 0.0454 1.4885 2.0096 55% 4.13% 28	TOTAL (ation(1 d) 0.0707 3.163 4.536 109.19% 116.79% 28	persent split 0.7 2.6 86.1 86.1
LAZY	I Algorithm	Ignored Class Unknown Instances	1 use transi set 0 9.099 28832.91 39248.30 16.0 8.2	201 supplie test se ng 04 (0 07 0.9 76 28832 41 39248, 25 186,0 25 18,0 26 8,3 28 1	1 1-19)P-TOT/ d cross validati 10 fold) 02 107 1067 0.5 118 467692 104 65.3 216 95.3 05 95.3 05 95.3 05 95.3 1	4 persenta split/66% 0 213 0.7 189 35187 195 47185.1 45 26. 096 32. 3	2 use trane) set 0 0 046 17081 9493 23475 41% 20 85% 2 10	(20) (2)) (2))	8-19jup-TO lied et cross valid 0 fold 0.02 3.9964 11.431 703 5.357 2686 6.46% 8.60% 28 1	1 TAL perspective po 0.3344 71.6748 255.7103 255.7103 255.7103 255.7103 21 21 21 21 21 21 21 21 21 21	0 0.7574 196.6809 133.4089 27.95% 34.10% 10	1 use tranning set 0.845 1.488 2.009 555 54,135 24,135 2	1 supp test 0 5 1 6 2 6 5 8 1	(2018-19)s- plied cross set valie 0 fol 0.03 0.0454 1.4885 5.0096 5.55% 2.3 2.3 2.3 2.3 2.3 2.3 2.3 2.3 2.3 2.3	TOTAL (ation(1 d) 0.0707 3.163 4.536 109.19% 116.79% 28 1	persent split 0.7 2.6 86.1 86.1
LAZY	LWL	Inner Ekken to test model Conrelation cardinisati Conrelation cardinisati And manual search error Root relative sogared error	1 use transi set 0 28832.91 39248.30 16.0 8.2	201 suppli- text se 04 (0 67 0.9 76 28832 41 39248. 25 16.0 76 8.2 28 1	1 1-19)P-TOTA d cross validati 10 fold) 02 04 10 fold) 05 05 28 28 1 1	4. persentaj on splitj66% 0 213 0.1 189 33187 195 47185.1 4745 236. 1	ge use trane) set 0 7232 0 0.046 17083 9493 223475 4495 2475 4495 2	(20) (2)) (2))	18-19jup-T0 lied cross valid 0 foli 0.02 0.9964 11.431 702 75.337 2686 6.40% 8.60% 28 1	1 TAL per sp tion[1 0 0.5344 71.6748 21 55.7103 21 65.93% 25.01% 28 1 1	0.7574 196.609 133.4089 27.55% 34.10%	1 use tranning set 0.845 1.488 2.009 551 54.135 2 2	1 test 6 2 6 5 1	(2018-19)s- plied cross set valic 0.03 0.0454 2.0096 55% 4.13% 28 1	10TAL ation(1 d) 0.0707 3.163 4.536 109.19% 28 1	persent split 0.7 2.6 86.7 86.7
LAZY	twt	Time taken to test model Correlation certificate Mara absolute error Root man squared error Root man squared error Root man squared error Root man squared error Total thinking of instances Total studies of instances Time taken	1 use transi set 0.9/ 28832.9/ 39248.30 16.00 8.2	201 suppli- test se 04 (0 07 0.9 76 28832 41 39248 25 156.0 75 8.2 28 1 1 01 (0	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	4. persenta; or split(66% 0 223 0.7 189 35187 195 47185, 195 35187 195 47185, 195 35187 195 3555 195 35555 195 35555 195 355555 195 355555555555555555555555555555555555	2 2 3 3 3 3 3 3 3 3 3 3 5 3 5 3 5 3 5 3	(20) (2)) (2))	18-19 up-TO lied cross valid 0.62 0.9944 703 11.431 703 75.337 2686 0.40% 28 1 0.02	1 TAL person po po 10.5344 21.67488 21.6748 21.6748 21.6748 21.6748 21.67	0 0.7574 196.6809 27.95% 34.10% 10 0.02	1 use tranning set 0.845 1.488 2.009 555 54.135 2 2 0.0	1 test 6 5 1 6 5 1 1 1	(2018-19)s- plied cross set valic 0.03 0.0454 1.4885 0.0056 55% 4.13% 28 1 0.02	totAL attion(1 d) -0.0707 3.163 4.536 109.19% 116.79% 28 1 1 0.06	persent split 0.7 2. 2.6 86.3 86.1
LAZY	twi	Inner taken to test model Correlation caeffraient Maan aboute error Root relative souarde error Root relative souarde error Root relative souarde error Root relative souarde error Inne taken Correlation caeffraient	1 use transi set 0 0.91 39248.30 16.0 8.2 9248.30 16.0 8.2 9248.30 10.0 9 0.91 0.91	(201 suppli- test se 04 (67 0.9 76 28832 41 39248 28 1 1 05 8.2 28 1 0 01 (97 0.9	1 1) 1) 1) 1) 1) 1) 1) 1) 1) 1	4 persenta on[split[66% 223 0. 223 0. 225 0. 