Data Mining for Engineering Schools

Predicting Students' Performance and Enrollment in Masters Programs

Chady El Moucary
Department of Electrical, Computer and Communication Engineering, Faculty of Engineering
Notre Dame University – Louaize (NDU)
North Lebanon Campus – P.O. Box 87, Tripoli – Municipality Street, Barsa – El Koura, Lebanon

***Abstract*— the supervision of the academic performance of engineering students is vital during an early stage of their curricula. Indeed, their grades in specific core/major courses as well as their cumulative General Point Average (GPA) are decisive when pertaining to their ability/condition to pursue Masters' studies or graduate from a five-year Bachelor-of-Engineering program. Furthermore, these compelling strict requirements not only significantly affect the attrition rates in engineering studies (on top of probation and suspension) but also decide of grant management, developing courseware, and scheduling of programs. In this paper, we present a study that has a twofold objective. First, it attempts at correlating the aforementioned issues with the engineering students' performance in some key courses taken at early stages of their curricula, then, a predictive model is presented and refined in order to endow advisors and administrators with a powerful decision-making tool when tackling such highly important issues. Matlab Neural Networks Pattern Recognition tool as well as Classification and Regression Trees (CART) are fully deployed with important cross validation and testing. Simulation and prediction results demonstrated a high level of accuracy and offered efficient analysis and information pertinent to the management of engineering schools and programs in the frame of the aforementioned perspective.**

Keywords-component; Educational Data Mining; Classification and Regression Trees (CART); Relieff tool; Neural Networks; Prediction; Engineering Students' Performance; Engineering Students' Enrollment in Masters' Studies.

I. INTRODUCTION

Data mining has attracted exceptionally diversified businesses for both the descriptive and predictive capabilities it promises, one of which is Education in its broad fields and organizational hierarchies [15] [36] [50]. In fact, Education, nowadays, not only involves the *ancestral* information- and/or knowledge- communication and transfer, but has also become a standalone and comprehensive *business* with excessive demands in information handling and analysis, as well as the management of a deeply spread tree of positions and interrelated functions [49]. Indeed, Education's features and attributes have dramatically shifted and augmented to a point where the integration of technology became inevitable in the attempt of sustaining a good position and thriving in a highly competitive and merciless market. This technology not only involves new teaching methodologies but also osculates with every single aspect of management of such institutions.

Furthermore, management of large amount of data has become undeniably forbearing to even expert staff; it even requires more powerful computational and specifications requirements when referring to machines and/or algorithms. Diversified challenges face Education and which fortunately attracted researchers from different fields of expertise who keep straining in order to achieve innovative but also *intelligent* techniques to help keep up with the pressure and find astute and reasonable answers to multifaceted questions. Globalization, International Accreditation, and e-learning have only added more threads to the pile.

Data mining, which is the science of digging into databases for information and knowledge retrieval, has recently developed new axes of applications and engendered an emerging discipline, called Educational Data Mining or EDM. This discipline seems to be a lot promising. EDM carries out tasks such as prediction (classification, regression, and density estimation), clustering, relationship mining (association, correlation, sequential mining, and causal data mining), distillation of data for human judgment, and discovery with models [1]. Moreover, by exercising EDM, educators and administrators can tackle both traditional and ad hoc educational issues and benefit from a good decision-making tool when facing challenges and/or exploring new horizons in their specialties. The list is long and requires wide and long tables to fit in. Nevertheless, in a non-exhaustive list, we can enumerate the most frequently inquiring subject matters such as predicting students' performance, developing courseware, students' behavioral modeling, strategic planning and scheduling of programs, supervising attrition rates, and grant management, etc. In other words, EDM aims at enhancing the understanding and supervision of learners', teachers', and administrators' domain representation, pedagogical engagement and behaviors [5] [18] [32] [37] [38] [39].

Remarkable amount of EDM endeavors have been conducted and published in many journals and conference proceedings related to, but not limited to, Artificial Intelligence, Learning Systems, Education, and others. In July 2011 the International Educational Data Mining Society [2] was founded by the International Working Group on Educational Data Mining with the main objective of capturing contributions from the EDM community and offering a forum for practitioners to impart their labor and exchange their competencies. It has so far organized four international conferences where recognized work can be archived in the

Journal of Educational Data Mining or JEDM (*ISSN 2157-2100*) [28].

One can find many definitions for data mining in books, journal papers, and e-articles [11] [12] [13] [14]. They all refer to data mining as a young and interdisciplinary field in computer science which is described as an *interactive* and *iterative* process aiming at *sundering out/revealing hidden/unobvious*, but *existing*, *patterns*, *trends* and/or *relationships* amidst data using statistical and mathematical procedures with a prime objective of providing decision support systems with information and knowledge. Furthermore, data mining is being recently interchangeably used with what is referred to as Knowledge Discovery in Database or KDD namely when excessively large data repositories is being involved. Fig. 1 shows Data Mining exercise as a step towards KDD [10].

A typical example of inexorable flood of data is the Europe's Very Long Baseline Interferometry (VLBI), which has 16 telescopes, each of which produces one Gigabit/second of astronomical data over a 25-day observation session [6]. An interesting overview of the largest databases in the world can be found in [7] [8]. The top-ten lists include The Library of Congress (LC) with over 130 million items, 530 miles of shelves, 5 million digital documents, and 20 terabytes of text data. The Central Intelligence Agency (CIA) possesses comprehensive statistics on more than 250 countries and entities (unknown number of classified information). Amazon, the world's biggest retail store, maintains over 59 million active customers ending up with over 42 terabytes of data. YouTube, the largest video library, observes more than 65,000 videos added each day and encompasses at least 45 terabytes of videos. ChoicePoint, the business of acquiring information about the American population (addresses, phone numbers, driving records, etc.) possesses a database that extends to the moon and back 77 times and holds over 250 terabytes of personal data. Sprint, one of the world's largest telecommunication companies, offers its services to more than 53 million subscribers with a 2.85 trillion database rows and 70,000 call detail-record insertions per second. Google, the famous search engine and industry, is subjected to 91 million searches per day (accounting to 50% of all internet search activity) and holds more than 33 trillion database entries, etc. It is spectacularly evident that traditional warehousing techniques and querying algorithms cannot cope with such colossal amount of data, thus, new techniques for information retrieval urged tons of research papers in the data mining/KDD field. Many strains have been deployed to manipulate and handle considerable amount of data [20] [35] [47] [48], but in many applications, the data available only covers parts of the inference chain from evidence to actions. However, a versed miner can extrapolate a small amount of initial knowledge into more knowledge using proficient mining.

