International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

# Prediction of Student Performance using Hybrid Classification

**A.Dinesh Kumar, R.Pandi Selvam, V.Palanisamy**

***Abstract:*** ***Data mining technologies allow collection, storage and processing huge amounts of data and carrying a large variety of data types and samples. Predicting academic performance of student is the most successive research in this era. Previous research work researchers are used different classification algorithm to predict the student performance. There is lot of research work to be taken in the field of educational data mining and big data in education to increase the accuracy of the classification algorithm and predict the academic performance of student. In this research work we used hybrid classification algorithm for predicting the performance of students. Two Popular classification algorithms ID3 and J48 were applied on the data set. To make hybrid classification voting technique is applied using weka machine learning tool. In this work we tested how the hybrid algorithm accurately predicts the student data set. To check the predicted result classification accuracy was computed. This hybrid classification algorithm gives accuracy with 62.67%.***

***Keywords: Big Data, Data Mining, Educational Data Mining (EDM), Hybrid Classification, Prediction.***

## I. INTRODUCTION

Data mining is one of the most cardinal areas in recent technologies for retrieving valid information from huge amount of unstructured and distributed data using parallel processing of data [7]. Data mining techniques are applied in various fields to find the novel information from huge data set. Nowadays data mining techniques are mostly used in educational field. Most of the researcher has taken data mining techniques to find the useful information from educational field. Applying data mining in education field is the most prominent research area of today's researcher. The researchers are seeking interest in the educational filed to investigate new research [19].

Manuscript published on November 30, 2019.
* Correspondence Author
**A.Dinesh Kumar***, Research Scholar, Department of Computer Applications, Alagappa University, Karaikudi, India.
**R. Pandi Selvam**, Assistant Professor and Head, PG Department of Computer Science, Ananda College, Devakottai, Tamilnadu, India.
**V. Palanisamy**, Professor and Head, Department of Computer Aplications, Alagappa Univeristy, Karaikudi, Tamilnadu, India.



Data mining has been applied in various applications. Recently data mining technology has been used in educational filed to extract the hidden information from educational data sets.

*Big Data in Education*

Big data has feature to revolutionize not just research, but also education.

Big data can support the classical educational system facilitating teachers to analyze what students know and what techniques are most effective for each student. In this way, teachers also able to learn new techniques and teaching methods about their work [12]. Big data can easily apply at online education. The online education has a very large development at recent years. It has a very escalating impact of education field [13].

Education community has found plentiful way to benefit from big data. Educational data mining and learning analytics employ technologies from statistics, computer science and machine learning to extract useful information from collected educational data, gain valuable insight into learning, and find out solutions to improve learning performance and teaching effectiveness.


**Fig.1. Big Data in Education**

Fig.1 explains big data an application was used in various field in educational data mining, such as improve student results, reduced dropouts. In recent years, big data is the hottest topic in the world of science and technology. In terms of big data a "3Vs" model is interpreted as three important characteristics such as volume, velocity and variety. Volume: Organization gathers data from variety of sources including business and social media. It is the primary characteristics for immense chunk of datasets. The size of data has been changed dramatically from the scale of KB, MB to GB in units respectively. Each of scale is thousands times larger. The data size has been exponentially growth [18].

Velocity: This is the second characteristics. It alludes to the high development rate of datasets. The speed of development infers two focuses. The first indicates the speed of generation of data is relatively high. The second denotes the speed of data processing is high.
Variety: The third illustrative concentrates the difference of data regarding distinctive data channels, configurations, and structures, which are past the present ability of structuring of data processing. Data comes in various kinds of groups from organized, numeric data in conventional databases to unstructured content records, email, video and sound.
This paper organized as follows: Section II describes the related work based on predicting student performance. Section III discusses techniques and tools used in big data and educational data mining. Section IV presents the proposed work of this paper. Section V concludes this paper.

