

1.ABSTRACT

Data mining is the evolving process of identifying and extracting the hidden information from a data warehouse. Data Mining is widely used in business, medical, engineering and educational areas for analyzing existing data, identifying measures for improvement and also forecasting the future prospects. This study covers the application of data mining in education for predicting the academic performance of the students. Educational Data Mining(EDM) plays a dominant role in the data mining era. There is an essential need to identify effective algorithms for predicting the student's performance. The dataset used in this research is taken from the University of Minho, Mathematics department which consists of 33 attributes and 650 observations. The algorithms used in this study to predict the academic performance of the students are Decision tree, Random Forest, Gradient Boosted tree, Logistic Regression and Multilayer Perceptron. Ensemble model is created by combining the Decision tree, C5.0 and Random Forest along with cross validation to improve the accuracy of the algorithms. The use of effective EDM techniques and tools would enable educators to improve the process by identifying any existing lacunae. EDM helps in developing a warning system for identifying weak student's prior and give adequate training to improve the academic performance of the students.

Keywords:

Student Performance Prediction, Educational Data Mining, Data Mining Technique, Academic Performance, Decision tree, Random Forest, Gradient Boosted tree

2. INTRODUCTION

Data Mining is process of analyzing the important information from a large set of data and come up with the prediction model. Data Mining is also called as Knowledge Discovery in Databases (KDD). The data mining plays an important role in all the fields like medical, airline, banking sector, movies, scientific information and numerous new data types. Data mining can be used to solve real time problems. Educational Data Mining (EDM) is the emerging technique for developing the prediction model with the help of available dataset, and extract the prediction of students' academic performance using machine learning technique. The prediction model acts like a warning system which is used to identify the weak students. EDM is a new field of research in

data mining. The recent increase in online learning by the students have led to the progression in development of EDM. The highly reputed educational institution mainly focuses on improving the performance of the students in order to retain the standard rank of the institution, hence they train the students in such way that they perform well in academics and extra-curricular activities. The data mining is classified into:

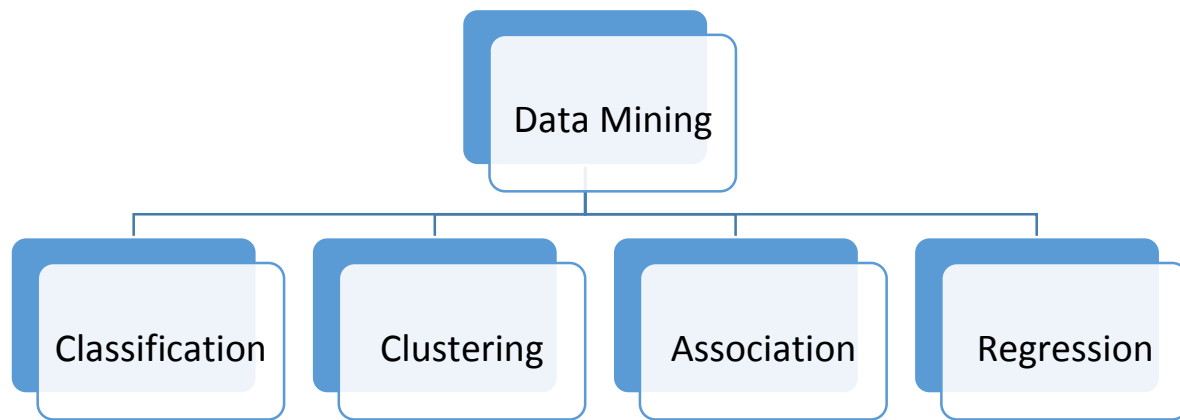


Figure 1: Data Mining Techniques

By using Educational data mining technique, the educational institutions can predict the performance of the students and identify low performing student early enough to overcome their difficulties in learning and improve their learning outcomes. Day by day the volume of the data is increased, hence there are different data mining algorithms which are used for predicting the performance of the students like supervised and unsupervised techniques to get the maximum accuracy. The supervised method is categorized into Classification or Categorization and Regression. The unsupervised method is categorized into Clustering and Association. Some of the algorithms which are popularly used in prediction are Decision tree, Multilayer Perceptron, Logistic Regression, Random forest, Gradient boosted trees, ID3 and J48. This study comprises of implementing different data mining techniques which are used in predicting the academic performance of the students. Two ensemble model is created, one model using Random Forest, Decision Tree and C5.0. Second model using Random Forest, Logistic Regression and Gradient Boosting. Major data mining techniques which are used for predicting the student's performance are shown below:

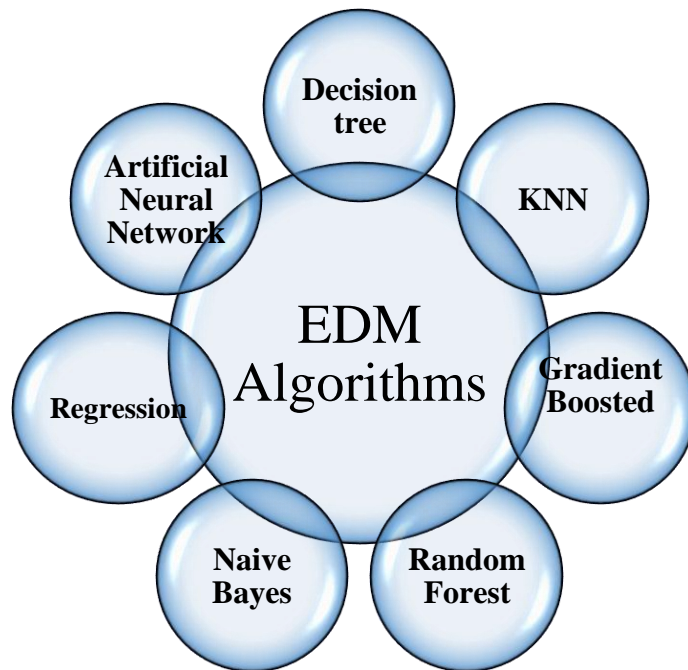


Figure 2: Data Mining Algorithm

RELATED WORK

Students' social activities and background details were used by Ching-Chieh Kiu [1] in the year 2018 for predicting the student's academic performance. This study used several data mining technique has been applied on predicting the accuracy of each dataset, of which decision tree J48 has given the best accuracy rate. The decision tree model has achieved 95% accuracy rate compared to other data mining technique. The dataset was divided into three subsets, students background details, student's social activities and student course work. The algorithms were implemented in WEKA tool. A comparative study on decision tree and random forest was done by Fergie Joanda and Reymon in the year 2018. The dataset was collected using questionnaires' from the computer science students. It had 249 instances out of which 3 different classes of data are used. The study proves that decision tree gives more accurate result of 66.9% compared to random forest which gave 61.44%. The comparison was implemented using WEKA tool [2].

A comparative study was done on Naïve Bayes, Decision Tree, K-Nearest Neighbor and Discriminant Analysis by Samuel, Nor Bahiah and Siti Mariyam in the year 2019. The study was done to identify the best data mining technique for predicting the student's academic performance. It used 10 datasets from the University of California Irvine Repository. The decision tree out performed with

the accuracy of 81.94% compared to other data mining techniques. The accuracy of other techniques was Naïve Bayes-73.61 %, KNN-80.56 % and Discriminant Analysis -77.78% accuracy rate. The tool used in the study was WEKA. In future hybrid metaheuristics algorithms will be for feature selection on the student data [3]. According to the study of Nongnuch Ketui, Warawut Wisomka and Kanitha Homjun in year 2019, Gradient boosted trees has given the best accuracy compared to other classification data mining techniques like Decision Tree, Weighted Decision Tree, Iterative Dichotomiser 3 (ID3) and Random Tree. WEKA is the data mining tool which is used for implementing the data mining techniques. A raw dataset was collected from the Rajamangala University of Technology Lanna Nan. The gradient boosted tree and decision tree gave good accuracy rate of 92.31% and 91.03% compared to other techniques. The Weighted Decision Tree 84.14%, ID3-89.66% and Random Tree-84.14% accuracy rate. The classification technique is widely used in predicting the student's performance [4].

