ANALYSIS OF THE IMPACT OF SOCIAL AND EMOTIONAL FACTORS ON STUDENT PERFORMANCE USING DATA MINING TECHNIQUES

[1]Polam Bharat Kumar, [2]Vikas Boddu
[1] UG Student, [2]Assistant Professor
[12] Computer Science and Engineering, GITAM Institute of Technology
[12] GITAM, Visakhapatnam, India.

Abstract : In present world's education system, secondary school education is considered as the main phase in the student's education life and high importance is given to it. Although, importance for the secondary school education is high many students are failing due to many reasons. Taking this into consideration many educational institutes are acquainting new methods to make students develop their grades. Educational Data Mining is a booming field of research which is used to predict the student's performance. Many data mining classification algorithms are used to predict the performance of the students which help the management to take necessary measures for the development of the student. Non-academic factors like parent's job, student activities, family relationships are considered. Data Mining models like Naive Bayes, Neural Networks, Support Vector Machine, Decision Tree, Random Forest are used to predict student's performance. Various features are evaluated using Data Mining models and the performance of each student is assessed. As the outcome of this research, necessary guidance is given to the students whose performance is low compared to others. This develops the education quality and standards of the school.

IndexTerms **- Data Mining, Classification, Prediction, Performance.**

I. INTRODUCTION

Education has been considered as an essential need in a person's life. Knowledge from raw data sets are obtained by using different tools offered by Business intelligence and Data Mining which aid the educational institutes[1]. Every educational institute wants its students to perform well. A country's economic growth relies on the success of its student's performance in their academics [2]. Many methods and programs are being ensued to decrease the failure rate. So, there is drastic use of data mining techniques by many educational institutions [3].
Educational data mining has been a boon to the educational institutes as it can be used to predict many useful patterns and extract information [4] about the students such as their performance and the factors that effect their education.

In this paper, we use various data mining algorithms like Naive Bayes, Neural Networks, Decision Tree, Random Forest, Support Vector Machine to predict whether the student pass or fail in the Term3. We then perform boosting on the dataset to increase the accuracy of the prediction and for the better performance of the DM algorithms [5]. Here, we consider three cases. In the first case A, we consider Term1 and Term2 marks and only Term1 marks in the second case B and neither Term1 nor Term2in the third case C. Thus, we evaluate by considering various features.

Paper is organized as follows. Section II describes about the related work that has been done related to Educational Data Mining by many other researchers. Section III describes about the methodology which include collection of data, data processing, data description, data visualization and attribute selection methods. Section IV describes about the result tables which include before and after boosting values. Finally, Section V gives the conclusion of the paper.

II. RELATED WORK

As Educational Data Mining is an emerging technology, many researchers have done a lot of research in this field comparing different feature selection algorithms.

Minaei-Bidgoli B, Kashy D, Kortemeyer G, and Punch W, 2003, predicted the final grade of the students by using the features that are extracted from logged data in a educational web based system. He used genetic algorithm to improve the prediction [6].

Ma Y, Liu B, Wong C, Yu P. and Lee S, 2000, In Proc. of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, USA, 457–464, used Association rule mining to target a student who is weak in a single or more subject [7].

Kotsiantis S, Pierrakeas C, and Pintelas P, 2004, used classification algorithms to predict the performance in distance learning [8].

Cortez, P.A.Silva. Using Data Mining to Predict Secondary School Student Performance. EUROSIS. A. Brito and J. Teixeira, Eds. 2008, 5-12, attempted to predict the failure by using classification algorithms like Decision Tree, Random Forest, Neural Networks, SVM [9].

III. RESEARCH METHODOLOGY

3.1 Collection of Data:

The data set that is used in this paper is taken from UCI Repository. We have taken two data sets of students of two high schools in Alentejo region of Portugal. The data sets are of two subjects namely, Mathematics and Portugal. This data set contains other social and emotional features[10] such as family size, mother's job, father's job, health, internet etc. It also has marks of term1(G1), term2(G2) and term3(G3). There are a total of 396 student data for the Mathematics subject and 650 students for Portugal subject which are then modelled into two categories that is Binary(pass/fail) and the second one is 5-level classification(from A very good or excellent to F - insufficient).

