Prediction of University Desertion through Hybridization of Classification Algorithms.

Carol Francia Rocha, Yuliana Flores Zelaya,
David Mauricio Sánchez, Armando Fermín Pérez
Universidad Nacional Mayor de San Marcos
Facultad de Ingeniería de Sistemas e Informática
Av. Germán Amezaga s/n, Lima 1, Perú
11200074@unmsm.edu.pe, 11200073@unmsm.edu.pe
dmauricios@unmsm.edu.pe, fferminp@unmsm.edu.pe

Abstract

At present time, the problem of university desertion in Peru is a social phenomenon that involves loss of Peruvian public investment in higher education (not less than a hundred of millions of dollars per year) and also the investment of their parents. For that reason, the aim of this research is to develop a prediction modeling of the dropout of Peruvian university students that allows us to identify those at greater risk to leave their studies, and giving a possibility to take preventive measures which help to maintain the rate of desertion and in the long term it might be reduced. In relation to the solution, we have identified the most influential factors (twenty-four). Additionally, the methodology used was KDD, and we worked with three classification algorithms: Naive Bayes, Multilayer Perceptron and C4.5 Decision Tree separately, and at the same time forming a hybrid prediction algorithm. Each algorithm has chosen based on its greater frequency of use in diverse researches, and its high precision in the prediction. The case study was the School of Systems Engineering of the National University of San Marcos; we used 840 student data from 2008 to 2013.

Keywords: Prediction, University Dropout, Desertion Factors, Data Mining.

1 Introduction

University dropout is a social and economic problem in different societies, Shieu-Hong (2012) define it in several ways: as a total abandonment of a university career, or as a change from one career to another. In this work, the university dropout is the partial abandonment of studies, that is, the student does not enroll in one or more academic cycles.

Currently, this social phenomenon is growing, Pal (2012) establishes that in according to the Organization for Economic Cooperation and Development (OECD), in 2010, the dropout rate in the United States was 50%, and in France, between 36 and 75%, while in Germany was 20% and in Finland, 10%. From the economic point of view, in accordance to Pal (2012), university desertion means a loss of public investment in education and the investment of parents in the higher education of their children. In Peru, between 40 and 50 thousand young people drop out of college each year, representing no less than $ 100 million of the family budget. In addition, the society suffers the loss of professional labor and the increase of unskilled labor, so it is important an early identification of students at high risk of deserting in order to take preventive measures.

Since 1958, there are many studies to predict university dropout using different methods in order to help the early identification of potential dropouts. Thus Mustafa et al (2012) describe the K-means clustering algorithm in the prediction of activities of student learning. On the other hand, Asif et al (2015) present a set of data mining models to predict students' performance using GINI index decision trees, decision trees with gain information, precision decision trees, induction rules with Gain information, neural networks, Naive Bayes and the nearest K-neighbor. Chandra and Pawar (2016) use the K-means clustering method to discover the knowledge that comes from the educational environment; while Kelley (2010) studied classification and logistic regression trees to identify subgroups of students most likely to be retained. In all these studies, variables are academic or family factors,

without considering other factors such as economic type. Single prediction techniques as ARTMAP neural network or ID3 and J48 Decision Trees were used, but also knowing that hybrid methods provide better results in several types of problems, a predictive hybrid model of university desertion using C4.5 Decision Tree, a Naive Bayes algorithm and a multi-layer perceptron have been proposed here based on others researches where they got better results.

This research deals with the prediction of university desertion of students of the Faculty of Systems Engineering and Computer Science of the National University of San Marcos, by identifying those students with the possibility of deserting, through a predictive model implemented with mining techniques.

The rest of the article is organized as follows, section 2 contains a review of university dropout papers, section 3 develops the research proposal, section 4 explains the experiment and then section 5 analyzes the results and shows the conclusions.

2 Related work

There are lots of researches about dropout of university students using data mining techniques, the first study that talks on the subject of consequences of desertion belongs to Tito (1975) and provides the most commonly referred theoretical model proposal.

About this issue, Affendey et al (2010) used the Bayesian approach to classify student academic achievement according to the sites or academic consultations they perform within a university's systems.

Porcel et al (2010) analyzed the relation of the academic performance of the incoming students to the Faculty of Exact and Natural Sciences and Surveying of the National University of the Northeast (FACENA-UNNE) in Corrientes, Argentina, during the first year of career with the same socio-educational characteristics. The performance was measured by the passing of the partial or final exams of the first Mathematics subject. A binary logistic regression model was fitted, which adequately classified 75% of the data.

