Efficient prediction of Student Performance Using hybrid SVM Classifier

Ms.Vineetha K.R
Ph.D Research Scholar,
Department of Computer Applications
Nehru College of Management,
Coimbatore, Tamil Nadu
vpvprakash@gmail.com

Dr.E.Chandra Blessie
Associate Professor
Department of Computer Applications
Nehru College of Management,
Coimbatore, Tamil Nadu
chandra_blessie@yahoo.co.in

Abstract — **Academic data processing (ADP) and Learning Analytics (LA) research have emerged as attention-grabbing areas of analysis, which are development helpful information from academic databases for many functions like predicting students' success. the power to predict a student's performance are often helpful for actions in fashionable academic systems. Existing strategies have used features that are principally associated with educational performance, family financial gain and family assets; whereas options happiness to family expenses and students' personal info square measure usually unnoticed. during this paper, a trial is created to research aforementioned feature sets by assembling the scholarship holding students' information from completely different.**

Learning analytics, differentiative and propagative classification models are applied to predict whether or not a student are going to be able to complete his degree or not. Experiments confirms that proposed research significantly outperforms existing approaches due to exploitation of family expenditures and students' personal information feature sets. Outcomes of this ADP/LA analysis will serve as policy improvement technique in educational activity

Keywords - Academic data processing (ADP) and Learning Analytics, performance, propagative classification models

1. INTRODUCTION

Student performance prediction during a case technique course is also assessed on a range of dimensions as well as category participation, individual written work on papers and exams, and cluster activities like comes and shows. Our focus here is on category participation, that is integral to the case technique and sometimes accounts for a major portion of a student's grade. Knowledgeable case coaches estimate class contribution maintained a student's contribution to the collective learning throughout category discussions. Establishing objective assessments of those contributions are often difficult. the standard of individual contributions relates not solely to the content, however additionally the delivery and temporal order of comments at intervals the flow of the category discussion. additional frequent participation is usually a positive issue, though excessive tries to comment could result in lower quality contributions and should mirror a bias toward speaking over listening. In assessing participation, instructors ought to remember the crucial role they play in shaping student performance through line patterns and also the sorts of queries and follow-ups they use with individual students. Also, the standard of the instructor's participation chase system could considerably have an effect on the dependableness of the general performance analysis.

From a student perspective, the participant-centred nature of the case technique generates bigger expectations and opportunities for feedback as compared to lecture-based pedagogies. As students participate in school discussions, they receive immediate feedback within the kind of teacher and student responses to their contributions. this kind of feedback, however, is also ambiguous and indirect, going away students unsure of the impact of their participation and the way they could enhance their effectiveness. To some extent, this is often not a nasty issue, since it encourages students to develop their own capabilities for reflection and self-assessment. Students could

actively ask for extra feedback from peers and also the teacher outside of sophistication. Ideally, instructors are ready to give each critical and organic process feedback in a manner that helps students discover additional insights concerning their strengths and opportunities for improvement.

II. EXISTING SYSTEM

The analysis drawback of students' performance prediction is analysed through numerous angles. within the current literature, a variety of complementary approaches offer a baseline for such Associate in the Nursing analysis. In a perfect situation, a chic dataset with student identity together with various characteristics may be the idea for advanced learning analytics. the matter is that in most of the cases, not all the info area unit obtainable for the dynamic construction of the code identity, additional restricted by lack of access to numerous sources. we tend to shortly gift a number of the foremost representative strategies of Applied academic data processing and Learning Analytics supported a comprehensive literature review.

Student performance prediction possesses lots of attention from the academic data processing researchers. Typical data processing strategies are utilized to alter totally different tasks associated with the scholars. A survey of information mining techniques for ancient academic systems like adaptational web-based and content management systems is conferred.