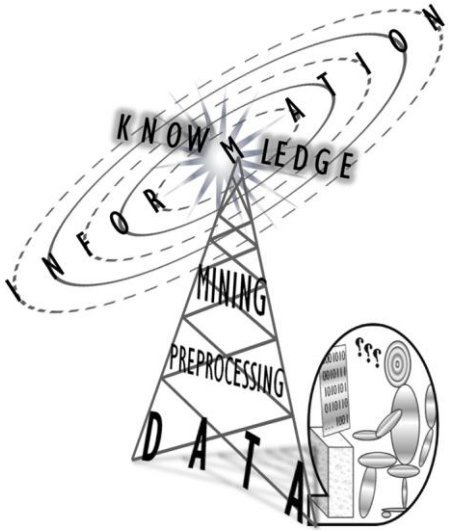

Figure 1 - Data Mining, a Mandatory Step towards KDD

In this paper, we will present a study that aims at offering a reliable and predictive tool for academicians and administrators working in engineering schools and universities to monitor students' performance at an early stage of their educational path. The goal is to link this observation/data with students' chances to either finish (succeed) a five-year Bachelor-of-Engineering program (BE) or enroll in Masters' program in a BS/MS track.

In the section to come, Data Mining will be reviewed and presented from different perspectives with emphasis on its various categories, tasks and implementations. In Section III, we will elaborate on the tool developed and underline data preparation and attributes' selection. Furthermore, the use of Neural Networks and Classification and Regression Trees (CART) will be explained and applied with cross-validation and pruning [40]. Error Histogram and ROC curves will also be studied in section III. Finally, section IV will portray a thorough analysis and discussions of the results. The core objective of the paper will be summarized and a conclusion is presented in this section as well.

II. DATA MINING

A. A Multifaceted Discipline

Data mining is a twofold discipline in the sense that it subtends two high-level primary objectives: prediction and description [10]. ***Prediction*** involves using some variables or fields in the database to predict unknown or future values of other variables of interest. ***Description*** focuses on finding human-interpretable patterns describing the data. The relative importance of prediction and description for particular data mining applications can vary considerably. While in the context of KDD description tends to be more important than prediction, prediction is often the primary goal in pattern recognition and machine learning applications.

B. Learning Approaches and Techniques

Data mining can be described as either supervised or unsupervised [9] [48] as shown in Fig. 2. Unsupervised data mining is rather a bottom-up approach that makes no prior assumptions and aims at discovering relationships in the data. In this sense, data are allowed to speak for themselves; there is no distinction between attributes and targets. In this context, unsupervised data mining is a descriptive approach. Typical methods and applications are clustering, density estimation, data segmentation, smoothing, etc. Supervised data mining, also called direct data mining, aims at explaining those

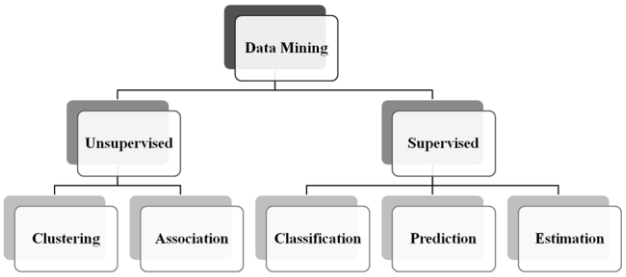

Figure 2 – Data Mining Learning Techniques

relationships once they are found. It is rather a predictive approach where the variables involved are classified as explanatory and dependent ones, and where the main goal is to achieve a liaison between them as in Regression Analysis. Typically, the target variable has to be well specified in advance and three important steps follow before achieving data mining or KDD purposes, as illustrated in Fig. 3.

Specifically, data is to be subdivided for **training**, **testing,** and **validation**. The objective of the training step is to construct a provisional model that attempts to subtend and/or engender the hidden relationship between the attributes and the target variable. The validation step plays a key role in reducing the overfitting traits of the model in the sense that it helps reduce the amount of unsatisfactory results or avoid patterns that are not present in the general dataset (sometimes called flat file). This would occur if the model has excessive number of attributes relative to the amount of data collected and available; the model will exaggerate minor fluctuations in the data and thus, have poor predictive capacity.

Another typical hitch could be an acquired *memory* characteristic that downsizes the model to specific cases in the training phase. For instance, assume that in the training data all students who have passed pre-calculus course have succeeded their graduate studies; we do not want the model to remember this liaison and create the pattern "if the student succeeds the pre-calculus course, then he/she will succeed the graduate studies". Instead, the model should apply *all* patterns found in the training phase to the future data and thus, acquire a generalization characteristic/capability. A number of statistical techniques can be applied to assess the model such as the ROC

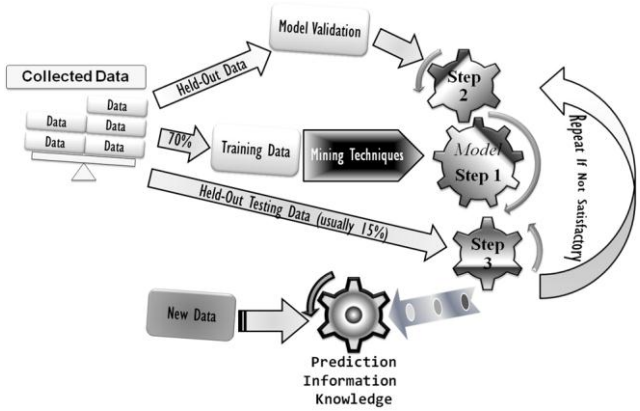

Figure 3 - Roadmap for an Efficient Predictive Tool

(Receiver/Relative Operating Characteristic) curves, which depict the true positive rate vs. the false positive one as the threshold of discrimination is modified in the case of a binary classifier. Finally, testing is used to evaluate, revisit and/or retrain the constructed model by using new *untrained* data that have been held out from the original complete dataset. At this point, the data miner should be able to ascertain whether the learned patterns meet the desired standards. If the outcome is satisfactory, the data miner shall transform the model into information and knowledge, otherwise, re-training, changing the pre-processing and/or the data-mining algorithm is inevitable. This is where the expertise of the data miner plays a decisive role. In fact, not only this expertise is crucial in rectifying the path of the mining process, but also in *converting* the information into knowledge in the sense of filling the gap between the ***past*** (existing data) and the ***future*** (prediction) and drawing an efficient roadmap (decision making and strategic planning) using the outcome of the KDD process.

C. Data Mining Tasks

Data mining objectives can be carried out by means of various procedures, frequently called tasks. Thus a further categorization of data mining is obtained:

- *Classification*: Typical supervised-learning task where information is arranged into predefined classes according to some learnt rules. The learning process aims at developing a model for predicting/assigning the class of a new instance. In other words, it is the generalized application of a known structure to a new data for mapping and classification purposes. The most widely used classifiers are Decision Trees, Bayesian and Neural Networks, etc. This task is applied to diversified fields of expertise such as, Speech and Handwriting Recognition, Web Search Engines, Geostatistics (remote sensing), etc. [42] [44]
- *Regression*: The goal of this task is to achieve a function (regression function) of the independent variables that allows computing the conditional expectation of a dependent variable for prediction and forecasting exercises based on the minimization of a certain type of error via an iterative procedure. Practically, Classification and Regression Trees (CART) summarize these tasks; classification is referred to when the target is nominal, whereas regression is used for continuous values of the target (infinite number of values).
- *Clustering*: It is a descriptive and typical unsupervised machine-learning task common to statistical data analysis wherein a finite set of groups and clusters are identified to describe the data. It is referred to in applications such as Image Analysis, Machine Learning, Biology and Medicine, Pattern Recognition, Education, Crime Analysis, etc. The most reputed approaches related to this task are the K-Means [26] and Fuzzy Clustering techniques [25] [33].
- *Summarization*: Various methods are formulated to describe the set of data and information in a more

compact representation. It also includes report generation.

