Journal of Universal Computer Science, vol. 25, no. 10 (2019), 1301-1322
submitted: 15/12/18, accepted: 19/3/19, appeared: 28/10/19 © J.UCS

Decision-making Model at Higher Educational Institutions based on Machine Learning

Yuri Vanessa Nieto
(University of Oviedo, Oviedo, Spain
uo250052@uniovi.es)

Vicente García-Díaz
(University of Oviedo, Oviedo, Spain
garciavicente@uniovi.es)

Carlos Enrique Montenegro
(District University Francisco José de Caldas, Bogotá, Colombia
cemontenegrom@udistrital.edu.co)

Abstract: At Higher Educational Institutions (HEI) the high hierarchical managers and directors face many challenges during the decision-making process, that sometimes are rely on intuition, and past experiences, leading not just to delays but the low impact in the whole academic community. A decision-making model for managers and administrator of HEIs is presented. We propose a detailed methodology when academic prognosis is taking place. The comparison between five robust Machine Learning algorithms is executed accomplishing outperformed results by Support Vector Machine. As a validation experiment, we executed the proposed decision model in a face-to-face public university in Colombia, showing the results in a developed web platform prototype with its correspondent architecture. Moreover, we discuss the social implication of low graduation rates.

Keywords: Decision Support System, Decision-making, Machine Learning, Support Vector Machine, Classification algorithms
Categories: I.2, I.5, L.2, L.3

1 Introduction

The primary mission of a Higher Educational Institution is to educate highly competent professionals that support the advancement of the region where they developed. Hence, the cohort graduation rate has been a worldwide accepted indicator of the students and university success [Guilbault 2017]. Education plays an important role to prevail a society, therefore when a student graduate the whole society is benefited.

Deciding the high hierarchical management position of an HEI has significant implications in the whole academic community. Policymakers, managers, and subordinates of HEIs are disengaged from students when deciding [La, Bajzíková and Dedze 2017]. Since managers and directors' decisions impact student success and most of their decisions rely on past experiences or intuition [Stefanova and Kabakchieva 2018], our research goal is to support this complex task through the development and test of a decision-making model that includes the latest and high reliable computational algorithms to this aim.

Machine Learning (ML) has been a growing trend when analysis information. Its popularity is due to its high accuracy reached, lowest processing time and the range of available algorithms that fit either classification or regression problems. Although many related works have been conducted using ML to identify and classify students, they collect information about distance education, and their prediction is reduced to single subjects or courses. Moreover, they lack the implementation and actions taken about their results probably, because they are focussed on students and teachers as their primary stakeholders.

Firstly, we propose a methodology for data-driven academic prognosis that identifies the steps required when aiming taking action over data. Moreover, in order to include the outperformed algorithm in our decision-making model, we compared Support Vector Machine, Decision Trees, Random Forest, Artificial Neural Network and Logistic Regression. These five classifiers are popular in related literature due to its high accuracy reached and efficiency as classification techniques in the educational field.

In addition, we considered HEIs' managers' academic concerns that were identified in our previous work, as long as their information visualization needs, with the aim of providing a decision-making model that support strategic decisions at HEIs, based on predictions made on graduation rates by the machine learning algorithm presented. The insightful information provided by this model leads to various potential actions. For instance is useful to i) increase students retention, ii) prioritize intervention efforts, iii) create strategies to mitigate early failure and strategic plans through the knowledge of the futurity, iv) diminish students dropout, v) increase the HEI quality indicators among others.

This paper is structured as follows: Section 2 presents the methodology overview, the model for decision-making, the algorithm of the classification technique, and the data used in the experiment. Section 3 exhibits the results where we compare the Machine Learning algorithms used, the toolkit developed for decision making and the social implication of graduation rates. In Section 4 we revise related works. Finally, in Section 5 and 6 we conduct discussions, conclusion and future work.

2 Machine Learning Model to predict graduation of Students

2.1 Problem statement

Decisions made at the strategic level of Higher Educational Institutions (HEIs) are affected in face-to-face educational model mainly because of (a) the disengagement of the stakeholders, (b) difficulties in data acquisition, formatting and centralization and (c) the lack of using efficient computational algorithms to support these complex processes[Nieto, García-Díaz, Montenegro and Crespo 2018].

Therefore, managers and HEIs directors struggle during the decision-making process leading to rely their decisions randomly and not based in logical analysis [Saeed and Dixit 2015], or based on intuition or past experiences [Stefanova and Kabakchieva 2018], delays, limited observation of the “whole picture”, limited academic impact of the decision, among others.

2.2 Methodology overview

The model proposed in this paper is focused on support academic decision-making through the prognosis of the number of graduate students, which will enable high level authorities to take action over data through the obtention of a greater understanding when making a strategic decision regarding to dropout, students' retention, resource planning, curriculum design and teacher's management. Moreover, this approach reduces processing time, increases reliability on predictions and promotes hypothesis formulation and hidden patterns understand.

The overall methodology showed in Figure 1. used in this study is built upon the standard process cycle of data mining and data-driven approach devising a set of steps along a workflow.

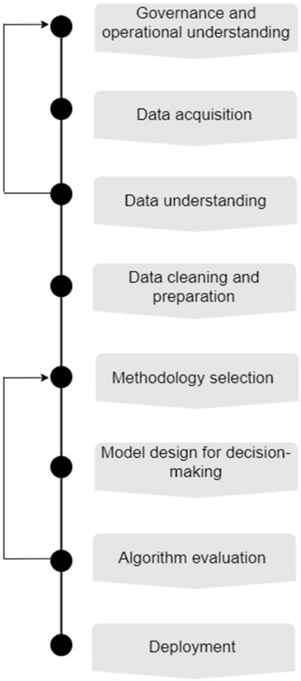

Governance and operational understanding
Data acquisition
Data understanding
Data cleaning and preparation
Methodology selection
Model design for decision-making
Algorithm evaluation
Deployment

Figure 1: Methodology for data-driven academic prognosis

(1) *Governance and operational understanding:* The objectives of the academic prediction should be set in order to identify the stakeholder needs and the possible algorithms to execute. The environment where the HEIs is developed should be evaluated to understand how extern policies affect their decisions. Their current academic situation along with their strategies, processes, people in charge, times and sources should be examined.

(2) *Data acquisition*: Data sources supply the information needed in any analytical purpose. Several HEIs with a face-to-face educational model finds the data accessing

a handicap for decision-making. Data should be harvest, integrated and storage in a coherent database, although it might be in different formats or store in various silos.

(3) *Data understanding:* According to the HEI' educational model and processes operation, data is organized, formatted and stored in different manners. Data should be review and describe within a context. The prognosis relies on the data quantity and quality.

(4) *Data cleaning and preparation:* Data should be selected, cleaned and formatted towards the feature selection. Important characteristics are extracted from the data. The infrastructure that will store and manage the data should be set.

(5) *Methodology selection*: The Machine Learning algorithm is selected. This step is the core functionality that learns from past data and generalizes into the future [Stecto, Dinmohammadi, Zhao and Robu 2019]. Following the main goal of the prediction, if the output variable is categorical, a classification algorithm must be selected. Otherwise, a regression algorithm should be used.

