Volume 4, Issue 9, September 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Prediction of Primary Pupil Enrollment in Government School Using Data Mining Forecasting Technique

Manisha Sahane*[1], Sanjay Sirsat[2], Razaullah Khan[3], Balaji Aglave[4]

[1, 2]Department of Management Science, Dr. B.A.M. University, Aurangabad (M.S.), India

[3]Department of Commerce & Management Science, Maulana Azad College, Aurangabad (M.S.), India

[4]Florida Ag Research, 3001 N. Kingsway Road, Thonotossassa, FL 33592 USA

Abstract— This research concentrates upon predictive analysis of enrolled pupil using forecasting based on data mining technique. The Microsoft SQL Server Data Mining Add-ins Excel 2007 was employed as a software mining tool for predicting enrollment of pupils. The time series algorithm was used for experiment analysis. U-DISE (Unified District Information System for Education) Dataset of Aurangabad district in Maharashtra (India) was obtained from SSA (Sarva Shiksha Abhiyan) and was used for analysis. Data mining for primary pupil enrollment research provides a feasible way to analyze the trend and to addresses the interest of pupil. The dataset was studied and analyzed to forecast registered pupils in government for upcoming years. We conclude that for upcoming year, pupil strength in government school will be dropped down and number of teachers will be supersizing in the said district.

Keywords— Data mining, Microsoft SQL Server Data Mining Add-ins Excel 2007, Time series algorithm, U-DISE

I. INTRODUCTION

Data Mining has its origins in various disciplines, of which the two most important are *statistics* and *Machine Learning* [15]. For examples, if we want to know how many houses person owns, basically we can start looking at data like how many firms he has, how far his firms from one another, how many children he has, what does his annual income? So forth. Accordingly we come to situation to start predicting certain things but this cannot be generally possible by firing some sort of queries, you need to apply some sort of sophisticated algorithms and that is where mining comes into pictures so by using data mining one can actually start predicting values by creating model which predict some values that can be incurred by using past data.

1.1 Data Mining

Data mining is process which exact hidden interesting patterns from pre-processed data by applying various tasks like classification, decision trees, regression, artificial neural network, support vector machine, association rules, forecasting technique, clustering. Data mining is a process of discovering various models, summaries, and derived values from a given collection of data [15]. Before going to start mining, data should go through data cleaning, data integration, data selection, data transformation processes. Pattern evaluation is done after applying mining algorithms. Data mining is a multi-billion global market that is gaining popularity. Data mining is the process of extracting hitherto unknown and potentially useful patterns, trends, anomalies and rules from stored historical data for business promotion, decision making or classification [5]. Similar meaning to data mining – for example, knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging [11].

1.2 Machine Learning

Main research area is for computer program to automatically learn to recognize complex patterns and make intelligent decision based on data [11]. Machine learning algorithms that can apply to task to solve real world problem. Machine learning finds how computer can learn based on data. Figure 1 show four types of learning that are highly related to data mining.

1. Supervised learning (or learning with a teacher) is basically related with classification. This model uses labeled training data and may require additional input during the training phase. As compare to unlabeled training data, labeled training data is hard and time consuming.
2. Unsupervised learning (or learning without a teacher) is synonym for clustering and uses unlabeled training data and does not require addition parameters settings.
3. Semi- supervised learning make use of both labeled and unlabeled training data..
4. Active learning is a machine approach that lets users (e.g. domain expert) plays an active role in the learning process.

Machine learning can often be successfully applied to problems, improving the efficiency of systems and the designs of machines. There are several applications for Machine Learning (ML), the most significant of which is data mining [16].

The application of DM forecasting technique in school education sector is the most challenging task. For accurate forecasting time series is an essential to running almost any business. Prediction from Data Mining offers the government school an opportunity to act before a pupil drops out or to plan for resource allocation with confidence gained from having complete records of all pupils reflecting their tracks of activities. U-DISE (Unified District Information System for Education) data is enough to create a reasonable forecasting model. Time series algorithm used general patterns discovered in the model to the U-DISE data to predict strength of pupils and number of surplus teachers for upcoming year.

