PERFORMANCE CRITIQUE FROM THE PERSPECTIVE OF THE RURAL AREA USING DATA MINING ALGORITHMS

Amandeep Kaur
Assistant Professor, Computer Science and Engineering, Guru Nanak Dev University, RC Kapurthala,Punjab,India
aman.cheema.2k12@gmail.com
ORCID:https://orcid.org/0000-0002-1947-0104

Dr.Karanjeet Singh Kahlon
Professor, Computer Science, Guru Nanak Dev University, Amritsar, Punjab, India
karankahlon@gndu.ac.in
ORCID:https://orcid.org/0000-0002-7897-1451

ABSTRACT

Advanced education has accomplished the most extreme need as it assumes a basic part of the socio-economic development of the nation. The objective of the administration is to deliver an instructive domain with Knowledge, Skills, Desired Values, Innovations. The measures are adopted to expand the number of educational foundations with the goal that no youngster is deserted in gaining instruction. But, there still exists a provincial divergence in the Indian Education System. A gigantic hole has been made in metropolitan and provincial region instructive foundations. The state governments are attempting their level best to minimize the gap by opening the number of instructive foundations of the most extreme quality in a rustic region so training ought to be uninhibitedly accessible at the doorsteps. This paper manages to take a contextual analysis of a rustic zone instructive establishment that assesses the exhibition of the understudies in a graduation course and targets perceiving the most persuasive credits that influence the performance. It additionally examines the explanations behind the dropouts and discovers the timeline when the dropouts are greatest. This is accomplished by applying various classification algorithms to the dataset. It establishes that the Multi-Layer Perceptron Model outstands all by accomplishing 100% precision and Kappa Statistics estimation of 1.

INTRODUCTION

Basic education in India manages the four essential issues of quality, access, equity, excellence.Whereas, Access and Equality are the difficulties to be gone through if higher education has to be uplifted. To address these issues it gets important to open organizations for advanced education in educationally restricted regions. Taking the scenario of Punjab, after its revamping in 1966, the legislature has made very much arranged procedures to create and extend educational facilities. These timely endeavors have prompted the opening of numerous colleges, universities, and schools territory-wise. As a significant activity to support training, different colleges have likewise evolved local grounds in rustic territories with the goal that education can be reached at the doorsteps of the students living in faraway spots. In recent years industries have emerged in Punjab and are the main source for providing job opportunities to both technical and non-technical manpower. To close the gap, the technical education and industrial training systems have been expanded, modernized, and reoriented.

Taking a case study of a rural area regional campus developed by the state government university near Sultanpur Lodhi. This university campus has been set up in 2014 for providing educational facilities to the economically and educationally poor sections of the society. Most of the population belonging to this area is Scheduled Castes (SC). The governments have put up many efforts to help these weaker sections by proposing various welfare schemes for them like Post Matric Scholarship Schemes for SC/ST in which no fee is charged from this section and even book banks free of cost and other facilities are provided. This study deals insight into why the university regional campuses are not doing fairly good even though these campuses are providing high-quality infrastructure, well-qualified faculty coming through selection criteria, well-equipped laboratories housing research and project labs separately, concessions and scholarships given to the needy and brilliant students, welfare schemes for SC Students and minority class, different departments for youth welfare and sports. This paper considers students enrolled in one of the courses offered on this rural area campus, and uses data mining techniques to forecast student success, analyze dropout rates, and compare school-level performance to college-level performance in examinations. And attempting to deduce the reasons for it.

This paper addresses the following research questions:
R1: How accurately the performance of the students in a graduation course in rural areas can be measured?
R2: To analyze the most influential attributes that affect the performance of the students.

R3: To evaluate the time duration when the dropout rates are at a peak level.

WEKA 3.8.1, an open-source data mining tool, was used to answer these research questions. The data sets for the students enrolled in a 3-year graduation course have been taken as input. The data sets consist of three years. The input data is firstly pre-processed by applying various filters and then data mining algorithms are tested like Naïve Bayes, J48, ZeroR, MLP, Simple CART. For performance enhancement, various ensemble algorithms as Bagging, Voting, and Stacking are used. The accuracy, Kappa statistics, Mean Absolute Error, and Relative Mean Squared Error, among other parameters, are used to test the results.

The following is a breakdown of the paper's structure:
Section II provides an overview of relevant educational data mining research.
Section III describes the research analysis, as well as the technique used and the output parameters considered.
Section IV analyses the results from the experimental setup.
Section V gives the conclusions and future work.

LITERATURE SURVEY

Ibrahim et al. use the Cumulative Grade Point Average to measure academic success using three classification models: Artificial Neural Networks (ANN), Decision Trees, and Linear Regression Model (CGPA). It shows that ANN has the highest level of accuracy, (Ibrahim & Rusli, 2007)

Quadri et.al. finds the dropout rate of the students by choosing the decision trees technique which performs the best. (Quadri & Kalyankar, 2010).

For measuring academic results, Ali et al. use both dependent variables (like grades) and independent variables (like age, economic status, education, and so on). Three models were used: Linear Regression Model, Correlation Analysis, and Descriptive Analysis, with the linear model proving to be the most efficient(Ali et al., 2013).

Radaideh et.al. uses a decision tree technique to find which attributes affect the most while measuring academic performance. (Andrew Braunstein, Michael McGrath, 2015)

Affendey et.al. uses Naive Bayes, AODE, and RBF Network for ranking of the courses that contribute to academic performance. (Affendey et al., 2010)

Baradwaj et.al. uses the Decision Tree Technique for predicting student performance. (Kumar & Pal, 2011)

In a Credit Based Continuous Evaluation System, Kaur et. al. use boosting algorithms to improve the performance of classification algorithms for early estimation of the student's marks in Major Tests (CBCES)(Kaur & Kaur, 2016)

EIGamal et.al. uses the decision tree technique for identifying the variables that predict student programming performance. (F.ElGamal, 2013)

According to Asif et al., graduation performance can be predicted using pre-university marks and 1st and 2nd-year marks without taking into account any socio-economic or demographic factors, and the results show that Naive Bayes performs the highest. (Asif et al., 2014)

Kovacic et al. use classification trees to work out how enrolment data will help determine who will succeed and who will fail. With a 60.5 percent classification score, the Classification and Regression Tree Model provided the best prediction. (J. Kovacic, 2010)

For the first year of study, Oancea et al. use neural networks to predict students' grades based on their grade point average. (Oancea et al., 2017)