Kovacic (2010) used data mining techniques (decision trees and neural networks) to identify the variables that define the profile of students who are susceptible to drop out, based on the identification of socio demographic and environmental variables such as ethnicity, course program and course block, which differ with the profile of successful students. All these researches were about high school dropout. About classification approaches, the regression tree classification (CART) was the most successful with a global percentage of correct classification of 60.5%.

Later, Shah (2012) proposed data mining models to identify student performance based on various factors such as: family background, social integration, academic integration, individual characteristics, satisfaction with the characteristics of the University, individual motivation. He studied different models of decision trees, neural networks, rules and functions. His study concludes that decision trees, specifically the RandomForest algorithm, implemented in the WEKA tool, is the one that presents a better performance based on its precision (92.42%).

Also, Shieu-Hong (2012) applied learning algorithms to analyze and extract information from existing data and to establish predictive models; some of them are: Decision Tree CART (Classification and Regression), Decision Tree J48 and Decision Tree ADT (Abstract Data Type), using WEKA software. His data set involved 934 students with 22 attributes, obtaining 83.9% of prediction accuracy using the decision tree ADT.

And more recently, Asif (2015) presented a set of data mining models to predict student achievement by completing four years of high school certification using only academic, non-socio-economic, and demographic factors. It uses the decision tree with GINI index, decision tree with gain information, precision decision tree, induction rule with gain information, neural networks, Naive Bayes and the nearest K-neighbor. It works with a set of balanced and unbalanced data, obtaining different precisions. The Naive Bayes algorithm presents an accuracy of 83.65%, which is the best result after having balanced the data.

3 Proposal

The present paper seeks to predict with a reasonable accuracy, the end of student's degree in the first year of studies based on mainly knowledge discovery method (KDD), the structure is shown in Figure 1. Educational, family, economic and personal factors have been considered, all of them contrasts with other researches that address the same social issue. At the same time, we decided to use a Hybrid in order to improve our results of prediction. Therefore, if a reasonable prediction can be reached, it

will evidence that hybrid works better than a single algorithm. Furthermore, if some of the main factors are identify that not affect in student's decision to leave the university, they will be separated, making the implementation of system intelligent, easier.

4 Methods and Experiment

As a result of comparing different machine learning techniques which resolve the problem of desertion, three classification algorithms were selected, with the highest precision in the prediction (see Table 4). They are Naive Bayes, Multilayer Perceptron and C4.5 Decision Tree. On the other hand, the classification model was implemented using WEKA 3.9 taking account KDD (see Figure 1).

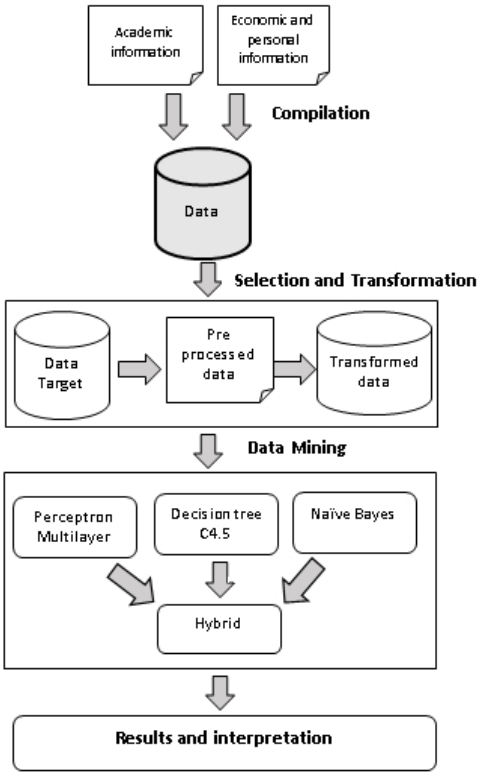

Figure 1: KDD methodology of the proposal.

This case of study was implemented in San Marcos University in the Faculty of System Engineer. The data belongs to the academic years between 2008 and 2013, being in total 1,154 records with 89 possible desertion factors, at the beginning, as shown in Table 5.

4.1 Methods and experiment

The institution provides the information from internal sources, one of them is the Unified System of Enrollment (SUM), which contains academic information, and the second one is the database of the Central Admissions Office (OCA) that saves family, economic and personal factors. From both, the information was given in .CSV format. Table 1 shows the distribution of students per year.

