A Regression Model to Predict Key Performance Indicators in Higher Education Enrollments

Ashraf Abdelhadi, Suhaila Zainudin, Nor Samsiah Sani
Center for Artificial Intelligence Technology (CAIT)
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi, Selangor, Malaysia

***Abstract*—Key Performance Indicators (KPIs) are essential factors for the success of an organization. KPIs measure the current performance and identify the ongoing progress for specified business objectives. The Ministry of Higher Education (MoHE) in Palestine used established formulas to predict the KPI. These KPIs are vital for charting the organization aims. This study applies regression models for student enrollment data sets to predict accurate KPIs that can be used and adapted for any higher education system. The predictive engine will determine the KPI based on linear regression techniques such as Lasso, Elastic Net, and non-linear regression such as Support Vector Regression (SVR), and K-Nearest Neighbor (KNN). The Ministry of Higher Education (MoHE) in Palestine provided the datasets related to enrollments and graduations data for different Higher Education Institutions (HEIs). The regression algorithms were evaluated by mean absolute error, mean square error (MSE), root mean square error (RMSE) and the R Squared. The experiment demonstrates that the 40% training with 60% testing splitting using linear regression shows the best result.**

Keywords—Data mining; KPI; regression; higher education; prediction model

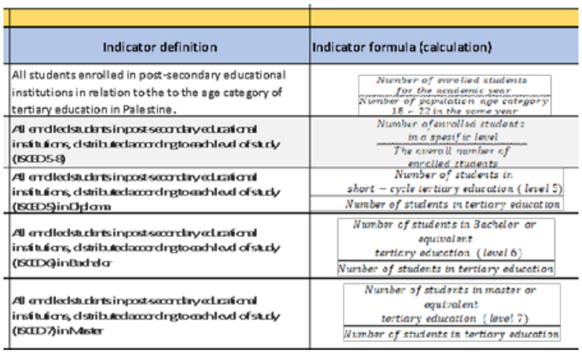
I. INTRODUCTION

Key Performance Indicators (KPIs) are the critical signs of development in the direction of a meant result. KPIs afford a focal point for strategic and operational improvement, create an analytical foundation for decision-making, and assist awareness interest on most topics. KPI performs a critical element given that it is given fast and specific data through evaluating present-day overall performance in opposition to a goal required to fulfil commercial enterprise desires and objectives [1].

Businesses adopted frameworks such as the balanced scorecard (BSC) [2] as a strategic performance metric to improve internal business functions and their outcomes. Correspondingly, education centers, knowledge creation and worker centers such as ministries or learning institutions also benefited from utilizing BSC to chart the KPIs for Higher Education Institutions [3]. On a global scale, the United Nations Educational, Scientific, and Cultural Organization (UNESCO) published a practical guide for educational planners who wish to construct an indicator system [4]. The author in [4] included examples of HEIs, notably the University of Edinburgh and University Technology Malaysia, that planned their strategic development plans alongside a monitoring system, such as BSC.

The structure and content of education systems around the world vary greatly. As a result, they compare national education systems with other countries or benchmark progress toward national and international goals. Hence, UNESCO designed the International Standard Classification of Education (ISCED) to serve as a framework to classify educational activities as defined in programs and the resulting qualifications into internationally agreed categories. The basic concepts and definitions of ISCED are internationally valid and comprehensive of the full range of education systems [5].

However, there is no solid data mining framework and model to predict the higher education (HE) KPI across the world and at MoHE Palestine in particular. For instance, the current MoHE practice to extract and predict KPI is manual. The staff collect the data from different resources by phone and emails, then record it into an excel sheet, as shown in Fig. 1. The formula miscalculated will lead to a wrong decision.



