ISSN 1870-4069

Modeling Students' Dropout in Mexican Universities

Noel Enrique Rodríguez-Maya[1], Carlos Lara-Álvarez[2], Oscar May-Tzuc[3], Brian Alison Suárez-Carranza[1]

[1] Departamento de Sistemas y Computación, Instituto Tecnológico de Zitácuaro Michoacán, Mexico

[2] CONACYT Research Fellow - Centro de Investigación en Matemáticas (CIMAT), Zacatecas, Mexico

[3] Facultad de Ingeniería, Universidad Autónoma de Yucatán, Mérida, Yucatán, Mexico

nrodriguez@itzitacuaro.edu.mx, carlos.lara@cimat.mx, maytzuc@gmail.com, alison@hotmail.com

Abstract. Nowadays, the student dropout rate in Mexican higher education institutions have increased, affecting diverse life aspects such as economic, social, academic, and among others. The students dropout prediction is a challenge because of numerous interacting factors; an accurate prediction is useful for universities to implement strategies that reduce student failure –e.g., to implement tutorial action plans. The present research work proposes a model to predict dropout based on the student self-reported information and scores on the university entrance exam. A case of study to evaluate the proposed model was performed, results show a precision of about 86%. This model can serve to support decisions on strategies to reduce students' failure.

Keywords: students' dropout, predictive models, machine learning, data mining.

1 Introduction

The scholar dropout is a problem present in the majority of institutions of higher education in Mexico, affecting their terminal efficiency and academic performance. In addition, the scholar dropout has been increasing in the last years with repercussions in social, economic, and academic aspects. According to the National Association of Universities and Institutions of Higher Education (ANUIES), in Mexico, 25% of first-year students leave the university without completing their courses, that percentage increases to around 50% as they advance in their studies [1]. Students leave their studies for a variety of reasons

that may be personal, academic, physical, economic, and emotional in nature [2,3,4,5].

Although the problem is not new, it acquires special relevance in the so-called 'knowledge society', in which the knowledge is particularly important for the economic development of the nations, and people seek to improve their educational levels in order to be able to compete for better jobs and lifestyle [6].

Institutions of higher education in Mexico select new students based on entrance exams such as the EXANI II (National New Applications Exam for Universities) designed by the National Center of Evaluation for the Superior Education (CENEVAL, for its acronym in Spanish). This exam brings information about previous knowledge for the new possible students in predictive areas that represents the possible academic performance in the university.

In the last years, the amount of data has increased considerably. Data Mining (DM) has proven to be a powerful tool to identify interesting and hidden patterns through the search for existing relationships within the available data. Different objectives can be achieved with DM, among which prediction and description are the most important. Prediction is responsible for the creation of models that can approximate a set of given observations. In general, there are two predictive approaches: classification and regression. In both cases there are a set of predictive attributes and an objective attribute. The main difference is that in classification, the objective is discrete, while in the regression the objective is to learn a continuous function [7].

This paper presents a classification model to predict students' dropout in universities. The model is intended to accurately and promptly identify those students at risk of dropout, so that universities can generate plans and strategies for retention. The model uses a DM methodology to learn models based on decision trees. A case study was used with information from the 2010-2015 and 2011-2016 generational cohorts of the Technological Institute of Zitacuaro. The data collected include: qualifications (performance) of students, and results of the EXANI II, which contemplates different personal, economic, social, cultural and academic factors.

The model was generated with the integration of a dataset, with the most important attributes of the students (predictors) as well as their corresponding level of academic performance (objective). The task consisted of creating a classification model in which each tuple belongs to a class or category. Recently, the majority of works have proposed the use of different local academic attributes to predict the scholar dropout [18,19,20,21,22]. This contribution proposes the use of some attributes obtained from EXANI II to generate learning models that predict the drop out. On the other hand, the establishment of the academic performance is an important aspect to take into consideration. This work proposes to use the number of academic credits in an specific generational cohort as the performance of students. The dataset is composed by the sequences $\mathcal{M}$ and $\mathcal{P}$ represented by Eq. 1:

$$\mathcal{D} = \{(m_i, p_i) \mid i \in 1, \ldots, k\}, \tag{1}$$

where $m_i \in \mathcal{M} \subset \mathbb{R}^N$ are academic attributes from the EXANI II, $p_i \in \{false, true\}$ is the performance obtained by students in a specific generational cohort, and k is the number of instances (information of students).