According to Romero, Cristóbal, et al [5] SMO gives more accurate result compared to other techniques like BayesNet, Naïve Bayes Simple and EM. The paper used four different datasets and each dataset produced its own accuracy. The algorithms were implemented using WEKA tool. The SMO produced 82.4% accuracy result and BayesNet produced accuracy result of 81.5% accuracy. The Naïve Bayes Simple produced an accuracy of 82.4% and EM has 80.7% accuracy rate. All the algorithms performed equally good. In the year 2014 [6], Hu, Ya-Han, Chia-Lun Lo, and Sheng-Pao Shih made a study on predict students' online learning using C4.5, CART and LGR. The WEKA tool was used for implementation. The dataset used was learning portfolio data and C4.5 produced more accurate result of 93.4%. Yu, Liang-Chih, et al [7] used Sentiment analysis in order to predict the student's academic performance in the year 2018. The study used unstructured dataset. The WEKA tool was used for implementing the sentiment analysis which produced an 76% accuracy rate.

According to Deepika, K., and N. Sathvanaravana Support Vector Machine produces more accurate result compare to Linear Regression and Random forest. The study used student dataset of various academic disciplines of higher educational institutions in Kerala, India. The Linear Regression produced 89.96% accuracy and Random forest produced 89.98% and Support Vector Machine produced 91.43% accuracy result [8]. A comparative study was done in the year 2018 by Uzel and Vahide Nida on Multilayer Perceptron, Random Forest, Naïve Bayes, Decision Tree and Voting classifiers. The study used an educational dataset (xAPI) which is generated from an e-learning system includes 480 instances and 16 attributes. The Voting classifiers has highest accuracy of 80.6% [9]. The Artificial Neural Network outperformed the decision tree and Naïve Bayes. The study used dataset from Kalboard 360 e-learning system with 500 instances and 17 attributes. The Decision Tree has

71.1% accuracy, Naïve Bayes has 67.5% accuracy and ANN has highest accuracy of 78.1% [10]. According to Rawat, Keshav Singh, and I. V. Malhan a hybrid classification gives more accurate result for predicting the student's academic performance. The study used data set of Department of Computer Science with 27 instances and 11 attributes. The Decision Tree produced 86.7% and KNN produced 87.5%, ANN produced 81.3% and NB produced 87.5%, Hybrid produced highest accuracy rate of 93.3% [11].

PAPER	EDM TECHNIQUE	TOOLS	DATASET	RESULT ACCURACY
[1] Kiu, Ching-Chieh. "Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities." <i>2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)</i> . IEEE, 2018.	1.Naïve Bayesian(NB) 2.Multilayer Perceptron 3. Decision Tree(DT) 4.J48 5. Random Forest	WEKA	Used 395 instances with 33 attributes that described performance in Mathematics subjects	DT-95% NB-76%
[2] Kaunang, Fergie Joanda, and Reymon Rotikan. "Students' Academic Performance Prediction using Data Mining." <i>2018 Third International Conference on Informatics and Computing (ICIC)</i> . IEEE, 2018.	1.Decision Tree(DT) 2. Random Forest(RF)	WEKA	Used 249 records with 3 different classes	DT -66.9% RF-61.14%
[3] Ajibade, Samuel-Soma M., Nor Bahiah Ahmad, and Siti Mariyam Shamsuddin. "An Heuristic Feature Selection Algorithm to Evaluate Academic Performance of Students." <i>2019 IEEE 10th Control and System Graduate Research Colloquium (ICSGRC)</i> . IEEE, 2019.	1.Naïve Bayes (NB) 2.Decision Tree (DT) 3. K-Nearest Neighbor (KNN) 4. Discriminant Analysis (DISC)	WEKA	Used 10 different datasets that are gotten from the University of California Irvine (UCI) Repository.	NB-73.61% DT-81.94 % KNN-80.56 % DISC-77.78%

PAPER	EDM TECHNIQUE	TOOLS	DATASET	RESULT ACCURACY
[4] Ketui, Nongnuch, Warawut Wisomka, and Kanitha Homjun. "Using Classification Data Mining Techniques for Students Performance Prediction." <i>2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON)</i> . IEEE.	1.Decision Tree 2.Weighted Decision Tree(WDT) 3.Iterative Dichotomiser 3 (ID3) 4.Random Tree 5.Gradient Boosted Trees	WEKA	Education Division of Rajamangala University of Technology Lanna Nan (RMUTL Nan) for gave the raw dataset	DT-91.03% WDT-84.14% ID3-89.66% RT-84.14% GBT-92.31%
[5] Romero, Cristóbal, et al. "Predicting students' final performance from participation in on-line discussion forums." <i>Computers & Education</i> 68 (2013): 458-472.	1.SMO 2.BayesNet 3.NaiveBayesSimple 4.EM	Meerkat ED SNAPP	Used four different student dataset	SMO -82.4% BayesNet - 81.5% NaiveBayes82.4% EM -80.7%
[6] Hu, Ya-Han, Chia-Lun Lo, and Sheng-Pao Shih. "Developing early warning systems to predict students' online learning performance." <i>Computers in Human Behavior</i> 36 (2014): 469-478.	1.C4.5 2. CART 3. LGR.	WEKA	Used learning portfolio data	C4.5-93.4% CART-76.9% LGR.-95%
[7] Yu, Liang-Chih, et al. "Improving early prediction of academic failure using sentiment analysis on self-evaluated comments." <i>Journal of Computer Assisted Learning</i> 34.4 (2018): 358-365.	1.Sentiment Analysis	WEKA	Used unstructured data	Sentiment Analysis-76%

PAPER	EDM TECHNIQUE	TOOLS	DATASET	RESULT ACCURACY
[8] Deepika, K., and N. Sathvanaravana. "Analyze and Predicting the Student Academic Performance Using Data Mining Tools." <i>2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)</i> . IEEE, 2018.	1.Linear Regression 2.Random forest 3.SVM	WEKA	Used student dataset of various academic disciplines of higher educational institutions in Kerala, India	LR-89.96% RF -89.98% SVM-91.43%
[9] Uzel, Vahide Nida, Sultan Sevgi Turgut, and Selma Ayşe Özel. "Prediction of Students' Academic Success Using Data Mining Methods." <i>2018 Innovations in Intelligent Systems and Applications Conference (ASYU)</i> . IEEE, 2018.	1.Multilayer Perceptron(MP) 2. Random Forest 3.NaïveBayes 4.Decision Tree 5.Voting classifiers(VC)	WEKA	Used an educational dataset which includes 480 instances and 16 attributes	MP -78.3 % RF-76.6% NB -67.7% DT -75.8% VC-80.6%
[10] Siddiqui, Isma Farrah, and Qasim Ali Arain. "ANALYZING STUDENTS' ACADEMIC PERFORMANCE THROUGH EDUCATIONAL DATA MINING." <i>3C Tecnologia</i> (2019).	1.Decision Tree 2.Naïve Bayes 3.Artificial Neural Network	WEKA	Used dataset from Kalboard 360 e-learning system with 500 instances and 17 attributes	DT-71.1% NB -67.5% ANN -78.1%
[11] Rawat, Keshav Singh, and I. V. Malhan. "A Hybrid Classification Method Based on Machine Learning Classifiers to Predict Performance in Educational Data Mining." <i>Proceedings of 2nd International Conference on Communication, Computing and Networking</i> . Springer, Singapore, 2019.	1.Decision tree 2.KNN 3.Artificial neural network 4.Naïve Bayes 5.Hybrid	WEKA	Used data set of Department of Computer Science with 27 instances and 11 attributes	DT-86.7% KNN-87.5% ANN-81.3% NB-87.5% Hybird-93.3%

Decision tree has accuracy level from 66.9% to 95% and Random Forest has accuracy level from 61.14% to 89.98%. Naïve Bayes has accuracy level from 67.5% to 82.4% and KNN has accuracy level from 80.56% to 87.5%. ANN has accuracy level from 78.1% to 81.3% and Discriminant Analysis has accuracy level from 77.78%. ID3 has accuracy level from 89.66% Weighted Decision Tree has accuracy from 84.14%. Gradient Boosted Trees has accuracy level of 92.31% and Sentiment Analysis has accuracy level 76%. Linear Regression has accuracy level 89.96% and Support Vector Machine has accuracy level 91.43%. Multilayer perceptron has 78.3% accuracy and Voting classifier has 80.6% accuracy. CART has 76.9% accuracy and LGR has 95% accuracy.

