Analysis of Student's Academic Performance Using Associative Classification

Illagiya Vijayalakshmi D[1], **Sowmya T**[2], **Sownthariyaa G**[3]
[1, 2, 3] Department of Information Technology
[1, 2, 3] Dr. Mahalingam College Engineering and Technology, Pollachi-642003, Tamil Nadu, India

Abstract- *Data mining is the most powerful tool for the analysis and the detailed extraction of hidden information from the dataset. This project works in the field of educational data mining in which the student's performance is analyzed using weighted ensemble algorithm. The single classification algorithms such as j48, decision stump, naive Bayes, IBK, simple cart, rep tree is applied for the dataset and weights are allocated for each instances based on their accuracy obtained using netbeans ide. The obtained instance of classification is then associated using the associative classification [ASC] algorithm in the user friendly interface of .NET Framework aiming at better analysis with improved accuracy of data. The analyzed data is then clustered based on the algorithmic result. Finally, the performance of the student is measured and the accuracy is improved.*

Keywords- Educational data mining, weighted ensample, j48, decision stump, naïve Bayes.

I. INTRODUCTION

Data Mining is a predominant technology with an ultimate potential to help the organization to focus on the most important information in their datasets. It can also be stated as the detailed and accurate analysis of complex datasets, thereby discovering the significant patterns which might be left unnoticed.

Educational Data Mining

Educational data mining refers to mining of data applied in the educational field. It is an emerging technology, which identifies the uniqueness of the dataset. It also relates to the identification of hidden patterns and relationships which enhances the interests and focus of the students in their academics.

Classification in Educational Data Mining

Data Mining deals with techniques such as classification and prediction in order to extract the data models which create the pattern for future analysis of data. Classification deals with the prediction of results for the target class. This prediction is done with the help of test data through which the training pattern is been created. This training pattern is then applied to the certain set of data to extract the targeted result. The extracted result is then compared with the sample result.

The algorithm discovers a relationship between the attributes which would make it possible to predict the outcome. Next the algorithm is given to a data set that not seen before, called prediction set, it holds the same set of attributes, except the prediction attribute which is not yet known. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how good the algorithm is.

Classification is a two-step process.

STEP 1: A classifier is build describing a predetermined set of data classes or concepts. This is the learning step, where classification algorithm built the classifier by analyzing or learning from a training set made up of database tuples and their associated class labels.

STEP 2: The model is used for classification. The predictive accuracy of classifier is estimated. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.

Algorithms in Classification:

Decision Tree:

Decision tree is a tree structure with flow chart, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test and each leaf node (terminal node) holds a class label. The topmost node in the tree is the root node. A choice between a number of alternatives, and each leaf node represents a decision is branch node in decision tree.

Decision tree are commonly used for gaining information for the purpose of decision -making. Decision tree starts with a root node on which it is for users to take actions. From this node according to decision tree learning algorithm, users split each node recursively. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome.

Ensemble

Ensemble algorithm can be trained and then used to make predictions, it is called supervised learning algorithm. The trained ensemble, therefore, represents a single hypothesis. Combination of models to increase the accuracy of the classification is used as ensemble classifier. It combines a series of learning models with the aim of creating new improved model. These ensemble classifiers are based on the weighting scheme. Basically there are two types of ensemble classifiers. They are boosting and bagging.

Boosting

Boosting is a ensemble based classifier which can improve the performance of any weak classifier. Boosting assigns a weight to each classifiers vote, based on how well the classifier performed. The lower a classifiers error rate, the more accurate it is. The class with the highest sum is the winner and is returned as the class prediction. Adaboost can significantly enhance the accuracy of classifier algorithm such as neural network learning and decision tree learning. Bagging is another ensemble method that averaging the prediction over a collection of classifiers. It is significantly better than a single classifier.

Bagging

The bagging algorithm each model gives an equally weighted prediction by creating an ensemble of models for a learning scheme. Sampled with replacement from the original set of tuples for each iteration, tuples is. Bagged classifier, the original training data has greater accuracy than a single classifier. It will not be considerably worse and is more robust to the effects of noisy data. The composite model reduces the variance of the individual classifiers since increased accuracy occurs.