22	0 use traini 0 set 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	(20) (20) test (0,01 (9964 (4308 1708 3569 2341 (46% 1 1,60% 28 1 0,05 (9997 (18-19jup-T0 lied et cross valid 0.0 folo 0.9964 11.431 703 5.357 2686 6.40% 28 1 0.02 0.997	1 FAL Person (1) 0 0.5344 71.6748 21.55.7103 21.65.7103	rsentage it(66%) 0.7574 196.6809 27.95% 34.10% 10 0.02 0.9034	1 use tranning set 0.845 1.488 2.009 555 54,135 2 2 0.0 0.0 0.98	1 supp test 6 2 6 5 1 6 2 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5	(2018-19)- plied cross set valic 0 fol 0.03 0.0454 1.4885 2.0096 55% 2.8 1 0.00 0.905	107AL ation(1 d) -0.0707 3.163 4.536 109.19% 116.79% 28 1 0.06 0.1115	persent split 0.7 2. 86.3 86.1 0.5
LAZY	I Algorithm	Trime taken to test model Correlation cselficient Root realized error Root mean squared error Root mean squared error Root means squared error Root realized sublick error Root realized sublick error Root realized sublick error Root realized error	1 use transi set 0.99 28832.91 39248.32 16.0 8.2 0 0 0 0.99 45048.5	1 supplie test se ng 04 00 07 0.9 76 28832 41 39248. 25 16.0 76 8.2 28 1 1 01 0 97 0.9 42 45068.	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	4 persenta on (split(66% 213 0.1 189 35187 195 47185, 1 189 35187 195 47185, 2 105 003 0.1 495 81027	2425 2455 2455 2455 2455 2455 2455 2455 2455 2455 2455 2455	(20) 100 (20)	III-19jup-TO III-0 III-0 Valid O fold 0.02 3994 11.431 705 5.307 28 6.40% 8.00% 1 0.02 9997 198.31	1 FAL PP 59 10 10 10 10 10 10 10 10 10 10 10 10 10	0 0.7574 196.6809 133.4089 34.1055 34.1055 10 0.02 0.9034 046.9889	1 use tranning set 0.845 1.488 2.009 555 54,135 2 2 0.0 0.0 0.98 0.92	1 supp test 6 2 6 5 7 1 1 5 5 2 2	(2018-19)s- plied cros set valic 0 fol 0.03 1.0454 1.4885 2.0096 5.5% 2.8 2.0096 1.28 2.0096 0.02 0.985 0.922	107AL ation(1 d) 0.0707 3.163 4.536 109.19% 116.79% 28 10.066 0.1115 2.6515	persent split 0.7 2.2.6 86.1 86.1 0.5 1.
LAZY	I Algorithm	Time taken to test model Correlation coefficient Correlation coefficient Maar absolute error Relative advolute error Relative advolute error Rodor relative spaced error Total Student Instances (general Class Univers) Instances (general Class Univers) Instances (Correlative spaced error Total Studen Instances Correlation spaced error Maar absolute error Maar absolute error Maar absolute error Maar absolute error	1 use transi set 0 39248.3 139248.3 139248.3 139248.3 10.0 0 0 0 0 0 0 138113.22	201 supplie test se 1 204 007 0.9 208 21 20 21 20 21 20 20 21 20 20 20 20 20 20 20 20 20 20	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	4 persenta on(split/66% 223 0.1 195 47185,1 195 47185,1 0% 32.2 1 0% 32.3 0.5 0% 32.3 0.5 10% 10% 10% 10% 10% 10% 10% 10%	2 1 2 2 2 3 2 4 2 4 2 4 2 4 2 4 2 4 2 4 2 4	(20) ing supp test 1 0.01 9964 C 4308 1706 3569 2345 445% 1 1 8.80% 28 1 0.05 9997 C 3995 265 1679 7958	18-19jup-T0 lied cross valid 0 fold 0.02 0.9964 11.431 703 75.357 2088 6.46% 8.60% 28 1 0.02 0.9997 198.31 825 3.9997	1 TAL period sp sp stion(1) 0 0.3344 71.6748 23 55.7103 28 0.055 28 1 1 0 0.0902 0.0902 1 1 1 1 1 1 1 1 1 1 1 1 1	0 0.7574 196.6809 27.95% 34.10% 10 0.02 0.9034 046.9889 3556.6368	1 use tranning set 0.845 1.488 2.009 551 54,135 2 2 0.0 0 0.98 0.92 1.353	1 test 6 5 5 1 1 5 2 2 9 1	(2018-19)s- filed cross set valid 0 fol 0.03 0.0454 4.4355 23 4.1355 23 0.02 0.902 0.902 1.3539	TOTAL ation[1 d] 0.00707 3.163 4.536 109.19% 28 1 1.6.79% 28 1 0.06 0.1115 2.6515 3.7361	0.7 2. 2.6. 86.1 86.1 0.5 1. 2.5