PROPOSED WORK

In this experiment, the dataset used consists of 650 instances with 33 attributes that describes the performance in Mathematics subjects. The attributes of the dataset are divided into four subsets:

- 1) Students background with 18 attributes
- 2) Student social activities with 12 attributes
- 3) Student coursework results with 2 attributes
- 4) Important values with 18 attributes

These subsets attributes will be used to predict final grade(G3). G3 is a Numeric datatype with range of 1 – 10 used to measure student performance on their final grade. The subset attributes will be evaluated under models: 2-level classification (Pass / Fail). Important attributes are selected using variable Importance function in order to enhance the accuracy of the algorithm. Ensembling is applied on the Important attributes. The algorithms which are used in study are listed below

DECISION TREE

It's one of the most powerful classification algorithm with tree like structure. Decision trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. A decision tree is a tree-shaped diagram used to determine a course of action. Each branch of the tree represents a possible decision, occurrence, or reaction. It has a root node,

sub node and leaf node. The root node is the starting node of the tree followed by sub node which is used to make decisions and finally the leaf node which gives the end result of the classification.

RANDOM FOREST

Random forest is a supervised learning algorithm which is used for both classification as well as regression. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

LOGISTIC REGRESSION

Logistic regression is a statistical algorithm and it is mainly used for Binary classification problems (problems with two class values). Logistic regression is used to describe data and to explain the relationship between one dependent Binary variable and one or more Nominal, ordinal, interval or ratio-level independent variables.

GRADIENT BOOSTING

Gradient boosting is one of the most powerful techniques for building predictive models. The idea used in gradient boosting is a weak learner can be modified to become better. It uses a collection of decision tree which is built sequentially one after the other based on the result of the first tree the next tree performance is improved. Trees are constructed in a greedy manner, choosing the best split points based on purity scores like Gini or to minimize the loss

MULTILAYER PERCEPTRON

In the Multilayer perceptron, there can be more than one linear layer (combinations of neurons). If we take the simple example the three-layer network, first layer will be the input layer and last will be output layer and middle layer will be called hidden layer. The input data is provided into the input layer and take the output from the output layer. The number of the hidden layer can be increased as much as needed, to make the model more complex according to our task.

ENSEMBLE WITH TREE

Ensembling is a technique of combining two or more algorithms of similar or dissimilar types called base learners. This is done to make a more robust system which incorporates the predictions from all the base learners. Ensemble methods allows to produce better predictions compared to a single model. Popular technique used in Ensembling is Boosting and Bagging. This study uses two different Ensemble model. One ensemble model is created using Random Forest, logistic Regression and Gradient Boosting, other created using Decision tree, Random Forest and C5.0. Three types of concepts are used in Ensembling to combine the result which are listed below.

AVERAGING

It's defined as taking the average of predictions from models in case of regression problem or while predicting probabilities for the classification problem.

Model1	Model2	Model3	AveragePrediction
45	40	65	50

MAJORITY VOTE

It's defined as taking the prediction with maximum vote / recommendation from multiple models predictions while predicting the outcomes of a classification problem.

Model1	Model2	Model3	VotingPrediction
1	0	1	1

WEIGHTED AVERAGE

Different weights are applied to predictions from multiple models then taking the average which means giving high or low importance to specific model output.

	Model1	Model2	Model3	WeightAveragePrediction
Weight	0.4	0.3	0.3	
Prediction	45	40	60	48

3. SYSTEM ANALYSIS

ABOUT THE TOOL

R is a language and environment for statistical computing and graphics. R is a programming language developed by Ross Ihaka and Robert Gentleman in 1993. R possesses an extensive catalog of statistical and graphical methods. It includes machine learning algorithm, linear regression, time series, statistical inference to name a few. Most of the R libraries are written in R, but for heavy computational task, C, C++ and Fortran codes are preferred. R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- An effective data handling and storage facility
- A suite of operators for calculations on arrays, in particular matrices,
- A large, coherent, integrated collection of intermediate tools for data analysis
- Graphical facilities for data analysis and display either on-screen or on hardcopy
- A well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

Many users think of R as a statistics system. We prefer to think of it as an environment within which statistical techniques are implemented. R can be extended (easily) via packages. There are about eight packages supplied with the R distribution and many more are available through the CRAN family of Internet sites covering a very wide range of modern statistics. R is used for Statistical inference, Data analysis and Machine learning algorithm. Data science is shaping the way companies run their businesses. Without a doubt, staying away from Artificial Intelligence and Machine will lead the company to fail. The big question is which tool/language should you use? They are plenty of tools available in the market to perform data analysis. Learning a new language requires some time investment. The picture below depicts the learning curve compared to the business capability a language offers. The negative relationship implies that there is no free lunch. If you want to give the best insight from the data, then you need to spend some time learning the appropriate tool, which is R. On the top left of the graph, you can see Excel and PowerBI. These two tools are simple to learn but don't offer outstanding business capability, especially in term of modeling. In the middle, you can see Python and SAS. SAS is a dedicated tool to run a statistical analysis for business, but it is not free. SAS is a click and run software.

REQUIREMENT SPECIFICATION

HARWARE

The windows64 bit operating system with x64 based processor and 4GB RAM

SOFTWARE

R studio is the software tool which is used for implementing the machine learning algorithms. R is an open source software which has lots of statistical packages installed. It also has different packages for implementing specific algorithms.

DATASET

The dataset used in this study is taken from the University of Minho, Mathematics department which consists of 33 attributes and 650 observations. This dataset includes G1, G2, G3 major attributes and G3 containing the average grade of G1 and G2.

4.SYSTEM DESIGN

TAXONOMY OF THE SYSTEM



DATABASE DESIGN

STUDENT BACKGROUND			
Attribute	Description	Type	Value
sex	gender of student	Binary	male female
school	school of student		Mousinho da Silveira Gabriel Pereira
address	type of student's home address		rural urban
Pstatus	cohabitation status of parent		living together apart
famsize	size of family		≤ 3 > 3
schoolsup	extra educational school support		yes no
famsup	educational support from family		yes no
Mjob	job of mother	Nominal	- at home - civil services - teacher - health care related - other
Fjob	job of father		
reason	reason to choose this school		- close to home - school reputation - course preference - other
guardian	guardian of student		- father mother other
Medu	education of mother	Numeric	0# none 1# primary education 2# 5th to 9th grade 3# secondary education 4# higher education
Fedu	education of father		
famrel	quality of family relationships		very bad (1) to excellent (5)
age	age of student		15 - 22
traveltime	travel time from home to school		< 15 min 15 to 30 min 30 min. to 1 hour > 1 hour
studytime	weekly study time		< 2 hours 2 to 5 hours 5 to 10 hours > 10 hours
failures	number of failures in past class		n if $1 \leq n < 3$, else 4

STUDENT SOCIAL ACTIVITIES

Attribute	Description	Type	Value
activities	extra-curricular	Binary	yes no
higher	plans for higher education		
internet	home internet access		
nursery	nursery school attended		
paidclass	extra paid classes		
romantic	in romantic relationship		
absences	absences from school	Numeric	very low (1) to very high (5)
health	status of current health		
freetime	free time after school		
goout	outing with friends		
Dalc	consume alcohol in weekday		
Walc	consume alcohol in weekend		0 - 93

STUDENT COURSEWORK RESULT

Attribute	Description	Type	Value
G1	1st grade period	Numeric	0 - 20
G2	2nd grade period		

IMPORTANT ATTRIBUTES

Attribute	Description	Type	Value
failures	number of failures in past class	Numeric	n if $1 \leq n < 3$, else 4
studytime	weekly study time	Numeric	1# < 2 hours 2# 2 to 5 hours 3# 5 to 10 hours 4# > 10 hours
G2	2nd grade period	Numeric	0 – 20

Attribute	Description	Type	Value
absences	absences from school	Numeric	very low (1) to very high (5)
goout	outing with friends	Numeric	very low (1) to very high (5)
Wal-c	consume alcohol in weekend	Numeric	0 – 5
Dalc	consume alcohol in weekday	Numeric	0 – 5
traveltime	travel time from home to school	Numeric	0# < 15 min 1# 15 to 30 min 2# 30 min. to 1 hour 3# > 1 hour
famrel	quality of family relationships	Numeric	very bad (1) to excellent (5)
Fedu	education of father	Numeric	0# none 1# primary education 2# 5th to 9th grade 3# secondary education 4# higher education
reason	reason to choose this school	Nominal	- close to home - school reputation - course preference - other
guardian	guardian of student	Nominal	- father mother other
Fjob	job of father	Nominal	- at home - civil services - teacher - health care related - other
Mjob	job of mother	Nominal	
higher	plans for higher education	Binary	yes no
internet	home internet access	Binary	yes no
paidclass	extra paid classes	Binary	yes no