Hardre et.al uses AMOS 4.0 to investigate the predictive relationship among student characteristics that influence motivation for learning and achievement. (Hardré et al., 2007)

Daud et.al. uses Bayes Network, Naive Bayes, C4.5, and Cart to measure academic performance by taking student personal information and family expenditure as attributes. (Daud et al., 2019)

For predicting students' academic risk, Vandamme et al. use Discriminant Analysis, Neural Networks, and Decision Trees. Discriminant Analysis produced the best results, with an overall classification rate of 57.35 percent. (Vandamme et al., 2007)

DATA AND METHODOLOGY

A. Data

The data that is used in the study comprises the dataset of a regional campus of a state-level university located in a rural area. A 3 -year data totaling 207 students have been taken. The variables in the data set (table 1) are associated with students' pre-admission marks (which are used to classify students for university admission) and the scores for all of the courses taught during the three years of the degree course. The degree course chosen for study is Bachelor of Computer Applications (BCA) which is a 3- year regular course. The data set is taken from a rural area campus, so it contains 75% data from the rural area and 25% from urban areas. The total marks are used as a performance indicator for testing the students' performance at the end of the degree program. The result attribute has two values 0 for fail and 1 for the pass. The value 0 is put in the data set when the student has left the course or has failed in the course, otherwise the value 1 for the pass when the student has completed the course and has cleared all the semesters. The base class taken is the performance, which is a nominal class and takes three values {Good, Bad or Average} depending on the attribute total marks attained by the student.

TABLE 1: DATA SET FOR STUDY

S no.	Attributes	Description
1	RNO	Roll number of students
2	NM	Name of the students
3	DOB	Date of Birth
4	GEN	Gender of the Student
5	FN	Father's name of the student
6	MN	Mother's Name of the student
7	FO	Father's occupation
8	MO	Mother's occupation
9	AI	Annual Income of the student
10	CAT	Category of students
11	REL	The Religion of the student
12	ADR	Address of the student
13	R10	Student 10th Result in percentage
14	R12	Student 12th result in percentage
15	STR	Subject opted by the student at school level like medical,non-medical, Commerce, Arts.
16	ARE	The Living area of student i.e. Urban or rural
17	PN	Phone no of the student
18	R1 -R6	1stSem -6th Sem result of BCA students
19	TOT	Total Marks of all semesters of a student
20	RES	Result show student result pass or fail
21	PER	Performance of student as good, average or bad

A data set of three years has been taken for the study. The year and no. of admissions done in each year are presented in Table II:

TABLE II. YEAR-WISE INTAKE OF STUDENTS

SNo	Year	No. of Students
1	Year1(2015)	90
2	Year 2(2016)	70
3	Year 3(2017)	47

B. METHODOLOGY

1) PROCEDURES

WEKA provides several classifier algorithms. which aims to classified the data according to the defined pattern and behavior of the data.

a) NaiveBayes: It is based on the Bayes theorem. It's a set of classifiers algorithms that make use of estimator classes. It assumes that the input values are nominal, although numerical inputs are supported by assuming a distribution. While numerical inputs are provided by assuming a distribution, it assumes that the input values are nominal. Bayes theorem find the probability from the results

P(A/B)= P(B/A)P(A)/P(B)

In the above equation, the probability for A is to be evaluated when the value of B is given.

It uses the kernel density estimators, which improves performance if the normality assumption is grossly incorrect; it can also handle numeric attributes using supervised discretization. The data is divided into two parts. Features matrix which defines the features and attributes and the Response matrix which defines the prediction or output.

b) MLP**:**

It is a classifier that classifies instances in a dataset using back-propagation. It is made up of a large number of neurons that are linked in a pattern. Neurons are divided into three categories: Input Neurons that receive and process information. Hidden Neurons, where the actual processing is performed by neurons, and the output neurons are the ones that generate the results after they've been processed. (Kaur & Kaur, 2016)(Chakraborty et al., 2020)

c) ZeroR **:**

It's the most basic classifier since it only refers to the target and ignores the prediction. The majority class is predicted by this classifier. It is useful for establishing baseline performance as a comparison point for other classification methods.

d) J48:

It is a successor of C4.5.It is developed by Ross Quanlan. It uses a greedy and top-down approach for decision making the dataset is partitioned into smaller partitions that use a recursive divide and conquer strategy. The partition of the dataset uses heuristics that choose the best partition on the dataset..(Bashir & Chachoo, 2017)

e) SimpleCART: It is known as a classification and Regression Tree. It generates a binary decision tree. The best splitting attribute is chosen from Entropy. It uses a learning sample with pre-assigned classes for all the observations for building decision trees. It gives the result as a classification or regression tree depending on the input data. By cross-validation, it selects the best tree from the sequence of trees in the pruning process. This algorithm uses a greedy algorithm and selects the best feature at each stage of the process. When implementing, the dataset is split into two subgroups, that are most different in outcome. This procedure is continued on each sub-grouping until the minimum subgroup size is reached. (Kalmegh, 2015).

f) REPTree: It is known as Reduced Error Pruning Tree. It is a fast decision tree learner who builds a tree using information gain as the splitting criteria and prunes it using reduced error pruning. It sorts numeric attributes only. It uses regression tree logic and creates multiple trees in different computations. Then it selects the best one from the generated trees(Kalmegh, 2015).

Performance Metrics

Different performance metrics for comparing different classification algorithms are elaborated below:

a) Kappa statistic*:*

Sometimes accuracy cannot be used as a measure for evaluating the performance of the unbalanced set. Then an important measure to be taken is Kappa statistics. It is an analog of the correlation coefficient. If the value is zero it means a lack of correlation and the value 1means a high correlation between class labels and attributes. It compares the observed accuracy with the expected accuracy.

To calculate Observed Accuracy**,** add the number of instances that the machine learning classifier agreed with the ground truth label, and divide by the total number of instances.

The Expected Accuracy is directly related to the number of instances of each class along with the number of instances that the machine learning classifier agreed with the ground truth label.

