IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876

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

Dr. Bhavana Narain
Professor
MATS School of Information Technology
MATS University, Raipur (C.G.), India
Email: narainbhawna@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 e-learning 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 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 K-means 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.

IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL 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).

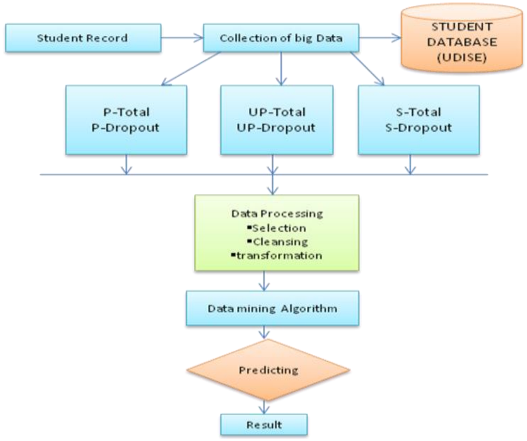

Fig.1 Prediction of student performance using UDISE

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

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

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

			28(97%)
StdDev	271452.2	2.513	

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

Table 3 . Show year 2018-19 statistical representation

(2018-19)

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

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

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

Table 4 . Show year 2017-18 statistical representation

(2017-2018)

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

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

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 28(97%)
StdDev	173589	3.566	

Table 5. Show year 2016-17 statistical representation

(2016-2017)

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

IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
, Journal Volume 11, Iss 11, 2022

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

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

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

Hypothesis$_1$: Is there no relation between transition rate, promotion rate and dropout rate?
Hypothesis$_2$: Are data gaps deteriorating data quality?
Hypothesis$_3$: 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 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.

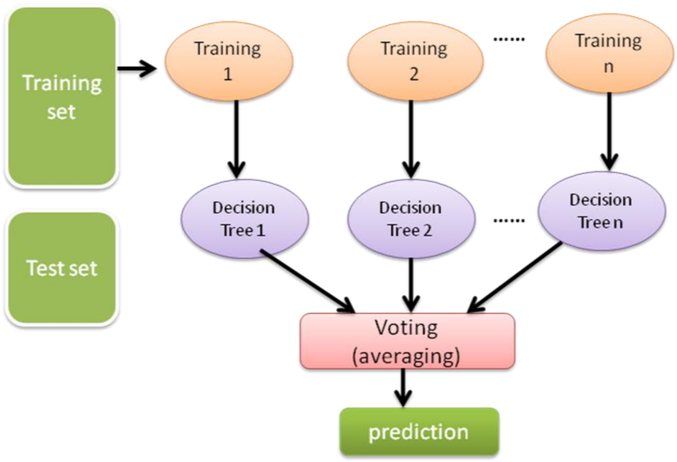

Fig.2 The above diagram explains the working of Random forest.

- **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.

IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876

Journal Volume 11, Iss 11, 2022

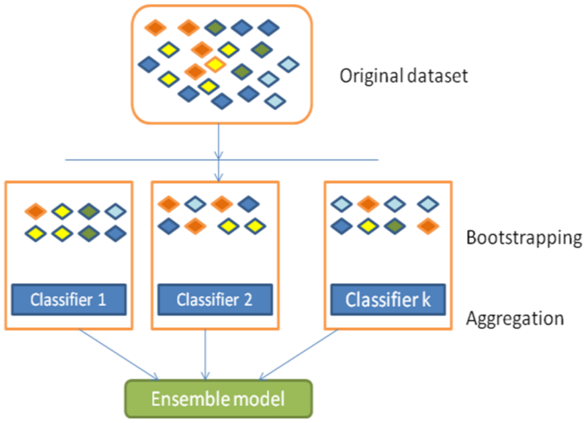

Fig. 3 This diagram presents the Bagging process.