The results yield a prediction model with accuracy greater than 86%. It is intended that this work can be easily generalized to be implemented in other universities. The generated models can be used as academic tool to prevent desertion and for the application of the necessary retention mechanisms since the selection of new students.

The rest of this paper is organized as follows: Section 2 presents the main related works; Section 3 presents the methodology and main used tools; Section 4 presents the Proposed Model; Section 5 shows the experiments and the main results; and the Section 6 presents the conclusions.

2 Related Work

The first aspect to be done in school dropout modeling is the discrimination of attributes that most affect this phenomenon. Rodríguez Echeverría et al. [8] performed an analysis on the characteristics of the admitted students to the Technological Institute of Sonora in 2002.

They shown the importance of the selection exam score and the performance obtained by the students in their previous degree. These parameters (qualification and performance) can serve as measures of academic behavior prediction in the university environment.

On the other hand, Artificial Neural Networks [14] have been proposed to predict scholar dropout at the University of Genova, Italy in the years 2008 and 2009. For the creation of the model, the authors used students' academic information from surveys, interviews, and students' characteristics before entering the university. The model was able to estimate those students with major scholar dropout incidence with an accuracy greater than 90%.

Scoot et al. [18] proposed the identification of the determinants of failure and subsequent dropout of students in the first year of a computer science course at Glasgow University. The information was collected through questionnaires, interviews, and data with academic students performance provided by the department of computer science. An analysis was made using different sets of data collected and DM techniques. It was concluded that the collected data were not sufficient to make a completely reliable prediction. However, interesting results were obtained that can serve as a point of reference to know the performance of the student and a possible identification of risk of scholar dropout.

Lam-On et al. [19] used the factors that would help to predict the students' performance by taking information such as pre-university characteristics, admission data and initial performance. Their proposed a method focuses on the transformation of data to a matrix where the relationships of the clustered sets were shown. It was hypothesized that multiple clusters of data may provide more useful information for classification. The results suggest that Principal Component Analysis leads to the most accurate prediction.

Márquez Vera et al. [20] collected information from three different sources: a survey designed to obtain personal and family data of the students, the results of the CENEVAL (EXANI I) exam and the scores obtained in the different subjects. The resulting information was 77 attributes of 670 students where 610 passed and 60 failed the course. At the end, a reduction was achieved to only 15 of the best attributes without losing the performance of the classification which was cost sensitive. The results showed an accurate classification in the minority class that is the one that matters in the research.

At the Eindhoven University of Technology in the Electrical Engineering career the scholar dropout rate was 40%, the students generally decide to dropout before the end of January, the goal of [21] was to build a model that predict the students' success, using as input data: information collected until the month of December of the 648 students of the institutional database, including pre-university and university data. Latter, different DM techniques were applied, such as Decision Trees, Bayesian Classifiers and Association Rules. The results indicate that an 80% accuracy was obtained with Decision Trees those proved to be better than other models applied.

Pal [22] applied an algorithm based on machine learning to analyze and extract information from the engineering students data of the Institute of Engineering and Technology of VBS Purvanchal University, Jaunpur, India. A predictive model was established with 1650 records, to identify the most likely students to dropout their studies. The collected data set was obtained through the enrollment form and includes demographic data, past performance data and personal data. Data were selected and transformed leaving only the fields required for DM. Before, four different classification techniques were applied which belong to the Decision Trees. Several experiments were carried out observing the resulting trees and the attributes with more relevance in the results. A precision percentage of more than 85% was obtained which is quite effective to reduce the school dropout rate.

In a previous work [25] we proposed experimentally a predictive model to determine the possible cases of scholar dropout. The model also is based on information collected from the selection exam and the academic performance of students. The main difference with this proposal is that in this research it is proposed a formal model for the creation of predictive models of students' dropout.