LAZY	Random fores	Ignored Class Unknown Instances Inner Salari to staf model Construitation (selffruid) Massa absolute error Ratative absolute error Ratative absolute error Ratative absolute error Time Islam Time Islam Root men spagned error Root men spagned error Root men spagned error Root men spagned error Root men spagned error	1 use transi set 2832.51 29243.36 3232.51 23243.36 3.6 3.6 0.99 4.5048.3 138113.2 25.0	201 suppli- test se ng 04 (0 67 0.9 76 28832. 41 39248. 25 186. 275 8.2 28 1 01 (0 97 0.9 42 45048. 3 138111 55 225.0	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	4 persenta of split(66% 0 223 0.1 195 477185,1 195 477185,1 195 477185,1 105 083 0.2 213 80602,2 213 80602,2 213 80602,5 213 907,5 213 907,5 213 907,5 213 907,5 214 907,5 215 907,	2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	(20) 1000 1000 1000 1000 1000 1000 1000 100	Image: 10 per sector Image: 10 per sector Image: 10 per sector Valid 0 folds 0.02 0.9964 1 1.431 702 0.60% 0.60% 0.60% 0.60% 0.02 1 0.02 1 0.9997 198.31 99977 153.36% 5.34% 27 5.34% 27	1 TAL per sp tion[1] 0 0.5344 77.6745 25.7103 265.935 25.935 28 1 0.05 0.0902 81.1059 41.175.263 44 77.7455	0 0.7574 196.6809 1133.4039 27.95% 34.10% 34.10% 0.02 0.9034 0.46.9889 556.6368 55.77%	1 use tranning set 2.009 557 54,135 2 0.0 0.98 0.92 1.353 33.867	1 test 6 5 1 1 1 1 5 2 2 9 1 6 3	(2038-19)s- plied cross set valic 0 fol 0.03 3.0454 4.4855 2.0096 55% 4.13% 2.28 0.02 0.922 0.922 1.5539 3.86%	TOTAL attion[1 0 0.0707 3.1636 109.19% 28 1 0.066 0.1115 3.7361 3.7361 91.53%	persent split 0.7 2.6 86.1 86.1 0.5 1.5 2.5 75.1
LAZY	LWL Random fores	Ignored Class Unknown instances	1 use transi set 0.99 28832.91 3948.30 3949.30	201 supplie test se 67 0.9 76 28832 41 39348 15 26.0 76 8.2 28 1 01 (97 0.9 42 45048. 39 138113 55 25.0 85 29.1 1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	4. on split(66%) 0 2213 0.1 185 33187 954 47185,7 185 33187 954 47185,7 1954 47185,7 105 105 105 105 105 105 105 105	0 0 0 0 7232 0 0 0 7234 0 0 0 0 0 <	(20) 100 100 100 100 100 100 100 100 100 100	I8-19jup-T0 lied cross valid 0 fold 0 fold 0 fold 0.02 5.357 2086 6.40% 8.60% 8.60% 8.60% 9.9997 198.31 829 17.168 273 5.34%	1 TAL person sp 0 0.5344 71.6745 23 55.7103 24 65.9355 25 1 0.055 0.0902 81.1059 41 175.263 44 175.263 44 175.263 44 175.768 45 177.7785 10 177.7785 10 177.7785 10 10 177.7785 10 10 10 10 10 10 10 10 10 10	0 0.7574 196.6809 27.95% 34.10% 10 0.02 0.9034 046.9889 556.676% 56.74%	1 use tranning set 0.845 1.488 2.009 551 54.135 2 2 0.0 0.98 0.92 1.355 33.860 36.471	1 test 0 5 5 1 6 5 5 2 9 9 1 1 5 5 2 9 9 1	(2018-19)6- pied cross set valic 0 fol 0.03 0.0456 1.4885 0.0056 5575 1. 28 1 1 0.02 0.985 0.922 1.3539 0.922 1.3539	TOTAL ation(1 d) 0 0.0707 3.163 4.536 109.19% 28 109.19% 21 10.066 0.1115 2.63756 2.63756 3.53765 9.53% 96.19%	persent split 0.7 2.2 86.1 86.1 9.5 1. 2.5 5.1 84.3 84.3
LAZY TREE	Random fores	Inner alkan to test model Constitution caefficient Constitution caefficient Constitution caefficient Relative abolistic error Relative abolistic error Rost relative spaared error	1 use transi set 0.91 28832.91 39248.32 139248.32 0.0.91 45048.3 138113.21 25.0 29.3	201 supplie test se 6 6 6 7 6 7 6 8 7 6 8 7 6 1 9 6 7 6 1 9 7 6 2 8 7 6 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	4 persentajon (splitj66% 0 223 0. 195 32187 195 47185, 195 32 195 32 105 082 0. 195 81007 213 80002 213 80002 213 80002 213 80002 213 80002 213 80002 213 80002 214 8007 215 99, 215 99, 215 99, 215 99, 216 99, 217 99, 218 90, 218 99, 218	0 use traine 0 set 0 set 0 202 0 064 0 20473 10 2035 0 2020 0 2020 0 2020 0 2020 0 2020 0 2020 9556 20 9557 22 10 201	(20: 1000 test 1 0.01 (0.000) 0.01 (0.000) 0.01 (0.000) 0.000 (0.000) 0.0000 (0.000) 0.000 (0.000) 0.0000 (0.000) 0.0000 (0.000) 0.0000 (0.000) 0	18-19)up-T0 Hed valid 0 fok 0.02 39964 1.431 703 5.337 2086 6.46% 8.60% 28 1 0.02 23 1.431 703 7.337 2086 6.46% 8.60% 23 1.431 829 1.431 829 1.4	1 TAL per sp tion(1) 0 0.3344 0.3344 55.7103 28 1 0.05 0.0902 1 1 0.05 0.0902 1 1 28 1 1 28 1 1 28 1 1 1 1 1 1 1 1 1 1 1 1 1	rsentage 0 0.7574 196.8809 133.4089 14.1055 14.1055 10 0.002 0.9034 10 0.002 0.9034 10 0.002 0.9034 10 0.002 0.9034 10 0.002 0.9034 10 0.002 0.9034 10 0.002 0.000	1 use transing set 0.8455 1.4888 2.0090 555 54.132 2 0.0.098 0.922 0.922 0.922 0.922 0.922 0.925 1.3586 1.6477 2 2	1 supp test 0 0 0 0 0 0 0 0 0 0 0 0 0	(2018-19)s- plied cross set valic 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.	