(6) *Model design for decision-making:* Apply learning algorithms to the model data. Depend on the algorithm selected in step above, extra steps as data normalization should be carried out. In this step the algorithm is trained on selected data, tested and validated.

(7) *Algorithm Evaluation*: Is highly recommended to perform algorithms comparison in order to choose the most suitable according to the objective goals. The validity of the model can be estimated through different metrics such as ROC, accuracy, F-measure, recall, and precision.

(8) *Deployment:* Human-Computer Interaction techniques should be considered in order to present the information to managers and directors in a friendly and easily understandable way. Prediction reports should be available on an online toolkit as defined in users' profile. The main results from this step are reliable, supported academic decisions, inducement to hypothesis formulations and a greater insight into the HEI situation.

2.3 Classification technique

Support Vector Machine is a widely used classification technique [Liu, Wang, Wang, Lv and Konan 2017] that, given a set of objects belonging to one of two categories, constructs a hyperplane in a high dimensional space that separates those categories [Miguéis, Freitas, Garcia and Silva 2018]. For nonlinear problems, by using Kernel functions, also known as the Kernel trick, the data is embedded to a higher dimensional space, where it becomes linearly separable, which makes SVM more potent due to is not restricted to linear decision surfaces [Wittek 2014].

Traditional kernel functions include linear, polynomial, Gauss, Sigmoid and Fourier series. The choice of kernel abruptly alters the nature of the decision boundary.

Notably, Gauss Kernel (radial formulation) is stated given its simplicity, high efficiency, easy access and less computation [Wang, Huang and Cheng 2016] [Huang, Maier, Hornegger and Suykens 2017]. Therefore, such a kernel is conveniently selected.

Radial Basis Function (RBF) is a commonly used kernelized learning algorithm in SVM. It has not explicitly defined embedding, operating in an infinite dimensional space. It has the form:

$$\boldsymbol{K}(\boldsymbol{x_m}, \boldsymbol{y_m}) = \boldsymbol{exp}(-\boldsymbol{\gamma}\|\boldsymbol{x_m}, \boldsymbol{y_m}\|^2) \quad (1)$$

Where $\gamma>0$ is the parameter that controls the radius of the basis function, which serves to control the dispersion of the kernel in the input space. The kernel function $\boldsymbol{K}(\boldsymbol{x_m}, \boldsymbol{y_m})$ intends to measure the "similarity" between $\boldsymbol{x_m}$ and $\boldsymbol{y_m}$ (the larger, the more similar). In terms of the square Euclidean distance, we assume how close are those points to each other.

The binary classification process (i.e., graduated, and not graduated students) imply to fitting a hyperplane (decision boundary) to only two separable classes. Thus, the hyperplane becomes in a straight line separating two classes [Athani, Kodli, Banavasi and Hiremath 2018]. γ and C parameters play an essential role in this RBF classification vector.

In a preliminary training phase, the parameter γ has to be determined [Fischetti 2016]. If this parameter is too small, the model might be under fitted, by contrast, if it is too large the model it might be overfitting.

In order to avoid any of these two phenomena, we take a portion of the 70% of the dataset to training. If we used the whole data to complete the training, the chance that the parameters capture noisy is enhanced, leading to overfitting. We determine the best values for C and γ using 5-fold cross validation, in each k-fold run we evaluate the performance of the algorithm. We used the best performing parameter for creating the final model and testing.

Moreover, we use the C SVM. This type of Support Vector Machine trades off correct classification of training examples using a parameter C. With the aim of helping to improve the accuracy of the output; this parameter implements a penalty on the misclassification that is performed while separating the classes, working as a regularization parameter. Konstanz Information Miner platform (KNIME) was the tool used to execute the Machine Learning algorithm. The extension to SVM along with the cross-validation, aid to encounter the most suitable values for γ (i.e., γ= 0.1) and C (i.e., C=10).

Support Vector Machines is the classification technique choose due to its superior accuracy, compact and comprehensive resulting [Stoean and Stoean 2014], it provides a better decision boundary in any classification problem [Kaneda, Pei, Zhao and Liu 2014][Wang et al. 2016], its remarkable accuracy obtained using similar academic features[Costa, Fonseca, Santana, de Araújo and Rego 2017] [Ghatasheh 2015] and are less prone to overfitting than other models [Czibula, Gergely and Gaceanu 2014]

[Singh, Taylor, Rahman and Pradha 2018]. For more details of the parameters set, observe the algorithm in the next subsection.

2.4 Model approach for decision-making at HEIs

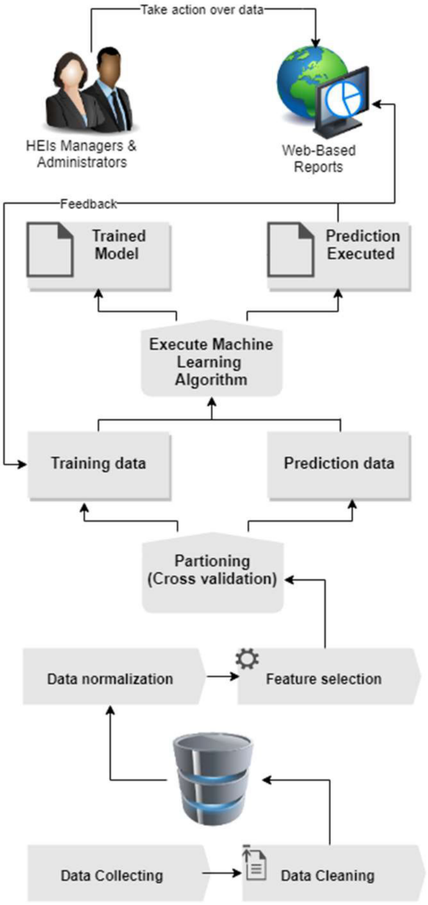


Figure 2: Model proposed for decision-making at HEIs

The model in Figure 2 presents the core goal of this study. The bottom-up approach depicts the integration of human and computer intelligence used within three main phases along with a HEIs' managers-centered objective function.

The first phase requires acquisition and data cleaning. A relatively large labeled data set is needed to create a reliable model [Ihalagedara, Kithuldeniya, Weerasekara and Deegalla 2015]. In academic prognosis, primary information arouses from systems of educational institutions[Tan and Shao 2015]. Educational models as distance and blended rely on more centralized information which differs with the struggles that face-to-face educational models have to face when gathering information.

The inputs listed in Table 1 comes from real data where our model is validated.