An association rule primarily based mining technique is applied for a choice of weak students in a very faulty and is found effective. Genetic algorithmic rule is employed to assign the weights for the modelling of students' grade for 3 levels (binary, 3-level, and 9-level). It shows that the mixture of multiple classifiers results in a big improvement in classification. A model is planned for predicting student performance mistreatment six machine learning techniques for distance learning education, that is kind of totally different from the normal academic system. The experimental results show that demographic and performance options area unit higher predictors for predicting student performance. A regression model is applied to predict the take a look at the score of a subject for varsity students. It concludes that mixed-effect models gift the best performance as compared to the Bayesian network.

A prediction model (CHAID) is developed to predict the performance of upper Gymnasium students, that is vital before obtaining admission into universities.

The grades of graduate student's area unit expected mistreatment Naïve Bayesian and Rule Induction classifiers. Clusters area unit made of students' information and therefore the outliers area unit with success known. A model is conferred to estimate the talents of scholars and ability of academics so as to predict the long run student outcomes. It shows that demographic profiles and temperament traits options area unit correlative and have a high impact on student performance. Equally, student performance analysis and engineering students' skills area unit analysed for improved accomplishment method by mistreatment data processing strategies.

III. PROPOSED RESEARCH

This paper addresses the problem of students' performance prediction by presenting new features, mostly related to family expenditure and student personal information. The researchers have used some basic characteristics related to the student personal information like family income and family assets information. Therefore, it is required to introduce some influential and effective features of students for performance evaluations in their studies. In this paper, family expenditures (electricity, telephone, gas bills, accommodation and medical) and student personal information features (e.g. self-employed and marital status) are explored. One of the greatest challenges for future Digital Learning Research in WWW is to investigate flexible and reliable methods for the extraction and integration of learners' data from diverse sources in order to support advanced educational decision making.

IV. METHODOLOGY

Two types of classification models (discriminative and generative) are used to learning the desired predictive function $\boldsymbol{F}$ **(.)**. Two generative and three discriminative models are used for experimental analysis. They are selected on the basis of their frequent usage in the existing literature. The list of methods are as follows:

1. Support Vector Machine (SVM) [discriminative]

A Support Vector Machine (SVM) is a discriminative classifier properly well-defined by a extrication hyper plane.

A. Set up the training data: The training data of this exercise is formed by a set of labelled 2D-points that belong to one of two different classes; one of the classes consists of one point and the other of three points.

B. Set up SVM's parameters: In SVM the training examples are spread into two classes that are linearly separable. However, SVMs can be used in a wide variety of problems (e.g. difficulties by non-linearly divisible data, an SVM using a kernel role to increase the dimensionality of the samples, etc). As a consequence of this, we have to define some parameters before training the SVM.

C. Regions classified by the SVM: The method which is used to classify an input sample is using a trained SVM. In this example we have used this method in order to colour the space depending on the prediction done by the SVM. In further arguments, an image is traversed construing its pixels as arguments of the Cartesian level. Each of the points is coloured depending on the class predicted by the SVM; in green, if it is the class with label 1 and in blue if it is the class with label -1

2. C4.5 [discriminative]

C4.5 and naive mathematician (NB) are 2 of the highest ten data mining algorithms because of their simplicity, effectiveness, and potency. it's accepted that NB performs alright on some domains, and poorly on others that involve correlate options. C4.5, on the opposite hand, generally works higher than NB on such domains. To integrate their benefits and avoid their disadvantages, several approaches, like model insertion and model combination, are projected. The model insertion approach like NBTree inserts NB into every leaf of the engineered call tree. The prototypical grouping method similar C4.5-NB shapes C4.5 subordinate degreed NB on a education dataset individually so associations their forecast outcomes for an unobserved illustration. during this paper, we have a tendency to specialise in a brand new read and propose a discriminative model choice approach. For detail, at the coaching time, C4.5 and NB square measure engineered on a coaching dataset severally, and therefore the most reliable one is recorded for every coaching instance. At the take a look at time, for every take a look at instance, we have a tendency to first off realize its nearest neighbour so select the foremost reliable model for its nearest neighbour to predict its category label.

