FORECASTING STUDENTS' ENROLLMENT IN MALAYSIA POLYTECHNICS USING PREDICTIVE ANALYSIS

Siti Zuhra binti Abu Bakar
Department of Information and Communication Technology
Politeknik Besut Terengganu
ctzuhra@gmail.com

'Afifah Nailah binti Muhamad
Department of Information and Communication Technology
Politeknik Ungku Omar
anailah@puo.edu.my

ABSTRACT

In the era of Industrial Revolution 4.0, it is crucial for higher institutions to provide quality education to their students. The impact of Covid-19 pandemic to the world has led a new norm in education sectors. Allocating students based on their choices and proposing the enrollment projection can be a challenging task as it requires many criteria to be considered. Implementation of data analytics and data science in Malaysia Polytechnic data is also minimal. Higher education institutions have created and collected huge amounts of data electronically for many years. Polytechnic department of Malaysia Higher Education Institution has a system for administration from student's enrollment, e-learning, examinations, industrial training, and graduate's employability. Through these systems, the polytechnic department has created a huge amount of data on students in the technical education environment. However, these data are not fully utilized to their full potentials and benefits. The main goal of this research is to predict students' enrollment in Malaysia Polytechnics. This research also gives a feasible solution that guides administrators to provide better education quality through analyzing the students' enrollment dataset by exploring Cross Industry Standard Process for Data Mining (CRISP-DM) using a descriptive data mining approach and proposed predictive model for the enrollment. A predictive model is generated for the next five years of student enrollment using Gaussian processes, Multi-Layer Perceptron (MLP) and SMOreg model.

Field of Research: *Predictive analysis, forecasting, data mining, data science, higher education institution, enrollment.*

1. Introduction

Higher education institution demands continually in upgrading their programs for job market and attracting prospective new students' enrollment. Malaysia government, under the administration of the Ministry of Higher Education (MOHE) objective is to make Malaysia as a centre for quality education. They also stated that there is high percentage of primary, secondary and tertiary educations are funded by the government (Ministry of Higher Education, 2009). Higher education institutions are constantly facing competitive environment; hence they should improve the quality of services by obtaining extensive and enough knowledge to improve services on assessment, evaluation, planning, and decision making (Siraj and

Abdoulha, 2011). Swamy and Hanumanthappa (2012) stressed that in order to improve the quality of education, data analysis plays an important role for decision support. Education sectors also provide large amount of data. These data also known as educational data mining has many potential and useful information to be extracted and can be used to solve many problems arises in education sectors either for government or private education providers. Aher (2011) discussed that educational data mining majorly concerned on developing and expanding methods in exploring the unique set and classification of data that come from education systems and setting in order to understand students need in their education setting. Despite the needs of upgrading the education quality admissions and education, there is lack of previous research on students' enrollments and placement into Malaysia polytechnic based on current data captured digitally using their system and warehousing capabilities. Placing and allocating students based on their choices can be challenging task as it requires many criteria to be taken care of such as past examination grades, admission criteria, locations of the programs' offered and so on. According to Terakegn and Sreenivasarao (2016), several issues can raise related to the department placement of students in different standpoint. Therefore, they suggested on educational data mining in predicting student's placement based on their order of preferences. Students who pursue admission to higher education institutions also experience a difficulty to select a program. They are facing difficulties in terms of wide variety of courses to choose from. Polytechnic administrator in Students Enrollment Department and Curriculum Development Department also facing the problems to suggest programs based on job market in currents to future demands. Michael et al. (2017) have gathered information from higher education institution members on the usage of predictive analytics mainly in pertaining the alignment of personnel allocation and financial resources, including the following criteria:

- Institutional commitment to increase undergraduate retention and improving enrollment management;
- Senior-level leadership encourages data-informed decision making;
- Strong partnership between campus functions, particularly information technology (IT) and institutional research;
- Adequate allocation of resources for staff to effectively address the findings produced from predictive models;
- Continuous training and support for personnel who collect, analyze, or utilize data;
- Capacity to connect data across systems or within one system; and
- Increased accountability metrics, such as performance-based funding.

Higher education institutions must take advantage of predictive analytics in their strategic management and planning. Bichsel (2012) have conducted a survey on educational data usage in various functional areas of corresponding institutions were using data at a level below the threshold identified in the definition of analytics which is, using data proactively or to make predictions but mostly were collecting data without using the data to make predictions. The goal of this research is to predict number of students enrollment for corresponding polytechnic in Malaysia for next five years.