The formulae used for calculating Kappa statistics is shown in equation (1)

Kappa Statistics = (observed accuracy - expected accuracy) / (1 - expected accuracy) (1)

b) Mean absolute error*:*

It calculates the average loss in the data set. The formulae to calculate is shown in equation(2)

MAE=$\boldsymbol{1/n\sum_{i=1}^{n}|xi-x|}$ (2)

Here x_i is a prediction value and x is a true value

c) *Root mean squared error:*

It calculates the difference between the predicted value and the actual observed value. It is a good measure of accuracy, but only to compare forecasting errors of different models for a particular variable and not between variables, as it is scale-dependent. It is also called the root-mean-square deviation, RMSD. shown in equation (3)

RMSE=$\sqrt{1/n\sum_{j=i}^{n}(y_j-\hat{y_j})^2}$ (3)

d) *Accuracy:*

The accuracy is defined as how well a given predictor can guess the value of the predicted attribute for new data as in equation(4)

Accuracy= number of sample predicted correctly/total number of samples (4)

e) *TP rate :*

It is a true positive rate which means Correctly classified are positive. given in equation (5)

TPR= TP/(TP+FN) (5)

f) *FP rate***:**

It is a false positive rate which means false classified is positive given in equation (6)

FPR=FP/(FP+TN) (6)

Or

FPR= 1-TNR

g) *TN rate***:**

It is a true negative rate, which means correctly classified as wrong. given in equation (7)

TNR=TN/(FP+TN) (7)

h) *FN rate***:**

It is a false negative rate, which means false classified as wrong given in equation (8)

FNR=FN/(FN+TP) (8)

Or

FNR= 1-TPR

i) *Precision***:**

It is classified items are truly classified given in equation (9)

Precision=TP/TP+FP (9)

j) Reca*ll***:**

It calculates, in actual item how many are classified given inequation (10)

RECALL=TP/TP+FN (10)

k) *F- Measure:*

It is a combination of precision and recall, providing a single measure. It measures the accuracy of the test. It is the harmonic mean of precision and recall given in equation (11)

$^{F1=2*}(Precision * Recall|Precision + Recall)$ (11)

l) *ROC area:*

It is the receiver operating characteristic curve**.** It examines the outcome of tested data. It reads the performance by creating a graph of TP vs. FP. It is useful to change the dataset that each instance is assigned a TP or FP class before the plot is made.

m) MCC:

It is known as the Matthews Correlation Coefficient, which measures the quality of binary classification given inequation (12).

$$\text{MCC}=\frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad (12)$$

n) *PRC AREA:*
It is a precision-recall curve**.** The PRC area is calculated separately for each class by treating instances of the class as “positive” instances and instances of all other classes as “negative” instances.

o) *Confusion matrix***:**
A confusion matrix is a technique for summarizing the performance of a classification algorithm shown in Table III.

TABLE III. CONFUSION MATRIX

	YES	NO
YES	TRUE POSITIVE(TP)	FALSE POSITIVE(FP)
NO	FALSE NEGATIVE(FN)	TRUE NEGATIVE(TN)

RESULTS AND DISCUSSIONS

To address the research Question 1:To project the academic success of students enrolled in a three-year regular graduation program.

To compare their outputs, various classification models such as J48, ZeroR, MLP, CART, and REPTree are used. Table IV shows the results of classifiers under the 10-fold cross-validation testing condition. The MLP classifier shows an accuracy of 99% and Kappa Statistics of 0.98 which means a very good correlation between class label and attributes. This model performs the best among all the models. But the time taken to build the model is 11.28 sec which is a higher amount. The J48 algorithm performs with 97.5% accuracy and 0.95 Kappa Statistics while the time taken to build the model is very less as compared to MLP. The ZeroR performs the worst with 58.4% accuracy and 0 Kappa values which means no agreement. For checking the evaluation results under Percentage Split where only 66% of data is used as training data as shown in Table V. The results show the improvement for all the classifiers. The MLP classifier achieved 100% accuracy and an excellent correlation value of 1.While ZeroR shows an increase to 60% but it is of no considerable use as the Kappa Statistics value is 0.

The results are also verified by combining the various classifiers with AdaBoostM1.It is a common classifier ensemble that can be integrated with other supervised learning techniques(Kaur & Kaur, 2016)With AdaBoostM1, the voting approach is used to combine the different classification algorithms. When classifiers are combined in a voting system, the class assigned to the test instance would be the one indicated by the majority of the ensemble's base-level classifiers.(Pandey & Taruna, 2016)(*The Stacking Ensemble Approach*, n.d.).The results are elaborated in Table VI and Table VII for Cross-validation and Percentage Split respectively. On investigating, Table IVand VI, it is found that the maximum increase in the accuracy is shown by the ZeroR from 58.4% to 93.3% and Kappa Statistics value from 0 to 0.95. Hence, ZeroR performs well when combined with the boosting algorithm. The other classifiers show the marginal increase in the accuracy and Kappa Statistics. While ADBoostM1+MLP shows a marginal decrease in performance. On analyzing, TableV and VII, the classification algorithms: J48, NaiveBayes, and CART show no change in values whereas AdaBoostM1+MLP shows a decline of 2% in the accuracy. But a huge impact on ZeroR shows an increase in accuracy to 98.6%. By considering all the parameters it is being observed that the MLP classifier performs the best among all the classifiers by achieving a 100% accuracy and Kappa Statistics value of 1. The Cost-benefit analysis curves are derived for each of the classifiers presented in Fig 1 to 6 for various classifiers. By minimizing the cost, the maximum gain is achieved by MLP classifiers of 82.61 and ZeroR performs the worst by having a gain of 0.

TABLE IV. COMPARISON RESULTS OF DIFFERENT CLASSIFIERS UNDER 10 FOLD CROSS-VALIDATION

Metrics **Models**	**Accuracy**	**Kappa statistics**	**ROC**	**Precision**	**Recall**	**F-measure**	**Time taken**	**Mean absolute error**	**Root mean squared error**
J48	97.5%	0.95	0.98	0.97	0.97	0.97	0.04	0.12	0.12
ZeroR	58.4%	0	0.84	0.34	0.58	0.43	0	0.37	0.43
Naviebayes	95.1%	0.9	0.99	0.95	0.95	0.95	0	0.03	0.16
MLP	99%	0.98	0.99	0.99	0.99	0.99	11.27	0.05	0.09
REPTree	89.3%	0.8	0.94	0.89	0.89	0.89	0.02	0.09	0.03
CART	97.1%	0.94	0.96	0.97	0.97	0.97	0.23	0.02	0.13

