Predictive Analytics on University Student Dropouts from Online Learning due to MCO

Md Erfan Sultan, Mohd Norshahriel Abd Rani, Nabilah Filzah Mohd Radzuan and Lim Huay Yen

INTI International University, Malaysia, {i17012589@student.newinti.edu.my, mnorshahriel.arani@newinti.edu.my, nabilah.radzuan@newinti.edu.my, huayyen.lim@newinti.edu.my}

ABSTRACT

COVID-19 has and have been affecting the lives of millions of people all around the world. Some of which involves in adapting to the work-from-home culture. It is applicable to all individuals who had to perform their tasks from the comfort of their home whether if it is a 9-to-5 job, running a business or students who are continuing their studies. Observations made shown that students have been facing problems when it comes to attending virtual classes. Hence, this study will be focusing on university students comprising of undergraduate, post-graduate and doctorate students who will be dropping out due to internal or external factors. In order to predict the number of students whom will be dropping out during their online studies; using data mining techniques such as classification techniques and prediction algorithms inclusive of K-Nearest Neighbor (KNN), Logistic Regression, Random Forest, Decision Tree, Naïve Bayes, Support Vector Machine and Genetic algorithms. Each of the algorithms will performed their unique features and percentage of accuracy when making predictions. Along with Association Rule mining such as the Apriori algorithm to identify the causes and effects between the identified factors. The testing data collection will be done manually via questionnaires distributed to students who are currently pursuing their online studies. Key factors involved in this study are age, financial, motivational factors and many more. The key beneficiary from this project will be educational institutions (major) concerned by their ratings, number of dropouts or even a significantly lesser number of student enrolling whether new or returning. The minor audience are the higher education students.

Keywords: Data Analytics, Predictive Analytics, Classification, Association Rule Mining, Data Mining, Algorithm, Higher Education, Student, Dropout.

I INTRODUCTION

COVID-19 has and have been affecting the lives of billions of people all around the world since the dawn of the year 2020. This had made people to adapt the work from-home culture. It is applicable to all individuals who had to perform their tasks from the comfort of their homes and students who are continuing their studies. Business meetings and classes have been conducted online during the time of quarantine. This has led to affect various industries by this pandemic some of which includes education. Businesses and services have come to a halt when the World Health Organization had declared COVID-19 as a pandemic and education was being affected as well. According to UNESCO, schools as a collective had started to close towards the end of February. About 298 million of students were labelled as affected learners, approximately 20% of them were enrolled over 5 countries. As of the second of April, 195 country's schools were closed, affecting about 1.6 billion of students (UNESCO, 2020).

Well known institutions such as Harvard University in the United States, admissions rates have shot up in the span of a year from 4.5% to 5.3% where they have been admitting more students as they were before (Delaney, 2019). An article from CNBC's forum states that the acceptance rates have a potential to skyrocket after the pandemic is over (Dickler, 2020). Campuses are taking in students for the upcoming intakes and have been continuously admitting students in advance as a precautionary measure to avoid losses in business and credibility as well as to stay in parity with the target tuitions. Amidst the COVID-19 pandemic, colleges and universities all around the globe has shut down and opted to conduct virtual classes to prevent the spread of the virus. According to Hess (2020), a significant number of students agree that due to health concerns along with the following of the public health officials, approved that face-to-face classes must come to a halt. This includes of the readiness of the management and students to attend online classes with the same amount of energy, (The College Finance Team, 2020).

Due to this radical shift, students have not been involved or anticipated in any virtual classes before. Therefore, students have started requesting reduction of fees from their institutions because they have not been using the institutions' facilities as they were before although the same amount of fees were being

charged. On the other hand, some students have opted to dropout straight away due to institutes demanding the same fees while students are not being to fully utilize their resources and experience. Small to medium-sized colleges and universities which rely fully on tuition to run the establishments are to face major difficulties according to Anderson, (2020). Colleges were warned that there would be less students attending because the worries of not receiving the same treatment as well as financial difficulties which will make students unable to bear the expenses according to the CEO of College Census, Jeremy Adler, as reported by Hess, (2020). However, institutes that have been operating online before the pandemic has been receiving more positive sentiments from students. Organizations such as Coursera, as a marketing strategy, had released a few of their courses free-of-charge for a certain period attracting more students to opt for that as an alternative.

