Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

Special Issue

SYSTEMATIC MAPPING REVIEW ON STUDENT'S PERFORMANCE ANALYSIS USING BIG DATA PREDICTIVE MODEL

S. M. Muthukrishnan[1,*], M. K. Govindasamy[2] and M. N. Mustapha[1]

[1]Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

[2]International Languages Teacher Education Institute, Malaysia

Published online: 05 October 2017

ABSTRACT

This paper classify the various existing predicting models that are used for monitoring and improving students' performance at schools and higher learning institutions. It analyses all the areas within the educational data mining methodology. Two databases were chosen for this study and a systematic mapping study was performed. Due to the very infant stage of this research area, only 114 articles published from 2012 till 2016 were identified. Within this, a total of 59 articles were reviewed and classified. There is an increased interest and research in the area of educational data mining, particularly in improving students' performance with various predictive and prescriptive models. Most of the models are devised for pedagogical improvements ultimately. It is a huge scarcity in producing portable predictive models that fits into any educational environment. There is more research needed in the educational big data.

Keywords: predictive analysis; student's performance; big data; big data analytics; data mining; systematic mapping study.

Over the last 10 years, schools and higher learning institutions have built immense databases that contain various student related information ranging from demographics information to exam results.

Author Correspondence, e-mail: cybersri@siswa.um.edu.my

doi: http://dx.doi.org/10.4314/jfas.v9i4s.41

Many have used this data retroactively to access students' performance and predict future outcomes, and engage in empirical analyses to determine underperforming students so that a proper intervention programs can be introduced. The massive collection of data in the education space has contributed to big data. The big data is currently used extensively to derive decisions from data analytics which popularly known as predictive analysis.

Educause [70], a nonprofit organization defines predictive analytics as "an area of statistical analysis that deals with extracting information from various technologies to uncover relationships and patterns within large volumes of data that can be used to predict behavior and events". Within the analytics, there are three categories: descriptive-description of data that we are dealing with; predictive-includes statistical modeling and data mining techniques that uses data to predict future events; prescriptive-provides dynamic model that will make recommendation about future events and how to address them.

The key component of a predictive analytics is the model formation for accurately predicting an event, in which for this paper is predicting the students' performance. The key issue that typically being addressed in this area is identifying and assist student-at-risk of failing or dropping out before they actually do so, ultimately improving students passing rate. In addition to addressing the drop-out cases, the predictive analytics have been used in many other areas within the educational space, mainly in schools and higher learning institutions. Among them are students' enrollments in terms of appropriate placement, behavioral monitoring [66], learning engagement, recommender systems for services and courses and many more.

In the last 4 years, many researches have been carried out in the area of Predictive Analytics to develop and improve the predictive models to accurately predict students' performance so that a proper intervention programs can be devised. One example is providing guidance to choose a right degree program based on the qualification one have [27]. Many modelling techniques and different approaches were tested extensively, chiefly clustering, regression, neural networks, decision trees, semantic, random forests and support vector machines [67].

The input or parameters for the models ranging from students' demographics, economic background, locality, previous results, class participation (both online i.e. VLE and offline), school and recently teachers' background is also included.

The main objective of this systematic mapping study is to gauge the existing predictive analytics models within the educational space of schools and other learning institutions. To the best of our knowledge, no systematic mapping study has been done on predictive analytics for student's performance improvement in a big data environment. However, there are similar researches worded in as 'research travelogue' [71] and surveys [3, 72-73]. But, these researches have not discussed or hardly discussed anything on big data and its implication on students' performance prediction model. Thus, the present study will be thorough on the various predictive models available within the educational data mining landscape and particularly with consideration to big data roles. Further, this study will identify previous researches that have employed features other than the common ones (grades etc.) in developing a predictive model like socio-demographic attributes. In [20] endorses there is strong relationship between students' performance and other various variables such as school grades, school type, school performance, socio-economic deprivation, neighbourhood participation, ethnicity, parents background and other areas.

2. METHODOLOGY

A review protocol specifies the methods that will be used to undertake a specific systematic review and reduces the possibility of researcher bias [4]. For this study, we on purpose chose systematic mapping study over systematic review study. A systematic mapping study provides an objective procedure for identifying the nature and extent of the research that is available to answer a particular research question. These kinds of studies also help to identify gaps in current research in order to suggest areas for further investigation. They therefore also provide a framework and background in which to appropriately develop future research activities [5]. In contrast, systematic reviews aim to address problems by identifying, critically evaluating and integrating the findings of all relevant, high-quality individual studies addressing one or more research questions [6]. While a systematic mapping review is a useful product in its own right and describes the kinds of research that have been undertaken within a particular field of study, it also provides an overview of a research area, highlighting areas in which empirical research has been conducted and aiding the identification of knowledge gaps.

Performing systematic mapping review for this study involves several discrete activities which can be grouped into three main phases namely 1-Research directives, 2-Data collection and 3-Results. This process is depicted in Fig. 1. The model was chosen based on previous study done by [7], which adapted model from [8].

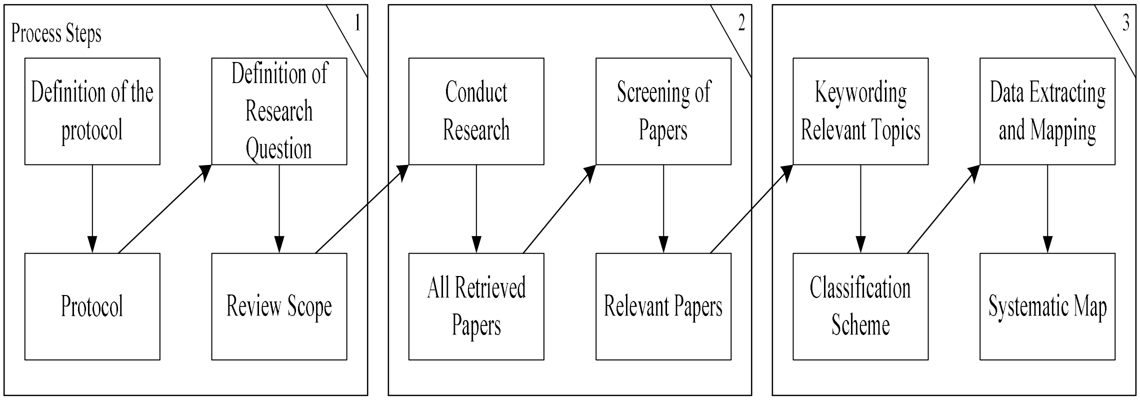

Fig.1. The systematic mapping process (adapted from [8]

3. SYSTEMATIC MAPPING PROCESS

In this section, we will discuss on the main process of our systematic mapping study as suggested by [8]. As per Fig. 1, the process of systematic mapping includes research questions, conducting the search and screening of papers, data extraction and mappings.