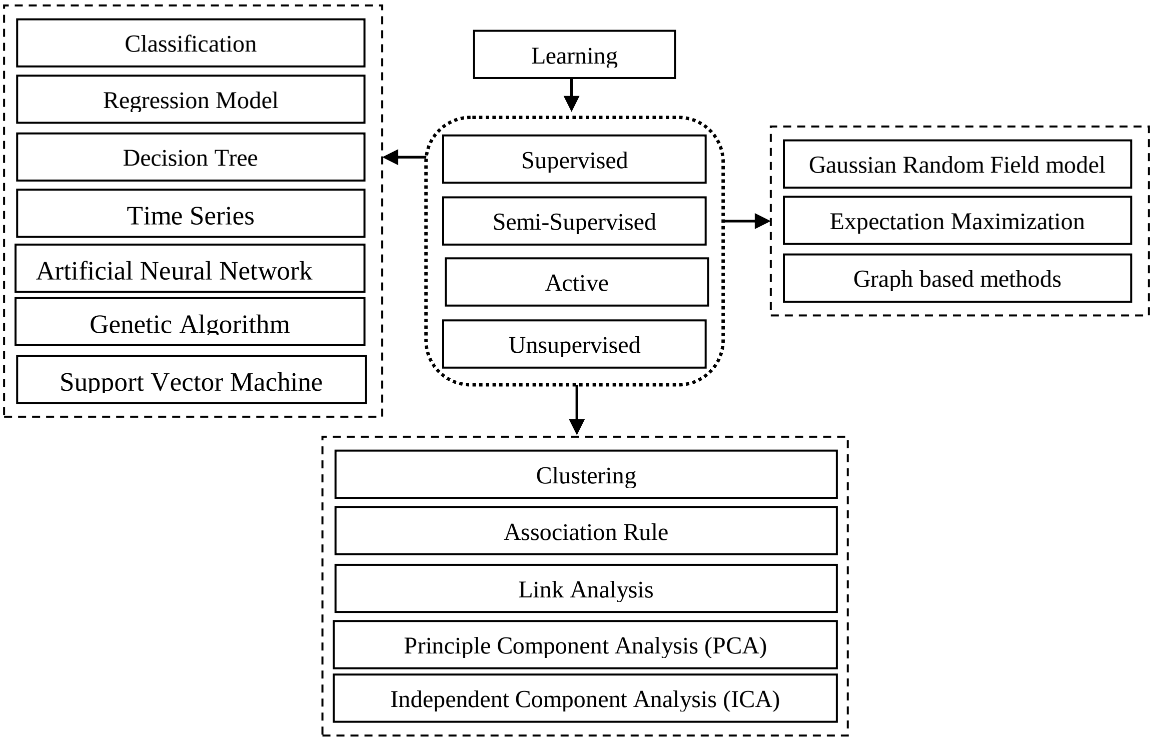

Fig. 1 Major Algorithms belongs to three learning types

II. Review of Educational Data Mining

Numerous studies were performed on the evaluation of the educational systems using educational data mining. Proposed a program evaluation framework using Educational data mining [14]. Presented a system that can be used for a performance improvisation of students in their academic studies. Division technique of Regression is used to differentiate between poor and good performers [19]. Clustering, Decision tree and neural networks are used to evaluate post-graduate student's performance and overcome the problem of low grades of post-graduate students [6]. A hybrid approach which uses EDM(educational data mining) and regression analysis to analyze live video streaming (LVS) students' online learning behaviors and their performance in their courses like students' participation, login frequency, the number of chat messages and questions that they submit to their instructors, were analyzed, along with students' final grades [1]. Data mining techniques are very much useful in the development and finding out the solutions of higher education in Madhya Pradesh state [4]. Data mining aided to identify some of the most important factors in teachers' own evaluation of the educational system [10]. Classification and cluster analysis tasks are applied to compare students' success from real data with predicted students' success and results show even if there is lack of attributes, one may still apply certain data mining algorithms over school data to gain knowledge on the mainstream flow [3]. Suitable prediction techniques using data mining tool WEKA to help in enhancing the quality of the higher educational system by evaluating student data to predict the student performance in courses during early period of study [21]. Analyzed the massive data sets generated by various processes used for predicting time series data using SPSS-Clementine and determine the feasibility and effectiveness of time series data [7]. Science enjoyment and frequent use of educational software play important roles in the academic achievement of Canadian students [24].

