IJASRM
International Journal of Advanced Scientific Research and Management, Volume 4 Issue 1, Jan 2019
www.ijasrm.com
ISSN 2455-6378

Predicting Student Performance Using Data Mining Techniques: A Survey Of The Last 5 Years

S.Menaka[1]*and G.Kesavaraj[2]

[1] PG & Research Department of Computer Science,
Vivekanandha College of Arts and Sciences for Women (Autonomous),
Tamil Nadu, INDIA

[2] PG & Research Department of Computer Science ,
Vivekanandha College of Arts and Sciences for Women (Autonomous),
Tamil Nadu, INDIA.

Abstract

Educational Data Mining (EDM) is the field by means of data mining techniques in learning environments. Applying data mining technique in inculcate setting is called as Educational Data Mining (EDM) and is a field that exploits statistical, machine learning, and data-mining (DM) algorithms over the variants of educational data. Existing methods have used features which are mostly associated to educational presentation, family income and family resources, while features belonging to family expenditures and students' individual information are usually ignored. In this article, an effort is made to examine aforementioned feature sets by collecting the scholarship holding students' data from different universities of India. Knowledge analytics, discriminative and generative classification models are applied to expect whether a student will be able to absolute his degree or not. We introduce the Weka tool and C4.5 algorithm to analyze replicating studies and discuss the importance of replicating and reproducing previous work. We describe the state of the art in collecting and sharing programming data. To better know the challenges concerned in replicating or reproducing presented studies.

Keywords: *Educational data mining, Programming, Learning analytics, Replication and Literature review*

1. Introduction :

Educational Data Mining (EDM) is the field of using data mining techniques in educational environments. Applying data mining technique in inculcate setting is called as Educational Data Mining (EDM) and is a field that exploits statistical, machine learning, and data-mining (DM) algorithms over the variants of educational data . There survive various methods and applications in EDM which can follow both applied research objectives such as improving and enhancing learning quality, as well as pure research objectives, which tend to improve our understanding of the learning process [1].

EDM is an emerging discipline, concerned with developing methods for exploring the unique and increasingly large scale data. Data mining is widely used in educational field to find the problems arise in this field. Student performance is of great concern in then educational institutes where several factors may affect the performance. The factors that describe student performance can be used for predicting students performance by using some algorithms such as J48, Naïve Bayes, KNN etc.,

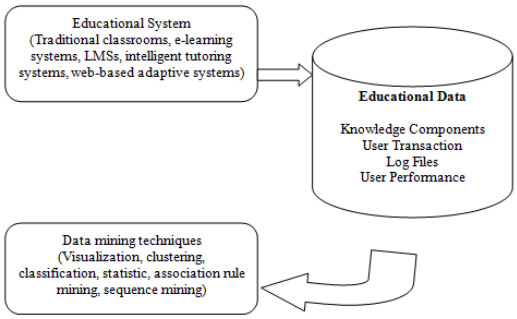

Fig.1.1 Shows the components of Educational Data Mining

Education is devoting increasing interest to digital technologies, as means to deliver contents, to advance learning, and to monitor advancements or students' behavior. The development nature of the careers filed demanding new procedures to prepare

the fresh practitioner for finding the appropriate job according to their Ambitions and orientations. This approach required the students to develop their professional identity while they acquiring their professional education [6].

2. Review Of Literature :

Algorithms and tasks in Educational Data Mining (EDM) can be categorized based on different properties. Several surveys of EDM exist, which have planned possible applications of EDM. We will seem into these surveys in more aspect in the literature assessment segment. In this list, we have tried to consider all the categories mentioned in previous surveys and in the literature, as well as new categories which we think need to be added. These new categories of algorithms can be explained by the growth of attention in EDM. In earlier studies, possible algorithms of EDM have been introduced sometimes in no exact order, sometimes based on the number of research papers completed in each of the categories. We try to collect possible algorithms of EDM into categories based on their final object. We have tried to set different algorithms mutually as much as possible to better highlight the similarities and differences.

2.1 Programming Replication

Petri Ihantola, Arto Vihavainen and et al., observed that there has been a significant increase in the amount of articles that are exploring student constructed solutions to programming problems. The underlying themes were often related to approaches for helping the student or the teacher, such as evaluating approaches for providing feedback, identifying at risk students, or extracting programming strategies or patterns that could, perhaps, be used during the instruction to inform students on their choices [3].

