Intelligent System to Predict University Students Dropout

https://doi.org/10.3991/ijoe.v18i07.30195

Hugo Vega(✉), Enzo Sanez, Percy De La Cruz, Santiago Moquillaza, Johny Pretell
Universidad Nacional Mayor de San Marcos, Lima, Peru
hvegah@unmsm.edu.pe

Abstract—The objective of this research is to reduce the dropout rate of students in the Faculty of Systems Engineering and Informatics of the Universidad Nacional Mayor de San Marcos (FISI-UNMSM), through the implementation of an intelligent system with a data mining approach and the autonomous learning algorithm (decision trees) that predicts which students are at risk of dropping out. It was developed in Python and the free software Weka. For this, the data of the students who entered the faculty from 2004 to 2014 have been considered. This solution increases the availability and the level of satisfaction of the faculty; in the learning process, an accuracy percentage of 90.34% and precision of 95.91% was obtained, so the data mining model is considered valid. In addition, it was found that the variables that most influenced students in making the decision to abandon their studies are the historical weighted average their grades, the weighted average their grades of the last cycle, and the number of credits of their approved courses

Keywords—intelligent system, machine learning, prediction, dropout

1 Introduction

Nowadays, students have great difficulties to carry out their studies, such as, economic problems, restrictions on access to the internet, family problems, among others. Universities aim to create knowledge and be able to transmit it to students to promote critical thinking and scientific research. The purpose of the student is often frustrated when the student, due to various factors, abandons his university studies, this abandonment is called "student desertion".

Figure 1 shows that according to the Ministry of Education of Peru (MINEDU) [1], in 2020 more than 300,000 university students abandoned their studies. The university dropout rate reached 18.6% in 2020, an indicator that is six points higher than that registered in 2019, equivalent to 12%.

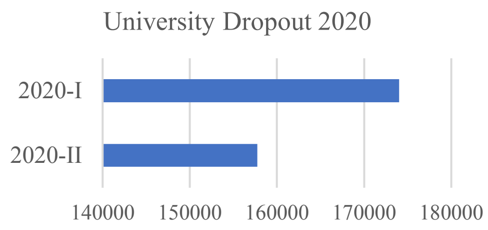

Fig. 1. University dropout in Peru in 2020 [1]

Student dropout at FISI-UNMSM can be evidenced by the low number of students who graduate in relation to the number of students who enter to the University. Table 1 shows the numbers of entrants versus graduates of FISI-UNMSM's systems engineering career from 2004 to 2014.

Table 1. Entrants vs. Graduates at FISI-UNMSM [2]

Year	Entrants	Graduates	%Graduates	%Dropout
2004	201	177	88%	12%
2005	196	144	73%	27%
2006	212	112	53%	47%
2007	200	115	58%	43%
2008	200	74	37%	63%
2009	156	115	74%	26%
2010	156	138	88%	12%
2011	147	120	82%	18%
2012	159	90	57%	43%
2013	150	117	78%	22%
2014	161	98	61%	39%
Total	1938	1300	68%	32%

Note. Adapted from UNMSM General Planning Office (2004-2014).

As we can see in Figure 2, from 2004 to 2014, 1938 students entered the faculty, of which only 1300 completed their studies, that is, 638 students, equivalent to 32% did not finish their degree. These data show the deficient process of detecting student dropout patterns at FISI-UNMSM. In summary, according to [1] in 2020 in Peru there was an increase in university dropout of 6% compared to 2019, while in the FISI-UNMSM in the period 2004 - 2014 the dropout rate was 32%.

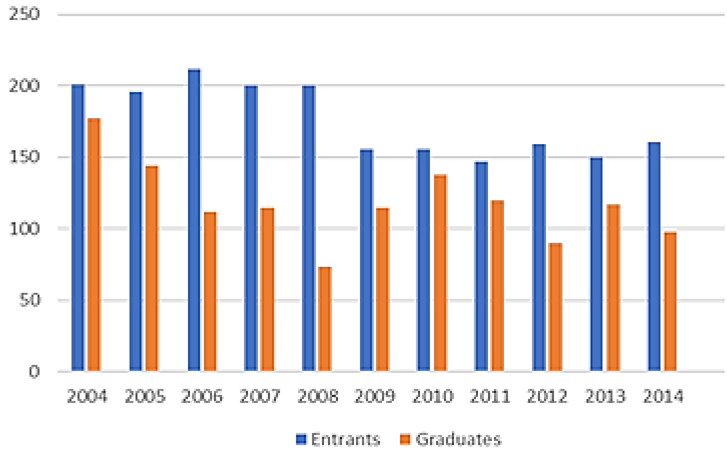

Fig. 2. Entrants vs. graduates at FISI-UNMDM (2004-2014)

2 Background

Table 2 shows the results of a study conducted by [2] which compared different autonomous learning techniques to achieve a model to predict students with a high probability of desertion, in which the J48 decision tree algorithm obtained an accuracy percentage of 94.3%. WEKA's J48 is an implementation of the C4.5 algorithm (used to generate a decision tree) that uses the concept of information entropy for the selection of variables that best classify the studied class.

