Innovations

Prediction of School Student Performance Using Classification Models

Khushbu Agrawal
Research Scholar
MATS School of Information Technology
MATS University, Raipur (c.g.), India

Dr. Bhavana Narain
Professor
MATS School of Information Technology
MATS University, Raipur c.g.), India

Abstract
The educational performance plays a vital role in classifying the student in higher education. The student performance affects various factors like the learning process, personal and social. This paper demonstrates the impact of student positive or negative performance on student success which is very helpful in the education domain. Here the most commonly used prediction algorithms were LWL, random forest and bagging. After applying these three algorithms we present a novel model Student prediction ratio (SDR). Comparison of both SDR model and giving three classification techniques shows the prediction of student performance and we predict the student dropout rate related to giving dataset, which is collected from questionnaire, Google form and circulated in many schools. For this we also have to collect big and authentic data which can be done through the uniform district information system for education (UDISE). And present we take here five year datasets session 2018-2023. This is helpful for academic progress.
Keywords: *Student Performance, Prediction, School, Student Dropout Ratio (SDR), Classification Algorithms*.

1. Introduction

Educational data mining is a scientific research area, it uses the multiple algorithms 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 used 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 mining 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. In this paper we create a model SDR for student performance prediction purpose and compare this model to three algorithms they are Bagging, Random forest and LWL. We have to try to present a result which technique give better accuracy. We take here five sessions starting 2018 to 2023 ends and take three levels of school student's primary, upper primary or secondary. This paper presents statistical result of our model so that we observe clearly schools conditions or in a future we try to remove problems and make better it for every student, here we take three levels primary level, upper primary level or secondary level. In our paper we are not including higher secondary level because the complete basic development of the student up to the secondary level has been done, which is very important for each and every child. After that they select different subjects and choose their aim. This paper highlights all the information regarding to the school student performance prediction.

2. Materials and Methods for Data Preprocessing

The steps of overall layout of the methods collecting datasets of school student and preprocessed that and converting in clean excel file then convert it into .CSV file format than we apply this file in WEKA platform and select classification algorithm for getting result. This methodology use to process and find the result in two ways, on the basis of comparison of taken three algorithms and create a novel model called Student Dropout Ratio (SDR) model. We have to apply same datasets in SDR model. They all have taken student dataset in five years 2018-2023 from different schools in primary, upper primary and secondary level through circulating Google form and give dropout ratio with level wise like primary level ,upper level and secondary level result and school code also show in our result. For that we have to find the performance of any school student enrollment or structure wise, teachers wise, students over all report and all that. Also show which school academic result is good, best and worst condition according to their dropout ratio.

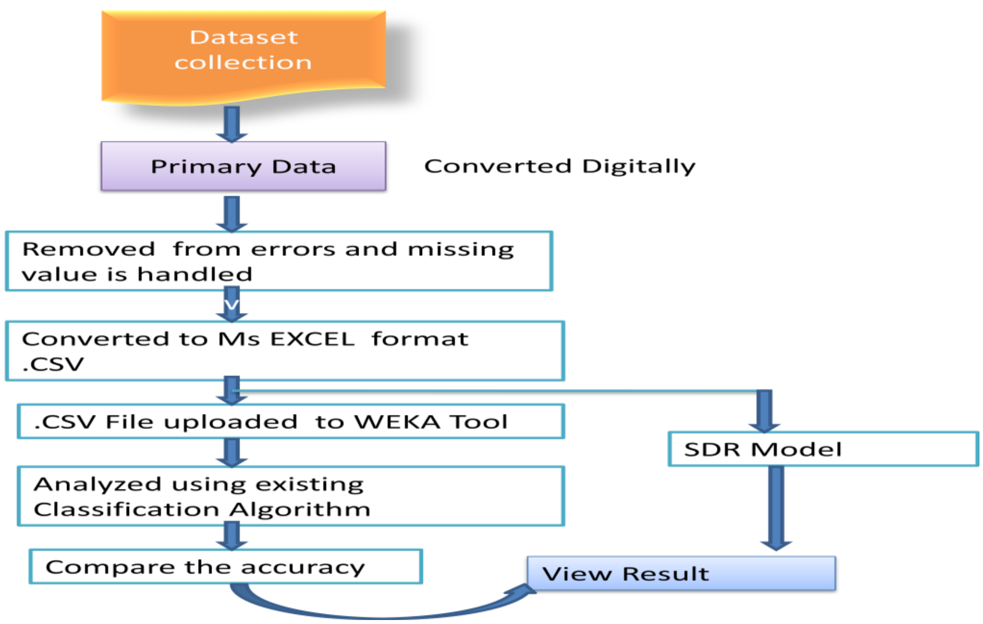


Fig. 1: Steps of collecting information show Preprocessing.