3.1. Research Questions

The main goal of this paper is to determine the type of predictive modelling that has been used to predict student's performance. We review articles on predictive modelling in big data analytics. By what we have agreed upon, we formulated these research questions:

- **RQ1**: What are the predictive models in data mining that has been used to predict students' performance? With so many models, we want to know what are the proven models for educational data mining and why are they chosen.
- **RQ2**: What are the precise purposes of the predictive model used for within student's performance improvement context? This will provide insights into what exactly the models used for and how they will benefit the stakeholders
- **RQ3**: Is big data considered in student performance prediction model? We want to know how many papers consider the big data role and impact when looking at students' performance predictions.
- **RQ4**: What are the key attributes mainly used in the predictive model and is there a selection process? With so many students' parameters how are they organized, categorized and chosen for optimal performance prediction model.

3.2. Search Strategy and Data Sources

From the research questions, we extracted the keywords to perform a primary query in digital library. The purpose of the keywords is to make the searching task easier to cover larger proportion of published papers in the area that we are looking for. For the purpose of this study, we used electronic procedures instead of manual procedures as the electronic mode of search offered free access to online databases and user can search through thousands of papers within short period of time [9]. There are a lot of online databases available such as IEEEXplore, ACM Digital Library, Elsevier Science Direct, Scopus, Web of Science and etc. We decided to search in only two databases which are IEEEXplore and Elsevier Science Direct because these two databases are among the important databases in computer science research.

By using the online databases, we then filtered papers which were published since 2012 to 2016. This is because we need to look into the most recent trend in prediction model or strategy in analyzing student's performance. The latest techniques could have resolved many issues thus the predictive analytics in measuring student's performance need to evolve. Also to note is that the papers in 2017 were not included as part of the search strategy as this paper was finalized in the beginning of 2017. The search strategy was further developed into reviewing the data needed to answer each of the research questions. The search keywords have to be constructed to avoid too many findings which would not be relevant to the research objectives.

We used several keywords in combination. These are Educational Data mining, Big Data Predictive Model, Predictive Analysis, Model and Student's Performance. Initially we omitted the word "Educational" in the "Data Mining" which resulted in very fewer results that are relevant to this current research which emphasize on data mining in educational institutions. However, upon adding "Educational" into "Data Mining", many results were returned that are relevant to the current research. As such, it is important to search using "Educational Data Mining (EDM)" as the keywords in the 2 databases. It is a popular key word for any research with regards to education. By mentioning 'student's performance' it is obvious that the 'education' will come into the picture. We decided to exclude 'education' keyword because we would like to focus on the technique or model or framework from computer science perspective. Putting 'education' would likely skew the research into transforming the education system. It will be like providing Big Data education or students' education.

In order to sort the searching technique into more systematic, we used the Boolean operator such as 'AND' and 'OR' or combination of both operator. Using these Boolean operator, it could improve the completeness of the result and we could size down to the relevant papers. The complete list of the search string is depicted in Table 1.

Table 1. List of research strings

No.	Research Strings
1	Big Data AND Data Analytic
2	Big Data AND Predictive Analytic
3	Educational Data Mining AND Data Analytic
4	Predictive Analytic AND Student's Performance
5	(Predictive Analytic OR Data Analytic) AND "Student's Performance"
6	Educational Data Mining AND "Student's Performance"
7	Data Analytic AND "Student's Performance"
8	(Educational Data Mining OR Data Analytic) AND "Student's Performance"
9	Predictive Analysis AND Modelling
10	"Student's Performance" AND Predictive Analysis AND Modelling
11	Prediction AND "Student's Performance"
12	(Predictive OR Prediction) AND "Student's Performance"
13	Educational Data Mining AND (Technique OR Model OR Framework OR Method)
14	Predictive Analytic AND (Technique OR Model OR Framework OR Method)
15	Predictive Analytic AND (Technique OR Model OR Framework OR Method) AND "Student's Performance"

3.3. Screening

Duplicates of papers were removed during this phase. The selected papers were further shortlisted using a set of inclusion/exclusion criteria. Firstly, we check for the relevancy of title and abstract. The paper was included if it:

- introduces a model or framework of predictive modelling.
- explain in terms of why they choose specific modelling or analytic techniques.
- specifies the sample use in the study.
- predict student's performance based on the model or framework introduced.
- interpret the results of the predictive analysis.
- studies are in the field of computer science.

Paper was excluded if:

- it does not introduce any model or framework about predictive modelling.
- no abstract or full text is available.
- complement more to education area rather than computer science study.

- The result of the predictive analysis was not being elaborate based on the techniques chosen.
- The prediction model or framework does not relate to student's performance.
- The paper is too brief and only has a keynote abstract.

Each paper was screened by the research team to ensure the paper meet the inclusion/exclusion criteria. Any differing opinions were discussed to reach a consensus decision. The selected papers from this phase were then examined one more time to ensure that they absolutely fit into the mapping study criteria. The results of this phase will be elaborated in section 5.

3.4. Classification Scheme

Classification of the papers was done by using two methods; one is looking from top down approach, where we define the general knowledge of the current research area. Second method is the bottom-up approach consists of extracting the classification theme by reading the selected papers. The followings are the classification properties that were retained:

- Investigation Type: This determines which method of empirical study was used: experiment, case study or survey.
- Scope: This specifies the scope of the paper whether the predictive model used for school or higher learning institutions and the presence of big data.
- Predictive model: This determines the method used in the study, some paper used multiple methods or models and this will make the classification task more complex.
- Purpose: This specifies the purpose to do prediction within the educational setup, whether to detect early drop-outs, improving the grade point and few others.

3.5. Data Extraction and Systematic Map

This phase determines the actual mapping process for each relevant article to categorize the classification scheme. The authors reviewed each article based on the classification strategy, the frequencies of publication, predictive model and accuracy, big data consideration and other metrics from the resulting classification. Section V presents the results about the findings reviewing the many papers considered in this study.

4. SELECTION PROCESS

The real selection process was performed during this phase. For the first filter, we identified relevant studies using defined search items. The search process was then refined further by applying the exclusion criteria to the study style. Total number before this process was 320 papers. One of the main reason some papers were rejected at this step was there is no specific prediction model mentioned in the paper but only result was described. After applying the first filter, total number of papers was reduced to 114. From here, the second filter process carried out to exclude studies on the basis of exclusion criteria applied to abstract and conclusion. Some papers have very brief abstract that it does not portray the thorough works of data mining, thus affecting the conclusion given. This has been the reason we rejected some of the papers. Total number of papers was down sized to 80 after the 2nd filter. The third step during this phase was to exclude the studies on the basis of exclusion criteria applied to abstract and conclusion. The number of papers was then reduced to 59. Selection processes became tougher as we had to select only very relevant papers to be included in the studies. For this last step, we came out with a proper tool to make the selection easier and more relevant to the studies.

The tool that we used did not involve complicated technique nor advance software. We simply used a collaborative spreadsheet in Google Docs to input relevant information so that the team members can see the progress. The field that was extracted were:

- The title of the paper.
- Year of paper published
- Publisher name
- Predictive model/framework/method use
- Attributes used for modelling
- Purpose of prediction use
- Big data presence
- Institution

By comparing the information as above, we managed to see the significant differences from each paper and finalized the 59 papers. Each paper selected comply the information that we needed. The steps taken in this phase can be seen in Fig. 2.

