Classification Models for Determining Types of Academic Risk and Predicting Dropout in University Students

Norka Bedregal-Alpaca[1], Víctor Cornejo-Aparicio[2], Joshua Zárate-Valderrama[3], Pedro Yanque-Churo[4]
Universidad Nacional de San Agustín de Arequipa, Arequipa, Perú

***Abstract*—Academic performance is a topic studied not only to identify those students who could drop out of their studies, but also to classify them according to the type of academic risk they could find themselves. An application has been implemented that uses academic information provided by the university and generates classification models from three different algorithms: artificial neural networks, ID3 and C4.5. The models created use a set of variables and criteria for their construction and can be used to classify student desertion and more specifically to predict their type of academic risk. The performance of these models was compared to define the one that provided the best results and that will serve to make the classification of students. Decision tree algorithms, C4.5 and ID3, presented better measurements with respect to the artificial neural network. The tree generated using the C4.5 algorithm presented the best performance metrics with correctness, accuracy, and sensitivity equal to 0.83, 0.87, and 0.90 respectively. As a result of the classification to determine student desertion it was concluded, according to the model generated using the C4.5 algorithm, that the ratio of credits approved by a student to the credits that he should have taken is the variable more significant. The classification, depending on the type of academic risk, generated a tree model indicating that the number of abandoned subjects is the most significant variable. The admission scan modality through which the student entered the university did not turn out to be significant, as it does not appear in the generated decision tree.**

Keywords*—Educational data mining; ID3 algorithm; C4.5 algorithm; artificial neural network; classification algorithms; student desertion; academic risk

I. Introduction

Education is fundamental to the development and well-being of a society, so students are the fundamental asset of any educational institution. The social and economic development of a country is directly related to the academic performance of its students [1]. In this context and because of the implications it has, university desertion is a problem that raises concerns in the managers of higher education institutions; on the one hand, the university's finances are affected and on the other hand, the efficiency of the higher education system is questioned, because only a small number of young people who start university studies manages to complete them.

One way to address this problem is to have timely information about the possibility of student dropout, hence the interest in analyzing the possible factors that may lead a student to enter conditions of dropout. Today, there are multiple applications from artificial intelligence to education, data mining techniques are used to discover important patterns and obtain useful information from academic-based information systems [2]. These techniques are known as Data Mining, which analyzes and understands student-related information; so it can transform this information into a useful way that can improve an institution's decision-making to improve the quality of education.

In [3] it is used artificial neural networks, specifically with a Multilayer Perceptron topology, to predict student performance. The attributes selected are mainly of two types, first academic attributes related to the academic details of the students and the personal attributes among them are: study interest, unit test notes, assignment, extracurricular activities, residency, parent education and family status. As a result, 91% accuracy was achieved with MLP training.

In [4] the ID3, J48, Bayes Net, and Naive Bayes algorithms are applied to a sample of 577 students to predict their performance on the first semester exam. Comparative analysis of the results of applying these algorithms has helped students improve and focus more on different courses.

The work of [5] focuses on the development of mining models to predict student performance, based on personal, pre-university and university performance characteristics. Well-known classification algorithms were applied to the data set including a "rule learner", a decision tree classifier, a neural network, and the nearest neighbor classifier. Rated accuracy is obtained between 67.46% and 73.59%. The highest accuracy is obtained for the neural network model (73.59%), followed by the decision tree model (72.74%) and the k-NN model (70.49%).

In [6], the C4.5 decision tree algorithm is used to apply it to students' internal assessment data to predict their performance on the final exam. The decision tree result predicted the number of students who are likely to fail or approve. It is obtained that the C4.5 algorithm is more efficient than ID3 in terms of accuracy and the time needed to train the tree and build it.

In [7] it is conducted an experiment on eight actual datasets using five C4.5 algorithm methods based on entropy from Shannon, Havrda - Charvt, Quadratic, R'enyi and Taneja, making a comparison between them and building a model that takes the entropies of Shanon, quadratic and Havrda-Charvt in parallel and produce more accurate classifications for the dataset and a result of this classification is comparable to the

other machine learning techniques. This approach based on entropies can be applied to real-world classification problems.