TABLE V. COMPARISON RESULTS OF DIFFERENT CLASSIFIERS UNDER PERCENTAGE SPLIT

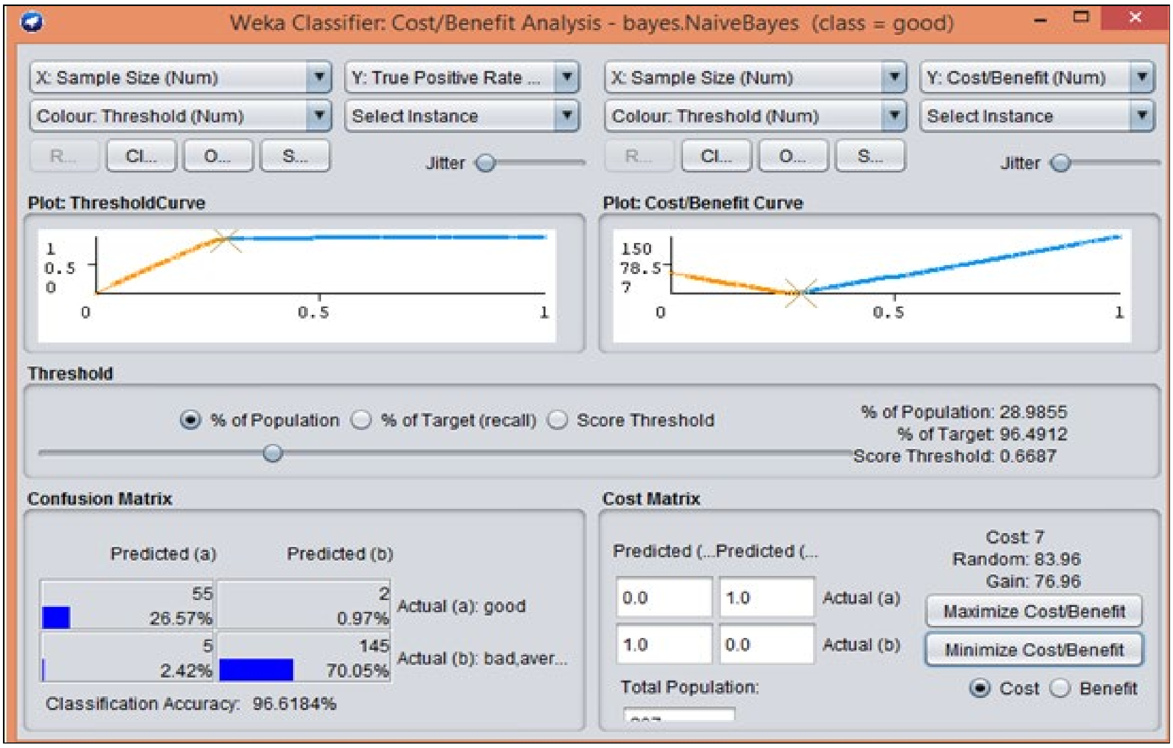
Metrics **Models**	**Accuracy**	**Kappa statistics**	**ROC**	**Precision**	**Recall**	**F-measure**	**Time taken**	**Mean absolute error**	**Root mean squared error**
J48	98.5%	0.97	0.98	0.98	0.98	0.98	0	0.01	0.09
ZeroR	60%	0	0.5	0.36	0.60	0.45	0	0.37	0.42
Naviebayes	95.7%	0.92	0.99	0.95	0.95	0.95	0	0.03	0.16
MLP	100%	1	1	1	1	1	10.36	0.06	0.07
REPTree	92.8%	0.86	0.98	0.93	0.92	0.92	0	0.06	0.18
CART	98.61%	0.97	0.98	0.98	0.98	0.98	0.24	0.01	0.09

TABLE VI. COMPARISON RESULTS OF DIFFERENT CLASSIFIERS WHEN COMBINED WITH ADABOOST UNDER CROSS-VALIDATION

Metrics **Models**	**Accuracy**	**Kappa statistics**	**ROC**	**Precision**	**Recall**	**F-measure**	**Time taken**	**Mean absolute error**	**Root mean squared error**
Adaboost+J48	97.6%	0.95	0.99	0.97	0.97	0.97	0.02	0.04	0.13
Adaboost+ ZeroR	93.3%	0.87	0.99	0.93	0.93	0.92	0.02	0.22	0.27
Adaboost+Naviebayes	97.6%	0.95	0.99	0.97	0.97	0.97	0.01	0.05	0.13
Adaboost+MLP	98.1%	0.96	1	0.98	0.98	0.98	10.52	0.06	0.12
Adaboost+REPTree	95.1%	0.91	0.99	0.95	0.95	0.95	0.01	0.08	0.17
Adaboost+CART	97.1%	0.94	0.99	0.97	0.97	0.97	0.28	0.05	0.14

TABLE VII. COMPARISON RESULTS OF DIFFERENT CLASSIFIERS WHEN COMBINED WITH ADABOOST UNDER PERCENTAGE SPLIT

Metrics Models	Accuracy	Kappa statistics	ROC	Precision	Recall	F-measure	Time taken	Mean absolute error	Root mean squared error
Adaboost+J48	98.6%	0.97	0.99	0.98	0.98	0.98	0.02	0.04	0.11
Adaboost+ ZeroR	98.6%	0.97	0.99	0.98	0.98	0.98	0.01	0.22	0.26
Adaboost+Naivebayes	95.7%	0.92	0.99	0.95	0.95	0.95	0.04	0.05	0.13
Adaboost+MLP	98.6%	0.97	1.0	0.98	0.98	0.98	10.28	0.07	0.10
Adaboost+Reptree	98.6%	0.97	1.0	0.98	0.98	0.98	0.08	0.07	0.13
Adaboost+CART	98.6%	0.97	0.99	0.98	0.98	0.98	0.24	0.04	0.11



● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 28.9855
% of Target: 96.4912
Score Threshold: 0.6687

Predicted (a)	Predicted (b)	
55 26.57%	2 0.97%	Actual (a): good
5 2.42%	145 70.05%	Actual (b): bad,aver...