3. Classification and Regression Tree (CART) [discriminative]

Decision Trees are an important type of algorithm for predictive modelling machine learning.

The classical decision tree algorithms have been around for decades and modern variations like random forest are among the most powerful techniques available.

In this post you'll discover the decision tree formula familiar by its additional fashionable name CART that stands for Classification and Regression Trees. once reading this post, you'll know:

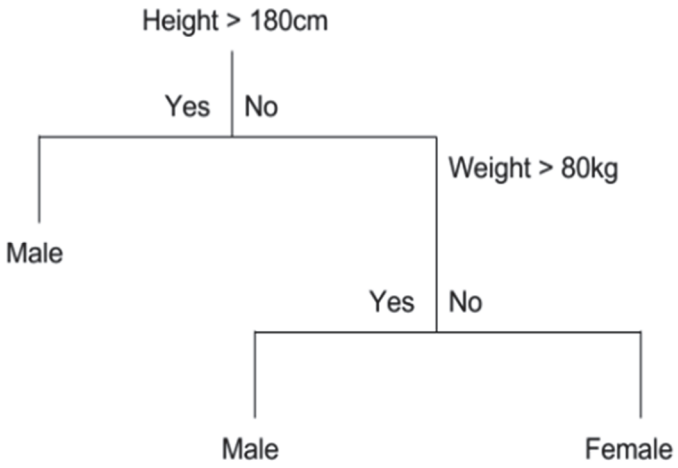
- The several names accustomed describe the CART formula for machine learning.
- The illustration utilized by learned CART models that's truly keep on disk.
- How a CART model may be learned from coaching information.
- How a learned CART model may be accustomed create predictions on unseen information.
- Additional resources that you simply will use to find out additional regarding CART and connected algorithms.

If you have got taken associate degree algorithms and information structures course, it'd be laborious to carry you back from implementing this easy and powerful formula. And from there, low step removed from your own implementation of Random Forests.

3.1. CART Model Representation

The illustration for the CART model could be a binary tree.

This is your binary tree from algorithms and information structures, nothing too fancy. every root node represents one input variable (x) and a split purpose on it variable (assuming the variable is numeric). The leaf nodes of the tree contain associate degree output variable (y) that is employed to form a prediction.
Given a dataset with 2 inputs (x) of height in centimetres and weight in kilograms the output of sex as male or feminine, below could be a crude example of a binary call tree (completely fictitious for demonstration functions only).



With the binary tree illustration of the CART model delineated on top of, creating predictions is comparatively simple. Given a replacement input, the tree is traversed by evaluating the precise input started at the foundation node of the tree. A learned binary tree is truly a partitioning of the input house. you'll consider every input variable as a dimension on a n-dimensional house. the choice tree split this up into rectangles (when p=2 input variables) or some reasonably hyper-rectangles with additional inputs. New information is filtered through the tree and lands in one among the parallelograms and therefore the output price for that rectangle is that the prediction created by the model. this offers you some feeling for the sort of selections that a CART model is capable of creating, e.g. three-dimensional call boundaries.

3.2 Learn a CART Model From Data

Creating a CART model involves choosing input variables and split points on those variables till an appropriate tree is built. The selection of that input variable to use and therefore the specific split or cut-point is chosen employing a greedy formula to reduce a price perform. Tree construction ends employing a predefined stopping criterion, like a minimum range of coaching instances appointed to every leaf node of the tree.

4. Bayes Network (BN) [generative]

probabilistic directed acyclic graphical model may be a probabilistic graphical model that represents a collection of variables and their conditional dependencies via a directed acyclic graph (DAG). for instance, a theorem network might represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network will be wont to figure the possibilities of the presence of assorted diseases.