This work focuses on classification techniques, techniques that use a set of pre-classified examples to develop a model that can classify a population of similar records [8]). Classification techniques are applied under two different approaches: the first related to student desertion which will be represented by two classes: the student ABANDON the career and the student DON'T ABANDON the career; and the second that identifies the type of risk of students from those who in the first approach the resulting class was that they were leaving the race; This is intended to identify whether the student presents an INCIPIENT, MODERATE, HIGH or VERY HIGH risk of dropping out of university studies.

It makes use of an own-build application where the functionality of creating classification models for different algorithms such as artificial neural networks and decision trees was implemented; of the latter specifically used the algorithm ID3 and C4.5. An analysis of the variables chosen as inputs to the classification algorithms will be performed and the performance metrics of the models generated for each of the proposed approaches will be compared. Subsequently the best models can be used to classify the desertion or risk type of future students.

II. Theoretical Mark

A. Classification

It is the most commonly applied data extraction technique; it can predict the value of a categorical attribute (discrete or nominal). Uses a set of preclassified examples to develop a model that can classify the overall record population [9]; uses a decision tree or neural network-based classification algorithms. The process involves two phases, learning and classification. In the learning phase, the classification algorithm analyzes the training data. In the classification phase, test data is used to estimate the accuracy of the classification rules. If accuracy is acceptable, rules can be applied to new data tuples. The classifier training algorithm uses the preclassified examples to determine the set of parameters required for proper discrimination, and then encodes those parameters into a model called classifier.

B. Artificial Neural Networks (ANN)

They are computing algorithms that can solve complex problems by mimicking animal brain processes in a simplified way [10]. They are based on a directed graph structure, composed of a set of neurons that interconnect through arcs directed with an associated weight that determines the force and sign of the connection. Neurons are organized by levels or layers and have two defined functions: activation and output. Their main advantage is that they can solve problems that are very complex for conventional technologies; problems that do not have an algorithmic solution or for which an algorithmic solution is very complex to be defined. Another advantage is that they are very robust with respect to noise in the data.

The best-known ANN is Multilayer Perceptron (MP), which comprises an input layer, a hidden layer, and an output layer, this type of network is being used to perform data mining processes [11]. MP networks consist of neurons arranged in layers and interconnected by synaptic weights and can filter and transmit information, in a supervised manner, to build a predictive model that classify data stored in memory.

C. Decision Trees

They are tree-shaped structures whose nodes represent a choice between several alternatives, and each leaf node represents a decision [4], [8]. They are generally used for classification, because they are simple hierarchical structures that facilitate user understanding. The training algorithms are simple and their computation time is short, as they only require traversing the tree until they reach the leaf node to perform the classification. They use real data extraction algorithms to help with classification. They are used as support for the choice between various lines of action, allowing to explore the possible results for various options, and evaluate the risk and rewards for each possible course of action. These decisions generate rules, which are then used to classify the data. Decision tree learning algorithms include ID3, C4.5 and ASSISTANT.

D. Algorithm ID3 – Iterative Dichotomizer 3

It was invented by Ross Quinlan, iteratively divides the attributes into two groups: the most dominant and the others to build the tree from top to bottom. ID3 uses information theory to determine the most informative attribute, for this there are two concepts involved: entropy and information gain. If the entropy of the attribute is zero, it is a homogeneous node and does not need to be classified; if your entropy is 1, it is a heterogeneous node and you need to continue to classify it. Built the tree, it is applied to a tuple in the database, resulting in the classification of that tuple. The sample dataset must consist of a series of tuples of values, each called attributes, in which one of them (the attribute to classify) is the target, which is of binary type (positive or negative, yes or no, valid or invalid).

Applies to discrete attributes, at each node an attribute is selected and a value is selected, for example, Average - High. The ID3 algorithm never produces trees that are too large, which makes it easier for the user to read and interpret it.

