ISBN: 978-628-95207-8-1. ISSN: 2414-6390. Digital Object Identifier: 10.18687/LACCEI2024.1.1.1316

Predictive modeling based on machine learning strategies to forecast student dropout at a Peruvian university: A case study

Kristelly Magdalena Aguilar Lopez[1], Yuri Carbajal Ortega[1], Daril Giovanni Martinez Hilario[1] and Sol Rodriguez[1]
[1] Facultad de ingeniería, Universidad Tecnológica del Perú, Perú, U71308332@utp.edu.pe
[1] Facultad de ingeniería, Universidad Tecnológica del Perú, Perú, U18204792@utp.edu.pe
[1] Facultad de ingeniería, Universidad Tecnológica del Perú, Perú, c26589@utp.edu.pe
[1] Facultad de ingeniería, Universidad Tecnológica del Perú, Perú, c19588@utp.edu.pe

***Abstract*– University students experiment different factors that bring as a consequence the abandonment of his professional career. In Perú, the dropout rate becomes a critical point of attention due to its increase since COVID-19. Despite the fact that the institutions join forces to improve student retention, these seem to be insufficient because of the root causes of the problem are not analyzed. Hence, this study aims to analyze the main causes associated to student dropout of a population of students from the academic period 2022-2 of a private university. For this purpose, three predictive models (random forest, logistic regression and decision tree) were designed to identify the main risks associated to abandonment of students. The predictive models were designed with the automatic learning method (Machine Learning) through Google Collab programming, obtaining a comparison of predicted dropout versus real dropouts, performing a model accuracy of 93% for the logistic regression model. Weighting the main risks identified, different retention strategies can be proposed to reduce the desertion rate.**

***Keywords*-- University dropout, desertion, machine learning, predictive model.**

I. INTRODUCTION

Higher education is an important aspect for modernization and development of societies. It is well-known that developed countries have invested in higher-education programs to create innovation atmospheres, knowledge clusters and it contributes to the formation on human capital [1], [2]. This have been seen by lower-income countries as an opportunity due to these countries are experimenting a continuous increase of young population that are able to learn and to create knowledge in universities [3]. In spite of the great difference and obstacles present in lower-income countries, higher education is under expansion. On the other hand, universities are obliged to review their offer and their programs in order to adapt the to the current markets demands (paradigms, technologies, etc.) [4].

Student's dropout is a problem that have been evaluated in previous literature. Student dropout or student desertion is understood as the abandonment of academic training due to different conditions. According to reference [5], desertion is a personal decision motivated by several factors related to perception and feeling of the students. Moreover, the socio-economic environment of the families and the lack of access of funding programs difficult the student permanence. In addition, other factors such as sexism, bullying, pregnancy and illness are associated to particular cases but must be taking in account by higher institutions [6–8]. For that reason, it is important for institutions and universities to find the particular factors associated to the student dropout, keeping in mind that each student community has its own problems [9]. To face this problematic, the prediction of university dropout becomes important to analyze and weight the factors that influence dropout aided with the risk prediction [10], although multifactorial analysis are difficult to interpret [3].

Previous works have explored the causes of desertion in different countries, evaluating different factors between students and institutions. Reference [11] evaluated the university dropout in an institution of Bogotá (Colombia) to develop a robust prediction model in 2019. The most important causes are socio-economical (gender, rural-urban, type of funding and type of admission). The model developed by the authors suggests that the models are more effective for students from third semester and older. Reference [3] evaluate the student desertion in a public university of Ecuador in 2019. The authors evaluate different external and internal factors, but they considered that student's personal circumstances (pregnancy, teacher's commitment to the student, limited knowledge of software, bullying, sexism, addictions, number of children, adaptation to the university learning, institution ranking perception and perspective about the labor market) were related to student desertion. The logistic regression method was used to predict university dropout, achieving an accuracy of 95% and delimiting the most important factors to elaborate new strategies of retention. Reference [5] studied the student dropout in an engineering program of a Peruvian public university (UNMSM) in 2022. According to the authors, it was found that the most relevant factors for a student to tend to dropout are the historical average of their grades, the average of their grades of the last cycle and the number of credits of their approved courses. The authors elaborated two models based on autonomous learning process, obtaining an accuracy percentage of 90.34% and precision of 95.91%. Similarly, Reference [12] evaluated the student dropout in an education program of a Peruvian public university in 2023 using a multifactorial analysis. The authors concluded that the most important factors of desertion are

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

associated to personal factors (motivation and personal goals) and academic factors (mainly academic credits), meanwhile economic factors and institutional factors were not determinant. As mentioned above, different factors are associated to university dropout and the institutions must know the factors associated to their students, taking in account that even in a particular region the student characteristics could be totally different.