3.2 Processing of Data:

The data set is pre-processed after collecting it as it has many unnecessary and blank data in it. So, it goes through a series of steps in the pre-processing like Data Reduction, Data Cleaning, Data Discretization, Data Transformation, Data Cleaning and Data Integration where the incorrect data is corrected, converting or partitioning continuous attributes, features or variables to discrete or nominal attributes/features, numerical or alphabetical digital information is transformed into corrected, ordered and simplified form.

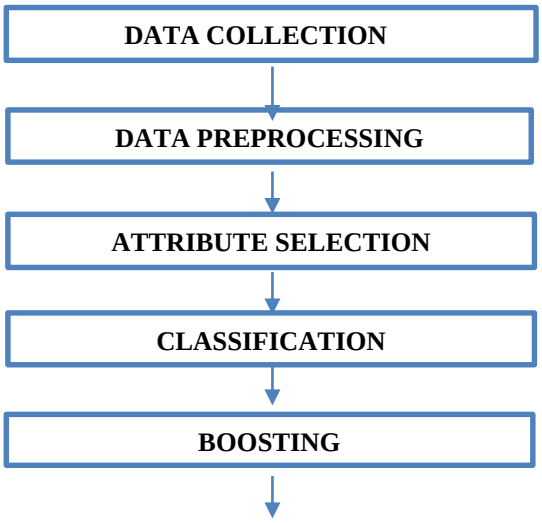
3.3 Data Description:

This section gives the brief description of the attributes in the dataset. We did not mention some attributes like reason, travel time, activities, family education support and many more which does not affect much compared to the following attributes.

ATTRIBUTE	DESCRIPTION
Sex	Student's sex
Activities	Extra circular activities
Fjob	Father's job
Fedu	Father's education
Medu	Mother's education
Mjob	Mother's job
Age	Student's age
School	Student's school
Famsize	Family size
Famrel	Family relation status
Failures	Failures in the past
Study time	Extra study time
Internet	If he uses internet
Walc	If he drinks alcohol
Health	Present health condition

Table 1: Description of attributes

The below given figure explains about the step by step process of the research



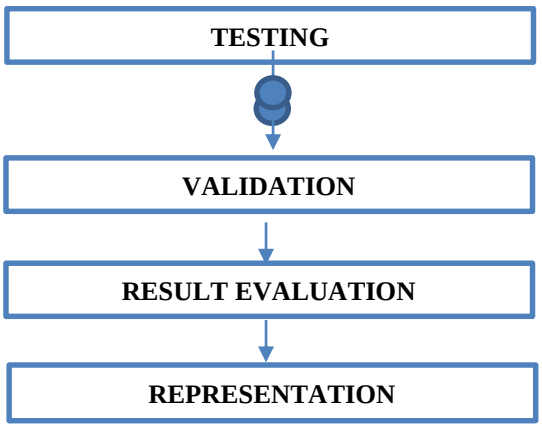


Figure 1: Steps in the methodology

3.4 Visualization of Data:

Data visualization is a process where we represent the data which is in the raw form into graphical or pictorial representation for a better understanding and analysis. There are many ways to visualize the data such as Bar graphs, Pie charts, Histogram, Plot graphs etc.

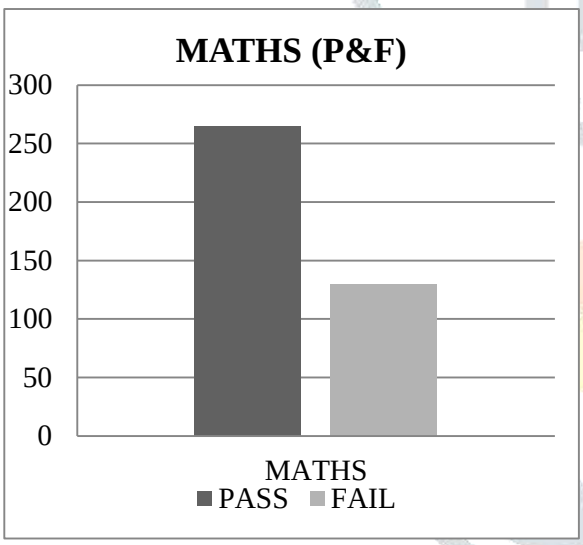


Figure 2: Mathematics fail and pass.