Hence, this study will be focusing on university students comprising of undergraduate, post-graduate and doctorate students who will be dropping out due to internal or external factors. In order to predict the number of students which and whom exactly will be dropping out during their online studies by using data mining techniques such as classification techniques and prediction algorithms. Each of the algorithms having their unique features and percentage of accuracy when making predictions. Along with Association Rule mining algorithms, to identify the causes and effects between the identified factors. Key factors involved in this study are age, financial, motivational factors and many more. The key beneficiary from this project will be educational institutions concerned by their ratings, number of dropouts or even a significantly lesser number of student enrolling whether new or returning as well as higher education students as a minor audience.

II BACKGROUND OF STUDENT ATTRITION

Dropouts are referred to students who voluntary leave their course or programme that they were studying in an institution before completion. The act of abandoning involves various factors which lead students to leave their academic journey. Student attrition is a term used to describe number of students reducing over time. Dropouts emerge from multiple factors which leads them to do so. The primary factor being; students' financial issues. It occurs when students are not able to continue paying the high course fees demanded from the particular institutions.

The other contributions to the dropout rate could be the student is facing issues in terms of their performance caused by various other factors such as health both mentally and/or physically. There have been various studies conducted to examine this phenomenon of studies either student leaving their academic voluntary or failing to uphold their academic performances in their respective educational institute. The studies that had been carried out by the authors, identified the factors that affected students to dropout as well as retaining the students; the ways to make the students stay and continue. Still, it has been an issue throughout the most institutions (Burke, 2019). This had been strongly affecting universities that fully rely on fees from students to run the establishment. The consequences resonate from that particular individuals' actions to affect the institution's name and ranking as well as society.

One of the earliest studies conducted by Vincent Tinto, an author that has been mentioned by various researchers created the "Tinto's Model", on student retention and dropout published in 1975, (Tinto, 1975). It is one of the first successful studies, conducted to understand and to study on student retention and dropping out. This model had included student characteristics such as demographics, ethnic, communal, household, educational upbringing, socioeconomic status, psychological profile, and academic progress. The model suggested that social and academic incorporation into the educational institution acts as the foremost determinant of a student to reach graduation. Tinto's model had also found a correlation between the family background, personal characteristics, former schooling, previous academic performance, and interactions amongst the respective faculty and student as mentioned in the journal of (Yaacob, et al., 2020) which uses data mining techniques to forecast student dropout at Universiti Teknologi Mara which the dataset was being obtained from the Department of Academic Affairs and Internationalization of the institution. Decades have passed and researchers had elaborated that (Tinto, 1975) model was flawed, tested by (Brunsden, et al., 2000) and concluded that the model was not proper for dropout and attrition research due to lack of significance between the factors which were correlated by the author.

Coming back to recent years, an identification of explanatory factors for students from an accounting program of bachelor's level from a public university in Brazil. The number of participants participated in the research were almost 400 students involves in both qualitative and quantitative methodologies. The quantitative were logistic regression and qualitative were semi-structured interviews, (Durso & Cunha, 2018). Another study conducted in 2018, (Balraj & Maalini, 2018), demonstrates the use of Naïve Bayes algorithm for prediction and its significant accuracy of 83%, was conducted to find the factors causing students to dropout. The dataset collection was done