## II. RELATED WORK

Sagardeep and Shailendra[3] have discussed about emerging trends in application of big data in educational data mining and learning analysis. Nowadays the big data technology has entered in the education field, to mine huge changes of data. They also discussed the latest tools and technologies of big data in the field of learning analytics and educational data mining.Xinguo and Shuang [2] have addressed that the typical applications of big data in education. They have introduced relevant application of educational data mining to show how big data in education help to solve education issues. They used three applications to solve educational issues in big data such as performance prediction, performance presentation and understanding the students learning activity.Maryam Zaaffar et al. [5] conducted research in Performance Analysis of Feature Selection Algorithm for Educational Data Mining. Feature Selection technique used to select the proper attribute for the classification. In that research work they have applied Six Feature Selection Algorithms for classification. Based on their results we found that principal components have shown better results by using it random forest classifier.Sujith and Jaiganesh [6] made a comprehensive survey on academic progression of students in territory education using classification algorithm. They intensively discussed about various factors influencing the progression of students like student pre-enrollment data, continuous assessment, psychological data and socio economic status. From that paper they concluded a recommender system has to be implemented to analyse the student performance.Ravinder et al. [8] have applied naïve bayes, decision trees and logistic regression algorithm to predict the probability of students' degree completion that work was implemented in R language. Based on their research they concluded that k-nearest neighbor algorithm is least preferable amongst all.

Wattana punlumjeak et al. [9] conducted experiment on student performance prediction using feature selection. They applied a large student data set, as a big, to find a prediction model to classify the student's performance on Microsoft Azure platform. Their result shows that mutual information in feature selection method with neural network classifier gave the best accuracy at 90.60% for student's data.

Pratiyush et al. [10] have predicted academic course preference of student using hadoop inspired MapReduce in big data. Their results shows that large volume of course combinations given in the form of input dataset after passed through the mapper function in map reduce framework the maximum of students have shown key interest towards "C", ,"C++" and java courses.

S.Rajeshwari and R.Lawrence [11] conducted research on predict the learner's academic performance using classification model in big data. They predicted the student's semester grade and final year student campus placemen using ID3, C4.5 and C5.0. They concluded C5.0 is the highest speed and pre-pruning algorithm for predicting the student's performance.

## III. TECHNIQUES USED IN EDM

### *A. Association Rule Mining*

Association rule mining algorithms are intended to discover pertinent relationships between the variables of the data set. These associations are then revealed by if…then rules. Association rules have a probability of occurrence, that is, if condition is met, and then there is a certain possibility of occurring result. Association rule mining algorithms are mine only strong rules. Strong rules satisfy a minimum support threshold and a minimum confidence threshold [20].

### *B. Clustering*

Clustering is an unsupervised learning paradigm; it may reveal interesting unknown relations in the data. It is a system of grouping sets of objects in groups in a way that objects from a cluster have more similarities than objects from different clusters. Each cluster may be considered as a class with no label, and thus, clustering is sometimes referred to as automatic classification [20].

### *C. Classification*

Classification is used to classify the data based on the training set and then uses that pattern to classify the new data which is known as the training set. It is called as supervised learning technique because the classes are predefined before extracting patterns on the target data [17]. Some popular classification methods used in EDM are Decision Trees, Bayesian Classifier, Artificial Neural Network, Support Vector Machine, K-Nearest Neighbor, Linear Regression and Density Estimation.

### *D. Linear Regression*

This is a prediction technique that predicts a numeric. Various attributes like sales, age and weight. It is statistical methodology [17]. The aim of the task is to achieve a function of the independent variables that allows computing conditional expectations of a dependent variable for prediction.

### *E. Support Vector Machine*

SVM is a method for the classification of both linear and nonlinear data. It uses a nonlinear mapping to transform the original training data into a higher dimension.

The significant benefit of SVM is the features of global optimization and high generalization ability. In addition, it removes over fitting issues and provides a sparse solution when comparing with traditional approaches like Artificial Neural Network (ANN) [14].

