Turkish Journal of Computer and Mathematics Education Vol.12 No.7 (2021), 2415-2426

ResearchArt

Analysis of factors affecting Student performance Evaluation using Education Dataminig Technique

J.Malini
Research scholar, Dep. of information technology
Vels Institute of Science Technology and Advanced Studies
Chennai,Tamil nadu, India.
malinicomp@gmail.com

Dr.Y.Kalpana
Professor, Dep. of information technology
Vels Institute of Science Technology and Advanced Studies
Chennai,Tamil nadu, India.
ykalpanaravi@gmail.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 16 April 2021

Abstract

Every year students success rate was analysed by the Educational Institutions to develop their Academic standard. To identify the success rate many kinds of techniques are used such as statistics, physical examination and currently ongoing datamining techniques. Data mining Techniques was widely used in many fields, it is also used in the Educational environment known as Educational Data Mining (EDM). Educational data mining generate prototype in solving the research problems in students data and used to locate the unseen patterns in the students detailed dateset. This paper uses the EDM to characterize the distinct factors affecting the students performance by making predictions with efficient algorithms. Educational professionals have to identify the causes for the student failure in academic performance and the students not succeed in completing their education which becomes a social problem these days. The machine learning algorithms help the researchers for evaluation of student's learning habits, their academic performance and added enhancement if required. This paper would discuss different kinds of algorithms to analyse the economic background of the students which mainly affects the students performance. The dataset was utilized from the UCI Repository of secondary school students performance and analysed using the Weka tool for the datamining process.

Keywords: Educational datamining, dataset, students performance, attributes, features, Machine learning.

I. INTRODUCTION

The most challenging part of higher education is to ensure that most of the students who come into the education complete their required courses. This is because many get affected in various factors one of the factor discussed in this paper was economic background of students, which act as a important attribute in the students data. Currently many [1] researchers have executed to work on educational datasets to analyse those factors which are the intention for creating issues in student's performance. [2] The core objective is to dig out useful data about students from databases and after deriving the features, find out those

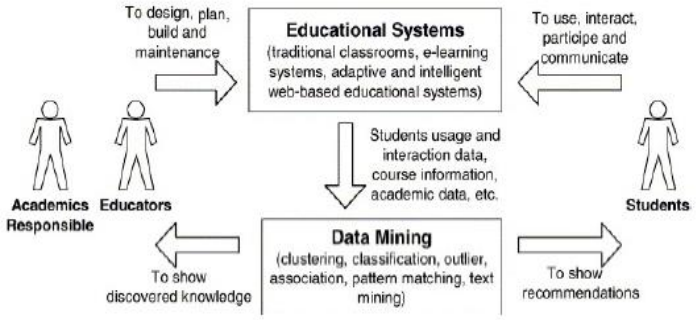


Figure 1 Educational Datamining System

attributes that are disturbing the student's behaviour. Educational Data Mining was applied while using it in developing a model of user thought, work and trial .Figure 1 depicts the Educational datamining system. Mining data or knowledge detection has acquired the fame that it has become the emerging technology as it was very helpful in developing different data models and reduce it into purposeful data[3]. EDM uses many Machine learning techniques like, decision trees, neural networks, k-NN, SVM, naive bayes, and many more. Many open source tools are available for analysing the datasets using Machine learning algorithms such as knime, rapid miner, weka, orange, SSDt (SQL Server data Tools) developed for patterns exploration and to get logical arrangement for later use. In this research paper, WEKA (Waikato Environment for Knowledge Analysis) tool was used for the investigation of data and to create data models to obtain analytical results. To bring the objective , this paper uses a standard dataset from UCI repository of secondary school students.

II. LITERATURE SURVEY

Big data in Education and Educational analytics methods are used in merge Calculus course for early evaluation of students' academic performance by Lu et. al [4]. The results show that seven main factors are affecting the students' performance. From which three were traditional attributes and four of them are online features that disturbs the students' performance.

Suhem Parack et. al [5] used Education data mining in his work to group students data and profiling to predict student performance. Correlations was discovered between set of factors by using Apriori algorithm, then student grouping was estimated using K-means clustering technique by converting a some of observations into subsets.