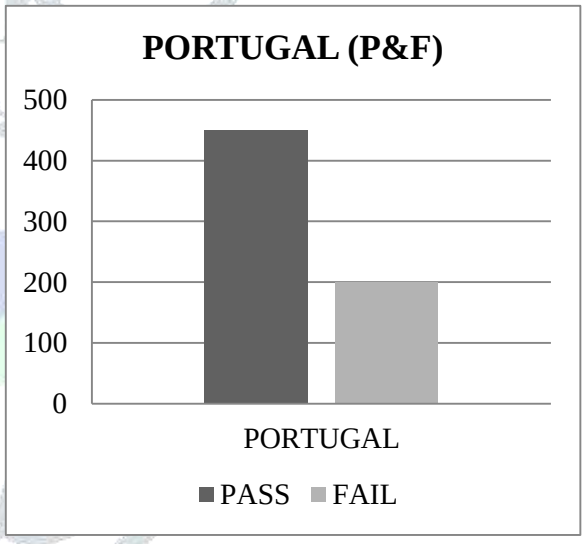
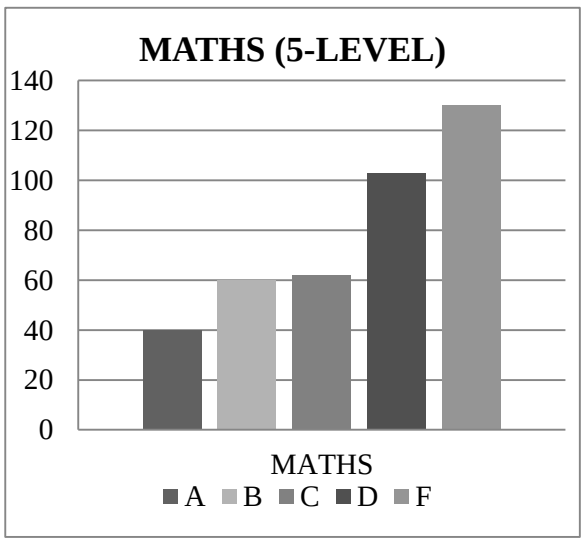
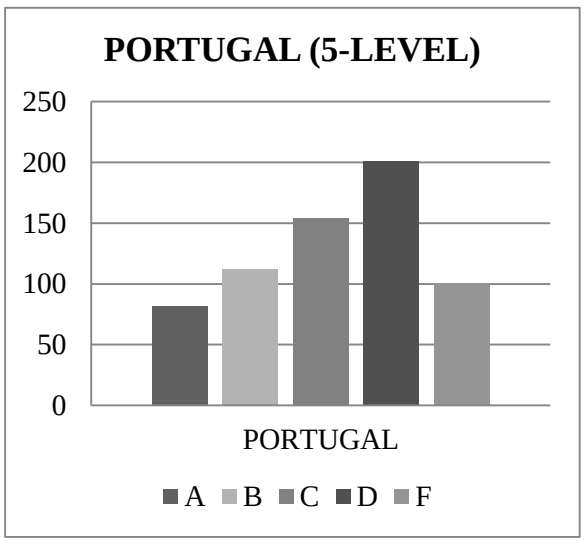


Figure 3: Portugal fail and pass.

The graph in Fig.2 represents the number of students that passed and failed in the mathematics subject in the form of a binary model(P&F). We have 265 students that had passed in the subject and 130 students that failed in the subject. The graph in Fig.3





represents the number of students failed and passed in the Portugal subject in the binary model(P&F). From the graph we can say that there are 450 students passed the subject and 199 students that got failed in the subject.

Figure 4 Mathematics 5-level model **Figure 5 Portugal 5-level model**

The graph in Fig.4 is a 5-level model representation of the pass and fail students in the mathematics subject. In this representation there are 5 levels A, B, C, D and F. So, we have 40 students in group A, 60 students in group B, 62 students in group C, 103 students in group D and 130 students in group F. The graph in Fig.5 is a 5-level model representation of the pass and fail students in the Portugal subject. In this representation there are 5 levels A, B, C, D and F. So, we have 82 students in group A, 112 students in group B, 154 students in group C, 201 students in group D and 101 students in group F.