III. Time Series Algorithm

A time series is set of statistical observation arrange in chronological order [9]. The nature of time series data includes: large in data size, high dimensionality and necessary to update continuously. Moreover time series data, which is characterized by its numerical and continuous nature, is always considered as a whole instead of individual numerical field [8]. It consists of consistent historical collected data over successive period of time. Any sequence of numbers examined over some period of time creates a time series. Microsoft time series algorithm presents patterns in two ways regression formula can tell you how strong or weak any factor is in forecasting, decision tree presents more descriptive rules about data. For example, in agriculture, you are running a retail store of agricultural products and you are managing inventory of each of the products you stock. Retailer knows that in the middle of May, farmers buy more agricultural products, so he put an extra stock. Retailer has an idea that in between October to February, there may be chance of crop pest so that he orders remedial Agricultural products from his supplier. He guesstimates how much of crop seeds of particular brand to order as well according to people's interest. In each of these cases, he is doing time series analysis. Retailer using past sales history of agricultural products to forecast future product requirements, and volume of products. Time series data are defined as a sequence of pairs,

$T= ([p_1, t_1], [p_2, t_2],\ldots,[p_n, t_n])$, where $t_1<t_2<\ldots<t_n$.

Each p_i is a data point in a d-dimensional data space, and each t_i is time stamp at which p_i occurs [15]. Four models namely Multiple Regression in Excel, Multiple Linear Regression of Dedicated Time Series Analysis in Weka, Vector Autoregressive Model in R and Neural Network Model using NeuralWorks predict models accurately predict the exchange rates, but Multiple Linear Regression of Dedicated Time Series Analysis in Weka outperforms the other three models [20]. Framework that enables to make class predictions about industrial stock performances using decision trees, neural networks, logistic regression models [12].

Following time series components are used to predict future patterns:

1. Secular trend or long term estimation (T)
2. Periodic movement or short term fluctuations
   a. Seasonal patterns (S)
   b. Cyclical variation (C), for example, business cycle
3. Random or Irregular variations (R or I) representing outliers

Method for measuring trend component in a time series:

1. Graphic or free-hand curve fitting
2. Semi-Averages
3. Curve fitting by the principle of Least squares
4. Moving Averages
   a. On linear trend
   b. On curvilinear trend
   c. On polynomial trend
   d. On irregular fluctuations

Microsoft time series algorithm presents patterns in regression formula and decision tress. Time series provides mainly two algorithms. ART_{XP} (Auto Regression Tree with cross-prediction) model is characterized by scientific analysis of history and to predict the result over time. It is used for short term accuracy and ARIMA (Auto-Regression Integrated Moving Averages) is used for long term predictive stability. In general, when the model prediction is more accurate, the results predicted by the model are more trustable. In this study, the time series method is used to analyze the statistical properties of schooling (U-DISE) data. Time series prediction method based on previous year data shown the regularity and influencing factors of data, and then makes predictions. This feature has very important significance for the taking action towards fluctuating pupil enrollment trend.

IV. District Information System For Education (DISE)

It is developed by National University of Educational Planning and Administration (NUPA). For successful implementation of educational program Sarva Shiksha Abhiyan is developed separately; concerted efforts have been made towards strengthening of Educational Management Information System (EMIS) in India. At the time of initiating District Primary Education Programme (DPEP) in 1994, it was felt that a sound information system is essential for successful monitoring and implementation of the programme. U-DISE software is now operational in 580 districts in 29 states and UTs in India. Central government has lunched U-DISE for collection of statistical data. In U-DISE system, each school must have to fill the data year (Student, teacher, class room, and other facilities) till 30th September in each. This information is valid for different central government schemes like mid-day meal , Sarva Siksha Abhiyan, Rashtriya Madhyamik Shiksha Abhiyan, and ICT – Information & Communication Technology in Schools [23].

V. Selections of Trained Sample Data of Aurangabad (M.S.) District

We chose the schooling data from 2004 to 2013 government of primary school as the sample data to analyze the pupil enrollment change situation. The dataset covers ten years of schooling data in all primary section. It provides time-series data at school, village, cluster, block and district levels. Data sample has taken from [22]. Dataset is in the interleaved format, it is a little bit difficult to interpret. We chose this data series to analyze. And then, we predicted the pupil enrollment for the year 2014.