Year	Number of students
2008	230
2009	152
2010	308
2011	149
2012	150
2013	165

Table 1: Distribution of students per year

4.2 Data collection

At the beginning the data was not in a suitable condition to be used by the algorithms, for that reason, it needs to be prepared before using them. The process of cleaning involves removing null or anomalous values and trying to reduce the number of categories for nominal attributes. Some of these factors that underwent changes are detailed below:

- "Own computer", the null value was replaced by NOT COMPUTER.
- "Entrance year", the number of process was eliminated. For instance, 2008–II was changed from 2008.
- "Disability", It contains null values and was eliminated.
- "Department", the null values were replaced by the mode.
- "Province", the null values were replaced by the mode.

As well, SAS MINER was used to create a program which could select the main factors. Internally, Step Forward was configured as a selector technique, giving as a result the list of the 24 final factors shown in the Table 2.

Variable	Values
Enrollment year	2008, 2009, 2010, 2011, 2012, 2013
cod_vcCodigo	4 digits or more
Gender	F (female), M (male)
pos_iattemps	0,1,2,3,4,5,6,7,8,9,10
pos_ilastschoolyear	before 2008
Live with	Parents, alone …
Family type	Both parents, mother, father
School type	< 100 (public), 101-300 (private), > 401 (private)
fic_iHouse Type	Own, rented, other
fic_iHouseMaterial	Bricks, other
fic_iComputer	Yes, no
cal_iEAPPlace	1 – 100
Enrollment Type	Direct, other
Last academic term	2008-1 - 2014-2
Actually works	Yes, no
Old years	17 or more
fic_iFamily Income	< 400, 801-1200, 1201-1600, 401-800, > 1601 soles
fic_iStudent Income	0, < 300, > 301
Score	0 - 2000
Order of academic merit	1 - 60
Weight score	0 - 20
Number of courses Approved	0 - 63
Number of courses Enrollment	0 - 209
Number of courses failed	0 - 63

Table 2: Final dataset for the proposal.

Fifteen of the most of the attributes including the predicted class variable are nominal, and only nine of the attributes are numeric.

4.3 Data mining

The classification model used in the intelligent system was generated through WEKA, which has implemented Naive Bayes, C4.5 Decision Tree and Multilayer Perceptron algorithms. The data was divided into two groups, the first for the training set; it represents 75% of data, and the other 15% for validation. The study was tested several times by adjusting the values, choosing the correct value and comparing the precision value. The most appropriate one will be selected for being used for the model. The way that hybrid works is that, it gets de results of prediction the other techniques and they are added to a list of variables which becomes a new list. Naive Bayes (it has been implemented inside) works with that and shows the results. Additionally, researchers attempted to add a new method, which is Vector Support Machine; however, the results were lower than expected, 85%, being the main reason to be excluded.

5 Results

The Table 3 shows that the hybrid technique has an accuracy of 94.47%, however the Naive Bayes technique has better precision 95.09% and it is better than the other two techniques (91.41% for C4.5 Decision Tree and 92.02% for Multilayer Perceptron) with the same dataset. This is how the feasibility of the Naive Bayes technique is accepted as a predictor of university desertion, but also the possibility of using the hybrid technique is allowed, since the % accuracy rate could improve according to the training.

	% Accuracy	
	Training	Validation
Naive Bayes	97.8	95.09
Hybrid	87.5	94.47
Multilayer Perceptron	99.7	92.02
C4.5 Decision Tree	99.3	91.41

Table 3: Percent of accuracy for each technique.

The Table 4 shows the results of other researches, which have been reviewed. In comparison with that, these three techniques got a better percentage of precision taking account the amount of registers that have been used in the experiment (training and validation). On the other hand, it was very difficult to find a research, which implements a hybrid technique for solving the same problem, because of that there is no similar paper to compare the results.

6 Conclusions

This study has developed an intelligent system based on a model the hybridization of the Neural Network Multilayer Perceptron, C4.5 Decision

Tree and Naive Bayes Classifier. The software integrated for using those algorithms was WEKA. Those techniques with the input variables selected as influential allow having a high degree of precision in the prediction process of student dropout, although Naive Bayes demonstrated that is more accurate than hybrid method. Finally, the KDD methodology allows a correct development of data mining models and their implementation in this case of study.