GRADES	MARKS
A	16-20
B	14-15
C	12-13
D	10-11
F	0-9

Table 2: 5-level model description

A-Excellent B-Good C-Satisfactory D-Satisfactory F-Fail

3.5 Attribute selection:

Attribute or Feature selection is a method of reducing the number of attributes by selecting the attributes that are more helpful in increasing the accuracy of the classifier. So, in this way by choosing the features that help in giving better results we can make an accurate predictive model.
In this paper we have used three attribute selection algorithms namely ReliefAttributeEval which is aided with Ranker search method, Principal Component Analysis which is also aided with Ranker search method and the last one is CfsSubsetEval which is aided with Best First Search.

After the feature selection process we will have a clean data set when compared to before. Now we use the classification algorithms to predict the outcomes and accuracy is measured. We used Naive Bayes, Neural Network, Support Vector Machine, Decision Tree, Random Forest classification algorithms in this paper. We apply Boosting method to increase the accuracy percentage of each classifier.

WEKA tool is used to predict the outcomes and measure the accuracy of the classification algorithm and apply Boosting to it [11].

IV. RESULTS AND DISCUSSION

4.1 Principal Component Analysis:

Curse of Dimensionality is where a classifier tend to overfit the training dataset in a high dimension. So, there should be a feature selection algorithm to select which feature should be removed and which feature should be used.
Principal Component Analysis is an algorithm where all the correlated features are divided into uncorrelated features called Principal Components. This help in removing the Curse Of Dimensionality theory.

	A		B		C	
	ACTUAL	BOOSTED	ACTUAL	BOOSTED	ACTUAL	BOOSTED
NV	81.2%	83.2%	76.4%	77.4%	69%	69.8%
NN	88.8%	89.1%	80%	80.7%	67.3%	67.3%
SVM	87.6%	90.8%	78.9%	79.4%	71.1%	71.4%
DT	75.4%	76.4%	60.7%	68.8%	62.5%	65%
RF	78.7%	79.2%	71.3%	72.9%	67.5%	68.3%

Table 3: Mathematics Binary model Performance evaluation.

	A		B		C	
	ACTUAL	BOOSTED	ACTUAL	BOOSTED	ACTUAL	BOOSTED
NV	81%	81.2%	80.4%	81%	74.5%	75.6%
NN	84.8%	85%	82.2%	84.1%	77%	78.4%
SVM	86.2%	86.4%	82.2%	84.1%	79.8%	79.9%
DT	76.4%	81.3%	72.7%	78.2%	77.9%	77.9%
RF	82.7%	83.2%	80.1%	81.2%	78.2%	78.3%

Table 4: Portugal Binary model Performance evaluation.

	A		B		C	
	ACTUAL	**BOOSTED**	**ACTUAL**	**BOOSTED**	**ACTUAL**	**BOOSTED**
NV	44.5%	48%	40.5%	41.7%	30.6%	30.6%
NN	58.9%	59.7%	43.7%	44%	29.3%	32.9%
SVM	56.7%	66.4%	41.2%	41.2%	30.6%	30.6%
DT	40.2%	44.5%	28.30%	33.10%	27.5%	28.3%
RF	44.5%	46.5%	39%	39.40%	30.6%	35.6%

Table 5: Mathematics 5-Level model Performance evaluation.

	A		B		C	
	ACTUAL	**BOOSTED**	**ACTUAL**	**BOOSTED**	**ACTUAL**	**BOOSTED**
NV	47.7%	48%	41.7%	41.7%	33.1%	33.1%
NN	60.7%	61.6%	41.4%	45.4%	31.8%	32.2%
SVM	60.4%	64.2%	48.6%	48.6%	32.2%	35.4%
DT	42.6%	47.4%	32.8%	36.8%	30.5%	31.2%
RF	48.3%	48.5%	43%	43.4%	36.2%	36.6%

Table 6: Portugal 5-Level model Performance evaluation

The above tables are the results before boosting and results after boosting for each classifier with three cases A, B, C. By analysing the above table's we can conclude that we can get the highest accuracy with the case A where we consider both the term exams and Support Vector Machine algorithm derives the more accuracy compared to other classifiers.