E. Algorithm C4.5

Based on Hunt's C4.5 algorithm, it handles categorical and continuous attributes to build a decision tree. The C4.5 algorithm is proposed in [12] as an improvement of the ID3 algorithm, which eliminates many of its limitations, for example:

- Allows the possibility to use continuous data by separating the possible results into two branches according to a selected threshold, so that all values above the threshold are assigned to one son and the rest to the other son. For example, Average >15 and Average <-15.
- Allows you to use attributes with missing values, where the example of unknown value is given the value that appears most in the other examples.
- Has the ability to use attributes with different weights.

III. Materials and Technicals

A. Materials

The data correspond to the academic records of 970 students belonging to the Professional School of Systems Engineering. The input variables considered for the study are described in Table I.

The output variables or classes in which a student will be classified are: for the first approach ABANDON AND NOT ABANDON and for the second approach are INCIPIENT RISK, MODERATE RISK, HIGH RISK and VERY HIGH RISK.

The application used to implement, execute and compare algorithms and their results was developed using the Java programming language, the MySQL database engine. JPA was also used, it is a useful Java framework for handling relational data.

B. Data Mining Techniques

In this research, the data mining techniques proposed for classifying students and predicting the risk of desertion (Abandon, Don't Abandon) are: artificial neural networks, C4.5 algorithm and ID3 algorithm.

TABLE. I. Variables and their Values

Variable	Data type	Value
Gender	Nominal	0 – Male 1 – Femaleo
Average Grades of Career Subjects	Numeric and Nominal	0 a 1 (Normalized with min=0 y max=20) Very bad (0 - 7), Bad (7-11), Regular (11-14), Good (14-16), Very good (16-20)
Average Grades General Subject	Numeric and Nominal	0 a 1 (Normalized with min=0 y max=20) Very bad (0 - 7), Bad (7-11), Regular (11-14), Good (14-16), Very good (16-20)
School of Origin	Nominal	1 - State 2 - Particular 3 - Parochial
Career Subject Credit Ratio	Numeric and Nominal	0 a 1
General Subject Credit Ratio	Numeric and Nominal	0 a 1
Admission test score	Numeric and Nominal	0 a 1 (Normalized with min=0 y max=100) Regular (0-50), Good (50-75), Very Good (75-100)
Admission type	Nominal	1 – Ordinary 2 – Extraordinary
The credit relationship the student has with respect to those he should have	Numeric and Nominal	0 - 1
Number of Abandoned Subjects	Numeric and Nominal	None, One, Two, More than two
Number of Disapproved	Numeric and Nominal	None, One, Two, More than two

C. Techniques for Evaluating Classifiers

The evaluation of classification techniques is important, as it allows to validate the goodness of fit of the model in relation to the training set. In addition, the evaluation allows to compare between several classification techniques and select the one that provides the most precision. For the evaluation of the data mining techniques used in this work, 3 performance measures were calculated: accuracy, precision and sensitivity.

IV. The Proposal

A. Knowledge Discovery in Databases (KDD)

For the development of this proposal was followed the process of Knowledge Discovery in databases that allows to identify valid and potentially useful patterns to generate knowledge and make decisions from it. The steps in this process are shown in Fig. 1.

B. Data Selection and Pre-Processing

It was used as input the data of students from the Professional School of Systems Engineering of the National University of San Agustín (UNSA) in Arequipa, Peru. The information was provided by the UNSA Institute of Computing, using 3 files with an .xlsx extension. Archives contained:

- Basic information of students such as: first and last names, C.U.I. (unique identification code), date of birth, gender, and type and place of the school of origin.
- Record of final grades of students in the subjects considered in the curriculum. It is considered the name and code of the subject that a student took (identified by his C.U.I.), the group, the enrollment number, the grade, the condition at the end of the course (approved, disapproved or abandoned the subject), the year and academic cycle in which the student took the subject.
- Score obtained by the student in the entrance test, the type of admission and the position in which he/she entered.

The data in the training set must be pre-processed before being evaluated by the algorithms. Initial data were for 970 students, data cleansing was performed to eliminate noise and isolated data or some inconsistencies. For example, there were students who did not have information about the entrance exam or lacked information about the school of origin or the subjects in which they were registered. Those cases were removed from the final dataset. Some typographical or formatting errors were also fixed. The pre-processed data set consisted of 451 students, of which 338 students were taken, representing 75% to be able to train and generate the ranking models. Of these 338 students, there were 215 instances with the class "DO NOT ABANDON" and 123 instances with the class "ABANDON". The remaining 25%, which is equivalent to 113 students were used for the trial phase, where 88 of these instances have the "DO NOT ABANDON" class and 25 with the "ABANDON" class.