Table 2. Performance of autonomous learning algorithms

Algorithm	TP rate	TN Rate	Accuracy	GM
Jrip	96.2	93.3	96.0	94.6
OneR	96.1	70.0	93.7	80.5
Prism	99.5	69.7	94.4	54.0
ADTree	98.1	81.7	96.6	89.0
J48	95.7	80.0	94.3	87.1
SimpleCart	97.2	90.5	96.6	93.6

Note. True Positive (TP), True Negative (TN), Geometric Mean (GM)

In turn [3] states that there are 2 problems with datasets containing student data. The problem of high dimensionality, because when there are many attributes, some of them may not be significant for classification and some of the attributes are likely to be correlated, and the problem of unbalanced data because often little importance is given to the minority or low-frequency class (dropouts). The result is that the classifier may not classify the data instances in a proper way.

Regarding the variables that influence students, [4] shows that the transformation of continuous variables into discrete variables considerably improves the effectiveness of autonomous learning algorithms because they work better with discrete variables.

There are many methods developed to predict the university student dropout such as [5], that using Sas Enterprise software could predict with a percentage of 70% the students who tended to drop out. These methods are very expensive because the Sas Enterprise license is high, and that is why we chose to use the Weka software, an open-source software developed by the University of Waikato that contains algorithms for data analysis and predictive modeling, as well as a graphical user interface to access its functionalities easily and quickly.

3 Previous concepts

3.1 Student dropout

For [6] desertion is a personal decision, which can be the result of factors related to the perceptions and feelings of the student, but which can also be the result of factors determined by the socio-economic environment in which he develops his daily activities, in which permanently or temporarily causes the student to leave the university classrooms, regardless of the effect it has on their future life. According to [7], student desertion is understood also like as the abandonment of academic training, regardless of the conditions and modalities of attendance, is the personal decision of the subject and does not obey a forced academic retirement (expulsion for low academic average) or withdrawal for disciplinary reasons.

According to [8], the lack of technological tools such as online support tools for students, both academic and psychological and everything related to the digital environment, generates impotence in students becomes a key factor of dropout.

For [9] the most important causes of the phenomenon of student dropout are: Personal, when individuals are not mature enough to manage the responsibilities that the university entails, they do not have a certainty that the degree chosen at first is really the desired one and / or do not identify with the university in which they are studying; Socio-economic, referring to the lack of resources, absence of scholarship programs or limitations for access to them; Institutional and pedagogical, linked to the lack of an institutional policy of induction, for the student, to the new system of higher education, as well as to the lack of vocational guidance before entering a bachelor's program; affective, referring to motivation, personal situations and health problems they face.