- *Association Rule Learning or Dependency Modeling*: This task aims at identifying frequent itemsets in a database and deriving association rules. A local model is identified and which describes the important dependencies between variables and datasets. It has been mainly used in applications involving decisions about marketing activities in supermarkets. It is also applied in the fields of Web Mining, Bioinformatics, etc. The most popular and famous algorithm used in this field is the *Apriori* [23] [24] [30] [31].
- *Outlier/Deviation Detection*: Important and significant deviations in the dataset are reported. It finds application in Fraud Detection, Intrusion Detection (Data Security), etc. It is used to increase accuracy by removing anomalous data from the dataset (supervised). Popular algorithms are based on K-Nearest Neighbor, Support Vector Machines, etc. [21] [22] [34].
- *Link Analysis*: Find relationships amongst richly structured databases where hidden patterns are somewhat difficult to be discerned using traditional statistical approaches [20].

Finally, it is noteworthy to mention that *estimation* and *prediction* are sometimes interchangeably used when dealing with classification and regression problems. The reality is that *estimation* is used when a ***continuous*** value is to be forecasted while *prediction* is referred to when new data is classified into one of predefined ***classes***, which are predetermined when building the model.

D. Data Miner Role

It should be noted that efficient and plausible mining exercises remain highly dependent from data preprocessing such as gathering, cleaning, representation, etc. Indeed, although data mining tools and tasks can be very appealing, domain-specific skills are required as a prerequisite before embarking on the trip towards KDD. In other words, human interface plays a decisive role in somewhat deploying or *transforming* data from an opaque entity into a transparent one in order to be resourcefully processed [3]. The data miner has a fundamental role in formulating the problem and preparing the suitable and relevant data. Additionally, data mining can turn out non-satisfactory results at first attempts and miners are to integrate their expertise into the model before starting another iteration of the process. Moreover, astutely adjusting some parameters or trying out different algorithms not only reveals necessary but also requires relevant and consistent justifications. Particularly, the choice of apposite and pertinent attributes can grow intractable, intricate and strenuous namely with large and/or complex classes or datasets. In this sense, adept miners would simplify this task at an early stage and at low computational cost by refining and consolidating the raw data. A typical example would be of David Heckerman [4] about Hot-Dogs and Barbecue-Sauce false inference.

III. Predicting Engineering Students Success and/or Enrollment in Masters Programs

In this paper, we will deal with a particular concern that considerably affects engineering programs in various types of Higher Education Institutions. Pondering over the engineering students' primordial performance demonstrated imperative for an efficient supervision of the attrition rate (on top of probation and suspension rates) and students' chances to further enroll in Masters' studies or to simply achieve (succeed) their engineering degrees [45] [46] [53] [54]. We will, at a first stage, aim at discovering the relationship between the most affecting factors and the aforementioned issues, then, we will try to construct a predictive model that will endow both advisors and administrators with a powerful decision-making tool. The predictive tool or model, as we will call it here, will help tackle such issues as predicting students' GPA, the attrition rate in the Engineering program, and have an insight of the enrollment rate in Masters' studies. Consequently, it will decisively help in planning the courseware and designating needed faculty members, amongst many other pertinent and resulting matters [16].

Another benefit of this tool is to help advisors and instructors have an insight about their students, namely the weak ones. This would help advisors know the capabilities of their advisees and thus, have a better decision when choosing their courses during registration. It would help instructors pay attention to these students during various in-class activities and team forming. Furthermore, special recommendations could be prescribed such as doing extra work or having office-hours visits, etc.

Engineering degrees are mostly offered in two different curriculum structures. One of them is the 150-credit Bachelor of Engineering program (BE) and the other one is the BS (107cr.)/MS (43 cr.) program. In either case, students are to fulfill strict requirements in order to graduate and hold a degree in the Engineering profession. Generally, the engineering program consists of different categories of courses to be completed by the students to fulfill the graduation requirements. Engineering students at Notre Dame University-Louaize (NDU) in Lebanon accounts for approximately 1,200 students (25% of the total number) repartitioned into three departments and four majors (Electrical, Computer and Communication, Civil and Environmental, and Mechanical). Courses are split into four categories: General Education, Core, Major, and Technical Elective courses. Currently, NDU offers the Bachelor of Engineering degree but it is also studying the prospect of launching the BS/MS program. The study undertaken in this paper applies to both cases since the main objective is to predict the performance of engineering students at an early stage of their residency for it affects various factors related to their academic path such as probation, suspension, attrition, graduation, and enrollment in further more-advanced tracks.

The purpose of displaying Fig. 4 below is to first show that a very strong correlation exists between the performance of a student in Major courses and his/her cumulative GPA.

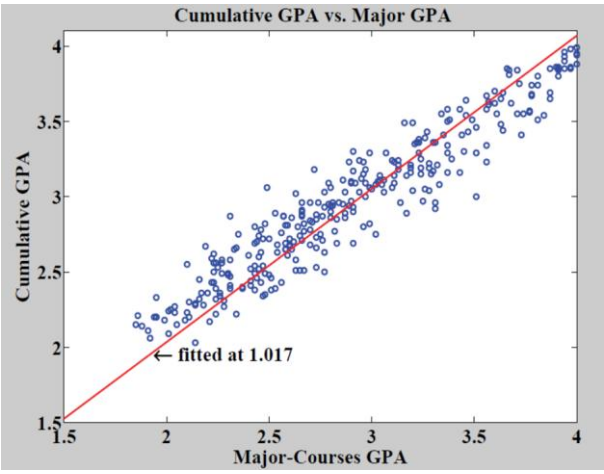

Figure 4 - Relationship between Performance in Major courses and the Cumulative GPA

Certainly, the latter data is unswervingly related to the issues stated at the beginning of this paragraph and therefore, it would be beneficial to use this indicator for it can be obtained, or as we will see later on, is related to attributes obtained at early stages of the engineering curricula.

Most of the core courses are usually taken during the first year. They comprise essentially Math, Physics, and Chemistry courses. *These courses are the prerequisites of almost all major courses* since students are exposed to the fundamental and basic concepts required to pursue specialized theories on a later stage. It is straightforward and safe to assume that if a student has weakly achieved a course, he/she will have fewer chances to *excel* or have superior performance in a higher-level, directly dependent course to which it is a prerequisite. Indeed, the material covered in the higher-level course generally relies on the one covered in the prerequisite and is sometimes a natural/further continuation and advancement in same or similar concepts and topics. Consequently, core courses convincingly play a decisive role in the students' performance in major courses.