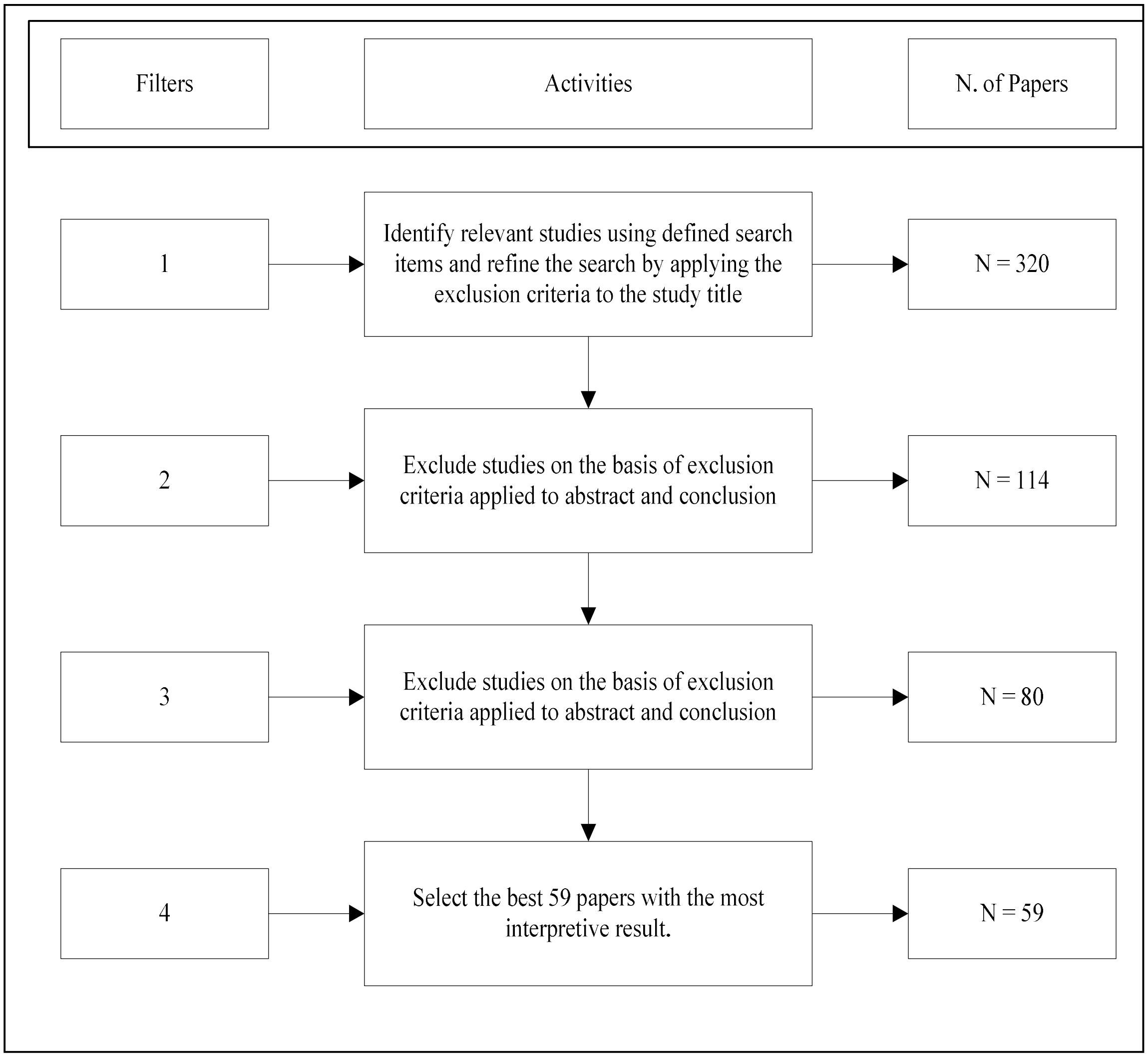
Filters
Activities
N. of Papers
1
Identify relevant studies using defined search items and refine the search by applying the exclusion criteria to the study title
N = 320
2
Exclude studies on the basis of exclusion criteria applied to abstract and conclusion
N = 114
3
Exclude studies on the basis of exclusion criteria applied to abstract and conclusion
N = 80
4
Select the best 59 papers with the most interpretive result.
N = 59

Fig.2. Stages of the selection process.

5. RESULTS AND DISCUSSION

5.1. Prediction Model (RQ1)

Fig. 3 shows the publication year of the selected paper of this mapping study. The selection of papers mostly concentrated in the last 3 years though there were about 8 papers from previous 2 years (2012-2013). This is to observe the technology evolution and new findings in the current research area. The earliest paper found matching to this study is 2012. Some of the papers within those two years were excluded due to not meeting with the criteria specified for this study.

There are 14 papers published in 2014 and were selected for this study as the predictive analysis used on student's performance is relevant. There are 22 papers found in 2015 showing an increasing interest trend. Lastly, only 15 papers in 2016 were selected for this study although there were quite a number of papers published regarding predictive analysis on student's performance.

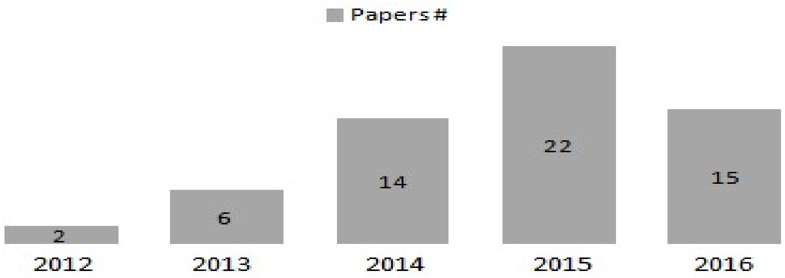

Fig.3. Publication by year

As mentioned in section 3, a thorough data extraction was done to finalize 59 papers that have been selected for this mapping study. Fig. 4 depicts the list of model or technique or method used in the study to predict students' performance. Each circle represents a method used by various papers. Bigger circles represent more papers that had used the particular techniques in their performance prediction. There could be multiple papers exist in multiple circles when a paper uses more than one method for comparison study.