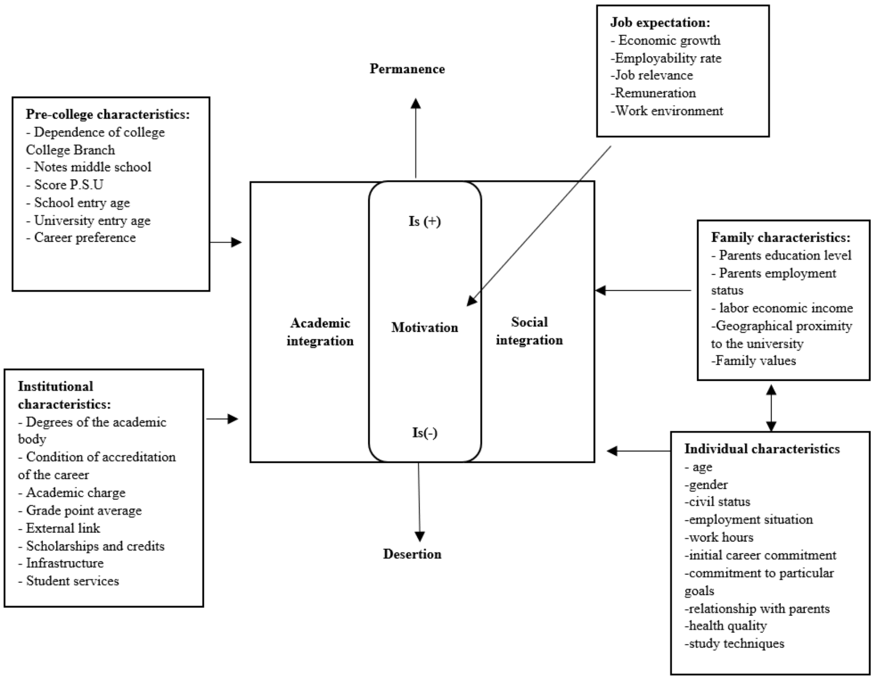

Fig. 3. Factors that most influence university student dropouts [9]

3.2 Data mining|

According [10] it is understood as data mining as the process in which relevant patterns are found from the extraction of large amount of data with previously unknown information, understandable to process them and make strategic business decisions. For the author, the data mining process consists of six important steps that are described in Figure 4.

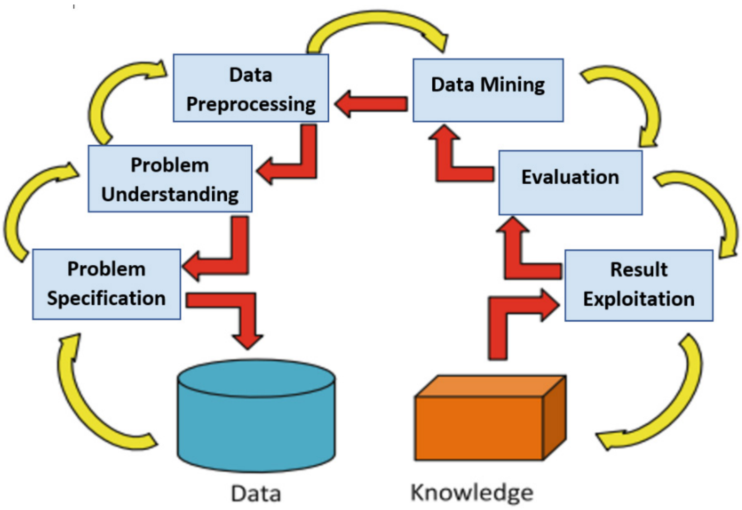

Fig. 4. Phases of data mining

Problem specification. In this stage, the domain of the application, the previous relevant knowledge obtained by the experts and the final objectives pursued by the end user are designated and organized.

Problem understanding. This stage includes the understanding of the selected data approach and the associated expert knowledge in order to achieve a high degree of reliability.

Data preprocessing. This stage includes operations for data cleaning (such as handling noise removal and inconsistent data), data integration (where multiple data sources can be combined into one), data transformation (where data is transformed and consolidated into forms that are appropriate for specific data mining tasks or aggregation operations) and data reduction, including selecting and extracting features and examples from a database.

Data mining. It is the essential process where methods are used to extract valid data patterns. This step includes choosing the most suitable data mining task. (Such as classification, regression, clustering, or association), the choice of the data mining algorithm itself, belonging to one of the previous families, and finally, the use and accommodation of the selected algorithm for the problem, by adjusting essential parameters and validation procedures.

Evaluation. In this stage, the extracted patterns are estimated and interpreted based on measurements of interest.

Result exploitation. The last stage may involve the direct use of knowledge, in this stage the knowledge is incorporated into another system for subsequent processes or simply report the knowledge discovered through visualization tools.

3.3 Autonomous learning

According to [11], autonomous learning is based on statistical models that use computer systems to generate a specific task without being cleanly programmed. This autonomous learning is based on different algorithms to solve problems, each one depends on the conditions of the problems you are trying to solve, the number of variables, the environment, among other factors. Figure 5 shows the different existing autonomous learning algorithms.