Martínez [13] developed a model based on descriptive techniques as Principal Component Analysis and Linear Discriminant Analysis, to detect possible cases of students that require tutoring to reduce the dropout at Technological Institute of Morelia in 2011. Using this tool, four profiles of students were formed: applicants, training, follow-up and closing, as well as three profiles of tutors: training, follow-up and closure. The results show that this tool is useful to generate precise profiles of students and tutors.

In [16] a model based on Chi-square Automatic Interaction Detector (CHAID) was constructed by means of which the most important predictor variables are determined in an Indonesian University. The results showed an accuracy greater

than 80%, determining the accuracy of models, the depth of the Decision Trees.

Many studies have used the DM approach to develop models in scholar dropout prediction [5,9,10,11,12]. Those models are based on Decision Trees and Artificial Neural Networks techniques using personal, economic and academic factors. The results reached precisions greater than 70%. Márquez [3] realized a DM process using information from the first period of students of a high school in the city of Zacatecas in 2012. The WEKA software was used to make the classification process: class balancing, attribute selection and the generation of models based on decision trees and rules of induction. The result was an early detection methodology for possible dropouts.

Cedano et al. [15] proposed a model to predict scholar dropout in an University at La Paz, Baja California Sur, Mexico based on the methodology Cross Industry Standard Process for Data Mining (CRISP-DM). They compared different techniques for model generation such as: Decision Trees, Artificial Neural Networks and Cluster k-means. The results showed precision values of 68% for the Decision Trees and k-means clusters, and 64% for the Artificial Neural Networks.

3 Methodology and Tools

This research was guided through the CRISP-DM methodology, that is conformed of six stages [24]:

1. *Business Understanding.* The aim of this stage is the comprehension of the project objectives from the point of view of the business perspective using a data mining scenario. The students dropout phenomenon was established as an institutional problem, therefore, it was considering all the possible variables (information from the selection exam and the academic performance).
2. *Understanding of data.* In this phase a data collection was performed, this by the understanding and identification of the main elements of the data that can be related with the problem; e.g., the establishment of the student performance.
3. *Data preparation.* Once the main data elements were determined, it begin with the construction of the dataset considering the raw data: this phase comprises a data cleaning and transformation, the selection of attributes, among other data activities.
4. *Modeling process.* In this phase various classifiers were selected according to the performance reported in literature and the selection of the proper software to make the experiments.
5. *Data Evaluation.* Once the most accurate model was selected, a data evaluation was performed; the model is examined in a depth way to determinate possible bias and to make assumptions about the model performance.
6. *Deployment.* Finally, the created model is deployed in a easy access platform to the final user; e.g., in a web platform.

The main tools used in this work are described below:

- *Data mining*: is a set of tools and techniques that allow the exploration of large information in order to find patterns that can explain the behavior of the data. The DM allows an in-depth analysis of data for the purpose to create models that allow, for example, to predict phenomena in the educational environment [4,5].
- *Weka Software*: the Weka platform is a collection of the state-of-the-art machine learning algorithms and data preprocessing tools [7]. It provides extensive support for the whole process of experimental DM that includes: preparing the input data, evaluating learning schemes statistically, and visualizing the input data and the outcome of learning. This diverse and comprehensive toolkit is accessed through a common interface. Therefore, its users can compare different methods and identify the most appropriate for the problem. Weka was developed at the University of Waikato in New Zealand. The system is written in Java and distributed under the terms of the GNU General Public License. It runs on almost any platform and has been tested under Linux, Windows, and Macintosh operating systems-even in a personal digital assistant.
- *Resample technique*: it is applied to a set of data to balance the number of patterns of each class defined in them [17]. Resample produces a random subsample of a dataset using either sampling with or without replacement. An example to configure it and that is of vital importance is to specify "NoReplacement" so that at the time of execution it does not return to select patterns that already have been selected. In Weka, the Resample technique adds instances to a class, this is done simply by adding instances of the class that has only a few instances multiple times in the result dataset. Thus, the resulting dataset is heavily biased in terms of a class for which only a few samples are available.
- *Random Forest*: this technique denotes an improved method for classification, by means of the creation of a large number of classifiers. Each one is constructed using a deterministic algorithm, the generated trees are different for two reasons; first, for each node, the best division is choosen from a random subset of predictors rather than the whole set of them. Secondly, each tree is constructed using a sample of the observations, approximately one-third of these are used to estimate the accuracy of prediction. Unlike other trees, in these there is no possibility of pruning or trimming [23].