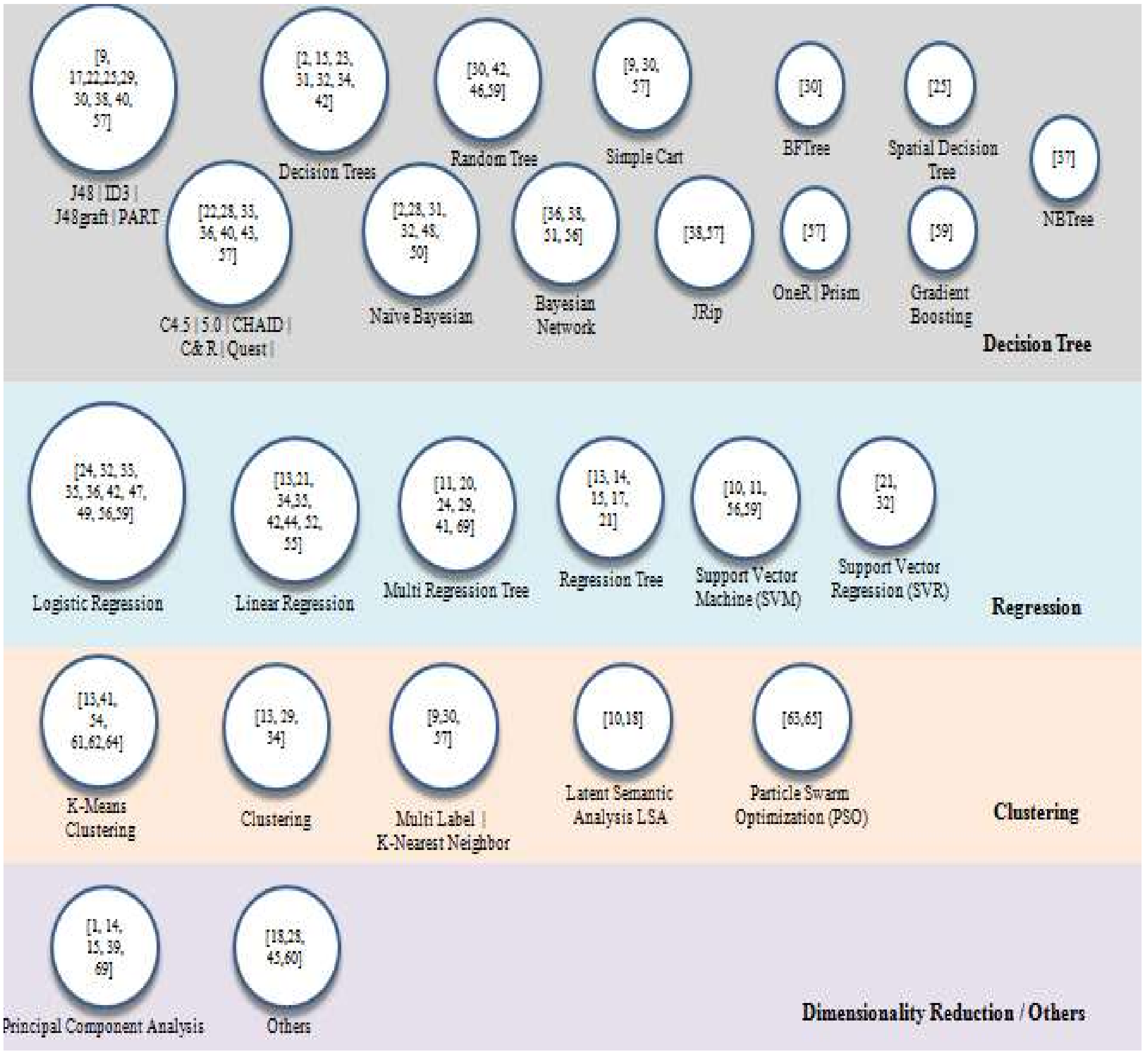

Fig.4. List of models / techniques / method used

From the 59 articles that were analyzed, the predictive models used are varied but concentrated on one particular model which is the regression model. The other models used were random forest, decision trees, support vector machine (SVM), artificial neural networks (ANN) and classification models. The regression is the most commonly used statistical method for modeling by making use of a linear combination. Twenty eight of the 59 papers have used the regression based model as their main model to predict student performance. Within the regression model, many other variations were used like logistic, support vector and multi regression. A semi-automated process of building the model is also employed called Stepwise Backward Regression.

It is also noticeable that some models used with other models for better prediction. [6] Uses regression combined with classification to predict students' grades for both subjects. In [7] uses both Multiple Linear Regression (MLR) and Support Vector Machine models to identify, which is the suitable model for a smaller sample size. According to [43], C5.0 is one of the best decision tree classification algorithms because it can handle continues and categorical values. Further, substantial papers used decision tree variations which are provided by most of the tools. Some of the key variance in decision trees are: J48, ID3, PART, J48graft, C4.5, C5.0, Naïve Bayesian, JRip, Prism, Gradient Boosting, NBTree, SimpleCart and Random Tree / Forest.

5.2. Purpose of Prediction Model (RQ2)

To answer this research question, we used our data extraction information to understand each purpose of the prediction model. Some of the prediction models shared the same purpose, therefore we managed to group the purpose for easier reference. The result can be seen in Fig. 5. From the figure, we can see that most of predictive models were used to predict students' final grade performance with 56% or 33 papers with this purpose. The second most purpose of prediction model use was to predict students' grade / performance on the MOOC or VLE platforms. This purpose carries 19% or 11 papers. About 12% of the papers dedicate their research in addressing student-at-risk of dropping-out from studies. While, the remaining 14 percent of papers for other purposes.

Another important element that we have to look in the study was the target sample. Although all papers carried a predictive model, the target sample could be either university students or school students. There are however some papers target the student's performance from an e-Learning channel such as Massive Open Online Course (MOOC).

Generally, all the researches are meant to predict student performance at the schools or colleges. However, with the increased adoption of the Internet based learning, we can see more Virtual Learning Education (VLE) systems being analyzed. Massively Open Online Course (MOOC) is a component of VLE that changes the learning and teaching landscape. MOOC provides more insight into online learning for researchers to help learn about learners' behaviour and their study patterns. Study by [16] concentrates on dropout prediction, or identifying students at risk of dropping out of a MOOC course. The issue is becoming important as more MOOC adoption is evident for tertiary learning and also due to the high attrition rate commonly found on many MOOC platforms.

In [10] in their study tried to understand students' behavior, so that an intervention program can be introduced to improve their learning activities by evaluating students' comments data and giving feedback at the end of the session. Similarly, a study by [9] analyze students' learning pattern and predict their final performance for a course in a university. In [13] identifies group of learners based on their answers and then guide them for future learning activities. In this study, different level of learner based on their competency received different set of learning activities. In [17] looked into predicting students' performance in final test by analyzing on-going evaluation or exams. Similarly, in [8, 10, 12] also looked into predicting students' performance using various models.

For early strategic guidance, in [2] used random forest model to predict those students who are the most at-risk of failing the introductory mathematics and physics courses with acceptable accuracy. The model provides an integrated evaluation of the current programs and offer strategic guidance to incoming students by better placing them in the appropriate academic sessions within STEM realm. For a limited size of sample, in [7] provides a model based on MLR and SVM which initially conducted by [16] which predicts students' academic performance in engineering subjects based on data collected in advance from students.

Finally, in [14] uses predictive model more holistically by looking into not only in-class students' parameters but their other predictors such as the place they live in, socio-economic status, environment and personality. They argue it is possible to conditionally predict student performance based on self-efficacy, socio-economic background, learning difficulties and related academic test results.