2.2 Student Evaluation Test

Carmela A. White, Daniela Wong Gonzalez evaluate the results of the meta-analysis provide strong support for the validity of student ratings as a measure of teaching effectiveness. SET ratings as a measure of teaching effectiveness that students learn more from highly rated professors has been accepted as the established fact in various research summaries and widely disseminated to faculty members, administrators, and general public. A Meta analysis necessity, at minimum, express the search strategies for main studies, provide the basic expressive information including produce size and sample size for all major studies, and ensure that the extracted primary study level data are perfect [4].

2.3 Naïve Bayes and Rule Induction Classifiers

Ali Daud, Naif Radi Aljohani, and et al., his research says a prediction model is developed to predict the performance of higher secondary school students, which is critical before getting admission into universities. The grades of graduate students are predicted using Naïve Bayesian and Rule Induction classifiers. Clusters are made from students' data and the outliers are successfully identified. A model is presented to estimate the abilities of students and competence of teachers in order to predict the future student outcomes. It shows that demographic profiles and personality traits features are correlated and have high impact on student performance. Similarly student performance evaluation and engineering students' abilities are analyzed for improved recruitment process by using data mining methods [7].

2.4 Machine Learning and Pattern Recognition

Nobuhiko Kondo, Midori Okubo and et al., his research says machine learning is the approach to give computers the ability to learn automatically like human beings. The machine learning methods have some sort of algorithms that discover patterns or rules from actual data, and the model learned appropriately can predict unseen data properly. The methods are often used in several fields such as pattern recognition, medical diagnostics, search engine, robotics, and so on. In the fields of analytics in education, the machine learning methods are frequently used for the construction of predictive model of learners [9].

3. Materials and Methods

3.1 Classification Model

Classification is one of the most general purpose domains of data mining. The aim of classification is to exactly predict the target class for each case in the data [7]. Classification is the problem of identifying to which of a place of categories a new study belongs, on the basis of

instruction set of data containing explanation whose category relationship is known. After knowledge phase, the classification, presentation the classifier form built is evaluated on a self-determining test set earlier used.

In classification, there are special types of techniques and algorithms likely to use for building a classifier representation. Classification model contains various algorithms as Decision tree, Artificial Neural Network (ANN), Support Vector Machine (SVM), Logistic Regression, Discriminate Analysis (DA), Rule Based System and Bayesian Belief Networks. In this paper the decision tree algorithm are used [8]. The decision tree approaches more powerful for classification problems. There are two steps in this techniques structure a tree and applying the hierarchy to the dataset. Decision tree algorithm is one of the predictive modeling approaches used in statistical or impurity, data mining and machine learning. Its goal to generate a model that predicts the value of a target variable based on several input variables. The impurity measures that are used to get the homogeneity of instances in a node of the tree, Information Gain, Gain Ratio and Gini Index are the mainly identified ones. Mostly Information Gain is used in Iterative Dichotomiser (ID3), but Gini Index value is used in Classification and Regression Trees. A sample representation of a decision tree was described in [9].

This survey has been extremely useful as a reference for this paper, as it provides many examples for each of the introduced categories as well as methods and techniques used in them. The categories of applications introduced in this survey are:

- Analysis and Visualization of Data
- Providing Feedback for Supporting Instructors
- Recommendations for Students
- Predicting Students Performance
- Student Modeling
- Detecting Undesirable Student Behaviors
- Grouping Students
- Social Network Analysis
- Planning Schedule

These applications are all mentioned in our list of applications with a few changes and additions. In earlier studies, possible algorithms of EDM have been introduced sometimes in no exact order, sometimes based on the number of research papers completed in each of the categories.

A) Prediction Tools

Waikato Environment for Knowledge Analysis (Weka) is a suite of machine knowledge software. Weka contains a compilation of visualization tools and algorithms for data analysis and predictive modeling, mutually with graphical user interfaces for simple contact to these functions. The unique non-Java version of Weka was a Tcl/Tk front end to (regularly third party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and to create file based system for running machine learning experiments.

Weka has a huge figure of regression and sorting tools. Native packages are the ones integrated in the executable Weka software, while other non-native ones can be downloaded and used within Weka environment. Among the native packages, the most prominent device is the M5p model tree package.