- **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 × 100

Specificity = TN/ N × 100

Precision = TP/ TP + FP × 100

Recall = TP/ P × 100

F1 − Measure = (2 × precision × recall /precision + recall) × 100

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

Matthews Correlation Coefficient (MCC) = TP × TN − FP × FN/√ (TP + FP) × P × N × (TN + FN)

Finally, the obtained result is statistically validated using z-test 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

Classification	Algorithm		(2019-20)P-dropout				(2019-20)up-dropout				(2019-20)s-dropout			
			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
		Time taken to test model	0 second	0.06	0 second	0 second	0.01 second	0.03	0 second	0 second	0 .03 second	0.02 second	0 second	0 second
		Correlation coefficient	0.9519	0.9519	0.9391	0.8679	0.8858	0.8858	0.3478	0.4587	0.7279	0.7279	-0.1718	-0.604
LAZY	LWL	Mean absolute error	0.7754	0.7754	0.8843	1.0043	0.9513	0.9513	2.1195	2.4933	1.2199	1.2199	1.7676	2.629
		Root mean squared error	1.0323	1.0323	1.1576	1.4588	1.159	1.159	2.6382	2.6982	1.5773	1.5773	2.3286	3.071
		Relative absolute error	31.52%	31.52%	34.52%	42.47%	50.86%	50.86%	110.92%	118.64%	72%	72%	101.19%	170.11%
		Root relative squared error	30.66%	30.66%	32.85%	50.85%	47%	46.97%	104.0618	108.50%	68.84%	68.84%	99.59%	171.68%
		Total Number of Instances	28	28	28	10	28	28	28	10	28	28	28	10
		Ignored Class Unknown Instances	1	1	1	null	1	1	1	null	1	1	1	null
		Time taken	0.02		0.05	0	0.01	0	0.04	0	0	0	0.03	0
		Correlation coefficient	0.9922	0.9922	0.9447	0.8902	0.9904	0.9904	0.4837	0.4004	0.9742	0.9742	-0.1113	-0.492
TREE		Mean absolute error	0.5215	0.5215	1.5645	1.4891	0.8044	0.594	1.5392	1.902	0.606	0.606	1.7486	1.921
	Random forest	Root mean squared error	0.6932	0.6932	2.1382	1.7725	31.76%	0.8044	2.1944	2.309	0.8496	0.8496	2.3555	2.263
		Relative absolute error	21.20%	21.20%	61.07%	62.97%	32.60%	31.76%	80.55%	90.50%	35.57%	35.57%	100.11%	124.32%
		Root relative squared error	20.59%	20.59%	60.67%	61.79%	28	32.60%	86.56%	92.85%	37.08%	37.08%	100.74%	126.49%
		Total Number of Instances	28	28	28	10	1	28	28	10	28	28	28	10
		Ignored Class Unknown Instances	1	1	1	null		1	1	null	1	1	1	null
		Time taken	0	0.02	0.04	0	0	0.01	0.02	0	0	0	0.01	0.02
META		Correlation coefficient	0.9607	0.9607	-0.5341	0	0.9559	0.9559	-0.4461	0	0.9258	0.9258	-0.231	0
		Mean absolute error	0.7893	0.7893	2.5352	2.1199	0.6369	0.6369	1.9022	2.0929	0.6771	0.6771	1.7123	1.495
	Bagging	Root mean squared error	1.136	1.136	3.5224	2.8215	0.8731	0.8731	2.5436	2.4748	1.1592	1.1592	2.3325	1.780
		Relative absolute error	32.09%	32.09%	98.96%	89.64%	34.06%	35.39%	99.55%	99.5894	39.75%	39.75%	98.03%	96.72%
		Root relative squared error	33.74%	33.74%	99.95%	98.36%	28	28	100%	99.5174	50.59%	50.59%	99.76%	99.528
		Total Number of Instances	28	28	28	10	1	1	28	10	28	28	28	10
		Ignored Class Unknown Instances	1	1	1	null			1	null	1	1	1	