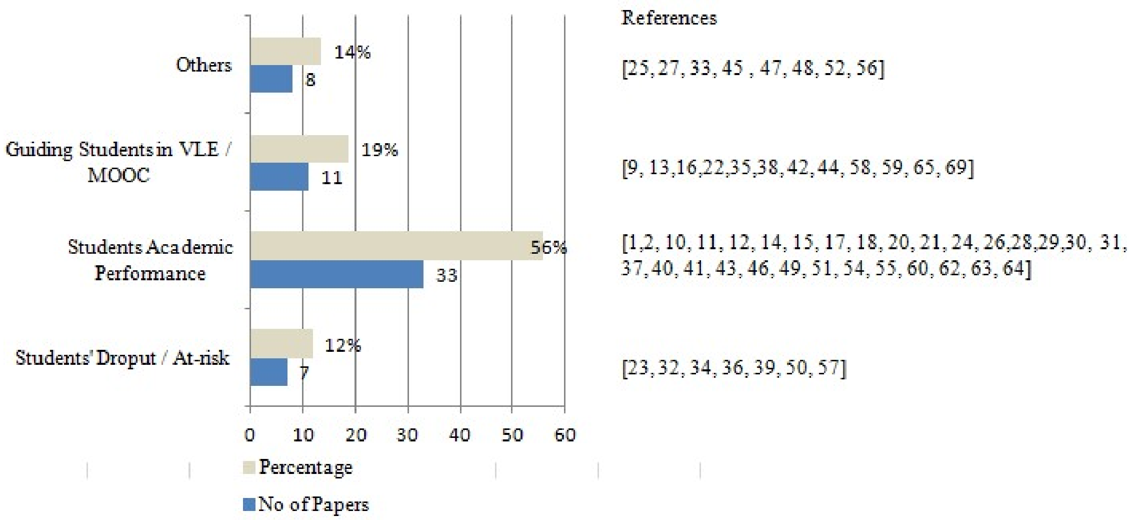

Fig.5. Purpose of the predictive model study

5.3. Big Data Consideration in Student Performance Prediction Model (RQ3)

This is the most important research question for this study. Big data is becoming integral part of the education with more data being collected from the educational ecosystem. Hence, it is important to include or to consider big data roles and impact in predicting students' performance. From this mapping study, only a handful considered big data when developing their predictive models [43]. There is a huge research gap due to inconsideration of big data in students' performance prediction. In the coming years, educational institutions will gather immense data on students, educators and all the supporting educational eco systems data. The prediction models in future must consider the big data roles which bring large data sets. These large data sets are complex, multi-faceted and nonlinear. This introduces processing performance issues to multi-co linearity challenges. Only 19% of the papers considered big data in their research as shown in Fig. 6. All other papers' recommendations will perform well on data sets that do not fit into the big data category. However, the same models used in these papers will face performance and accuracy challenges as very much larger set of data, the frequency of data arrival rate and processing complexity of continuous data is introduced in big data environments [63]. Fig. 6 shows the vast research difference when big data is considered in the students' performance predictive modeling.

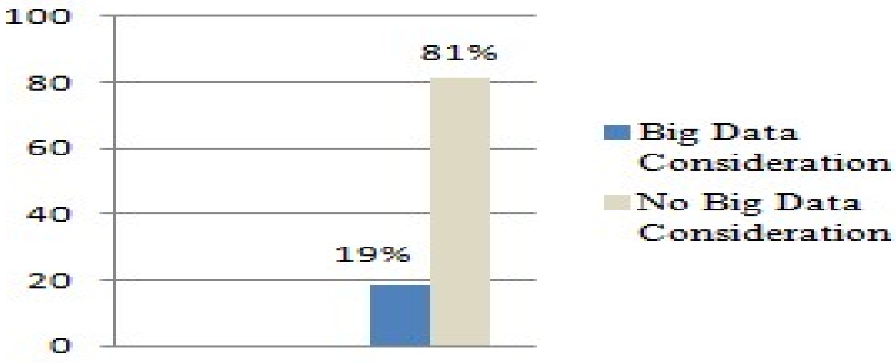

Fig.6. Big data consideration in research

Limited number of papers considered the big data challenges in educational setting [19-20, 27, 41, 43, 55, 58-60, 63, 65]. A bold statement by [60] shows most of the papers researched need to revisit their findings if big data was introduced in their environment, in which eventually it will be the case. Most of the methods are based on shallow architecture which only implements one or two layers of feature representation. These types of models cannot capture all the relationship among attributes in relatively large and correlated data set. Therefore, in [60], a prediction system called Students Performance Prediction Network (SPPN) was introduced. SPPN is capable of training millions of parameters which require massive computation power.

Additionally, in [60] used graphical processing unit (GPU) for faster execution and training. In [63], it address big data challenges in students' performance predictive model through Parallel Swarm Optimization (PPSO) based clustering mechanism. PPSO will reduce the processing time of clustering of students based on their ability, quality and efficiency. In [41], parallel K-Means algorithm based on MapReduce is introduced for classification and prediction of students' performance. Table 2 provides the complete list of papers that considered big data in their research.

Table 2. Big data incorporation in predictive model

References	Title	Contribution Type
[19]	Towards Conceptual Predictive Modelling for Big Data Framework	Model
[20]	Predicting Student Performance Using Personalized Analytics	Method
[27]	Typical Applications of Big Data Education	Open Item
[41]	A Big Data Approach for Classification and Prediction of Student Result Using MapReduce	Method
[43]	Classification Model to Predict the Learners' Academic Performance using Big Data	Model
[55]	Progression Analysis of Students in Higher Education Institution using Big Data Open Source Predictive Modeling Tool	Method
[58]	An Overview of Studies About Students' Performance Analysis and Learning Analytics in MOOC	Open Items
[59]	Big Data Application in Education: Dropout Prediction in Edx MOOCs	Model
[60]	Predicting Students Performance in Educational Data Mining	Method
[63]	Performance Analysis of Parallel Particle Swarm Optimization Based Clustering of Students	Model
[64]	Performance Analysis of Student Learning Metric using K-Means Clustering Approach	Model
[65]	Continuous Clustering in Big Data Learning Analytics	Method

5.4. Key Attributes and Feature Selection (RQ4)

The key attributes chosen by the papers mainly dependent on the objective of the papers.

Most of the papers surveyed used predictive models for students' performance prediction. The attributes revolve around students' personal information, grades and some basic demographics information. From the analysis of the papers, there are 3 types of attributes categories:

i) Basic Information-students profile, current and past grades, basic family details

ii) Extended Information-basic + parents' education, income level, siblings and etc.

iii) Holistic Information-basic + extended + locality, school / university status, social recognition and rest of attributes.

Limited papers fall under extended [19-20, 27, 41, 43, 55, 58-60, 63-65] and holistic [14, 22-24, 39] categories. In [25] included distance of schools from a district office to predict school's performance and accreditation within the vicinity of the district office. It is also noticeable that number of attributes in a model will affect the accuracy. In [26] has demonstrated that by removing few attributes the accuracy has increased. Well devised method using feature selection will help to reduce high dimensionality and improve accuracy.

Key attributes selection or popularly known as feature selection is a preprocessing step in data mining and predictive modeling. When predicting the students' performance or any other related issues, the chosen features should provide nuance into the problem that one solves. Feature selection eliminates irrelevant and redundant information. This will improve the quality of learning and accuracy of a model [50]. It is also an effective way of reducing dimensionality and increasing learning accuracy [68]. It is a technique to produce a subset of the most useful features without losing much originality of a data. Effectiveness of a feature selection refers to the quality of a subset returned, while efficiency explains the time required for the selection process. A very good feature selection method will provide a subset of data to a manageable level without losing its originality that bring sense and meanings for research goals (effective), increased predictive accuracy, lowering computational complexity and its storage, building generalizable models and finally producing it at an acceptable timing. This was concurred by [23, 31-32].