*Tools used in big data*

There are many tools which are open source that helps in hold back big data. Few of them are

Hadoop: Hadoop is an open source framework that allows distributed storage and processing of massive datasets. It is used for storing and processing big data in a distributed manner on large clusters of commodity hardware. It was developed based on the paper written by the Google on mapreduce system and it applies concepts of functional programming,

Weka: Weka tool is a collection of machine learning algorithm used for data classification, clustering, and association rule. Weka tool is developed by using java code. This tool contain preprocessing, cluster, associate, select attributes and visualize. In education, weka has been widely used to make prediction due to its efficiency in exploring, analyzing and predicting student performance [6].

Orange: Another free open source tool which is written in Python for processing and mining big data. Its interface is relatively easy and simple with move and customizes functionalities with collection of additional items.

Mango DB: It is free, cross-platform, open source document oriented non-relational database management system. It falls under NO-SQL database category.

Python: It is used for data exploration, analytics, and prediction of an educational data. To perform data exploration and analysis, Python Numpy, scikit learn and other packages are used which helps in speeding up the exercise and establish interfaces with other packages in the Python Ecosystem [15].

## IV.PROPOSED WORK

In this proposed work we collected data from various departments of UG students. Totally 500 data set were collected from the student, from this we have taken only 300 data set for predicting the performance of student. To predict the student performance we applied the data set in weka tool.

*Data Set*

The data set we used for better prediction is given in the Table-1.

**Table-1: Attribute Description**

| Attribute | Description |
|---|---|
| FI | Father's Income |
| ME | Mother's Education |
| MW | Mother Working Status |
| SH | Student's Study Hours |
| TU | Tuition |
| SN | Social Network Usage |

FI- Fathers Income is mainly correlated with the student performance. So this attribute is taken for predicting the student performance.

ME- Educated mother can help their children studies. So this attribute also considered by as for prediction.

MW- If mother goes for work they can't help their student's studies. That's why we considered this attribute for predict the performance of the student.

SH- Apart from the above attributes student's study hours are mainly related to the performance of the student.

TU- Tuition attribute is considered by us to check whether the student's performance is improved by the tuition or not.

SN- Nowadays most of the students are using mobile. We consider this attribute to check whether student is affected by social network usage or not.

*Data Preprocessing*

Predicting the student performance student.arff data was given to the weka machine learning. This contains 13 attributes. This 13 attributes were applied to data preprocessing for better prediction. For better classification we applied CfsSubsetEval from attribute selection method in data preprocessing stage. Based on the above preprocessing step, 7 attributes are selected for classification step. Fig.2 shows the attribute selection method using CfsSubsetEval.

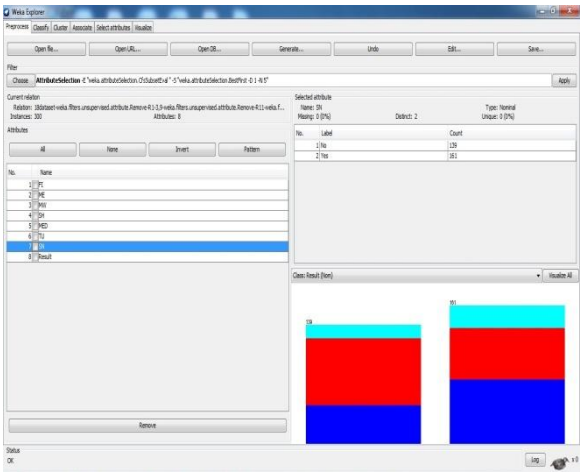

**Fig. 2. Attribute Selection**

*CfsSubsetEval*

CfsSubsetEval Estimates the significance of a subset of features by taking into individual predictive ability of each attribute of each attribute along with the degree of redundancy between them [5]. Using these attribute evaluation step proper attributes was selected for the classification step.