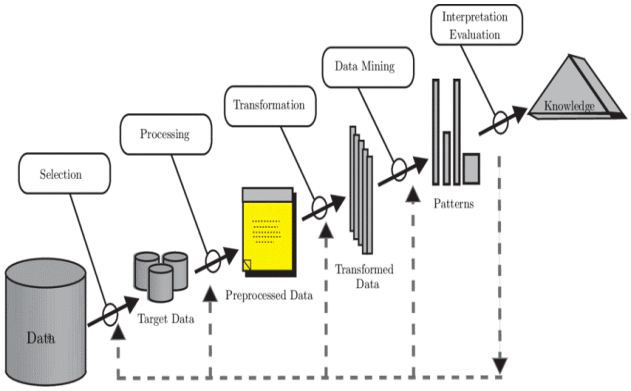

Fig. 1. Steps to Extract Knowledge from the Data (Source: Baradwaj [12]).

For the second approach, a training and testing set consisting of those instances of the previous set of 338 students where the already known class is "ABANDON" was considered to determine the type of risk for each: INCIPIENT RISK, MODERATE RISK, HIGH RISK AND VERY HIGH RISK. Because of this, a set of 160 student instances was used for this approach, 137 belong to the training set and 23 to the test set. Fig. 2 shows a part of this last set (the information appears in Spanish, as it is the language in which the information was worked on).

One point to keep in mind when splitting datasets is that it has enough information about each of the input variables and their classes but that it's also not too much not to over-adjust the training algorithm so it requires to be found adequate proportions. Fig. 3 and 4 show statistical graphs on the percentage of occurrence of each class of a variable in the set and the number of students belonging to a variable class, respectively.

C. Data Transformation

Model training requires inputs to the values of the variables corresponding to each of the students within the training set. The average grades obtained by a student is one of the most commonly used variables in other research around academic performance. Other variables were also considered: the number of credits that the student should possess at the end of a semester, number of subjects disapproved or abandoned during the semester, type of school of origin, gender and notes related to his admission to college.

	A	B	C
1	CUI	CLASE	TIPO
2	20110189	Riesgo moderado	ENTRENAMIENTO
3	20110343	Riesgo moderado	ENTRENAMIENTO
4	20110589	Riesgo incipiente	ENTRENAMIENTO
5	20110729	Riesgo incipiente	ENTRENAMIENTO
6	20111426	Riesgo bajo	ENTRENAMIENTO
7	20111427	Riesgo incipiente	ENTRENAMIENTO
8	20111428	Riesgo incipiente	ENTRENAMIENTO
9	20111429	Riesgo moderado	ENTRENAMIENTO
10	20111430	Riesgo moderado	ENTRENAMIENTO
11	20111431	Riesgo incipiente	ENTRENAMIENTO
12	20111432	Riesgo incipiente	ENTRENAMIENTO
13	20111433	Riesgo moderado	ENTRENAMIENTO
14	20111434	Riesgo moderado	ENTRENAMIENTO

Fig. 2. Data File Provided to the System (Student Representation, Exit Class, and Set Type).

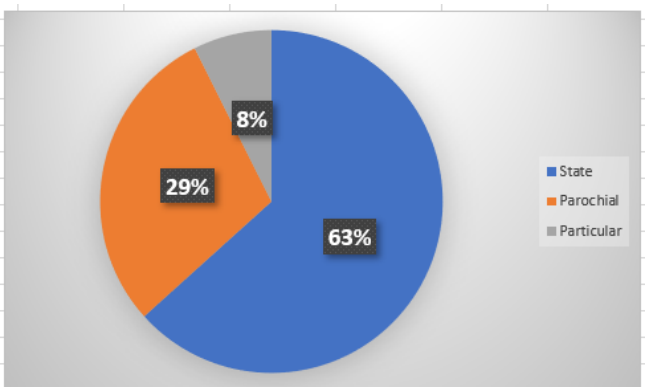

Fig. 3. Data Sharing for the Variable "Scool of Origin" in the Training Set.