4 Proposed Model

A supervised learning strategy is used to predict the student dropout. That is, from a set of training examples

$$\mathcal{D} = \{(m_i, p_i) \mid i \in 1, \ldots, k\}, \tag{2}$$

where $m_i \in \mathcal{M} \subset \mathbb{R}^N$ are academic attributes from the EXANI II; $p_i \in \{false, true\}$ is the performance obtained by students in a specific generational cohort, and k is the number of instances (information of students), we want to

learn a function F that maps a set of academic predictors $\mathcal{M}$ to indicators of student desertion I:

$$F : \mathcal{M} \rightarrow I, \tag{3}$$

where $I \in \{true, false\}$.

4.1 Predictive Variables

Every year CENEVAL defines the most representative variables to determinate the most appropriate students selection; i.e., some attributes can change. The EXANI II applications 2010 and 2011 contain some identical set of attributes, The first part of the model is the unification of predictive attributes taken from the EXANI II. The attributes for EXANI II 2010 and 2011 are[1]:

$EXANI{-}II\_2010$=$\{tipo\_exa$, opc_apli, ano_ver, $tipo\_reg$, $tipo\_resp$, cve_bpm, $apli$, $fecha\_apli$, cve_inst, $identifica$, $desc\_ident$, $lpos\_img$, $folio$, $matricula$, ape_pat, ape_mat, $nombre$, dia_nac, mes_nac, ano_nac, $sexo$, li_mad, li_pad, edo_proc, nom_proc, ciu_proc, cve_proc, reg_proc, mod_bac, $prom\_sec$, $prom\_bac$, exa_extr, dej_est, rec_beca, fal_dia, $porc\_ptar$, fre_tar, can_tar, cal_acl, cal_pun, cal_etar, $trab\_inv$, exa_dept, con_mrl, act_depo, act_cul, act_sal, hrs_trab, est_alca, no_lic, si_lic, si_pos, acc_con, acc_dud, acc_est, int_eje, fre_eje, ses_eje, act_pot, act_paa, act_ipf, fam_exa, pre_exa1, pre_exa2, apo_ami, apo_pro, apo_cur, apo_otr, uti_gui, $enlace$, $vive\_mad$, $vive\_pad$, $vive\_sup$, sup_cali, $trab\_mad$, $trab\_pad$, $esco\_mad$, $esco\_pad$, $cuan\_lib$, $cuan\_peli$, exp_pad, ser_per, ser_rev, ser_inte, ser_cabl, ser_luga, $cine$, $museo$, $espec$, con_lib, ex_ag, ma_pi, ser_tele, ser_lav, ser_ref, $bien\_mic$, ser_dvd, $bien\_pc$, ser_tv, ser_auto, pad_tarj, vac_rm, edo_rep, hab_cop, hab_elim, hab_vir, hab_ptex, hab_pres, hab_fbas, hab_int, hab_core, hab_baj, hab_adj, li_inter, li_taca, li_noti, pos_sel, $icne$, $percen$, $porcecne$, $pcne$, $prlm$, $pmat$, prv, $pesp$, $ptic$, $irlm$, $imat$, irv, $iesp$, $itic\}$.