Indicator definition	Indicator formula (calculation)
All students enrolled in post-secondary educational institutions in relation to the to the age category of tertiary education in Palestine.	$\frac{\text{Number of enrolled students for the academic year}}{\text{Number of population age category 18 - 22 in the same year}}$
All enrolled students in post-secondary educational institutions, distributed according to each level of study (ISCED5-8)	$\frac{\text{Number of enrolled students in a spesific level}}{\text{The overall number of enrolled students}}$
All enrolled students in post-secondary educational institutions, distributed according to each level of study (ISCED5) in Diploma	$\frac{\text{Number of students in short} - \text{cycle tertiary education (level 5)}}{\text{Number of students in tertiary education}}$
All enrolled students in post-secondary educational institutions, distributed according to each level of study (ISCED6) in Bachelor	$\frac{\text{Number of students in Bachelor or equivalent tertiary education (level 6)}}{\text{Number of students in tertiary education}}$
All enrolled students in post-secondary educational institutions, distributed according to each level of study (ISCED7) in Master	$\frac{\text{Number of students in master or equivalent tertiary education (level 7)}}{\text{Number of students in tertiary education}}$

Fig. 1. Example of KPIs Formulation.

II. PROBLEM STATEMENT

Although the MoHE has computerized most of its services and automated most operations, the ministry is still facing some issues in the reporting system and predication, which affects the strategic plan for the upcoming years, for instance, predicting wrong enrollment students' number for the forthcoming academic year in government Tertiary Education Institutes (TEIs) can cause in improper budget allocation which means wasting of resources. Also, extracting knowledge from complex data sets takes a long time and a human effort to drill deeply into the big data sets. Therefore, the main worthwhile problem that needs to be addressed is to discover a new fast, efficient, and incredibly accurate computerized approach or data mining algorithm to resolve the KPI extraction and prediction problem primarily for our case study (MoHE) based on the database for the benefit of the higher education management.

Data availability, especially for the education sector, has spurred interest in data-driven decision making [6]. The process of making organizational decisions based on actual data rather than intuition or observation alone is known as data-driven decision making (or DDDM). Therefore, DDDM offers the opportunity to discover a new fast, efficient, and incredibly accurate computerized approach or data mining algorithm to resolve KPI extraction and prediction for the MoHE case study.

III. Related Work

Data mining includes many techniques from other domains such as statistics, machine learning, pattern recognition, database, data warehouse systems and visualization [7]. Most organizations monitor their operation performance and achievement through dashboards and Business Intelligence (BI) [8]. However, in many institutes, this is limited to standard reports which cannot measure the unknown KPIs and in most cases, it is difficult to predict future performance. Most top managers rely on their intuition in order to select their potential KPIs that will lead to redundant KPIs. Managers also focus on the results rather than on the actual indicators that can be used [1].

The author in [1] built a model to predict key performance indicators for Massive Online Open Courses (MOOC) that is very similar to the Cross-Industry Standard Process for Data Mining (CRISP-DM). The model consisted of six stages from defining the business strategy model, definition of KPIs and the multidimensional model. The multidimensional model is composed of two analysis cubes: Enrollment and Activity. The enrollment analyzes the students’ features such as country, interests and expectations and whether these features represent specific patterns. Data mining techniques are used to extract and predict KPIs. These techniques analyze the KPIs to mine the relationships identified during the business strategy modelling. The author in [1] used different algorithms such as Support Vector Machines (SVM), a Random Forest of Decision Trees (DT) and Neural Networks.

In 2015, [9] proposed a framework for predicting students' academic performance. The primary purpose is to discover hidden information and knowledge from the students' data so that the model can predict the student grades in a specific subject based on independent parameters such as GPA, race, gender, family income, university entry mode. The model proposed in [9] used three different classifications algorithms: Decision Tree (DT), Naïve Bayes (NB), and Rule-Based (RB) through the WEKA software tool. The model allows users to categorize the students under two or three categories; good, poor, and average. If this framework and model can be modified to use a regression algorithm, the output can be numbers or percentages, which is more accurate.

The author in [10] built a model to classify attrition among B40 students in bachelor's degree programs in Malaysia's public universities. The machine learning model indicates that the Random Forest algorithm is the best model in predicting student attrition compared to Neural Network and Decision Tree.

The author in [11] applied different machine learning techniques to qualitatively predict the whole project KPIs in critical construction project stages. Artificial neural network (ANN) and the neuro-fuzzy method using fuzzy C-means (FCM) and subtractive clustering to predict project KPIs. The models map the KPIs of three critical project stages to the whole project KPIs. Validation used the data of actual projects to confirm models' effectiveness and compare the results of the employed machine learning techniques.