SCREEN DESIGN

- The implementation of Decision tree is given below

```
Untitled1* x DecisionTree.Rmd x
46 data_test <- create_train_test(stud, 0.8, train = FALSE)
47 dim(data_train)
48 dim(data_test)
49
50
51 prop.table(table(data_train$class))
52
53 #STEP4: Build the model
54
55 tree<-rpart(class~sex+school+Pstatus+failures+studytime+Medu+Fedu+traveltime+Fjob+Mjob+
56 famrel+reason+Pstatus+famsup+schoolsup+famsup+age+guardian, data=data_train,
57 method='class')
58 rpart.plot(tree,extra=106,cex=.7,roundint = FALSE)
59
60 #STEP5: Make a prediction
61 pred<-predict(tree, data_test, type = 'class')
62
63 table_mat <- table(data_test$class, pred)
64 table_mat
65
66 #STEP6: Measure performance
67 #ACCURACY=TP+TN/TP+TN+FP+FN
68
69 accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)
70 print(paste('Accuracy for train', accuracy_Test))
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645
2646
2647
2648
2649
2650
2651
2652
2653
2654
2655
2656
2657

```


- The implementation of Multilayer Perceptron is given below

```
43 stud <- stud[sample(1:nrow(stud), length(1:nrow(stud)), 1:nrow(stud))]
44 StudentValues <- stud[,1:18]
45 StudentTargets <- decodeClassLabels(stud[,19])
46 stud <- splitForTrainingAndTest(StudentValues, StudentTargets, ratio=0.20)
47 na.omit(stud$inputsTrain)
48 na.omit(stud$targetsTrain)
49
50 #STEP4:BUILT THE MULTILAYER PERCEPTRON MODEL
51
52 model <- mlp(
53   stud$inputsTrain, stud$targetsTrain, size=10, learnFuncParams=c(0.1),
54   maxit=50, inputsTest=stud$inputsTest, targetsTest=stud$targetsTest)
55
56 model
57 weightMatrix(model)
58 extractNetInfo(model)
59 par(mfrow=c(2,2))
60 plotIterativeError(model)
61
62 predictions <- predict(model, stud$inputsTest)
63
64 plotRegressionError(predictions[,2], stud$targetsTest[,2])
65
66
67 confusionMatrix(stud$targetsTrain, fitted.values(model))
68 table_mat<-confusionMatrix(stud$targetsTest, predictions)
69 table_mat
70 accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)
71 print(paste('Accuracy for test', accuracy_Test))
72
73:1 | Chunk 1 | R Markdown
```

- The implementation of Gradient Boosting is given below

```
42   return (data[-train_sample, ])
43 }
44 }
45
46 data_train <- create_train_test(stud, 0.8, train = TRUE)
47 data_test <- create_train_test(stud, 0.8, train = FALSE)
48 dim(data_train)
49 dim(data_test)
50 prop.table(table(data_train$class))
51
52
53
54 model <- train(class~sex+school+address+Pstatus+failures+studytime+Medu+Fedu+traveltime
55   +Fjob+Mjob+famrel+reason+Pstatus+famsup+schoolsup+famsup+age+guardian, data =
56   data_train, method = "xgbTree", trControl = trainControl("cv", number=10))
57
58 model$bestTune
59
60 # Make predictions on the test data
61 predicted.classes <- predict(model, data_test)
62 head(predicted.classes)
63
64 # Compute model prediction accuracy rate
65 mean(predicted.classes == data_test$class)
66
67
```

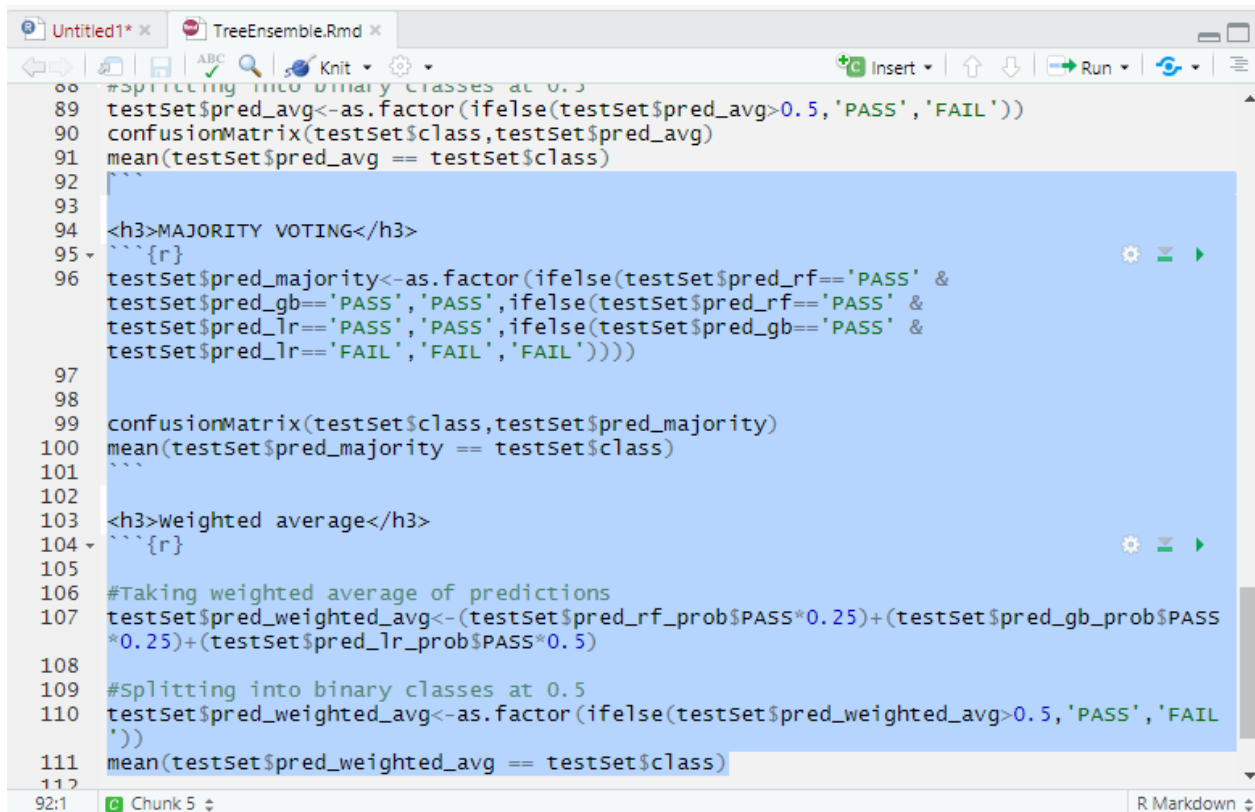
➤ The implementation of Logistic Regression is given below

```
logisticRegression.Rmd
41 test_pass <- pass[-pass_train, ]
42 test_fail <- fail[-fail_train, ]
43 test<- rbind(test_pass, test_fail) # row bind the pass and fail
44
45 table(test$class)
46 table(train$class)
47
48 mymodel<-glm(class~sex+school+address+Pstatus+failures+studytime+Medu+Fedu+traveltime+F
job+Mjob+famrel+reason+Pstatus+famsize+schoolsup+famsup+age+guardian,
family='binomial', data=train,maxit=100)
49
50 mymodel
51 summary(mymodel)
52
53 retest<-predict(mymodel,test,type="response")
54
55 ROCRPred<-prediction(retest,test$class)
56 ROCRPref<-performance(ROCRPred,"tpr","fpr")
57
58 plot(ROCRPref,colorize=TRUE,print.cutoffs.at=seq(0.1, by=0.1))
59
60 confmatrix<-table(Actual_value=test$class,Predicted_value=retest>=0.5)
61 confmatrix
62 accuracy<-sum(diag(confmatrix))/sum(confmatrix)
63 print(paste("Accuracy of the test",accuracy))
64
65 ```
66
67
```