A recommendation to other researchers is to integrate this intelligent system with a centralized database which store the information of all students, in order to avoid manual work of uploading and cleaning the data before processing.

References

Samy Abu Naser, Ihab Zaqout, Mahmoud Abu Ghosh, Rasha Atallah, Eman Alajrami. 2015. Predicting student performance using artificial neural network in the faculty of engineering and information technology. *International Journal of Hybrid Information Technology*, 8(2):221-228.

Yegni Amaya, Edwin Barrientos, Diana Heredia. 2013. Modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos. *Artículos 2014. Conferencia TICAL 2014.* Cooperación Latino Americana de Redes Avanzadas, RedCLARA.

Deanna Kelley-Winstead. 2010. New Directions in Education Research: Using Data Mining Techniques to Explore Predictors of Grade Retention. *PhD dissertation*. Department of Statistics, George Mason University, Fairfax, VA, U.S.A.

Eduardo Adolfo Porcel, Gladys Noemí Dapozo, María Victoria López. 2010. Predicción del rendimiento académico de alumnos de primer año de la FACENA (UNNE) en función de su caracterización socioeducativa. *Revista Electrónica de Investigación Educativa*, 12(2).

Sonia Formia, Laura Lanzarini, Waldo Hasperué. 2013. Caracterización de la deserción universitaria en la UNRN utilizando Minería de Datos. Un caso de estudio. *TE & ET*, Vol. 11, pages 92 – 98.

Necdet Güner, Abdulkadir Yaldır, Gürhan Gündüz, Emre Çomak, Sezai Tokat, Serdar İplikçi. 2014. Predicting Academically At-Risk Engineering Students: A Soft Computing Application. *Acta Polytechnica Hungarica,* 11(5):199-216.

Marisa Fabiana Haderne. 2012. Uso de Tecnologías de la Información para Detectar Posibles Deserciones Universitarias, *VII Congreso de Tecnología en Educación y Educación en Tecnología,* Red de Universidades con Carreras en Informática, Argentina.

Hitesh Chandra Mahawari1, Mahesh Pawar, 2016. A Study of Various Methods to find K for K-Means Clustering. *International Journal on Computer Sciences and Engineering*. 4(3):45-47.

Kabakchieva, D. 2012. Student Performance Prediction by Using Data Mining Classification Algorithms. *International Journal of Computer Science and Management Research,* 1(4):686–690.

Sarwar Kamal, Linkon Chowdhury, Sonia Farhana Nimmy. 2012. New Dropout Prediction for Intelligent System. *International Journal of Computer Applications*, 42(16):26-31.

Lin Shieu-Hong. 2012. Data Mining for student retention management. *Journal of Computing Sciences in Colleges*. 27(4): 92-99.

L.S. Affendy, I.H.M. Paris, N. Mustapha, Nasir Sulaiman, Z. Muda. 2010. Ranking of influencing factors in predicting students´ academic performance. *Information Technology Journal.* 9(4): 832-837.

María Alejandra Malberti, Graciela Elida Beguerí, Raúl Oscar Klenzi. 2013. Reconociendo factores resilientes en alumnos de informática, mediante la aplicación de TIC. *TE&ET*, Vol. 11, pages 24-34.

Rose Marra, Demei Shen, Kelly Rodgers, Barbara Bogue. 2012. Leaving Engineering: A Multi-Year Single Institution Study. *Journal of Engineering Education,* Vol. 101, pages 6-27.

Mohammad Nurul Mustafa, Linkon Chowdhury, Sarwar Kamal. 2012. Students dropout prediction for intelligent system from tertiary level in developing country. *IEEE/OSA/IAPR International Conference on Informatics, Electronics & Vision*, pages 113–118. https://doi.org/10.1109/ICIEV.2012.6317441

Najmus Saher Shah. 2012. Predicting factors that affect student's academic performance by using data mining techniques. *Pakistan Business Review*, pages 631-668.

Oficina General de Planificación – UNMSM 2007. Seguimiento a los ingresantes y su nivel alcanzado, Vol. 1. *OGP-UNMSM*, Lima, Perú.

Raheela Asif, Agathe Merceron, Mahmood Pathan. 2015. Predicting student academic performance at degree level: A case of study. *International Journal of Intelligent Systems and Applications*, Vol. 7, pages 49–61. https://doi.org/10.5815/ijisa.2015.01.05

Sweta Rai, Ajit Kumar Jain. 2013. Students' Dropout Risk Assessment in Undergraduate Courses of ICT at Residential University – A Case Study. *IJCA International Journal of Computer Applications*, 84(14):31-36.