Classification Accuracy: 96.6184%

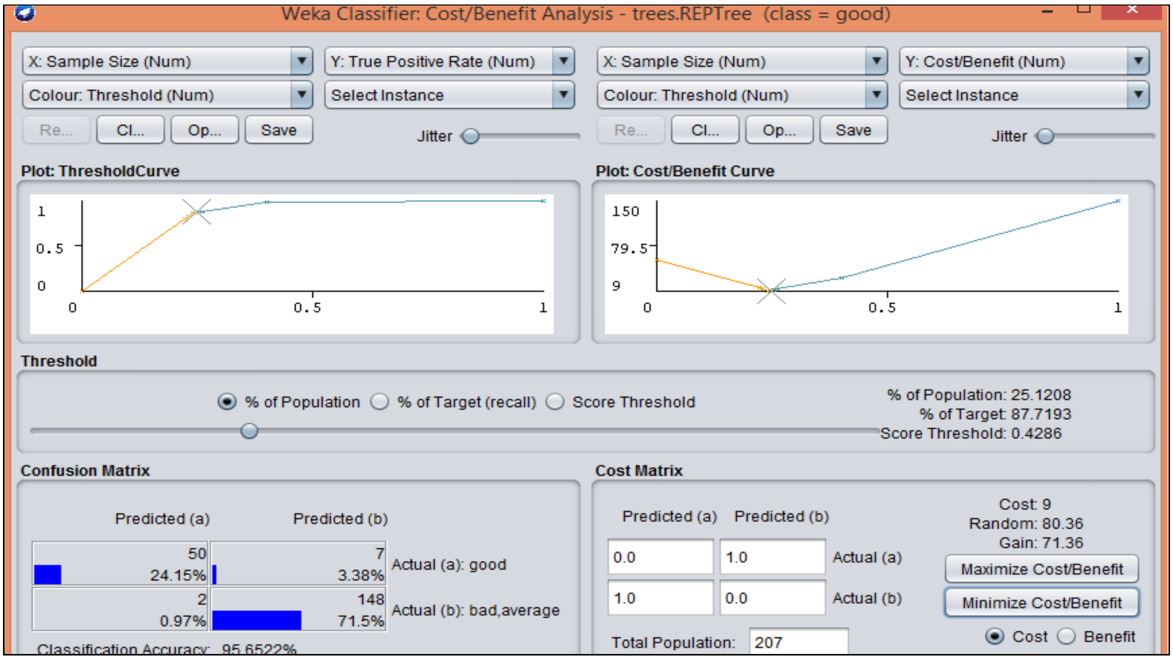
Predicted (...	Predicted (...	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Cost: 7
Random: 83.96
Gain: 76.96

Total Population:

● Cost ○ Benefit

Figure1: Cost-benefit Analysis for Naïve Bayes



Weka Classifier: Cost/Benefit Analysis - trees.REPTree (class = good)

X: Sample Size (Num) | Y: True Positive Rate (Num) | Colour: Threshold (Num) | Select Instance

X: Sample Size (Num) | Y: Cost/Benefit (Num) | Colour: Threshold (Num) | Select Instance

Threshold

● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 25.1208
% of Target: 87.7193
Score Threshold: 0.4286

Confusion Matrix

Predicted (a)	Predicted (b)	
50 24.15%	7 3.38%	Actual (a): good
2 0.97%	148 71.5%	Actual (b): bad,average

Classification Accuracy: 95.6522%

Cost Matrix

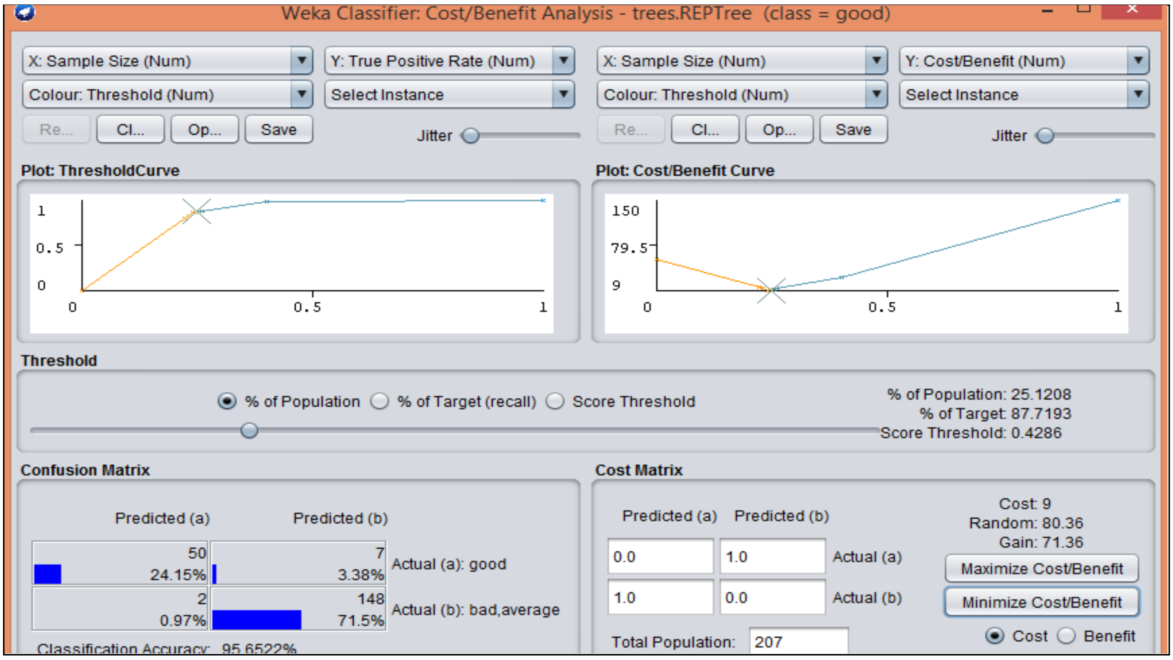
Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 207

Cost: 9
Random: 80.36
Gain: 71.36

● Cost ○ Benefit

Figure2: Cost-benefit Analysis for REPTree



Weka Classifier: Cost/Benefit Analysis - trees.REPTree (class = good)

X: Sample Size (Num) | Y: True Positive Rate (Num) | Colour: Threshold (Num) | Select Instance

X: Sample Size (Num) | Y: Cost/Benefit (Num) | Colour: Threshold (Num) | Select Instance

Threshold

● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 25.1208
% of Target: 87.7193
Score Threshold: 0.4286

Confusion Matrix

Predicted (a)	Predicted (b)	
50 24.15%	7 3.38%	Actual (a): good
2 0.97%	148 71.5%	Actual (b): bad,average