from UCI Machine Learning Repositories which included various variables such as demographic, social, school, grades, etc. 2 datasets within the selection were on the subjects of Computer Science programme: Mathematics and Portuguese at the Residential University. A sample of 220 students were conducted and the results shown that universities could use this study in reducing dropouts and increasing their enrolment rates with the help of the findings. A prominent study on predicting virtual learning dropouts by (Yukselturk, et al., 2014), mentioned in the previous study by (Balraj & Maalini, 2018) and many others had involved data mining techniques as well. This study had used below a sample size of 200 students where application of data mining techniques such as k-Nearest Neighbour, Decision Tree, Neural Network and Naïve Bayes was demonstrated. 10-fold cross validation was applied to perform the training and testing process for prediction where k-Nearest Neighbour showed a significantly higher detection sensitivity than the rest. The dataset consisted of not only the demographics from the previously mentioned studies but also data on self-efficacy, locus of control, readiness, and prior knowledge on the course.

Institutes that are highly dependent on tuition fees for maintenance and salaries to its staff are the ones strongly affected during this Covid-19. These are the institutions that do not received funding from the governmental organization and being in an organization where stakeholders have a chance of opting out from universities which do not yield much profits. Reduction in number of students leads universities to receive lesser amount of fees which leaves the institutes unable to continue being in the education industry. This would lead to retrenchment of staffs and management and it would lead to bankruptcy. According to OECD (Organization for Economic Co-operation and Development), about 40% of students pursuing their Undergraduate actually graduated within the duration of study while another 28% graduated outside their actual period as per their study plans (Guerra & Coates, 2019). As per an electronic publication by the World Bank, it had declared that this phenomenon brings a strong negative effects to countries with developing economies and also this has also failed to aid the poverty reduction (World Bank Group, 2015).

David Laude, a professor from the University of Texas undertook a study upon realizing that the professor himself was responsible for the dropouts of hundreds of students. In his study, he had looked into the profiles of students that had dropped out. The results came out to neither be their academic performance nor their determination (how hard they had studied), instead it was found to be correlated to their household income. He suggested that 30% of the students were prone to graduate in time who were in economic need. Students from a high-income household tend to graduate at double the rates in comparison to ones from a poor income regardless of their SAT/pre-university course (PBS NewsHour, 2015). From his observations, he states that student had a tendency of not feeling that they belong in the respective communities. In this situation, the student would feel the emotions of being left out of place due to not being able to cope with many other of their peer's mannerisms and lifestyles. Students with qualifications are the future contributions to the nations' economy. Therefore, the need to obtain an education qualification is important and the focus of retaining them in the institutions must be addressed by the management. According to the World Bank Group, 2016, the employee of a higher educational qualification will obtain a higher wage in any companies.

III METHODOLOGY

The institutions will be aided by applying a data-driven system from student's enrolment, academic and personal (socioeconomic) data. This system will be used to predict students' unique ways in order to increase the retention rates and decrease the rates of dropouts. Big data has become a buzz word and has taken interest by various organizations all around the world. Nowadays, generating massive amounts of data every day, more than 1000 petabytes to be more precise (Lackey, 2019), and has become more valuable than oil (Silva, 2019). The gain in popularity was achieved due to its capabilities of obtaining hidden patterns and insights from data sources previously deemed as futile. The hidden knowledge includes identifying previously unknown business opportunities. It involves in a complex procedure of analyzing large amounts of data. The ability of generating reports which helps decision makers to make strategic, smart, and informed business decisions and performing analysis on various markets. It also helps in improving operational efficiency. The system allows managers to in developing strategies to gain competitive advantages against competitors. It also has the ability to improve business requirements and provide better customer service (Rouse, et al., 2019). Big data analytics supports in decision making, suggestions, and identifying useful information through various techniques from one organization to another.