4.2 ReliefAttributeEval:

ReliefAttributeEval is a feature selection algorithm where each feature is assigned with a score which is also called weights and selects the feature which has the highest score. This is mainly meant for binary classification with discrete values or numeric values. It filters the features and selects the best from of it, so it is called filter based approach.

	A		B		C	
	ACTUAL	**BOOSTED**	**ACTUAL**	**BOOSTED**	**ACTUAL**	**BOOSTED**
NV	86.3%	88.8%	79.7%	81.7%	67%	69.1%
NN	88.8%	88.8%	81%	81.7%	65.3%	65.8%
SVM	89.6%	89.8%	81%	81.7%	68.1%	68.5%
DT	89.3%	90.6%	84.3%	84.5%	66.3%	68.6%
RF	90.6%	91.6%	83.7%	84.3%	69.1%	69.8%

Table 7: Mathematics Binary model Performance evaluation

	A		B		C	
	ACTUAL	**BOOSTED**	**ACTUAL**	**BOOSTED**	**ACTUAL**	**BOOSTED**
NV	85.3%	86.4%	90.4%	90.9%	78.7%	79.1%
NN	85.2%	85.8%	81.5%	82.4%	73.3%	73.3%
SVM	85.2%	86.1%	84.5%	84.5%	78.4%	78.7%
DT	88%	89%	85%	85.2%	76.1%	77.4%
RF	90.4%	90.9%	84.7%	85.3%	78.2%	78.2%

Table 8: Portugal Binary model Performance evaluation

	A		B		C	
	ACTUAL	**BOOSTED**	**ACTUAL**	**BOOSTED**	**ACTUAL**	**BOOSTED**
NV	70.3%	71.2%	52.4%	53.4%	30.8%	30.8%
NN	60.7%	61%	46.8%	47.8%	31.8%	32%
SVM	54.1%	59.4%	45.5%	46%	30.6%	31.1%
DT	71.9%	72%	52.9%	53.6%	27.5%	32.6%
RF	72.5%	74.6%	49%	49.4%	33.6%	34.1%

Table 9: Mathematics 5-Level model Performance evaluation

	A		B		C	
	ACTUAL	**BOOSTED**	**ACTUAL**	**BOOSTED**	**ACTUAL**	**BOOSTED**
NV	68.2%	68.2%	53.4%	53.4%	35.9%	40%
NN	57.3%	57.9%	47%	47.4%	30.5%	36%
SVM	56%	62.2%	47.1%	48.5%	35.4%	35.5%
DT	68.2%	69%	54.1%	54.8%	28.1%	33.4%

RF	72.7%	73%	55%	55.4%	34.3%	35%

Table 10: Portugal 5-Level model Performance evaluation

The above tables are the results before boosting and results after boosting for each classifier with three cases A, B, C. By analyzing the above tables we can conclude that, we can get the highest accuracy with the case A where we consider both the term exams and Random Forest algorithm derives the best accuracy compared to other classifiers.

4.3 CfsSubsetEval:

Set of attributes are selected through a correlated based approach using a feature selection algorithm. The subsets which are highly correlated with the class are given the most priority. It comes under wrapper class.

	A		B		C	
	ACTUAL	**BOOSTED**	**ACTUAL**	**BOOSTED**	**ACTUAL**	**BOOSTED**
NV	86.5%	89.3%	85.3%	85.3%	72.1%	72.5%
NN	88.6%	89.1%	84%	84%	69.6%	69.8%
SVM	89%	89.6%	84.3%	84.3%	70.6%	71.1%
DT	91.8%	92%	83.2%	84.1%	70.6%	70.6%
RF	90%	91.2%	82.7%	83%	69%	70.3%