Classification Accuracy: 95.6522%

Cost Matrix

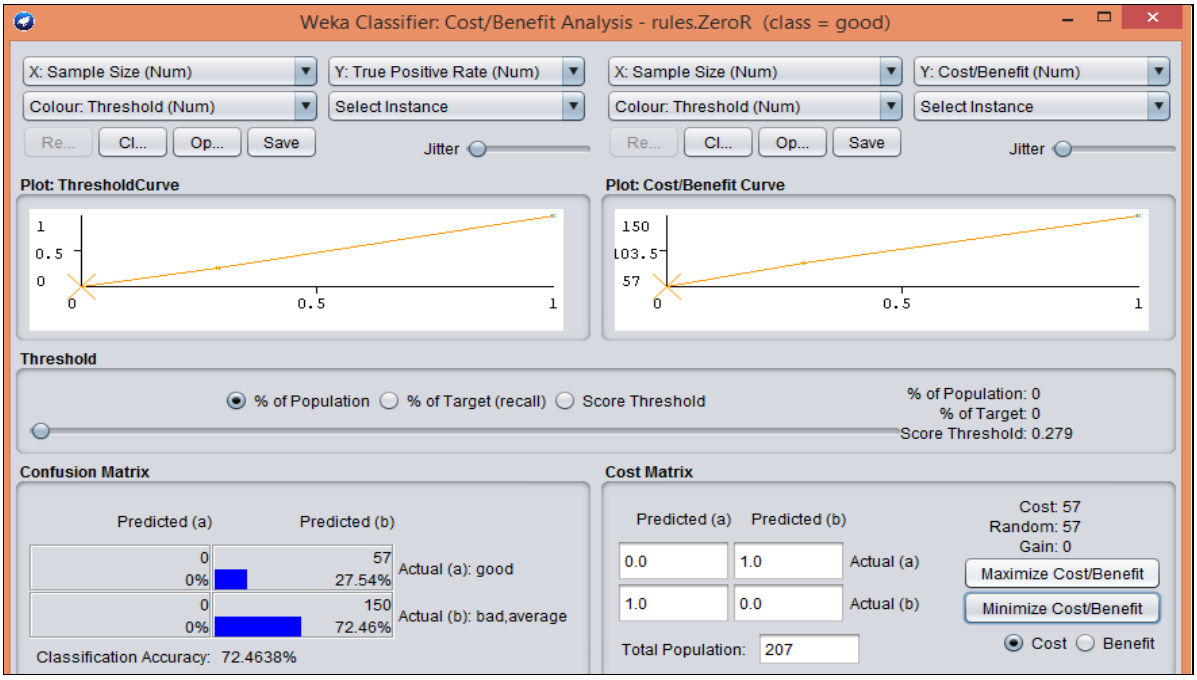
Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 207

Cost: 9
Random: 80.36
Gain: 71.36

● Cost ○ Benefit

Figure3: Cost-benefit Analysis for CART


Weka Classifier: Cost/Benefit Analysis - rules.ZeroR (class = good)

X: Sample Size (Num) | Y: True Positive Rate (Num) | X: Sample Size (Num) | Y: Cost/Benefit (Num)

Colour: Threshold (Num) | Select Instance | Colour: Threshold (Num) | Select Instance

Plot: ThresholdCurve

Plot: Cost/Benefit Curve

Threshold

● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 0
% of Target: 0
Score Threshold: 0.279

Confusion Matrix

Predicted (a)	Predicted (b)	
0 0%	57 27.54%	Actual (a): good
0 0%	150 72.46%	Actual (b): bad,average

Classification Accuracy: 72.4638%

Cost Matrix

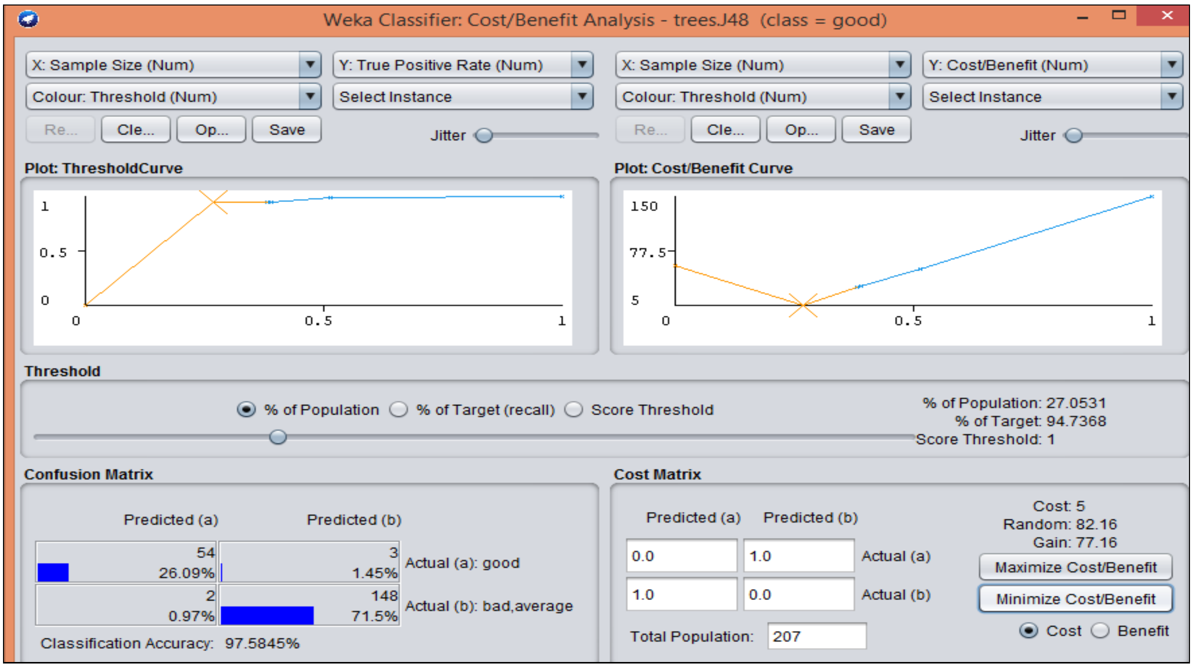
Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 207

Cost: 57
Random: 57
Gain: 0

● Cost ○ Benefit

Figure 4: Cost-benefit Analysis for ZeroR


Weka Classifier: Cost/Benefit Analysis - trees.J48 (class = good)

X: Sample Size (Num) | Y: True Positive Rate (Num) | X: Sample Size (Num) | Y: Cost/Benefit (Num)