2.1 Data acquisition, augmentation and transformation

Table 1.In this paper we have collected five years sample datasets from questionnaire and Google form circulating in various school. Sample of this data set have school id, student id, student name, gender, Dropout score, and dropout rate. Here dropout score of students is dependent variable. School id is also called UDISE code, which is uniquely provided to each and every school. By this here we have to know about schools

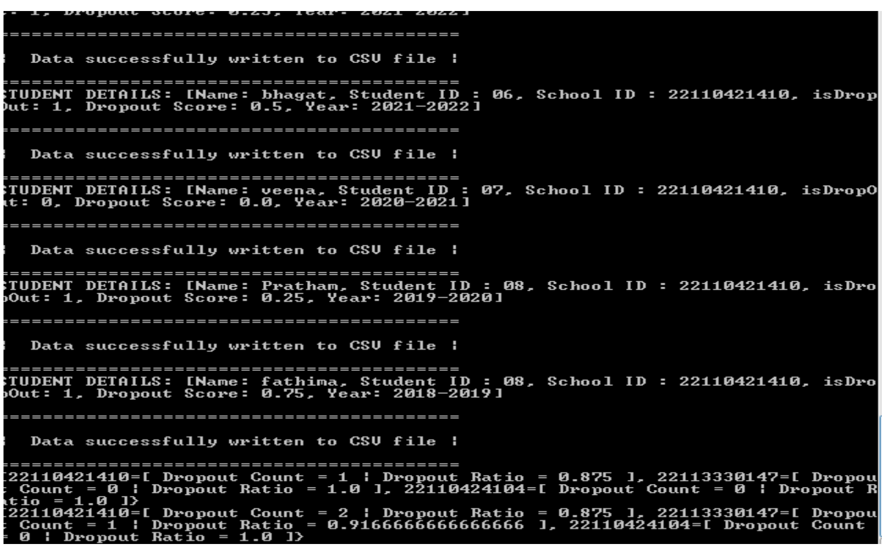
progress report and with the help of this we try to improve school performance. This dataset we use in both classification method and SDR model in weka platform after that comparing and generated result in both methods and present after analysis which model or classification give better accuracy. In the following fig show dataset, this is use in our paper.

SID	SchoolId	Name	Sex	Group	Dropout	Year	DropoutScore
11	22110424104	Divya	0	2	0	2022-2023	0
12	22110424104	payal	0	1	0	2022-2023	0
13	22110424104	aakash	1	2	0	2022-2023	0
14	22110424104	neha	0	0	0	2021-2022	0
16	22110424104	deena	0	2	0	2019-2020	0
17	22110424104	Ramsigh	1	2	1	2020-2021	1
17	22110424104	priya	0	2	1	2020-2021	1
18	22110424104	meet patel	1	2	0	2020-2021	0
19	22110424104	payal	0	1	0	2020-2021	0
20	22110424104	vasu verma	1	1	1	2019-2020	1
101	22113330147	ishita	1	0	1	2022-2023	2
102	22113330147	Pramod	1	1	1	2022-2023	2
103	22113330147	Gajendra	1	2	0	2022-2023	0
104	22113330147	mahak	0	2	1	2021-2022	1
105	22113330147	mansi	0	2	0	2021-2022	1
106	22113330147	hashim	1	0	1	2020-2021	2
2	22110421410	Mihir	1	0	0	2022-2023	0
3	22110421410	nilay	1	1	1	2022-2023	0
4	22110421410	Seeta	0	0	0	2022-2023	0
5	22110421410	Diva	0	2	1	2021-2022	1
6	22110421410	bhagat	1	2	1	2021-2022	2
7	22110421410	veena	0	2	0	2020-2021	0
8	22110421410	Pratham	1	2	1	2019-2020	1
8	22110421410	fathima	0	1	1	2018-2019	3

Fig. 2: Datasets collected by SDR model.