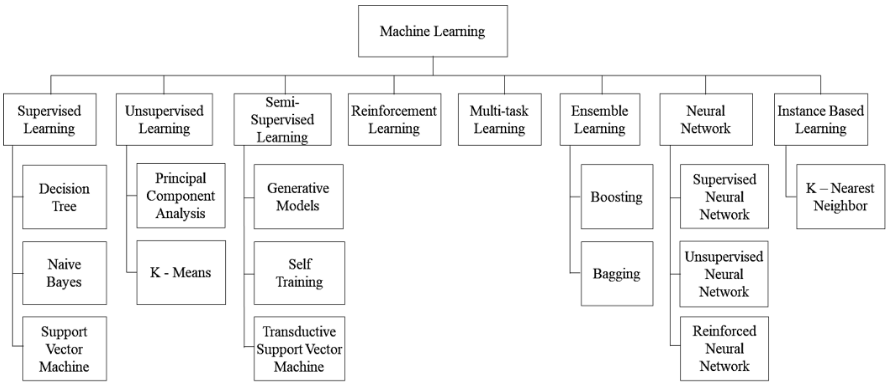

Fig. 5. Autonomous learning algorithms [11]

3.4 Decision trees

According to [12] the decision tree is a classifier with a spatially instantiated recursive partition. It consists of nodes that form the root of the root tree, the other nodes have exactly one leading edge. The node with protruding edges is called the internal node, the other nodes are the leaves. Each test node of the tree is used to partition each instantiated space into two or more subspaces bound to a discrete function with respect to its input values. For the simplest cases each test uses only one attribute, so the instantiated space is divided according to the value of the attribute, but in the multiple attribute case the condition is bound to a range. Each sheet corresponds to a class that represents the best target value. The sheet may contain the value of target probability. See Figure 6.

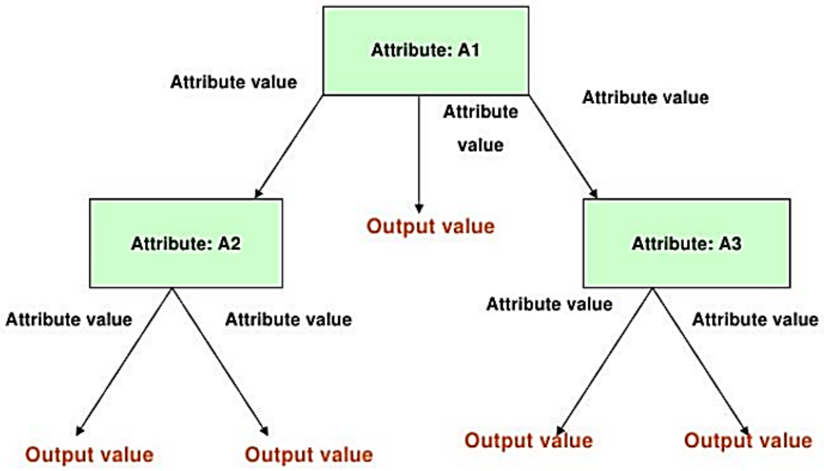

Fig. 6. Decision tree model [12]

4 Methodological framework

4.1 Data collection and integration

For the development of the study, the enrollment unit of the Faculty of Systems Engineering provided demographic and academic data of the students corresponding to the last 6 years (from 2004 to 2014), said information consisted of 1300 instances with 40 attributes, which corresponded to information recorded by the student at the time of enrollment, as well as socio-economic survey that it carries out and some that is generated during its study cycle such as the weighted average, previous average, etc. Table 3 shows the 40 attributes of this information, 32 of which are demographic attributes and 8 are academic attributes.

Table 3. List of attributes of the information collected

Attribute	Description
cod_alumno	Student Code
ape_paterno	Last name
ape_materno	Middle Name
nom_alumno	Student Name
dir_ubi_dist	District of residence
dir_ubi_prov	Province of residence
dir_ubi_depa	Residence Department
tel_alumno	Student landline
coe_alumno	Student Email
did_alumno	Student ID
tel_alu_movil	Student Cell Phone

sex_alumno	Sex of the student
est_civil	Civil Status of the Student
Age	Student Age
col_procedencia	University of origin
cambio_residencia	Change of residence?
dep_padre_tutor	Dependency on parents/guardians
num_hijos	Number of children of their parents
situ_laboral	Employment Status
des_vactual	Family you live with
des_tip_vivienda	Type of Housing
re_transporte	Do you have your own transport?
re_libro_estudio	Do you have study books?
re_dinero_alimentacion	Do you have money to feed yourself?
re_acc_internet	Do you have internet access?
des_tip_transporte	Type of transport
tiempo_transporte	Time to get to university
Disability	Disability you have
regimen_ssocial	Type of Insurance
otro_idioma	Second language
nivel_otro_idioma	Second language level
cod_plan	Curriculum
ppd_hist	Historical Weighted Average
Situation	situation
anio_ingreso	Year of entry
anio_estudio	Years of study at university
creditos_aprob	Total appropriations approved
promedio_anterior	Average of the last cycle
ultima_matricula	Last Registration

To improve the accuracy of the algorithm, some attributes were transformed from numerical to categorical. For example, Table 4 shows the classification of the Weighted Average attribute according to its category.