TOTAL ation(1 0 0 0.0707 3.163 4.536 119.19% 28 1 0.066 0.11155 2.65155 3.7366 1.538 96.19% 28	persent split 0.7 2.2 86.3 86.4 0.5 1. 2.5 7 7.5 84.1
LAZY	Random fores	Ignored Class Unknown instances Time taken to first model Conrelation quefficient Mass abouts entry and entry Read abouts entry Read abouts entry Read abouts entry Read abouts entry Read	1 use transi set 0.99 2832.25,1 239248.36 1.6.0 8.2 0.99 45048.3 138113.2 25.0 29.3	281 supplie test se 67 0.9 76 2832 41 39248. 10 01 (97 0.9 42 45068. 10 01 (97 0.9 42 45068. 11 10 10 (97 0.9 42 45068. 13 138111 55 226. 1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	4. persentajo splitj66% 0 0 223 0.2 195 35185, 195 47185, 195 47185, 195 47185, 195 35185, 195 47185, 195 35185, 195 35185,	0 use training 0 set 0 set 0 2475 10 2475 110 2455 0 0	(20) (2)) (2))	18-19]up-TO lied cross valid 0 fold 0 fold 0.02 5:357 2088 6.40% 8.40% 2.8 1 1 0.02 2.8 9.9964 1 1.431 702 5.357 2088 6.40% 8.40% 1 1 1 2.8 9.9964 1 1 1 2.8 9.9964 1 2.8 9.9964 1 2.8 9.9964 1 2.8 9.9964 1 2.8 9.9964 1 2.8 9.9964 1 2.8 9.9964 1 2.8 9.9964 1 2.8 9.9964 1 2.8 9.9964 1 2.8 9.9967 1 2.8 9.997 1 2.5 3.57 2088 6.40% 8.40% 9.022 1 1 2.8 9.9967 1 2.8 9.9977 2.8 9.1857 2.8 9.156 1.4 1.4 1.4 1.4 1.4 1.4 1.4 1.4	1 TAL perspective sp 1 0 0 0 0 0 0 0 0 0 0 0 0 0	rsentage nt[66%] 0 0.7574 196.8009 133.4089 27.55% 34.10% 34.10% 34.10% 34.10% 34.10% 34.10% 34.10% 35.6368 35.6368 35.6368 35.6368 35.75% 36.8368 36.75% 36.8368 36.75% 36.8368 36.75% 36.8368 36.75% 36.8368 36.75% 36.8368 36.75% 36.8368 36.8368 36.75% 36.8368 36.93688 36.93688 36.93688 36.93688 36.936888	1 use transing set 0.845 555 54.135 2 0.0.98 0.022 1.355 3.3.660 18.647 2	1 supp test 0 0 0 0 0 0 0 0 0 0 0 0 0	(2018-19)s- plied cross of of 0.03 3.0454 4.135 2.0095 1 1 0.02 0.985 1 1 0.02 0.985 1.5539 3.88% 6.47% 2.28 2.29 2.29 2.29 2.29 2.29 2.29 2.29	TOTAL 4 4 4 4 4 6 100,100 100,000 200 200 200 200 200 200 20	persent split 0.7 2.8 86.1 86.1 9.5 9.5 1.1 2.5 7 5.1 84.1
LAZY	Random fores	Time taken to test model Correlation caefficient Correlation caefficient Correlation caefficient Correlation caefficient Relative aboute error Relative error Relative aboute er	1 use transi set 0.99 28832.93 39244.31 180.0 8.2 0.99 45048.1 118113.21 229.3	1 1 2011 supplify test se 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	4 persentajan 6 splitj6656 1213 0.0 1213 0.0 1215 0.		(20: 1000 1000 1000 1000 1000 1000 1000 100	18-19]up-TO lied cross valid 0 folc 0.02	1 TAL Person Sp tion(1 Sp 0 0.5344 71.6748 21 55.7103 28 0.055 28 0.055 28 1 0.055 28 1 0.059 41 175.263 44 175.263 44 175.263 44 175.77485 28 28 28 28 28 28 28 28 28 28	rsentage 0 0.7574 133.4089 27.35% 34.10% 0.02 0.9034 10066.889 56.75% 556.486 56.77%	1 transing set 0.845 2.009 555 54.135 2 0.0.98 0.02 0.98 0.02 0.98 0.02 0.355 33.860 13.551 33.861 18.427 2	1 supp test 6 5 2 2 9 1 6 3 3 1 1 6 3 3 1 1 6 3 3 1 1 6 3 3 1 1 6 5 1 1 1 6 5 1 1 1 6 5 1 1 1 6 5 1 1 1 6 5 1 1 1 6 5 1 1 1 1	(2018-19)s- plied cross set valic of of 0.03 0.055 0.0056 1.4885 0.0056 1.4885 0.005 0.922 1.5539 0.922 1.5539 0.922 1.5539 0.922 1.5539 0.4275 2.88% 0.4275 2.88%	TOTAL atlon[1 0 0 0.0707 3.163 4.536 109.19% 28 109.19% 28 109.19% 28 116.79% 28 15.79% 28 28 26 25.55 27.55% 28 26 26 27.55% 28 26 26 25.55% 28 28 28 28 29 28 28 28 28 28 28 28 28 28 28	persent split 0.7 2.2 86.1 86.4 86.4 86.4 86.4 86.4 86.4 86.4 86.4