Formally, theorem networks area unit DAGs whose nodes represent variables within the theorem sense: they'll be noticeable quantities, latent variables, unknown parameters or hypotheses. Edges signify restricted dependencies; nodes that aren't associated represent variables that area unit not absolutely freelance of every other. every node is related to a likelihood operate that takes, as input, a selected set of values for the node's parent variables and provides (as output) the likelihood (or likelihood distribution, if applicable) of the variable described by the node. for instance, if m parent nodes represent m mathematician variables then the likelihood operate might be described by a table of 2^ entries, one entry for every of the 2^ potential combos of its folks being true or false. Similar concepts could also be applied to planless, and probably cyclic, graphs; like Markov networks.

Efficient algorithms exist that perform illation and learning in theorem networks. theorem networks that model sequences of variables (e.g. speech signals or macromolecule sequences) area unit known as dynamic theorem networks. Generalizations of theorem networks that may represent and solve call issues beneath uncertainty area unit known as influence diagrams.

5. Naive Bayes (NB) [generative]

Naïve Thomas Bayes doesn't model chance directly. It models the chance, and subsequently it calculates p(y|x). We're inquisitive about the p(y|x) wherever y will take associate degree example whether or not an e-mail is spam or not spam, x vector denotes the words during a specific document. From Thomas Bayes Formula, p(y|x) = p(x|y)p(y)/p(x). thus, if you've got all those stuff in your hand, you'll generate the info. Here is that the generative story of this model: we have a tendency to initial choose a y, that indicates our generating e-mail is whether or not spam or not. Bearing in mind y's worth, we have a tendency to generate words per conditional distribution p(x|y). Assume that we have a tendency to generate few words. once {do we have a tendency to stop? Whenever x word that we generate is adequate STOP_EMAIL word, we have a tendency to end choosing word for that e-mail. As a result, we will generate AN e-mail. as a result of it computes p(x,y)p(x,y) instead of p(y|x)p(y|x).

Although it appears like you are examination p(y|x)p(y|x) on totally different yy's, you are truly calculative it by dividing the joint by the previous distribution of xx:

p(y|x)=p(x,y)p(x)p(y|x)=p(x,y)p(x)

The discriminative counterpart of NBC is logistical regression.

V. EXPERIMENTAL RESULTS

The results show that the approach of exploitation multiple information sources alongside heterogeneous ensemble techniques is extremely economical and correct in prediction of student performance yet as facilitate in correct identification of student in danger of attrition.

A. Practical implications

The approach projected during this study can facilitate instructional, the tutorial and the academic directors and policy manufacturers operating at intervals educational sector within the development of latest policies and information on teaching that square measure relevant to student retention. additionally, the overall implications of this analysis to observe is its ability to accurately facilitate in early identification of scholars in danger of quitting of HE from the mix of information sources in order that necessary support and intervention may be provided.

B. Originality / value

The analysis through empirical observation investigated and compared the performance accuracy and potency of single classifiers and ensemble of classifiers that create use of single and multiple information sources. The study has developed a unique hybrid model which will be used for predicting student performance that's high in accuracy and economical in performance. Generally, this analysis study advances the understanding of the appliance of ensemble techniques to predicting student performance exploitation learner information and has with success self-addressed these elementary questions: What combination of variables can accurately predict student educational performance? what's the potential of the employment of stacking ensemble techniques in accurately predicting student educational performance

This analysis work presents the code tutorial prediction ways that use four differing types of options namely: family expenditure, family financial gain, student personal data and family assets. It additionally adapts the method of feature set choice so as to spot the foremost effective determinants for student tutorial performance prediction. it's evident from the comparative analysis that our planned options square measure necessary predictors and achieved F1-score of eighty-six on real world college man students' knowledge.

Table 1 Features Distribution.