ID	Name	Description	Measurement
Inputs provided by the university administration			
1	Residential Stratum	Socio-economic status according to the students' residence address	{1,2,3,4,5,6}
2	Average of the taken subjects	Arithmetic mean of the final grade of all subjects coursed	Grade from 0.0 to 5.0
4	Subjects took in the semester	Subjects enrolled in each semester	Quantity
5	Subjects approved in the semester	Subjects satisfactory approved in each semester	Quantity
6	Subjects failed in the semester	Subjects failed in each semester	Quantity
7	Subjects validated in the semester	Subjects that were failed but after an extra summary exam become approved	Quantity
Inputs calculated from the data given to including in the study			
8	Median grade	Value of the grade separating the higher half from the lower half of the grades	Grade from 0.0 to 5.0
9	Maximum grade	Maximum value from all over final grades	Grade from 0.0 to 5.0
10	Minimum grade	Minimum value from all over final grades	Grade from 0.0 to 5.0
11	The range of the grades	Difference between maximum and the minimum grade	Quantity
12	First quartile of the grades	The middle grade between the smallest grade and the median of the data set	Quantity
13	Second quartile of the grades	Median grade	Quantity
14	Third quartile of the grades	The middle grade between the median and the highest grade of the data set	Quantity
15	The standard deviation of the grades	The square root of the average of the squared deviations of the values subtracted from their average value	Quantity
16	Accumulated subjects took in the whole career	Summary of the subjects enrolled during the whole career	Quantity
17	Accumulated subjects approved in the whole career	Summary of the satisfactory approved subjects during the whole career	Quantity
18	Accumulated subjects failed in the whole career	Summary of the failed approved subjects during the whole career	Quantity
19	Accumulated subjects validated in the whole career	Summary of the subjects that become approved after the validation exam during the whole career	Quantity

Table 1: Description and values of the student features used in the study

In the second phase, we constructed a Machine Learning algorithm to predict the number of students to graduate. Learning in a model translates into fitting a model's parameter to a specific dataset, interactively them updating with several passes through the data [Stecto et al. 2019]. From ML perspective, we use a supervised learning model that rely the prognosis on a dataset of fault events held in the past, from which a learning algorithm finds patterns. It correlates the captured data from the monitored asset to a target variable. [Diez-Olivan, Del Ser, Galar and Sierra 2019].

The main steps carry out to build the model are the following:

(1) Normalize data given them values from [0-1] with the transformation: $p^n = (p - p^{mean})./p^{std}$ where $p^{mean}$is the average of the input vectors in the data set, and $p^{std}$is the vector containing the standard deviations of each element of the input vectors.

(2) Set the accuracy desired (i.e., β>83%)

(3) Set initial settings of the Support Vector Machine (SVM)

  (3.1) Set type of SVM, i.e., C-SVC

  (3.2) Set type of kernel function, i.e., Radial Basis Function (RBF)

  $K(x_m, y_m) = exp(-\gamma\|x_m, y_m\|^2)$ where $\gamma>0$

  Settings of the Kernel:

  (a) Set gamma in kernel function, i.e., $\gamma=1.0$

  (b) Set C, i.e., C=10.0

(4) Split dataset into two subsets 70% for training and 30% for prediction using a stratified sampling technique to obtain homogenous groups

(5) Split the training subset from step (4) into 5-fold groups to later execute 5-fold cross-validation

(6) Compute the algorithm with initial settings in step (3) Search for $K(x_m, y_m)$ where (x_m, y_m) are Boolean variables according to the original problem formulation, ie., x=graduated, y=not graduated

(7) Store of the accuracy obtained in each k-fold, as well as the values of γ and C, entered in each run

(8) Repeat step (6) to find the best parameters of γ and C

  (8.1) Adjust settings in each run until the accuracy is satisfactory for the researches

(9) Compute the algorithm in step (6) to train the whole **training set** using the best values obtained for γ and C

(10) Compute $K(x_m, y_m)$ in step (5) on the **prediction set**

(11) If the accuracy desired is reached approved the last saved solution, if not such a threshold exists, reduce β

Is important to note that our work driver is Managers and Directors of HEIs concerns about graduation rates when making an academic decision [Nieto et al. 2018]. Therefore, instead of determining the exact student performance in a semester, we instead search to segment them, by classifying their academic performance at the time taking to conclude their degree into two groups *(x,y),* graduated or not graduate students respectively. Processing the students' graduates prediction using ML has managerial implications by enhancing reliability on results and reducing processing time.

We propose in phase three to expose prediction results in a web platform toolkit where HEIs deans and administrative managers can access the information in an easy to understand manner. We believe that the insightful information provided by this model opens an opportunity for different potential actions. For instace to a) create strategic plans through the knowledge of the futurity [Hu, Liu, Chen and Qin 2017] [Martínez et al. 2009], b) prioritize intervention efforts [Aguiar et al. 2015], c) diminish students dropout [Hamoud, Hashim and Awadh 2018], d) increase students retention, e) create strategies to mitigate early failure, f) better manage resources[Medina, García and Olguín 2018], g) increase the HEI quality indicators, among others.

2.5 Data

To illustrate our methodology in a real case study, we perform an experiment using data from an engineering faculty, belonging to a face-to-face educational model in a public University in Colombia. The engineering faculty has approximately 6000 students, 400 professors, whom more than 150 are researches, and offers five undergraduate programs and more than ten graduated programs (specialization, masters, and doctorate) in several fields such as Electrical Engineering and Cadastral Engineering.

The primary data set includes in this study refers to undergraduate students' information for those enrolled between 2004 to 2014, i.e., 12477. After disregarding cases of missing data, our final sample counts 6103 students from whom we have 55220 records. In the data cleaning phase, we consider that engineer programs last ten semesters. Thus, students enrolled after 2009 would not graduate before 2014. Therefore, their data was removed since they are not worthy to train a supervised algorithm. The description of the attributes used in this study is presented in Table 1.

While the model proved to be promising there are limitations that we come through during the data collection phase:

(1) The data provided by the institution used as a case study was very limited because of their data-protection policies. Therefore, relevance attributes such as socio-demographic data, socio-economic status, high school background, and enrollment process information are missing. Students' academic records are evaluated for the most part.

(2) Consequently, the feature selection process is limited by the short number of attributes. Thus, to cope with this limitation, we derivate data that come from the Oracle Database accessed by calculating extra attributes that are shown in the second part of Table 1.

(3) Despite the fact that the data structures were provided, the governance and operational understanding were necessary to process and clean data. For instance, grades of each subject were provided, but policies to approved or validate a subject were necessary to understand in order to classify students into the objective variable. Moreover, the changes on the syllabus in each career during the observation window represented a challenge to train the data set.

From the limitations stated above, a challenging question arises: Can one predict whether a student graduated or not graduated with these attributes? Classification learning algorithms used in our study address this problem, despite the limitations given by the institutions' data policy. In section 4 we observe the notorious accuracy that these algorithms achieve and how this data highlights the academic situation of a student.

2.6 Evaluation criteria

A 5-fold-cross-validation method was used to obtain error metrics. The data set was split in 5-fold at random, and using 4-folds to train the model, predicting the remaining fifth fold, considered as new data, and calculating the prediction error metrics. The process is repeated five times to predict each fold.

Moreover, as we propose in the methodology above, is highly recommended to perform algorithms evaluation and comparison to choose the one that fits and provides better results according to the objective goals using generalized methods.