1.1 Research Objectives

The aim of this research is to propose predictive model based on predictive analysis using Gaussian Processes, MLP and SMOreg for students' enrollment in higher education institutions of Malaysian Polytechnics and to compare the performance accuracy of proposed predictive model. In this research, R programming was used to provide descriptive analysis on the dataset.

WEKA is used as predictive analytics tools using machine learning algorithm which is Decision Trees (J48), Naive Bayes and Neural Network (Multi-layer perceptron with Back Propagation) for predictive analysis on offered course and Gaussian processes, MLP and SMOreg for predictive analysis on students' enrollment. There are many tools for data processing and analysis in data science, but each has come with different capacities, capabilities, criteria, complexity and contains tools for data pre-processing, classification, regression, clustering, association rules, forecasting and even data visualization. The key feature of WEKA provides many different algorithms for data mining and machine learning. WEKA is an open source that is freely available and easily usable by people who are not data mining specialists (Nur Syahela et al., 2015). Meanwhile, other tools such as Tableau is used as visualization tools and dashboard analytics.

This research is beneficial for higher education institution in providing better programs based on resources and facilities provided in polytechnic Malaysia. This also includes program allocation based on students' choices and preferences. Based on the research result of the proposed predictive model, statistical report for student's enrollment for the next five years can be generated. Hence, the outcome of this research will help polytechnic administrator for their strategic planning in resource allocation, student's successes, and finance management. Moreover, this research outcome will help Students Enrollment Department of Polytechnic to formulate proper and suitable programs and total of students' enrollment for each polytechnic.

2. LITERATURE REVIEW

2.1 Data Science and Education Data Mining

Data science has become widely used with data mining and big data. Data science refers to a study of the generalizable extraction of knowledge from data. Dhar (2013) described data science might imply a focus around data extended with statistics which is systematically study on organization, properties, and analysis of data with their role in inference. Data science process consists of data mining, machine learning and prediction. Furthermore, Provost and Fawcett (2013) explained data science as a package of basic principles used to support and conduct extraction of the data that follows these fundamental concepts:

i. Generate valuable and useful knowledge from raw data to mitigate business issues and problems.
ii. In term of data science result evaluation, it also considers many components and elements.
iii. The correlation between business and analysis can be decomposed into sub problems.
iv. Information technology tool and software assist organization to get an insight of the data.
v. The data maybe not standardized from the source of data.
vi. To make conclusion in decision making, another aspect also takes into consideration.
vii. Not only focus on the results.

Data mining is one of the processes in data science. Hence, data mining is a technique to gain useful information from extensive of data to support decision making in various aspect. This is support by Ajay and Saurabh (2013) stated that data mining is the process of analysis and extracting of data to discover useful information from it. In recent years, data mining has become widely used in many areas such as education, banking, agriculture, engineering, medical and many more. There are many related studies using data mining. Data mining can

be defined as gaining useful and insight knowledge discovery from various of datasets from different methods such as classification, clustering, prediction, association rules and many more while education data mining (EDM) is an application of data mining (Ramanathan et al., 2014).

EDM is designed and build for the analysis processes of data from educational setting to understand students and their education and learning environment situation. Figure 2.2 shows the application of data mining to the design of educational systems is an iterative cycle of hypothesis formation, testing and improvement illustrated by Saeed et al., (2014). The potential and benefits of education data mining has offered to higher education institutions making it is gaining its popularity (Alshareef et al., 2015). Higher education institution can benefit from data mining by improving quality of education, student's enrollment and admissions, employability rates and the overall practice (Tarekegn and Sreenivasarao, 2016).

2.2 Cross Industry Standard Process for Data Mining (CRISP-DM)

Cross Industry Standard Process for Data Mining, also known as CRISP-DM was introduced by Chapman et al. (2000) involved six phases which are business understanding, data understanding, data preparation, modeling, evaluation, and deployment as shown in Figure 1.0. This method is one of popular methodology for analytics, data mining or data science projects (Gregory, 2014). Hence, CRISP-DM is used in this research to extract useful information in predicting students' enrollment.