In a big data environment, the feature selection will play an important role to address the various multi-dimensional challenges. For the research conducted for this paper, only a handful used the feature selection method [23, 26, 31-32, 36, 43, 50, 52, 57, 60-61, 64]. In [43] conducted a survey on feature selection techniques and discovered that when feature selection is used in high dimensional dataset, the prediction will produce best accuracy at a faster rate. In [50] provides a comprehensive approach on feature selection by producing a framework. Therefore, feature selection is strongly recommended in developing an efficient prediction model particularly in big data environment. Table 3 lists the papers that uses feature selection in attributes selection process.

Table 3. Feature Selection (FS) incorporated papers

References	Title	Type of FS/Contribution
[19]	Towards Conceptual Predictive Modelling for Big Data Framework	Method
[23]	Predictive Analytics Using Data Mining Techniques	Weka-cfsSubsetEval, InfoGainAttributeEval, GainRatioAttributeEval, FilteredSubsetEval, Principal Components
[26]	Attributes Selection for Predicting Students' Academic Performance using Education Data Mining and Artificial Neural Network	Multilayered Perceptron
[31]	A Comparative Study of Feature Selection Techniques for Classify Student Performance	Genetic Algorithm, SVM, Information Gain, Min and Max Redundancy
[32]	Models for Early Prediction of at-risk Students in a Course Using Standards-based Grading	Correlation FS
[36]	Using Data Mining Techniques to Predict Students at Risk of Poor Performance	Chi-square algorithm
[43]	Classification Model to Predict the Learners' Academic Performance using	Chi-squared, Gain ratio, Information gain ratio, Chi-squared

	Big Data	ration
[50]	Feature Extraction Model to Identify At-Risk Level of Students	Feature selection framework
[52]	A Theory-Driven Approach to predict Frustration in an ITS	Goal-blocking-based theory
[57]	Students Dropout Factor Prediction Using EDM Techniques	Weka-cfsSubsetEval, GainRatioAttributeEval, FilteredSubsetEval, ChiSquared, Consistency-SubsetEval, Filtered, OneRAttributeEval, ReliefAttributeEval
[60]	Predicting Students Performance in Educational Data Mining	Layered approach
[61]	Student Performance Analysis Using Clustering Algorithm	Manual method
[64]	Performance Analysis of Student Learning Metric using K-Means Clustering Approach	Manual method

6. LIMITATIONS

As with any systematic mapping study, many considerations can limit the validity of the drawn conclusions. In this section, we discuss the most important ones.

6.1. Selection Bias

The major limitation of our study is the absence of an exhaustive search. We select databases that allow us to export the result in a format that is of convenience to us and easier to process in our machines. This eliminates other potential source of databases that could provide better research in prediction of students' performance. Predictive analysis model has gained increasing interest since 2011, therefore an exhaustive search using both techniques; manual and electronic databases should have been done rigorously. Initial searches for primary studies can be undertaken initially using electronic databases, but this is not sufficient. Other sources of evidence such as manual searching must also be searched. The emerging trend using big data analytic has seen significant improvement on the techniques or model use by researcher to measure students' performance.

6.2. Screening

The screening activities were performed by the research team for each article and consulted to avoid any divergent opinion. Most of the papers are screened thoroughly using the abstracts and conclusions to ensure the key valiant points for the current research are addressed. This may have resulted in rejection of poor abstract and / or conclusions, but with valid content. However, the rejection rates under this condition is very minimal.

6.3. Terminology

It is quite noticeable that throughout this paper we refer the predictive model as technique or method or framework. For some of the papers we select, there is no proposed terminology to refer the predictive analysis model used. Some researchers used the term framework to portrait the complexity of the model to predict student's performance. We did not do proper comparison to dig deeper for each model used to differentiate which one is framework or method or technique. The definition of the term itself has to be clear to avoid confusion referring to the meaning. We understand that we need to study the terminology in our future works and perhaps take into consideration to be put in our screening step later. Our main goal is to review predictive model available to predict students' performance. However, the list of method used that we have put up in Fig. 4 will give an idea which terminology is suitable for the model used.

6.4. Classification

Another element that we did not do for this study is to classify each paper to whether it is a case study, a survey or a review. We understand the goal of this study classification is different, but somehow we managed to extract the data from each paper to understand our topic better. Nevertheless, there are some borderline cases where an article could be classified in more than one category.

6.5. Impact Factor

To add to the limitation to our study, we did not look for an indexed journal listed in Web of

Science or Scopus. Although IEEEXplore and Elsevier are among reliable indexed journal, further steps need to take into consideration to do thorough research within indexed journal only as we did not want to fall into unreliable journal especially the one listed in Beall's List of Publisher. Looking for the journal impact factor is also considered important. This impact factor plays an important role as the citation rate is high, making the journal as an important source among researchers.

7. CONCLUSION AND FUTURE WORKS

In this paper, we report on a systematic mapping study of predictive model used in students' performance research. Our study uses two online databases namely IEEE Xplore and Science Direct. This study which covers the period of year 2012 and 2016 was conducted following the systematic mapping process. We at first queried databases and collected publications. Then, we screened them using their abstract and conclusion to ensure they are within our research area and analysis.

The ability to predict student's academic performance will entail a number of potential implications. Such a predictive analytic can be integrated into an online assessment system, so that educators can prioritize teaching for students whose performances are predicted to be low. The accuracy rates among the models employed varies and there is no clear guide on which model will be suitable for better prediction due to some key variables or attributes in a model. Based on the 59 articles identified on predictive model for student performance, most of the researches have their own model and predominantly used clustering and regression based models. However, all of the articles have their own methodology and parameter that decides the predictive accuracy level. It is highly recommended to conduct research on portable predictive modelling so that there is some kind of standard used when predicting student performance.

Further, there is a huge potential research area on the student performance predictive modelling using big data as most of the models do not consider big data characteristics that employs huge amount of data, continuous streaming of data and other characteristics of 3Vs. Most of the papers produce models that ride on shallow architecture with readily available predictive algorithms without considerations on hardware implications. The future is about gigantic size of students' data that need total different approach when devising a predictive model. A comparison study of different predictive algorithm performance with big data consideration is highly recommended.

From the limitations explained in section VI, we further note that some improvements have to be in place in our future works. At the early stage before conducting the review, we need to define the general concepts based on Population, Intervention, Comparison, Outcomes and Context (PICOC). The databases selection could be expanded to various databases indexed by Web of Science and Scopus while at the same time consider their high impact factor. The prediction model to predict students' performance analysis should be a study in depth to understand how it really works according to the various factors such as sample size, model use, number of attributes and most importantly in a big data environment.

8. ACKNOWLEDGEMENTS

We would like to thank the Malaysian Ministry of Education for providing the grant for this research under the Fundamental Research Grant Scheme (FRGS), Grant No FRGS/1/2015/ICT04/IPGM/03/1.