B) C4.5 Algorithm

C4.5 is an algorithm used to create a decision tree developed by Ross Quinlan. C4.5 is an addition of Quinlan's earlier ID3 algorithm. The decision trees generate by C4.5 can be use for classification. C4.5 builds decision trees from a set of education data in the same way as ID3, using the model of information entropy. The training data is a set {\displaystyle S={s_{1},s_{2},...}} of previously confidential samples. Each sample {\displaystyle s_{i}} consists of a p-dimensional vector{\displaystyle (x_{1,i},x_{2,i},...,x_{p,i})}, where{\displaystyle x_{j}} stand for aspect values or features of the sample, as well as the class in which {\displaystyle s_{i}}falls.

At each node of the tree, C4.5 chooses the quality of the data that most efficiently splits its set of samples into subsets enriched in one class or the other. The splitting state the normalized in sequence gain (difference in entropy). The attribute with the maximum normalized information gain is selected to create the decision. The C4.5 algorithm then recursion the partitioned sub lists.

IJASRM International Journal of Advanced Scientific Research and Management, Volume 4 Issue 1, Jan 2019
www.ijasrm.com
ISSN 2455-6378

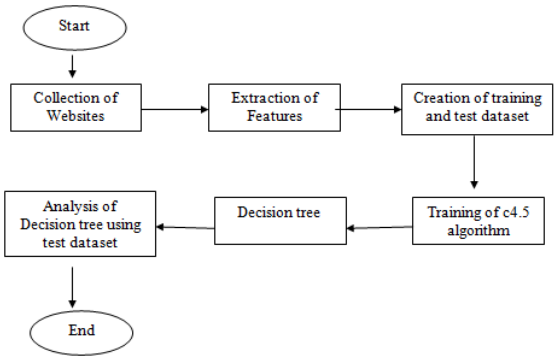

Fig.3.1. shows the flow of methodology

This algorithm has a few base cases.

- All the samples in the list belong to the equivalent class. When this happens, it just creates a leaf node for the decision tree saying to select that class.
- None of the features give any information expands. In this case, C4.5 creates a decision node advanced up the tree using the estimated value of the class.
- Instance of previously-unseen class encountered. C4.5 creates a decision node superior up the tree using the predictable value.

C) C5.0 Algorithm

C5.0 algorithm is used to create a decision tree which can be used for classification so it referred to as a statistical classifier. This model works through splitting the model based on the field that provides the maximum information gain. Every subsample defined in the first split is then split again, generally based on a different field, and the process repeats until the subsamples cannot be split any further [10]. Finally, the lowest-level splits are reexamined, and those that do not add considerably to the value of the model are indifferent or pruned.

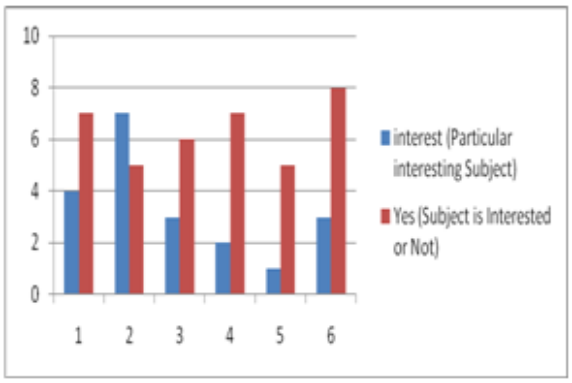

Fig.3.2.shows the result of Student Performance using C5.0 Techniques

In C5.0 Algorithms, there are some improvements in terms of handling the bias toward tests with many outcomes presentation and pruning. Gain Ratio is used for these algorithms and its extension to ID3Calculations which has a kind of normalization to Information Gain using Spilt Info values [9].

D) Naive bayes

Naïve bayes classification technique is based on Bayes Theorem with an assumption of independence among predictors. It is very simple, which assumes that the classification attributes are independent and they do not have any connection between them. A lot of researchers have found that this assumption of liberty do not workin the entire cases for which other different methods is proposed to raise the presentation. The creative Naïve Bayesian technique is based on the conditional probability and the maximum likelihood incidence. The Naïve Bayesian algorithm based on the description provided in [13].

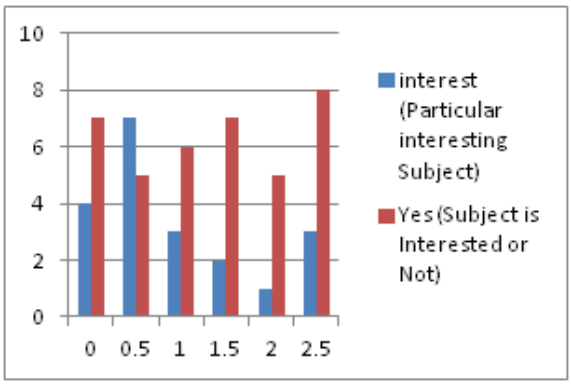

Fig.3.3. shows the result of Student Performance using Naïve Bayes Techniques

This process goes on until all the data classification is perfectly classified or run out of attributes. The knowledge represented in the form of IF-THEN rules are Pruning technique was executed by removing nodes with less than preferred quantity of substance.