Algorithm		(2019-20)P-Total				(2019-20)up-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	0.01	0.03	0	0	0	0.03	0	0
	Correlation coefficient	0.998	0.998	0.4021	0.6249	0.9978	0.9978	0.4094	0.6636	0.9978	0.9978	0.4174	0.692
LWL	Mean absolute error	21911.0069	21911.0069	114298.5835	28360.3592	12269.3743	12269.3743	64026.5426	15612.4633	7565.054	7565.054	40264.6235	9397.8754
	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	15459.4617
	Relative absolute error	12.09%	12.09%	61.29%	21.02%	12.08%	12.08%	61.29%	21.19%	12%	12%	60.42%	20.35%
	Root relative squared error	6.37%	6.37%	95.04%	29.34%	7%	7%	94.79%	31.01%	6.61%	6.61%	94.60%	30.49%
	Total Number of Instances	28	28	28	10	28%	28%	28	10	28	28	28	10
	Ignored Class Unknown Instances	1	1	1	null	1%	1%	1	null	1		1	null
	Time taken	0	0	0.03	0	0.01	0.01	0.05	0	0	0	0.06	0
	Correlation coefficient	0.9995	0.9995	0.0165	0.6256	0.9997	0.9997	0.0243	0.271	0.9997	0.9997	0.0888	0.97
	Mean absolute error	46109.5493	46109.5493	149243.5269	94743.6603	24344.5772	24344.5772	83388.6934	55020.7297	15952.015	15952.015	52005.1127	27742.8899
Random forest	Root mean squared error	139442.726	139442.7263	478473.021	104031.124	77663.35%	77663.35%	267531.727	64113.2619	49582.105	49582.105	169964.8179	30549.2603
	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%	60.07%
	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%	60.24%
	Total Number of Instances	28	28	28	10	28	28	28	10	28	28	28	10
	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	0	0	0	0	0
	Correlation coefficient	0.9938	0.9938	-0.5194	0	0.9935	0.9935	-0.5164	0	0.9935	0.9935	-0.5134	0
	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	56551.1619
Bagging	Root mean squared error	243610.99	243610.9897	496268.179	172663.465	136127.245	136127.245	277409.4748	95949.3983	86736.6317	86736.6317	176754.8264	60307.3569
	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%	122.45%
	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%	118.92%
	Total Number of Instances	28	28	28	10.00%	28	28	28	10	28	28	28	10
	Ignored Class Unknown Instances	1	1	1	null	1	1	1	null	1	1	1	null

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

Classification	Algorithm		(2018-19)P-dropout				(2018-19)up-dropout				(2018-19)s-dropout			
			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
		Time taken to test model	0.01	0.03	0	0	0	0.03	0	0	0	0.02	0.02	0
		Correlation coefficient	0.9448	0.9286	0.9203	0.9041	0.9062	0.9062	0.3711	0.2389	0.8458	0.8458	-0.0707	0.7217
LAZY	LWL	Mean absolute error	1.9025	303838	2.2322	2.1032	1.2741	102741	2.88	2.6558	1.4885	1.4885	3.163	2.082
		Root mean squared error	2.2268	5.0338	2.6526	2.4334	1.5884	1.5884	3.6429	3.3596	2.0096	2.0096	4.563	2.6508
		Relative absolute error	41.03%	79.63%	45.68%	53.40%	48.00%	48.00%	102.86%	102.35%	55%	55%	109.19%	86.71%
		Root relative squared error	32.85%	67.80%	37.22%	45.68%	42.39%	42.39%	93.18%	109.76%	54.13%	54.13%	116.79%	86.65%
		Total Number of Instances	28	28	28	10	28	28	28	10	28	28	28	10
		Ignored Class Unknown Instances	1	1	1		1	1	1		1	1	1	
		Time taken			0.06	0		0.01	0.06	0	0.02	0.02	0.05	0
		Correlation coefficient	0.9954	0.9954	0.9318	0.9334	0.9976	0.9976	0.6576	0.6428	0.985	0.985	0.1115	0.3912
TREE		Mean absolute error	1.1287	1.1287	2.9704	2.4611	0.8421	0.8421	2.2748	2.1371	0.922	0.922	2.6515	1.803
	Random forest	Root mean squared error	1.5141	1.5141	4.1428	2.9279	1.2159	1.2159	3.2054	2.5645	1.3539	1.3539	3.7361	2.5801
		Relative absolute error	24.34%	24.34%	60.79%	62.48%	31.72%	31.72%	81.25%	82.36%	33.86%	33.86%	91.53%	75.09%
		Root relative squared error	22.33%	22.33%	58.12%	54.97%	32.45%	32.45%	81.99%	82.78%	36.47%	36.47%	96.19%	84.37%
		Total Number of Instances	28	28	28	10	28	28	28	10	28	28	28	10
		Ignored Class Unknown Instances	1	1	1		1	1	1		1	1	1	
		Time taken	0	0	0	0	0.01	0.01	0.03	0	0	0	0.02	0
META		Correlation coefficient	0.95553	0.95553	-0.5933	0	0.9648	0.9648	-0.5337	0	0.9665	0.9665	-0.5717	0
		Mean absolute error	1.6284	1.6284	4.8397	3.5722	0.9935	0.9935	2.773	2.5106	1.0565	1.0565	2.9254	2.4175
	Bagging	Root mean squared error	2.5321	2.5321	7.1055	5.2455	1.5098	1.5098	3.9007	2.9905	1.6577	1.6577	3.9157	3.0791
		Relative absolute error	35.12%	35.12%	99.69%	90.69%	37.43%	37.43%	99.04%	96.75%	38.80%	38.80%	100.99%	100.68%
		Root relative squared error	37.35%	37.35%	99.69%	98.44%	40.30%	40.30%	100%	97.70%	44.66%	44.66%	100.82%	100.69%
		Total Number of Instances	28.00%	28.00%	28	10	28	28	28	10	28	28	28	10
		Ignored Class Unknown Instances	1	1	1		1	1	1		1	1	1	