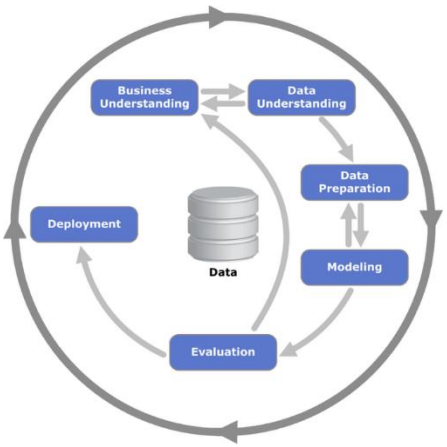

Figure 1: CRISP-DM Methodology (Chapman et al., 2000)

2.3 Machine Learning and Prediction in Education Data Mining

Machine learning is an artificial intelligence field where it gives the computer to learn without being programmed. The process of machine learning is close to data mining. The purpose of machine learning is to generate predictions models and analytics. There are many previous studies used machine learning algorithm to perform prediction. Hence, there are many categories of machine learning algorithm such as Decision Trees, Multi-Layer Perceptron and so on. Multi-layer Perceptron (MLP) with Back Propagation is a supervised learning network that has been widely implemented Neural Network topologies. Mohammad et al. (2015) described that this Neural Network is a popular learning algorithm in a sense that Neural Network knows the required output and adjusting of weight coefficients is done in such way, that the calculated and desired outputs are as close as possible. The reason why MLP is chosen for classification in this research is due to its generality, computational simplicity, reachability,

robustness, flexibility, and high computational rates without requiring so much expertise (Lim et al., 2007).

2.4 Predictive Analysis using Time Series Analysis

The world is embracing prediction problems in many areas given it is always seen as it is behaving in other times and places. For example, population prediction in ten years as prediction problems are inescapable. Michael et al. (2017) discussed on predictive analytics has become essential part in education as the value of higher education has increasing where higher education institutions are turning to business intelligence practices to improve outcomes. Hence many institutions have adopted data analytics practices to forecast operational needs and enrollment trends and are now applying the use of predictive analytics directly to student success initiatives. In this research, predictive analytics is used to utilize students' enrollment data to predict future enrollment for polytechnics across Malaysia. Figure 2 illustrates the main steps for predictive analytics in educational systems (Jindal and Dutta, 2015). Jindal and Dutta (2015) also discussed predictive analytics can assists decision-making process for higher education stakeholders.

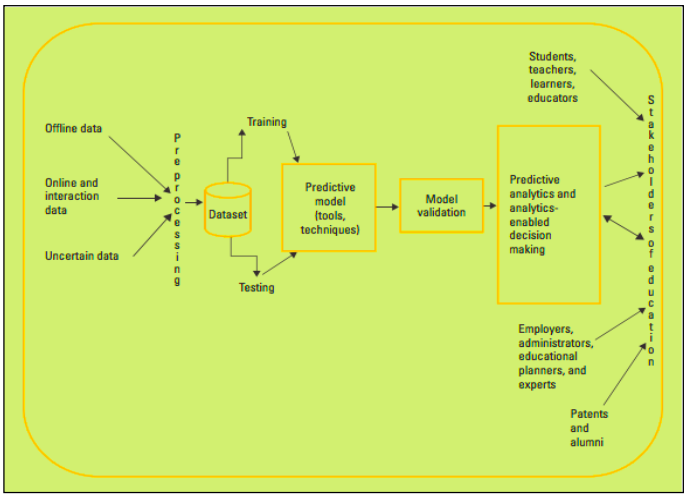

Figure 2: The main steps for predictive analytics in educational systems (Jindal and Dutta, 2015)

There is a necessity for forecasting to determine the possible demand in the future when analyzing demand data that are considered time series and represent the historical market demand in each period which can reduce the decision risk (Arnis and Arkady, 2012). Predicting students' enrollment using the time series analysis can be beneficial to polytechnic stake holders to make decision on future program allocation. Malaysia Polytechnic Department has divided into several sections: Department of Administration Services, Department of Policy, Department of Polytechnic Planning, Department of Polytechnic Research and Innovation, Department of Industrial Relation and Employability, Department of Curriculum Development, Department of Examination and Assessment, Department of Instructional and Digital Learning, Department of Student Development and Department of Students Admission. These departments have their own Key Performance Index (KPI) to achieve, hence requires better planning and analysis to achieve better performances. Department of Students Admission have flow chart for each Students Admission Officials in each polytechnic across Malaysia to be followed.

3. Research operational framework

An input dataset will be prepared in data preparation for statistical analysis. The research framework is based on data science process of pre-analytics and post analytics methods. Pre-analytics consists of business understanding, data understanding and data preparation. On the other hand, post analytics consists of modeling, evaluation, and deployment, involving process of splitting the data into training and testing to be applying with machine learning algorithm to build a predictive model. The results were compared with each other to gain insight prediction accuracy. In this research, machine learning tools which are R programming and WEKA were used for both pre-analytics and post-analytics due to its suitability for data analysis, modeling, prediction, and graphical presentation.