II. LITERATURE REVIEW

1]. Classification Model of Prediction for Placement of Students

Author : Ajay Kumar Pal, Saurabh Pal
Concept: Classification approach to find an enhanced evaluation method for predicting the placement for students. This model can determine the relations between academic achievement of students and their placement in campus selection

2].Performance Analysis of Data Mining Techniques for Placement Chance Prediction

Author : V.Ramesh, P.Parkavi, P.Yasodha
Concept: Predicting the performance of a student is a great concern to the higher education managements. The scope of this paper is to investigate the accuracy of data mining techniques

3].Mining Educational Data for Students' Placement Prediction using Sum of Difference Method

Author : Ramanathan.L, Swarnalatha.P, Ganesh Gopal.D
Concept: Using the attributes such as academic records, age, and achievement etc., EDM has been used for predicting the performance about placement of final year students. Based on the result, higher education organizations can offer superior education to its students.

4].Association Rule Generation for Student Performance Analysis using Apriori Algorithm

Author : D. Magdalene Delighta Angeline
Concept: Students are classified based on their involvement in doing assignment, internal assessment tests, attendance etc., which helps to predict the performance of the student based on the pattern extracted from the educational database. This would help to identify the average and below average students and to improve their performance to provide good results. This analysis further helps matching organizations requirement with students profile to provide placement for the students.

III. PROBLEM DEFINITION

The numerous data leads to lots of confusion while predicting the result of data for future processing. The large amounts of student dataset are available and it is difficult to analyze those data and obtain accurate result for it. The classification algorithm of weighted ensemble is applied for the data source and categorization of students from data source is done. But obtaining high accuracy for the result is a big challenge, to overcome this, the algorithm of [ASC] Associative Classification is developed to obtain the maximum accuracy for the student dataset.

IV. PROPOSED SYSTEM

The weighted ensemble algorithm is used for applying classification to the student academic performance dataset. The ensemble algorithm possessing single classifier algorithm such as j48, decision stump, naive Bayes, ibk, simple cart, rep tree are applied to extract the overall accuracy, accuracy for each classes referred in the Table 4.1 and the instance of the dataset. The extracted result is given as input for the ASC algorithm, which works by allocating the weights for the result acquired by the ensemble algorithm. This ASC algorithm aims in increasing the accuracy to be as high as possible and the results are analyzed.

AdaBoost Algorithm:

Input:

D, a set of d class-labeled training tuples;
k, the number of rounds (one classifier is generated per round);
a classification learning scheme.

Output:

Confusion Matrix

Procedure:

```
Step 1: initialize the weight of each tuple in D to 1=d;
Step 2: for i D 1 to k do // for each round:
Step 3: sample D with replacement according to the tuple weights to obtain Di ;
Step 4: use training set Di to derive a model, Mi;
Step 5: compute error in Mi // the error rate of Mi
Step 6: repeat until n classifier
Step 7: weight for the classifier is initially assigned to 1
Step 8: then the weight is assigned based on the accuracy provided by the classifier
Step 9: calculate confusion matrix and accuracy
```

Table 4.1

CLASSES	REFERRED TO
Class 1	NEEDS IMPROVEMENT
Class 2	AVERAGE
Class 3	EXCELLENT

Modules Description

The project involves three different modules such as:
1]. Data Selection and Preprocessing
2]. Applying Associative Classification [ASC]
3]. Analysis

Data Selection and Preprocessing

The module1 deals with preprocessing of the data collected i.e. removing the noisy data, filling the missing data or the incorrect data and thereby cleansing it for the further use. This is the primary data construction model dealt in the data mining.

This module deals with selecting the appropriate data from the entire dataset. This module also includes the replacement of missing data, filling the data with the constant term/value, correcting the noisy data and thereby reducing the flaws in the data.

The sample dataset of student's academic performance is collected from the following website: https://archive.ics.uci.edu/ml/datasets.html

The dataset contains the following attributes list as:

- School
- Gender
- Age
- Address
- Family_size
- Mother_education
- Father_education
- Mother_job
- Father_job
- Reason
- Guardian
- Travel_time
- Study_time
- Failure
- Extracurricular
- Higher_studies
- Free_time
- Internet
- Absences
- Work_time
- Grade

Implementation of Ensemble Algorithm

The preprocessed data models are then imported in .arff format in net beans IDE that is connected with weka for implementing the single classifier algorithm. The ensemble algorithm is found under the tree package of classifier. The processed results are then exported into the excel file. These resultant data are then imported to find the association between the filtered dataset.