Classification	Algorithm		(2018-19)P-TOTAL				(2018-19)up-TOTAL				(2018-19)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
		Time taken to test model	0.04	0.02	0	0	0.01	0.02		0	0	0.03	0	[illegible]
		Correlation coefficient	0.9967	0.9967	0.5213	0.7232	0.9964	0.9964	0.5344	0.7574	0.8458	0.8454	-0.0707	0.7217
LAZY	LWL	Mean absolute error	28832.9176	28832.918	467692.189	35187.046	17081.4308	17081.431	70371.6748	21196.6809	1.4885	1.4885	3.163	2.082
		Root mean squared error	39248.3041	39248.304	65.39%	47185.9493	23475.3569	23475.357	268655.7103	20133.4009	2.0096	2.0096	4.536	2.6508
		Relative absolute error	16.02%	16.02%	95.24%	26.41%	16.46%	16.46%	65.93%	27.95%	55%	55%	109.19%	86.71%
		Root relative squared error	8.20%	8.20%	28.00%	32.85%	8.60%	8.60%	95.01%	34.10%	54.13%	54.13%	116.79%	86.68%
		Total Number of Instances	28	28	28	10	28	28	28	10	28	28	28	10
		Ignored Class Unknown Instances	1	1	1		1	1	1		1	1	1	
		Time taken	0.01	0.01	0.05	0	0.05	0.02	0.05	0.02	0.01	0.02	0.06	0
		Correlation coefficient	0.9997	0.9997	0.0083	0.7969	0.9997	0.9997	0.0902	0.9034	0.985	0.985	0.1115	0.5912
TREE		Mean absolute error	45048.542	45048.542	147363.495	81027.031	26298.3095	26298.31	82981.1059	43046.9889	0.922	0.922	2.6515	1.803
	Random forest	Root mean squared error	138113.2139	138113.21	475959.218	86062.502	79587.1679	79587.168	273175.263	46656.6368	1.3539	1.3539	3.7361	2.5801
		Relative absolute error	25.03%	25.03%	79.58%	60.81%	25.34%	25.34%	77.74%	56.77%	33.86%	33.86%	91.53%	75.09%
		Root relative squared error	29.14%	29.14%	96.92%	59.91%	29.15%	29.15%	96.61%	56.44%	36.47%	36.47%	96.19%	84.37%
		Total Number of Instances	28	28	28	10	28	28	28	10	28	28	28	10
		Ignored Class Unknown Instances	1	1	1		1	1	1		1	1	1	
		Time taken	0	0	0.01	0	0	0	0.01	0	0	0.01	0.03	[illegible]
META		Correlation coefficient	0.9938	0.9938	-0.5194	0	0.9935	0.9935	-0.5162	0	0.9934	0.9934	-0.5123	0
		Mean absolute error	73033.3317	73033.332	196435.681	162099.8766	41968.1419	41968.144	113063.8748	92720.0224	26121.62	26121.62	70584.1372	55677.5602
	Bagging	Root mean squared error	242195.8603	242195.86	493389.924	170740.7636	139421.8095	139421.81	284108.7713	98246.8055	86983.41	86983.41	177200.6834	59508.6401
		Relative absolute error	40.58%	40.58%	106.08%	121.65%	40.44%	40.44%	105.93%	122.27%	40.30%	40.30%	105.90%	122.44%
		Root relative squared error	51.11%	51.11%	100.47%	118.87%	51.07%	51.07%	100%	119.09%	51.06%	51.06%	100.46%	118.80%
		Total Number of Instances	28.00%	28.00%	28	10	28	28	28	10	28	28	28	10
		Ignored Class Unknown Instances	1	1	1		1	1	1		1	1	1	