Table 11: Mathematics Binary model Performance evaluation

	A		B		C	
	ACTUAL	**BOOSTED**	**ACTUAL**	**BOOSTED**	**ACTUAL**	**BOOSTED**
NV	88.7%	90.4%	83.8%	83.9%	79.6%	79.6%
NN	89.3%	89.5%	84.4%	84.4%	78.4%	78.5%
SVM	88.1%	89%	85.2%	85.2%	79.8%	80.1%
DT	88.1%	89.2%	85.6%	86%	77.1%	77.9%
RF	90.1%	91.3%	84.2%	85.1%	78.2%	78.2%

Table 12: Portugal Binary model Performance evaluation

	A		B		C	
	ACTUAL	**BOOSTED**	**ACTUAL**	**BOOSTED**	**ACTUAL**	**BOOSTED**
NV	70.6%	70.6%	59.2%	59.2%	33.9%	33.9%
NN	73.4%	74.4%	57.4%	57.6%	34.1%	34.1%
SVM	66.3%	67.5%	54.6%	55.4%	34.9%	34.9%
DT	76.4%	76.9%	54.1%	54.1%	36.4%	36.4%
RF	71.8%	72.4%	53%	55.6%	34.1%	34.5%

Table 13: Mathematics 5-Level model Performance evaluation

	A		B		C	
	ACTUAL	**BOOSTED**	**ACTUAL**	**BOOSTED**	**ACTUAL**	**BOOSTED**
NV	71.4%	71.4%	58.5%	58.5%	34%	34%
NN	72.7%	72.8%	57.4%	58.3%	34.6%	34.6%
SVM	69.4%	69.4%	51.3%	53.7%	33.8%	33.8%
DT	75.3%	76%	62.2%	62.5%	34.5%	34.6%
RF	71%	72%	56.8%	57.1%	31.7%	32.5%

Table 14: Portugal 5-Level model Performance evaluation

The above tables are the results before boosting and results after boosting for each classifier with three cases A, B, C. By analyzing the above tables we can conclude that, we can get the highest accuracy with the case A where we consider both the term exams and Decision Tree algorithm derives the best accuracy compared to other classifiers.

V. RESULTS AND DISCUSSION

Educational Data Mining is booming in the world of education as it is helping the educational institutes in improving their student's performance. Institutes can predict their student's performance and help them in improving their marks in the upcoming exam by providing them with extra support.

In this paper, we used various classifiers like Naive Bayes, Neural Networks, Support Vector Machine, Decision Tree, Random Forest and later applied an ensemble method called Boosting, which helps in increasing the accuracy of the classifier. After we

evaluate the results we can observe that the Binary model derives high accuracy compared to 5-Level model and the case A derives more accuracy compared to B and C.