As a result, the sought predictive tool will be based on the performance of students in the core-requirement courses, which can be tracked or depicted early. The study will subtend all types of students for the purpose of generalization: weak, average and good performers are included. Table 1 displays the distribution of students with respect to their overall performance.

A. *Data Preparation*

Five hundred Computer and Communication Engineering students' records have been gathered covering a period of almost seven years. These records consist of comprehensive transcripts with the students' grades in all courses taken throughout their academic path. These records also include the number of times students went on probation, those who have been suspended and the cumulative General Point Average (GPA) upon graduation.

The records were preprocessed and cleansed in the sense that records with non-consistent structures were eliminated if adjustment revealed not possible. In fact, over the period of seven years, the engineering curricula have slightly changed; courses have been deleted, modified and/or replaced by new courses. Additionally, straightforward 4.0 or near 4.0 GPA have also been discarded in order to avoid misleading the prediction procedure. This significantly reduced the outlier presence and noisy data. This part of the data mining process resulted in 305 clean records with no missing data and seamless consistency amongst attributes.

Table 1 below displays the distribution of students in relationship with the cumulative GPA. As mentioned earlier, various types of learners have been enrolled in this study; weak, moderate and superior performers have been chosen as to try enhancing the generalization capacity of the model as well as it accuracy.

Table 2 shows a snapshot of the students' records and transcripts as obtained from the Registrar's Office of NDU.

B. *Choosing The Most Pertinent Attributes Using Relieff Algorithm*

The core requirement pool consists of 39 credits comprising the following courses: CEN 201, ENG 201, ENG 202, MAT 211, MAT 213, MAT 215, MAT 224, MAT 235, MAT 326, MAT 335, CHM 211, PHS 212, and PHS 213. For more details regarding a description for each course, please refer to NDU's online catalog cited in [29]. The core requirement pool represents a blend of some chemistry, physics, statics, and mostly math courses, which are mandatory and fundamental. Furthermore, they introduce the students to the most important topics and concepts necessary to pursue courses in Electric Circuits and Electronics, Microprocessor Systems, Electromagnetism, Signal Processing, Communication, Programming, Database, Networking, etc.

In order to produce an effective tool, namely with such database size (not *very* large), we opted for underlying the most influential attributes prior to exercising data mining CART. This helped achieving some sort of pre-pruning before even training the Decision Tree.

Matlab ***Relieff*** algorithm computes the ranks and weights of attribute (predictors) for an input data matrix and response vector for classification and regression with K-Nearest neighbors. When applied to the 305-record data matrix, the importance of the Math courses outperformed the one of Physics courses and finally, ENG 201, ENG 202, CHM 211, and CEN 201 came at the bottom of the list. Particularly, the following courses were retained for our classification and regression study based on the outcome of the *Relieff* algorithm: MAT 213, MAT 224, MAT 235, MAT 335, PHS 212, and PHS 213. Furthermore, weights of the latter Math courses were very close to each other. A similar observation was

Table 1 - Students Distribution vs. GPA

Overall Performance	*Number of Students*	*Cumulative %*	*Relative %*
GPA ≤ 2.0	0	0.00%	**0.00%**
2.0 < GPA ≤ 2.7	104	34.10%	**34.10%**
2.7 < GPA ≤ 3.3	128	76.07%	**41.97%**
3.3 < GPA ≤ 4.0	73	100.00%	**23.93%**