Category	Name	Description	Status	Info. Gain	Gain Ratio	Features Used
Family Income	Father Income	Per month income of father/guardian of student	Old	0.29	0.04	√
	Mother Income	Per month income of mother of student	Old	0.02	0.03	√
	Land Income	Per month income from land of family of student	Old	0.02	0.05	√
	Miscellaneous Income	Per month miscellaneous income of family of student	Old	0.08	0.03	√
	Earning Hands	Total number of Earning hands of student's family	Old	0.007	0.005	
	Father Status	Status of father of student: alive or deceased	New	.0008	0.001	
	Father Retired	Father retired or in service	New	0.002	0.003	
	Guardian Alive	Is student's guardian alive	New	0.003	0.004	
Student Personal Information	Gender	The gender of the student (male or female)	Old	0.004	0.005	
	Marital Status	Marital status of student (married or unmarried)	New	0.003	0.01	√
	House Owner Ship	Student have his/her own house	New	0.08	0.10	√
	Previous Program Scholarship	Scholarship received or not in previous academic program	New	.0002	.0003	
	Previous Institution Type	Type of student previous institution	Old	0.001	0.002	
	Self Employed	Is student is self employed	New	0.06	0.04	√
Family Assets	Land Value	Current value of lands belongs to student's family	Old	0.04	0.02	√
	Bank Balance	Bank balance of student's family	Old	0.05	0.07	√
	Stock Value	Value of Shares/Bonds belong to student' s family	Old	0.01	0.08	√
	House Value	Value of house belong to student's family	Old	0.14	0.03	√
	House Condition	Structure of house belong to student's family	New	0.06	0.04	√
	Miscellaneous Asset Value	Any other assets related to student	Old	0.04	0.02	√
	Location	Type of Location where student resides; urban or rural	Old	0.03	0.04	√
	No of Vehicles at home	How many vehicles belong to family of a student	Old	0.005	0.008	

TABLE 1. FEATURES DESCRIPTION

The options associated with family expenditure like fossil fuel, electricity, telephone, water, accommodation, miscellaneous expenditures, and most significantly family expenditure on education square measure found to be handiest in predicting tutorial performance. Most of those options are neglected by the baseline ways and previous studies. the simplest prognosticative performance is obtained once family expenditure based mostly options square measure combined with alternative options (hybrid features). it's been ascertained that family expenditures have an effect on the students' performance and scale back their concentration and interest in studies. The claims created on the premise of experimental outcomes are verified by twenty-five students learning on scholarships. Most of those students comply with the subsequent discussions.

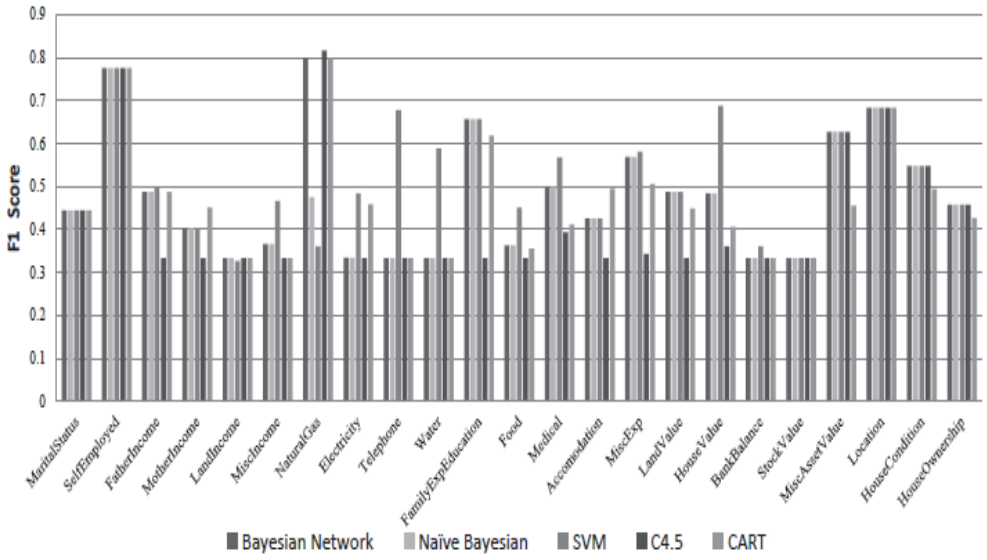
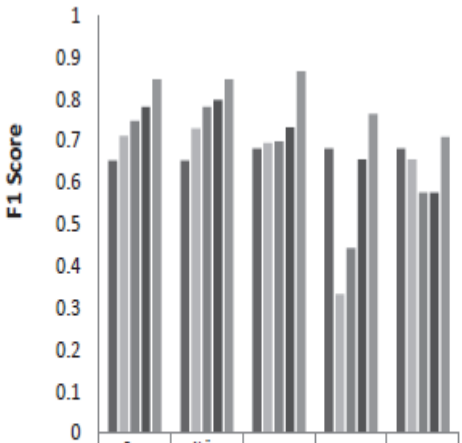