The author in [12] created a model to predict and identify factors that influence graduates' employability. Seven years of data (from 2011 to 2017) from Malaysia's Ministry of Education were used to test and evaluate the model. They applied three different algorithms; Decision Tree, Support Vector Machines and Artificial Neural Networks. The results show the decision tree (J48) produces higher accuracy compared to other techniques. Also, according to this study, three factors, attribute age, industrial internship, and faculty, contain the most information and affect the final class, which is employability.

TABLE I. Summary of related work

Reference	*Theme (concept)*	*Findings/Conclusions*
[1]	Data mining framework and KPI Predictive model	It is a good model but without a clear framework that can cover the whole KPIs prediction process.
[9]	Data mining framework and KPI Predictive model	The model can predict the student grades (dependent parameter) in a specific subject based on independent parameters such as GPA, race, gender, family income, university entry mode. The study focused on being more comparative between three algorithms. The result is a lack of graphs and charts that clearly show the output and the output discrete, not a continuous number.
[11]	KPI Predicative model	All KPIs were measured qualitatively by designing a questionnaire, and there is no database containing accurate records. Also, The research measures project performance from the owner's point of view.
[12]	Predictive model	Created a model to predict and identify factors that influence graduates' employability.
[10]	Predictive model	Built a model to classify attrition among B40 students in bachelor's degree programs in Malaysia's public universities.

IV. Materials and Methods

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a methodology model with six stages describing the information technology existence cycle. It will help plan, organize, and enforce data science (or machine learning) tasks Fig. 2. It standardizes data mining techniques throughout industries, analytics, and data science projects.

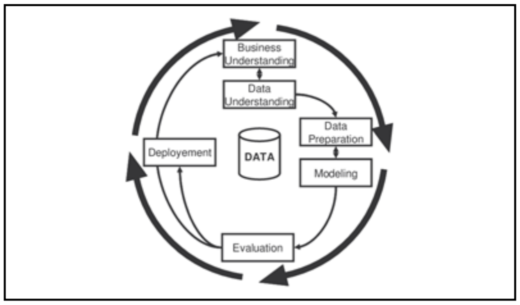


Fig. 2. CRISP-DM Diagram.

The six CRISP-DM Phases are Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment.

The research experiment will apply the six CRISP-DM Phases on the Ministry of Higher Education and Scientific Research (MoHE) in Palestine [13] focus on the KPIs related to students' enrollments according to INTERNATIONAL STANDARD CLASSIFICATION OF EDUCATION (ISCED) [14] such as enrolled students in post-secondary educational institutions, distributed according to each field of study:

1) According to Education Program.
2) Arts and Humanities.
3) Social Science, Journalism and information.
4) Business, Administration and law.
5) Natural Science, Math and Science.
6) Information and communication technology.
7) Engineering Manufacturing and Constructions.
8) Agriculture, Forestry, Fisheries and Veterinary.
9) Health and Welfare.
10) Services.

UNESCO designs ISCED to serve as a framework to classify educational activities as defined in programs and the resulting qualifications into internationally agreed categories. Therefore, the basic concepts and definitions of ISCED are intended to be internationally valid and comprehensive of the full range of education systems [14].

The details of the methodology followed in this study is explained below.

Phase 1: Business Understanding

This phase concerns determining the business goals which is to predict a set of KPIs that has been defined in higher education and the best practice to measure those KPIs.

Phase 2: Data Understanding and Data Resources Analysis

At this phase, the data resources have been prepared for modelling, including several activities such as data selection, data cleaning, data construction, data integration, combining data from multiple sources, and re-formatting data as necessary. Data Source identification (databases, schema names, tables, view, spreadsheets), SQL scripts performed to create specific views in the staging database, combine data from multiple sources to one repository pre-processing data stage, including data selection, cleaning and integration.

The enrollment and graduation data form the core sources [13] for our data mining experiments. For instance, the original enrollment table consists of 50 attributes and 3,895,158 instances as it contains the historical data since the MoHE establishment. The enrollment attributes (fields) were identified to contain 34 features and 3862763 instances as some fields duplicated for both English and Arabic values. The graduation data sets have 461,598 instances and 24 attributes.