Therefore, it is essential to establish the metrics to compare the algorithms. To this aim, in this experiment, we include the measure of F_1-score, due to if provides us the harmonic mean between precision and recall [Costa et al. 2017] described in equation 3 and 4 respectively.

$$F_1 score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (2)$$

$$Precision = \frac{TP}{FP + TP} \quad (3)$$

$$Recall = \frac{TP}{FN + TP} \quad (4)$$

Where TP (True Positives) is the number of positive instances correctly classified as positive. FP (False Positives) is the number of negative instances incorrectly classifies as positive. FN (False Negatives) is the number of positive instances incorrectly classified as negative.

Recall denotates the classifier performance concerning false negatives (the quantity we miss). On the other hand, precision uses the false positive the quantity we caught) to give us information about the classifier performance.

F_1-score also known as F-measure represents in a single score recall and precision. Although the harmonic mean behaves like an average when x is equal to y, when they are different F_1-score is closer to the smaller number as compared to the larger number, giving the algorithm and appropriate score rather than just an arithmetic mean.

3 Results

3.1 Comparing Machine Learning Algorithms

Our primary objective is to design a model for decision making at HEI supported on Machine Learning algorithms. First, to define the algorithm for our model, we conducted in previous works a comparison and analysis of the most used and relevant algorithms for classification in the educational field.

In order to get better predictability, we consider a total of five robust classifiers: Support Vector Machine (SVM), Artificial Neural Network (ANN), Decision Trees (DT), Random Forest (RF) and Logistic Regression (LR). Thus, using the effectiveness metrics defined, we identified and chose the algorithm that best predicts the number of graduated and not graduated students.

However, a previous step before comparing the algorithms is exposed to illustrate and support how we manage the handicaps presented during the experiment.

Since we have a reduced the attributes' quantity given by the case study, and we are addressing a face-to-face educational model where data is more limited in contrast to distance, and blended education were researchers have to clean data by removing, for instance, loggings or time expended in the platform, we tested if data given was enough to represent students' academic situation by running the algorithms mentioned above with a different quantity of data features.

The differences across the scenarios where the number of features (N) vary are shown in Figure 3.

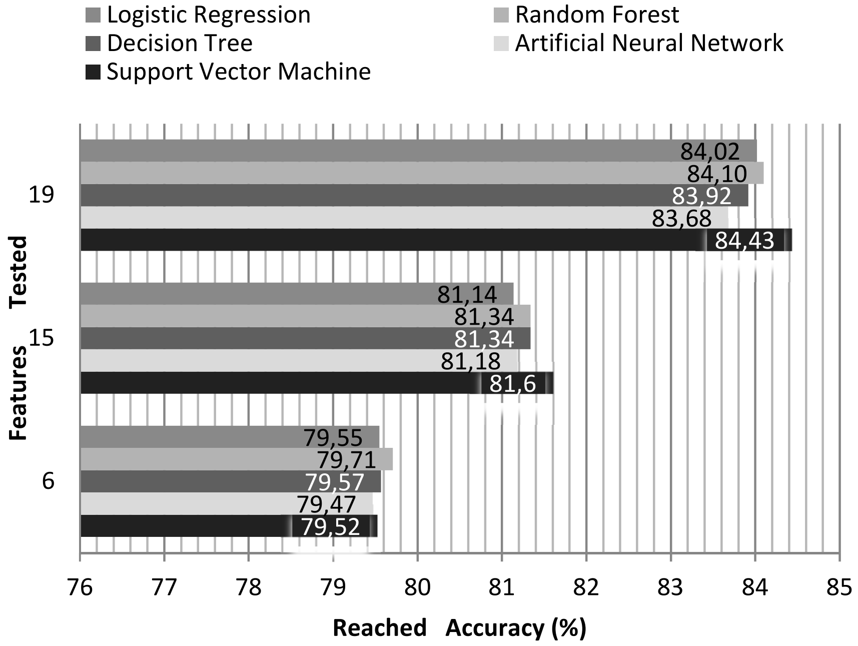


Figure 3: Accuracy reached by the algorithms with different quantity of features

The x-axis in Figure 3 indicates the percentual reached accuracy by the algorithms tested, while the y-axis indicates the three scenarios executed where the number of features varies. From the accuracy results of this experiment, we observe it was relevant to add more features to the dataset by calculating them. The accuracy increases as more features are integrated into the model. The accuracy enhancement is approximately 5% when using 19 features comparing when just 6 features are used, which regarding students correctly classified, it means more than 300 students, becoming a significant finding.

The results listed in Table 2 show the evaluation metrics. Although our evaluation criteria are based on the F-Score, we exposed the results given by metrics such as accuracy, precision, recall, Area Under the Curve (AUC), and Cohen Kappa.

	Support Vector Machine	**Artificial Neural Network**	**Decision Tree**	**Random Forest**	**Logistic Regression**
F-Score	0,8951569	0,8866559	0,889378	0,8910678	0,8894826
Precision	0,8599858	0,8598123	0,866170	0,8644938	0,8704747
Recall	0,9333276	0,9152296	0,913863	0,9193273	0,9093392
Accuracy	0,8454062	0,8345406	0,839249	0,8410600	0,8402148
Error	0,1545937	0,1654593	0,160750	0,1589399	0,1597851
AUC	0,8972	0,8986	0,883	0,8994	0,9028
Cohen Kappa	0,601	0,596	0,596	0,598	0,602

Table 2: Description and values of the student features used in the study

When comparing SVM against ANN, RF, LR, and DT, we found SVM provide slightly better results regarding F-score (values vary between 0.00577 and 0.00408). Every classifier was tested using the same data set and data features using the open analytic platform KNIME. SVM stands out with better most of the metrics. It shows his superior precise reaching the best F- Score (89.51%).

Feasibly, the reason for SVM attains the best performance is because classifies efficiently non-linear separable data when using the appropriate kernel function, is highly tolerant of overfitting and highly accurate (i.e., 89.72%). Similarly occurs with RF which is known to be a better classifier when multiple categorical variables are presented. The combination of these two algorithms could enhance the overall accuracy of the model.

In terms of Area Under the Curve (AUC) that represents the expected performance of the classifier LR achieve the best result (i.e., 0.9028) with 0.0056 points over SVM and Kappa value (i.e., 0.602). By contrast, Artificial Neural Network had the worse overall metrics with the lowest kappa value (i.e.0.596) and accuracy (i.e., 83.45%).

Since both of our binary outcomes (graduated and not graduated) are equally important for our model goal, the F-Score is the suitable metric to seek a balance between Precision and Recall. Thus, and for the outstanding results in the other metrics evaluated, Support Vector Machine is the Machine Learning algorithm used in our proposed model.

3.2 Toolkit for decision-making

The results of the prediction themselves are difficult to access for Higher Educational managers and directors because they would have to get the technical knowledge to search for the outcomes into the programs or platforms where the analysis is held. Besides, the results presented by these platforms in tables and few graphics are not easy to visualize, understand and manage.