Romero et. al [6] used the dataset collected from online discussion forum and predicted academic performance of students based on it. Weekly basis the data were separated into data subsets. Many data-mining techniques are applied on the each data subset for predicting the accuracy. The student interaction prior to a midterm exam was predicted using Sequential minimal optimization algorithm to findout the affecting factors of student performance.

An early warning system was developed by [7] to find weak leaners at risk using Decision tree classifier. A data includes of 300 students with 13 online attributes was used to create a model. [8] This author has proposed a student performance prediction with new data features. The performance was predicted using the traditional algorithms of datamining and also included ensemble methods like RF, bagging and boosting algorithms and showed 22.1% improvement. Achieved more than 80% accuracy in the model created. [9] This author has used college students dataset collected from BSIT course information. In that study three algorithms were used to analyse the dataset collected from which is deeplearning classifier predicts 95% accuracy higher than the other two algorithms.This used to investigate the students academic performance. [10]He has introduced a new model to overcome the problem in achieving the performance of students in class. The main feature chosen as the student absence days in class and parents' involvement in the learning process. Then used three datamining algorithms such as ANN, NB, and Decision tree and achieved the accuracy of 10% more than the existing methods.

III. RESEARCH METHODOLOGY

After investigating the data and formulating the objectives of the study, this paper specifies the tools and data cleaning methods, techniques used in this research. This study explains the dataset in the expression of source, type of data and techniques used in that dataset. Moreover, techniques that were applied to the data to prepare it are also described with limitations.

A. DATASET

From the online UCI Repository, the student performance dataset was selected which Predicts student performance in secondary education (high school). Two Portuguese schools' secondary education dataset was found in that data repository. The data features present are social, demographic, student grades and school related features which were accumulated by using school reports and questionnaires. Both the datasets have the performance of two different subjects such as Mathematics (mat) and Portuguese language (por). In [11], the two datasets were analysed under binary/five-level classification and regression technique. It consists of 649 profiles and 33 attributes. Table 1 shows the attributes of the data.

TABLE 1 ATTRIBUTES OF THE DATASET

Attributes	Details
1. school - student's school	binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira
2. sex-student'ssex	binary: 'F' - female or 'M' - male
3. age - student's age	**numeric: from 15 to 22**
4. address - student's home address type	binary: 'U' - urban or 'R' – rural
5. famsize - family size	binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3
6. Pstatus - parent's cohabitation status	binary: 'T' - living together or 'A' - apart
7. Medu - mother's education	(numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
8. Fedu - father's education	(numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
9. Mjob - mother's job	nominal: 'teacher', 'health' care related, civil 'services' e.g. administrative or police, 'at_home' or 'other'
10. Fjob - father's job	nominal: 'teacher', 'health' care related, civil 'services'
11. reason - reason to choose this school	nominal: close to 'home', school 'reputation', 'course' preference or 'other'
12. guardian - student's guardian	nominal: 'mother', 'father' or 'other'
13. traveltime - home to school travel time)	numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour
14. studytime - weekly study time	numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours
15 failures - number of past class failures	numeric: n if 1<=n<3, else 4