9. REFERENCES

[1] Gamulin J, Gamulin O, Kermek D. Data mining in hybrid learning: Possibility to predict the final exam result. In 36th IEEE International Convention on Information and Communication Technology Electronics and Microelectronics, 2013, pp. 591-596

[2] Guarín C E, Guzmán E L, González F A. A model to predict low academic performance at a specific enrollment using data mining. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, 2015, 10(3):119-125

[3] Shahiri A M, Husain W. A review on predicting student's performance using data mining techniques. Procedia Computer Science, 2015, 72:414-422

[4] Kitchenham B. Procedures for performing systematic reviews. Technical report TR/SE-0401, Staffordshire: Keele University, 2004

[5] Clapton J, Rutter D, Sharif N. SCIE Systematic mapping guidance. London: Social Care Institute for Excellence, 2009

[6] Brereton P, Kitchenham B A, Budgen D, Turner M, Khalil M. Lessons from applying the systematic literature review process within the software engineering domain. Journal of Systems and Software, 2007, 80(4):571-583

[7] Neto P A, do Carmo Machado I, McGregor J D, De Almeida E S, de Lemos Meira S R. A systematic mapping study of software product lines testing. Information and Software Technology, 2011, 53(5):407-423

[8] Petersen K, Feldt R, Mujtaba S, Mattsson M. Systematic mapping studies in software engineering. In 12th International Conference on Evaluation and Assessment in Software Engineering, 2008, pp. 68-77

[9] Grivokostopoulou F, Perikos I, Hatzilygeroudis I. Utilizing semantic web technologies and data mining techniques to analyze students learning and predict final performance. In IEEE International Conference on Teaching, Assessment and Learning, 2014, pp. 488-494

[10] Sorour S E, Mine T, Godaz K, Hirokawax S. Comments data mining for evaluating student's performance. In 3rd IEEE International Conference on Advanced Applied Informatics, 2014, pp. 25-30

[11] Mativo JM, Huang S. Prediction of students' academic performance: Adapt a methodology of predictive modeling for a small sample size. In IEEE Frontiers in Education Conference, 2014, pp. 1-3

[12] Taruna S, Pandey M. An empirical analysis of classification techniques for predicting academic performance. In IEEE International Advance Computing Conference, 2014, pp. 523-528

[13] De Morais A M, Araujo J M, Costa E B. Monitoring student performance using data clustering and predictive modelling. In IEEE Frontiers in Education Conference, 2014, pp. 1-8

[14] Poh N, Smythe I. To what extend can we predict students' performance? A case study in colleges in South Africa. In IEEE Symposium on Computational Intelligence and Data Mining, 2014, pp. 416-421

[15] González-Nucamendi A, Noguez J, Neri L, Robleda-Rella V. Predictive models to enhance learning based on student profiles derived from cognitive and social constructs. In IEEE International Conference on Interactive Collaborative and Blended Learning, 2015, pp. 5-12

[16] Fei M, Yeung D Y. Temporal models for predicting student dropout in massive open online courses. In IEEE International Conference on Data Mining Workshop, 2015, pp. 256-263

[17] Kaur K, Kaur K. Analyzing the effect of difficulty level of a course on students performance prediction using data mining. In 1st IEEE International Conference on Next Generation Computing Technologies, 2015, pp. 756-761

[18] Sorour S E, Luo J, Goda K, Mine T. Correlation of grade prediction performance with characteristics of lesson subject. In IEEE 15th International Conference on Advanced Learning Technologies, 2015, pp. 247-249

[19] Kim J S, Kim E S, Kim J H. Towards conceptual predictive modeling for big data framework. International Journal of Software Engineering and Its Applications, 2016, 10(1):35-42

[20] Elbadrawy A, Polyzou A, Ren Z, Sweeney M, Karypis G, Rangwala H. Predicting student performance using personalized analytics. Computer, 2016, 49(4):61-69

[21] Siddiqui M A, Gemalel-Din S. Evaluation of academic plans of study using data mining techniques. In IEEE 13th International Conference on Advanced Learning Technologies, 2013, pp. 224-228

[22] Rubiano S M, Garcia J A. Formulation of a predictive model for academic performance based on students' academic and demographic data. In IEEE Frontiers in Education Conference, 2015, pp. 1-7

[23] Gulati H. Predictive analytics using data mining technique. In 2nd IEEE International Conference on Computing for Sustainable Global Development, 2015, pp. 713-716

[24] Thiele T, Singleton A, Pope D, Stanistreet D. Predicting students' academic performance based on school and socio-demographic characteristics. Studies in Higher Education, 2016, 41(8):1424-1446

[25] Giri E P, Arymurthy A M. Model prediction for accreditation of public junior high school in Bogor using spatial decision tree. In IEEE International Conference on Advanced Computer Science and Information Systems, 2014, pp. 333-338

[26] Borkar S, Rajeswari K. Attributes selection for predicting students' academic performance using education data mining and artificial neural network. International Journal of Computer Applications, 2014, 86(10):25-29

[27] Yu X, Wu S. Typical applications of big data in education. In IEEE International Conference of Educational Innovation through Technology, 2015, pp. 103-106

[28] Mayilvaganan M, Kalpanadevi D. Comparison of classification techniques for predicting the performance of students academic environment. In IEEE International Conference on Communication and Network Technologies, 2014, pp. 113-118

[29] Jacob J, Jha K, Kotak P, Puthran S. Educational data mining techniques and their applications. In IEEE International Conference on Green Computing and Internet of Things, 2015, pp. 1344-1348

[30] Sa C L, Hossain E D, bin Hossin M. Student performance analysis system (SPAS). In 5th IEEE International Conference on Information and Communication Technology for the Muslim World, 2014, pp. 1-6

[31] Punlumjeak W, Rachburee N. A comparative study of feature selection techniques for classify student performance. In 7th IEEE International Conference on Information Technology and Electrical Engineering, 2015, pp. 425-429

[32] Marbouti F, Diefes-Dux H A, Madhavan K. Models for early prediction of at-risk students in a course using standards-based grading. Computers and Education, 2016, 103:1-5

[33] Jamil N I, Ahmad S N. Mining operational databases to predict potential donors among University Alumni. In IEEE Business Engineering and Industrial Applications Colloquium, 2013, pp. 922-925

[34] Hung J L, Wang M C, Wang S, Abdelrasoul M, Li Y, He W. Identifying at-risk students for early interventions-A time-series clustering approach. IEEE Transactions on Emerging Topics in Computing, 2017, 5(1):45-55

[35] Conijn R, Snijders C, Kleingeld A, Matzat U. Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. IEEE Transactions on Learning Technologies, 2017, 10(1):17-29

[36] Alharbi Z, Cornford J, Dolder L, De La Iglesia B. Using data mining techniques to predict students at risk of poor performance. In IEEE Science and Information Organization Computing Conference, 2016, pp. 523-531

[37] Christian T M, Ayub M. Exploration of classification using NBTree for predicting students' performance. In IEEE International Conference on Data and Software Engineering, 2014, pp. 1-6