Building Database Repository using SQL server:

The database repository is built based on main tables such as enrollment, graduations, ISCED levels, programs, and degrees, in addition to many lookup tables such as high schools' lists, districts, nationalities, universities lists. Data views were created to focus on the data from 2014 to 2018, including 25 attributes and other attributes from other tables containing the ISCED data, which is essential for data mining. Some repeated features (fields) such as the Arabic values have been eliminated because it's considered duplicate values, other values replaced with null values excluded.

Data Cleaning and Transformation:

Any noisy and inconsistent data were removed to handle the missing data fields, transform data into forms appropriate for the mining task, for instance, the area code to numbers from 1 to 16, the high school types coded from 1 to 5 and the high school stream coded to numbers as well (Tables II, III and IV) The data is split into 60% training and 40% testing sets.

TABLE II. AREA CODE DATA TRANSFORMATION

CODE	***Area***
1	Quds
2	Hebron
3	Ramallah
4	Bethlehem
5	Nablus
6	Tukaram
7	Qalqilya
8	Sal fit
9	Jenin
10	Jericho
11	Gaza
12	Middle Gaza
13	Khan Younis
14	DerAlbalah
15	Rafah
16	Tubas

TABLE III. HIGH SCHOOL TYPE DATA TRANSFORMATION

CODE	***HS_Type***
1	Gov. High School
2	Bajrout
3	GCE
4	IB
5	SAT

TABLE IV. HIGH SCHOOL STREAM DATA TRANSFORMATION

CODE	***HS-Stream***
1	Humanities
2	Literature
3	Science
4	Industry
5	Economic
6	Agriculture
7	Nursing
8	Hospitality
9	Islamic Study
13	Applied Industry
14	Applied Agriculture
15	Vocational
19	IT
20	Entrepreneurship
21	Technology

Phase 3: Modeling

Regression predicts a range of numeric values or continuous values. For example, a regression model that predicts KPI values could be developed based on observed data for many other factors such as enrolled students, specific programs, number of graduates throughout history.

Numerous models were constructed and assessed primarily based on numerous techniques. In this study employed Linear Algorithms: Linear Regression (LR), Lasso Regression (LASSO) and Elastic Net. The study also applied nonlinear algorithms such as Support Vector Regression (SVR), and K-Nearest Neighbors (KNN) using Python. In terms of parametrization, the variable "ISCED_Level1_Id" is assigned as a target to be predicted. To generate the training, the random_state variable is assigned to 1 to replicate results with frac=0.6. Then select any data, not in the training set and include it in the testing set based on the index, test = df.loc [~df. index. isin(train.index).

Predicting ISCED KPIs:

In this study, the first experiment is to predict the KPIs for enrolled students in post-secondary educational institutions, distributed according to every ten fields of study based on the first level of (ISCED) and the general studies. So, the model will predict KPIs according to the 10 identified field of study. The second experiment is to predicts find the ratio between enrollment and graduation based on the graduates data sets.

Phase 4: Evaluation

There are three metrics for evaluating predictions in regression; Mean Absolute Error, Mean Squared Error, and R2. The Mean Absolute Error (or MAE) is the sum of the absolute differences between predictions and actual values. It gives an idea of how wrong the predictions were. The measure provides a picture of the magnitude of the error but no idea of the direction (e.g., over or under predicting). A value of 0 indicates no error or perfect predictions.

The Mean Squared Error (or MSE) is just like the implied absolute mistakes in that it affords a gross concept of the significance of the mistakes. Taking the rectangular root of the implied squared mistakes converts the units lower back to the unique units of the output variable and may be significant for description and presentation. This is referred to as the Root Mean Squared Error (or RMSE). So, for instance, if MSE= -34.705 and SD =45.574, this metric is inverted to increase the outcomes.

The R2 (or R Squared) metric illustrates the goodness in the shape of a fixed of predictions to the actual values. In statistical literature, this degree is referred to as the coefficient of determination. This is a value among zero and 1 for no-match and best match, respectively. For example, if R2 = 0.2, the predictions have a negative match to the real values with a value toward 0 and much less than 0.5. The last stage is the deployment with the task of plan deployment and tracking, produce the final report, and review tasks by conducting an assignment retrospective approximately what went well, what might have been better, and a way to enhance it [1].