The student's dropout is a worrying situation for Peruvian universities and institutions due to its effects. According to Ministry of Education (MINEDU) of Perú, 300 000 university students abandoned their studies [5] due to COVID-19 (2020), and in 2021 the desertion rate reduced in 11.5% [13]. The consequences of student dropout are diverse. For students, university dropout affects their personal goals and their insertion in the labor market [4]. In most cases, the desertion is absolute and students do not complete a professional career, representing a low value labor [14]. For institutions, the student dropout affects, in the short term, their financial projections in each cycle [2]; but in the long term, the student dropout affects the prestige and the public perception of the institutions because higher institutions are perceived as weak organizations that do not executed strategies to attract talented people [3], [8], [15], [16]. In 2021, the Peruvian government invested 520 million of soles to assure the continuity of studies in public institutions and this financial aid continues as part of a political decision [13]. Therefore, it is important to analyze the main factors and to elaborate strategies to promote the student retention. Hence, the aim of this paper is to design a predictive model to monitor and forecast the student desertion in a Peruvian university. For this purpose, the authors will analyze student data form a Peruvian university during the period 2022-2. The dropout models presented in this paper are: decision tree, logistic regression and random forest algorithm.

II. Methodology

The present research uses methodologies based on Machine Learning models for a detailed study of the identification process of the variables that may be relevant in the decision made by a student to abandon their higher education [17]. The population is comprised of 10401 students, who belonged to the academic period 2022-2 from August to December 2022 of a higher-education institution. In a first stage, a collection of information on variables available to the institution was considered, preparing a list of reports and a calendar of availability of reports (variables). These reports were extracted from the institutional CRM software, which consists of a customer control and management system (students). The reports extracted from the system contain both personal and institutional historical information of the student, which are downloaded in Excel formats. It is important to mention that this calendar only considers 16 academic weeks out of a total of 18, because the last 2 weeks are considered as a process of closure of dropouts.

Data Collection Technique

Five types of report were extracted from the institutional CRM, are the following:

- Attendance report 2022-2.
- Report of notes 2022-2.
- Debt Report 2022-2.
- Dropout Report 2022-2.
- General information report of enrolled students 2022-2.

Data Mining Technique

The data of each report was reviewed and normalized, identifying irregularities in its format. It was important to order and summarize the relevant information of the variable contained in the extracted report. In this case a total of 5 reports were obtained where we worked in the same way. Subsequently, all the information was grouped in a single Excel document, following an order, and summarizing with the values of each variable. Each variable was acquired cumulatively over the 16 academic weeks, in order to process and assemble a single information board that was converted and imported into csv format, similar to the process performed by Reference [18].

The analysis of the behavior of the dropout rate was carried out with respect to the grouping by weeks, which allowed to verify the hypothesis on certain data according to what was expected. In order to compare the expected values and the real values, the statistical chi-square test was executed to know if the null hypothesis was true [19].

The predictive level analysis of the main variables was carried out by applying WOE and Information Value techniques:

Information Value (IV), it is a technique of selection of variables for predictive models, which allows to classify the variables according to their relevance. In addition, this allows the measurement of predictive power by grouping the attributes of each variable. Equation (1) represent the IV calculation, where k is the number of categories:

$$IV = \sum_{i=1}^{k}(Distribution\ of\ good\ events_i - Distribution\ of\ bad\ events_i) \times WOE_i \quad (1)$$

This allowed to determine the predictive power of the variables, according to what is established in the standard predictive rule specified in Table 1.

Table 1: Prediction levels established by the results of the Information Value

Information Value (IV)	*Prediction levels*
< 0.02	Useless for prediction
0.02 al 0.1	Low predictor
0.1 al 0.3	Medium predictor
0.3 al 0.5	High predictor
> 0.5	Very high predictor

WOE indicates the weight determined for the IV which indicates the predictive power of an independent variable in

relation to the dependent variable. It is calculated by grouping the independent variables according to equation (2).

$$WOE = \left[\ln\left(\frac{Relative\ frequency\ of\ good\ events}{Relative\ frequency\ of\ bad\ events}\right)\right] \times 100 \quad (2)$$

In the analysis, it is considered as "good event" the students who is enrolled (attending regularly) and "bad event" is determined to that student who is at risk of dropping out.

Data mining techniques allow classifying student profiles to determine the risk of dropout, through the application of three predictive models: Logistic Regression, Decision Tree, and Random Forest. This implementation was achieved through the free tool Google Colab, which has environments that allow tests based on Jupiter Notebook, achieving programming in Python language, as well as the training of the proposed models. Within its environment, libraries were configured, which allowed certain functions to be grouped and allowing the training of the models [11, 12].

Technique for Evaluating Classifiers

Through this evaluation, the fitting level of the model with the training grouping was determined. Its main objective is to compare and analyze the efficiency between each classifier to determine which model was able to recognize the profile of dropout students. To this, three performance measures were used: accuracy, precision and sensitivity. For this, correlation analysis plots, cut-off point, confusion matrix and ROC curve analysis were used. The correlation graphs made it possible to identify whether two variables are related to each other or not. The cut-off point was used to convert the prediction probability into positive or negative cases of a given training model. The confusion matrix allowed to capture the performance of the model to be executed. The ROC curve allowed the evaluation of the capacity of the predictive models, followed by a discrimination of the models to determine the appropriate model of greater capacity to be executed [21].