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

IJFANS International Journal of Food And Nutritional Sciences

IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

 Journal Volume 11, Iss 11, 2022

Algorithm		(2017-18)P-dropout				(2017-18)up-dropout				(2017-18)s-dropout			
		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
LWL	Time taken to test model	0.03	0.01	0	0	0.01	0.03	0	0	0.01	0.02	0.02	0
	Correlation coefficient	0.7842	0.7842	-0.1137	0.2463	0.7581	0.7581	0.1453	-0.1697	0.8849	0.7581	0.1453	-0.1697
	Mean absolute error	0.543	0.543	1.0079	1.2621	0.9979	0.9979	1.5727	2.3155	1.3869	0.9979	1.5727	2.3155
	Root mean squared error	0.6992	0.6992	1.2315	1.3938	1.3245	1.3245	2.0867	2.8138	1.7171	1.3245	2.0867	2.8138
	Relative absolute error	62.70%	62.70%	108.81%	94.12%	66.26%	66.26%	101.76%	135.21%	53%	66%	101.76%	135.21%
	Root relative squared error	63.83%	63.83%	105.74%	94.69%	66.22%	66.22%	100.61%	131.71%	49.04%	66.22%	100.61%	131.71%
	Total Number of Instances	28	28	28	10	28	28	28	10	28	28	28	10
	Ignored Class Unknown Instances	1	1	1			1	1		1	1	1	
Random forest	Time taken	0.01	0.02	0.09	0	0	0.01	0.02	0	0	0.02	0.06	0
	Correlation coefficient	0.9846	0.9846	-0.1102	0.1581	0.9935	0.9935	0.176	0.5849	0.9932	0.9932	0.248	0.5021
	Mean absolute error	0.2944	0.2944	0.9181	1.2831	0.5651	0.5651	1.5312	1.64	0.956	0.956	2.5671	2.4349
	Root mean squared error	0.3875	0.3875	1.1377	1.4426	0.7562	0.7562	1.9706	2.0932	1.2445	1.2445	3.3969	2.763
	Relative absolute error	34.00%	34.00%	99.11%	95.68%	37.52%	37.52%	99.07%	95.76%	36.22%	36.22%	94.60%	92.07%
	Root relative squared error	35.37%	35.37%	97.69%	98.00%	37.81%	37.81%	95.01%	97.98%	35.54%	35.54%	95.00%	90.41%
	Total Number of Instances	28	28.00%	28	10	28	28	28	10	28	28	28	10
	Ignored Class Unknown Instances	1	1	1		1	1	1		1	1	1	
Bagging	Time taken	0	0	0.02	0	0	0	0.02	0	0	0	0	0
	Correlation coefficient	0.97	0.97	-0.6628	0	0.9715	0.9715	-0.4011	0	0.9411	0.9411	-0.2581	0
	Mean absolute error	0.3159	0.3159	0.9139	1.3257	0.6004	0.6004	1.547	1.9016	0.9443	0.9443	2.7146	3.0317
	Root mean squared error	0.4283	0.4283	1.1457	1.4557	0.8756	0.8756	2.0529	2.3957	1.4773	1.4773	3.5543	3.444
	Relative absolute error	36.47%	36.47%	98.66%	98.86%	39.86%	39.86%	100.09%	111.04%	35.78%	35.78%	100.04%	114.64%
	Root relative squared error	39.10%	39.10%	98.38%	98.89%	43.78%	43.78%	99%	112.14%	42.19%	42.19%	99.40%	112.69%
	Total Number of Instances	28	28	28	10	28	28	28	10	28	28	28	10
	Ignored Class Unknown Instances	1	1	1		1	1	1		1	1	1	