LAZY	Algorithm	Ignored Class Unknown Instances	1 use tranel set 0.99 2832.51 2832.52 2832.52 2832.52 2832.52 0.99 45048.31 138113.22 25.0 0.99 45048.31 138113.22 25.0	[201] (201] (upplication) (100) (1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	4 persenta of plate66% of 223 0.0 185 35187 195 47125.1 47	ge use trann 0 set 0	(20) (2)) (2))	18-19jup-TO lied cross valid 0 folo 0.02 0.5364 13.431 703 55.357 208 6.40% 28 1 28 1.9992 28 1.99937 28 1.7.168 273 5.34% 28 1 0 9.13% 28 1 0 9.935 1	1 TAL per sp tion[1] 0 0.3344 7.6748 23 53.015 28 0.05 0.0902 81.1059 41.175.263 41.175.263 41.175.263 41.175.263 41.175.263 41.175.263 41.177.4% 36.0155 28 1 0.015 28 1 0.015 28 1 0.015 28 1 0.015 28 1 0.05 0.0902 81.1059 41.174% 28 1 0.015 0.05 0.0902 81.1059 41.174% 28 1 0.015 0.05 0.0902 81.1059 1 0.05 0.0902 81.1059 1 0.05 0.0902 81.1059 1 0.05 0.0902 81.1059 1 0.05 0.0902 1 0.05 0.0902 1 0.05 0.0902 1 0.05 0.0902 1 0.05 0.0902 1 0.05 0.0902 1 0.05 0.0902 1 0.05 0.0905 0.0902 1 0.05 0.0902 1 0.05 0.0902 0.05 0.0002 0.0	rsentage it[665() 0.7574 196.8809 27.95% 34.10% 34.10% 34.10% 356.81868 556.81868 556.81868 556.81868 556.81868 556.81868 0.0000 556.81868 556.81868 0.0000 0.00000 0.0000000000000000000	1 transing set 0.8455 1.4886 2.009 5555 54.133 2 2.009 5555 54.133 2 2.009 5555 54.135 2 2.009 5555 54.135 2 2.009 5555 54.135 2 2.009 5555 54.135 2 2.009 5555 54.135 2.009 5555 54.135 2.009 5555 54.135 2.009 5555 54.135 2.009 555 54.135 2.009 555 54.135 2.009 555 54.135 2.009 555 54.135 2.009 555 54.135 2.009 555 54.135 2.009 555 54.135 2.009 555 54.135 2.009 555 54.135 2.009 555 54.135 2.009 555 54.135 2.009 555 54.135 2.009 555 54.135 2.009 5.000 5.000 5.000 5.000 5.000 5.000 5.0000 5.0000 5.0000 5.0000 5.0000 5.0000 5.0000 5.00000 5.00000 5.00000000	1 supp test 6 5 2 9 1 1 5 5 2 9 1 1 5 5 2 9 1 1 5 5 2 9 1 1 5 5 2 9 1 1 5 5 2 9 1 1 5 5 7 7 9 1 1 9 1 9 1 9 1 9 1 9 1 9 1 9 1 9	(2018-19)+ plied cross of of 0.003 0.0454 1.4885 2.0096 5575 1.28 0.002 0.985 0.992 1.3539 0.922 1.3539 2.88% 6.47% 2.8 1 0.992 1.3539 1.28% 1.29% 1.2	TOTAL ation[1 0 0 0.0707 3.163 109.19% 28 111.6.79% 28 3.7361 11.6.79% 28 96.19% 28 96.19% 28 10.006 0.01115 28 10.006 0.007 0.007 11.6.79% 28 10.006 0.007 11.6.79% 28 10.006 11.6.79% 28 10.006 11.6.79% 28 10.006 11.6.79% 28 10.006 10.007	persent split 0.7.2 2.6 86.6 86.6 0.5 1. 2.5 5.5 1. 84.1 84.1
LAZY TREE META	Random fores	Ignored Class Unknown instances Time taken to test model Correlation coefficient Correlation coefficient Mara abolute arror Roor reality squared arror Roor Roor Roor Roor Roor Roor Roor Roor Roor	1 use tranel set 0.9 39248.3 136.0 8.2 0.9 45048.3 138113.2 29.3 29.3 29.3 29.3 29.3	2011 2012 2014 2014 2014 2014 2014 2014	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	4 persentajon (splitje655 0 0 0 115 31187 115 115 115 115 115 115 115 11	ge use trans set 0 0 0 046 17081 9993 224475 100 100 1011 26298 3055 20 100 1011 26298 3055 20 100 1012 26298 3055 22 310 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	(20) (2)) (2))	18-19]up-TO lied cross valid, 0 fold 0 fold 0.02 5.9594 1.431 703 75.337 2086 6.46% 2.8 1.431 703 75.337 2086 6.46% 2.8 1 0.02 9.9997 198.31 829 7.158 273 5.46% 2.8 9.15% 2.8 1 0 0 0 5.9935 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	TAL per per per per per per per per	rsentage 0 0,7574 196,8809 133,4089 34,1055 34,1055 36,405 356,8388 3556,8388 3556,8388 3556,8388 355775 36,4459 10 0 0,022 0,020 0,00000000	1 use transing set 2.009 555 54,135 2 2 0.0,0 9 2 1.351 33,867 2 2 0.099 2 0.092 1.351 16,477 2 0.0931 1,6 10 10 10 10 10 10 10 10 10 10 10 10 10	1 supp test 0 0 0 0 0 0 0 0 0 0 0 0 0	(2018-19)s- plied cross set valie 0 fol 0.03 3.8455 2.0096 5.5% 4.13% 0.096 0.922 1.5539 0.922 1.5539 0.922 1.5539 1.388% 6.47% 2.888% 1.155 2.0096 0.922 1.5539 1.1539 2.0096 1.539 2.0096 2.0097 2.0096 2.0006 2.0	TOTAL ation[1 0 0 0.0707 3.163 109.19% 28 11 0.066 0.11155 2.6515 3.7361 9.13% 12 9.13% 13.7361 13.7361 13.7361 14.5866 14.5866 14.5866 14.5866 14.5866 14.586	persent split 0.2 2.6 86.1 86.1 2.5 75.5 84.1 84.1