Colour: Threshold (Num) | Select Instance | Colour: Threshold (Num) | Select Instance

Plot: ThresholdCurve

Plot: Cost/Benefit Curve

Threshold

● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 27.0531
% of Target: 94.7368
Score Threshold: 1

Confusion Matrix

Predicted (a)	Predicted (b)	
54 26.09%	3 1.45%	Actual (a): good
2 0.97%	148 71.5%	Actual (b): bad,average

Classification Accuracy: 97.5845%

Cost Matrix

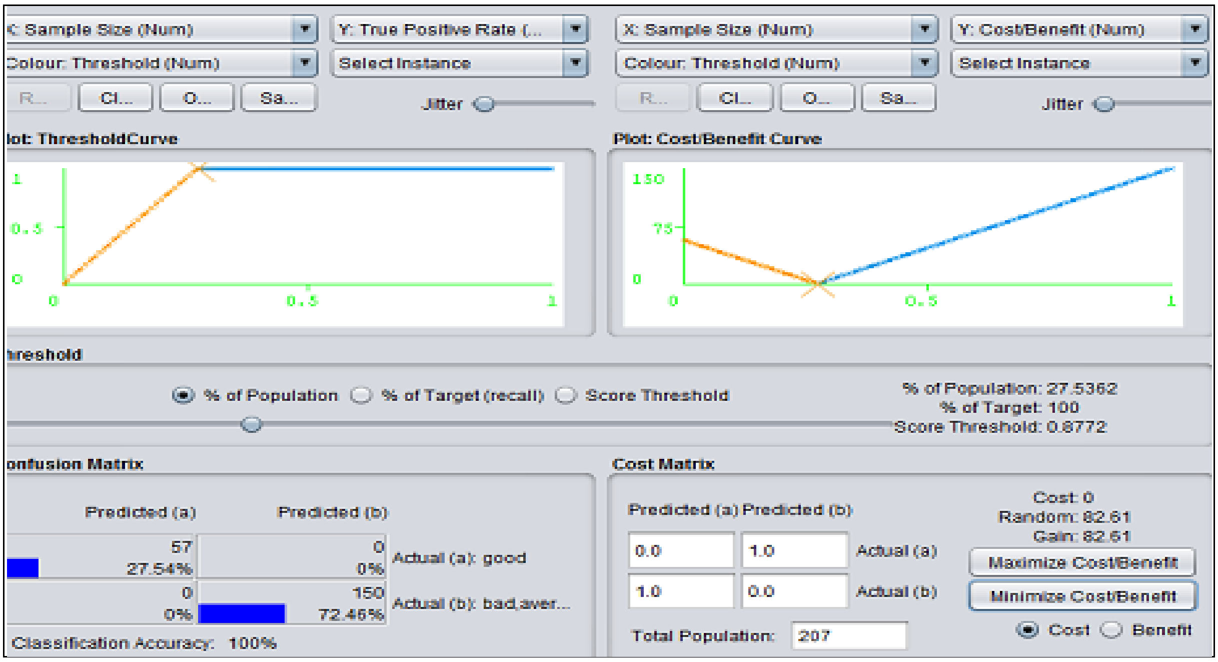
Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 207

Cost: 5
Random: 82.16
Gain: 77.16

● Cost ○ Benefit

Figure 5: Cost-benefit Analysis for J48


X: Sample Size (Num)
Y: True Positive Rate (...
Colour: Threshold (Num)
Select Instance

X: Sample Size (Num)
Y: Cost/Benefit (Num)
Colour: Threshold (Num)
Select Instance

Threshold
● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 27.5362
% of Target: 100
Score Threshold: 0.8772

Confusion Matrix

	Predicted (a)	Predicted (b)	
	57 27.54%	0 0%	Actual (a): good
	0 0%	150 72.46%	Actual (b): bad,aver...

Classification Accuracy: 100%

Cost Matrix

Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 207

Cost: 0
Random: 82.61
Gain: 82.61

● Cost ○ Benefit

Figure 6: Cost-benefit Analysis for MLP

To address research question 2: Which attributes rank the most important among the all in finding the Academic Performance.

For this Select Attributes tab is selected in WEKA and Attribute Evaluator is chosen as Classifier AttributeEval which evaluates the worth of an attribute by using a user-specified classifier. The Ranker Approach was chosen to rank the attributes based on their individual assessments(Affendey et al., 2010). By using the method and combining it with the Naive Bayes classifier to rank the attributes (as shown in Table VIII), it was discovered that background details and parameters, as well as required academic attributes like total points, play a significant role.

TABLE VIII: RANKING OF THE ATTRIBUTES

S no.	Attributes	Description	Ranking
1	RNO	Roll number of students	0.26
2	NM	Name of the students	0.001
3	DOB	Date of Birth	0.39
4	GEN	Gender of the Student	-0.0004
5	FN	Father's name of the student	0.11
6	MN	Mother's Name of the student	0.12
7	FO	Father's occupation	0.49
8	MO	Mother's occupation	0.23
9	AI	Annual Income of the student	0.51
10	CAT	Category of students	0.2
11	REL	The religion of the student	-0.0002
12	ADR	Address of the student	-0.0045
13	R10	Student 10th Result in percentage	0.47
14	R12	Student 12th result in percentage	0.45
15	STR	Subject opted by the student at school level like medical,non-medical, Commerce, Arts.	0.10
16	ARE	The Living area of student i.e. Urban or rural	0.59
17	PN	Phone no of the student	0

18	R1	1stSem result of BCA students	0.68
19	R2	2ndSem result of BCA students	0.61
20	R3	3rdSem result of BCA students	0.66
21	R4	4thSem result of BCA students	0.58
22	R5	5thSem result of BCA students	0.57
23	R6	6thSem result of BCA students	0.70
24	TOT	Total Marks of all semesters of a student	0.72
25	RES	Result show student result pass or fail	0.71
26	PER	Performance of student as good, average or bad	Base Class

To address research question 3: When are the dropout rates of the student's maximum.

For the answer to this, by observing the dataset it has been found that whose result in initial entry points to the institution is below the average are the most to get dropout in the graduation class. For this J48 tree was analyzed. The reasons for the drop-outs can be linked easily to the fear of new institutions, new courses, and a whole changed paradigm of the education environment. It can also be seen as a trend that the alarming number of students after the completion of their secondary examination is going abroad to pursue higher education. From the reports, Punjab has shown a decline of 30% in admissions from July 2017 and most of the students belong to rural areas. This number is sky-rocketing as youth is unable to find suitable jobs after the completion of the study. Moreover, the poor educational qualifications of the family are also the signaling parameter found in the study. More awareness campaigns, lectures should be delivered to motivate the youth. Also, a well-qualified faculty is a major requirement of these institutes to morale up the students to work hard.