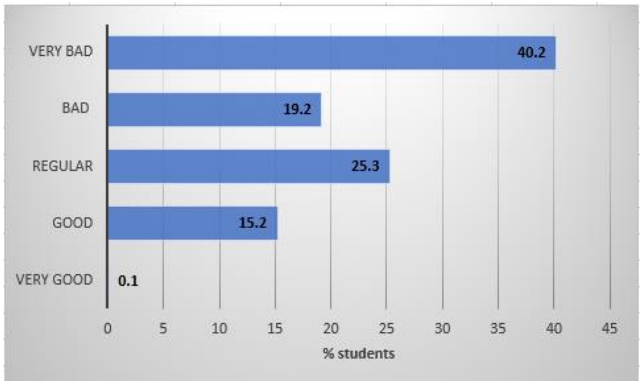

Fig. 4. Data Sharing for the "Califications Average" in Training Set.

To generate classification models for each of the implemented algorithms (neural networks, ID3 and C4.5) the values of the mentioned variables must be provided as inputs, however; each algorithm receives these values differently so it is necessary to transform these values to an appropriate representation.

For example, in the case of the ID3 algorithm, it can only work with nominal attributes, i.e. if working with the gender variable, the values 'FEMALE' and 'MALE' would be used. On the other hand, the C4.5 algorithm supports both nominal and continuous data, i.e. for the case of the average note variable could work with the continuous values which in this case would be a decimal value between 0 and 20; or use nominal or categorical values as very good (16 to 20), good (14 to 16), regular (11 to 13), bad (7 to 10), very bad (0 to 6). For neural networks it is necessary to normalize the data to values between 0 and 1 due to the activation function used, sigmoidal, which has a range of values between 0 and 1.

D. Training Phase

After having the transformed data, the models are trained. The system was provided with a file with the input variables (gender, average notes, credit ratio, etc.) and a field classified as output (abandon or not) of the model. It was also provided with a file listing the student codes that made up the training set along with the class tag to which they belonged, a classification that was previously made based on real information. This action is justified in the training being performed using known data and when the actual classification is performed, it is done on unworked data either in the training

phase or in the test phase, so the results thus obtained are a prediction.

A matrix of Student/Variables is then constructed which contains, for each student within the dataset, the value of each of the selected variables (Table II). The Student/Variables matrix serves as input to the classification algorithm that will initiate your training.

If the chosen option is a neural network, a multilayer perceptron network trained with the backpropagation algorithm is built. To stop the training, two different criteria were established: reach an error rate of 0.01% or reach a maximum of 100000 iterations. At the end of the training, the weights of the connections of the neurons were saved in a file, which would allow the network to be reassembled in case the model wanted to be used again.

In the case of the ID3 algorithm, the entropy and information gain of each attribute is calculated, and in this way the most dominant attribute can be found. The most dominant attribute is placed in the tree as a decision node. After that, entropy and profit scores would be calculated again among the other attributes. Thus, the next most dominant attribute is found. This procedure continues until a decision is reached for that branch. Once it is built tree, it is applied for a tuple in the database and this results in the classification for that tuple.

For C4.5 training is very similar to ID3, but in this case the gain ratio of each attribute is calculated.

The output variables will depend on the approach used where they will be "ABANDON" and "DO NOT ABANDON" if try to determine the desertion of students; whereas it will be classified into the four different types of risk if the second approach is used.

After the training phase, the respective tests must be performed with the test set. For each trained model, 3 performance measures were calculated: accuracy, precision and sensitivity.

To perform the training phase, the system was provided with the set of records that belonged to the test suite. Similarly to the training phase, the class to which the student belongs was specified the goal was to apply the classification model to each record and compare the output with the actual class to which the record belongs and apply the performance measures as it is shown in Fig. 5, where in the Student/Variable matrix, it has a column indicating the actual class to which the student belongs and another column indicating the classification made by the model used. Performance metrics are subsequently generated.