*Hybrid classification*

After preprocessing step the selected attributes were applied to classification process to predict the student performance. In the classification step we applied hybrid classification algorithm.

To make hybrid classification we use voting technique in weka machine learning. Through voting technique two best classification algorithms are selected. Such algorithms are J48 and ID3.

*J48*

J48 algorithm is developed by Quinlan Ross that generates the decision trees which can be used for classification problems. It is the successor of ID3 algorithm by dealing with both categorical and continuous attributes to build a decision tree. It is also based on Hunt's algorithm. To handle the continuous attributes, J48 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values [16].

*ID3*

ID3 is a simple decision tree algorithm introduced by Ross Quinlan in 1986. It is based on Hunts algorithm. The basic idea of ID3 algorithm is to construct the decision tree by employing a top- down, greedy search through the given sets to test each attribute at every tree node. The tree is constructed in two phases. The two phases are tree building and pruning [16]. Fig. 3 shows hybrid classification applied through voting technique.

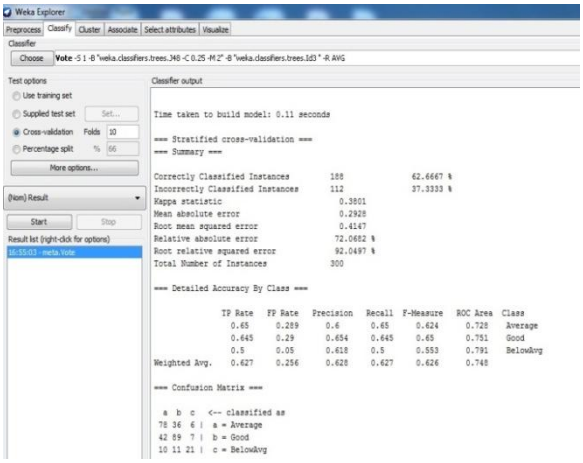


**Fig.3. Hybrid Classification**

**Table- 2: Experimented Results**

| Total No. of Instances | Instances Classified Correctly | Instances classified Incorrectly | Classification Accuracy |
|---|---|---|---|
| 300 | 188 | 112 | 62.67% |

Table- 2 shows the experimented result. The hybrid classification algorithm classified 188 instances are correctly and classified 112 instances are incorrectly. The classification accuracy was 62.667%. So this hybrid algorithm accurately predicted the given data set. This classification accuracy will helpful for predict the future data set.

## V.CONCLUSION AND FUTURE WORK

In this Paper we used two best classification algorithms for the prediction of student performance. Prediction of student performance is one of the most prominent researches filed in educational data mining. Using voting technique ID3 and J48 classification algorithms were applied for predicting the student performance. The classification accuracy was computed by hybrid classification is 62.667%. In future we wish to apply different classification to improve the classification accuracy. There is numerous research has to be done in the field of education to increase the classification accuracy using different hybrid classification algorithm. This helps to the teacher as well as institution to take proper decision to improve the performance of the student. So based on the above research work, in Future we are going to investigate new hybrid classification technique to improve the classification accuracy for better prediction.