Table 2 - Snapshot of the Students' Records

Core										CSC					Major Courses											Major Labs				GPA	Standing		
PHS212	PHS213	CHM211	MAT211	MAT213	MAT225	MAT226	MAT335	MAT336	CHM261	CSC212	CSC213	CSC312	CSC414	CSC425	EEN201	EEN202	EEN220	EEN230	EEN311	EEN324	EEN331	EEN340	EEN344	EEN443	EEN203	EEN221	EEN312	EEN325	GPA	Dean's List	Suspension	Probation
2.3	3.3	3.7	4	3.3	3	4	3.7	4	4	4	4	3.7	3	3.3	3.3	4	4	3.7	3.3	3.3	2.7	3.3	3.7	3	4	4	4	3.3	3.58	YES	NO	NO
4	3	3.7	4	3.7	4	4	4	4	4	4	4	4	3.7	3.7	4	4	3.3	4	4	4	2.3	3.7	3.3	3	3.7	3.7	3.7	3	3.58	YES	NO	NO
2.3	3	3.7	3.7	3.7	3.7	3	3	4	3.3	4	3.7	3	3.3	4	4	2.3	3.3	3	3	3.3	3.3	3.7	3.7	3	3	3.3	3.7	3.3	3.7	3.49	YES	NO	NO
2	1	1.7	2	3	1.7	2	3	1.3	2.3	1.3	3.3	1.7	1.7	1.7	1.7	2.3	3	2	1.7	1.7	3.3	1.7	1.7	2.3	1.7	2.3	2	1.7	2.7	2.03	YES	YES	YES
1	1.7	1	2	3	3.7	3	4	3.7	2.7	2.7	3.3	3.3	3.3	2	3	3.3	3.3	2.7	1.7	2	1.7	2	4	4	3	2	3	3	2	2.5	NO	NO	NO
1.3	1.7	3	2	1	2.3	3.3	2	1.7	2.3	2	2	2	1.7	2.3	3	1.7	1.7	2	1.7	1.7	2.3	1.7	2	2	2.3	3	2.7	3.3	3	2.11	NO	YES	YES
3.3	2.3	1.3	1.7	1	2.7	2	1	1.7	2.3	3.3	3	2.3	2	2.7	2.7	2.3	2.3	2.3	2.7	3	2.3	2.3	2.3	2.7	4	4	3.7	3	2.6	NO	NO	NO
3.3	3.3	3.3	4	2.7	2	4	3	3	2.7	2.3	3.7	4	4	3.7	3.7	3.7	3	4	3.7	2.7	3.3	4	3	3	3.3	3.7	3.7	3.3	4	3.29	YES	NO	NO
2.7	2	3.7	4	3.7	3	3	3.7	2.7	3	3.7	3.7	3.3	2.3	2.7	2.7	3.3	1.7	2.3	2.3	3	2	3	2	2	3.3	4	4	3.3	3	3.06	YES	NO	NO
2.3	1	4	4	2	3	2.7	3	2	2.3	3	3.7	2.7	2.3	3.3	3.3	3	1.7	4	2.7	3	1.7	2.3	2.3	2.3	2.3	4	4	3.3	3.7	2.89	YES	NO	NO
1.7	2.3	3	1	1	2.3	1.3	1.3	2	1.7	1	2.3	2.7	1.7	2.3	3	3.3	2	2.7	1.7	2	4	1.7	1.7	2.3	2	2	2.7	3	3	2.22	NO	NO	NO
1.3	1.7	2.7	2.3	2	3	3	1.3	1.3	3	1.7	2	1.7	2.7	2.7	2	1.7	2	2.3	1.7	2.7	1.7	1.7	1.7	2.3	1.7	3	3	3	2	2.2	NO	NO	NO
2.7	2.3	4	3	3.3	3.3	4	4	3	3.3	3	3.7	4	3.7	4	3.7	3.7	3.3	4	3.3	3.7	3	4	4	3.7	3.7	3.7	4	3.7	4	3.44	YES	NO	NO
2.3	1.3	1	4	3.3	4	3.3	1.3	1.3	2.7	2.3	4	4	3.7	3	3.7	2.7	2.7	3.3	3.3	2	4	3.3	3.3	2.7	3.3	3	4	4	3.7	3.03	YES	NO	NO
3	2.3	3	3	2	3	1.7	2.3	3.7	3.3	3	3.7	3.3	3.3	3.3	3	2.7	2.3	4	3	2.7	3.7	2	3.3	3	2.3	4	4	3.7	4	2.97	YES	NO	NO
3.3	4	2.3	3	2.3	3	2.7	3	1	2.3	3	3.3	4	3	4	4	2.3	2.7	3.7	3.7	3.7	2.7	3.3	2.3	3.3	3.7	4	4	4	3.03	YES	NO	NO
4	4	4	4	4	3.7	4	3.7	3.3	3.3	4	4	4	3.7	4	4	4	3.3	4	4	3.7	2.7	4	3	4	4	4	4	4	3.84	YES	NO	NO
2.7	2	3	3.7	3	3	2.7	2.7	3	2.3	3	2.7	3	3	2	3.3	2.7	3	3	3	2.3	2	3	2.3	3.3	3	3.3	3.7	4	4	2.87	YES	NO	NO
2.3	2	2.7	1	2.7	3	2	1	1.7	1.7	3	3.7	2.3	3.3	3.3	2.7	3.7	3	2.3	2	3	3.3	2	2.3	3	2.7	2.7	3	3.3	4	2.85	YES	NO	NO
1.7	3.3	4	3.3	3.7	3	3.7	3.7	3	3.7	2.7	3	3.3	3.7	3	3	3.3	3	3.7	2	2.7	2.3	4	2.3	2	1.7	3.7	2.7	4	2.3	2.96	YES	NO	NO
1.3	1.7	3	1	1	3.3	1.3	1.3	1.7	2.3	2.7	2.7	2.3	2	3	2.3	2	1.7	2.3	2	2.3	3	1.7	2.3	2	2.3	2.7	2.7	3.3	3.7	2.32	NO	NO	NO
1.3	2.3	1.7	2.7	1	2.3	2	3	1.3	3	2.3	2.7	2.7	2	1.7	2.7	3	1.7	2.3	2	1.7	2	3.3	2.7	2.7	2.3	4	3.3	3.3	3	2.4	NO	NO	NO
2.3	3	1.3	3.7	2.7	3	2	1	2.3	3.3	3	3.7	3	3	3.7	4	2	3	3.3	2	3.3	3	1.7	2.7	3	2.3	3.7	3	2.3	3.7	3.02	YES	NO	NO
2.7	2	2	3.3	2.7	3	2.3	2.7	4	3	2.3	3.7	3.7	3	2.3	2.7	2.3	3	3.3	2.3	2	1.7	3	1.7	3	2	3.3	3.7	4	4	2.8	YES	NO	NO
3.7	4	4	2.7	1.7	3	2.3	2.7	2	2.7	4	2.3	3.3	2.7	3	3.3	2.3	3.3	4	3.3	3.7	3	3.3	4	3.3	3	4	4	3.3	4	3.21	YES	NO	NO

depicted for the Physics courses. Consequently, and for other practical reasons such as keeping a reasonable tree size and achieving a good balance between the number of attributes and records, we decided to design a classification tree with two highly predominant attributes: computed average performance in Math courses and computed average performance in Physics courses [43] [51] [52].

It is noteworthy to mention that students' grades are ordinal and they belong to the following set of Letter-Grade/GPA: A/A^+ (4.0), A^- (3.7), B^+ (3.3), B (3.0), B^- (2.7), C^+ (2.3), C (2.0), C^- (1.7), D^+ (1.3), D (1.0) and F (0.0).

C. Creating the Model – Training, Validating, and Testing

We used Matlab Classification and Regression Trees (CART) to achieve two objectives. The first one is to elaborate a ***regression*** tree to predict students' GPA upon graduation based on the GPAs they obtained for the abovementioned Math and Physics courses. The second objective was to create a ***binary classification*** tree that dictates the possibility of a student to either enroll in an Engineering Masters' track or graduate (succeed) the Bachelor of Engineering (BE) program since the decision is based on same academic standards and conditions [17].

For the training stage, 75% of the records were used. The remaining records were equally distributed between cross-validation and testing. We also used Matlab Neural Network Pattern Recognition Tool (NNPRT) to derive the ROC curves and test the performance of our created model.

D. Neural Networks for Pattern Recognition

Matlab Neural Networks tool for Pattern Recognition is applied to the 305 students' records and Fig. 5 and 6 show some performance indicators obtained in our case. Seventy percent of the data have been used for training, 15% for cross-validation and 15% for testing [19] [41]. In order to apply the tool to the data, the nominal classes were transformed as follows:

- *If the student succeeds the BE or enrolls in Masters, a '1' is assigned to represent the class "YES".*
- *If the student did not succeed the BE (failed, dropped out, or suspended) or did not enroll in Masters (failed to meet the academic standards), a '0' is assigned to represent the class "NO".*

Fig. 5 shows the Error Histogram while deploying the model; the results are promising and reveal a reliable tool since the great majority of the error is substantially low.

Fig. 6 shows the ROC curves (Receiver/Relative Operating Characteristic) obtained. These curves also confirm a highly performant operating model in the frame of our application.

E. Classification and Regression Trees

Matlab offers an algorithm described by ***classregtree***, which allows inducing classification and regression trees for different types of attributes. This algorithm creates a *binary* decision tree for predicting the class and/or estimating the value as a function of the predictors (attributes). Each branching node within the tree is split based on the values of a column of the attributes [27]. This algorithm is endowed with powerful control parameters such a minimum splitting criterion and a pruning level, which prunes branches giving less improvement in error cost using a *resubstitution* method by turning some branch nodes into leaf nodes and thus, removing the leaf nodes under the original branch. Furthermore, using the ***test*** function, Matlab allows a 10-fold cross-validation to compute the cost vector and returns a vector containing the standard error of each *cost* value and a scalar value for the best level of pruning. For a classification tree, the cost of a node is the sum of the misclassification costs of the observations in that node whereas for a regression tree, the cost of a node is the average squared error over the observations in that node. The ***classregtree*** also induces an *if-then* structure of rules for a different visualization of the model.

Finally, the ***eval*** function of Matlab produces a vector of predicted response values based on a matrix of new, untrained data predictors.