$EXANI{-}II\_2011$=$\{tipo\_exa$, opc_apli, ano_ver, $tipo\_reg$, $tipo\_resp$, cve_bpm, $apli$, $fecha\_apli$, cve_inst, $identifica$, $desc\_ident$, $lpos\_img$, $folio$, $matricula$, ape_pat, ape_mat, $nombre$, dia_nac, mes_nac, ano_nac, $sexo$, con_dce, con_impe, con_esc, con_ver, con_len, li_mad, li_pad, edo_proc, nom_proc, ciu_proc, cve_proc, reg_proc, ano_bac, mod_bac, $prom\_bac$, bec_bdac, bec_bne, bec_bhd, por_cce, por_obap, por_bib, ppr_acm, ppr_eje, ppr_tar, ppr_pun, ppr_dud, ppr_asi, fev_eno, fev_rude, fev_ties, fev_tefe, niv_idat, niv_erel, niv_ride, niv_rpun, cua_lib, niv_coh, niv_err, niv_resl, niv_ensa, niv_cart, niv_repo, niv_doco, niv_exp, niv_duda, niv_deba, hrs_trab, est_alca, fam_exa, pre_exa1, pre_exa2, $vive\_mad$, $vive\_pad$, $trab\_mad$, $trab\_pad$, $esco\_mad$, $esco\_pad$, $cuan\_lib$, $cuan\_peli$, exp_pad, ser_tele, ser_lav, ser_ref, ser_hor, ser_inte, ser_cabl, ser_dvd, $bien\_pc$, ser_tv, ser_auto, ser_bano, $cine$, $museo$, $espec$, vac_rm, edo_rep, li_inter, li_taca, li_noti, hab_ptex, hab_pres, hab_fbas, hab_baj, pos_sel, $icne$, $percen$, $porcecne$, $pcne$, $prlm$, $pmat$, prv, $pesp$, $ptic$, $irlm$, $imat$, irv, $iesp$, $itic$.

To homogenize the information it is necessary a process to intersect the common attributes for both applications (2010 and 2011). The intersection set

[1] For more information http://www.ceneval.edu.mx/exani-ii

is defined by Eq. 4:

$$EXANI-II = EXANI-II\_2010 \cap EXANI-II\_2011. \tag{4}$$

Once the $EXANI-II$ set is established, it is necessary a process to determine the most representative attributes; i.e., the selection of the most correlated attributes with the objective.

4.2 Objective Attribute

The objective value is determined from the cumulative academic efficiency of student s_i after ten semesters, expressed as the ratio:

$$c(s_i) = \frac{\text{number of credits earned by } s_i}{\text{credits required for graduation}}, \tag{5}$$

therefore, the performance of student s_i is categorized according to

$$p_i = p(s_i) = \begin{cases} \text{false} & \text{if } c(s_i) \geq 1, \\ \text{true} & \text{otherwise,} \end{cases} \tag{6}$$

where 'true' represents a possible dropout, and 'false' the completion of the academic program.

5 Experiments and Results

Experiments were performed in the Technological Institute of Zitacuaro (ITZ for its Spanish abbreviation). ITZ is a superior educational institute with a technological orientation, placed at the east of the Michoacan State in Mexico. Actually, the ITZ offers nine bachelors degrees, increasing the number of students yearly in a constant way. In the last period, more than 2000 students signed up in the different academic programs; hence, it requires a complex academic and administrative planning. In the generational cohorts 2010-2015 and 2011-2016 were reported 16.2% and 19% of desertion, respectively, which affects directly the students' success rates.

Dataset $\mathcal{D}$ is composed by sequences $\mathcal{M}$ defined by the set $EXANI-II$ (Eq. 4), and the objective values $\mathcal{P}$ that correspond to performance values defined by Eq. 6. Data was collected in a separated by commas file (CSV) containing personal and academic information, and EXANI II results of ITZ students. A total of 671 records, from the generational cohorts 2010-2015 and 2011-2016, were prepared and processed. The dataset was structured in a two dimensional matrix containing the EXANI II information and its corresponding success or fail attribute for each student. The attributes are[2]: *cine*, *pcne*, *mes_nac*, *ser_inte*, *li_mad*, *nom_proc*, *fam_exa*, *hab_baj*, *tipo_resp*, *icne*, *ano_nac*, *cuan_peli*, *matricula*,

[2] The complete description of attributes can be found in: http://ceneval.edu.mx/web/guest/exani-ii

Table 1. Correlation of the selected attributes with the student performance.