Results Interpretation Technique:

In this process, the results of the model were exported, obtaining greater accuracy, precision, and sensitivity. The extraction of the results contains the detail per student and the categorization values that at the end were processed to measure their predictive force, for which the WOE statistical formula was used, which allowed to measure the effectiveness of the groupings of variables (categories), in a predictive way, determining the marginalization of positive or negative variables for the model, obtaining the result of the value of the probability to desert. Fig. 1 shows the diagram that explains the methodology applied for the execution of the predictive model.

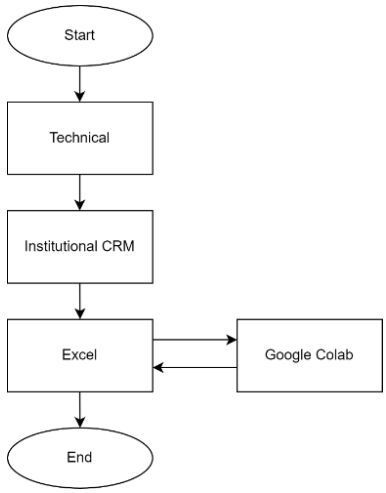

Fig. 1. Process flow diagram of the methodology described above.

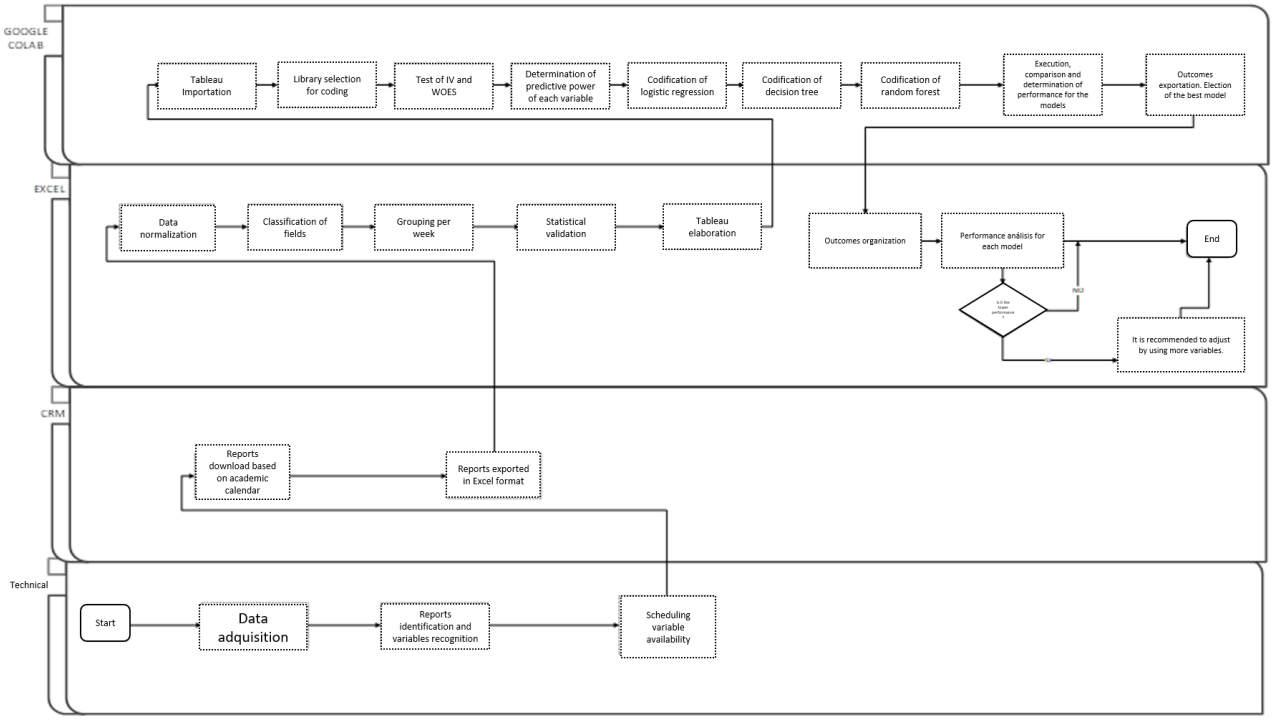

Fig. 2. Flowchart for methodology.