Thus, we go beyond the model itself, presenting a web-based toolkit for managers and directors that enables to predict the number of graduated and not graduated students, exposing the results in a friendly and easy to understand manner. The main challenge is to allow the upload of new data and predictions and make it transparent to users. Although the machine learning algorithm is still executed in KNIME, in our prototype we propose a three-layer architecture to integrate the infrastructure as shown in Figure 4.

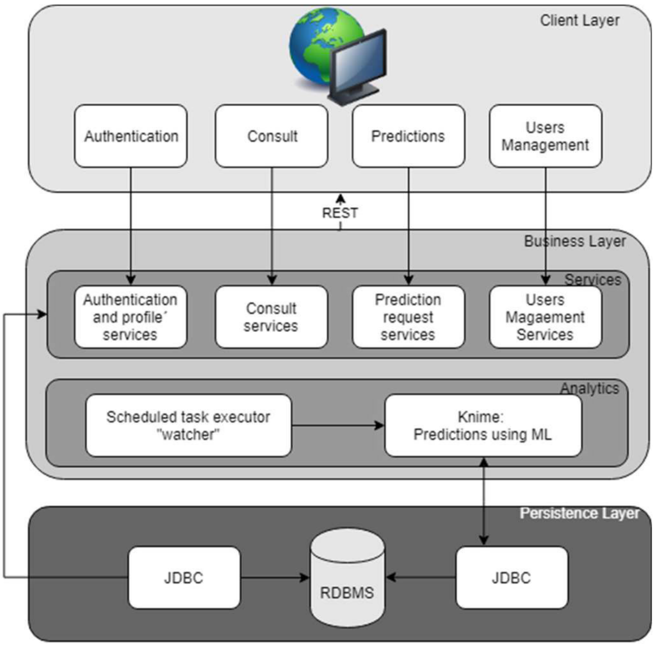


Figure 4: Overall architecture of the prototype

(1) Client Layer: After a secure logging, the web toolkit offers four functional modules: *Authentication* (manage profile settings and updating), *Consult* (allow the visualization of the requests done such as tables, filters, and graphics rendered from the business layer), *Predictions* (manage the prediction request and obtain the data) and *Users management* (Allows to create, update and delete users). The modules access to the business layer through REST scheme based on the HTTP protocol, which is part of the business layer. The development is done in java using the libraries React and Flux to create the user interface. The library d3 is used to visualize graphics,

(2) Business Layer: Comprises the business logic and is represented in two sub-layers: the services layer and the analytics layer.

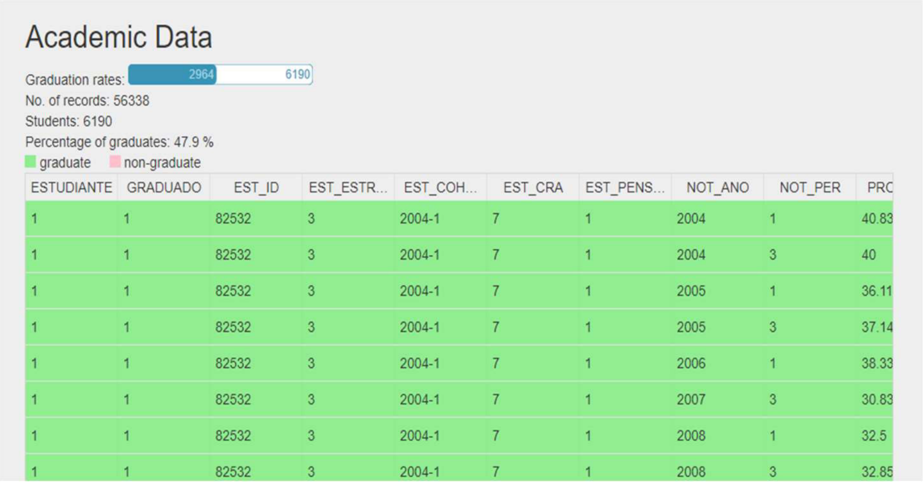
   a. Service sub-layer: As well as the client layer, is divided into modules that expose the four services according to its functionality. The Object Relational Mapping Hibernate applies the persistence interface of Java JPA. Liquibase is used as a migratory system to fill the base information in the persistence layer.

   b. Analytics sub-layer: The server application called “watcher” uses the spring-schedule extension to execute an instance task periodically from

Knime. This last contains the workflows to process the training and prediction functionalities. The communication between these two components is done with command lines that call on the Knime instance specifying the correspondent workflow to run.

(3) Persistence layer: It uses MySql as the Relational Data Base Management System (RDBMS). The data sources supply the information required in the toolkit. Users can upload new data in CSV format. A template is provided to ensure the data contains all the features required to execute the prediction.

With the aim of providing a friendly web platform to access the predictions done by or model, support the specific needs of the Higher Educational Institution that serves as a case study and, overcome the high costs of Knime server, we developed this prototype. Some of the results are shown in Figure 5 and 6.



Academic Data

Graduation rates: 2964 6190

No. of records: 56338

Students: 6190

Percentage of graduates: 47.9 %

🟩 graduate 🟥 non-graduate

ESTUDIANTE	GRADUADO	EST_ID	EST_ESTR...	EST_COH...	EST_CRA	EST_PENS...	NOT_ANO	NOT_PER	PRC
1	1	82532	3	2004-1	7	1	2004	1	40.83
1	1	82532	3	2004-1	7	1	2004	3	40
1	1	82532	3	2004-1	7	1	2005	1	36.11
1	1	82532	3	2004-1	7	1	2005	3	37.14
1	1	82532	3	2004-1	7	1	2006	1	38.33
1	1	82532	3	2004-1	7	1	2007	3	30.83
1	1	82532	3	2004-1	7	1	2008	1	32.5
1	1	82532	3	2004-1	7	1	2008	3	32.85

Figure 5: Consultation of academic information in the prototype

A consultation on the graduation rate is shown in Figure 5. Green color represents graduated students while red color represents not graduated students. The graduation rate is calculated in a 47.9% with more than 2940 graduated students.

Furthermore, between the various graphics the prototype plots, we consider a potential outcome from Figure 6. That was described by the District University engineer Dean and programs directors as “suitable towards retention decisions” because it is “a simple way to understand information that was unrevealed before.”