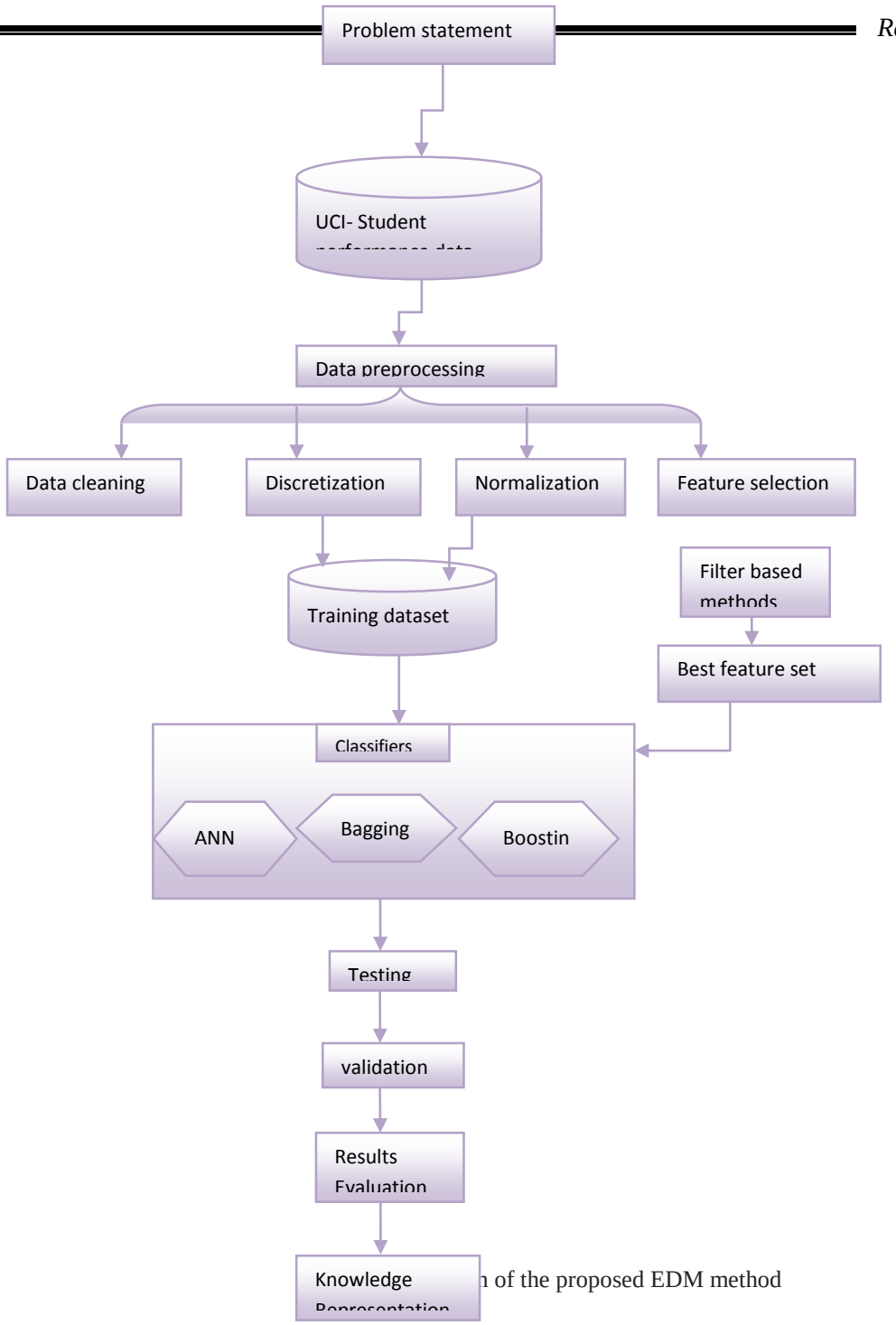
16 schoolsup - extra educational support	binary: yes or no
17 famsup - family educational support	binary: yes or no
18 paid - extra paid classes within the course subject (Math or Portuguese)	binary: yes or no
19 activities - extra-curricular activities	binary: yes or no
20 nursery - attended nursery school	binary: yes or no
21 higher - wants to take higher education	binary: yes or no
22 internet - Internet access at home	binary: yes or no
23 romantic - with a romantic relationship	binary: yes or no
24 famrel - quality of family relationships	numeric: from 1 - very bad to 5 - excellent
25 freetime - free time after school	numeric: from 1 - very low to 5 - very high
26 goout - going out with friends	numeric: from 1 - very low to 5 - very high
27 Dalc - workday alcohol consumption	numeric: from 1 - very low to 5 - very high
28 Walc - weekend alcohol consumption	numeric: from 1 - very low to 5 - very high
29 health - current health status	numeric: from 1 - very bad to 5 - very good
30 absences - number of school absences	numeric: from 0 to 93

31 G1 - first period grade	numeric: from 0 to 20
32 G2 - second period grade	
33 G3 - final grade	numeric: from 0 to 20, output target

Data was processed before applying the Machine learning algorithms. The data was cleaned checked for null values and the data type was changed accordingly. The dataset was analysed to remove unused data so that accuracy could be improved than the existing methods.