IAZY TREE	Random fores	Inner Eslan to test model Control to test model Restance about error Restance about error Root mana signande error Root mana signand	1 use tranel set 0.99 2883.2; 39348.3; 39348.3; 39348.3; 8-2 0.99 45048.3; 138113.2; 225.0 0.99 45048.3; 138113.2; 229.3 2019.5; 0.97 77003.3; 292195.4; 202	1 1 201 201 201 201 201 201 201 201 201	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	4 persentag of of particless of of of of of of of of of of	ge use trane 0 set 0	(20:00	L8-19]up-T0 Hed valid of old 0 fold 0 fol	TAL store(1 0 0 0.5344 1 0.0 0.55101 2 85.935 1 1 0.051 2 85.935 1 1 0.052 2 85.0156 2 2 8 0.0962 2 8 0.0962 1 0 0 0 0 0 0 0 0 0 0 0 0 0	rsentage 0 0 0.7574 196.8809 133.0095 34.10% 34.10% 34.10% 34.10% 556.8398 34.10% 556.8495 10 0 0 0 0 0 0 0 0 0 0 0 0 0	1 use tranning set 0.845 1.488 2.009 555 54.135 2 2.055 54.135 2 2.059 8.092 1.355 3.865 3.8.865 3.8.865 3.8.865 3.8.855 3.855 3.8.855 3.8.855 3.855 3.8.855 3.8.855 3.8.855 3.8.855 3.8.855 3.8.855 3.8555 3.8.855 3.8.855 3.8555 3.8.855 3.8555 3.8555 3.8555 3.8555 3.8555 3.8555 3.85555 3.85555 3.855555 3.85555555555	1 supp test 0 0 0 0 0 0 0 0 0 0 0 0 0	(2018-19)6 plied cross of all of fol 0.003 0.005	TOTAL attion[1 0 0 0.07070 3.163 4.536 109.19% 116.79% 28 13 0.066 0.1155 2.6515 2.6515 3.7361 91.53% 96.19% 11 0.068 0.068 0.068 0.070 0.00	persent split 0.7.2 2.8 86.4 86.4 9.5 9.5 75.6 75.6 9.4 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5
LAZY TREE META	Random forest	Ignored Class Unknown instances Time taken to field model Controllation quell'Instant Controllation quell'Instant Controllation quell'Instant Controllation quell'Instant Controllation quell'Instant Controllation quell'Instant Generation could' Class Unknown Instances Ignored Class Unknown Instances Controllation seglificant Root mans signated error Root mans signated error Root mans signated error Class Unknown Instances Unstantann Controllation coefficient Maria abolothe error Class Unknown Instances Unstantan Controllation coefficient Maria abolothe error Class Unknown Instances Unstantan Controllation coefficient Maria abolothe error Class Unknown Instances Unstantan Controllation gentificant Maria abolothe error Maria error	1 use tranel est 0.99 136.0 8.2 0.91 45041 138113.2 29.3 29.3 29.3 29.3 225.0 255.0	2011 2012 2014 2014 2014 2014 2014 2014	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	A persentaj splitj66% 0 223 0.0 189 3518795 47185,25, 9% 47185,25, 9% 47185,25, 9% 47185,25, 9% 47185,25, 9% 47185,25, 105 083 0.0 283 60,02 283 60,02 284 60,02 285 60,020 285 60,020 295 60,020 295 60,020 205 60,0200 205 60,0200 205 60,020000000000000	2 2 g# use trans 0 set 0 0 7232 0 0 0 0 17033 493 224475 203 23055 201 262588 202 75651 203 2555 203 2555 203 255 203 255 203 255 203 255 204 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	(20:0) (20:0)	Image: Constraint of the second sec	TAL FAL FAL FAL FAL FAL FAL FAL F	rsentage 0 0 75744 196.6809 133.4089 27.95% 134.10% 10.022 356.6368 56.77% 56.44% 10 0 0 0 720.0224	1 use tranning set 0.8454 2.009 5555 555 555 555 555 555 555 555 555	1 test test 1 1 test 1 1 5 5 2 9 1 6 32 6 32 1 1 5 5 2 9 1 6 32 1 1 5 5 1 1 6 32 1 1 5 5 1 1 5 1 1 5 5 1 1 5 5 1 1 5 5 1 1 5 5 1 1 5 5 1 1 5 5 1 1 5 5 1 1 5 5 1 1 5 5 1 1 5 5 1 1 5 5 5 5 1 1 5 5 5 1 1 5 5 5 5 5 5 5 5 5 5 5 5 5	(2018-19)- piled cross valid 0 fol 0 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.	TOTAL ation[1 0 0 0.0707 3.163 109.19% 116.79% 285 116.79% 285 116.79% 285 13.7361 91.53% 91.53% 90.19% 20.0000 20.0000	persent split 0.7 2. 2.6. 86. 86. 86. 1 2.5. 75. 75. 75. 75. 84. 85. 75. 84. 85. 75. 84. 85. 75. 84. 85. 75. 75. 75. 75. 75. 75. 75. 75. 75. 7