FIGURE 1. IMPACT OF SELECTED INDIVIDUAL FEATURES ON CLASSIFICATION ACCURACY

An increase within the expenditures of family reduces the opportunities for a student to become older and surpass in their studies as a result of time and cash square measure necessary factors in life and square measure directly connected with the family expenditure. the rise in expenditure, particularly on medical treatments and accommodations dominantly affects the performance of scholars. a lot of expenditure on medical relates to health problems and a lot of expenditure on accommodation could have an effect on the budget (for education) of a social class family.



	Bays Network	Naive Bayesian	SVM	C4.5 (j48)	CART
Tair and Al Haleej(2012)	0.653	0.653	0.683	0.683	0.683
Osmanbegovic and Suljic (2012)	0.713	0.731	0.695	0.333	0.657
Sree, et al.(2013)	0.748	0.782	0.699	0.444	0.577
Ramaswami and Bhaskaran (2010)	0.782	0.799	0.733	0.656	0.577
Hybrid Features (Proposed)	0.848	0.848	0.867	0.766	0.71

FIGURE 2. COMPARISON WITH BASELINE AND PROPOSED FEATURES.

On the opposite hand, some personal characteristics of scholars also are necessary predictors for his or her performance analysis, e.g., married students higher focus on their studies as compared to bachelor students may be due to emotional stability in their personal lives. constant is that the case with students UN agency themselves or their oldsters have their own property. Families having their own house positively saves cash by not paying house rent and may pay these savings for the education of their youngsters. They additionally don't get to keep ever-changing the rented homes which can waste time and energy of a student. Similarly, the self-employment standing of a student permits him/her to higher schedule time for studies in AN economical means as a result of less worries concerning finances end in comfort and satisfaction. additionally, to the current, self-employment develops exhausting operating perspective within the personalities of the scholars, each these factors square measure terribly useful for college students in achieving higher performance in their studies. Last however not the smallest amount, the house condition of a student is additionally an important issue as a result of having a comfortable living accommodation permits him/her to higher utilize his/her skills in studies. On the opposite hand, if the house condition isn't smart, the student's time and energy could also be wasted in repairs or in serving to his/her oldsters to induce the repairs done. Previous studies have additionally explored the gender and establishment kind characteristics, that we tend to don't contemplate.

VI. CONCLUSION

In this analysis, an attempt is created to search out the impact of our projected options on student performance prediction with the assistance of generative and discriminative classification models. A feature house is made by considering characteristics of family expenditure, family financial gain, personal info and family assets of scholars. The potential/dominant options choice is un-avoidable because it provides U.S. with set of options. SVM classifier is found effective for our projected options of family expenditure and student personal info classes. It will be ended from the results that family expenditure and private info options have important impact on the performance of the scholar attributable to intuitive reasons provided in discussions.

In future, World Wide Web analysis on Digital Learning and Learning Analytics ought to be targeted on the subsequent directions:

- that sorts of strategies and versatile applications allow the development of essential learners' knowledge from the World Wide Web, e.g. mining of social media content will be a basis for private expenditure?
- that are the chances to proceed to dynamic identification of non-public characteristics of scholars from the Deep Web?
- that are the standards for codifying essential students' info within the World Wide Web, and the way this will envision future World Wide Web based mostly learning services?
- Learning analytics of mobile and present learning environments from the angle of human laptop interaction conjointly need careful exploration additionally to said ancient and net based mostly options.