TABLE. II. STUDENT/VARIABLE MATRIX

Student	**Variable**			
	V1	**V2**	**…**	**Vm**
E1	E1V1	E1V2	…	E1Vm
E2	E2V1	E2V2	…	E2Vm
E3	E3V1	E3V2	…	E3Vm
…	…	…	…	…
En	EnV1	EnV2	…	EnVm

CUI	GENDER	CAL AVERAGE	CREDITS RELATION	SCHOOL ORIGIN	ADMISIO N SCCORE	CLASIFICA TION	RESULT
20110189	FEMALE	12.0	92.0	PARTIC	65.0	NO ABAN	NO ABAN
20110343	MALE	12.0	80.0	PARTIC	68.0	NO ABAN	NO ABAN
20110689	FEMALE	12.0	71.0	PARTIC	64.0	NO ABAN	NO ABAN
20110789	MALE	9.0	9.0	STATE	60.0	NO ABAN	NO ABAN
20111426	MALE	13.0	100.0	STATE	65.0	NO ABAN	NO ABAN
20111325	MALE	12.0	76.0	PARROQ	58.0	ABAN	ABAN

ACCURACY	PRECISION	SENSITVITY
0.831858407079646	0.87777777777778	0.9080459770114943

Fig. 5. Classification Performed on Test Set and Performance Metrics.

Considering the values obtained for the established performance measures, it is possible to determine which model is the most appropriate when classifying.

V. RESULTS AND DISCUSSION

As shown in Fig. 6, six different models were implemented. The structure of the model was stored in the database along with the information of its performance metrics, the sizes of the training and test sets; in addition to the variables that were used.

A different model was first created for each algorithm (neural network, ID3 and C4.5) where performance metrics were obtained that were around 80% minimum. For these models, the training and test set of 338 and 113 instances, respectively, was used. These models were aimed at classifying students according to whether or not they abandoned the career.

Fig. 7 shows the decision tree built by the C.45 classifier, which is the algorithm with higher performance measures with respect to accuracy, precision and sensitivity. From this tree model it is possible to determine which credit ratio was the most significant variable, followed by the admission exam score and the average of grades, on the other hand, the type of admission exam is not included within the tree by which is not a significant variable for the model. Preliminary results, related to this approach, have been presented in [13].

Three other models were implemented to classify students according to their type of risk: Incipient Risk, Moderate Risk, High Risk and Very High Risk.

First, the model for the two decision tree algorithms was constructed, neural networks were excluded for this case due to the transformation of the data since the attributes had very varied ranges of values and the different values were not going to be significant.

NAME	MODEL	ACCURACY	PRECISION	SENSITIVITY
NETWORK MODEL	NEURONAL	0.796460176911505	1.0	0.79646017699115
ID3 DESERTION	ID3	0.823008849557221	0.822222222222	0.94871794871794
C45 DESERTION	C4.5	0.831858407079646	0.877777777778	0.90804597701149
ID3 RISK TYPE	ID3	0.585217391304347	0.585217391304	1.0
C45 RISK TYPE	C4.5	0.652173913043478	0.652173913043	1.0
ID3 RISK TYPE	ID3	0.707964601769911	0.707954601769	1.0

Fig. 6. Classification Models Generated.

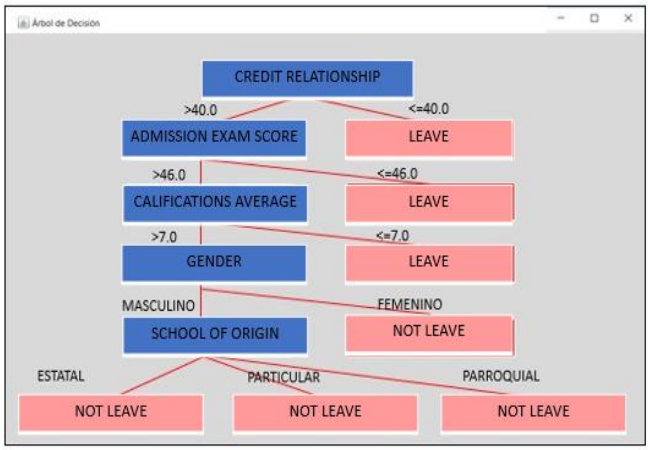

Fig. 7. Decision Tree Generated by the Application to Classify Student Dropout.