2.2Feature extraction and feature selection

In this Fig 3: we can see our model SDR generated report for school performance prediction in different session wise. Here student details student name, student id, school UDISE ID, year and student dropout cause is filing here under four cause like financial ,family, academic, admin .if student choose one cause then the dropout score shown. After applying these details we get desired output which is show in last three –four lines. 22110424104, 22110421410 and 22113330147 present three different school codes. This is unique UDISE ID we can access secondary datasets from this portal. This last three line show student dropout ratio from which school in which section the improvement of schools conditions.



```
========================================
  Data successfully written to CSV file !
========================================
TUDENT DETAILS: [Name: bhagat, Student ID : 06, School ID : 22110421410, isDrop
Out: 1, Dropout Score: 0.5, Year: 2021-2022]
========================================
  Data successfully written to CSV file !
========================================
TUDENT DETAILS: [Name: veena, Student ID : 07, School ID : 22110421410, isDropO
t: 0, Dropout Score: 0.0, Year: 2020-2021]
========================================
  Data successfully written to CSV file !
========================================
TUDENT DETAILS: [Name: Pratham, Student ID : 08, School ID : 22110421410, isDro
Out: 1, Dropout Score: 0.25, Year: 2019-2020]
========================================
  Data successfully written to CSV file !
========================================
TUDENT DETAILS: [Name: fathima, Student ID : 08, School ID : 22110421410, isDro
Out: 1, Dropout Score: 0.75, Year: 2018-2019]
========================================
  Data successfully written to CSV file !
========================================
22110421410=[ Dropout Count = 1 | Dropout Ratio = 0.875 ], 22113330147=[ Dropou
 Count = 0 | Dropout Ratio = 1.0 ], 22110424104=[ Dropout Count = 0 | Dropout R
tio = 1.0 ]}
22110421410=[ Dropout Count = 2 | Dropout Ratio = 0.875 ], 22113330147=[ Dropou
 Count = 1 | Dropout Ratio = 0.9166666666666666 ], 22110424104=[ Dropout Count
= 0 | Dropout Ratio = 1.0 ]}
```

Fig 3: Statically report generated by our model.

Table 1: show the overall statistic value in school ID showing by Weka tool.

Name :School Id Type: Numeric
Missing: 0(0%) District:4 Unique: 1(4%)

Statistic	Value
Minimum	22110421410
Maximum	22113330147
Mean	22111121937
StdDev	1266512.297

Table 2: show the overall statistic value in Dropout score and Dropout generated by weka tool.

Name :Dropout score Type: Numeric
Missing: 0(0%) District:4 Unique: 1(4%)

Statistic	Value
Minimum	0
Maximum	3
Mean	0.72
StdDev	0.891

2.3 Hypothesis of our Work

H_1: Is there no correlation between enrolment, number of teachers and location of school with overall enrolment?

H_2: is there no relation between passing rate and dropout rate?

H_3: Are data gaps deteriorating data quality?

H_4: Are private schools imparting poor quality of education than government schools?

H_5: Did the model successfully classify the U-DISE data or not, in terms of location wise enrolment and learning performance?

3. Models and Tools used for Classification

In this paper we have to create a novel model Student Dropout Model (SDR) for accessing better accuracy and use weka tool. Here we use three data mining classification algorithm such as Bagging, Random forest and LWL.

3.1 Model(Student Dropout Ratio)

Student Dropout Ratio (SDR) model developed in this paper for student performance prediction. We categories SDR model in two p phases - (1) School Panel and phase is (2) Student panel and submit all information regarding this student panel help to identify the school which is give poor performance by session wise. Student fills all information like their id number; level means he/she gets the drop in primary or upper primary and secondary level on that session. They also fill their gender and most important part of this SDR model they choose the reason behind that and select the reason after that school Id also available in this model which is very close to our research work. This school ID is called UDISE code stands for Unified District Information System for Education, which is uniquely available and all the schools are provided through government which is shows the uniqueness of that school. By this code we know how many student drops on that session. Here we take five years dataset and five sessions for finding student performance year 2022-2023, 2021-2022, 2020-2021, 2019-2020, and 2018-2019. This model surely helps to the improvement of every academic progress and success.