V. EXPERIMENTAL RESULTS AND DISCUSSION

Before applying different algorithms for different datasets based on the academic years, the three linear regression algorithms were tested with varying percentages of splitting of training and testing data (10% to 90%) (Table V).

According to [15] (scikit-learn, 2021), Linear Regression fits a linear model with coefficients $w = (w_1, \dots, w_p)$to minimize the residual sum of squares between the observed targets in the dataset and the targets predicted by the linear approximation. Mathematically it solves a problem of the form;

$$\underset{w}{min}||Xw - y||_2^2$$

TABLE V. SHOWS LINEAR REGRESSION RESULTS FOR DIFFERENT SPLIT

Algorithm Linear Regression	***Time (S)***	***MAE***	***MSE***	***RMSE***	***R Squared***
10% to 90%	9.6	0.00000006210088990	0.00000099729	0.000998644080	0.98771090
20% to 80%	9.95	0.00000006220099000	0.00000500290	0.002236716340	0.97871090
30% to 70%	9.95	0.00000005018788899	0.00000149280	0.001221801959	0.9680980
40% to 60%	9.87	0.00000004218000023	0.00000098129	0.000990600837	0.99871990
50% to 50%	9.9	0.00000007657778899	0.00000145280	0.001205321530	0.97771440
60% to 40%	10.49	0.00000004656890001	0.000000997522	0.000998761990	0.98871090
70% to 30%	9.95	0.00000005865412345	0.00000090020	0.000948788709	0.977087155
80% to 20%	11.21	0.00000006643210008	0.00000172280	0.001312554765	0.95571870
90% to 10%	10.15	0.00000008090087799	0.00000242280	0.001556534613	0.90871090

Lasso Regression:

The Lasso is a linear model that estimates sparse coefficients. It is helpful to prefer solutions with fewer non-zero coefficients, effectively reducing the number of features the given answer depends on. For this reason, Lasso and its variants are fundamental to the field of compressed sensing. Under certain conditions, it can recover the exact set of non-zero coefficients.

$$\min_{w} \frac{1}{2n_{\text{samples}}} ||Xw - y||_2^2 + \alpha ||w||_1$$

The Lasso estimate thus solves the minimization of the least-squares penalty with $\alpha||w||_1$ is the l_1The implementation in the class Lasso uses coordinate descent as the algorithm to fit the coefficients [15] (scikit-learn, 2021). Table VI shows the experimental results for lasso regression when applying different splitting.

Elastic Net is a linear regression model trained with both l_1 and l_2-norm regularization of the coefficients. This combination allows for learning a sparse model where few of the weights are non-zero, like Lasso Elastic-net, which is beneficial for multiple features correlated with each other, such as high school average and high school stream. Lasso is likely to pick one of these at random, while elastic-net is likely to determine both [15] (scikit-learn, 2021).

$$\min_{w} \frac{1}{2n_{\text{samples}}} ||Xw - y||_2^2 + \alpha\rho ||w||_1 + \frac{\alpha(1-\rho)}{2} ||w||_2^2$$

Table VII shows the experimental results for Elastic Net regression when applying different splitting.

When comparing the three linear algorithms (Fig. 3), the Linear algorithm score the lowest error compared to the Lasso and Elastic Net.

TABLE VI. LASSO REGRESSION FOR DIFFERENT SPLITTING

Algorithm (Lasso Regression)	***Time (S)***	***MAE***	***MSE***	***RMSE***	***R Squared***
10% to 90%	9.65	0.22752037948	0.124891937503	0.3534005341	0.982836699366
20% to 80%	9.73	0.228279447022	0.127591253438	0.3571991789	0.982481802987
30% to 70%	9.47	0.227984559434	0.125763102058	0.3546309378	0.982732256546
40% to 60%	9.65	0.220175387930	0.126210860621	0.3552616790	0.989648213930
50% to 50%	9.2	0.228149889028	0.126973529679	0.3563334529	0.982525680214
60% to 40%	9.72	0.227763064940	0.126232959929	0.3552927805	0.982659892778
70% to 30%	9.43	0.226355016909	0.125727079609	0.3545801455	0.982683063141
20% to 80%	9.8	0.229283720738	0.128320218693	0.3582181160	0.982302143850
90% to 10%	9.7	0.229283720738	0.128320218693	0.3582181160	0.982302143850