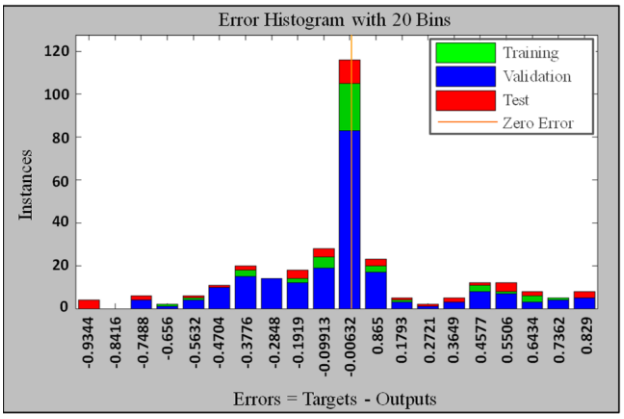

Figure 5 - Error Histogram

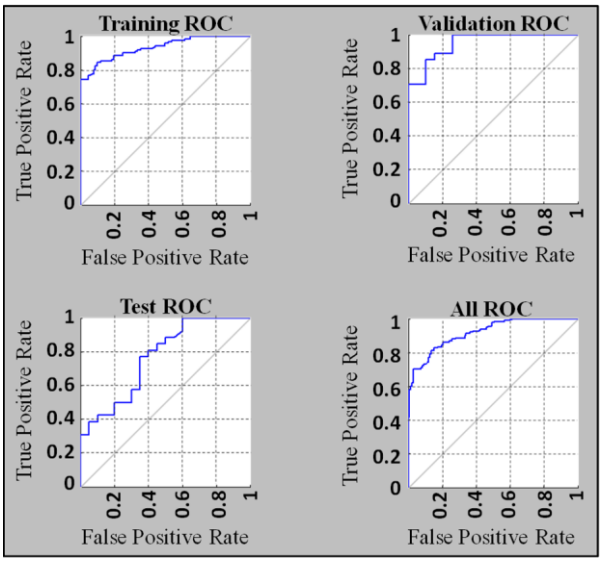

Figure 6 - ROC Curves

Fig. 7 shows how the cumulative GPA is estimated based on the performance of students in Math and Physics courses from the core-requirement pool with a best pruning level reached at 72. It is obvious that the cumulative GPA significantly worsens when students do not highly perform in these courses. Nonetheless, it is indubitably **not reliable** for accurate prediction because of both the size and type of the data. On the other hand, it constitutes another indicator that confirms our postulations and hypotheses.

To use the decision tree effectively and make use of the results from the Neural Networks Error-Histogram and ROC curves, a binary classification tree will be deployed and which will *accurately* predict the ability of students to either succeed their bachelor of engineering or enroll in Masters' programs. As shown in Fig. 8, the classification tree designates one class, "YES" or "NO", to a record based on the attributes. The former class infers achievement of a BE or enrollment in Masters, while the latter one signifies that the student would most probably fail or not be allowed to enroll in Masters, as previously detailed.

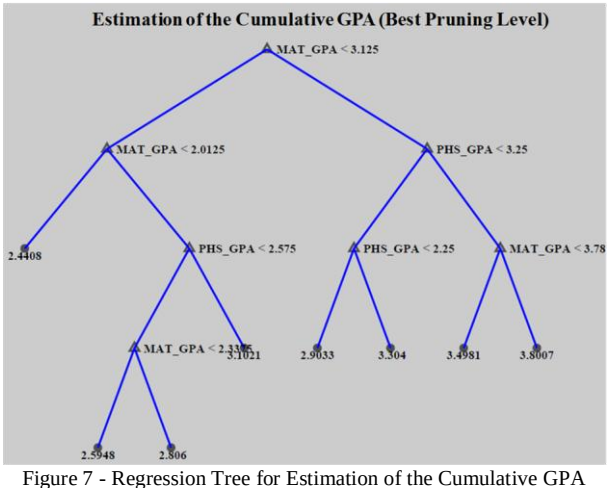

Figure 7 - Regression Tree for Estimation of the Cumulative GPA

Fig. 8 portrays two different pruning levels: Fig. 8-b exhibits the best-pruning level while Fig. 8-a exhibits a pruning level that is less by one than the best level. The purpose is to have a deeper and wider bifurcation when a student presents close attributes with respect to some deciding node values and has to be evaluated by the model.

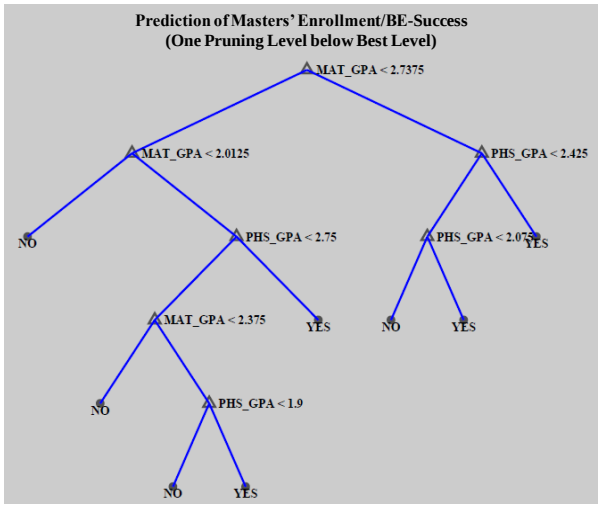

Figure 8.a - Classification Tree (Best Level minus one)

The induced if-then structure as given by Matlab is shown below. This structure is related to the decision tree obtained for the best pruning level.

```
Decision tree for classification
1  if MAT_GPA<2.7375 then node 2 else node 3
2  if MAT_GPA<2.0125 then node 4 else node 5
3  class = YES
4  class = NO
5  if PHS_GPA<2.75 then node 6 else node 7
6  class = NO
7  class = YES
```

After examining the classification tree, we can summarize the following results:

- *If a student's GPA in Math courses is above 2.7375, then he/she is most likely to succeed and/or enroll in Masters.*
- *If a student's GPA in Math courses is above 2.0125, then the condition for a positive outcome is to have a GPA on the Physics courses of at least 2.75.*
- *Otherwise, the chances for students who have a GPA in Math courses less than 2.0125 are scarce to enroll in Masters or accomplish a Bachelor of Engineering.*

IV. Results Analysis, Conclusion

In this paper, we carried out a study to find a reasonably accurate and reliable predictive tool that enables academicians (instructors and advisors) and administrators to decide about the enrollment of engineering students in Masters' studies or to succeed a Bachelor-of-Engineering program.

The study has been conducted in different stages and on different levels. The strong correlation between students' performance in major courses, which are usually taken during the last three years of a Bachelor-of-Engineering program or during the Masters' curriculum, and their cumulative GPA was

demonstrated. The objective of this part of the study was to enable linking core-requirement courses with the GPA and thus, allow predicting students' overall performance at an early stage of their studies.

At first, Matlab ***Neural Networks Pattern Recognition*** tool was applied in order to examine and decide of the accuracy of the predictive tool, which was to be eventually developed. The Error Histogram as well as the ROC curves demonstrated high level of satisfaction and reliability, namely given the size of the data.