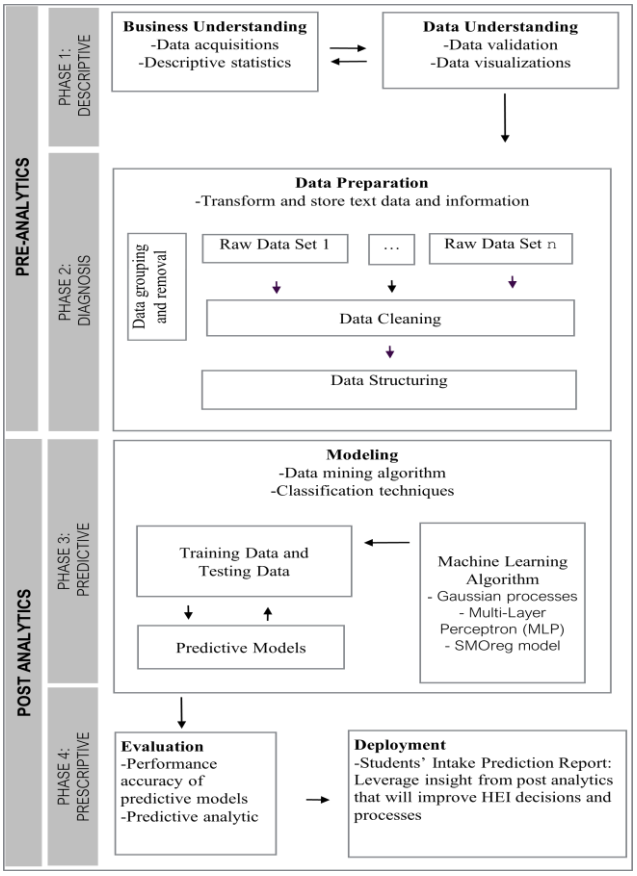

Figure 3: Research framework

4. Predictive analytics on students enrollment data using machine learning algorithm

Predictive analytics include many disciplines such as statistical techniques, machine learning and data mining that analyze current and historical dataset to make prediction or forecast. The overall approach in predicting students' enrollment for polytechnic is to choose an appropriate method. The method used to predict students' enrollment is Gaussian Processes (GP), MLP and SMOreg as for non-linear regressions since the data has uncertainty for students' enrollment by each state. This process is also known as time series analysis. Predictive artificial intelligence model recommend the enrollment prediction for five years from 2017 until 2022. For this experiment, the method that has been applies as follows:

- Grouping the data according to respective polytechnic state location and year from 2012 until 2016.
- Install Forecast packages in Weka.
- Load data into the package.
- For Target Selection, select the instances and values to predict. For this option, all states are chosen.
- For the parameter, insert the number of time units to forecast. In this research, 5 years.
- Time Stamp: Year.
- Periodicity: Yearly.
- On advance configuration, select a base learner configuration (Gaussian Process (GP), Multi-layer perceptron (MLP) and SMOreg)
- Click Start to begin the time-series analysis and prediction.

5. Result and discussions

Gaussian Processes is a time series analysis model that implements Gaussian Processes for regression on fixed dataset without hyperparameter tuning. Hyperparameters are defined as the parameters of the covariance or kernel function. Gaussian Processes is a stochastic model that perform probability distribution over functions. This means a GP can predict the total enrollment for a course given the term and year the course is offered.

Through implementing a time series analysis, Mean Absolute Percentage Error (MAPE) is used to evaluate the accuracy of the predictive model. The formula of MAPE is describe as: sum (abs ((predicted - actual) / actual)) / N where abs is absolute number that eliminates negative value and N is the total. Lower values of MAPE indicates that the forecasted values representing better prediction. Table 1 demonstrate the evaluation of training data with MAPE, MAE and RMSE of GP, MLP and SMOreg.