LAZY TREE META	Random forest	Time taken to test model Correlation caefficient Correlation caefficient Correlation caefficient Correlation caefficient Relative aboute error Relative aboute error Root relative seguered error Root relative seguered error Root relative seguered error Root relative seguered error Root ness seguered error	1 use transi set 0 0 19348.3 193048.3 193048.3 0 0 0 0 0 0 0 0 9 193048.3 10 0 0 9 193048.3 10 0 0 9 193048.3 10 0 193048.3 10 0 193048.3 10 10 193048.3 10 10 193048.3 10 10 10 10 10 10 10 10 10 10	201 201 201 201 201 201 201 201 201 201	1 1 1 cross validati 10 ford) 10 cross validati 10 ford) 10 cross validati 10 ford) 10 cross validati 10 cross validati	A persentary or split(66% 0 213) 0. 213) 0. 213) 0. 213) 0. 213) 2512 214 215] 2512 215] 2512 25	c c c c	(200) (2	18.53/up-TO Hed cross vitil 0.62 9.996 1 0.02 9.997 28 0.02 9.997 28 0 1.15% 28 0 1.15% 23 0 1.315% 0 1.315% 0 1.215% 1.215%	TAL TAL FAL FAL FAL FAL FAL FAL FAL F	rsentage it[65%] 0 0.7574 196.8809 133.4095 133.4095 133.4095 133.4095 133.4095 133.4095 133.4095 133.4095 133.4095 133.4095 133.4095 10.002 10.002 10.002 12.27%	1 use transing set 0.845 1.4888 2.000 0.928 0.928 0.028 0.988 0.028 0.988 0.028 0.988 0.028 0.988 0.028 0.988 0.045 0.988 0.045 0.985 0.045 0.045 0.985 0.045	1 supp test test 6 2 6 3 6 3 6 3 7 9 1 6 3 8 1 1 1 1 1 1 1 1 1 1 1 1 1	(2018-10)- piled cross of valid of of 0.03 0.03 0.0454 1.055 0.045 0.045 0.045 0.045 0.045 0.045 0.045 0.045 0.045 0.047	TOTAL attion(1 0 0 0.0707 109.19% 119.19% 0.1115 3.7361 3.7361 3.13752 0.033 0.1115 3.13752 0.033 0.033 0.033 0.033 0.033 0.031 0.035 0	persent split 0.7 2.8 86.1 86.1 9.5 75.5 75.5 75.5 75.5 75.5 75.5 75.5

Table 7.Represented data in the year 2018-19 total anddropout value by running weka platform.



SCIENCES ISSN PRINT 2319 1775 Online 2320 7876 Research Paper © 2012 IJFANS. All Rights Reserved, Journal Volume 11, Iss 11, 2022



Table 8.Represented data in the year 2017-18 total and dropout value by running weka platform.

Table 6,7 and 8 show here three years big data classify represented here from 2019-2020, 2018-2019, 2017-2018 and three data mining techniques LAZY, TREE, META present 3 data mining algorithm like LWL, Random forest, Bagging.They process primary total, dropout, upper primary total dropout and secondary total dropout data in four test option like use training data set, supplied test set, 10 fold cross validation and 66% percentage split. After running this test option they all show different results as time taken to test model, correlation coefficient, mean absolute error, Root mean squared error, Relative absolute error, Root relative squared error, Total Number of Instances, Total Number of Instances.

A. Analysis And Discussion

Given below are the result based on Statistical Parameters of the Year 2019 – 2020, 2018-19, 2017-2018, 2016-2017.





ISSN PRINT 2319 1775 Online 2320 7876

Research Paper © 2012 IJFANS. All Rights Reserved, Journal Volume 11, Iss 11, 2022

Fig.4Graphical representation on P-total, dropout, UP- total, dropout and S-total, dropout three year dataset Year 2019 – 2020, 2018-19, 2017-2018, 2016-2017.