To generate these models, only those students who had the "ABANDON" class were filtered from the previous data set, which reduced the test set to 137 and 23, respectively.

The results in the metrics were not very good showing values between 56 and 65%. Analyzing the data, there was a great difference in the distribution of the values of the variables and because of this it did not train the model correctly.

Fig. 8 shows the low percentage of cases for some classes of the variable "DISAPPROVED SUBJECTS" where the greatest distribution is in the class of "NO SUBJECT DISAPPROVED".

A last model was built using the ID3 algorithm and the same input and output variables as in the previous approach with the difference that the dataset was used for the first 3 cases. This time there was a better distribution of training and testing examples, which yields of 0.77, 0.76 and 0.94 were obtained for accuracy, precision and sensitivity, respectively.

Unlike the first approach, in this one the most significant variable is the number of ABANDON SUBJECTS followed by the DISAPPROVED SUBJECTS.

Fig. 9 shows the decision tree generated by the application to classify students according to the types of risk previously defined.

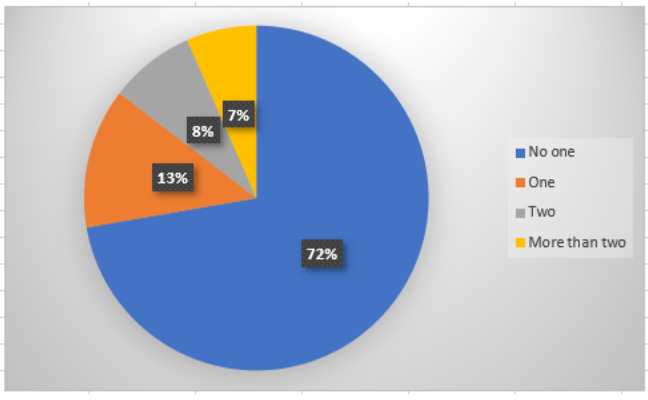

Fig. 8. Distribution of Classes of the Disapproved Subjects (Training Set: 113 Registers).

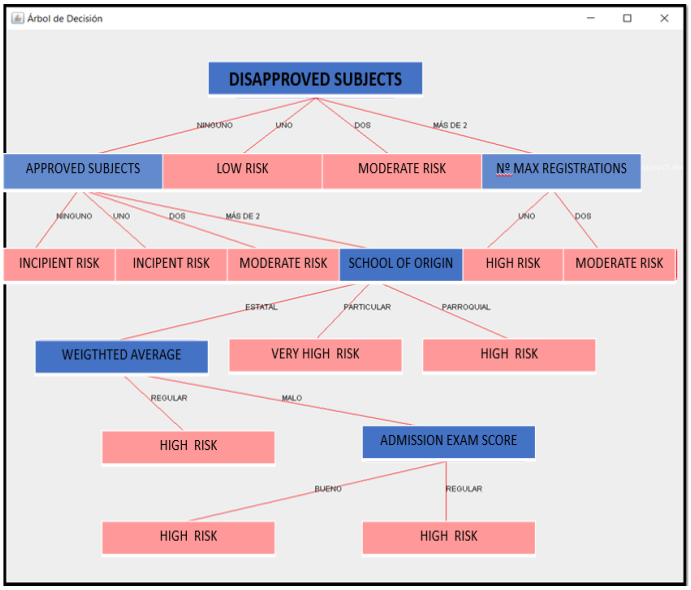

Fig. 9. Generated Decision Tree to Classify Risk Type.

Data Mining Techniques have been shown to be effective tools for obtaining models to predict the permanence of students enrolled in an engineering career, as well as to more specifically determine the risk of dropping out of the university.

VI. Conclusions

With the development of this work, it has been possible to determine those factors that affect academic performance, for this purpose different variables of the personal and academic type were used for two different approaches.

The first approach was used to determine student desertion and three classification models were created (one per algorithm), for which the C4.5 algorithm presented performance improvements with respect to the neural network and the ID3 algorithm (because ID3 did not you can work with continuous data).