TABLE VII. ELASTIC NET REGRESSION FOR DIFFERENT SPLITTING

Algorithm (Elastic Net Reg.)	***Time (S)***	MAE	MSE	RMSE	R Squared
10% to 90%	11	0.228353991688426	0.124519448633917	0.352873133907807	0.983113221419126
20% to 80%	10	0.228482139545430	0.124887485646334	0.353394235445818	0.983121343318927
30% to 70%	12	0.228194487547074	0.125319877380630	0.354005476483952	0.983100147069539
40% to 60%	9.43	0.221302607034649	0.124985776002300	0.353533274250530	0.989003814519843
50% to 50%	11.35	0.228598585503690	0.124864650076610	0.353361925052220	0.983053758064118
60% to 40%	12	0.228957846923130	0.124948105589363	0.353479993195319	0.983181872293709
70% to 30%	11.77	0.228222951386659	0.123474484082960	0.351389362506835	0.98337124158077
80% to 20%	9.45	0.229529023543800	0.128214370176883	0.358070342498346	0.982728184185912
90% to 10%	9.65	0.228076229470795	0.120055825721761	0.346490729633219	0.983591490829027

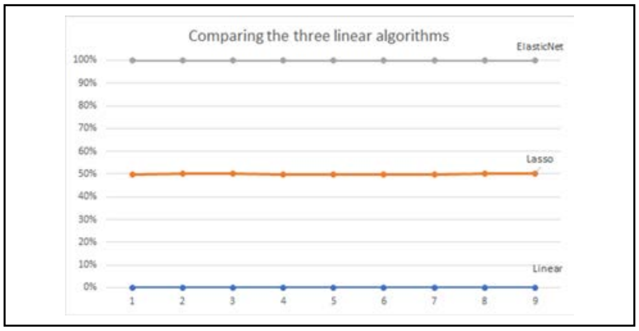


Fig. 3. Comparison between the three different Linear Algorithms.

Moreover, the slightest error margin was 40% training and 60% testing data sets (Fig. 4). Therefore, the rest of the algorithms tested for the exact sampling percentages (40% training and 60% testing) for the same academic year. Then, we look at the different iterations for three algorithms with varying percentages of data sampling (training and testing). There is no significant difference using the same algorithm for further selection, but there is a difference when it comes to the non-linear algorithms.

The experiment conducted for the same academic year, the enrollment KPI was based on ISCED level one for five different algorithms, as shown in Table VIII.

The ISCED KPIs predicted values for enrolled students in post-secondary educational institutions, distributed according to the 10 fields of study. Fig. 5 shows that the values are very close to the actual values.

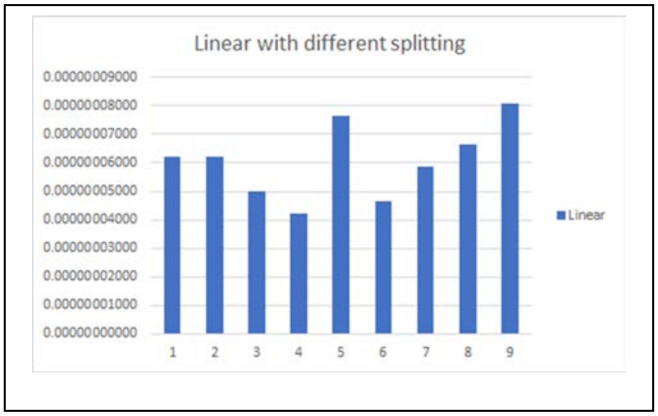


Fig. 4. Margin Error based on % Splitting Sampling.