A. Data Mining

The journey of data analysis begins by collecting data from various sources. These raw data will be combined by cleaning, manipulating, and pre-processing the data. After the processed being done, it is ready to be analyzed to obtain valuable information. The findings are then presented in the

form of visualizations i.e. graphs, charts, etc. It would be written in a report to be dedicated for the respective departments in an organization (Rouse, et al., 2019). Data mining acts as a part of big data analytic, which is the process applied to perform data analytics as mentioned earlier. It is used to search for valuable evidence and identify patterns hidden within a large raw dataset by applying algorithms that are then used to describe the dataset. After which, a machine learning or predictive model is created using various techniques and later deployed for ETL (Extract-Transform-Load) processes. This then automatically starts to execute the manually performed tasks by the analyst prior to deployment (Twin, 2019).

When it comes to machine learning, there are actually three techniques in total: (i) Supervised, (ii) Unsupervised, and (iii) Reinforced Learning. In this context, we will be only looking into the first two types. Supervised Learning applies to when data from a given dataset is labelled or classified and the output is not known. However, both the inputs and outputs are labelled, the learning takes place by the help of mapping. From the known input data, the output is derived. It is usually used for Classification and Regression problems. Where the out of the Classification problems are categorical and Regression problems involving real values are outputs. Classification problems may include Random Forest along with Support Vector Machines. Whereas Unsupervised Learning refers to the output or the results being unknown from known or labelled input data. In this technique, the algorithms used determines which possible labels are able to produce the known results and the learning takes place from trial and errors (Brownlee, 2019). In the learning process, data is split into training and test data in either an 80:20, 60:40, or 70:30 splitting ratio of training and testing data, respectively.

Therefore, this research will apply Knowledge Discovery in Databases methodology (KDD). KDD is referred to the complex yet illustratable process in finding valuable knowledge from large amounts of data. This process involves various disciplines and techniques including Computation and Statistics to depict hidden patterns (Fayyad, et al., 1996). Figure 1 shows the step of KDD. It starts off with obtaining data from various sources along with studying and understanding the domain of interest. After obtaining the appropriate datasets the suitable ones are selected to be cleaned. Cleaning involves removing and manipulating the data to make it standardized for databasing are known to be containing inconsistent, incomplete, complicated, and problematic records. This stage is known to be the most time-consuming step throughout the entire process. The cleaned data is then transformed by identifying the features based on the goal. It is transformed into the suitable form for analysis. After which data mining tasks take place. It involves in choosing the data mining algorithm such as classification and association for identifying the patterns then to decide which models to apply or whether to build a customized model. At the end, from the findings the analyst evaluates and represents the knowledge found in the forms of reports. It then becomes an automated process of knowledge gathering up until further tweaks are required.

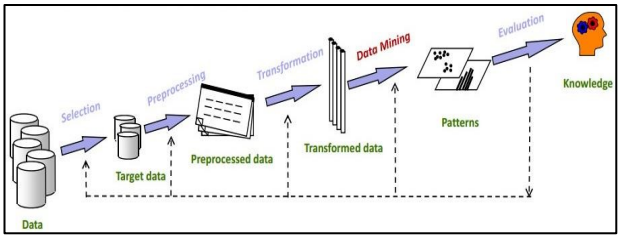

Figure 1. Steps in KDD (Fayyad, et al., 1996)

B. Strengths and Weaknesses

Data mining helps in identifying the hidden patterns and obtain knowledge from a given dataset by following the steps of KDD. This knowledge can then be used to understand the domain of interest, generate strategies, and make smart and informed decisions. It also helps in making predictions and identify suggestions to a given problem. Using data mining in educational domain will help institutions to understand the needs and requirements of their students as well as to understand what factors exactly inhibits to affect the dropout rates from their student demographics. It helps in finding the knowledge required for institutions to take precautionary measures and avoid the dropout issues as an entirety.

Data mining will be beneficial for institutions from various perspectives. The underlying issues with data mining are that during the process of knowledge extraction adds on operating costs as well as hardware and software resources along with manpower which can be bothersome. This process requires certified and skilled professionals which may not be affordable for many companies. The pre-processing stage requires most of the time for analysts to be spent on and having severely unclean data may cause analysts to spend the time unnecessarily.