Saurabh Pal. 2012. Mining Educational Data to Reduce Dropout Rates of Engineering Students. *I.J. Information Engineering and Electronic Business,* Vol. 2, pages 1-7. https://doi.org/ 10.5815/ijieeb.2012.02.01

Saurabh Pal. 2012. Mining Educational Data Using Classification Decrease Dropout Rate of Students. *International Journal of Multidisciplinary Sciences and Engineering*, 3(5):35-39.

Ricardo Timarán Pereira, Andrés Calderón Romero, Javier Jiménez Toledo. 2013. La Minería de Datos como un método innovador para la detección de patrones de deserción estudiantil en programas de pregrado en Instituciones de Educación Superior. *World Engineering Education Forum 2013*. Cartagena, Colombia.

Valquiria Ribeiro De Carvalho Martinho, Clodoaldo Nunes, Carlos Roberto Minussi. 2013. An intelligent system for prediction of school dropout risk group in higher education classroom based on artificial neural networks. *IEEE 25th International Conference on Tools with Artificial Intelligence,* pages 159-166.

Valquiria Ribeiro De Carvalho Martinho, Clodoaldo Nunes, Carlos Roberto Minussi. 2013. Prediction of school dropout risk group using neural network. *3013 Federated Conference on Computer Science and Information Systems (FedCSIS),* pages 111-114.

Vincent Tito, 1975. Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*, 45(1): 89-125.

Appendix A: Table 4.

Method	References	Dataset	Precision
Learning	Kabakchieva (2012)	10067	67.45
C4.5 Decision Tree	Hong Lin (2012)	5943	68.8
	Pal (2012)	300	80.8
	Timarán (2013)	7924	75
	Kabakchieva (2012)	10067	72.74
	Chee-waprakobkit (2013)	1600	85.2
	Formia (2013)	11102	71.65
K-nearest neighbours	Kabakchieva (2012)	10067	70.47
	Asif (2015)	347	74.04
Multilayer Perceptron	Kabakchieva (2012)	10067	73.59
	Chee-waprakobkit (2013)	1600	83.8
	Abu Naser (2015)	1407	84.6
ID3 Decision Tree	Pal (2012)	300	85.7
ADT Decision Tree	Hong Lin (2012)	5943	83.9
	Pal (2012)	300	72.4
CART Decision Tree	Hong Lin (2012)	5943	83.9
	Pal (2012)	300	72.4
Decision Tree with GINI	Asif (2015)	347	68.27
Decision Tree with precision	Asif (2015)	347	60.58
Induction rule with gain information	Asif (2015)	347	55.77
Naive Bayes	Hong Lin (2012)	5943	77.9
	Pal (2012)	165	91.7
	Asif (2015)	347	83.65
ARTMAP FUZZY Neural Network	Abu Naser (2015)	499	85

Table 4: Data mining techniques for predicting university dropout.

Appendix B: Table 5.

Factor	References
Assistance	Praveen (2013)
Loans	Hong Lin (2012)
Teaching	Pal (2012), Marra (2012)
Subjects studied	Amaya (2013)
Annual income	Rai (2013), Amaya (2013), Timarán (2013), Formia (2013)
Lost materials	Timarán (2013)
Socioeconomic level of the family|	Güner (2014), Timarán (2013), Valquiria (2013)
Weighted average	Cheewaprakobkit (2013), Haderne (2012), Timarán (2013), Abu Naser (2015)
Califications average	Cheewaprakobkit (2013), Amaya (2013)
Type of admission	Pal (2012), Rai (2013)
Failed test	Haderne (2012)
Time spent studying	Rai (2013)
English	Cheewaprakobkit (2013)
Family income	Pal (2012), Saher (2012) Valquiria (2013)
Best weighted average	Cheewaprakobkit (2013)
Type of house	Timarán (2013)
Location	Pal (2012), Valquiria (2013)
Continent	Cheewaprakobkit (2013)
Bedroom	Cheewaprakobkit (2013)
Native English	Cheewaprakobkit (2013)
First generation	Saher (2012)
Type of job	Formia (2013)
Children	Formia (2013)
State	Cheewaprakobkit (2013), Amaya (2013), Timarán (2013), Valquiria (2013)
Type of family	Güner (2014), Rai (2013), Saher (2012)
Occupation father	Pal (2012), Saher (2012), Cheewaprakobkit (2013), Timarán (2013), Formia (2013)
Year of high school graduation	Güner (2014), Formia (2013)