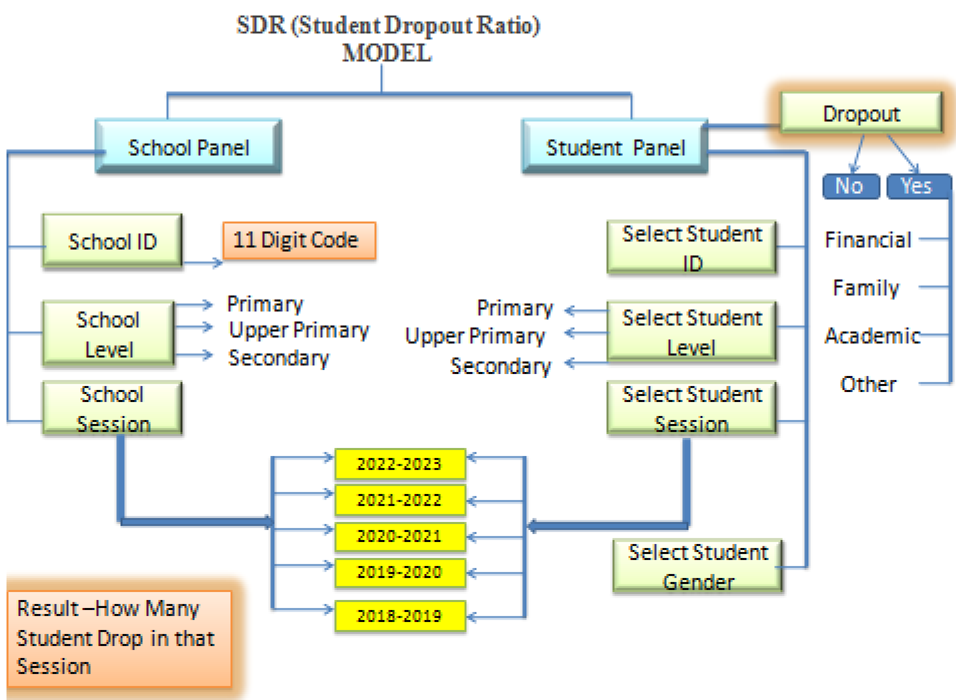


Fig. 4: working process of SDR Model.

3.2 Tools used in this research

We use Weka tool in this paper. It is use for data mining techniques and it is machine learning algorithm developed by the University of Waikato in New Zealand. The data file use in weka in ARFF file format (.CSV format).

3.3 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.

A. 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.

B. Random Forest

Random forest is 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 creates the random forest by combining N decision tree and second phase is to make predictions for each tree created in the phase.

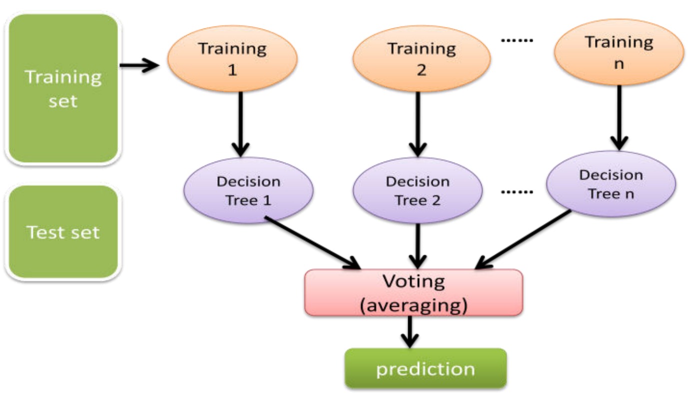


Fig 5. The above diagram explains the working of Random forest.

C. 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.

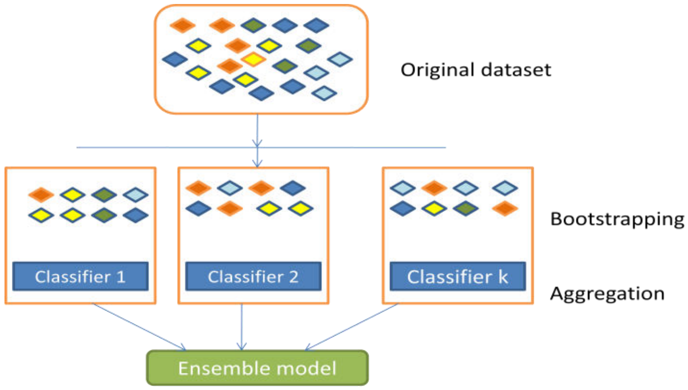


Fig 6.This diagram presents the Bagging process.

3.4 Performance evaluation metrics

(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

The following figure 7 present here sequential model have different abbreviations: LWL, Random Forest, Bagging, SDR model which is called training supervised model. It has two phases as phase-I and phase-II. Phase-I take three attribute primary, upper primary and secondary level with classification algorithm and phase –II define SDR model to evaluate the performance of the student or check schools behavior . Here we have to calculate performance evolution measure.