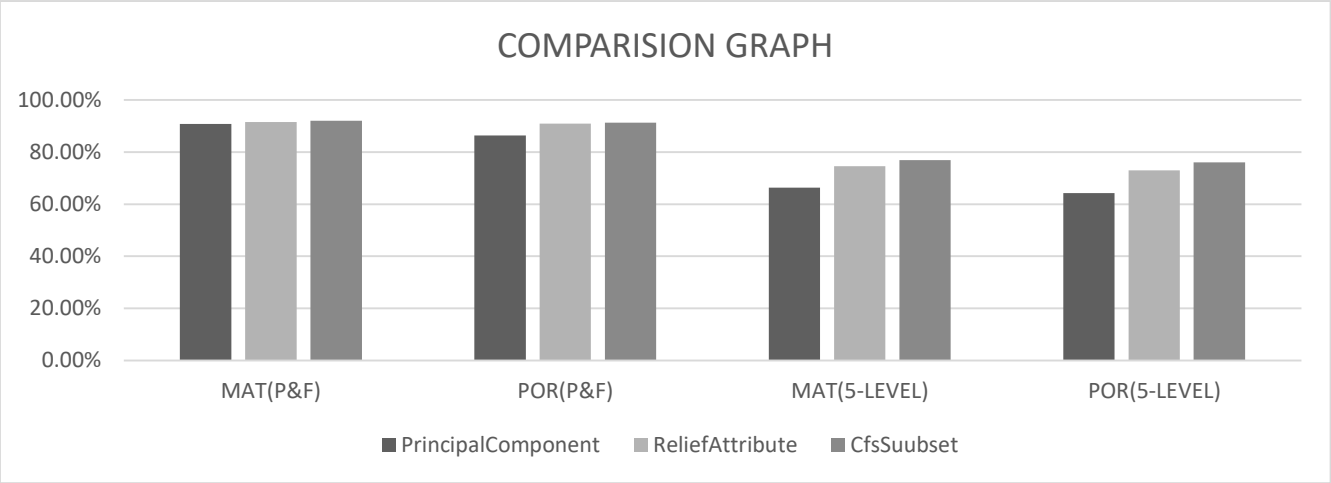


Figure 6: Attribute Comparison Graph.

We can observe that the actual accuracies got increased when we do boosting to it. In the subject of Mathematics, accuracy of 87.6% is derived in Principal Component Analysis using Support Vector Machine which is increased to 90.8% after applying boosting. Accuracy of 90.6% is obtained in ReliefAttributeEval using Random Forest which is later increased by applying boosting to 91.6% and accuracy of 91.8%is obtained in CfsSubsetEval using Decision Tree which is later increased by applying boosting to 92%. On the other hand, in the subject of Portugal, accuracy of 86.2% is derived in Principal Component Analysis using Support Vector Machine which is increased to 86.4% after applying boosting. Accuracy of 90.4% is obtained in ReliefAttributeEval using Random Forest which is later increased by applying boosting to 90.9% and accuracy of 90.1%is obtained in CfsSubsetEval using Random Forest which is later increased by applying boosting to 91.3%.
From the above results we can derive a conclusion that in Mathematics subject we get the highest accuracy of 92% in CfsSubsetEval using Decision Tree and in the subject Portugal we get the highest accuracy of 91.3% in CfsSubsetEval using Random Forest. This study can be further enhanced by adding other emotional and social features and applying other ensemble method to increase the accuracy and derive other interesting patterns.

VI. Acknowledgment

Sincere gratitude is hereby extended to Prof K.Thammi Reddy, Professor and Head , Dept. of Computer Science and Engineering, GIT,GITAM for providing the necessary resources and for the encouragement given by him for completing this work.

REFERENCES

[1] Paulo Cortez and Alice Silva,Using Data Mining To Predict Secondary School Student Performance,(2008).
[2] Gaviria, A. Los quesuben y los quebajan: educacion y vilidad social en Colombia. Fedesarrollo, Alfaomega.(2002).
[3] A. M. Shahiri and W. Husain, A review on predicting student's performance using data mining techniques, Procedia Computer Science,72:414-422, 2015.
[4] E. Osmanbegović, M. Suljić, and H. Agić, Determining Dominant Factor For Students Performance Prediction By Using Data Mining Classification Algorithms, Tranzicija,16:147-158, 2015.
[5] E.Prabhakar and Dr.C.Nalini, Boosted Adaboost to Improve the Classification Accuracy(2012).
[6] Minaei-Bidgoli B.; Kashy D.; Kortemeyer G.; and Punch W., 2003. Predicting student performance: an application of data mining methods with an educational web-based system. In Proc. of IEEE Frontiers in Education. Colorado, USA, 13–18.
[7] Ma Y.; Liu B.; Wong C.; Yu P.; and Lee S., 2000. Targeting the right students using data mining. In Proc. of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, USA, 457–464.
[8] Kotsiantis S.; Pierrakeas C.; and Pintelas P., 2004. Predicting Students' Performance in Distance Learning Using Machine Learning Techniques. Applied Artificial Intelligence (AAI), 18, no. 5, 411–426.
[9] Cortez, P., A.Silva. Using Data Mining to Predict Secondary School Student Performance. EUROSIS. A. Brito and J. Teixeira, Eds. 2008, 5-12.
[10] Pritchard M. and Wilson S., 2003. Using Emotional and Social Factors To Predict Student Success. Journal of College Student Development, 44, no. 1, 18–28.
[11] M. Hall, E. Frrank, G.Holmes, B.Pfahringer, P. Reutemann, and I. H. Witten, The WEKA data mining software: an update, ACM SIGKDD exploration newsletter, 11:10-18,2009.