VI. Methods and Materials

Microsoft provides nine algorithms i.e. Association, clustering, decision tree, liner regression, logistic regression, naïve bayes, neural network, sequence clustering, and time series algorithm and can be used for building models. Microsoft provides two main tools

- Microsoft SQL Server Management studio: it is easier to use and it is generally used by executives.
- Business Intelligence development studio: This is actually start creating mining model by defining training data.

Microsoft SQL Server DM Add-ins Excel 2007 tool was used as the data mining tool. In this tool, most of the mining algorithm becomes more applicable and easy to use without professional statistics and data mining background. It covers the all set of SQL server data mining and it provides a much simpler client interface for performing data mining. Classification, decision tree, neural networks, regression, and other data mining algorithms are not available in standard Excel but in the Microsoft SQL Server Data Mining Add-ins Excel 2007 allow users to build complex statistical models in Excel. It also provides different methods for getting data into Excel, preparation of data, defining model, validation of model, applying model etc. [17].

In fig. 2 the x-axis of the graph shows time period in years and the y-axis of the graph shows enrollment. The Forecast tool under the Analyze tab performed forecasting over a selected schooling table. The forecasted values are added to the original table and highlighted. A chart is generated (in a separate sheet) displaying the present value and the forecasted evolution of the series. To run this task, table should have at least 10 rows.

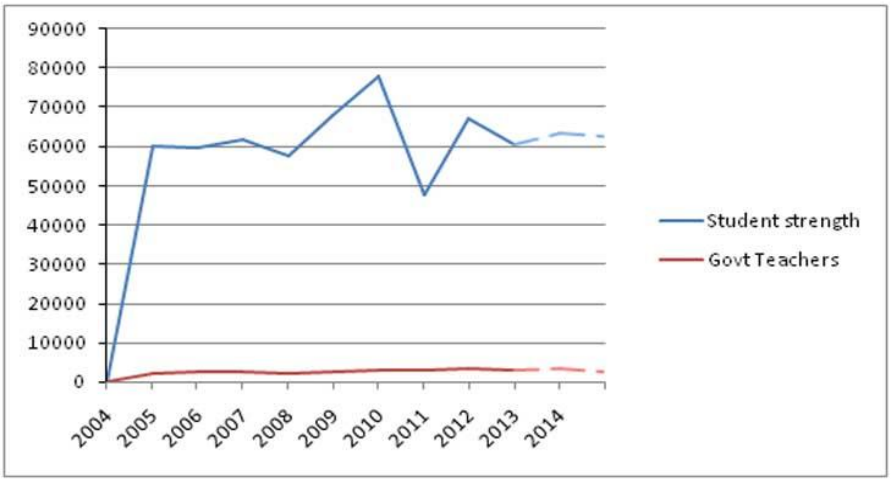

Fig. 2 The forecasted values was calculated by analyze forecast

We picked Time series Microsoft mining algorithm to perform the data mining. Time series models provide with enough information to see how data interacts by using tree and histogram so that one can make sound decision. In figure 3, the x-axis of the graph shows time period in years and the y-axis of the graph shows enrollment. Time series showing number of government teachers and pupil enrollment at government school for last 10 years with discovered patterns in the model. A forecasting model of MS SQL Server DM 2007 Add-ins detects patterns in a series of cells. It uses these patterns to forecast the evaluation of these series of cells.

VII. Decline Pupil Enrollments in Government Primary School

In the 2001 census, the total population size in Aurangabad district was 2,902,602 and in 2011 census, it was 3,701,282. From this one can conclude, the population size is increased in said district from 2001 to 2011 [13]. In fig. 4 shown the total population of male and female children of age 0-6 years old and supersizing in the population from 2001 to 2011, so there has to be constantly inclined strength of students in government schools but in figure 2 student strength trends is fluctuating over the past years. In 2013, total primary classes from 1st standard to 5th standard are 3063 in Aurangabad district. Out of that government primary classes are 2161 and private primary classes are 902 [22].