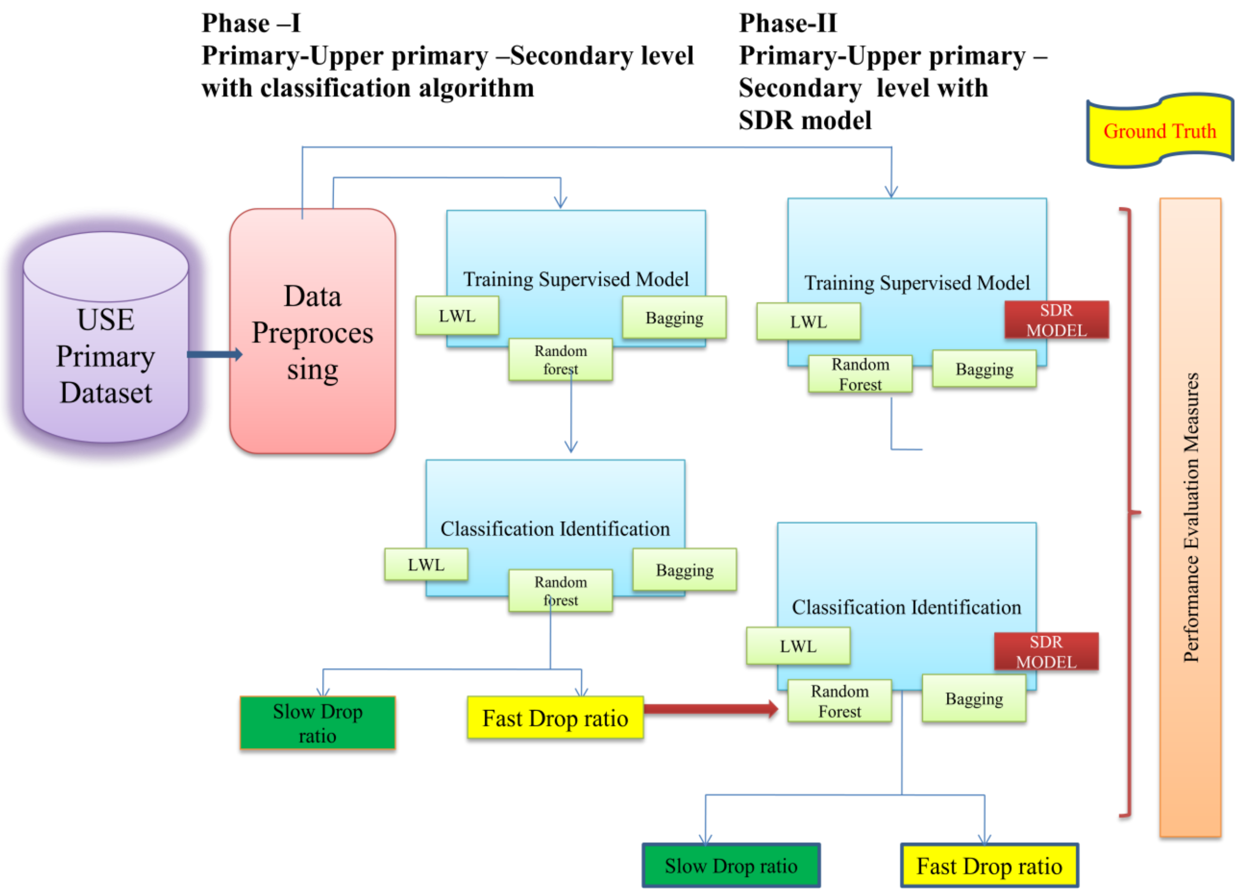


Fig 7.The Prototype of the proposed automatic sequential Model (Abbreviations: LWL, Random Forest, Bagging, SDR Model)

3. Experimental results and discussion

Experiment 1 — (Hypothesis 1).
Experiment 2 — (Hypothesis 2).
Experiment 3 — (Hypothesis 3).
Experiment 4 — (Hypothesis 4).
Experiment 5 — (Hypothesis 4).

In the following fig number 8 represented here form for collection of information regarding the different-different schools of different students. Because we present real or authentic schools or students dataset from Google form or questionnaire which is play a very important part of our related research work and we get desire output for student performance prediction. Every year many students take a drop and discontinue their study so this is our work to motivate that student for their study and try to know about their problem or solve it. So these SDR models help each and every school also session wise and their self assessment for their improvement. It is developed by java user defined package.