➤ Implementing Ensemble

```
Ensemble.Rmd
80
81 ```
82 <h3>AVERAGING</h3>
83 ```{r}
84 #Predicting the probabilities
85 testset$pred_rf_prob<-predict(object = model_rf,testSet[,predictors],type='prob')
86 testset$pred_gb_prob<-predict(object = model_gb,testSet[,predictors],type='prob')
87 testset$pred_lr_prob<-predict(object = model_lr,testSet[,predictors],type='prob')
88
89 #Taking average of predictions
90 testset$pred_avg<-(testset$pred_rf_prob$PASS+testset$pred_gb_prob$PASS+testset$pred_lr
_prob$PASS)/3
91
92 #Splitting into binary classes at 0.5
93 testset$pred_avg<-as.factor(ifelse(testset$pred_avg>0.5,'PASS','FAIL'))
94
95 confusionMatrix(testset$class,testset$pred_avg)
96 mean(testset$pred_avg == testset$class)
97
98 ```
99
100 <h3>MAJORITY VOTING</h3>
101 ```{r}
102 testset$pred_majority<-as.factor(ifelse(testset$pred_rf=='PASS' &
testset$pred_gb=='PASS', 'PASS', ifelse(testset$pred_rf=='PASS' &
testset$pred_lr=='PASS', 'PASS', ifelse(testset$pred_gb=='PASS' &
testset$pred_lr=='FAIL', 'FAIL', 'FAIL'))))
103
104
```

➤ Implementing Ensemble with Tree algorithms



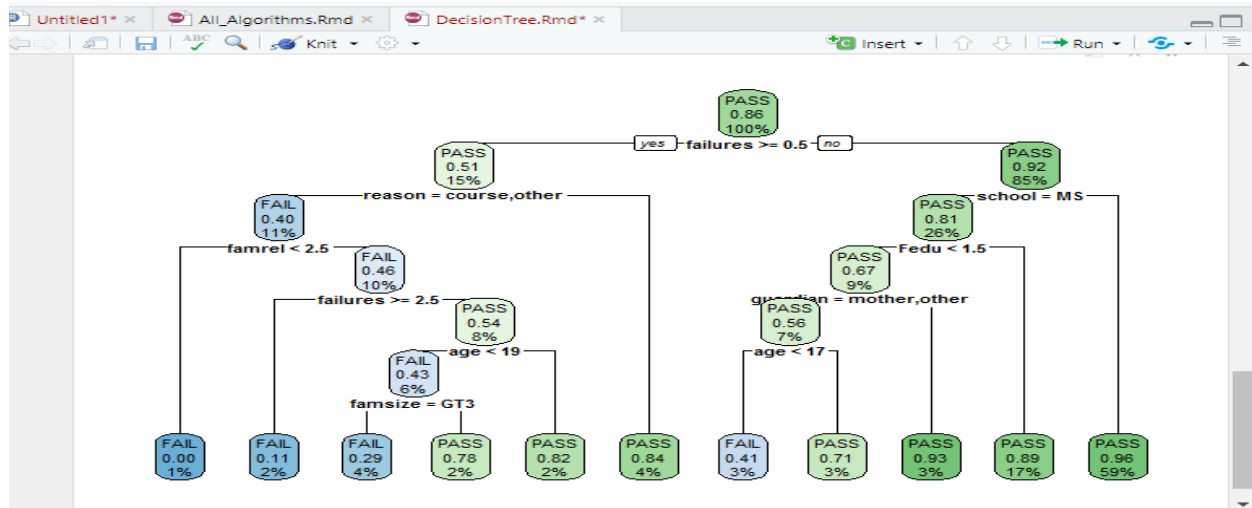
```
88 #splitting into binary classes at 0.5
89 testSet$pred_avg<-as.factor(ifelse(testSet$pred_avg>0.5,'PASS','FAIL'))
90 confusionMatrix(testSet$class,testSet$pred_avg)
91 mean(testSet$pred_avg == testSet$class)
92
93
94 <h3>MAJORITY VOTING</h3>
95 ```{r}
96 testSet$pred_majority<-as.factor(ifelse(testSet$pred_rf=='PASS' &
97 testSet$pred_gb=='PASS', 'PASS', ifelse(testSet$pred_rf=='PASS' &
98 testSet$pred_lr=='PASS', 'PASS', ifelse(testSet$pred_gb=='PASS' &
99 testSet$pred_lr=='FAIL', 'FAIL', 'FAIL'))))
100
101 confusionMatrix(testSet$class,testSet$pred_majority)
102 mean(testSet$pred_majority == testSet$class)
103
104 <h3>weighted average</h3>
105 ```{r}
106 #Taking weighted average of predictions
107 testSet$pred_weighted_avg<-(testSet$pred_rf_prob$PASS*0.25)+(testSet$pred_gb_prob$PASS
108 *0.25)+(testSet$pred_lr_prob$PASS*0.5)
109 #splitting into binary classes at 0.5
110 testSet$pred_weighted_avg<-as.factor(ifelse(testSet$pred_weighted_avg>0.5,'PASS','FAIL'))
111 mean(testSet$pred_weighted_avg == testSet$class)
```

REPORT

The implementation of decision tree with background details gave 73% accuracy, with social activities attributes 78% accuracy, with course work attributes 93% accuracy and with important attributes it gave 93% accuracy. The implementation of Random forest with background details gave 76% accuracy, with social activities attributes 78% accuracy, with course work attributes 91% accuracy and with important attributes it gave 91% accuracy. The implementation of Logistic Regression with background details gave 87% accuracy, with social activities attributes 83% accuracy, with course work attributes 93% accuracy and with important attributes it gave 92% accuracy. The implementation of Gradient Boosting with background details gave 79% accuracy, with social activities attributes 74% accuracy, with course work attributes 93% accuracy and with important attributes it gave 92% accuracy. The implementation of Logistic Regression with background details gave 83% accuracy, with social activities attributes 83% accuracy, with course work attributes 90% accuracy and with important attributes it gave 93% accuracy.

EXPERIMENTAL RESULTS

➤ Output of the decision tree



➤ Output of Random forest

```
41 result<-data.frame(testing$class, predict(rf, testing[,1:32], type="response"))
42 result
43 plot(result)
44
45
46 #ACCURACY OF THE MODEL
47 prediction<-predict(rf,testing,type="class")
48
49 ConfusionMatrix<-table(prediction,testing$class)
50 ConfusionMatrix
51
52 accuracy<-sum(diag(ConfusionMatrix))/sum(ConfusionMatrix)
53 print(paste("Accuracy of the test",accuracy))
54
55 #STEP5:VARIABLE IMPORTANCE
56 varImpPlot(rf,sort = T,main="Variable Importance",n.var=5)
57
58 var.imp <- data.frame(importance(rf,type=2))
59 # make row names as columns
60 var.imp$variables <- row.names(var.imp)
61 var.imp[order(var.imp$MeanDecreaseGini,decreasing = T),]
62
63
64 ...
```

[1] "Accuracy of the test 0.769230769230769"

➤ Output of Gradient Boosting

```
Untitled1* x GradientBoosting.Rmd x
49 dim(data_test)
50 prop.table(table(data_train$class))
51
52
53
54 model <- train(class~sex+school+address+Pstatus+failures+studytime+Medu+Fedu+traveltime
+Fjob+Mjob+famrel+reason+Pstatus+famsize+schoolsup+famsup+age+guardian, data =
data_train, method = "xgbTree", trcontrol = trainControl("cv", number=10))
55
56 model$bestTune
57
58 # Make predictions on the test data
59 predicted.classes <- predict(model, data_test)
60 head(predicted.classes)
61
62 # Compute model prediction accuracy rate
63 cat("Accuracy of Gradient Boosting", mean(predicted.classes == data_test$class))
64
65
66
67
68
69
70
Accuracy of Gradient Boosting 0.7923077
```

➤ Output of Multilayer Perceptron

```
Untitled1* x MultilayerPerceptron.Rmd x
50
51 model <- mlp(
52   stud$inputsTrain, stud$targetsTrain, size=10, learnFuncParams=c(0.1),
53   maxit=50, inputsTest=stud$inputsTest, targetsTest=stud$targetsTest)
54
55 model
56 weightMatrix(model)
57 extractNetInfo(model)
58 par(mfrow=c(2,2))
59 plotIterativeError(model)
60
61 predictions <- predict(model, stud$inputsTest)
62
63 plotRegressionError(predictions[,2], stud$targetsTest[,2])
64
65
66 confusionMatrix(stud$targetsTrain, fitted.values(model))
67 table_mat <- confusionMatrix(stud$targetsTest, predictions)
68 table_mat
69 accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)
70 print(paste('Accuracy for test', accuracy_Test))
71
72
73
```