Table 4. Categorical weighted average

Category	Weighted Average
D1	20 - 13.142
D2	13.141 - 12.322
D3	12.321 - 11.561
D4	11.560 - 0

4.2 Solving the high dimensionality problem

To deal with the problem of high dimensionality, the selection of contents is carried out in the preprocessing and data processing phase, which consists of eliminating the less relevant attributes. For this the dataset was loaded into the Weka software and variable selection algorithms were used. We used 6 attribute selection methods that are available in version 3.8.5. of the Weka. Table 5 allows us to appreciate the most relevant attributes were selected from a total of 40 attributes presented after the application of the six algorithms indicated above.

Table 5. Variables obtained by each selection method applied

CfsSubsetEval	ChiSquaredAttributeEva1	OneRAttributeEval	GainRatioAttributeEval	lnfoGainAttributeEval	ClassifierAttributeEval
prom_hist	prom_hist	prom_hist	prom_hist	prom_hist	prom_hist
prom_ant	prom_ant	prom_ant	prom_ant	prom_ant	prom_ant
Situación	edad	anio_estudio	situacion	edad	Edad
creditos_aprob	creditos_aprob	edad	creditos_aprob	creditos_aprob	est_civil
dir_ubi_prov	regimen_ssocial	dir_ubi_dist	re_transporte	regimen_ssocial	col_procedencia
sex_alumno	situacion	creditos_aprob	nivel_otro_idioma	situacion	num_hijos
col_procedencia	situ_laboral	situacion	re-libro-estudio	situ_laboral	cambio_residenc
dep_padre_tutor	col_proc_edencia	regi men ssocial	anio_estudio	col_procedencia	sex_alumno
num_hijos	dir_ubi_dist	tiempo_transport	re_dinero_aliment	des_vactual	dir_ubi_depa
situ_laboral	re_transporte	situ_laboral	dep_padre_tutor	dir_ubi_dist	dir_ubi_prov
re_transporte	des_tip_vivienda	des_vactual	regimen_ssocial	re_transporte	dir_ubi_dist
re_accinternet	dep_padre_tutor	col_procedencia	situ_laboral	des_tip_vivienda	situacion
des_tip_transport	re_libroestudio	re_transporte	sex_alumno	dep_padre_tutor	creditos_aprob
tiempo_transport		dep_padre_tutor	edad	re_libro_estudio	dep_padre_tutor
discapacidad					
regimen_ssocial					
nivel_otro_idiom			-		

The attributes were then classified according to the number of times they were chosen by the algorithms. Table 6 shows the respective frequencies.

Only the attributes that were chosen 3 times or more because they were the most relevant according to the Weka algorithms were considered, so the number of attributes per student was reduced from 40 to 17, thus solving the high dimensionality problem.

Table 6. Frequencies of the attributes obtained by each Weka selection method

ATTRIBUTE	FREQUENCY
creditos_aprob	6
dep_padre_tutor	6
prom_ant	6
prom_hist	6
Situacion	6
anio_estudio	5
col_procedencia	5
des_vactual	5
re_transporte	5
situ_laboral	5
des_vactual	4
dir_ubi_dist	4
re_libro_estudio	4
regimen_ssocial	4
des_tip_vivienda	3
Edad	3
sex_alumno	3
dir_ubi_prov	2
nivel_otro_idioma	2
num_hijos	2
tiempo_transporte	2
des_tip_transporte	1
dir_ubi_depa	1
Discapacidad	1
re_acc_internet	1

4.3 Solving the data balancing problem

The problem of unbalanced data arises because the majority class (students who did not drop out) is higher than the minority class (dropout students), which would cause the model to not be able to correctly predict the students who will drop out of the career. To solve this problem, the balancing algorithm (SMOTE) is used, which, according to [13], solves the balance problem between the majority class (students who do not drop out) and the minority class (students who drop out) creating individuals synthetics from the minority class.

4.4 Predictive model training

Once the data to be processed is obtained, it is divided into an input set (70%) and a validation set (30%). With the training set a predictive model is trained using the Decision Tree algorithm and with the validation set the performance of the model is evaluated according to the metrics of performance, accuracy, sensitivity, etc. Once the model is validated, a prediction is made with the data of the new students to determine if they are at risk of deserting. The results obtained show a corresponding prediction rate of 90.43%. Which means that the proposed model is adequate in terms of quality and effectiveness. See Table 7.