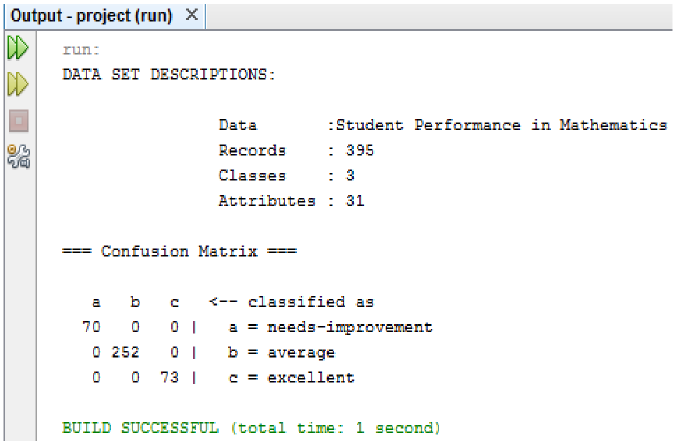
Applying Associative Classification [ASC]

The ASC algorithm is applied to the filtered dataset in which the weights are allocated for the result acquired and analyzed to predict the output. The result obtained from the applied classification algorithm produce incorrect accuracy. So in order obtain the improvised accuracy, further association is applied to the results. As a result of applying associative classification, the aim of achieving improved accuracy of the given dataset is done successfully.

Analysis

The filtered dataset are analyzed for its accuracy and can be separated into different categories as needed for the training process done by the organization. This helps the students in way of approaching their preparation and improvising their performance to attain the next level.

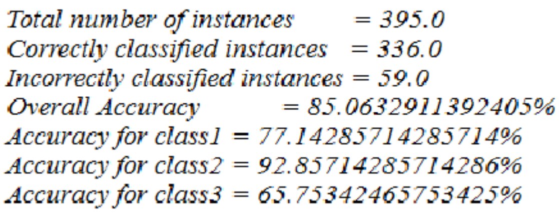
V. PERFORMANCE METRICS



```
run:
DATA SET DESCRIPTIONS:

                    Data        :Student Performance in Mathematics
                    Records     : 395
                    Classes     : 3
                    Attributes  : 31

=== Confusion Matrix ===

   a   b   c   <-- classified as
  70   0   0 |   a = needs-improvement
   0 252   0 |   b = average
   0   0  73 |   c = excellent

BUILD SUCCESSFUL (total time: 1 second)
```

Figure 5.1

The figure 5.1 represents the result of applying REP Tree classification algorithm to the dataset which produces the confusion matrix with the classes specified as needs improvement, average, excellent and with the attributes details.



Total number of instances = 395.0
Correctly classified instances = 336.0
Incorrectly classified instances = 59.0
Overall Accuracy = 85.0632911392405%
Accuracy for class1 = 77.14285714285714%
Accuracy for class2 = 92.85714285714286%
Accuracy for class3 = 65.75342465753425%

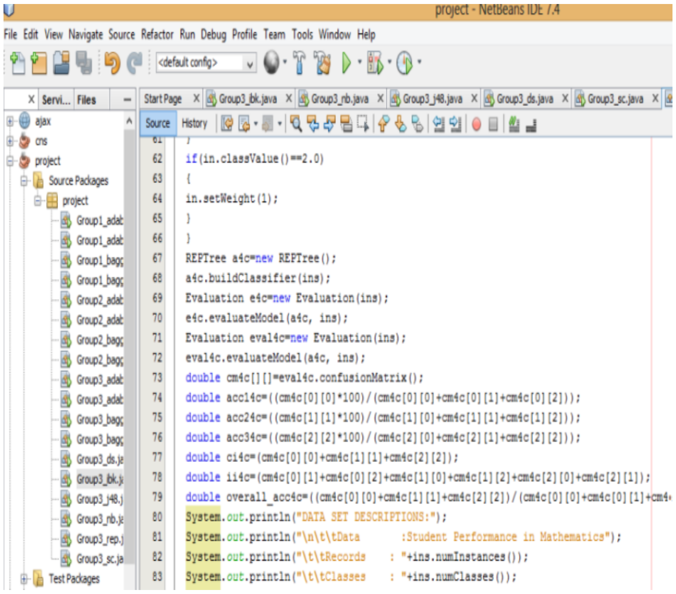
Figure 5.2

The figure 5.2 represents the accuracy details and other specific details on applying REP Tree classification algorithm to the dataset, which is exported to the text file and is then processed to produce the association results.