➤ Output of Logistic Regression

```
44
45 table(test$class)
46 table(train$class)
47
48 mymodel<-glm(class~sex+school+address+Pstatus+failures+studytme+Medu+Fedu+traveltime+F
job+Mjob+famrel+reason+Pstatus+famsize+schoolsup+famsup+age+guardian,
family='binomial', data=train,maxit=100)
49
50 mymodel
51 summary(mymodel)
52
53 retest<-predict(mymodel,test,type="response")
54
55 ROCRPred<-prediction(retest,test$class)
56 ROCRPref<-performance(ROCRPred,"tpr","fpr")
57
58 plot(ROCRPref,colorize=TRUE,print.cutoffs.at=seq(0.1, by=0.1))
59
60 confmatrix<-table(Actual_value=test$class,Predicted_value=retest>=0.5)
61 confmatrix
62 accuracy<-sum(diag(confmatrix))/sum(confmatrix)
63 print(paste("Accuracy of the test",accuracy))
64
65 [1] "Accuracy of the test 0.871794871794872"
```

➤ Output of Ensemble

MAJORITY VOTING

```
testSet$pred_majority<-as.factor(ifelse(testSet$pred_rf=='PASS' & testSet$pred_gb=='PASS','PASS',ifelse(testSet$pred_rf=='PA
SS' & testSet$pred_lr=='PASS','PASS',ifelse(testSet$pred_gb=='PASS' & testSet$pred_lr=='FAIL','FAIL','FAIL'))))

confusionMatrix(testSet$class,testSet$pred_majority)

## Confusion Matrix and Statistics
##
##      Reference
## Prediction FAIL PASS
##      FAIL    14     6
##      PASS     0    109
##
##      Accuracy : 0.9535
##      95% CI : (0.9015, 0.9827)
##      No Information Rate : 0.8915
##      P-Value [Acc > NIR] : 0.01065
##
##      Kappa : 0.7977
##
##      Mcnemar's Test P-Value : 0.04123
##
##      Sensitivity : 1.0000
##      Specificity : 0.9478
##      Pos Pred Value : 0.7000
##      Neg Pred Value : 1.0000
##      Prevalence : 0.1085
##      Detection Rate : 0.1085
##      Detection Prevalence : 0.1550
##      Balanced Accuracy : 0.9739
##
##      'Positive' Class : FAIL
##
```

➤ Output of Tree Ensemble

MAJORITY VOTING

```
testSet$pred_majority<-as.factor(ifelse(testSet$pred_rf=='PASS' & testSet$pred_gb=='PASS','PASS',ifelse(testSet$pred_rf=='PASS' & testSet$pred_lr=='PASS','PASS',ifelse(testSet$pred_gb=='PASS' & testSet$pred_lr=='FAIL','FAIL','FAIL'))))
```

```
confusionMatrix(testSet$class,testSet$pred_majority)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  FAIL  PASS
##      FAIL    17     3
##      PASS     2    107
##
##              Accuracy : 0.9612
##              95% CI : (0.9119, 0.9873)
##      No Information Rate : 0.8527
##      P-Value [Acc > NIR] : 6.419e-05
##
##              Kappa : 0.849
##
##  Mcnemar's Test P-Value : 1
##
##      Sensitivity : 0.8947
##      Specificity : 0.9727
##      Pos Pred Value : 0.8500
##      Neg Pred Value : 0.9817
##      Prevalence : 0.1473
##      Detection Rate : 0.1318
##      Detection Prevalence : 0.1550
##      Balanced Accuracy : 0.9337
##
##      'Positive' Class : FAIL
```

5.IMPLEMENTATION

The dataset used in this study is taken from the University of Minho, Mathematics department which consists of 33 attributes and 650 observations. This dataset includes G1, G2, G3 major attributes and G3 containing the average grade of G1 and G2. The response attribute class is created using G3 attribute. This class attribute contains only the pass and fail values, the students who got more than 10 in G3 they are placed into pass class and students who got less than 10 in G3 are placed in fail category. The dataset is divided into three groups: Background details, Social activities and Course work. The Background details contain Sex, School, Address, Pstatus, famsize, schoolsup, famup, Mjob, Fjob, Fedu, Medu, Reason, Guardian, Famrel, Age, Travel time, Study time and Failures. The Social activities contain Activities, Higher, Internet, Nursery, Paid Class, Romantic, Absences, Health, Free time, Gout, Dalc and Walc. The course work contains G1 and G2 attribute. From the 33 attributes 14 attributes are selected as important attributes by using the Variable Importance function in order to improve the accuracy of the algorithms.

SCREENSHOT

Untitled1* x logisticRegression.Rmd x dataset x

Filter

	failures	studytime	famrel	travelttime	guardian	higher	absences	goout	Dalc	Walc	internet	paid	Fedu	Mjob	Fjob	reason	G2
1	0	2	4	2	mother	yes	4	4	1	1	no	no	4	at_home	teacher	course	11
2	0	2	5	1	father	yes	2	3	1	1	yes	no	1	at_home	other	course	11
3	0	2	4	1	mother	yes	6	2	2	3	yes	no	1	at_home	other	other	13
4	0	3	3	1	mother	yes	0	2	1	1	yes	no	2	health	services	home	14
5	0	2	4	1	father	yes	0	2	1	2	no	no	3	other	other	home	13
6	0	2	5	1	mother	yes	6	2	1	2	yes	no	3	services	other	reputation	12
7	0	2	4	1	mother	yes	0	4	1	1	yes	no	2	other	other	home	12
8	0	2	4	2	mother	yes	2	4	1	1	no	no	4	other	teacher	home	13
9	0	2	4	1	mother	yes	0	2	1	1	yes	no	2	services	other	home	16
10	0	2	5	1	mother	yes	0	1	1	1	yes	no	4	other	other	home	12
11	0	2	3	1	mother	yes	2	3	1	2	yes	no	4	teacher	health	reputation	14
12	0	3	5	3	father	yes	0	2	1	1	yes	no	1	services	other	reputation	12
13	0	1	4	1	father	yes	0	3	1	3	yes	no	4	health	services	course	13
14	0	2	5	2	mother	yes	0	3	1	2	yes	no	3	teacher	other	course	12
15	0	3	4	1	other	yes	0	2	1	1	yes	no	2	other	other	home	14
16	0	1	4	1	mother	yes	6	4	1	2	yes	no	4	health	other	home	17

Showing 1 to 17 of 649 entries, 18 total columns

```

40
41 test_pass <- pass[-pass_train, ]
42 test_fail <- fail[-fail_train, ]
43 test<- rbind(test_pass, test_fail) # row bind the pass and fail
44
45 table(test$class)
46 table(train$class)
47
48 mymodel<-glm(class~sex+school+address+Pstatus+failures+studytime+Medu+Fedu+travelttime+Fjob+Mjob+famrel+reason+Pstatus+famsize+schoolsup+famsup+age+guardian, family='binomial', data=train,maxit=100)
49
50 mymodel
51 summary(mymodel)
52
53 retest<-predict(mymodel,test,type="response")
54
55 ROCRPred<-prediction(retest,test$class)
56 ROCRPref<-performance(ROCRPred,"tpr","fpr")
57
58 plot(ROCRPref,colorize=TRUE,print.cutoffs.at=seq(0.1, by=0.1))
59
60 confmatrix<-table(Actual_value=test$class,Predicted_value=retest>=0.5)
61 confmatrix
62 accuracy<-sum(diag(confmatrix))/sum(confmatrix)
63 print(paste("Accuracy of the test",accuracy))
64
65 ```
66