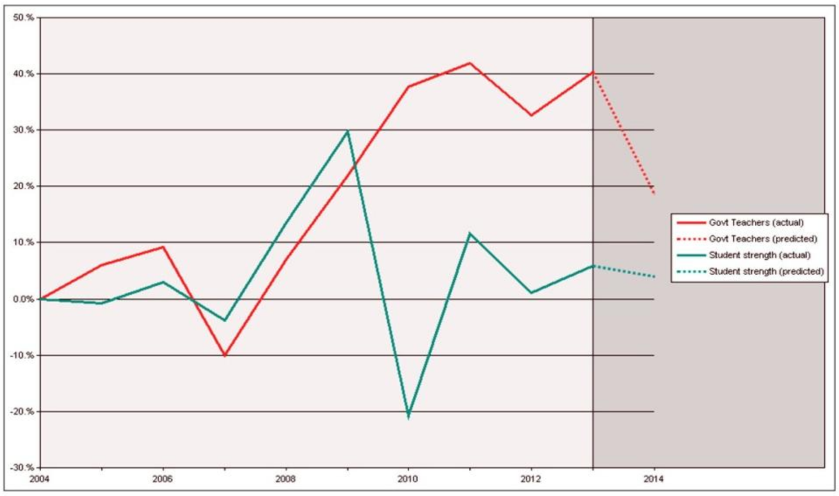

Fig. 3 The forecasted values was calculated by Microsoft Time series algorithm

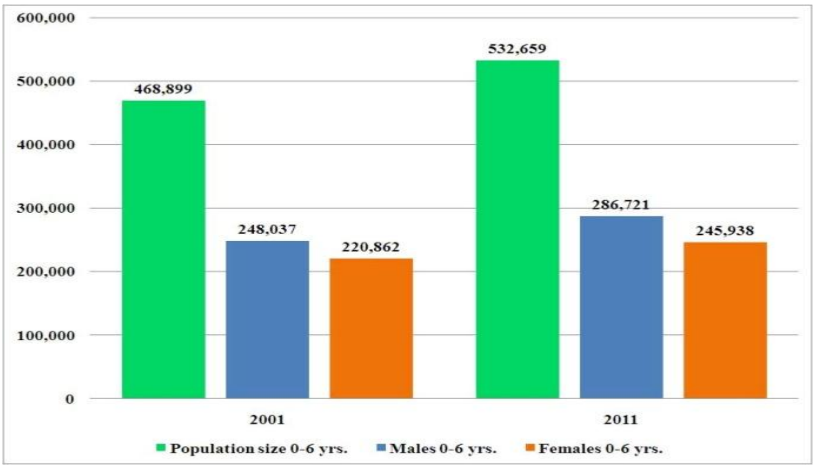

Fig. 4 Population size 0-6 years in Aurangabad district

Enrollment in the government and private primary schools, the preference was clearly for the private. General reasons behind fluctuating trend of pupil enrollment can be similar to the following analysis done on primary education in Delhi by Yash Aggarwal. He had done following detailed analysis of issues were identified as far as provision of primary education in Delhi.

1. Inadequate/absence of access to a comparable quality of education.
2. Overcrowding in the existing government schools.
3. Mismatches between demand and supply of schooling facilities.
4. Dilapidated condition of class rooms, particularly those running in rented buildings. Repairs o rented buildings cannot be undertaken under Rent Control Act.
5. Lack of sanitation and water facilities in old school buildings.
6. Excessive reliance on centrally sponsored schemes has also created its own problems. The states seldom initiate programs of educational development at their own initiatives and wait for central government initiatives.
7. The educational planning for UEE in the urban context requires special emphasis. The traditional methods of removing supply side constraints would not succeed in achieving UEE objectives [2].

According to RTE [18] act, the strength of teachers to be sanctioned on number of students. The Maharashtra state government has come out with a Government Resolution (GR) which makes it compulsory for primary schools to have a teacher for every 30 students from primary section (class I to V).

VIII. CONCLUSION

In this paper, the analyzed U-DISE data set for applying to the problem of primary schooling enrollment in government schools in Aurangabad district. It is wake up call for local authority that the enrollment is not uniformed throughout the period of the study i.e. 2004-2013. During this period the strength has shown a fluctuating trend on year on year basis. Similarly the decline in strength resulted in surplus of teachers in some years. In 2012-13 the pupil enrollment value is 63555 and predicted pupil enrolled in 2013-14 is 62410. The present adopted algorithm helps in predicting primary pupil enrollment in the current year based on previous year trend. The tool used Microsoft SQL Server Data Mining Add-ins Excel 2007 can provide useful predictions based on which corrective majors may be taken and it allows building a variety of powerful models very quickly and reduces effort required to extract information from data set. This study provided vital information for policy formulation and it may help in preparation of district elementary education plans.