CONCLUSION AND FUTURE SCOPE

This paper investigates the possibility to predict accurately the performance of the graduate students studying in the rural area with the help of a contemporary data mining tool as WEKA 3.8.1. The results show that MLP classifiers outperform all the classifiers picked for investigation. It gives 100% accuracy and a very good Kappa Statistics value of 1.This paper additionally brings to focus that not only accuracy but Kappa Statistics and ROC curves produce a huge impact on calculating the performance of the classifier. By using the ensemble approach, AdaBoost was ensembled with the chosen classifiers for the study. The results produce a picture that all the classifiers have shown an improvement in the performance of the classifier. However, MLP shows a marginal decline in the values whereas ZeroR classifiers perform the worst among all. When investigating the major attributes that affect the performance of the student, the most influential were the marks of the different semesters in graduation, Area(Rural or Urban), Parental Income, and Occupation. It was also seen from the results that most of the dropouts are occurring in the first year of the study in the course. This paper deals with many reasons for the alarming number of dropouts and how to increase the retention of the students. More suitable jobs, well-qualified faculty, and awareness campaigns are areas that need to be focused on to tackle this issue.

These results can be useful to investigate the keyholes and will help the management, teachers, and students to fill the gaps and to boost education in the rural areas. It also put pressure to look upon an alarming rise in the dropout rates in rural areas and what measures to be adopted to increase the performance of the students.

The research can be further extended by taking a dataset from an urban area college and comparing it with a rural area based on various parameters.

REFERENCES

Affendey, L. S., Paris, I. H. M., Mustapha, N., Sulaiman, M. N., & Muda, Z. (2010). Ranking of influencing factors in predicting students' academic performance. In *Information Technology Journal* (Vol. 9, Issue 4, pp. 832–837). https://doi.org/10.3923/itj.2010.832.837

Ali, S., Haider, Z., Munir, F., Khan, H., & Ahmed, A. (2013). Factors Contributing to the Students Academic Performance: A Case Study of Islamia University Sub-Campus. *American Journal of Educational Research*, *1*(8), 283–289. https://doi.org/10.12691/education-1-8-3

Andrew Braunstein, Michael McGrath, and D. P. (2015). Mining Student Data Using Decision Trees. *Expert Systems with Applications*, *40*(2), 1–18. https://doi.org/10.1017/CBO9781107415324.004

Asif, R., Merceron, A., & Pathan, M. K. (2014). Predicting Student Academic Performance at Degree Level: A Case Study. *International Journal of Intelligent Systems and Applications*, *7*(1), 49–61.

https://doi.org/10.5815/ijisa.2015.01.05

Bashir, U., & Chachoo, M. (2017). Performance Evaluation of J48 and Bayes Algorithms for Intrusion Detection System. *International Journal of Network Security & Its Applications*, *9*(4), 01–11. https://doi.org/10.5121/ijnsa.2017.9401

Chakraborty, M., Biswas, S. K., & Purkayastha, B. (2020). Data Mining Using Neural Networks in the form of Classification Rules: A Review. *4th International Conference on Computational Intelligence and Networks, CINE 2020*. https://doi.org/10.1109/CINE48825.2020.234399

Daud, A., Lytras, M. D., Aljohani, N. R., Abbas, F., Abbasi, R. A., & Alowibdi, J. S. (2019). Predicting student performance using advanced learning analytics. *26th International World Wide Web Conference 2017, WWW 2017 Companion*, 415–421. https://doi.org/10.1145/3041021.3054164

F.ElGamal, A. (2013). An Educational Data Mining Model for Predicting Student Performance in Programming Course. *International Journal of Computer Applications*, *70*(17), 22–28. https://doi.org/10.5120/12160-8163

Hardré, P. L., Crowson, H. M., Debacker, T. K., & White, D. (2007). Predicting the academic motivation of rural high school students. *Journal of Experimental Education*, *75*(4), 247–269. https://doi.org/10.3200/JEXE.75.4.247-269

Ibrahim, Z., & Rusli, D. (2007). Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision tree, And Linear Regression. *Proceedings of the 21st Annual SAS Malaysia Forum, September*, 1–6. https://www.researchgate.net/profile/Daliela_Rusli/publication/228894873_Predicting_Students'_Academic_Performance_Comparing_Artificial_Neural_Network_Decision_Tree_and_Linear_Regression/links/0deec51bb04e76ed93000000.pdf

J. Kovacic, Z. (2010). Early Prediction of Student Success: Mining Students Enrolment Data. *Proceedings of the 2010 InSITE Conference*, 647–665. https://doi.org/10.28945/1281

Kalmegh, S. (2015). Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart, and RandomTree for Classification of Indian News. *International Journal of Innovative Science, Engineering & Technology*, *2*(2), 438–446.

Kaur, K., & Kaur, K. (2016). Analyzing the effect of the difficulty level of a course on students' performance prediction using data mining. *Proceedings on 2015 1st International Conference on Next Generation Computing Technologies, NGCT 2015*, *September*, 756–761. https://doi.org/10.1109/NGCT.2015.7375222

Kumar, B., & Pal, S. (2011). Mining Educational Data to Analyze Students Performance. *International Journal of Advanced Computer Science and Applications*, *2*(6). https://doi.org/10.14569/ijacsa.2011.020609

Oancea, B., Dragoescu, R., & Ciucu, S. (2017). *network*. *72041*.

Pandey, M., & Taruna, S. (2016). Towards the integration of multiple classifiers pertaining to the Student's performance prediction. *Perspectives in Science*, *8*, 364–366. https://doi.org/10.1016/j.pisc.2016.04.076

Quadri, M., & Kalyankar, D. (2010). Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer*, *10*(2), 2–5. http://computerresearch.org/stpr/index.php/gjcst/article/viewArticle/128

The stacking ensemble approach. (n.d.). 82–95.

Vandamme, J. -P., Meskens, N., & Superby, J. -F. (2007). Predicting Academic Performance by Data Mining Methods. *Education Economics*, *15*(4), 405–419. https://doi.org/10.1080/09645290701409939