IV. CONCLUSION

Educational data mining play an important role in higher education system, the use of rising technology need to largest dataset. With the help of U-DISE (unified district information system for education) we get overall type of big data as related to school information like students, faculty members and dropout student etc. it is government authorized nationalized platform, which is use in data mining concept to predict student performance. By this study provide improvement in public and private field to improve academic performance. It can be concluded that classification techniques provide better accuracy compared to other approach.

References

- Zambrano J.L., Lara Torralbo J.A., and Romero C., "Early Prediction of Student Learning Performance Through Data Mining: A Systematic Review", ISSN 0214 - 9915 CODEN PSOTEG, Vol. 33, No. 3, 456-465doi: 10.7334 /psicothema 2021.62 (2021)
- [2] Viswanathan S., Vengatesh S.K., "Study Of Students' Performance Prediction Models Using Machine Learning", Turkish Journal of Computer and Mathematics Education ,Vol.12 No.2 (year-2021), 3085 – 3091
- [3] Ingale N.V., Dr. Sivakkumar M., Dr. Namdeo V., "Survey on Prediction System for Student Academic Performance using Educational Data Mining", Turkish Journal of Computer and Mathematics Education Vol.12 No.13 (2021), 363-369
- [4] Mangat P. and Saini K., "Educational Data Mining Tools and Framework for Predicting Students Academic Performance", International Journal of Advance Science and Technology Vol. 29, No. 10S, (2020), pp. 2525-2533.
- [5] Zhou N., Zhang Z, Li J., "Analysis on Course Scores of Learners of Online Teaching Platforms Based on Data Mining", Ingenierie des Systemes d'information vol 25, No.5, pp.609-617.October 2020
- [6] Ahmad N.B.B, Ajibade M.S, Shamsuddin M.S., "Educational data mining : Enhancement of Student Performance model using Ensemble methods" IOP Conf. Series: Materials Science and Engineering 551 (2019) 012061,doi:10.1088/1757-899X/551/1/012061
- [7] Hossain Z., Akhtar N., Ahmad R.B., Rahman M., "A dynamic Kmeans clustering for data mining". Indonesian Journal of Electrical Engineering and Computer Science.Vol. 13, No. 2, pp. 521~526 ISSN: 2502-4752, February 2019.
- [8] Ali N., Neagu D., Trundle P., "Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets", SN Applied Sciences (2019) 1:1559.
- [9] Sankari A., Masih S., Ingle M., "A Review On Research Areas In Educational Data Mining And Learning Analytics", International Journal of scientific & technology research volume 8, ISSUE 12, December 2019 ISSN 2277-8616.
- [10] Yaacob W.F.W., Nasir S.A.M., Yaacob W.F. W., Sobri N.M., "Supervised data mining approach for predicting student performance" Indonesian Journal of Electrical Engineering and Computer Science Vol. 16, No. 3, December 2019, pp. 1584~1592 ISSN: 2502-4752, DOI: 10.11591/ijeecs.v16.i3.pp1584-1592.
- [11] Majeed I, Naaz S., "Current State of Art of Academic Data Mining and Future Vision", Indian Journal of Computer Science and Engineering". e- ISSN: 0976-5166, p-ISSN: 2231-3850April 2018.

- [12] Qiao X. and Jiao H., "Data Mining Techniques in Analyzing Process Data: A Didactic", original research published: 23 November 2018, doi: 10.3389/fpsyg.2018.02231.
- [13] More SS, Narain B, Jadahv BT (2017) A comparative analysis of unimodal and multimodal biometric systems. In: International conference on innovative trends in engineering science and management (ITESM-2017).
- [14] Narain B, Zadgaonkar AS, Kumar S (2013) Impact of digital image processing on research and education. Natl Semin Work.
- [15] More S.S., Narain B., Jadhav B.T. (2021) Advanced Encryption Standard Algorithm in Multimodal Biometric Image. In: Rizvanov A.A., Singh B.K., Ganasala P. (eds) Advances in Biomedical Engineering and Technology. Lecture Notes in Bioengineering. Springer, Singapore. https://doi.org/10.1007/978-981-15-6329-4_7
- [16] Data encryption standard algorithm in multimodal biometric image, SS More, B Narain, BT Jadhav - Int. J. Comp. Sci. Eng, 2018.
- [17] Hybrid Support Vector Machine and Distance Classifier in Breast Tumor DetectionU Sharma, B Narain, V Nohria - SPAST Abstracts, 2021
- [18] Student Satisfaction in Educational Organization using Machine Learning, P Singh, B Narain NOVYI MIR, 2021.
- [19] Impact of digital image processing on research and educationB Narain, AS Zadgaonkar, S Kumar Natl Semin Work, 2013
- [20] The Online Retail Market Analysis for Social Development with Machine Learning M Nayak, B Narain - SPAST Abstracts, 2021.
- [21] On-Line Big Data Analysis using K-Mean and Modified K-Mean Algorithm with Machine Learning Techniques* B Narain 2021.
- [22] Big Data Mining Algorithms for Predicting Dynamic Product Price by Online Analysis* M Nayak, B Narain - Computational Intelligence in Data Mining, 2020.