The codes employed for each model and the work flowchart (Fig. 2) are shown as follow:

Logistic regression

```
#ENTRENAMOS LOS DATOS IMPORTANDO LA LIBRERÍA LOGISTICREGRESSION
FROM SKLEARN.LINEAR_MODEL IMPORT LOGISTICREGRESSION
LOGREG = LOGISTICREGRESSION(RANDOM_STATE=1) #CREAMOS LA VARIABLE LOGREG PARA EL MODELO DE REGRESIÓN LOGÍSTICA
LOGREG.FIT(X_TRAIN2, Y_TRAIN2) #ENTRENAMOS EL MODELO DE REGRESIÓN LOGÍSTICA USANDO LOS DATOS DE X_TRAIN, Y_TRAIN
LOGREG_PRED = LOGREG.PREDICT(X_TEST) #GENERAMOS LAS PREDICCIONES CON X_TEST USANDO EL MODELO DE REGRESIÓN LOGÍSTICA
LOGREG_PRED #MOSTRAMOS LAS PREDICCIONES GENERADAS 0 CUANDO NO SOBREVIVE 1 CUANDO SÍ SOBREVIVE
ARRAY([0, 0, 0, ..., 1, 0, 1])

#CALCULAMOS EL SCORE DE ACCURACY COMPARANDO LAS PREDICCIONES GENERADAS VERSUS Y_TEST IMPORTANDO LA LIBRERÍA ACCURACY_SCORE
FROM SKLEARN.METRICS IMPORT ACCURACY_SCORE
PRINT('EL ACCURACY PARA MI PRIMERO MODELO ES :{0:.8F}'.FORMAT(ACCURACY_SCORE(Y_TEST,LOGREG_PRED))) #USANDO LA EXPRESIÓN REGEX .8F PARA MOSTRAR 8 DECIMALES
EL ACCURACY PARA MI PRIMERO MODELO ES :0.90300053
#COEFICIENTES DE LA REGRESIÓN LOGÍSTICA
#MOSTRAMOS LOS COEFICIENTES DE LA ECUACIÓN OBTENIDA CON LA REGRESIÓN LOGÍSTICAS
COEFICIENTES = PD.DATAFRAME({'VARIABLES':X_TRAIN.COLUMNS.VALUES, 'BETAS': NP.ROUND(LOGREG.COEF_[0],4) })
COEFICIENTES
#CALCULAMOS LA PROBABILIDADES DE OBTENER 1 (SÍ SOBREVIVIÓ) CON EL MÉTODO PREDICT_PROBA
PROBA_PRED_TEST = LOGREG.PREDICT_PROBA(X_TEST)[:,1]
PROBA_PRED_TEST
ARRAY([0.05045149, 0.05816764, 0.03000154, ..., 0.58461543, 0.0309276 ,
    0.5857106 ])

#CREAMOS UNA FUNCIÓN PARA VARIAR EL PUNTO DE CORTE -POR DEFECTO 0.5- COMPARANDO CON LAS PROBABILIDADES OBTENIDAS Y CALCULAMOS EL SCORE EN CADA ITERACIÓN
LISTA_DE_ACCURACY=[]
FOR PUNTO_DE_CORTE IN RANGE(0,100):
  PRED_0_1 = [1 IF X >= PUNTO_DE_CORTE/100 ELSE 0 FOR X IN PROBA_PRED_TEST]
  LISTA_DE_ACCURACY.APPEND(ACCURACY_SCORE(Y_TEST, PRED_0_1))
#DIBUJAMOS LOS PUNTOS DE CORTE CON SUS RESPECTIVOS SCORE
XS = [X/100 FOR X IN RANGE(0,100)]
YS = LISTA_DE_ACCURACY
PLT.FIGURE(FIGSIZE=(8,6))
PLT.GRID(TRUE)
PLT.XLABEL('PUNTO DE CORTE')
PLT.YLABEL('SCORE')
PLT.PLOT(XS, YS)
#OBTENEMOS NUEVAS PREDICCIONES (0,1) A PARTIR DEL NUEVO PUNTO DE CORTE ÓPTIMO A PARTIR DE LA GRÁFICA OBTENIDA
PREDICCIONES_NUEVO_PC = [1 IF PROB>=0.5 ELSE 0 FOR PROB IN PROBA_PRED_TEST]
NP.ARRAY([PREDICCIONES_NUEVO_PC])
#MOSTRAMOS UN NUEVO SCORE COMPARANDO Y_TEST CON LAS NUEVAS PREDICCIONES A PARTIR DEL NUEVO PUNTO DE CORTE
PRINT('EL ACCURACY PARA MI PRIMER MODELO CON NUEVO PUNTO DE CORTE ES :{0:.8F}'.FORMAT(ACCURACY_SCORE(Y_TEST,PREDICCIONES_NUEVO_PC)))
#GRAFICAMOS UNA MATRIZ DE CONFUSIÓN COMPARANDO Y_TEST CON LAS NUEVAS PREDICCIONES IMPORTANDO CONFUSION_MATRIX
FROM SKLEARN.METRICS IMPORT CONFUSION_MATRIX
MATRIZ_CONFUSION = CONFUSION_MATRIX(Y_TEST,PREDICCIONES_NUEVO_PC)
PLT.FIGURE(FIGSIZE=(10,6))
AX=SNS.HEATMAP(MATRIZ_CONFUSION, ANNOT = TRUE, ANNOT_KWS={"SIZE": 10}, FMT="".1F"")
AX.SET_YLIM((0,2))
PLT.XLABEL('LO PREDICHO')
PLT.YLABEL('LO REAL')
PLT.PLOT()
```