ALGORITHM	NEEDS IMPROVEMENT [Class-1]	AVERAGE [Class-2]	EXCELLENT [Class-3]
J48	9	9	9
DECISION STUMP	6	1	9
NAIVEBAYES	9	9	9
REP TREE	10	10	10
SIMPLE CART	10	1	10
IBK	10	10	10

Figure 5.3

The figure 5.3 represents the weights allocated by the ASC algorithm for each instances corresponding to their classes in the dataset.

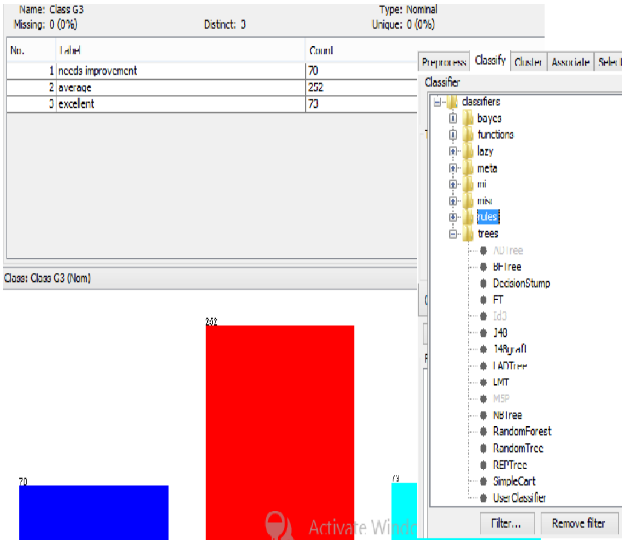


```
}
if(in.classValue()==2.0)
{
in.setWeight(1);
}
}
REPTree a4c=new REPTree();
a4c.buildClassifier(ins);
Evaluation e4c=new Evaluation(ins);
e4c.evaluateModel(a4c, ins);
Evaluation eval4c=new Evaluation(ins);
eval4c.evaluateModel(a4c, ins);
double cm4c[][]=eval4c.confusionMatrix();
double acc14c=((cm4c[0][0]*100)/(cm4c[0][0]+cm4c[0][1]+cm4c[0][2]));
double acc24c=((cm4c[1][1]*100)/(cm4c[1][0]+cm4c[1][1]+cm4c[1][2]));
double acc34c=((cm4c[2][2]*100)/(cm4c[2][0]+cm4c[2][1]+cm4c[2][2]));
double ci4c=(cm4c[0][0]+cm4c[1][1]+cm4c[2][2]);
double ii4c=(cm4c[0][1]+cm4c[0][2]+cm4c[1][0]+cm4c[1][2]+cm4c[2][0]+cm4c[2][1]);
double overall_acc4c=((cm4c[0][0]+cm4c[1][1]+cm4c[2][2])/(cm4c[0][0]+cm4c[0][1]+cm4
System.out.println("DATA SET DESCRIPTIONS:");
System.out.println("\n\t\tData        :Student Performance in Mathematics");
System.out.println("\t\tRecords     : "+ins.numInstances());
System.out.println("\t\tClasses     : "+ins.numClasses());
```

Figure 5.4

The figure 5.4 represents the implementation of REP Tree classification algorithm to the given dataset. Similarly some of the single classifier algorithm such as Decision Stump, J48, Simple Cart, Naïve Bayes, K Nearest Neighbor are applied to the same dataset to produce the ensemble output, which helps in improving the accuracy of the overall algorithm.

VI. RESULTS AND DESCRIPTION



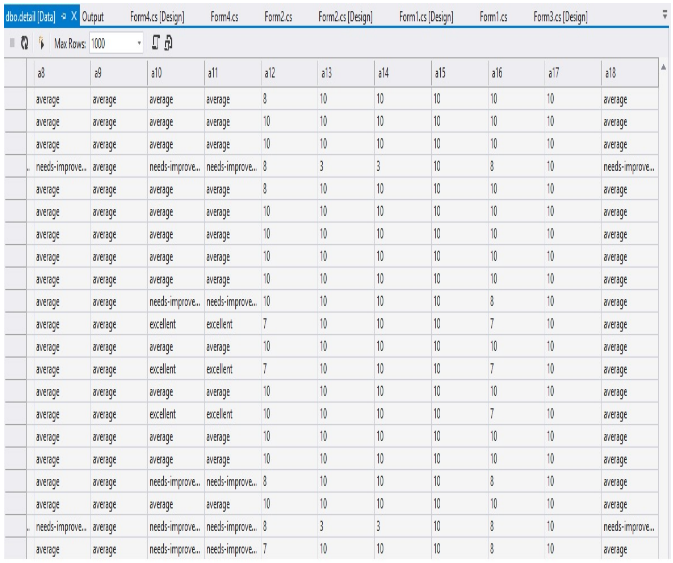
Name: Class G3 Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count
1	needs improvement	70
2	average	252
3	excellent	73