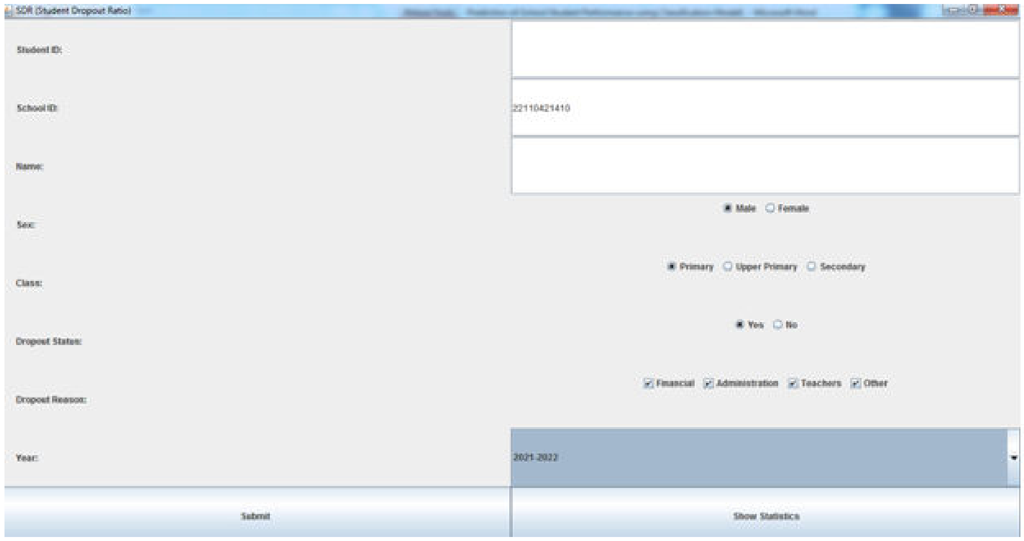


Fig: 8. Representation of SDR Model's form.

In the following fig number 9 Representing here statistics form. It is generating after fill all the school or students information or dataset, we submit information than this form is shown the top of the screen and with the help of this we generate our report after clicking the filter button for calculating dropout ratio. This model is surely helpful for all academic progress. In this form student details is mention here by gender wise, year wise.

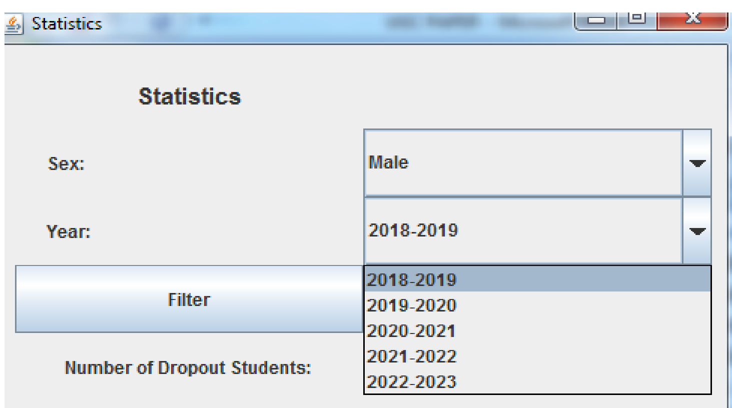


Fig: 9. Representation of Second SDR from for statistics.

For both the phases (Phase-I and phase-II),

2. Experimental results and discussion

Here we collected the datasets in three levels primary, upper primary and secondary level and calculate our datasets is dependent in dropout ratio. Here we find how many student take drop in given session with this we also find that which school give good or worst performance.

Table 4. Output generated by weka tool use in primary data set.

(Out of 8 attributes -student id, school id, name, sex, dropout, group, year, dropout score, we select dropout score attribute run in weka)

Meta		**Use tranning set**	**Supplied test set**	**Cross-validation**	**Percentage split**
	Tme taken to be test model	0	0	0	0.01
	Tme take to be built model	0.04	0.04	0.03	0.01
	Correlation coefficient	0.9386	0.9386	0.2271	0
Bagging	Mean absolute error	0.301	0.301	0.776	0.9721
	Root mean squared error	0.3933	0.3933	0.9206	1.1717
	Relative absolute error	0.4019	0.4019	0.9880	0.9793
	Root relative squared error	0.4507	0.4507	90.985	0.9683
	Total Number of Instances	25	25	25	8

Tree		**Use tranning set**	**Supplied test set**	**Cross-validation**	**Percentage split**
	Tme taken to be test model	0.01	0	0	0.01
	Tme take to be built model	0.09	0.04	0.04	0.04
	Correlation coefficient	0.9892	0.9892	0.5194	0.4541
Random	Mean absolute error	0.1991	0.1991	0.5809	0.9273
Forest	Root mean squared error	0.259	0.259	0.774	1.1824
	Relative absolute error	0.2659	0.2659	0.7396	0.9341
	Root relative squared error	0.2968	0.2968	0.8395	0.9772
	Total Number of Instances	25	25	25	8

Lazy		**Use tranning set**	**Supplied test set**	**Cross-validation**	**Percentage split**
	Tme taken to be test model	0.03	0	0	0
	Tme take to be built model	0	0	0	0
	Correlation coefficient	0.9098	0.9098	0.6703	0.6084
LWL	Mean absolute error	0.2345	0.2345	0.4239	0.7402
	Root mean squared error	0.3633	0.3633	0.6521	1.102
	Relative absolute error	0.3132	0.3132	0.5398	0.7456
	Root relative squared error	0.4163	0.4163	0.7073	0.9108
	Total Number of Instances	25	25	25	8

Table 5 and Table 6 present three years' big data classify represented here year 2019-2020, 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 options 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.