```


SAMPLE CODE

```
chisq.test(class,Fjob)
chisq.test(class,Mjob)
chisq.test(class,paid)
chisq.test(class,guardian)
cor(Fedu,G3)
cor(traveltime,G3)
cor(absences,G3)
cor(Walc,G3)
cor(Dalc,G3)
create_train_test <- function(data, size = 0.8, train = TRUE) {
  n_row = nrow(data)
  total_row = size * n_row
  train_sample <- 1: total_row
  if (train == TRUE) {
    return (data[train_sample, ])
  } else {
    return (data[-train_sample, ])
  }
}

data_train <- create_train_test(stud, 0.8, train = TRUE)
data_test <- create_train_test(stud, 0.8, train = FALSE)

rf<-randomForest(class~failures+studytime+G2+higher+absences+goout+Walc+famrel+reason+Fedu+internet+Fjob,data=data_train, mtry=12, ntree=500, importance=TRUE)


rf


mymodel<-glm(class~failures+studytime+G2+higher+absences+goout+Walc+famrel+reason+Fedu+internet+Fjob, family='binomial', data=data_train,maxit=100)


model <- mlp(stud$inputsTrain, stud$targetsTrain, size=5, learnFuncParams=c(0.1),
  maxit=50, inputsTest=stud$inputsTest, targetsTest=stud$targetsTest)
```

6.CODE REVIEW & TESTING

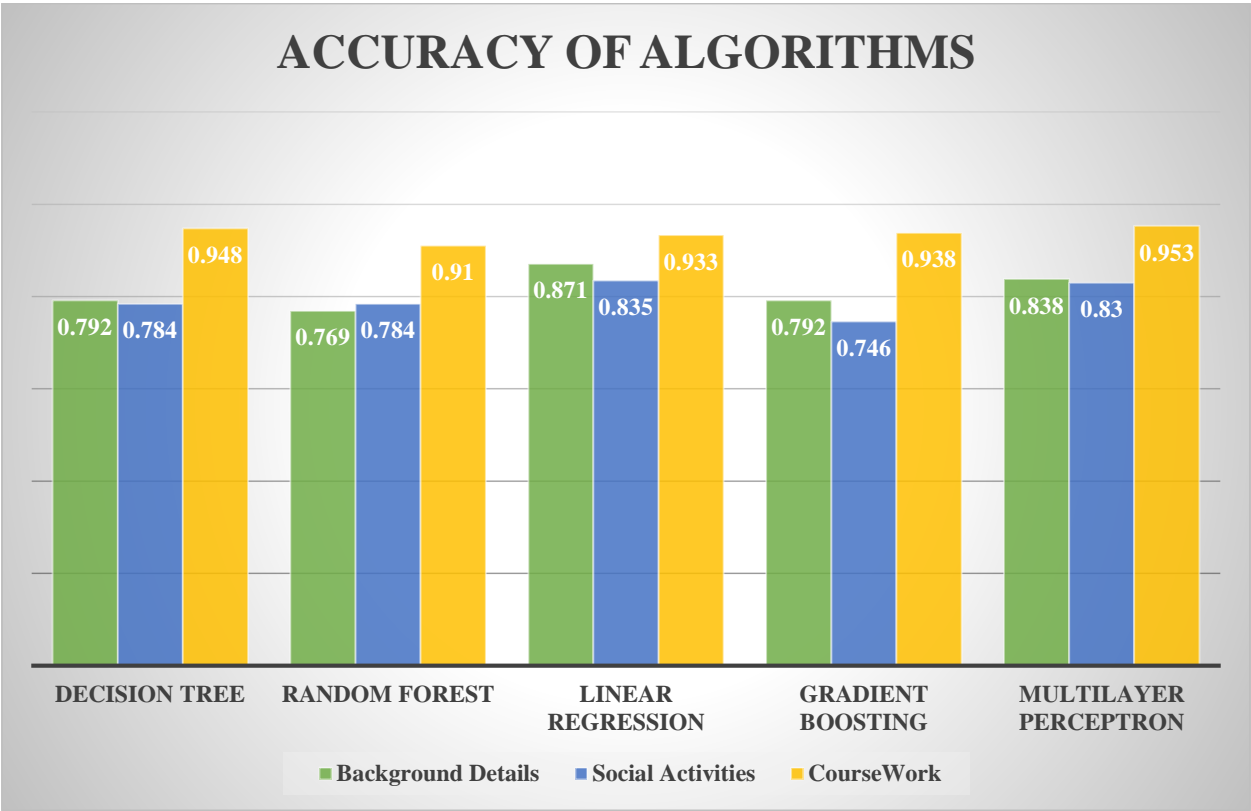
Individual Algorithms Accuracy				
S.no	Algorithm	Background Details	Social Activities	Course
1	Decision Tree	0.792	0.784	0.93
2	Random Forest	0.769	0.784	0.91
3	Linear Regression	0.871	0.835	0.933
4	Gradient Boosting	0.792	0.746	0.938
5	Multilayer Perceptron	0.838	0.83	0.9

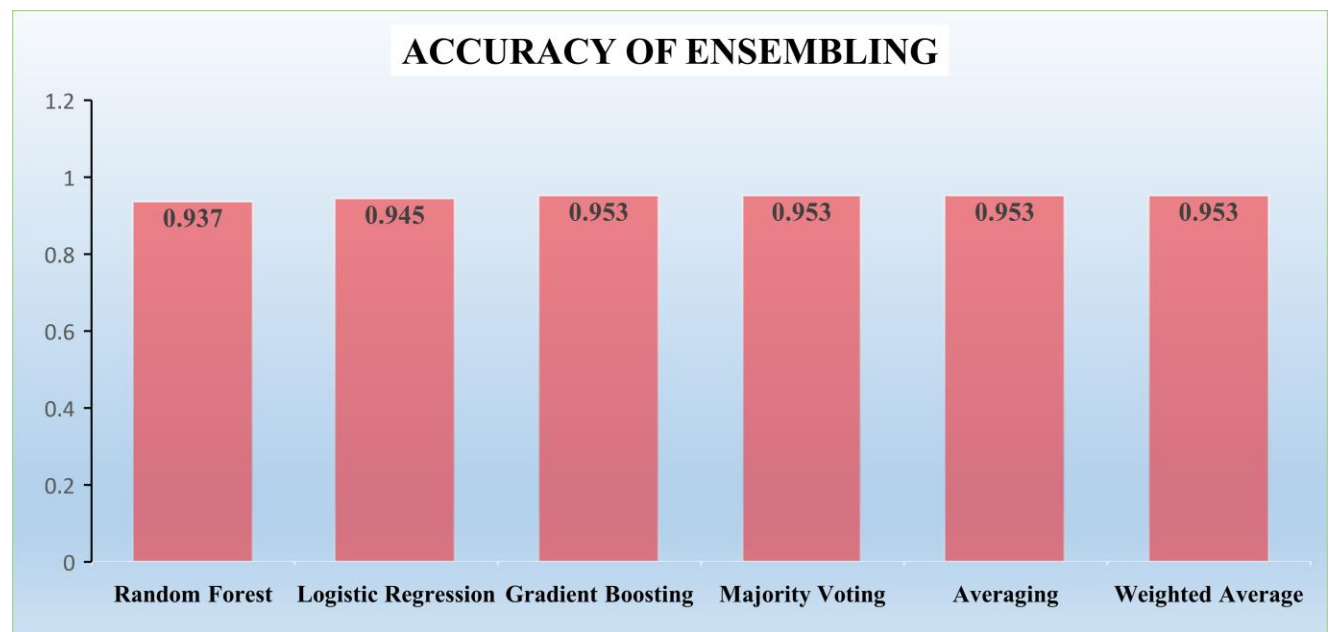
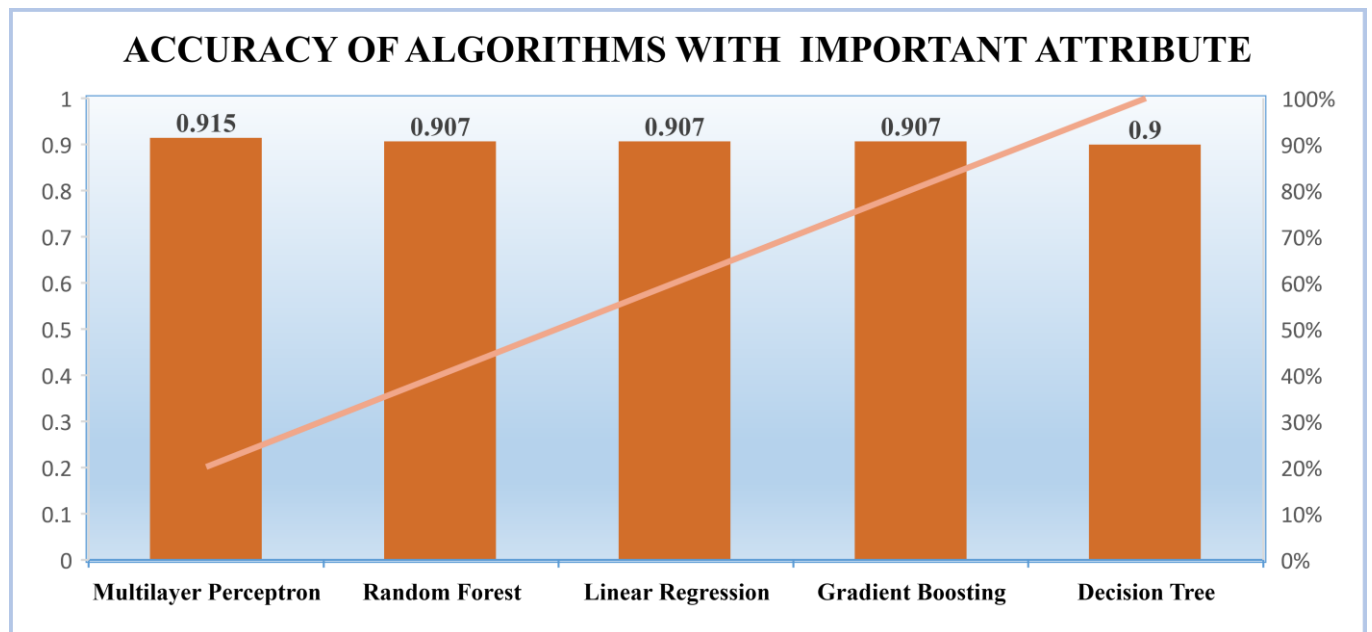
Individual Algorithms Accuracy with Important Attributes 		
S.no	Algorithm	Accuracy
1	Decision Tree	0.93
2	Random Forest	0.915
3	Linear Regression	0.923
4	Gradient Boosting	0.923
5	Multilayer Perceptron	0.93