IV. DESIGN OF THE PROPOSED EDM FOR STUDENT PERFORMANCE EVALUATION

The proposed study would find the hidden knowledge in the datasets for the objectives that are expected in this research[12]. The student economic background attributes that affects the student performance was analysed using the Machine learning algorithms to obtain good result. Then the ML algorithms choosed was combination of traditional and strong in order gain a proper model. The design of the proposed work is mentioned in the Figure 2 This proposed design have used a dataset with unformatted data which cannot apply machine learning algorithms on them because it need only clean data to predict accurate results. In order to train models effectively data must be precise so that models will calculate accurate results. To bring datasets into wanted shape, it should be fully analysed to find out all null values either to remove or replace them with proper value. In the next process, should look for those attributes which are co-related on each other so that it can be either spitted or removed in case they are not needed for the training purpose. This proposed method has concentrated on the relevant attributes for the dataset. The main contribution in the work is choosing the right attributes and predicting the economic background factor that affects the student's performance. [13] Then building a proper Machine Learning representation to predict the student performance for the further enhancement in their Education



n of the proposed EDM method

V. EXPERIMENTS AND RESULTS

Moreover, EDM utilizes wide range of techniques to examine data including the Supervised and Unsupervised model induction, parameter estimation, Relationship Mining, and other methods. This paper explains in detail the procedure followed to achieve different predictive models of students' academic performance. More intentionally, the steps and results of the following three main DM techniques will be explained: Bagging, Artificial Neural Networks (ANN) and Boosting**.**

A. TOOLS USED IN THE IMPLEMENTATION

The proposed test was performed in the [14]WEKA tool which was being used for Machine learning algorithms. This has been used by the Researchers to enhance their work without worrying about the coding part. Many new algorithms can be downloaded in weka tool which is also an Open source software created by University of Waikato in New Zealand.

B. EXPERIMENT

The experiment of the proposed work was carriedout in the Window 7 Compaq PC with a processor Pentium® Dual core CPU installed RAM 2.00GB @2.30GHz in 64 bit Operating system. The Weka tool was installed and the necessary packages and libraries were installed for the data mining process. Then dataset was downloaded for the experiment purpose from the UCI Repository, loaded into the Weka Tool as CSV file**.**

C. DATA PREPARATION

This is the pre-processing task that involves in the steps before applying the data mining algorithm. This converts the original data into appropriate shape prior to the algorithm application. Data pre-processing consists of various tasks as cleaning of dataset, attribute selection and data transformation.

- **Data cleaning** : Humans causes error while handling the Educational datasets so it is not normally handled with usual techniques[15] so this dataset has limited or zero noise, when compared with other form of dataset. To improve the visualization and clarity of data filtering, sorting, and labelling procedures were done. Important attributes could be easily chosen if it is cleaned effectively based on their usage.

D. **Data Transformation: Discretization** is done to clear the uneven values in the dataset. It is a procedure of converting unfinite data into set of small intervals. Continuous attributes are used in the real world data mining activities. **Normalization:** changing the data variable in to given range of values is the Normalisation. Normalizing the values gives the result in a perfect form. In pre-processing, the dataset will be converted to increase the value of the model result. [16]

E. FEATURE SELECTION

A research study [17] explained the feature selection as most important work in data preprocessing. The aim of this procedure is to choose few important and suitable subset of features from dataset to convert or reduce the number of attributes that can appear in the algorithm. Therefore reducing the few number of features will remove the repeatable and inappropriate data. Likewise the feature selection helps in improving the performance of the learning algorithm by enhancing the data quality. Two Feature selection methods are Wrapper Based methods and Filter Based methods. Filter method is used to discover applicable subset of features while restricts the remaining. This ranks the attributes by using variable ranking methods so that appropriate features can be chosen and applied to the learning algorithm. In this proposed work, we applied selection algorithms based on the gain ratio which used various feature scores to identify the most important features for building students' performance . Figure 3 shows the highly ranked features after filter based evaluation of the UCI student performance dataset.