Table 1: Evaluation results of training data

States	Mean Absolute Error (MAE)			Mean Absolute Percentage Error (MAPE)			Root Mean Square Error (RMSE)		
	GP	MLP	SMOreg	GP	MLP	SMOreg	GP	MLP	SMOreg
Perlis	166.71	0	2.65	7.35	0	0.15	257.67	0	2.66
Penang	231.85	0	2.22	9.01	0	0.09	254.22	0	2.29
Kedah	351.56	0	5.18	8.28	0	0.12	398.21	0	5.48
Perak	191.61	0	2.68	5.44	0	0.08	219.54	0	3.00
Selangor	173.61	0	2.83	5.37	0	0.09	210.81	0	2.97
Kuala Lumpur	28.23	0	0.25	8.68	0	0.09	39.60	0	0.30
Negeri Sembilan	278.07	0	3.53	8.40	0	0.10	325.40	0	3.75
Melaka	236.05	0	1.84	7.09	0	0.06	358.68	0	2.36
Johor	211.54	0	1.22	8.53	0	0.05	226.68	0	1.41
Pahang	432.37	0	3.44	8.40	0	0.08	599.08	0	4.03
Kelantan	280.44	0	1.80	8.33	0	0.06	344.61	0	2.08
Terengganu	240.26	0	2.77	9.64	0	0.10	269.96	0	3.11
Sabah	242.94	0	2.23	9.69	0	0.09	267.24	0	2.46
Sarawak	455.61	0	5.40	9.15	0	0.11	501.43	0	5.72

Figure 4 shows the forecast for polytechnic students' enrollment in each state in Malaysia. The model is generated using all the data that is grouped by states in Malaysia. There are 14 states in Malaysia including Kuala Lumpur.

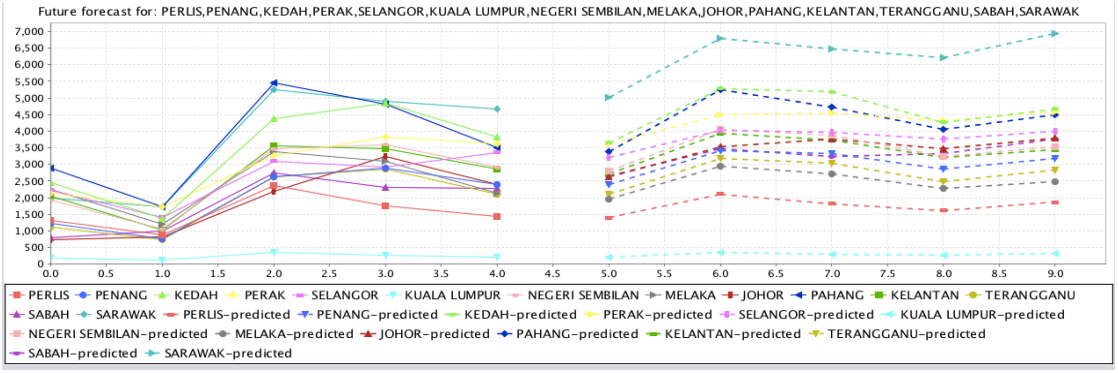

Figure 4: Polytechnic students' enrollment forecast for each state

Number of forecast enrollments for polytechnic students is shown in Table 2. The year with "*" indicates the forecast values for next consecutive years for students' enrollment for each state. Prescriptive analytics automatically synthesize information and applies machine learning to make predictions and then suggests decision options to take benefit of the predictions. Higher education institution can take advantages as prescriptive analytics can give information and advices on student's enrollment administration and management. From the analysis, prediction models are developed for student's enrollment data based on offered course where J48 shows higher accuracy with minimum errors among other classifiers. Using this prediction models, polytechnics can predict whether the current system that has been used is suitable in allocation polytechnic programs offering across all branches in Malaysia. Hence, administrator and all the stakeholders should improve the offering system taking account into students' choice considerations.

Table 2: Number of students' enrollment forecast for 2017 to 2021

YEAR	PERLIS	PENANG	KEDAH	PERAK	SELANGOR	KUALA LUMPUR	NEGERI SEMBILAN	MELAKA	JOHOR	PAHANG	KELANTAN	TERENGGANU	SABAH	SARAWAK
2017*	1394	2383	3657	3604	3219	224	2793	1943	2618	3396	2710	2114	2645	5021
2018*	2113	3410	5289	4483	4036	343	4045	2943	3530	5242	3933	3171	3480	6784
2019*	1817	3320	5177	4540	3964	301	3865	2717	3753	4735	3719	3033	3236	6477
2020*	1611	2859	4272	4306	3751	267	3245	2280	3473	4054	3212	2482	3338	6217
2021*	1860	3173	4668	4573	3988	311	3517	2475	3805	4502	3428	2816	3759	6927

6. Future works

This research is mainly contributing the education data mining fields with more prediction models build by using different dataset in terms of volumes and varieties. The importance of proposing the research can be implemented for the registrar office of Department of Polytechnic Education and other Higher education institutions for better enrollments and distributions. For analysis, more dataset can be added to support current dataset in terms of data understanding and data preparation. Hence, it is important for polytechnic to provide quality education and programs to their students. Furthermore, polytechnic decision maker and stakeholder can utilize the predictive analysis to optimize the facilities and staff allocation for each polytechnic. For future research, it is suggested to use more periodic data for students' enrollment such as semester or batch enrollments and test using another time series analysis algorithm such as Linear Regression to forecast on more historical data with various attributes.