Preprocessing the data was of high importance because it helped achieve a low level of outliers and present the data in a more efficient way to classification and regression trees. Particularly, the number of attributes was relatively high when compared to the number of records and thus, preliminary studies have been conducted and which efficiently, and without loss of information, reduced the size of attributes to two highly decisive ones. Matlab ***Relieff*** function was used to reveal the most influential attributes to be taken into account.

At the last stage, Matlab ***classregtree*** tool was used in two steps. The first one was to create a regression tree that estimates the cumulative GPA based on the attributes. This gave us another confirmation of the postulation we started from and then, at a later step, a binary classification tree was achieved after cross-validation and appropriate pruning. This decision tree created an if-then rule structure that enables the engineering staff to ponder over students' chances of succeeding their engineering studies.

The results revealed promising especially when discussed with Math and Physics courses' instructors and engineering advisors who all agreed on the *discovered/learnt* liaison and patterns. It confirmed that students, who are weak performers particularly in this pool of courses, exhibit difficulties in most of the cases in comprehending advanced engineering concepts and in achieving high performance in major courses. Furthermore, the study gave specific numbers and thresholds in courses taken at early stages that would alarm academicians about the situation of the concerned students.

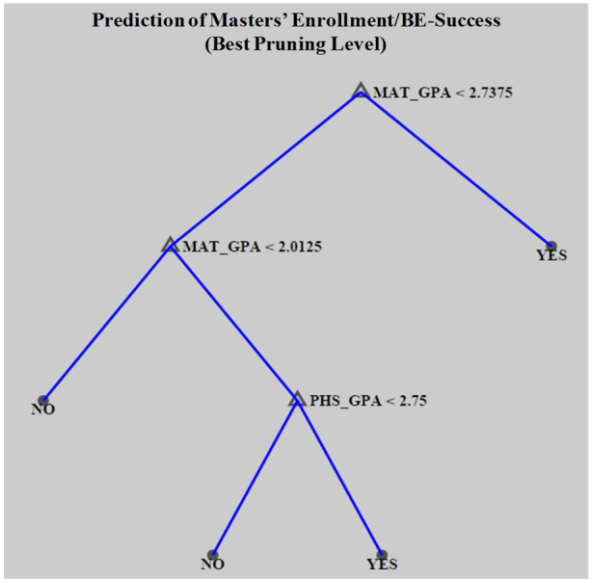

Figure 8.b - Classification Tree (Best Pruning Level)

Finally, this study will indubitably help in predicting the number of students who will reach the end of the engineering program and thus, constitutes a performant tool for decision-making and forecasting enrollment and courseware planning as well as pondering over attrition in engineering studies. It would also endow advisors and courses' instructors an anticipated estimation of their advisees and students capabilities during registration and in-class activities and a reliable platform for special recommendations.

REFERENCES

[1] R.S.J.d. Baker, "Data Mining for Education," In International Encyclopedia of Education, vol. 7, B. McGaw, P. Peterson, E. Baker (Eds.), 3e, Oxford, UK: Elsevier, 2010, pp. 112-118.

[2] http://www.educationaldatamining.org/

[3] J. Teresko, "Information Rich, Knowledge Poor?," Data Warehouses Transform Information into Competitive Intelligence, February 3, 1999. (http://www.industryweek.com/articles/information_rich_knowledge_poor_245.aspx)

[4] C. Silverstein, S. Brin, R. Motwani, and J. Ullman, "Scalable Techniques for Mining Causal Structures," Data Mining and Knowledge Discovery, volume 4, numbers 2-3, pp. 163-192, 2000, DOI: 10.1023/A:1009891813863

[5] P. Domingos, "Toward knowledge-rich data mining," In Proceedings of Data Min. Knowl. Discov., 2007, volume 15, issue 1, pp. 21-28, DOI: 10.1007/s10618-007-0069-7

[6] http://www.kdnuggets.com/data_mining_course/x1-intro-to-data-mining-notes.html (retrieved in September 2011)

[7] http://www.focus.com/fyi/10-largest-databases-in-the-world/

[8] http://beyondrelational.com/justlearned/posts/290/top-10-largest-databases-in-the-world.aspx

[9] M. Berry and G. Linoff, "Data Mining Technique: For Marketing, Sales, and Customer Support," New York: Wiley Computer Publishing, 1997

[10] U. M. Fayyad, G. Piaetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery: An Overview," In Advances In Knowledge Discovery and Data Mining, AAAI/MIT press, Cambridge mass, 1996.

[11] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Third Edition, Morgan Kaufmann, 2011.

[12] Jiawei Han, Micheline Kamber, and Jian Pei, Data Mining: Concepts and Techniques, Third Edition, Morgan Kaufmann, 2011.

[13] Krzysztof J. Cios, Witold Pedrycz, and Roman W. Swiniarski, Data Mining: A Knowledge Discovery Approach, Springer 2007.

[14] D. J. Hand, Heikki Mannila, and Padhraic Smyth, Principles of Data Mining, Massachusetts Institute of Technology, 2001.

[15] N. Delavari, S. Phon-Amnuaisuk, and M. R. Beikzadeh, "Data Mining Application in Higher Learning Institutions," Informatics in Education, vo. 7, no. 1, pp. 31-54, 2008.

[16] S. Sembiring, M. Zarlis, D. Hartama, S. Ramliana, and E. Wani, "Prediction of Student Academic Performance by an Application of Data Mining Techniques," International Conference on Management and Artificial Intelligence, IACSIT Press, IPEDR vol. 6, pp. 110-114, 2011.

[17] S. A. Kumar and M. N. Vijayalakshmi, "Efficiency of Decision Trees in Predicting Student's Academic Performance," First International Conference on Computer Science, Engineering and Applications, CS and IT 02, pp. 335-343, 2011.

[18] C. Ho Yu, S. DiGangi, A. Jannasch-Pennell and C. Kaprolet, "A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year," Journal of Data Science, vol. 8, pp. 307-325, 2010. (neural networks, cross validation)

[19] Fausett, Laurene (1994), Fundamentals of Neural Networks: Architectures, Algorithms and Applications, Prentice-Hall, New Jersey, USA.

[20] L. Getoor, "Linking Mining: A New Data Mining Challenge," ACM SIGKDD Explorations, vol. 5, no. 1, pp. 84-89, 2003. (large data)

[21] D. Denning, "An Intrusion Detection Model," Proceedings of the Seventh IEEE Symposium on Security and Privacy, pp. 119-131, May 1986.

[22] M. R Smith and T. Martinez, "Improving Classification Accuracy by Identifying and Removing Instances that Should Be Misclassified," Proceedings of International Joint Conference on Neural Networks, pp. 2690-2697, 2011.

[23] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases," Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 207-216, May 1993.

[24] G. Piatetsky-Shapiro, "Discovery, analysis, and presentation of strong rules," in G. Piatetsky-Shapiro and W. J. Frawley, eds, Knowledge Discovery in Databases, AAAI/MIT Press, Cambridge, MA, 1991.