Table 7. Prediction results table

ACCURACY	PRECISION	TP	TN	FN
90.43	95.91	93.48	59.26	6.52

Note. TP: True positive, TN: True negative, FN: False negative

5 System implementation

5.1 Conceptual model

Figure 7 shows the conceptual model of data mining used in the project to predict student dropout based on data of FISI-UNMSM's students from 2004 to 2014. In this model the most important and significant processes from the registration of data to the development of the predictive model of student dropout are presented.

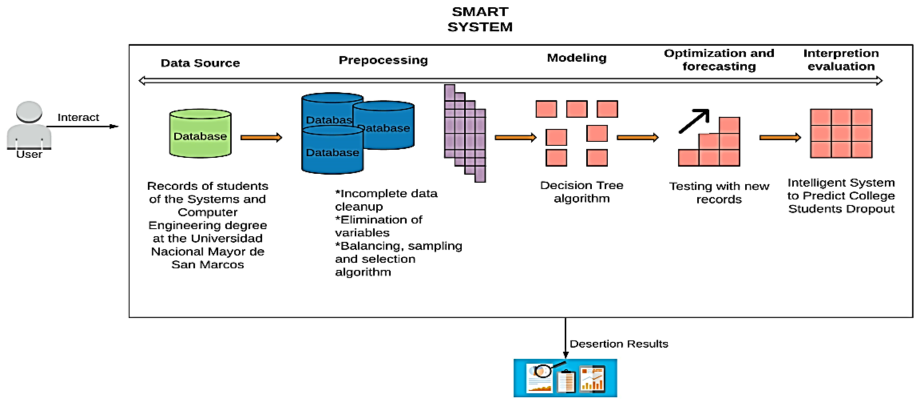

Fig. 7. Conceptual model of the project

5.2 System use case diagram

The project proposes a solution with several profiles that have permissions and functionalities so that they can carry out their corresponding processes. Figure 8 shows the use cases of the implemented system where the main processes carried out by users are appreciated, we can highlight the process of "Make prediction" available to the Administrative Staff.

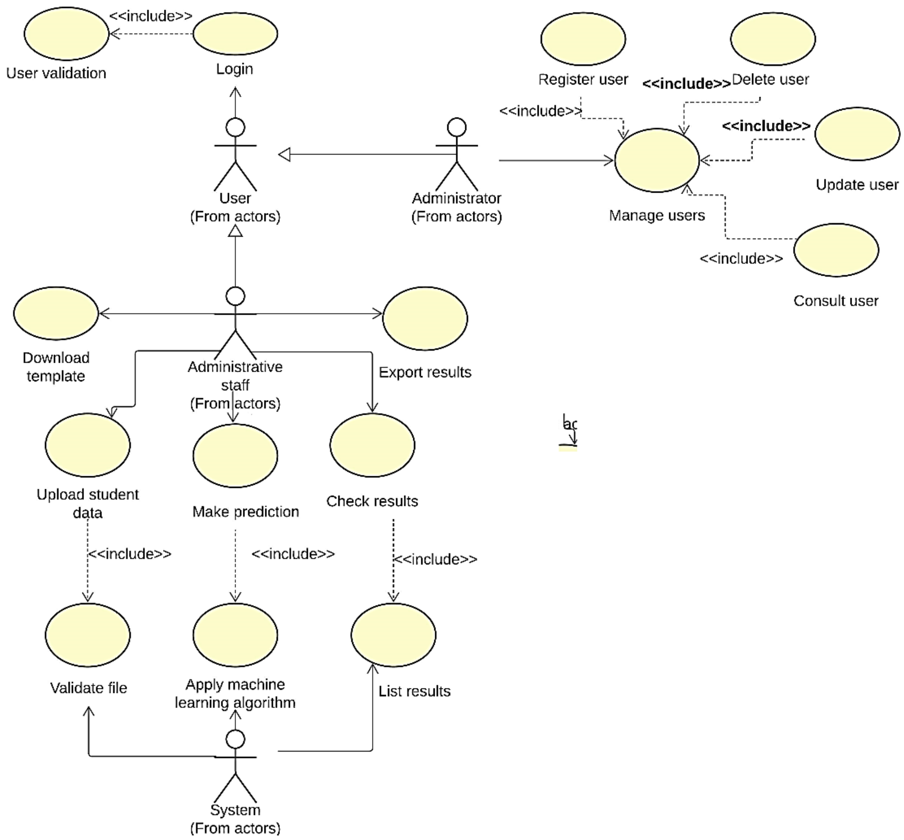

Fig. 8. System use cases

5.3 System architecture

Figure 9 presents the design of the architecture in 3 layers; this solution has been implemented in Python considering the multiple benefits that we find in its available libraries.