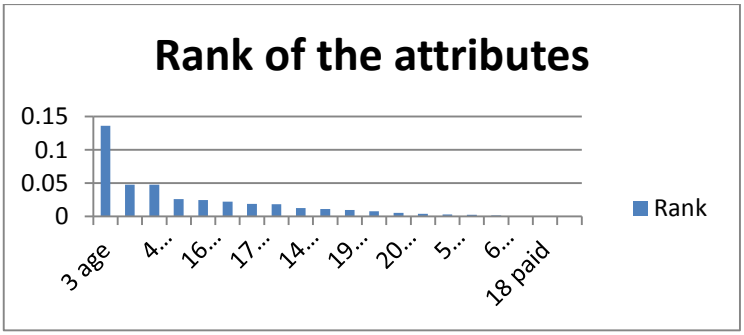


Figure 3 UCI Student performance dataset features

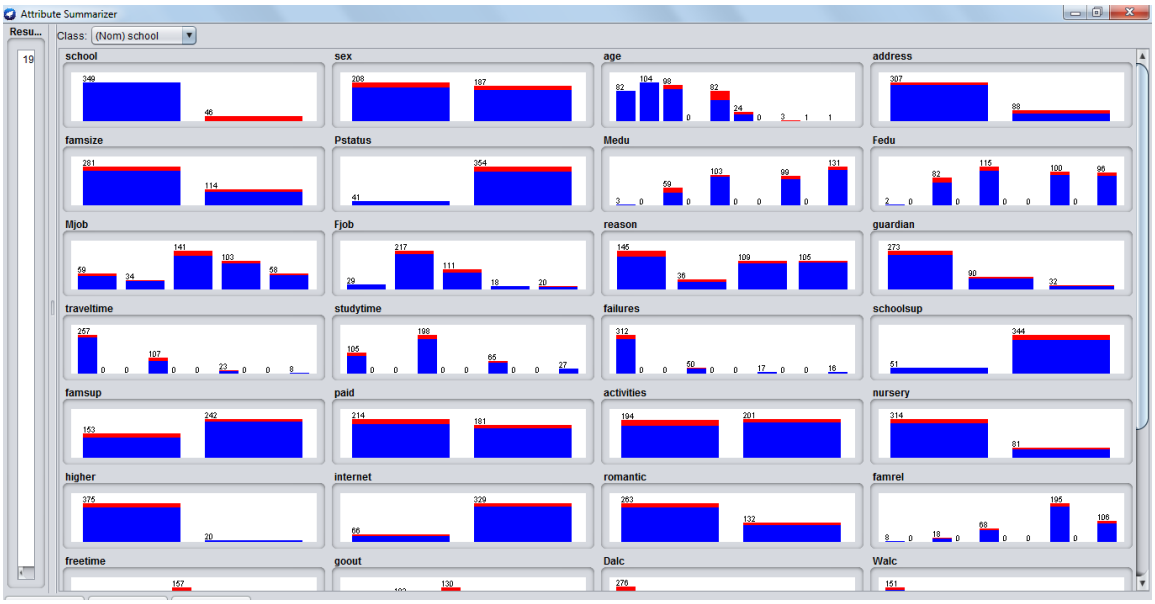


Figure 4 Attributes of UCI student dataset

The Figure 3 states the features selected using the ranking methods using the infogain ranking filter method. From which the features are categorized manually to get the desired result and to analyse the performance based on the three attributes types 1) Academic background 2) Personal attributes 3) Economic background .These are used together and separately to analyse the factors affecting the student's performance. Figure 4 explains the attributes of UCI dataset elaborately in a visual form[18].

We have categorized the academic background attributes as

Personal attributes- Sex, age, address, romantic, dalc, walc.

Academic background- Studytime, failures, paid, activities, nursery, higher, gout, absences

Economic background – pstatus , fedu , famsize, fjob, medu, reason, guardian, mjob, famsup, traveltime, freetime, internet,schoolsup, famrel, health.