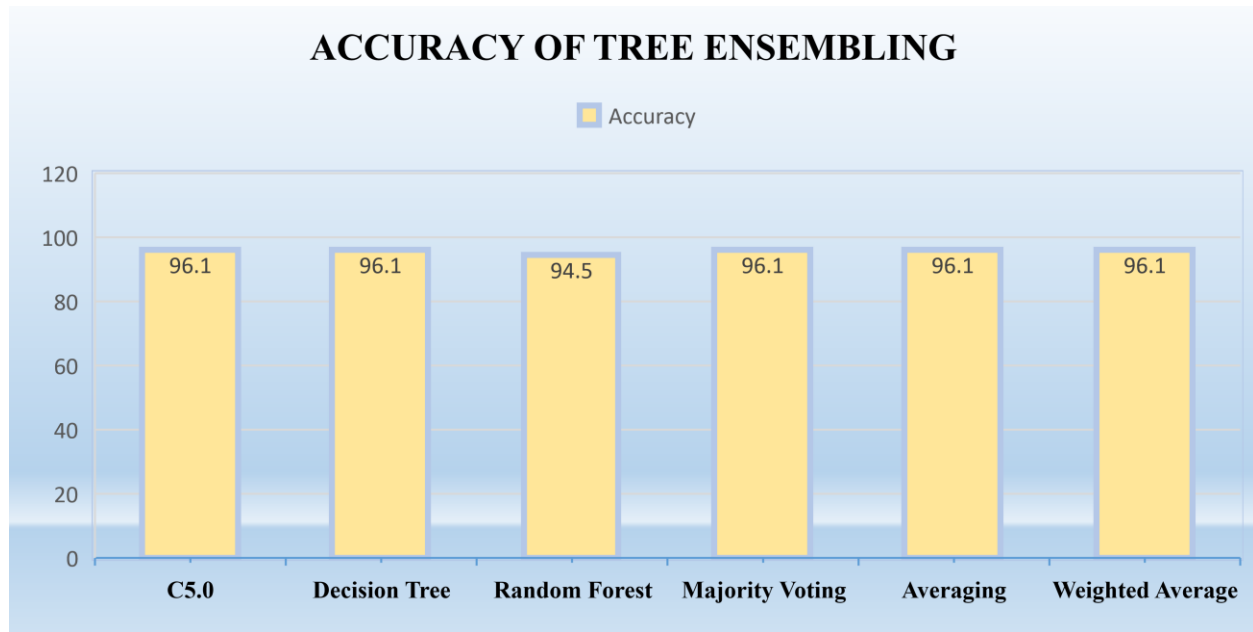
Ensemble with Important attributes 		
S.no	Algorithm	Accuracy
1	Random Forest	0.937
2	Logistic Regression	0.945
3	Gradient Boosting	0.953
Averaging		0.953
Majority Voting		0.953
Weighted Average		0.953

Ensemble Trees with Important attributes 		
S.no	Algorithm	Accuracy
1	C5.0	0.961
2	Decision Tree	0.961
3	Random Forest	0.945
Averaging		0.961
Majority Voting		0.961
Weighted Average		0.961

CHARTS







7.CONCLUSION

Educational data mining is the interesting field of research for educationalist. With the help of EDM the educational institutions can be benefitted by identifying the weak student's and give adequate training for improving the performance of the student. The classification technique is the popular data mining technique used in Educational Data Mining. This study implemented five classification algorithms which are Decision tree, Random Forest, Logistic Regression, Gradient Boosting and Multilayer Perceptron. The Logistic Regression out performed other algorithms with an accuracy of 87% using Background, Social with an accuracy of 83.5% and Course work attributes with 93%. And Multilayer Perceptron and Decision tree out performed other algorithms with an accuracy of 93% using Important attributes. Ensembling with Gradient Boosting, Random Forest and Logistic Regression gave accuracy of 95%. The Tree Ensembling using Decision tree, Random forest and C5.0 gave accuracy of 96%. As the age increases the failure rate also increases, the female students face less failures compared to male. If the study time is greater than equal to 3 hours the student can escape from failure.

8.REFERENCES

- [1] Kiu, Ching-Chieh. "Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities." *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*. IEEE, 2018.
- [2] Kaunang, Fergie Joanda, and Reymon Rotikan. "Students' Academic Performance Prediction using Data Mining." *2018 Third International Conference on Informatics and Computing (ICIC)*. IEEE, 2018.
- [3] Ajibade, Samuel-Soma M., Nor Bahiah Ahmad, and Siti Mariyam Shamsuddin. "An Heuristic Feature Selection Algorithm to Evaluate Academic Performance of Students." *2019 IEEE 10th Control and System Graduate Research Colloquium (ICSGRC)*. IEEE, 2019.
- [4] Ketui, Nongnuch, Warawut Wisomka, and Kanitha Homjun. "Using Classification Data Mining Techniques for Students Performance Prediction." *2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON)*. IEEE.
- [5] Romero, Cristóbal, et al. "Predicting students' final performance from participation in on-line discussion forums." *Computers & Education* 68 (2013): 458-472.
- [6] Hu, Ya-Han, Chia-Lun Lo, and Sheng-Pao Shih. "Developing early warning systems to predict students' online learning performance." *Computers in Human Behavior* 36 (2014): 469-478.
- [7] Yu, Liang-Chih, et al. "Improving early prediction of academic failure using sentiment analysis on self-evaluated comments." *Journal of Computer Assisted Learning* 34.4 (2018): 358-365.
- [8] Deepika, K., and N. Sathvanaravana. "Analyze and Predicting the Student Academic Performance Using Data Mining Tools." *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2018.
- [9] Uzel, Vahide Nida, Sultan Sevgi Turgut, and Selma Ayşe Özel. "Prediction of Students' Academic Success Using Data Mining Methods." *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 2018.
- [10] Siddiqui, Isma Farrah, and Qasim Ali Arain. "ANALYZING STUDENTS' ACADEMIC PERFORMANCE THROUGH EDUCATIONAL DATA MINING." *3C Tecnologia* (2019).
- [11] Rawat, Keshav Singh, and I. V. Malhan. "A Hybrid Classification Method Based on Machine Learning Classifiers to Predict Performance in Educational Data Mining." *Proceedings of 2nd International Conference on Communication, Computing and Networking*. Springer, Singapore, 2019.