Algorithm		(2017-18)P-TOTAL				(2017-18)up-TOTAL				(2017-18)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
LWL	Time taken to test model	0.01	0.02	0	0	0.02	0.03	0	0	0.02	0.02	0.02	0.02
	Correlation coefficient	0.999	0.999	0.4005	0.8875	0.9988	0.9988	0.3822	0.8827	0.9987	0.9987	0.3781	0.9276
	Mean absolute error	18230.1553	18230.1553	112952.5906	19413.6568	11459.3335	11459.3335	69648.9016	13131.7543	6790.6529	6790.6529	39910.3424	7556.3174
	Root mean squared error	21916.1598	21916.1598	472545.9015	25638.5987	14370.7987	14370.7987	288059.2839	16472.8841	8607.1068	8607.1068	166946.2167	8811.3599
	Relative absolute error	10.03%	10.03%	60.36%	14.43%	10.30%	10.30%	60.84%	16.12%	11%	11%	60.24%	16.86%
	Root relative squared error	4.56%	4.56%	~~94.97%~~	17.73%	4.90%	4.90%	94.76%	18.69%	5.05%	5.05%	94.62%	17.87%
	Total Number of Instances	28	28	28	10	28	28	28	10	28	28	28	10
	Ignored Class Unknown Instances	1	1	1		1	1	1		1	1	1	
Random forest	Time taken	0	0	0.01	0	0	0	0	0	0	0	0	0
	Correlation coefficient	1	1	0.1013	0	1	1	0.2241	0	1	1	0.2268	0
	Mean absolute error	0	0	184557.8027	40327.9412	0	0	98161.1295	26325.5471	0	0	57043.4231	15590.2
	Root mean squared error	0	0	478610.0901	53704.9873	0	0	288394.024	34180.5418	0	0	167486.0158	21401.5122
	Relative absolute error	0.00%	0.00%	98.62%	29.98%	0.00%	0.00%	85.75%	32.32%	0.00%	0.00%	86.10%	34.78%
	Root relative squared error	0.00%	0.00%	96.19%	37.14%	0.00%	0.00%	94.87%	38.79%	0.00%	0.00%	94.93%	43.40%
	Total Number of Instances	28	28	28	10	28	28	28	10	28	28	28	43.3969
	Ignored Class Unknown Instances	1	1	1		1	1	1		1	1	1	10
Bagging	Time taken	0.02	0	0.02	0	0	0	0.02	0	0	0	0.01	0
	Correlation coefficient	0.9938	0.9938	-0.5206	163738.8934	0.9935	0.9935	-0.5162	0	0.9933	0.9933	-0.5101	0
	Mean absolute error	73956.6988	73956.6988	198789.4829	172108.2676	44940.7407	44940.7407	121272.6126	99576.9916	26119.327	26119.3268	70431.2722	54887.2018
	Root mean squared error	245419.3395	245419.3395	499938.2753	121.719	149901.1381	149901.138	305421.744	105111.1317	87058.148	87058.1478	177254.2125	58608.9196
	Relative absolute error	40.67%	40.67%	106.22%	119.03%	40.38%	40.38%	105.94%	122.26%	40.53%	40.53%	106.31%	122.44%
	Root relative squared error	51.12%	51.12%	100.47%	10.00%	51.08%	51.08%	100%	119.27%	51.07%	51.07%	100.47%	118.84%
	Total Number of Instances	28	28	28		28	28	28	10	28	28	28	10
	Ignored Class Unknown Instances	1	1	1		1	1	1		1	1	1	