.

4.Conclusion

In Existing methods have used features which are mostly associated to educational presentation, family income and family resources, while features belonging to family expenditures and students' individual information are usually ignored. In this article, an effort is made to examine aforementioned feature sets by collecting the scholarship holding students' data from different universities of India. Knowledge analytics, discriminative and generative classification models are applied to expect whether a student will be able to absolute his degree or not. We introduce the Weka tool and C4.5 algorithm to analyze replicating studies and discuss the importance of replicating and reproducing previous work. We describe the state of the art in collecting and sharing programming data. To better know the challenges concerned in replicating or reproducing presented studies.

References :

[1]. Behdad Bakhshinategh, Osmar R. Zaiane and et al., "Educational data mining applications and tasks: A survey of the last 10 years", Springer Science + Business Media New York 2017, Educ Inf Technol, DOI 10.1007/s10639-017-9616-z, 3 July 2017.

[2]. Yun-En Liu, Travis Mandel, Emma and et al., "Towards Automatic Experimentation of Educational Knowledge", Session: Games and Education CHI 2014, One of a CHInd, Toronto, ON, Canada, Pp.3349-3358.

[3]. Petri Ihantola, Arto Vihavainen and et al., "Educational Data Mining and Learning Analytics in Programming: Literature Review and Case Studies", ITiCSE WGR'16, July 4–8, 2015, Vilnius, Lithuania, c 2016 Copyright held by the owner/author(s), ACM ISBN 978-1-4503-4146-2/15/07, DOI: http://dx.doi.org/10.1145/2858796.2858798, Pp.41-64.

[4]. Carmela A. White, Daniela Wong Gonzalez, "Meta-analysis of faculty's teaching effectiveness: Student Evaluation of teaching ratings and student learning are not related", http://dx.doi.org/10.1016/j.stueduc.2016.08.007 0191-491X/ã 2016 Elsevier Ltd.

[5]. Li Wei, Shen Xiaoling and et al., "Enhancing the Student Teamwork Ability and Innovation Ability in Blended Teaching", 978-1-5386-5495-8/18/$31.00 ©2018 IEEE , The 13th International Conference on Computer Science & Education (ICCSE 2018) August 8-11, 2018, Colombo, Sri Lanka.

[6]. Hutchatai Chanlekha, Jitti Niramitranon "Student Performance Prediction Model for Early-Identification of At-risk Students in Traditional Classroom Settings", MEDES'18, Sept, © 2018 Copyright held by the owner/author(s). 978-1-4503-5622-0/18/092018, Tokyo, Japan, Pp. 239-245.

[7]. Ali Daud, Naif Radi Aljohani, and et al., "Predicting Student Performance using Advanced Learning Analytics" WWW2017 Companion, April 3-7, 2017, Perth, Australia. ACM 978-1-4503-4914-7/17/04. DOI: http://dx.doi.org/10.1145/3041021.3054164, Pp.415-421.

[8]. George Karypis "Improving Higher Education—Learning Analytics & Recommender Systems Research", RecSys'17, August 27–31, 2017, Como, Italy, RecSys'17, August 27–31, 2017, Como, Italy © 2017 Copyright is held by the owner/author(s). ACM ISBN 978-1-4503-4652-8/17/08. DOI: http://dx.doi.org/10.1145/3109859.3109870.

[9]. Nobuhiko Kondo, Midori Okubo and et al., "Early Detection of At-Risk Students Using Machine Learning Based on LMS Log Data", 2017 6th IIAI International Congress on Advanced Applied Informatics, 978-01-76955386-61780621-36/17 $31.00 © 2017 IEEE, DOI 10.1109/IIAI-AAI.2017.51, Pp.198-200.

[10]. Chinchu ThomasDinesh Babu Jayagopi "Predicting Student Engagement in Classrooms using Facial Behavioral Cues", MIE'17, November 13, 2017, Glasgow, UK www.2017 Association for Computing Machinery. ACM ISBN 978-1-4503-5557-Pp.33-40.