Therefore, having an appropriate dataset is essential. The algorithms are required to be supervised by analysts due to at times, it may produce unreliable results. Same as before, there are violation concerns with privacy of the individuals' data that is used for the analysis. For which, establishments must take it into account and enforce tight security to safeguard the data (personal information). This entire process makes the lives of businesses as well as consumers

easier however, comes with consequences of protecting the user's privacy.

IV RESULT

In this section, the performance on a series of tests on the Student Dropout Prediction System is shown in order to evaluate and validate the outputs that it produces. The process of testing is performed to verify whether the system meets the previously determined requirements. Starting from the dataset to the models build itself had been tested along with the application which was built in order to serve the end users. The core purpose of testing was to validate the modules and components within the system whether it works as it was ought to. The series of tests included: Usability, Functionality, Reliability, Performance, and Security Testing.

Scikit-learn also known as Sklearn, is an open source Python library consisting of numerous tools for data analysis as well as predictive analysis and machine learning. It is built on the packages known as NumPy (numerical Python), SciPy (scientific Python), and matplotlib (Mat Lab plot). The Metrics module in the Sklearn package is used to perform assessments and also to check the quality of the predictive model by the help of various methods from it. The Accuracy Score, Confusion Matrix and Classification Report are imported as shown in Figure 2.

```
from sklearn.metrics import confusion_matrix, accuracy_score
```

Figure 2. Python Library

The accuracy_score function calculates the accuracy of the model by comparing the number of correct predictions in accordance with the total elements which are the test subset/validation dataset in this case, y_test. Therefore, the formula 1 is as follows:

$$\text{Accuracy Score} = \frac{\text{Number of elements correctly predicted(y\_pred)}}{\text{Elements(y\_test)}} \quad (1)$$

Since this dataset contain in equal number of positive and negative labels, the author had implemented this validation technique. Yet another function which is similar to K fold cross validation however, it takes in a proper combination of the target labels. It performs selection of the defined number of splits to be done in integers and allocates the feature, target labels accordingly. Then after fitting the model, predictions are performed but, most importantly, different, and equal portions of class labels from the dataset are selected.

Here as shown in Figure 3, the cross-validation score function had been imported from the model selection module of Sklearn. What this function does is that it takes in the model as estimator the entire dataset split into the features and target respectively and lastly the number of times it will be repeated. This method takes in different train and test sets every time it repeats that is, if there are 10,000 records in a dataset, the function will take 1,000 sets at random then perform the predictions. These predictions and model accuracy scores are then kept into account and repeated for the n-number of times determined by the user.

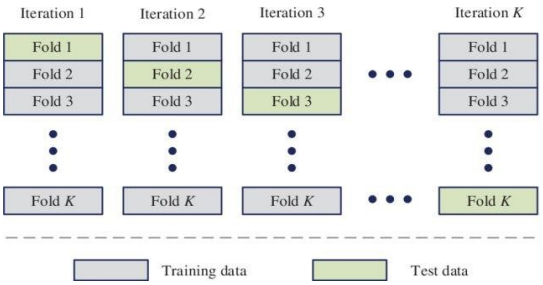

Figure 3. Illustration of K Fold CV, (Li, 2019)

The comparing of the models using the aforementioned accuracy scores along with other parameters such as precision, recall, f1-score, support macro and weighted average based on each target i.e. the correctness and performance of each model for model selection in order to deploy are applied as shown in Figure 4.

```
from sklearn.model_selection import StratifiedKFold
```

Experiment 0	Test	Train	Train	Train	Train	Classification Accuracy 0
Experiment 1	Train	Test	Train	Train	Train	Classification Accuracy 1
Experiment 2	Train	Train	Test	Train	Train	Classification Accuracy 2
Experiment 3	Train	Train	Train	Test	Train	Classification Accuracy 3
Experiment 4	Train	Train	Train	Train	Test	Classification Accuracy 4

Figure 4. Demonstration of Stratified K Fold CV (Automatic Addison, 2019)

It was found that the best performing models were Random Forest Classifier, followed by the Decision Tree Classifier. The performance from both these models were quite high compared to others being over 97% however, another model was to be selected as both of these are considered to be having similar characteristics.