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

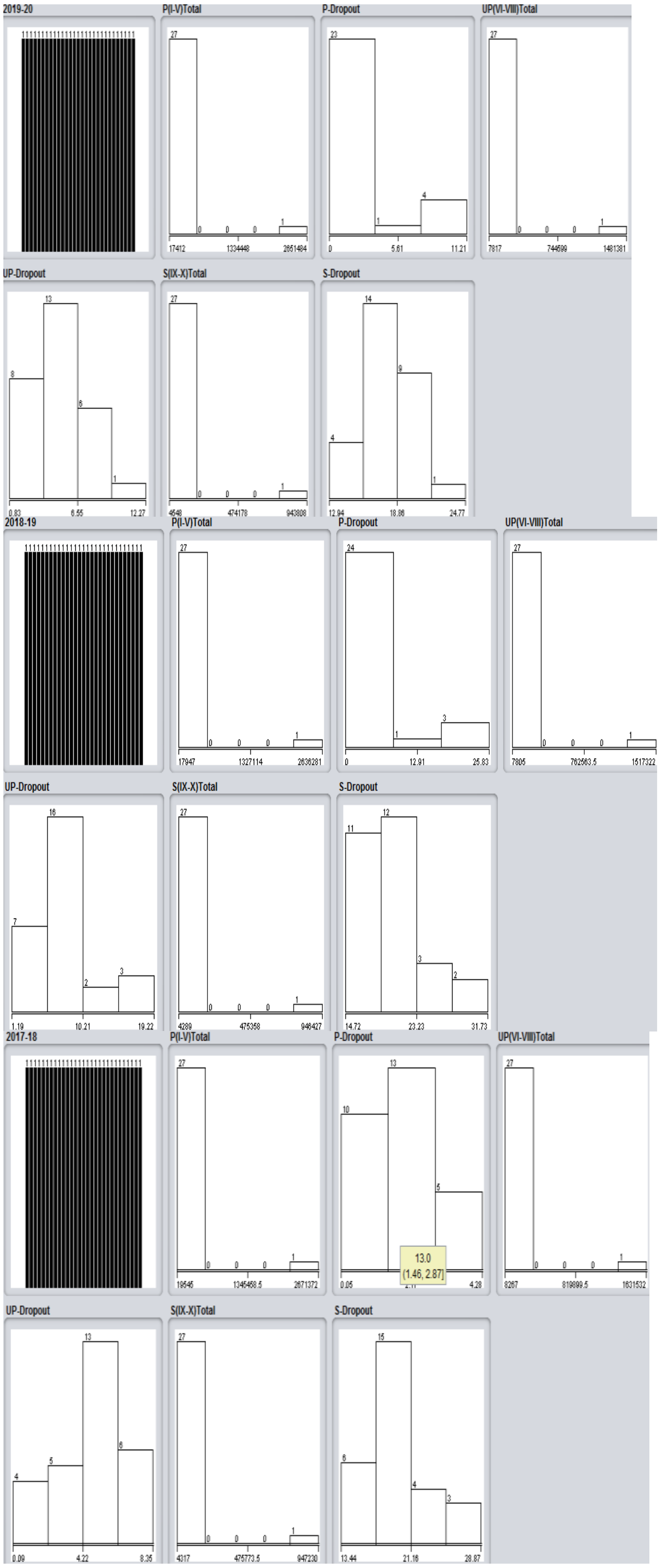

Table 6,7 and 8 show here three years big data classify represented here from 2019-2020, 2018-2019, 2017-2018 and three data mining techniques LAZY, TREE, META present 3 data mining algorithm like LWL, Random forest, Bagging.They process primary total, dropout, upper primary total dropout and secondary 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.

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

[1] 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 K-means 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.