[38] Sanchez-Santillan M, Paule-Ruiz M, Cerezo R, Nuñez J. Predicting students' performance: Incremental interaction classifiers. In ACM Conference on Learning@ Scale, 2016, pp. 217-220

[39] Sarker F, Tiropanis T, Davis H C. Linked data, data mining and external open data for better prediction of at-risk students. In IEEE International Conference on Control, Decision and Information Technologies, 2014, pp. 652-657

[40] Adhatrao K, Gaykar A, Dhawan A, Jha R, Honrao V. Predicting students' performance using ID3 and C4. 5 classification algorithms. International Journal of Data Mining and Knowledge Management Process, 2013, 3(5):39-52

[41] Mohan M M, Augustin S K, Roshni V K. A BigData approach for classification and prediction of student result using MapReduce. In IEEE Recent Advances in Intelligent Computational Systems, 2015, pp. 145-150

[42] Klüsener M, Fortenbacher A. Predicting students' success based on forum activities in MOOCs. In IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 2015, pp. 925-928

[43] Rajeswari S, Lawrance R. Classification model to predict the learners' academic performance using big data. In IEEE International Conference on Computing Technologies and Intelligent Data Engineering, 2016, pp. 1-6

[44] Ashenafi M M, Riccardi G, Ronchetti M. Predicting students' final exam scores from their course activities. In IEEE Frontiers in Education Conference, 2015, pp. 1-9

[45] Ratnaparkhi B, Katore L, Umale J S. Improved student psychology prediction and recommendation strategy using 2 state data analysis. In IEEE Global Conference on Communication Technologies, 2015, pp. 869-873

[46] Parmar K, Vaghela D, Sharma P. Performance prediction of students using distributed data mining. In IEEE International Conference on Innovations in Information, Embedded and Communication Systems, 2015, pp. 1-5

[47] Sharma A S, Prince S, Kapoor S, Kumar K. PPS-Placement prediction system using logistic regression. In IEEE International Conference on MOOC, Innovation and Technology in Education, 2014, pp. 337-341

[48] Saeed F, Dixit A. A decision support system approach for accreditation and quality assurance council at higher education institutions in Yemen. In IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education, 2015, pp. 163-168

[49] Park J Y, Luo H, Kim W H. Factors affecting students' completion: A study of an online master's program. In IEEE International Conference of Educational Innovation through Technology, 2015, pp. 275-278

[50] Singh M, Singh J, Rawal A. Feature extraction model to identify at-risk level of students in academia. In IEEE International Conference on Information Technology, 2014, pp. 221-227

[51] Itoh Y, Itoh H, Funahashi K. Forecasting future students' academic level and analyzing students' feature using schooling logs. In IEEE 4th Global Conference on Consumer Electronics, 2015, pp. 288-291

[52] Rajendran R, Iyer S, Murthy S, Wilson C, Sheard J. A theory-driven approach to predict frustration in an ITS. IEEE Transactions on Learning Technologies, 2013, 6(4):378-388

[53] Dietz-Uhler B, Hurn J E. Using learning analytics to predict (and improve) student success: A faculty perspective. Journal of Interactive Online Learning, 2013, 12(1):17-26

[54] Alfiani A P, Wulandari F A. Mapping student's performance based on data mining approach (a case study). Agriculture and Agricultural Science Procedia, 2015, 3:173-177

[55] Jose M, Kurian PS, Biju V. Progression analysis of students in a higher education institution using big data open source predictive modeling tool. In 3rd IEEE MEC International Conference on Big Data and Smart City, 2016, pp. 1-5

[56] Chen Y, Pan C C, Yang G K, Bai J. Intelligent decision system for accessing academic performance of candidates for early admission to university. In 10th IEEE International Conference on Natural Computation, 2014, pp. 687-692

[57] Pradeep A, Das S, Kizhekkethottam J J. Students dropout factor prediction using EDM techniques. In IEEE International Conference on Soft-Computing and Networks Security, 2015, pp. 1-7

[58] Duru I, Dogan G, Diri B. An overview of studies about students' performance analysis and learning analytics in MOOCs. In IEEE International Conference on Big Data, 2016, pp. 1719-1723

[59] Liang J, Yang J, Wu Y, Li C, Zheng L. Big data application in education: Dropout prediction in Edx MOOCs. In IEEE 2nd International Conference on Multimedia Big Data, 2016, pp. 440-443

[60] Guo B, Zhang R, Xu G, Shi C, Yang L. Predicting students performance in educational data mining. In IEEE International Symposium on Educational Technology, 2015, pp. 125-128

[61] Singh I, Sabitha A S, Bansal A. Student performance analysis using clustering algorithm. In 6th IEEE International Conference Cloud System and Big Data Engineering, 2016, pp. 294-299

[62] Man L, Ruisheng S. Mining the relation between dome arrangement and student performance. In IEEE International Conference on Big Data, 2015, pp. 2344-2347

[63] Govindarajan K, Boulanger D, Seanosky J, Bell J, Pinnell C, Kumar V S, Somasundaram T S. Performance analysis of parallel particle swarm optimization based clustering of students. In IEEE 15th International Conference on Advanced Learning Technologies, 2015, pp. 446-450

[64] Shankar S, Sarkar B D, Sabitha S, Mehrotra D. Performance analysis of student learning metric using K-mean clustering approach K-mean cluster. In 6th IEEE International Conference Cloud System and Big Data Engineering, 2016, pp. 341-345).

[65] Govindarajan K, Somasundaram T S, Kumar V S. Continuous clustering in big data learning analytics. In IEEE 5th International Conference on Technology for Education, 2013, pp. 61-64

[66] Baig A R, Jabeen H. Big data analytics for behavior monitoring of students. Procedia Computer Science, 2016, 82:43-48

[67] Schalk P D, Wick D P, Turner P R, Ramsdell M W. Predictive assessment of student performance for early strategic guidance. In IEEE Frontiers in Education Conference, 2011, pp. 1-5

[68] Chidambaram M, Umasundari R. A survey on feature selection in data mining. International Journal of Innovative Research in Computer Science and Technology, 2016, 4(1):13-14

[69] Pardo A, Han F, Ellis R A. Combining university student self-regulated learning indicators and engagement with online learning events to predict academic performance. IEEE Transactions on Learning Technologies, 2017, 10(1):82-92

[70] Van Barneveld A, Arnold K E, Campbell J P. Analytics in higher education: Establishing a common language. EDUCAUSE Learning Initiative, 2012, https://library.educause.edu/~/media/files/library/2012/1/eli3026-pdf.pdf

[71] Thakar P. Performance analysis and prediction in educational data mining: A research travelogue. International Journal of Computer Applications, 2015, 110(15):60-68

[72] Peña-Ayala A. Educational data mining: A survey and a data mining-based analysis of recent works. Expert Systems with Applications, 2014, 41(4):1432-1462

How to cite this article:
Muthukrishnan S M, Govindasamy M K, Mustapha M N. Systematic mapping review on student's performance analysis using big data predictive model. J. Fundam. Appl. Sci., 2017, *9(4S), 730-758.*