Formerly, carefully compared the performance from the remaining 3 models, when comparing the accuracy scores, Logistics Regression algorithm had shown to be providing the highest accuracy of 88.57%, followed by Support Vector Machine (88.24%) and Naïve Bayes (86.84%) which was clear enough to be deduced with the lowest accuracy as shown in Table 5. The compared techniques is towards the sensitivities of the remaining two

models; LR[1] and SVM[2]. Although both these models showcased a high accuracy, the precision score obtained in terms of predicting the dropout of a student were considerable amounts of 59% and 51% respectively. As for the Recall and F1-score, SVM had taken a lead though in a minor context however, the support counts of the LR model was higher than that of SVM. Assessing Figure 3, it was clear to select the LR model as it had a higher accuracy score, time taken to train, and support count.

Table 5. Overall Results of the Models

Model Name	Random Forest	Decision Tree	LR	SVM	Naïve Bayes
Accuracy Score	97.92 %	97.29 %	88.57 %	88.24%	86.65%
TP	829	870	229	380	634
TN	7375	7281	7122	7013	6642
FP	33	112	208	372	703
FN	141	115	749	613	399
Precision					
Did Not Drop Out	0.98	0.98	0.90	0.92	0.94
Dropped Out	0.89	0.89	0.59	0.51	0.47
Recall					
Did Not Drop Out	0.98	1.00	0.97	0.95	0.90
Dropped Out	0.88	0.85	0.29	0.38	0.61
F1-Score					
Did Not Drop Out	0.98	0.99	0.94	0.93	0.92
Dropped Out	0.88	0.91	0.38	0.44	0.54
Support					
Did Not Drop Out	7393	7408	7330	7385	7345
Dropped Out	985	970	1048	993	1033

V CONCLUSION

In conclusion, the conducted tests on the predictive models along with the system using various techniques. The testing included checking the metrics of the selected models whether they were able to perform as tested using the training dataset as compared to a newly record as per the user's requirement. All the test cases were met according to their expectations.

ACKNOWLEDGMENT

This work is acknowledged by INTI International University (IIU), Malaysia for financial support.

REFERENCES

Anderson, N. (2020, May 21). Virus crisis slams college admissions: Some schools extend deadlines for students to accept offers. Retrieved June 10, 2020, from https://www.washingtonpost.com/local/education/virus-crisis-slams-college-admissions-some-schools-extend-deadlines-for-students-to-accept-offers/2020/03/20/bce58a92-6927-11ea-b313-df458622c2cc_story.html

Automatic Addison. (2019, July 9). Five-Fold Stratified Cross-Validation. Retrieved October 10, 2020, from https://automaticaddison.com/five-fold-stratified-cross-validation/

Balraj, E. & Maalini, D.. (2018). A survey on predicting student dropout analysis using data mining algorithms. International Journal of Pure and Applied Mathematics. 118. 621-626.

Brownlee, J. (2019, August 12). Supervised and Unsupervised Machine Learning Algorithms. Retrieved July 5, 2020, from https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/#:~:text=Supervised%3A%20All%20data%20is%20labeled,structure%20from%20the%20input%20data.