Attribute	Correlation	Attribute	Correlation
prom_bac	0.2626275	*ptic*	0.0960172
percen	0.2028458	*ser_cabl*	0.0944656
porcecne	0.1573105	*itic*	0.0935362
prlm	0.1541754	*ser_tv*	0.0828841
pos_sel	0.1516624	*bien_pc*	0.0800759
prv	0.1497613	hab_fbas	0.0783265
irv	0.1330568	li_taca	0.0776661
irlm	0.1329894	hab_ptex	0.0757085
icne	0.1307960	$identifica$	0.0710744
pmat	0.1275899	$folio$	0.0703989
pesp	0.1194576	ser_dvd	0.0698548
imat	0.1165091	ser_lav	0.0657144
edo_rep	0.1089331	li_noti	0.0640612
li_pad	0.0992314	ser_auto	0.0585061
ser_inte	0.0987749	vac_rm	0.0552041
...	...	...	...

$esco\_pad$, prv, edo_proc, $tipo\_exa$, exp_pad, $folio$, $tipo\_reg$, $desc\_ident$, $prom\_bac$, $ptic$, li_noti, $esco\_mad$, irv, ser_lav, $pmat$, hrs_trab, $museo$, est_alca, $cuan\_lib$, $irlm$, $vive\_mad$, $espec$, hab_pres, pre_exa1, $percen$, $trab\_mad$, hab_fbas, cve_bpm, ape_pat, pos_sel, ser_tv, ser_cabl, reg_proc, $sexo$, edo_rep, dia_nac, $nombre$, vac_rm, $imat$, ciu_proc, $iesp$, mod_bac, opc_apli, ser_dvd, $porcecne$, cve_inst, ser_auto, ano_ver, pre_exa2, $trab\_pad$, $itic$, $apli$, li_inter, ser_ref, $pesp$, $bien\_pc$, li_taca, cve_proc, $vive\_pad$, $lpos\_img$, $prlm$, ser_tele, li_pad, hab_ptex, ape_mat, $identifica$, $fecha\_apli$, $aluctr$, $performance$.

The next step is to determine those attributes most correlated with the objective attribute; i.e., performance (Eq. 5). Table 1 shows the correlation between each attribute and the performance, and Table 2 shows the selected attributes.

Attributes with a correlation greater or equal than 0.08 were selected; then, the dataset is composed of the following set:

$$\begin{aligned} EXANI-II = \{&prom\_bac, percen, porcecne, prlm, pos\_sel, prv, irv, irlm, icne,\\ &pmat, pesp, imat, edo\_rep, li\_pad, ser\_inte, ptic, ser\_cabl, itic,\\ &ser\_tv, bien\_pc, performance\} \end{aligned}$$

These attributes are related to students' score in different subjects such as: mathematics, verbal and logical reasoning, and Spanish; but it also includes cultural and economic aspects such as: number of mexican states that the student has visited, if students' parents speak an indigenous dialect, if student has access to computer, internet, and cable, among others. It is clear that the most important variables to predict the university studies success are those related with the basic subjects, technologies and cultural aspects.

Table 2. Description of the most correlated attributes.

Attribute	Description	Attribute	Description
prom_bac	Mean of qualification of high school	*pesp*	Percent of qualification of spanish
percen	Percentile of the selection exam	*imat*	Qualification of mathematics (ceneval index)
porcecne	% > CNE of the selection exam	*edo_rep*	Visited states (México)
prlm	Percent of qualification of mathematical logical reasoning	*li_pad*	Father speaks some indigenous dialect
pos_sel	Reached position for the student in the exam	*ser_inte*	Internet availability (in home)
prv	Qualification in verbal reasoning in percent	*ptic*	Qualification in Technologies of information and communications in percent
irv	Qualification in verbal reasoning (CENEVAL index)	*ser_cabl*	Availability of pay cable service (in home)
irlm	Qualification of mathematics logical reasoning (CENEVAL index)	*itic*	Qualification in Technologies of information and communications (CENEVAL index)
icne	CENEVAL index qualification in the exam of selection	*ser_tv*	Number of televisions in home
pmat	Qualification in Mathematics (CENEVAL index)	*bien_pc*	Number of computers in home

Table 3. Performance of different classifiers.