Using these attributes the analysis was done for each group of attributes to find which affects the student performance.

F. CLASSIFIERS

The three different classifiers that are applied to assess the student's performance are Boosting algorithm, Bagging and Artificial Neural Network (ANN).

- **Multi-layer Perceptron (MLP)** It is a supervised learning technique that uses a function $f(\cdot):Rm \rightarrow Ro$, where m is the number of size for input and o is the number of size for output. The MLP uses a set of attributes like X=x1,x2,...,xm and a target y, which can learn a non-linear function estimator for either classification or regression techniques.
- **Boosting** – The accuracy of the data predictions could be enhanced by using the power of Boosting algorithm. The term 'Boosting' means changing the weak learner to strong learners in a family of algorithms. Boosting merges the weak learner and builds the base learner to form a strong rule. Different allocations of base learning algorithms are applied to discover a weak rule. A new weak prediction rule was created during each time the base learning

algorithms are applied. After numerous iterations, the boosting algorithm gathers these weak rules into a one strong prediction rule[19].

- **Bagging** : A simple and very dominant ensemble technique. Bootstrap Aggregation is known as Bagging[20]. An ensemble method is a technique that integrate the predictions from several machine learning algorithms jointly to make more precise predictions than any individual data model.

G. EVALUATION METRICS

There are many evaluation metrics are available but the metrics used in this research are the Accuracy, True Positive rate, False positive rate, precision, Recall , F-Measure and Confusion matrix.

H. RESULTS

This results are tabulated according to the different features of two Portuguese schools performing in the math subject.This would help to find out the students not performing well at earlier stage. The data was analysed from 10% training data to 80% where the results are keep increasing as the training set increases.As shown in the Table 3 it gives highest accuracy with all attributes with 80% data.

TABLE 3 TRAIN DATA-ACCURACY

Train data	**Accuracy**
10	88
20	86
30	84
40	83
60	90
70	90
80	91

TABLE 4 MLP RESULTS WITH DIFFERENT ATTRIBUTES

MLP CLASSIFIER						
ATTRIBUTES	ACCURACY	TPR	FPR	PRECISION	RECALL	F-SCORE
ECONOMIC BACKGROUND	72	79	66	87	79	83
PERSONAL	81	88	58	89	88	89
ACADEMIC	84	98	91	85	98.5	91

TABLE 5 BAGGING RESULTS WITH DIFFERENT ATTRIBUTES

BAGGING CLASSIFIER						
ATTRIBUTES	ACCURACY	TPR	FPR	PRECISION	RECALL	F-SCORE
ECONOMIC BACKGROUND	88	97	95	88	97	93
PERSONAL	83	96	93	86	96	91
ACADEMIC	86	98	98	88	98	93

TABLE 6 MULTIBOOST RESULTS WITH DIFFERENT ATTRIBUTES

MULTIBOOST CLASSIFIER

ATTRIBUTES	ACCURACY	TPR	FPR	PRECISION	RECALL	F-SCORE
ECONOMIC BACKGROUND	86	98	99	88	98	92
PERSONAL	86	99	99	86	99	92
ACADEMIC	82	98	98	88	98	93

Then the MLP classifier was analysed with the economic background features of the students which has the accuracy of 72.2% and with personal attributes alone accuracy was 81% and academic attributes gives 84% accuracy in results. As per the accuracy the MLP has the good accuracy with academic attributes but the economic features are 72% seems higher in early predicting of students at risk.Then same experiment with bagging classifier which shows the accuracy of economic background as 88% and the personal attributes as 83% and then academic as the 88%. From (table 5)this the early stage of predicting the students performance could be found using the economic background alone as the results of economic attributes as 88%.

Then again the same dataset with multiboost classifier which shows the accuracy of economic background as 86% and the personal attributes as 86% and then academic as the 88%. From (table 6)this the early stage of predicting the students performance could be found using the economic background alone as the results of economic attributes as 86%.