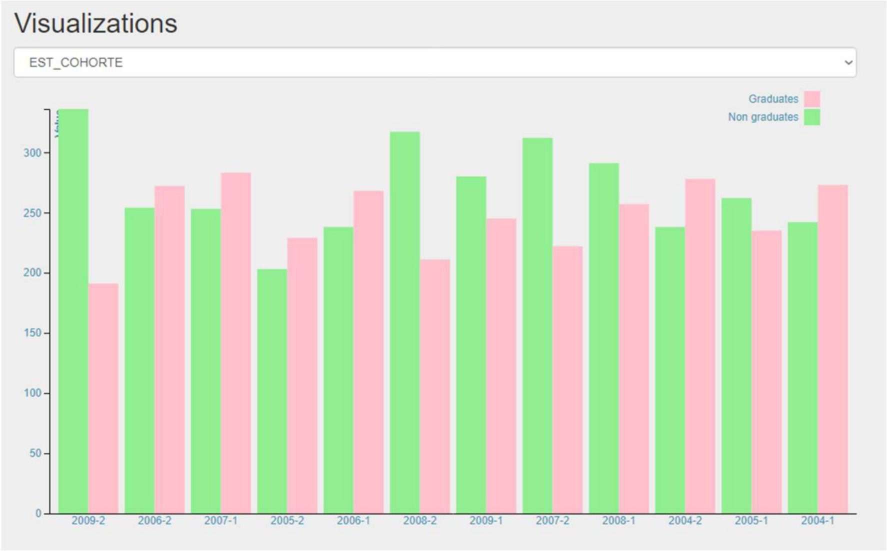


Figure 6: Graduated vs. Not graduated students by cohort

The bars in Figure 6 represent the graduate students (green bars) and not graduated students (red bars) in each cohort. The x-axis indicates the cohort, while the y-axis indicates the number of students. Although the green color prevails in most of the cohorts observed, the slight difference with the not graduated students should boost high hierarchical administrator of the HEIs to formulate hypothesis and take action on this concern.

From this point of view, the toolkit prototype shows the results of our proposed model and supports managers and directors given them insightful and highly reliable information. Thus, new opportunities emerge to achieve more favorable decisions.

3.3 Students graduation and social implications

District University Francisco José de Caldas is a public face-to-face Higher Educational Institution. We consider the implications of students' low rate graduations in our case study due to their remarkable implications in Colombian economy and society in general.

From high school graduate students, just 48% access to professional education and less than 10% have the opportunity to study in a public university. One student cost to the national state 3500 dollars semester, and accordingly to the World Bank and coherently to our study outcomes (i.e., 47.9%), in Colombia just 50% of students that access to professional education finish their studies and graduate [Granja 2017].

From a general society point of view, where all Colombian support with their taxes to the National General Budget, one can believe that the funds invested in Higher Education, half of it is not generating yields.

Education plays a vital role in a country development, and its reputation [Vo and Nguyen 2012], In Colombia low graduation rates affects a) the young people, due to the time expended with not tangible rewards and also have to accept underpaid jobs, b) working people that contribute from their salary to education without experimenting a national enhancement economy, c) industry sector that lack from graduates in some expertise areas, and invest less in scholarships due to the low trust factors, among numerous hidden implication that this phenomenon has in our society.

The cohort graduation rate is a crucial measure of student, and institutional success [Guilbault 2017] and our results expose alarming statistics from District University that are worth to mention.

4 Related work

Decision Support Systems (DSS) as a concept arose in the seventies as the systematization of the decision-making process. Since then, countless applications have been published related to it. Renowned decision analysis methods include influence diagrams, cost/benefits analysis, multi-attribute utility models, analytic hierarchy process, statistics, operation research, among other methods that work with dynamic, uncertain, and multi-criteria aspects of a decision.

In the educational domain, a further evolution link to the computational advances of the 2000s emerged by the name of Educational Data Mining. Recent work on DSS as the Mohamed [Abdahllah 2015] uses operation research to propose a Decision Support model that assist students for long course planning. His mathematical optimization balances students' preferences and advisors' recommendations while maintaining regulation policies.

Livereis [Livereis, Mikropoulos and Pintelas 2016] proposes a DSS that predicts students' performance concerning the final examination in Mathematics in a private secondary school. They utilized Naïve Bayes, Decision Trees, Sequential Minimal Optimization, Back Propagation, 3NN, and Ripper algorithm combine through the voting methodology. A decision support platform is presented as a prototype; however, methodology or a model itself is missing.

Several studies have been published in the literature that use Machine Learning algorithms to help place students within an organization [Thangavel, Bkaratki and Sankar 2017], predict students behavioural intentions , and mainly to identify and segment students that are likely to fail or dropout [Stimpson, Cummings and Member 2014] [Ghatasheh 2015] [Shanthini, Vinodhini and Chandrasekaran 2018] [Sandoval, Gonzalez, Alarcon, Pichara and Montenegro 2018].

The approach proposed at the University of Liege in Belgium [Hoffait and Schyns 2017] used Artificial Neural Network, Decision Trees and Random Forests to identify first-year students profiles with high failure risk. They propose to add an "uncertain class" to increase the accuracy of the predictions. Thus, their prediction results were distant from in just in 4% higher. However, they include many other data, besides academic sources. It is confirmed that some attributes such as socio-demographic data from students represent an enhancement on the prediction. Despite this, their contribution to Higher Educational Institutions are just the prediction numbers itself, due to not any action is proposed from these results. Probably because their focus is on students and their decisions related to this concern are limited.

The work in [Costa et al. 2017]follows a similar machine learning comparison. They compare Support Vector Machine, Decision Tree, Neural Network, Naïve Bayes, to predict students likely to fail in a specific introductory subject. Agreeing with our results, SVM outperforms the other algorithms significantly. However, it does not spcify data feature, and apart from the results, it does not propose any improvement because of the alarming statistics. Besides, it does not help during the decision-making process.

Even though, these studies have shown the high usability of Machine Learning algorithms in the educational field; they lack the implementation and actions taken about their results perhaps, because their stakeholders are mainly students and teachers. Therefore, their results conclude just with the statistics. Moreover, they extract information from the distance education arena, although their hold representative amount of data, their experiments are aimed at small courses or subjects.

Besides, from recent years publications review, and to our knowledge so far, we observed that Machine Learning had not been used to help decision-making process for high hierarchical managers of HEIs or even schools, neither there are focused on fulfilling managers or directors' requirements.

As discussed before, our novel contribution includes a step by step methodology to execute an academic prediction at HEIs. We exposed the features include in our model and specify the mathematical algorithm include in the decision- making model proposed. Additionally, we overcome the limitations of the University that serves a case study and present the results achieved in different scenarios. Finally, to exhibit the decision-making model scope, we developed a prototype to display results to managers and directors that help them to get a whole picture of the students' academic situation regarding graduation rates.

5 Discussions

In this study, we have used Support Vector Machine as the algorithm which our decision model relies on to make academic predictions. Despite its high accuracy and highlighted metrics results, this classifier can be inefficient and slow due to its computational difficulties as well as the model complexity with a more massive scale training set [Liu et al. 2017].

Time complexity in SVM is usually between $O(m^2n)$ and $O(m^3n)$ where m is the number of instances, and n is the number of features [Stecto et al. 2019], but none of the mentioned studies have used a data set more extensive than ours. Although, we believe that if a larger data set is analyzed, time processing can become in a handicap of the model without affecting its high reliability.

In the present study, we have overcome the lack of data attributes from the case study as our information provider. After various experiments, we found the algorithms reached a significant accuracy enhancement when more attributes were added by calculating them from the raw data.

The academic prediction was determined by the authors in previous research as our work driver after a survey conducted to Deans and higher management positions at HEIs where they stated their concern about graduation rates when making an academic decision.