TABLE VIII. KPI PREDICTION ERROR

Algorithm	***Time(M.S.MS)***	***MAE***	***MSE***	***RMSE***	***R2***
Linear Regression	9.87	0.00000004	0.000000981	0.000990454	0.99871090
Lasso Regression	9.65	0.22752037	0.123148555719873	0.3509252850	0.9831577460
Elastic Net Regression	9.43	0.22887746	0.125000076102056	0.3535534982	0.9830981775
Support Vector Regression (SVR)	47.71	0.66592015	6.968257047	2.6397456407	-16.320
K-Nearest Neighbors (KNN)	3.29.50	0.66537128	1.406274302	1.1858643694	0.7453734924

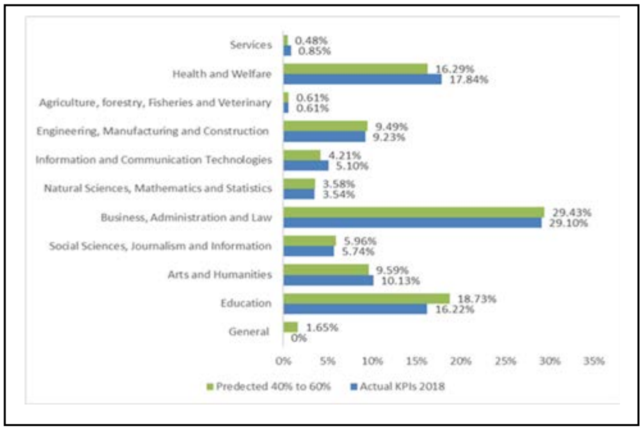


Fig. 5. Comparison between the Actual and Predicted ISCED KPIs.

The second Experiment was to find predicted ratio between enrollment and graduation. Fig. 6 shows the ratio between the predicted enrollment and graduation KPIs based on ISCED level 1.

The ISCED numbers in Fig. 6 can be translated as per Table IX.

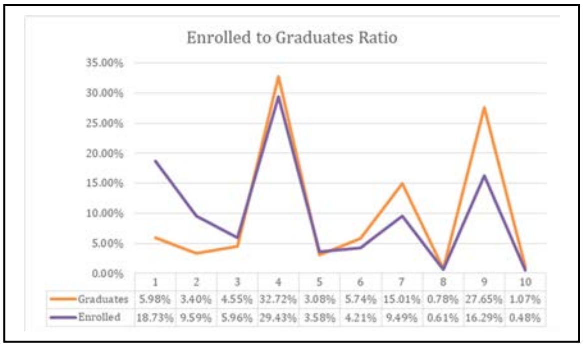


Fig. 6. Enrollment to Graduates Ratio.

TABLE IX. ISCED LEVEL 1 DESCRIPTION

ISCED_Level1_Description	***ISCED_Level1_Id***
Education	1
Arts and Humanities	2
Social Sciences, Journalism and Information	3
Business, Administration and Law	4
Natural Sciences, Mathematics and Statistics	5
Information and Communication Technologies	6
Engineering, Manufacturing and Construction	7
Agriculture, forestry, Fisheries and Veterinary	8
Health and Welfare	9
Services	10

VI. CONCLUSION

In conclusion, with clear and coherent strategies, figuring out the present-day situations, operation sector, unique varieties of competencies that generate, performance will lead to success. To create this kind of situation calls for the provision of strategic records to confirm the current situations, to outline the strategy [16]. Also, applying Cross-Industry Standard Process for Data Mining (CRISP-DM) process model as a research methodology to develop a data mining model that could help be adapted by individuals and HEIs, using machine learning algorithms can lead to good results and accuracy [17]. However, without clear KPIs, it's challenging to have a clear strategy for the upcoming years. It is crucial to create an analytical model to act as the basis for decision making and help focus attention on HE enrollment. This study provides a practical solution for such a problem by proposing a KPIs predicting model from available data at MoHE and integrating the data from different resources into a database repository from which KPIs will be predicted. This model tested different regression algorithms such as linear regression, Lasso, Elastic Net; non-linear Support Vector Regression (SVR) and K-Nearest Neighbors (KNN. However, the most successful predictive model and particularly in performance indicators used was Linear regression. The training and splitting data were tested from 10% to 90%, the targets values were compared from the historical data in the last few years. The regression algorithms were evaluated by mean absolute error, mean square error (MSE), root mean square error (RMSE) and the R Squared. The 40% training with 60% testing splitting using linear regression shows the best result. In the future, this model can be part of a complete HE Framework to predict the KPIs and act as the main engine for that Framework.