## REFERENCES


1. Sri Murugarasan, Mohd Yasin, Mallika Govindasamy, " Big Data Frame Work for Students' Academic Performance Prediction: A Systematic Literature Review", IEEE, 2018.
2. Xinguo Yu, Shaung Wu ,"Typical Application of Big data in Education",IEEE, 2015.
3. S.Roy , Dr.Shailendra " Emerging Trends in Application of Big Data in Educational Data Mining and Learning Analytics", 7th International Conference on Cloud Computing, Data Science & Engineering, IEEE,2017.
4. Lin Cen, Dymitr Ruta, Jason Ng," Big Education: Opportunities for Big Data Analytics", IEEE, 2015.
5. Maryam Zaffar, Manzoor Ahmed, K.S.Savita, "Performance Analysis of Feature Selection Algorithm for Educational Data Mining", IEEE Conference on Big Data and Analytics (ICBDA), 2017.
6. Azwa Abdul Aziz, Nur Hafieza Ismail, "Mining students academic performance"- Journal of theoretical and applied information technology, Vol.53-2013.
7. Pratiyush Guleria, Manu Sood ",Big Data Anlalytics: Predicting Academic Course Preference Using Hadoop Inspired MapReduce", IEEE, 2017.
8. Ravinder Ahuja, Yash Kankane," Predicting the Probability of Student's Degree Completion by Using Different Data Mining Techniques", IEEE, 2017.
9. Wattana Punlumjeak, Nachirat Rachburee, "Big Data Analytics: Student Performance Prediction Using Feature Selection and Machine Learning on Microsoft Azure Platform", Journal of Telecommunication, Electronic and Computer Engineering", ISSN: 2289-8231, Vol.9.
10. Pratiyush Guleria, Manu Sood, "Big Data Analytics: Predicting Academic Course Preference Using Hadoop Inspired MapReduce", Fourth International Conference on Image Processing (ICIIP), 2017.
11. S.Rajeshwari and R. Lawrence, " Classification Model To Predict the Learners' Academic Performance Using Big Data", IEEE,2016.
12. D.M West, "Big Data for Education: Data Mining, Data Analytics, and Web Dashboards," Government Stud.Brook US Reuters, 2012.
13. Athanasios S. Drigas, P. Lelipoulous , " The Use of Big Data in Education", Internal Journal of Computer Science Issues,Vol-11, September 2014.
14. J. John Kennedy, R. Pandi Selvam, "Cloud-Centric IoT based Decision Support System for Gestational Diabetes Mellitus using Optimal Support Vector Machine", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1, May 2019.
15. Pratiyush Guleria, Manu Sood " Predictive Data Modeling: Educational Data Classification and Comparative Analysis of Classifiers Using Python", 5th IEEE International Conference on Parallel, Distributed and Grid Computing(PDGC-2018).
16. Jiawei Han , Micheline Kamber, Jian Pei," Data Mining Concepts and Techniques", 3rd Edition.
17. A.Dinesh Kumar, R.Pandi Selvam, K.Sathesh Kumar, "Review on Prediction Algorithms in Educational Data Mining", International Journal of Pure Applied and Mathematics (IJPAM), Volume-118.

18. R.Swathi, N.Pavan Kumar, L.KiranKranth, "Systematic Approach on Big Data Analytics in Education Systems", International Conference on Intelligent Computing and Control Systems (ICICCS), 2017.
19. A.Dinesh Kumar, Dr.V.Radhika," A Survey on Predicting Student Performance", International Journal of Computer Science and Information Technologies (IJCSIT), Vol 5.
20. 0]Ricardo Mendes and Joao P.Vilela, "Privacy-Preserving Data Mining: Methods, Metrics and Applications", IEEE,2017.

## AUTHORS PROFILE

**A.DINESH KUMAR** is a Ph.D Scholar in the Department of Computer Applications, Alagappa University, Karaikudi. He received M.Phil., Degree in Sri Krishna College of Arts and Science, Coimbatore. His area of research work involves in Data Mining and its applications.

**Dr.R.PANDI SELVAM** received Ph.D Degree in Computer Science and Engineering, Alagappa University, Karaikudi. He is Currently Working as an Assistant Professor and Head, PG Department of Computer Science, Ananda College, Devakottai. His area of research interest includes Computer Network and Security and Data Mining.

**Dr.V.PALANISAMY** received Ph.D Degree in Computer Applications, Alagappa University, Karaikudi. He is a Professor and Head of the Department of Computer Applications, Alagappa University, Karaikudi. He has been in 30 years of teaching in the field of Computer Science and Applications. His research interests are Computer Algorithm, Network Security, Data Mining, Ad-Hoc Networks and Biometrics Security.