Class: Class G3 (Nom)

Figure 6.1

The figure 6.1 represents some of the single classifier algorithm that can be applied for the given dataset using Weka, the data mining tool.



a8	a9	a10	a11	a12	a13	a14	a15	a16	a17	a18
average	average	average	average	8	10	10	10	10	10	average
average	average	average	average	10	10	10	10	10	10	average
average	average	average	average	10	10	10	10	10	10	average
needs-improve...	average	needs-improve...	needs-improve...	8	3	3	10	8	10	needs-improve...
average	average	average	average	8	10	10	10	10	10	average
average	average	average	average	10	10	10	10	10	10	average
average	average	average	average	10	10	10	10	10	10	average
average	average	average	average	10	10	10	10	10	10	average
average	average	average	average	10	10	10	10	10	10	average
average	average	needs-improve...	needs-improve...	10	10	10	10	8	10	average
average	average	excellent	excellent	7	10	10	10	7	10	average
average	average	average	average	10	10	10	10	10	10	average
average	average	excellent	excellent	7	10	10	10	7	10	average
average	average	average	average	10	10	10	10	10	10	average
average	average	excellent	excellent	10	10	10	10	7	10	average
average	average	average	average	10	10	10	10	10	10	average
average	average	average	average	10	10	10	10	10	10	average
average	average	needs-improve...	needs-improve...	8	10	10	10	8	10	average
average	average	average	average	10	10	10	10	10	10	average
needs-improve...	average	needs-improve...	needs-improve...	8	3	3	10	8	10	needs-improve...
average	average	needs-improve...	needs-improve...	7	10	10	10	8	10	average

Figure 6.2

The figure 6.2 represents the result of six different single classifier algorithms and their prediction which is applied for the given dataset.

VII. CONCLUSION

The purpose of the project is to analyze the student dataset and predict the performance of the test data and apply to the future data for the prediction of output, as well as improving the accuracy of the data predicted. This aim is fulfilled by applying the weighted ensemble algorithm and filtering those results by the associative classification [ASC] algorithm developed. The filtered results are analyzed and verified.

REFERENCES

[1] Romero, C.Ventura, S. and Garcia, "Data mining in course management systems: Model case study and Tutorial". Computers & Education,Vol. 51, No. 1. pp.368-384. 2008.

[2] Samrat Singh and Dr. Vikesh Kumar, “Performance analysis of Engineering Students for Recruitment Using Classification Techniques”, IJCSET February 2013 Vol 3, Issue 2, 31-37

[3] Jai Ruby and Dr. K. David, “Analysis of Influencing Factors in Predicting Students Performance Using MLP – AComparative Study”, 10.15680/ijircce. 2015.0302070.

[4] Ritika Saxena, “Educational data Mining: Performance Evaluation of Decision Tree and Clustering Techniques using WEKA Platform”, International Journal of Computer Science and Business Informatics, MARCH 2015.

[5] Sunita B Aher and Mr. LOBO L.M.R.J, “Data Mining in Educational System using WEKA”, International Conference on Emerging Technology Trends, 2011.

[6] Srecko Natek and Moti Zwilling, “Data Mining for Small Student Data Set – Knowledge Management System for Higher Education Teachers”, Knowledge Management and Innovation, International Conference, 2013.

[7] Dorina Kabakchieva, “Predicting Student Performance byUsing Data Mining Methods for Classification”, Cybernetics And Information Technologies, Volume 13, No 1, 2013.

[8] Bhise R.B., Thorat S.S., and Supekar A.K, “Importance of Data Mining in Higher Education System”, IOSR Journal Of Humanities And Social Science (IOSR-JHSS), Jan-Feb, 2013.