[25] R. Nock and F. Nielsen, "On Weighting Clustering," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 28, no. 8, pp. 1-13, August 2006.

[26] J.H.Ward Jr., "Hierarchical Grouping to Optimize an Objective Function," Journal of the American Statistical Association, vol. 48, pp. 236-244, 1963.

[27] J. R. Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, Kluwer Academic Publishers, Boston, pp. 81-106, 1986.

[28] http://www.educationaldatamining.org/JEDM/

[29] http://www.ndu.edu.lb/administration/registrar/catalogs.htm

[30] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," in VLDB'94 Proceedings of the 20th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1994. (a priori)

[31] P. Fournier-Viger, U. Faghihi, R. Nkambou, and E. M. Nguifo, "CMRULES: An Efficient Algorithm for Mining Sequential Rules Common to Several Sequences," in Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010), AAAI Publications, pp. 410-415, 2010. (association task)

[32] A. Nandeshwar, T. Menzies, and A. Nelson, "Learning Patterns of University Student Retention," Expert Systems with Applications, vol. 38, issue 12, pp. 14984-14996, Nov.-Dec. 2011. (attrition)

[33] R. Nock and F. Nielsen, "On Weighting Clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 8, pp. 1-13, August 2006.

[34] M. A. Maalouf, Machine Learning and Data Mining for Computer Security: Methods and Applications, Springer-Verlag, Limited 2006. (outlier)

[35] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," in OSDI'04 Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation, USENIX Association Berkeley, CA, USA, vol. 6, 2004. (large data)

[36] J. Luan, "An Exploratory Approach to Data Mining in Higher Education: A Primer and a Case Study," Paper presented at the AIR Forum, Seattle, Wash., 2000. (Higher Education)

[37] R.S. Baker, A. T. Corbett, K. R. Koedinger, and A. Z. Wagner, "Off-Task Behavior in the Cognitive Tutor Classroom: When Students Game The System," in Proceedings of ACM CHI: Computer-Human Interaction, pp. 383-390, 2004. (higher education)

[38] R.S.J.d. Baker, S. K. D'Mello, M. M. T. Rodrigo, and A. C. Graesser, "Better to Be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners' Cognitive-Affective States during Interactions with Three Different Computer-Based Learning Environments," International Journal of Human-Computer Studies, vol. 68, no. 4, pp. 223-241, 2010. (higher education)

[39] R.S.J.d. Baker, and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," *Journal of Educational Data Mining*, vol. 1, issue 1, pp. 3-17, 2009.

[40] F. Thabtah, "Pruning Techniques in Associative Classification: Survey and Comparison," Journal of Digital Information Management, vol. 4, no. 3, pp. 197-202, September 2006. (pruning)

[41] Sarle, Warren S. (1994), "Neural Networks and Statistical Models," Proceedings of the Nineteenth Annual SAS Users Group International Conference, April, pp 1-13.

[42] J. Quinlan, "Simplifying Decision Trees," International Journal of Man-Machine Studies, vol. 27, no. 3, pp. 221-248, 1987. (Decision Trees)

[43] L.-X. Zhang, J.-X. Wang, Y.-N. Zhao, and Z.-H. Yang, "A Novel Hybrid Feature Selection Algorithm: Using Relieff Estimation for GA-Wrapper Search," in Proceedings of the Second International Conference on Machine Learning and Cybernetics, pp. 380-384, 2-5 November 2003.

[44] P. Shekhawat and S. Dhande, "A Classification Technique using Associative Classification," International Journal of Computer Applications, vol. 20, no.5, pp. 20-28, April 2011. (AC)

[45] G. P. Adanez and A. D. Velasco, "Predicting Academic Success of Engineering Students in Technical Drawing from Visualization Test Scores," Journal for Geometry and Graphics, vol. 6, no. 1, pp. 99-109, 2002. (engineering)

[46] V.O. Oladokun, A.T. Adebanjo, and O.E. Charles-Owaba, "Predicting Students' Academic Performance using Artificial Neural Network: A Case Study of an Engineering Course," The Pacific Journal of Science and Technology, vol. 9, no. 1, pp. 72-79, 2008. (engineering)

[47] F. Thabtah, P. Cowling, and Y. Peng, "Multiple Label Classification Rules Approach," Journal of Knowledge and Information System, vol. 9, pp. 109-129, Springer-Verlag, 2006. (large data)

[48] T.-M. Huang, V. Kecman, and I. Kopriva, Kernel Based Algorithms for Mining Huge Data Sets: Supervised, Semi-supervised, and Unsupervised Learning, Springer-Verlag Berlin Heidelberg, 2006. (large data + supervised)

[49] J. Meenakumari and R. Krishnaveni, "Transforming Higher educational institution administration through ICT," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 2, no. 8, pp. 51-54, 2011. (higher education as a business)

[50] V. Kumar and A. Chadha, "An Empirical Study of the Applications of Data Mining Techniques in Higher Education," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 2, no. 3, pp. 80-84, March 2011.

[51] Y. Wang and F. Makedon, "Application of Relief-F Feature Filtering Algorithm to Selecting Informative Genes for Cancer Classification Using Microarray Data," in Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference, IEEE Computer Society Washington, DC, USA, 2004 http://dx.doi.org/10.1109/CSB.2004.35 (relieff)

[52] I. Kononenko, E. Simec, and M. R. Sikonja, "Overcoming the Myopia of Inductive Learning Algorithms with Relieff," Journal of Applied Intelligence, vol. 7, no. 1, pp. 39-55, 1997. (relieff)

[53] A. H. Basha, A. Govardhan, S. V. Raju, and N. Sultana, "A Comparative Analysis of Prediction Techniques for Predicting Graduate Rate of University," European Journal of Scientific Research, vo. 46, no. 2, pp. 186-193, 2010.

[54] V. Ramesh, P. Parkavi, and P. Yasodha, "Performance Analysis of Data Mining Techniques for Placement Chance Prediction," International Journal of Scientific and Engineering Research, vol. 2, issue 8, pp. 1-6, August 2011.

AUTHOR'S PROFILE

Dr. Chady El Moucary graduated from the Lebanese University – Faculty of Engineering with a diploma in Electrical and Electronics Engineering. He pursued his postgraduate studies in Paris (France) with a scholarship from the French National Center for Scientific Research (CNRS) and received his PhD degree in Electrical Engineering from the University of Paris XI and the SUPELEC in 2000. Currently, Dr. El Moucary is a full-time Assistant Professor and Researcher at Notre Dame University -Louaize (NDU, Lebanon) and the Coordinator of the Faculty of Engineering in NDU's North Lebanon Campus. His research interests and publications are in Electric Machine Control, Digital Watermarking, and Data Mining. He is also a member of the Scientific Committee of the *Order of Engineers* in Tripoli (Lebanon) and a reviewer for diverse reputable worldwide conferences as well as an active member in many other educational/academic committees.