During the experiments, we used the analytic platform that in the beginning suits our needs, but some different analytics platforms are also possible. Alternatively to Knime, platforms such as TensorFlow, RapidMiner, Alteryx among others allows through nodes or libraries access to Machine Learning tools. Since we have to face a high cost for Knime server and also have to fulfill specific needs from the case study HEI, we develop our toolkit that allows to run predictions and visualize results.

Data used in the case study and the workflows of the algorithms executed in Knime are available at https://github.com/vicegd/decision.making.higher.education.

6 Conclusions and Future work

The choice of the Machine learning algorithm depends on the problem to solve. Therefore, the methodology proposed is an essential outcome of the research due to specifies important hints when attempting academic prognosis, starting from governance and operational understanding to the deployment.

Additionally, this paper provides the comparison of five robust Machine Learning algorithms (SVM, ANN, DT, RF, LR) and their ability to recognize a balance segmentation of graduated and not graduated students. It was revealed that Support Vector Machine has the best recognition F- Score (i.e., 89,51%) and Accuracy (i.e., 84,54%) among them.

In the near future, a hybrid model proposal between Support Vector Machine and Random Forest would be worth attempting when seeking graduation rates due to its outperformed at overall prediction classification problem.

One of the main limitations of this study was data acquisition. Thus, it would be interesting to execute the same model proposed with larger data attributes that include socio-economic, and demographic data from students. For instance, students' gender, tuition fee, graduated high school GPA, monthly income, people in charge, and city where they came from, to mention few attributes that would be worth to analyze. When having more input parameters, the use of an optimization parameter such as gradient descendent [Jeon, Park and Lee 2018] might help to optimize our problem.

Performing the model proposed offers reliable predictions' results and insightful information about the academic situation of a HEIs. Additionally, the achieved outcomes along the model' execution, and the positive feedback received from the case study Dean and program directors about the prototype ensures us that is possible to support academic decision-making through our model efficiently. Nevertheless, a web-based toolkit that meets Human-Computer Interaction requirements would represent a significant enhancement of our prototype and the way the model exposes information.

This research represents an innovative model that supports high hierarchical administrators of HEIs during academic decision- making regarding graduation rates. We care about stakeholder that were not addressed whom significantly impact the whole institutional community when deciding. Furthermore, hypothesis formulation, and decision respecting retention rates, students' exclusion policies, students' dropout rates, and strengthen programs, are encouraging to be held from the decision- making model proposed.

References

[Abdahllah 2015] Abdahllah, M.: 'A decision support model for long-term course planning'; Decision Support Systems, Vol. 74 (2015), pp. 33–45. https://doi.org/http://dx.doi.org/10.1016/j.dss.2015.03.002

[Aguiar, Lakkaraju, Bhanpuri, Miller, Yuhas and Addison 2015] Aguiar, E., Lakkaraju, H., Bhanpuri, N., Miller, D., Yuhas, B., Addison, K. L.: 'Who, when, and Why: A Machine Learning Approach to Prioritizing Students at Risk of Not Graduating High School on Time'; In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK '15. ACM (2015), pp. 93–102. https://doi.org/10.1145/2723576.2723619

[Athani, Kodli, Banavasi and Hiremath 2018] Athani, S. S., Kodli, S. A., Banavasi, M. N., Hiremath, P. G. S.: 'Student performance predictor using multiclass support vector classification algorithm'; In IEEE International Conference on Signal Processing and Communication, ICSPC 2017 (2018), pp. 341–346. https://doi.org/10.1109/CSPC.2017.8305866

[Costa, Fonseca, Santana, de Araújo and Rego 2017] Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., Rego, J.: 'Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses'; Computers in Human Behavior, Vol. 73 (2017), pp. 247–256. https://doi.org/10.1016/j.chb.2017.01.047

[Czibula, Gergely and Gaceanu 2014] Czibula, G., Gergely, I., Gaceanu, R.: 'A support vector machine model for intelligent selection fo data representations'; Applied Soft Computing, Vol. 18 (2014), pp. 70–81. https://doi.org/http://dx.dor.org/10.1016/j.asoc.2014.01.026

[Diez-Olivan, Del Ser, Galar and Sierra 2019] Diez-Olivan, A., Del Ser, J., Galar, D., Sierra, B.: 'Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0'; Information Fusion, Vol. 50 (2019), pp. 92–111. https://doi.org/https://doi.org/10.1016/j.inffus.2018.10.005

[Fischetti 2016] Fischetti, M.: 'Fast training of Support Vector Machines with Gaussian kernel'; Discrete Optimization, Vol. 22 (2016), pp. 183–194. https://doi.org/10.1016/j.disopt.2015.03.002

[Ghatasheh 2015] Ghatasheh, N.: 'Knowledge Level Assessment in e-Learning Systems Using Machine Learning and User Activity Analysis';, Vol. 6, No. 4 (2015).

[Granja 2017] Granja, S.: 'Colombia mejora en acceso a la educación superior pero falta calidad'; El Tiempo. Bogotá (2017, June 5), pp. 1–5. Retrieved from https://www.eltiempo.com/vida/educacion/acceso-y-calidad-de-educacion-superior-en-colombia-segun-el-banco-mundial-95456

[Guilbault 2017] Guilbault, M.: 'Students as customers in higher education : The (controversial) debate needs to end'; Journal of Retailing and Consumer Services (2017), pp. 8–11. https://doi.org/http://dx.doi.org/10.1016/j.retconser.2017.03.006

[Hamoud, Hashim and Awadh 2018] Hamoud, A. K., Hashim, A. S., Awadh, W. A.: 'Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis'; International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 5, No. 2 (2018), p. 26. https://doi.org/10.9781/ijimai.2018.02.004

[Hoffait and Schyns 2017] Hoffait, A., Schyns, M.: 'Early detection of university students with potential difficulties'; Decision Support Systems (2017). https://doi.org/http://dx.doi.org/10.1016/j.dss.2017.05.003

[Hu, Liu, Chen and Qin 2017] Hu, J., Liu, H., Chen, Y., Qin, J.: 'Strategic planning and the stratification of Chinese higher education institutions'; International Journal of Educational Development, , No. 2016 (2017), pp. 1–8. https://doi.org/10.1016/j.ijedudev.2017.03.003

[Huang, Maier, Hornegger and Suykens 2017] Huang, X., Maier, A., Hornegger, J., Suykens, J.: 'Indefinite kernels in least squares support vector machines and principal component analysis'; Applied and Computational Harmonic Analysis, Vol. 43 (2017), pp. 162–172. https://doi.org/http://dx.doi.org/10.1016/j.acha.2016.09.001

[Ihalagedara, Kithuldeniya, Weerasekara and Deegalla 2015] Ihalagedara, K., Kithuldeniya, R., Weerasekara, S., Deegalla, S.: 'Feasibility of Using Machine Learning to Access Control in Squid Proxy Server'; (2015), pp. 491–494.

[Jeon, Park and Lee 2018] Jeon, Y., Park, Y., Lee, S.: 'Machine Learning Optimization of Parameters for Noise Estimation'; Journal of Universal Computer Science, Vol. 24, No. 9 (2018), pp. 1271–1281.