Lastly the following fig number 10 representing here our model report. If dropout yes than count 1 or if no select by student than it is count 0 condition. School id 1234567835456 is giving firstly their dropout is one and last 12345646 school id their dropout is 2 there are the different outputs presenting here.

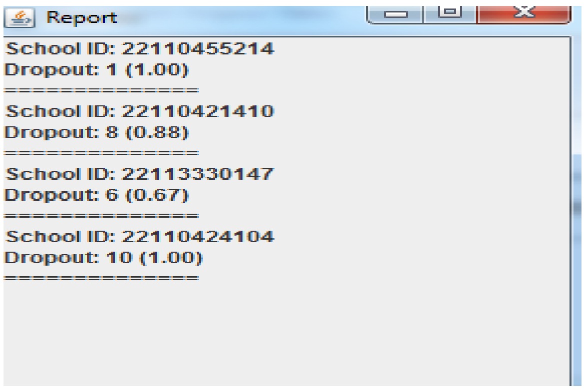
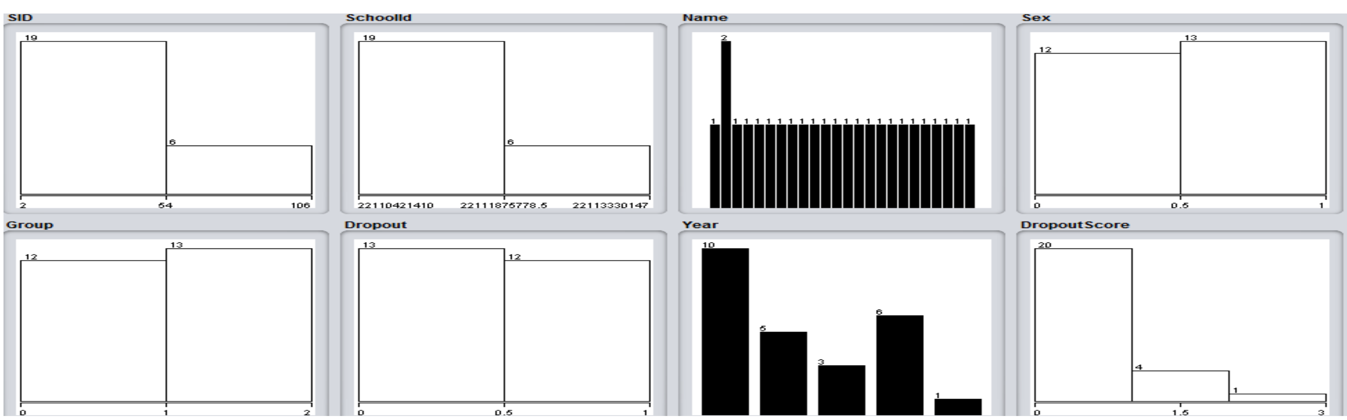


Fig: 10.This is Report generated by SDR model.

3.1. Statistical analysis

Result based on Statistical Parameters of the Year 2018 – 2023 for primary, upper primary or secondary attribute.



Result based on Statistical Parameters of the Year 2019 – 2020 for primary total, upper primary total or secondary total and primary dropout, upper primary drop out and secondary dropout.

3.2. Discussion

In this section we have done comparative study of all three algorithms in two phases. In I phase accuracy of our model with respect to primary dropout and secondary dropout and phase II accuracy of our model with respect to upper primary total and secondary total.

Table 7

Classifier	Phase-I				Phase-II			
	Accuracy		Accuracy		Accuracy		Accuracy	
	P-Dropout	S-Dropout	P-Total	S-Total	UP-Dropout	S-Dropout	UP-Total	S-Total
LWL	33%	49%	67%	56%	30%	40%	60.23%	76.21%
Random forest	28.09%	46%	70.08%	55.00%	29.15%	39%	71%	79.67%
Bagging	20%	19%	72.78%	87%	20.21%	19%	88.61%	89%
SDR model	29.19%	49%	72.80%	87.12%	22.02%	20%	89%	89.90%

Table8. Average ranking of classifiers based on different classification performance we sees bagging gave the best result compared to other algorithm.