Focusing on **Economic background attributes** such as famsize, pstatus, medu, fedu, mjob, fjob, reason, guardian, traveltime, schoolsup, famsup, internet, famrel, freetime, health are the important features that early predicts the students performance than the other. So these also affects their learning behaviour that economically background students are unable to perform the best. So the online learning could be a dilemma if the economic background attributes are not fulfilling to buy gadgets and internet at every home. Thus we could say the economic background plays a essential role in the students' accomplishment. As the accuracy with economic features are higher than 70% this may predict the students performance earlier for virtual learning. The figure 11 displays the internet attribute distribution.

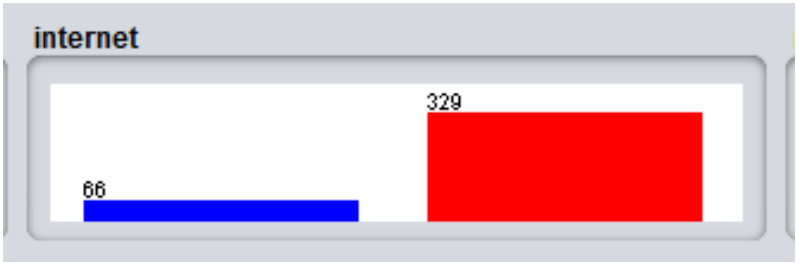


Figure 11 Histogram Of Internet Attribute Distribution

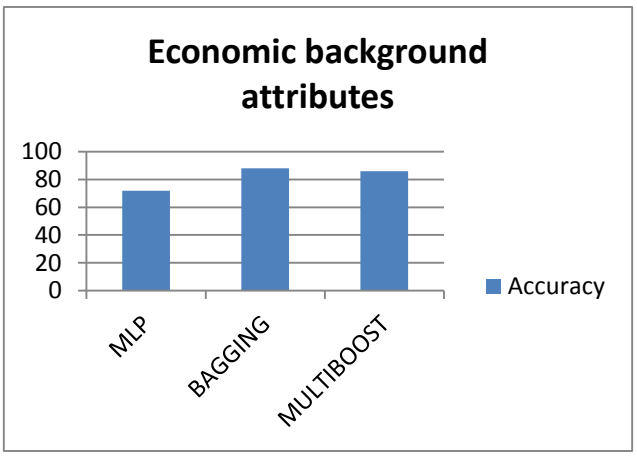


Figure 12 Accuracy of economic background attributes

VI. CONCLUSION

In this paper the UCI student performance dataset was analysed to detect the various element which affects the student performance. The dataset contains many attributes but only highly ranked attributes were used to reduce the dimensionality as well as the unwanted features. The economic background plays a important role in the student life which affects them in various ways are discussed. Thus the results show that only economic attributes affects the students performance. MLP gives 72% accuracy, Bagging classifiers shows 88% and MultiBoost classifiers gives 86% accuracy with the economic background attributes. Baaging classifiers shows the higher accuracy from which it is clear that economic background of the student also affects their learning behaviour in Educational system. This work could be useful for the Education people to increase the student performance by changing their learning methods that suits the individual student to gain confidence in their academic performance. The future work is to use different datasets to analyse other factors that are affecting the student performance.

REFERENCES.

[1] O. H. Lu, A. Y. Huang, J. C. Huang, A. J. Lin, H. Ogata, and S. J. Yang, "Applying Learning Analytics for the Early Prediction of Students' Academic Performance in Blended Learning", Journal of Educational Technology & Society, 21(2), 2018, pp.220-232.

[2] P. Suhem, Z. Zain, and M. Fatima, "Application of Data Mining in Educational Databases for Predicting Academic Trends and Patterns", 2012 IEEE International Conference on Technology
Enhanced Education (ICTEE), 2012, pp 1 – 4

[4] C. Romero, M. I. López, J. M. Luna, and S. Ventura, "Predicting students' final performance from participation in online discussion forums", Computers & Education, 2013, 68, pp.458-472.

[5] Y. H. Hu, C. L. Lo and S. P. Shih, "Developing early warning systems to predict students' online learning performance", Computers in Human Behavior, 2014, 36, 469-478.