REFERENCES

[1] Abdous, M., He, W., and Yen, C.-J. (2012). Using Data Mining for Predicting Relationships between Online Question Theme and Final Grade. *Educational Technology & Society, 15* (3), 77–88.

[2] Aggarwal, Y. (2013). Quality Concerns in Primary Education in India Where is the Problem? *National Institute of Educational Planning and Administration (NUPA),* 1-17.

[3] Ahmedi, L., Bytyci, E., Rexha, B., and Raca, V. (2012). Applying data mining to compare predicted and real success of secondary school students. *Advances in Applied Information Science*, 178-181.

[4] Arjariya, T., Kumar, S., Shrivastava, R., and Varshney, D. (2011). *International Journal of Soft Computing and Engineering (IJSCE),* 1(5), 238-240.

[5] Chattamvelli, R. (2009). Data Mining Methods. Narosa Publishing House Pvt. Ltd.

[6] Chuchra, R. (2012). Use of Data Mining Techniques for the Evaluation of Student Performance:A Case Study. *International Journal of Computer Science and Management Research*, 1(3), 425-433.

[7] Diwan, T., Chouksey, P., Thakur, R., and Lodhi, B. (2012). Exploiting Data Mining Techniques For Improving the Efficiency of Time Series Data. *International Journal of Computer & Communication Technology (IJCCT),* 3(3), 42-51.

[8] Fu, T. - C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence,* 24, 164–181.

[9] Gupta, S. (2009). Fundamental of statistics. 6th ed. Himalaya Publishing House Pvt. Ltd.

[10] Haisan, A.-A., and Bresfelean, V. (2013). A Data Mining Survey on Identifying the Factors that Influence Teachers' View of the Romanian Educational System. *Advances in Applied Information Science,* 7(3), 160-165.

[11] Han, J., Kamber, M., and Pei, J. (2012). Data mining concepts and techniques. 3rd ed. Boston: Morgan Kaufmann Publishers.

[12] Hargreaves, C. and Hao, Y. (2013). Prediction of Stock Performance Using Analytical Techniques. *JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE,* 5(2), 136-142.

[13] http://censusindia.gov.in/2011census/censusinfodashboard/index.html.

[14] Hung, J.-L., Hsu, Y.-C., and Rice, K. (2012). Integrating Data Mining in Program Evaluation of K-12 Online Education. *Educational Technology & Society,* 15 (3), 27–41.

[15] Kantardzic, M. (2011). Data mining concepts, models, methods, and algorithms. 2nd.ed. A John Wiley and sons, inc., publication.

[16] Kumar, R., and Verma, R. (2012). *International Journal of Innovations in Engineering and Technology (IJIET),* 1(2), 7-14.

[17] MacLennan, J., Tang, Z., and Crivat, B. (2008) Data Mining Microsoft SQL server 2008. Willey publication.

[18] Ministry of law and justice. (2009). The right of children to free and compulsory education. *The gazette of India.* Section 25 (1).

[19] Moroney, K., and Makh, S. (2012). Data mining Application to Design a System for Performance Improvisation of Students in Their Academic Studies. *IJCSET,* 2(9), 1396-1401.

[20] Saigal, S. and Mehrotra, D. (2012). PERFORMANCE COMPARISON OF TIME SERIES DATA USING PREDICTIVE DATA MINING TECHNIQUES. *Advances in Information Mining,* 4(1), 57-66.

[21] V.Ramesh, P. Thenmozhi, and K. Ramar. (2012). Study of influencing factors of academic performance of students: A data mining Approach. *International Journal of Scientific & Engineering Research,* 3(7), 1-5.

[22] www.dise.in

[23] www.maharashtra.gov.in. (14 November 2013). School Education and Sport Department. Maharashtra Government Resolution No: Index 2013/96/RTE. Government Resolution code: 201311141545022821.

[24] Yu, C., Kaprolet, C., Jannasch-Pennell, A., and DiGangi, S. (2012). A Data Mining Approach to Comparing American and Canadian Grade 10 Students' PISA Science Test Performance. *Journal of Data Science 10*, 441-464.