Decision tree

```
#Entrenamos los datos importando la librería DecisionTreeClassifier
from sklearn.tree import DecisionTreeClassifier
tree_clf = DecisionTreeClassifier(random_state=1) #creamos la variable tree_clf para el modelo de árbol de clasificación
tree_clf.fit(X_train2,Y_train2) #entrenamos el modelo de árbol de clasificación usando los datos de X_train, y_train
tree_y_pred = tree_clf.predict(X_test)#generamos las predicciones con X_test usando el modelo de árbol de clasificación
tree_y_pred #mostramos las predicciones generadas 0 cuando no sobrevive 1 cuando sí sobrevive

#Calculamos el score de accuracy comparando las predicciones generadas versus y_test importando la librería accuracy_score
Print ('El accuracy para mi segundo modelo es :{0:.8f}'.format(accuracy_score(Y_test,tree_y_pred))) #usando la expresión regex .8f para mostrar 8 decimales

#Variamos la profundidad de las ramas del árbol de clasificación obtenemos el valor óptimo
for i in range(1,10):
    tree_clf = DecisionTreeClassifier(max_depth=i)
    tree_clf.fit(X_train2,Y_train2)
    y_pred = tree_clf.predict(X_test)
    print("Mi árbol da un accuracy de:", accuracy_score(Y_test,y_pred), "cuando su max_depth es: ", i)account_circle
Mi árbol da un accuracy de: 0.7863477642841692 cuando su max_depth es:  1
Mi árbol da un accuracy de: 0.7801983869472513 cuando su max_depth es:  2
Mi árbol da un accuracy de: 0.9259293594141096 cuando su max_depth es:  3
Mi árbol da un accuracy de: 0.9339019189765458 cuando su max_depth es:  4
Mi árbol da un accuracy de: 0.9431414356787491 cuando su max_depth es:  5
Mi árbol da un accuracy de: 0.9105404653749885 cuando su max_depth es:  6
Mi árbol da un accuracy de: 0.952442755168258 cuando su max_depth es:  7
Mi árbol da un accuracy de: 0.9498161367077655 cuando su max_depth es:  8
Mi árbol da un accuracy de: 0.9576959920892433 cuando su max_depth es:  9

#Obtenemos las nuevas predicciones para el árbol de clasificación con max_depth óptimo
tree_clf = DecisionTreeClassifier(max_depth=8)
tree_clf_2=tree_clf.fit(X_train2,Y_train2)
predicciones_nuevo_md = tree_clf_2.predict(X_test)

#Graficamos una matriz de confusión comparando y_test con las nuevas predicciones importando confusion_matrix
from sklearn.metrics import confusion_matrix
matriz_confusion = confusion_matrix(Y_test,predicciones_nuevo_md)

plt.figure(figsize=(10,6))
ax=sns.heatmap(matriz_confusion, annot = True, annot_kws={"size": 10}, fmt=".1f")
ax.set_ylim((0,2))
plt.xlabel('Lo predicho')
plt.ylabel('Lo real')
plt.plot()
```

Decision tree

```
#Entrenamos los datos importando la librería RandomForestClassifier
from sklearn.ensemble import RandomForestClassifier
rnd_clf = RandomForestClassifier(n_estimators = 1000, n_jobs = -1,max_depth=3,random_state=1) #creamos la variable tree_clf para el modelo de random forest
rnd_clf.fit(X_train2,Y_train2) #entrenamos el modelo de random forest usando los datos de X_train, y_train
y_pred_rnd = rnd_clf.predict(X_test) #generamos las predicciones con X_test usando el modelo de random forest
y_pred_rnd #mostramos las predicciones generadas 0 cuando no sobrevive 1 cuando sí sobrevive

#Calculamos el score de accuracy comparando las predicciones generadas versus y_test importando la librería accuracy_score
print('El accuracy para mi tercer modelo es :{0:.8f}'.format(accuracy_score(Y_test,y_pred_rnd))) #usando la expresión regex .8f para mostrar 8 decimales

#Graficamos la importancia de cada variable con gráfica de barras usando el método feature_importances_
pesos = rnd_clf.feature_importances_
cols = X_train.columns
plt.figure(figsize=(8,12))
indices = np.argsort(pesos)
plt.barh(range(len(indices)), pesos[indices], align = 'center')
plt.yticks(range(len(indices)),[cols[i] for i in indices])
```

```
plt.show()
#Graficamos una matriz de confusión comparando y_test con las nuevas predicciones importando confusion_matrix
matriz_confusion = confusion_matrix(Y_test,y_pred_rnd)
plt.figure(figsize=(10,6))
ax=sns.heatmap(matriz_confusion, annot = True, annot_kws={"size": 10}, fmt=".1f")
ax.set_ylim((0,2))
plt.xlabel('Lo predicho')
plt.ylabel('Lo real')
plt.plot()
```