[Kaneda, Pei, Zhao and Liu 2014] Kaneda, Y., Pei, Y., Zhao, Q., Liu, Y.: 'Study on the effect of learning parameters on decision boundary making algorithm'; In EEE International Conference on Systems, Man and Cybernetics (2014), pp. 705–710. https://doi.org/10.1109/SMC.2014.6973992

[La, Bajzíková and Dedze 2017] La, A., Bajzíková, Ľ., Dedze, I.: 'Barriers and drivers of innovation in higher education : Case study-based evidence across ten European universities'; International Journal of Educational Development, Vol. 55, No. May (2017), pp. 69–79. https://doi.org/10.1016/j.ijedudev.2017.06.002

[Liu, Wang, Wang, Lv and Konan 2017] Liu, C., Wang, W., Wang, M., Lv, F., Konan, M.: 'An efficient instance selection algorithm to reconstruct training set for support vector machine'; Knowledge-Based Systems, Vol. 116 (2017), pp. 58–73. https://doi.org/10.1016/j.knosys.2016.10.031

[Livereis, Mikropoulos and Pintelas 2016] Livereis, I., Mikropoulos, T., Pintelas, P.: 'A Decision Support System for Predicting Student Performance'; Themes in Science and Technology Education, Vol. 9, No. 1 (2016), pp. 43–57. https://doi.org/10.15680/IJIRCCE.2014.0212015

[Martínez, Franco, Rodriguez, Crespo, G-Bustelo and Baena 2009] Martínez, O. S., Franco, E. T., Rodriguez, H. C., Crespo, R. G., G-Bustelo, B. C. P., Baena, L. R.: 'Viabilidad de la aplicación de Sistemas de Recomendación a entornos de e-learning'; In V Simposio Pluridisciplinar sobre Diseño y Evaluación de Contenidos (2009). Retrieved from http://www.web.upsa.es/spdece08/contribuciones/157_SPDECE.pdf

[Medina, García and Olguín 2018] Medina, M., García, C., Olguín, M.: 'Planning and Allocation of Digital Learning Objects with Augmented Reality to Higher Education Students According to the VARK Model'; International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 5, No. 2 (2018), p. 53. https://doi.org/10.9781/ijimai.2018.02.005

[Miguéis, Freitas, Garcia and Silva 2018] Miguéis, V. ., Freitas, A., Garcia, P. J. V, Silva, A.: 'Early segmentation of students according to their academic performance: A predictive modelling approach'; Decision Support Systems, Vol. 115 (2018), pp. 36–51. https://doi.org/https://doi.org.10.1016/j.dss.2018.09.001

[Nieto, García-Díaz, Montenegro and Crespo 2018] Nieto, Y., García-Díaz, V., Montenegro, C., Crespo, R. G.: 'Supporting academic decision making at higher educational institutions using machine learning-based algorithms'; Soft Computing (2018), pp. 1–9. https://doi.org/10.1007/s00500-018-3064-6

[Saeed and Dixit 2015] Saeed, F., Dixit, A.: ‘A decision support system approach for accreditation & quality assurance council at higher education institutions in Yemen’; In IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE). IEEE (2015), pp. 163–168. https://doi.org/10.1109/MITE.2015.7375308

[Sandoval, Gonzalez, Alarcon, Pichara and Montenegro 2018] Sandoval, A., Gonzalez, C., Alarcon, R., Pichara, K., Montenegro, M.: ‘Centralized student performance prediction in large courses based on low-cost variables in an institutional context’; The Internet in Higher Education, Vol. 37 (2018), pp. 76–89. https://doi.org/https://doi.org/10.1016/j.iheduc.2018.02.002

[Shanthini, Vinodhini and Chandrasekaran 2018] Shanthini, A., Vinodhini, G., Chandrasekaran, R. .: ‘Predicting students’ academic preformance in the university using meta decision tree classifiers’; Journal of Computer Science, Vol. 14, No. 5 (2018), pp. 654–662. https://doi.org/https://doi.org/10.3844/jcssp.2018.654.662

[Singh, Taylor, Rahman and Pradha 2018] Singh, S., Taylor, R., Rahman, M., Pradha, B.: ‘Developing robust arsenic awareness prediciton models using machine learning algorithms’; Journal of Enviromental Management, Vol. 211 (2018), pp. 125–137. https://doi.org/10.1016/j.jenvman.2018.01.044

[Stecto, Dinmohammadi, Zhao and Robu 2019] Stecto, A., Dinmohammadi, F., Zhao, X., Robu, V.: ‘Machine learning methods for wind turbine condition monitoring: A review’; Renewable Energy, Vol. 133 (2019), pp. 620–635. https://doi.org/https://doi.org/10.1016/j.renene.2018.10.047

[Stefanova and Kabakchieva 2018] Stefanova, K., Kabakchieva, D.: ‘Educational data mining perspectives within university big data environment’; In 2017 International Conference on Engineering, Technology and Innovation: Engineering, Technology and Innovation Management Beyond 2020: New Challenges, New Approaches, ICE/ITMC 2017. IEEE (2018), pp. 264–270. https://doi.org/10.1109/ICE.2017.8279898

[Stimpson, Cummings and Member 2014] Stimpson, A. J., Cummings, M. L., Member, S.: ‘Assessing Intervention Timing in Computer - Based Education Using Machine Learning Algorithms’; IEEE Access, Vol. 2 (2014), pp. 78–87. https://doi.org/10.1109/ACCESS.2014.2303071

[Stoean and Stoean 2014] Stoean, C., Stoean, R.: ‘Post-evolution of variable-length class prototypes to unlock decision majing within support vector machines’; Applied Soft Computing, No. 25 (2014), pp. 159–173. https://doi.org/http://dx.doi.org/10.1016/j.asoc.2014.09.017

[Tan and Shao 2015] Tan, M., Shao, P.: ‘Prediction of student dropout in E-learning program through the use of machine learning method’; International Journal of Emerging Technologies in Learning, Vol. 10, No. 1 (2015), pp. 11–17. https://doi.org/10.3991/ijet.v10i1.4189

[Thangavel, Bkaratki and Sankar 2017] Thangavel, S. K., Bkaratki, P. D., Sankar, A.: ‘Student placement analyzer: A recommendation system using machine learning’; In 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS). Colmbatore (2017), pp. 1–5. https://doi.org/10.1109/ICACCS.2017.801463

[Vo and Nguyen 2012] Vo, T. N. C., Nguyen, H. P.: ‘A Knowledge-Driven Educational Decision Support System’; In 2012 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future. IEEE (2012), pp. 1–6. https://doi.org/10.1109/rivf.2012.6169819

[Wang, Huang and Cheng 2016] Wang, X., Huang, F., Cheng, Y.: ‘Computational performance optimization of support vector machine based on support vectors’; Neurocomputing, Vol. 211 (2016), pp. 66–71. https://doi.org/10.1016/j.neucom.2016.04.059

[Wittek 2014] Wittek, P.: ‘Quantum Machine Learning’; Quantum Machine Learning (1st ed.). Elsevier (2014). https://doi.org/https://doi.org/10.1016/C2013-0-19170-2