Classifier	Correctly Classified	Precision	Recall
Naive Bayes	63.48%	0.628	0.635
Multilayer Perceptron	81.67%	0.816	0.817
J48	80.18%	0.801	0.802
Random Forest	**86.14**%	**0.862**	**0.861**
Random Tree	83.61%	0.836	0.836

As shown in Fig. 1, data is composed of 274 and 397 instances for the $true$ (dropout) and $false$ (success) classes, respectively. In this sense, data has more instances for the class $false$ than $true$; hence, it is necessary a balancing process to ensure the most distributed patterns into the dataset. In this case, the re-sample tool provided by WEKA was used (more details in Section 3). Table 3 shows the performance of different classifiers. The generated models were validated using cross-validation with ten folds, and the WEKA classifiers used its default parameters.

The models based on decision trees obtained the most accurate results, while the model based on Multilayer Perceptron obtained the third best result, and finally the model based on Naive Bayes obtained the worst performance. On the other hand, Table 4 presents the confusion matrix for the Random Forest model (the most accurate model).

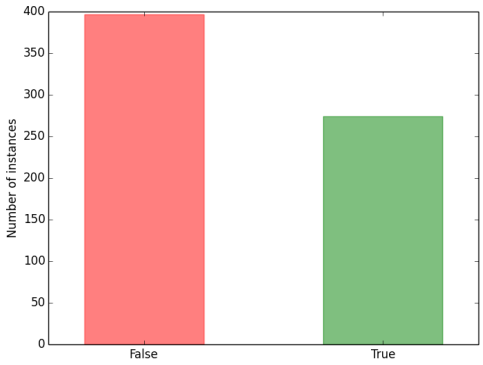

Fig. 1. Distribution of classes.

Table 4. Confusion matrix of model based on Random Forest.

true	**false**	
212	62	**true**
31	366	**false**

The model classifies the *true* instances with an accuracy of 77.4% while the *false* instances with an accuracy of 92.2%. It is clear that the model presents major problems when classify the possible students to dropout their studies.

6 Conclusions

This research presents a predictive model based on a classification task. The phase of knowledge extraction is based on a DM process, the Cross Industry Standard Process for Data Mining (CRISP-DM). The WEKA software was used to support the experiments. A model based on decision trees was the most accurate to predict student dropout in Mexican universities. The proposed model uses information obtained from the admission exam (EXANI II) and students' performance. The model was probed on a study case, particularly the Technological Institute of Zitacuaro in the generational cohorts 2010-2015 and 2011-2016. Results showed an accurate model with more than 86% of precision.

As a future work, the first step is the implementation of the model in a web based interface to support different mobile devices. Next, the prototipe will be implemented in a pilot proof-of-concept to determinate the success or possible dropout of students. The model will be completed with the academic information generated every period, and finally, the model will be implemented as a support to the tutorial works to guide the new ITZ students to increment the student terminal rates and the terminal efficiency (degree acquisition).