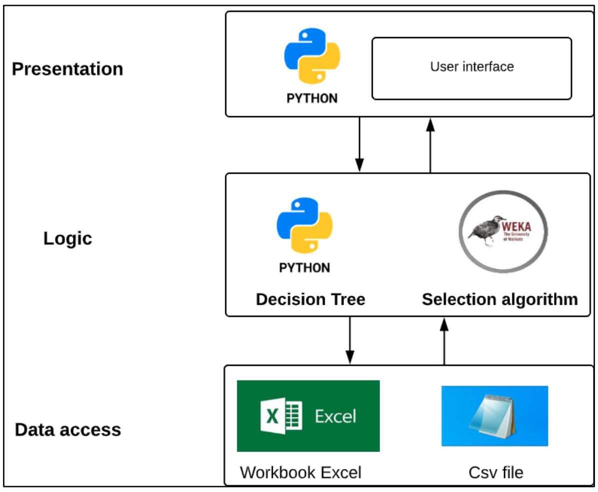

Fig. 9. System architecture

5.4 Prototypes

In [14] and [15] the authors present prototypes of informatic systems properly elaborated about similar approaches, therefore, we have taken these prototypes as a reference to develop the prototypes of our system. The Figure 10 and Figure 11 shows the most important prototypes of the processes presented in the Use Case Diagram in Figure 8 they are the Prediction Module and the Prediction Results Module.

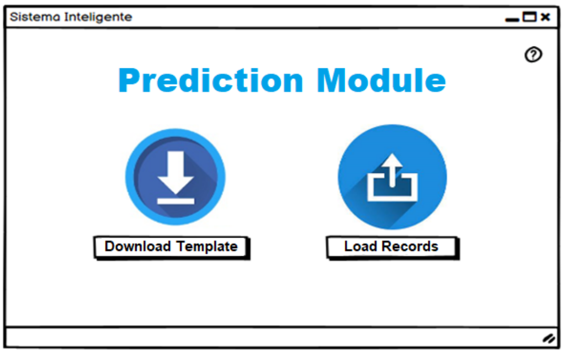

Fig. 10. Prediction module

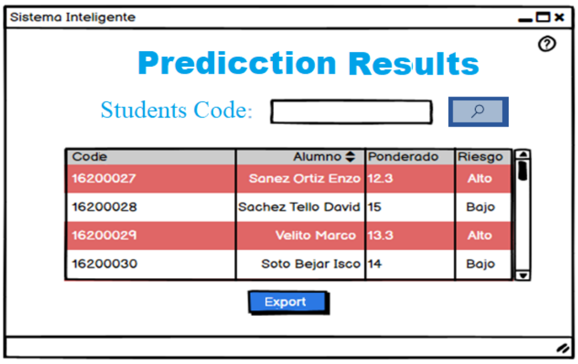

Fig. 11. Prediction results

6 Conclusions

- In this work it has been possible to implement an intelligent system that predicts with an accuracy of 90.43% and a precision of 95.91%, that students are at risk of dropping out.
- It has also made it possible to identify those factors that lead students to abandon the career. It was found that the most relevant factors for a student to tend to drop out are the historical weighted average their grades, the weighted average their grades of the last cycle, and the number of credits of their approved courses
- The information provided by the system is helping FISI-UNMSN to take preventive measures on students who are at risk of dropping out among which we can highlight the preventive analysis of personal and family problems to provide them with specialized academic tutoring, psychological, medical, economic, or other types of support.
- There are aspects to improve, which we can consider as future work, such as information security, the use of psychological variables, the application of encryption, the transfer of the system to the web, access protocols and the extension of the system to other faculties of the UNMSM.

7 Acknowledgments

We express our recognition acknowledgments and gratitude to the teachers who, as evidenced in [16] have been supporting the research work.