Curva ROC para los 3 modelos

```
#Importamos las librerías necesarias para dibujar las curvas ROC de todos los modelos actualizados
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score
clasificadores = [logreg, tree_clf_2, rnd_clf] #creamos la variable clasificadores para guardar los modelos actualizados
#Sin balanceo
#X_train2=X_train
#Y_train2=Y_train
#Con remuestreo
oversample = SMOTETomek(random_state=42)
X_train2, Y_train2 = oversample.fit_resample(X_train, Y_train)
#Creamos una función para obtener los parámetros de la curva ROC para cada uno de los modelos actualizados
tabla_resultados = pd.DataFrame(columns=['clasificadores', 'fpr','tpr','auc'])
for cls in clasificadores:
    model = cls.fit(X_train2, Y_train2)
  yproba = model.predict_proba(X_test)[:,1]
    fpr, tpr, _ = roc_curve(Y_test.astype(np.uint8), yproba)
    auc = roc_auc_score(Y_test.astype(np.uint8), yproba)
    #yproba = model.predict_proba(X_train)[:,1]
    #fpr, tpr, _ = roc_curve(Y_train.astype(np.uint8), yproba)
    #auc = roc_auc_score(Y_train.astype(np.uint8), yproba)
tabla_resultados =
tabla_resultados.append({'clasificadores':None,'fpr':fpr,'tpr':tpr,'auc':auc},
ignore_index=True)
#Mostramos la tabla_resultados declarando como índices el nombre de cada uno de los modelos
tabla_resultados['clasificadores'] =
['regresion_logistica','arbol_clasificacion','random_forest']
tabla_resultados.set_index('clasificadores', inplace=True)
tabla_resultados
```

```
#Dibujamos la curva ROC con los parámetros obtenidos para cada uno de los modelos actualizados
fig = plt.figure(figsize=(8,6))
for i in tabla_resultados.index:
    plt.plot(tabla_resultados.loc[i]['fpr'],
             tabla_resultados.loc[i]['tpr'],
             label="{}, AUC={:.4f}".format(i, tabla_resultados.loc[i]['auc']))
plt.plot([0,1], [0,1], color='orange', linestyle='--')
plt.xticks(np.arange(0.0, 1.1, step=0.1))
plt.xlabel("Falsos Positivos", fontsize=15)
plt.yticks(np.arange(0.0, 1.1, step=0.1))
plt.ylabel("Verdaderos Positivos", fontsize=15)
plt.title('Análisis de la curva ROC', fontweight='bold', fontsize=15)
plt.legend(prop={'size':12}, loc='lower right')
plt.show()
```

III. Results and Discussion

In order to achieve a successful data collection, variables available at the institution that could contribute to the development of our statistical model were investigated. According to Reference [22], academic variables and psychological characteristics are two of the most important aspects that has to be considered. Based on that, the variable disposition calendar (Table 2) was elaborated to describe the reports generated per week. The extracted values are exported from the institutional CRM in Excel format (Table 4), which add up to a total of five reports and showed the most relevant columns to be used with the following description.

Table 2. Academic reports from the institution.

Attendance Report Period 2022-2	▪ Student code ▪ Class number or course code ▪ Number of absences ▪ Number of Attendances ▪ Total number of scheduled classes ▪ % Total absences
Report Card Period 2022-2	▪ Student code ▪ Class number or course code ▪ Status of validity or withdrawal of the course ▪ Term in effect or withdrawn status. ▪ Paper 1 (Average 1) ▪ Paper 2 (Average 2) ▪ Paper 3 (Average 3) ▪ Final Average
Report of debts Period 2022-2	▪ Student code ▪ Fee No. ▪ Payment date ▪ Due Date ▪ Days past due ▪ Week
Dropouts Report Period 2022-2	▪ Student code ▪ Effective or withdrawal status of the academic period ▪ Date of withdrawal
General Enrollment Information Report Period 2022-2	▪ student code ▪ gender ▪ age ▪ type of student ▪ area ▪ career ▪ duration ▪ cycle ▪ distance ▪ possibility to graduate

Table 3 was then converted to CSV format, so that it could be uploaded to Google Collab to start the configuration of the libraries and instances that would allow coding to create the predictive models. Once the information was ready, a chi-square test was previously performed to analyze the information uploaded, to determine whether there is any difference in the dropout rate with respect to the week variable to be used, and to evaluate the dependence or independence between the two variables. The variable "week" was divided in 4 categories, which show the results of the number of students accumulated in force, the accumulated dropout rate per single week, totals, and the dropout rate per week (Table 4).

Table 3. Example of single information board (Excel).