ACKNOWLEDGMENT

The authors would like to thank Universiti Kebangsaan Malaysia (UKM) and DTK Grant (TT-2020-015) for supporting this research and the General Directorate of Research and Development at Ministry of Higher Education and Scientific research-Palestine for providing the data.

REFERENCES

[1] Peral, J., Maté, A. and Marco M. Application of Data Mining techniques to identify relevant Key Performance Indicators. Computer Standards and Interfaces Volume 54, Part 2, November 2017, Pages 76-85, 2017.

[2] Kaplan, R. S. 2009. Conceptual Foundations of the Balanced Scorecard. Handbooks of Management Accounting Research 3: 1253–1269. doi:10.1016/S1751-3243(07)03003-9.

[3] Weerasooriya, R. B. Adoption of the Balanced Scorecard (BSC) Framework as a Technique for Performance Evaluation in Sri Lankan Universities. SSRN Electronic Journal (November). doi:10.2139/ssrn.2223933, 2013.

[4] Martin, M. and Sauvageot, C. 2011. Constructing an indicator system or scorecard for higher educ Martin, M. and Sauvageot, C. Constructing an Indicator System or Scorecard for Higher Education. A Practical Guide. UNESCO International Institute for Educational Planning. Paris, 2011. ISBN: 978-92-803-1329-1. Pages: 83.

[5] I. S. The International Standard Classification of Education (ISCED). In Prospects (Vol. 5, Issue 2), 1975. [online]. Available: https://doi.org/10.1007/BF02207511.

[6] Ballou, Brian, Heitger, Dan L. and Stoel, Dale, (2018), Data-driven decision-making and its impact on accounting undergraduate curriculum, Journal of Accounting Education, 44, issue C, p. 14-24, https://EconPapers.repec.org/RePEc:eee:joaced:v:44:y:2018:i:c:p:14-24.

[7] Hartama, D., Windarto, A. P., and Wanto, A. (2019). The application of data mining in determining patterns of interest of high school graduates. Journal of Physics: Conference Series, 1339(1) doi:http://dx.doi.org/10.1088/1742-6596/1339/1/012042[8] Stefanovic, N. 2015. Collaborative predictive business intelligence model for spare parts inventory replenishment. Computer Science and Information Systems 12(3): 911–930. doi:10.2298/CSIS141101034S.

[8] Ahmad, F., Ismail, N. H. and Aziz, A. A. The prediction of students' academic performance using classification data mining techniques. Applied Mathematical Sciences 9(129): 6415–6426. doi:10.12988/ams.2015.53289, 2015.

[9] Sani, N. S., Nafuri, A. F. M., Othman, Z. A., Nazri, M. Z. A., and Nadiyah Mohamad, K. Dropout Prediction in Higher Education Among B40 Students. International Journal of Advanced Computer Science and Applications, 11(11), 550–559. https://doi.org/10.14569/IJACSA.2020.0111169, 2020.

[10] Fanaei, S. S., Moselhi, O., Alkass, S. T. and Zangenehmadar, Z. Application of Machine Learning in Predicting Key Performance Indicators for Construction Projects. International Research Journal of Engineering and Technology: 1450. Retrieved from www.irjet.net, 2018.

[11] Othman, Z., Shan, S. W., Yusoff, I., and Kee, C. P. Classification techniques for predicting graduate employability. International Journal on Advanced Science, Engineering and Information Technology, 8(4–2), 1712–1720. https://doi.org/10.18517/ijaseit.8.4-2.6832, 2018.

[12] Ministry of Higher Education and Scientific Research (MoHE) in Palestine (Arabic only), 2020. [online]. Available: http://www.mohe.pna.ps.

[13] I. S. The International Standard Classification of Education (ISCED). In Prospects (Vol. 5, Issue 2), 2012. [online]. Available: https://doi.org/10.1007/BF02207511.

[14] Machine learning in Python, 2020. [Online]. Available at http://www.Scikit-learn.org.

[15] Valdez, A., Cortes G., Castaneda, S., and Laura, V. Development and Implementation of the Balanced Scorecard for a Higher Educational Institution using Business Intelligence Tools. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10, 2017.

[16] Mahmud,Y., Shaeeali, N. S., Mutalib S,. Comparison of Machine Learning Algorithms for Sentiment Classification on Fake News Detection. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 10, 2021.