Brunsden, V., Davies, M., Shevlin, M., & Bracken, M. (2000). Why do HE students drop out? A test of Tinto's model. Journal of further and Higher Education, 24(3), 301-310. https://doi.org/10.1080/030987700750022244

Burke, A. (2019, May 16). Student Retention Models in Higher Education: A Literature Review. Retrieved June 1, 2020, from https://www.aacrao.org/research-publications/quarterly-journals/college-university-journal/article/c-u-vol.-94-no.-2-spring/student-retention-models-in-higher-education-a-literature-review. ISSN: ISSN-0010-0889

Dickler, J. (2020, May 12). Colleges acceptance rates may go higher as schools start aggressively courting applicants. Retrieved June 10, 2020, from https://www.cnbc.com/2020/05/12/college-acceptance-rates-rise-nearly-across-the-board-amid-coronavirus.html

Durso, S., & Cunha, J. (2018). DETERMINANT FACTORS FOR UNDERGRADUATE STUDENT'S DROPOUT IN AN ACCOUNTING STUDIES DEPARTMENT OF A BRAZILIAN PUBLIC UNIVERSITY. Educação em Revista, 34(10), 1-25. http://dx.doi.org/10.1590/0102-4698186332

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Advances in Knowledge Discovery and Data Mining. In A. Press (Ed.), From Data Mining to Knowledge Discovery: An Overview (pp. 1-34). CA: The MIT Press. https://doi.org/10.1145/240455.240464

Guerra, L. C., & Coates, K. S. (2019, November 4). What universities can do to keep students from dropping out. Retrieved July 2, 2020, from https://theconversation.com/what-universities-can-do-to-keep-students-from-dropping-out-123505

Hess, A. (2020, April 29). Some students are considering dropping out of college because of coronavirus. Retrieved June 10, 2020, from https://www.cnbc.com/2020/04/28/students-are-dropping-out-of-college-because-of-coronavirus.html

Lackey, D. (2019, March 28). How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. Retrieved July 4, 2020, from https://blazon.online/data-marketing/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#:~:text=There%20are2.5%20quintillion%20bytes,This%20is%20worth%20re%2Dreading!

Li, M. (2019, Feb). Tectonic discrimination of olivine in basalt using data mining techniques based on major elements: a comparative study from multiple perspectives. Big Earth Data, 10(1), 3. https://doi.org/10.1080/20964471.2019.1572452

PBS NewsHour. (2015, August 17). Why poor students drop out even when financial aid covers the cost. Retrieved July 2, 2020, from https://www.youtube.com/watch?v=1Sst2RPsQM8

Rouse, M., Labbe, M., Martinek, L., & Stedman, C. (2019, September). Big Data Analytics. Retrieved July 4, 2020, from https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

The College Finance Team. (2020). The Class of Coronavirus. Retrieved June 10, 2020, from https://collegefinance.com/blog/the-class-of-coronavirus

Twin, A. (2019, August 18). Data Mining. Retrieved July 5, 2020, from https://www.investopedia.com/terms/d/datamining.asp

[1] Logistic Regression

[2] Support Vector Machine

UNESCO. (2020, March 15). Education: From disruption to recovery. Retrieved June 10, 2020, from UNESCO Building peace in the minds of men and women: https://en.unesco.org/covid19/educationresponse

World Bank Group. (2015). Education Global Practice - Smarter Education Systems for Brighter Futures. Driving Development with Tertiary Education Reforms. Retrieved July 2, 2020, from http://documents1.worldbank.org/curated/en/613701468188661472/pdf/98454-REVISED-Box393212B-PUBLIC.pdf

Yaacob, W. W., Sobri, N. M., Nasir, S. M., Norshahidi, D. N., & Husin, W. W. (2020). Predicting Student Drop-Out in Higher Institution Using Data Mining Techniques. Journal of Physics: Conference Series, 1496, 1-2. doi:10.1088/1742-6596/1496/1/012005

Yukselturk, E., Ozekes, S., & Kılıç Türel, Y. (2014). PREDICTING DROPOUT STUDENT: AN APPLICATION OF DATA MINING METHODS IN AN ONLINE EDUCATION PROGRAM. European Journal of Open, Distance and e-Learning , 17(1), 118-129. https://doi.org/10.2478/eurodl-2014-0008