Variable	Result	Description of Variable
SEMANA	16	It's showed the week in which the information was segmented, from week 1 to week 16.
ID	0000033479	Unique code assigned to the student by the enrollment system.
RETIRO_TOTAL_SEMANA	0	Two values were mentioned in this column, zero indicated a "No" and 1 means "Yes". For those students who have withdrawn in the assigned week, which for the example the reading would be that in week 16 this student is still valid "No" has withdrawn.
RETIRO_ACUMULADO	0	There are 2 values in this column, zero means "Withdrawn" and 1 means "Still in force".
SEMANA_G	S15_S16	It showed the grouping of the weeks according to the report availability calendar, segmented into 4 groups of weeks by new variable entries.
GENERO_G	F	It showed the information of female gender "F" and male gender "M".
EDAD_G	21_30	This column was grouped by age range for a better distribution of information.
TIPO_ALUMNO_G	IENTRANTS	This column is displayed if the student is a new student or an old student.
SEDE_G	T0002	This column indicated the location in which you are enrolled, in order to cross-check information and determine the distance.
AREA_G	DD	This column showed the area to which the enrolled student's career belonged.
CARRERA_G	C2	This column shows the name of the career in which the student is enrolled.
DURACION_G	3	This column showed the length of the student's career.
CICLO_G	1	This column shows the student's academic cycle, which can be from 6 to 8 academic cycles.
POSIBLE_EGRESADO_G	STUDENT	This column shows the cumulative number of students who could possibly graduate if they pass all their courses within the current period.
ZONA_G	Lima North	This column shows the geographic area where the student lives.
INASISTENCIA_G	30_PCT	This column shows the % of absences obtained by the students.
CRET_G	25_PCT	This column shows the % of withdrawn courses taken by students.
TRABAJO1_G	13_17	This column showed the range of grades in which the student scored for his or her GPA 1.
TRABAJO2_G	13_17	This column showed the range of grades in which the student was obtained for his or her GPA 2.
TRABAJO3_G	10_13	This column showed the range of grades in which the student scored for his or her GPA 3.
TRABAJOJ_G	10_13	This column shows the range of grades in which the student obtained his final average.
DISTANCIA_G	Top	This column shows the level of risk with respect to the domicile and the assigned headquarters.
CUOTA2_G	1 mont debt	This column shows the delinquency status with respect to the quota that should have been paid in the assigned week.
CUOTA3_G	1 mont deb	
CUOTA4_G	1 mont debt	
CUOTA5_G	1 mont debt	
CUOTA6_G	1 mont debt	

Note 1. Reports extracted from the CMR system contain student information data that will be considered as variables. After the normalization process, we grouped all the results of each report in a single Excel sheet shown in Table 3. All results are organized by student, week, ranges, and cumulative results of each variable finally in a single Excel sheet as recommended by Reference [17].

Note 2. Table 3 shows the results of the information for one week and for one student only, as an example to better understand the analysis of the information of the board to be used.

Table 4. Students enrolled and desertions per week.

Grouping by weeks	Accumulated number of students in force	Accumulated desertion	Total, general	Desertion rate
S01_S05	51730	133	51863	0.00256
S06_S10	50458	281	50739	0.00554
S11_S14	39455	188	39643	0.00474
S15_S16	19483	77	19560	0.00394
Total	161126	679	161805	

For the chi-square test, two hypotheses were stablished:

- Null Hypothesis (H_o): There is no difference in the desertion rate with respect to weekly groupings.
- Alternative Hypothesis (H_1): There is a difference in the desertion rate for the groupings by weeks proposed.

As a general rule, the significance level was determined to be regularly 0.05, determining that the probability of obtaining differences is 5%. In order to calculate the expected values, equation (3) was used for each cell of the contingency table (Table 5). The evaluation of chi-square indicates that P value is 58.06 with 95% acceptance and 3 degrees of freedom. In summary, after grouping the variables it is concluded that, in the particularities of the variables for each week, similar dropout rates were sought for this exercise, but it resulted in a different behavior between each week.

$$\boldsymbol{Value\ expected} = \frac{(SSum\ of\ rows * Sum\ of\ columns)}{Sum\ of\ table} \quad (3)$$

Table 5. Results for expected cumulative frequency.

Grouping by weeks	Current cumulative frequency	Cumulative dropout frequency	$(OE)^2/E$ Cumulative frequency	$(OE)^2/E$ Cumulative dropout
S01_S05	51645.36	217.64	0.14	32.92
S06_S10	60526.08	212.92	0.09	21.77
S11_S14	39476.64	166.36	0.01	2.82
S15_S16	19477.92	82.08	0.00	0.31

Later, once the Excel consolidated data was converted into CSV format, it was imported to the Google Colab tool, in which the libraries were structured in the aforementioned tool, which is very similar to a Jupiter Notebook, but within Google drive, being its hosting in the cloud. The information was organized in different libraries for further analysis. WOES and "Information value" were calculated to evaluated the corresponding weight for each factor evaluated to develop the predictive models. Those factors and the value is shown in Table 6.