ACKNOWLEDGEMENT

We would like to thank the reviewers for their insightful and constructive Comments. We thank QBiT Technologies, Coimbatore, India, for the technological and creative support of our work. We also thank Mr. Mohana Maniganda Babu V and Mr. Sakthi Sivakumar V for the discussion on algorithms.

REFERENCES

[1] Ali Daud, Naif Radi Aljohani , Rabeeh Ayaz Abbasi , "Predicting Student Performance using Advanced Learning Analytics" International World Wide Web Conference Committee (IW3C2), WWW 2017 Companion, April 3-7, 2017, Perth, Australia.
[2] N. R. Aljohani and H. C. Davis, "Learning analytics in mobile and ubiquitous learning environments," in 11th World Conference on Mobile and Contextual Learning, 2012.
[3] N. R. Aljohani, H. C. Davis, and S. W. Loke, "A comparison between mobile and ubiquitous learning from the perspective of human-computer interaction," International Journal of Mobile Learning and Organization, vol. 6, no. 3/4, pp. 218-231, 2012.
[4] R. Asif, A. Merceron, and M. K. Pathan, "Investigating performance of students: a longitudinal study," in Fifth International Conference on Learning Analytics And Knowledge (LAK '15), New York, USA, 2015, pp. 108-112.
[5] M. A. Chatti, A. L. Dyckhoff, U. Schroeder, and H Thüs, "A reference model for learning analytics," International Journal of Technology Enhanced Learning (IJTEL), vol. 4, no. 5/6, pp. 318-331, 2012.
[6] N. Fournier, R. Kop, and H. Sitlia, "The value of learning analytics to networked learning on a personal learning environment," in 1st International Conference on Learning Analytics and Knowledge, 2011, pp. 104-109.
[7] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting students' performance in distance learning using machine learning techniques," Applied Artificial Intelligence, vol. 18, no. 5, pp. 411-426, 2004.
[8] E. Lotsari, V. Verykios, C. Panagiotakopoulos, and D. Kalles, "A Learning Analytics Methodology for Student Profiling," in Artificial Intelligence: Methods and Applications, 2014, pp. 300-312.
[9] Y. Ma, B. Liu, C. K. Wong, P. S. Yu, and S. M. Lee, "Targeting the right students using data mining," in 6th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD '00), New York, USA, 2000, pp. 457-464.
[10] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch, "Predicting student performance: an application of data mining methods with an educational Web-based system," in 33rd Annual Frontiers in Education (FIE 2003), Westminster, CO, 2003.
[11] T. Mishra, D. Kumar, and Sangeeta Gupta, "Students' Employability Prediction Model through Data Mining," International Journal of Applied Engineering Research, vol. 11, no. 4, pp. 2275-2282, 2016.
[12] E. Osmanbegović and M. Suljić., "Data mining approach for predicting student performance," Economic Review, vol. 10, no. 1, pp. 3-12, 2012.
[13] O. K. Oyedotun, S. N. Tackie, and Ebenezer O. Olaniyi, "Data Mining of Students' Performance: Turkish Students as a Case Study," International Journal of Intelligent Systems and Applications, vol. 7, no. 9, pp. 20-27, 2015.
[14] Z. A. Pardos, N. T. Heffernan, B. Anderson, C. L. Heffernan, and W. P. Schools, "Using fine-grained skill models to fit student performance with Bayesian networks," in Handbook of educational data mining., 2010, pp. 417-426.
[15] P. J. Piety, D. T. Hickey, and M. J. Bishop, "Educational data sciences: Framing emergent practices for analytics of learning, organizations, and systems," in 4th International Conference on Learning Analytics and Knowledge, 2014, p. 193.