References

1. ANUIES: Desertion, lag and terminal efficiency in Higher Education Institutions. Methodological proposal for its study. In: Collection Library of Higher Education, Research Series. Mexico: National Association of Universities and Institutions of Higher Education (2002)
2. Magaña Hernández, M.: Causes of School Failure. In: XIII Congress of the Spanish Society of Adolescent Medicine, Spain (2002)
3. Márquez Vera, C.: Prediction of failure and dropout through Data mining techniques. University of Cordoba (2015)
4. Timarán Pereira, R., Jiménez Toledo, J.: Pattern Detection Student Dropout in Undergraduate Programs of Institutions of Higher Education with CRISP-DM. In: Iberoamerican Congress of Science, Technology, Innovation and Education, Buenos Aires, Argentina (2014)
5. Amaya, Y., Barrientos, E., Heredia, D.: Student Dropout Predictive Model Using Data Mining Techniques. IEEE Latin America Transactions, Vol. 13, No. 9 (2015)
6. González González, M. T.: Absenteeism and Dropout: A Single Situation of Educational Exclusion. Electronic Magazine Iberoamerican on Quality, Efficiency and Change in Education, 4: 1–15 (2006)
7. Witten, I. H., Frank, E., Hall, M. A.: Data Mining: Practical Machine Learning Tools and Techniques. Third Edition, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2011)
8. Rodríguez Echeverría, M., Nereida Aceves López J.: Performance of university students and their relation to the average and Preparatory Course: ITSON case. In: VI International Congress of Organizational Analysis (2008)
9. Martinho, V. R. C., Nunes, C., Minussi, C. R.: A New Method for Prediction of School Dropout Risk Group Using Neural Network Fuzzy- ARTMAP. In: International Conference on Artificial Intelligence 2013 - ICAI'13, 2013, Las Vegas, USA, v. 1., pp. 359–365 (2013)
10. Tejas, S., Swarna, A. and Mathey, R.: Classification with Wekatool for Predicting Student Failure. International Journal of Engineering, Science and Computing, pp. 874–877 (2014)
11. Harilatha, U., Sudhakaryadav, N.: Predicting Educational Performance of a Student Failure and Dropout by using Data mining Techniques. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5(6) (2014)
12. Mendiola, J. L. A., Rosas, R. M. V., Velázquez, J. A. A., Eleuterio, R. A., Marcial-Romero, J. R.: Analysis of school dropout with mining data. Research in Computing Science, 93: 71–82 (2015)
13. Martínez Avila, G. I.: Intelligent Tutoring Model (MIT) in the Technological Institute of Morelia. Autonomous University of Queretaro, Faculty of Information Technology (2015)
14. Siri, D.: Predicting Students' Dropout at University Using Artificial Neural Networks. Italian Journal of Sociology of Education, 7(2), 225– 247 (2015)
15. Cedano, J. A. H., Castro, J. A.: Model of data mining for Identification of patterns that influence the academic. Thesis, La Paz, Baja California Sur, Mexico (2015)
16. Novita, R., Sabariah, M. K., Effendy, V.: Identifying factors that influence student failure rate using Exhaustive CHAID (Chi-square automatic interaction detection). In: Information and Communication Technology (ICoICT), 2015 3rd International Conference on, Nusa Dua, pp. 482–487 (2015)

17. Reyes, J. I., Artagaveytia, F.: Introduction to pattern recognition 2015: Classification of forests by cartographic information. Spain (2002)
18. Scoot, J., Graal, M.: Student Failure in First Year Modules in the Biosciences: An Interview Based Investigation. Bioscience Education (2007)
19. Lam-On, N., Boongoen, T.: Using cluster ensemble to improve classification of student dropout in Thai university. In: Soft Computing and Intelligent Systems (SCIS), 2014 Joint 7th International Conference on and Advanced Intelligent Systems (ISIS), 15th International Symposium on, Kitakyushu, pp. 452–457 (2014)
20. Marquez Vera, C., Romero, C., Ventura, S.: Predicting School Failure Using Data Mining. Zacatecas (2010)
21. Dekker, G., Pechenizkiy, M., VleeShouwers, J.: Predicting Students Dropout: A case study. Zacatecas (2010)
22. Pal, S.: Mining Educational Data to Reduce Dropout Rates of Engineering Students. I.J. Information Engineering and Electronic Business (2012)
23. Pizzuti, C., Ritchie, M. D., Giacobini, M.: Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. In: 8th European Conference, EvoBIO, Instambul, Turkey (2010)
24. Shearer, C.: The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, Vol. 5, No. 4 (2000)
25. Rodríguez-Maya, N.E., Jiménez-Alfaro, A.J., Reyes-Hernández, L.Á., Suárez-Carranza,B.A., Ruiz-Garduño, J.K. Minería de Datos: Modelo Predictivo de Deserción Escolar. En: congreso IEEE Mexican Humanitarian Technology Conference (MHTC) (2017)