[6] R. Asif, A. Merceron, and M. K. Pathan, "Predicting student academic performance at degree level: A Case study", International Journal of Intelligent Systems and Applications, 2014, 7(1), 49-61.
[7] Y. Meier, J. Xu, O. Atan, and M. Schaar, "Predicting grades". IEEE Transactions on Signal Processing, 2016, 64(4), 959-972.

[8] L. C. Yu, C. W. Lee, H. I. Pan, C. Y. Chou, P. Y. Chao, Z. H. Chen, S. F. Tseng, C. L. Chan, and K. R. Lai. "Improving early prediction of academic failure using sentiment analysis on self_evaluated comments." Journal of Computer Assisted Learning, 2018.

[9] Murugananthan, V. & ShivaKumar, B. L. (2016). An adaptive educational data mining technique for mining educational data models in elearning systems. Indian Journal of Science and Technology, 9(3), pp. 1–5. doi: http://dx.doi.org/10.17485/ijst/2016/v9i3/86392

[10] Amrieh, E. A., Hamtini, T. & Aljarah, I. (2015). Preprocessing and analyzing educational data set using X–API for improving student's performance. In 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT) (pp. 1–5). IEEE. doi: http://dx.doi.org/10.1109/ AEECT.2015.7360581

[11] Ching-Chieh Kiu, "Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities",978-1-5386-7167-2/18/$31.00 ©2018 IEEE

[12] Chitra Jalota, Rashmi Agrawal ,"Analysis of Educational Data Mining using Classification", 978-1-7281-0211-5/19/$31.00 2019 ©IEEE

[13] sana, "analyzing students' academic Performance through Educational data mining", 3C Tecnología. Glosas de innovación aplicadas a la pyme. ISSN: 2254–4143

[14] K. Deepika, " Comparison Of Student Academic Performance On Different Educational Datasets Using Different Data Mining Techniques", ISSN (e): 2250 – 3005 Volume, 08 Issue, 9 Sepetember – 2018 International Journal of Computational Engineering Research (IJCER) .

[15] Muhammad Faisal Masood, Aimal Khan, Farhan Hussain, Arslan Shaukat, Babar Zeb, Rana Muhammad Kaleem UllahTowards the Selection of Best Machine Learning Model for Student Performance Analysis and Prediction, 978-1-7281-4577-8/19/$31.00 ©2019 IEEE

[16] Paulo Cortez and Alice Silva, "Using Data Mining to Predict Secondary School Student performance", Dept.Information Systems, Algoritmi,2008, R&D Centre, University of Minho. http://www3.dsi.uminho.pt/pcortez .

[17] Romero C, Ventura S, Pechenizky M, Baker R. (2010). Handbook of Educational Data Mining. Data Mining and Knowledge Discovery Series. Boca Raton, FL: Chapman and Hall/CRC Press

[18] Karegowda, A. G., Manjunath, A. S. & Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. International Journal of Information Technology and Knowledge Management, 2(2), pp. 271–277.

[19] Chitra Jalota, Rashmi Agrawal," Analysis of Educational Data Mining using Classification", 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), India, 14th -16th Feb 2019, IEEE

[20] Luiz Carlos B. Martins, "Early prediction of college attrition using data mining", 0-7695-6321-X/17/31.00 ©2017 IEEE DOI 10.1109/ICMLA.2017.000-6

[21] Elaf Abu , Thair hamitini, Ibrahim Aljarah, "Mining Educational Data to Predict Student's Academic Performance using Ensemble Methods", Vol 9, no.8(2016), pp.119-136

[22] Mayreen V. Amazona, Alexander A. Hernandez, "Modelling Student Performance Using Data Mining Techniques: Inputs for Academic Program Development", https://doi.org/10.1145/3330530.3330544,pp.36-40.

[23] Sana, Isma Farrah Siddiqui, Qasim Ali Arain, "Analyzing Students' Academic Performance Through Educational Data Mining" Special Issue, May 2019, pp. 402–421. doi: http://dx.doi.org/10.17993/3ctecno.2019.