Table 5: Desertion Factors

Factor	References
Family Size	Timarán (2013), Valquiria (2013), Saher (2012)
Brothers in college	Güner (2014), Timarán (2013)
Mother	Güner (2014), Timarán (2013)
Father	Güner (2014), Timarán (2013)
Number of brothers	Güner (2014)
Time for extracurricular activities	Saher (2012)
Type of family problem	Rai (2013)
Father's education	Pal (2012), Güner (2014), Rai (2013), Saher (2012), Amaya (2013), Valquiria (2013),
Membership	Marra (2012)
School location	Rai (2013), Abu Naser (2015), Valquiria (2013)
Middle schooling	Rai (2013)
Best course in high school.	Rai (2013)
Average of the previous semester.	Praveen (2013), Saher (2012)
Top Rated	Hong Lin (2012)
Grade	Pal (2012), Haderne (2012)
Year of admission	Haderne (2012)
Counseling.	Marra (2012)
Study plan	Marra (2012)
Programming languages	Asif (2015)
College expenditure	Rai (2013)
Number of extracurricular courses	Cheewaprakobkit (2013)
Mother Occupation	Pal (2012), Formia (2013) Cheewaprakobkit (2013), Timarán (2013)
Credits	Cheewaprakobkit (2013), Amaya (2013), Abu Naser (2015)
Type of high school.	Hong Lin(2012), Güner (2014), Timarán (2013), Valquiria (2013), Abu Naser (2015)
Necessity	Hong Lin (2012), Kamal (2012)

Table 5: Desertion Factors (continued)

Factor	References
Deprecated Courses	Haderne (2012), Timarán (2013), Saher (2012)
Specialty	Pal (2012)
Semester	Haderne (2012), Amaya (2013)
Work in university	Güner (2014)
Enrollment value.	Timarán (2013)
School day	Timarán (2013)
Faculty	Timarán (2013)
Mother's education	Pal (2012), Güner (2014), Saher (2012), Rai (2013), Amaya (2013), Valquiria (2013)
Work hours	Cheewaprakobkit (2013), Formia (2013)
Average high school.	Hong Lin(2012), Haderne (2012), Saher (2012), Abu Naser (2015)
Enrolled in preferred course.	Saher (2012)
High School Ranking.	Hong Lin (2012), Güner (2014)
Age	Hong Lin (2012), Pal (2012), Saher (2012), Cheewaprakobkit (2013), Rai (2013), Haderne (2012), Kamal (2012), Valquiria (2013)
Taste for the university	Rai (2013), Saher (2012)
View of the university system.	Rai (2013), Saher (2012)
View of infrastructure of the university.	Rai (2013)
Extracurricular courses view	Rai (2013)
Entertainment on campus view	Rai (2013)
Participation in extracurricular activities	Praveen (2013), Rai (2013), Saher (2012)
Year of birth	Güner (2014), Haderne (2012), Formia (2013)
Category	Pal (2012)
Ethnicity	Hong Lin (2012), Kamal (2012), Valquiria (2013), Formia (2013)
Lives	Hong Lin (2012), Valquiria (2013)

Table 5: Desertion Factors (continued)

Factor	References
Origin	Hong Lin (2012), Amaya (2013), Valquiria (2013)
Admission Note	Güner (2014), Haderne (2012), Saher (2012)
Occupation	Amaya (2013), Valquiria (2013)
Has computer?	Valquiria (2013), Saher (2012)
Self-appraisal	Valquiria (2013)
Transport	Valquiria (2013)
Work	Valquiria (2013)
Home	Hong Lin (2012), Haderne (2012), Rai (2013), Timarán (2013), Valquiria (2013)
Age of admission	Timarán (2013)
Gender	Hong Lin (2012), Pal (2012), Güner (2014), Amaya (2013), Marra (2012), Haderne (2012), Timarán (2013), Malberti (2013), Kamal (2012), Valquiria (2013), Saher (2012), Abu Naser (2015)
Area materia	Timarán (2013)
Lost semesters	Timarán (2013)
Performance in internal examinations	Praveen (2013)
Performance in seminars	Praveen (2013)

Table 5: Desertion Factors (continued)