8 References

[1] Infomercado, "Minedu: 300 mil universitarios dejaron de estudiar en el 2020," *19/03/2021*, 2021. https://infomercado.pe/minedu-300-mil-universitarios-dejaron-de-estudiar-en-el-2020/

[2] B. Perez, C. Castellanos, and D. Correal, "Applying Data Mining Techniques to Predict Student Dropout: A Case Study," *2018 IEEE 1st Colomb. Conf. Appl. Comput. Intell. ColCACI 2018 - Proc.*, Oct. 2018. https://doi.org/10.1109/ColCACI.2018.8484847

[3] A. Pradeep, S. Das, and J. Kizhekkethottam, "Students dropout factor prediction using EDM techniques," in *IEEE Int. Conf. Soft-Computing Netw. Secur. ICSNS*, 2015, pp. 1–7. https://doi.org/10.1109/ICSNS.2015.7292372

[4] E. Drousiotis, P. Pentaliotis, L. Shi, and A. Cristea, "Capturing Fairness and Uncertainty in Student Dropout Prediction – A Comparison Study," *Artif. Intell. Educ.*, vol. 2, no. 2, pp. 139–144, 2021. https://doi.org/10.1007/978-3-030-78270-2_25

[5] R. Alkhasawneh and R. Hobson, "Modeling student retention in science and engineering disciplines using neural networks," in *(EDUCON), IEEE Global Engineering Education Conference*, 2011, pp. 660–663. https://doi.org/10.1109/EDUCON.2011.5773209

[6] J. C. Poveda Velasco, I. M. Poveda Velasco, and I. A. España Irala, "Análisis de la deserción estudiantil en una universidad pública de Bolivia," *Rev. Iberoam. Educ.*, vol. 82, no. 2, pp. 151–172, 2020. https://doi.org/10.35362/rie8223572

[7] G. Paramo and C. Correa, "Deserción estudiantil universitaria. Conceptualización," *Revista Universidad EAFIT*, vol. 35, no. 114. pp. 65–78, 1999.

[8] R. Saldarriaga, H. Vega, C. Rodriguez, and P. De La Cruz, "Academic approach about E-learning modules from the teacher/student perspective at the National University Mayor de San Marcos, Lima-Perú," *3C TIC Cuad. Desarro. Apl. a las TIC*, vol. 10, no. 3, pp. 121–139, 2021. https://doi.org/10.17993/3ctic.2021.103.121-139

[9] F. Dzay and O. Narváez, *La deserción escolar desde la perspectiva estudiantil*. 2012.

[10] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*. 2015. https://doi.org/10.1007/978-3-319-10247-4

[11] B. Mahesh, "Machine Learning Algorithms-A Review," *Int. J. Sci. Res.*, vol. 9, no. 1, pp. 1–6, 2018, doi: 10.21275/ART20203995.

[12] V. Nasteski, "An overview of the supervised machine learning methods," *Horizons.B*, vol. 4, no. December, pp. 51–62, 2017. https://doi.org/10.20544/HORIZONS.B.04.1.17.P05

[13] N. Hutagaol and Suharjito, "Predictive modelling of student dropout using ensemble classifier method in higher education," *Adv. Sci. Technol. Eng. Syst.*, vol. 4, no. 4, pp. 206–211, 2019. https://doi.org/10.25046/aj040425

[14] J. Cordova, H. Vega, and C. Rodriguez, "Firma digital basada en criptografía asimétrica para generación de historial clínico," *3C Tecnol.*, vol. 9, no. 4, pp. 65–85, 2021. https://doi.org/10.17993/3ctecno/2020.v9n4e36.65-85

[15] G. Martínez, H. Vega, C. Rodriguez, and Y. Guzmán, "Marketing de proximidad mediante aplicación móvil con dispositivos Beacon," *3C TIC Cuad. Desarro. Apl. a las TIC*, vol. 9, no. 4, pp. 89–111, 2020. https://doi.org/10.17993/3ctic.2020.94.89-111

[16] H. Vega, S. Moquillaza, O. Pacheco, and P. De La Cruz, "Support in research work for the increase of graduates by the thesis dissertation mode," *3C TIC*, vol. 11, pp. 171–189, 2022. https://doi.org/10.17993/3ctic.2021.104.17-31

9 Authors

Hugo Vega is with Universidad Nacional Mayor de San Marcos, Lima, Peru.

Enzo Sanez is with Universidad Nacional Mayor de San Marcos, Lima, Peru.

Percy De La Cruz is with Universidad Nacional Mayor de San Marcos, Lima, Peru.

Santiago Moquillaza is with Universidad Nacional Mayor de San Marcos, Lima, Peru.

Johny Pretell is with Universidad Nacional Mayor de San Marcos, Lima, Peru.

Article submitted 2022-02-14. Resubmitted 2022-04-17. Final acceptance 2022-04-17. Final version published as submitted by the authors.