References

Abraham, R. (1999). Emotional dissonance in organizations: conceptualizing the roles of self-esteem and job-induced tension, Leadership & Organization Development Journal, 20(1), 18-25.

Aher, S. B. (2011). Data Mining in Education System using WEKA. International Journals of Computer Applications, pp. 20-25.

Ajay, K. P. and Saurabh, P. (2013). Classification Model of Prediction for Placement of Students. I.J.Modern Education and Computer Science, 11, 49-56.

Alshareef, A., Ahmida, S., Abu Bakar, A., Hamdan, A. R. and Alweshah, M. (2015). Mining Survey Data on University Students to Determine Trends in the Selection of Majors. Science and Information Conference, July 28-30. London, UK.

Arnis, K. and Arkady, B. (2012). A Comparative Analysis of Short Time Series Processing Methods. VERSITA. Information Technology and Management Science. DOI: 10.2478/v10313-012-0009-4.

Bichsel, J. (2012). Analytics in Higher Education: Benefits, Barriers, Progress, and Recommendations (Research Report). Louisville, CO: EDUCAUSE Center for Applied Research, August 2012, Retrieve 14 August, 2017 from http://www.educause.edu/ecar.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Thomas, R., Shearer, C., and Wirth, R., (2000). CRISPDM 1.0 Step-by-step data mining guide. SPSS White paper-technical report CRISPWP-0800, SPSS Inc.

Department of Polytechnic Planning (2015). Projection Capacity of Polytechnic Students (2015-2025). Unpublished document.

Dhar, V. (2013). Data Science and Prediction. Communications of the ACM, Vol. 56 No. 12, pp. 64-73.

Gregory, P (2014). CRISP-DM, still the top methodology for analytics, data mining or data science project. Retrieve from: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Jindal, R. and Dutta, B. M. (2015). Predictive Analytics in a Higher Education Context. IEEE. IT Pro.

Lim, T. Y., Ratnam, M. M. and Khalid , M.A. (2007). Automatic Classification of Weld Defects using Simulated Data and an MLP Neural Network. Journal INSIGHT, Learned and Professional Society Publisher.

Michael, B., Amelia, P., Alexis, W. and Kevin, K. (2017) Predictive Analysis of Student Data. A Focus on Engagement and Behavior. NASPA. Student Affair Administrators in Higher Education. April, 2017.

Mohammad, M. A. M., Shovasis, K. B., Monalisa, C. U. and Abubakar, S. (2015). An Algorithm For Training Multilayer Perceptron (MLP) For Image Reconstruction Using Neural Network Without Overfitting. International Journal of Scientific and Technology Research. Volume 4(Issue 02).

Nur Syahiela, H., Sarina, S. and Siti Mariyam, S. (2016). Tools in Data Science for Better Processing. AIP Conference Proceeding.

Provost, F., and Fawcett, T. (2013). Data Science Its Relationship Data-Driven Decision Making, 1(1), 51–59.

Ramanathan, L., Swarnalatha, P. and Gopal, D. G. (2014). Mining Educational Data for Students' Placement Prediction using Sum of Difference Method. International Journal of Computer Applications (0975 – 8887), Volume 99(Issue.18).

Saeed, A., Hamidreza M., Ashish D., Teh Y. W. and Tutut H. (2014). An Approachable Analytic Study on Big Educational Data Mining. Computational Sciences and Its Application – ICCSA. Springer International Publishing, pp. 721-737.

Siraj, F. and Abdoulha, M. A. (2011). Mining Enrolment Data using Predictive and Descriptive Approaches.

Swamy, M. N. and Hanumanthappa, M. (2012). Predicting Academic Success from Student Enrollment Data using Decision Tree Technique. IJAIS. Volume 4(Issue 3).

Tarekegn, G. B. and Sreenivasarao, V. (2016). Application of Data Mining Techniques to Predict Students Placement in to Departments. International Journal of Research Studies in Computer Science and Engineering (IJRSCSE). Volume 3(Issue 2).