Classification Algorithm	Average ranking of classification algorithms			
	Phase-I Dropout	Phase-I Total	Phase-II Dropout	Phase-II Total
LWL	33.245%	69.8%	30.2%	60.61105%
Random forest	28.32%	70.355%	29.345%	71.3983%
Bagging	20.095%	72.215%	20.305%	89.055%
SDR model	22.98%	72.80%	22.02%	89.90%

4. Conclusion

Educational data mining plays 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. In this paper we use primary dataset collected through Google from and apply this data set in LWL, Random Forest and Bagging and student dropout ratio. It can be concluded that after comparing SDR model with classification techniques, Student Dropout Ratio (SDR) provide better accuracy compared to other approach. In future work we will take more datasets and more classification algorithms and try to present a new result with level wise such as primary, upper primary and secondary level.

Acknowledgement

I am thankful to our management chacellar Shri Raj Pagariya ji and Vice Chancellor K.P.Yadav sir ji and DG.Shri Priyesh pagariya ji for providing as research incubation lab in MATS University. I am also thankful to my guide Dr. Bhavana Narain mam and my family members.

References

1. *Kannan K.K., Abarna K.T. M., Vairachilai S.(2023) .Open Issues, Research Challenges, and Survey on Education Sector in India and Exploring Machine Learning Algorithm to Mitigate These Challenges.International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 11 Issue:*
2. *Sing and Dr.Singh Sodhi R (2023).Predicting school students' academic performance using data mining classification algorithm. Eur. Chem. Bull. (Special issue 12) 876-900 .*
3. *Ramrao Yadav N. and Deshmukh S.S.(2022).Prediction of Student Performance Using Machine Learning Techniques: A Review. ICAMIDA ACSR 105, pp. 735–741.*
4. *Rodríguez P., Villanueva A., Dombrovskaia L., Valenzuela P.J. (2022).A methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems: the case of Chile. Education and Information Technologies.*
5. *Samy Selim1 K., Saeed Rezk S.(2023).On predicting school dropouts in Egypt: A machine learning approach. Education and Information Technologies 28:9235–9266.*
6. *Teoh CW., Sin-Ban Ho, Dollmat K.S, and Hong C (2022). Ensemble-Learning Techniques for Predicting Student Performance on Video-Based Learning. International Journal of Information and Education Technology, Vol. 12.*
7. *Ugale D., Pawar J.,Yadav S., Dr. Raut C(2020). Student Performance Prediction Using Data Mining Techniques. International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 p-ISSN: 2395-0072*
8. *Yağc M.(2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. Yağcı Smart Learning Environments.*
9. *Leena H. Alamri , Almuslim S.S., S. Alotibi M., Dana K. Alkadi , Khan I.U., Aslam N(2020). Predicting Student Academic Performance using Support Vector Machine and Random Forest. ICETM 2020 London, United Kingdom.*
10. *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).*
11. *Narain B, Zadgaonkar AS, Kumar S (2013) Impact of digital image processing on research and education. Natl Semin Work.*
12. *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.*
13. *More SS, Narain B, Jadahv BT (2018) Data encryption standard algorithm in multimodal biometric image . Int. J. Comp. Sci. Eng,.*
14. *U Sharma, B Narain, V Nohria (2021). Hybrid Support Vector Machine and Distance Classifier in Breast Tumor Detection . SPAST Abstracts.*
15. *P Singh, B Narain (2021). Student Satisfaction in Educational Organization using Machine Learning, - NOVYI MIR.*
16. *B Narain, AS Zadgaonkar, S Kumar (2013). Impact of digital image processing on research and education. Natl Semin Work.*
17. *M Nayak, B Narain (2021).The Online Retail Market Analysis for Social Development with Machine Learning . SPAST Abstracts.*
18. *B Narain (2021). On-Line Big Data Analysis using K-Mean and Modified K-Mean Algorithm with Machine Learning Techniques.*
19. *M Nayak, B Narain (2020). Big Data Mining Algorithms for Predicting Dynamic Product Price by Online Analysis. Computational Intelligence in Data Mining.*

20. *Agrawal k.,Nariyan B.(2022). A Literature Review on School student Interest using Datamining Algorithm. Explore Journul of research Peer Reviewed Journal, ISSN 2278-0297(print)ISSN 2278-6414(online) Patna womens college Patna India.*
21. *Agrawal k.,Nariyan B(2023). A Bagging Arthimetic Approch for Prediction of School student Performance.Publisher-,IEEE Xplore:.*