A student's current credit ratio with respect to the credits they should possess turned out to be the most significant variable in the construction of the model, followed by the notes, while the type or modality of admission exam with which the student he entered college did not turn out to be significant as he did not appear in the decision tree generated.

The second approach was worked with two different types of training and testing sets. However, the same entry variables were worked on (number of abandoned subjects, number of subjects disapproved, weighted average, admissions test score, home school and maximum number of failed enrolments in a subject). It was also worked with the same output variables (incipient, moderate, high and very high risk).

This second approach achieved lower metrics than using the first approach. A misdistribution was determined in the different values of the student attributes for each set so the experiment was re-performed using the initial dataset which had a better distribution in the data and metrics of acceptable performance to be able to classify the type of risk. The

generated tree determined that the most significant variable was the number of abandoned subjects. The performance measures obtained could be improved, as well as the rate of success of the classification of the models by increasing new variables in the future, such as social, economic variables, etc.

VII. Future Work

To expand the results of the research it is proposed to consider institutional and socio-economic variables. At the same time, increasing the number of variables should consider working with a much larger data set, taking information from students from more professional schools at the university.

The elaborate system can be adapted to create models with different input and output variables (classes). This was demonstrated as different variables were used for each approach for both input and output variables (two variables for the first approach and four variables for the second).

The best performing models can be used in several professional schools or each school uses a different classification model that best suits its needs.

Acknowledgment

This work was carried out with the support of our house of studies, the National University of San Agustin, in which the Vice-Chancellor of Research channels the resources from the mining canon and convenes a set of insolvable financial schemes.

It is through one that the IBA 004-2016 project was funded, "Model of Academic Performance Assessment for the Detection of Outstanding Students and Students at Academic Risk".

References

[1] I. Mushtaq and S. Nawaz. “Factors Affecting Students’ Academic Performance”. Mohammad Ali Jinnah University Islamabad, Pakistan. 2012.

[2] S. Oloruntoba and J. Akinode. “Student Academic Performance Prediction using Support Vector Machine”. Computer Science Department, the Federal Polytechnic,ILARO,Ogun State, Nigeria. 2017.

[3] S. Sumam and J. Puthiyidam, “Evaluating Students Performance by Artificial Neural Network using WEKA”, International Journal of Computer Applications (0975 - 8887), June 2015.

[4] I. Ganiyu. “Data Mining: A Prediction for Academic Performance Inprovement of Science Student using Classification”. 2016.

[5] D. Kabakchieva, “Student Performance Prediction by Using Data Mining Classification Algorithms”, University of National and World Economy Sofia, Bulgaria, November 2012.

[6] S. Anupama Kumar and M. Vijayalakshmi, “Efficiency of Decision Trees in Predicting Student’s Academic Performance”, India 2011.

[7] S. Sharma, J. Agrawal, S. Sharma, “Classification Through Machine Learning Technique: C4.5 Algorithm based on Various Entropies”, International Journal of Computer Applications 82(16):28-32, October 2013. DOI: 10.5120/14249-2444.

[8] B. Baradwaj and S. Pal, “Mining Educational Data to Analize Students’ Performance”. International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6. 2011.

[9] C. Menacho, “Predicción del rendimiento académico aplicando técnicas de minería de datos”, Universidad Nacional Agraria La Molina, Lima – Perú, 2017.

[10] N. Zacharis, “Predicting Student Academic Performance in Blended Learning Using Artificial Neural Networks”. International Journal of Artificial Intelligence and Applications (IJAIA), Vol. 7, No. 5. 2016.

[11] K. Manchandia and N. Khare, “Implementation of Student Performance Evaluation through Supervised Learning Using Neural Network”. 2017.

[12] R. Bhardwaj and S. Vatta, “Implementation of ID3 Algorithm”. CSE, Bahra University, India, International Journal of Advanced Research in Computer Science and Software Engineering, Vol.3, June 2013.

[13] N. Bedregal-Alpaca, V. Cornejo-Aparicio and D. Aruquipa-Velazco, “Analysis of the academic performance of Systems Engineering students, desertion possibilities and proposals for retention”, INFONOR 2019 : 10th International Conference on Computing and Informatics in Northern Chile, 2019. In press.