The levels of prediction or relevance of each variable were presented by performing an internal logical search that was requested when coding the WOE function and the Information value in our Google Colab work environment. According to the list obtained, for example, the "distance variable" has an IV of 0.006, which if we compare it with Table 6, the risk prediction index indicates that it has an IV of less than 0.2 (Table 1), therefore, it would not be a good predictor and should be excluded. Then the analysis was performed between each variable, with each category remaining within each variable. This was done for each week. All the transformed fields were grouped together, excluding the gender and distance variable, since they obtained very low IV, giving rise to a new table with transformed results with their respective WOES week by week within the 4 categories. It was also pointed out that it was not possible to exclude all the variables, despite the fact that the risk levels were already defined, which indicated that they should be less than 0.5, because the institution did not have more variables, and if more variables were removed, the model to be implemented could weaken its performance. Likewise, a correlation analysis was performed, as shown in Figure 2, in order to determine that the variables are not correlated with each other, in order to ensure that the variables that we are going to consider to build our model do not redound in information, since, if this is not verified, the model would not have an efficient prediction result. According to Fig. 2, it indicates that the duration variable versus the career variable is highly correlated. The correct procedure is to choose one of them and not both. In order to that, it is important to verify the results of the IVs,

which will determine which one will have greater relevance. The analysis is performed one by one, which allows us to have an honest list of variables to be considered in the model. Finally, the model building, a two-part division will be made: training data with 80% and test data with 20%. Then, the Smotetomek function was used for the treatment of unbalanced data, which allows sampling and sub-sampling for weighting.

Table 6. Variables by Information Value.

Variable	Value
distancia	0.00629725
genero	0.00679502
área	0.01731376
sede	0.01929210
cuota6	0.02032528
zona	0.02138063
edad	0.02500843
cuota5	0.02830907
cuota2	0.03850329
duración	0.04509395
cuota3	0.05129106
cuota4	0.05531566
carrera	0.08982304
Trabajo1	0.09277476
semana	0.09433597
posibles egresados	0.12619280
trabajo3	0.17083575
tipo_alumno	0.64973450
ciclo	0.83424431
trabajo2	1.33400604
inasistencia	1.54868121
pct_cursos_ret	2.00560751
trabajo1	2.33413396

The logistic regression model was developed based on the normalized data. The functions that allow the training of the imported data in the logistic regression library were introduced. Then, the accuracy score was calculated by comparing the predictions generated versus the Y_ test in. Finally, the values called betas are obtained as detailed in Table 7, this as a result of the application of logistic regression. The results of the logistic regression equation were obtained: According to the previous analysis, the cut-off point is obtained with the purpose of being able to classify the probability, considering grouping them in categories of a variable, therefore if the variable achieves a weighted greater than the PDC (cut-off point), it is determined that this variable is admitted within that category, obtaining the following PDC comparison graph (Fig. 3).

Then, the authors proceeded to code the functions in order to run the second model, which is the decision tree model, where we also ran a confusion matrix detailed in Fig. 4, which also tells us that out of a total of 147 dropouts, only the model can correctly predict 89 dropouts [23].

At this stage, the third and last Random Forest model was executed, which consists of applying the decision tree and making replications where the levels of importance of each variable were obtained, obtaining as a result the results shown in Fig. 5, according of the analysis of Reference [24].

Table 7. Ordering of variables by betas obtained from WOES.

Variable	Betas
AREA_W	-1.6063
TRABAJO3_W	-1.3574
CUOTA4_W	-0.9905
POSIBLE_EGRESADO_W	-0.9783
CUOTA3_W	-0.7800
INASISTENCIA_W	-0.7116
CICLO_W	-0.6561
PCT_CURSOS_RET_W	-0.6178
TRABAJOJ_W	-0.6095
TRABAJO2_W	-0.6073
TRABAJO1_W	-0.5653
SEDE_W	-0.2922
CARRERA_W	-0.1448
EDAD_W	-0.0465
SEMANA_W	0.0573
ZONA_W	0.0694
CUOTA2_W	0.5427

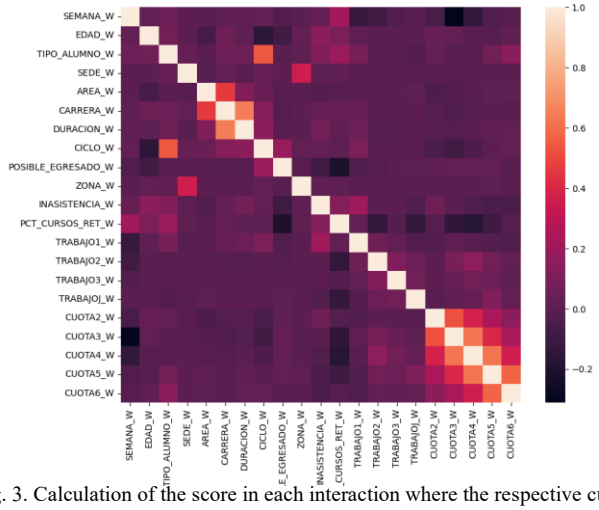

Fig. 3. Calculation of the score in each interaction where the respective cut-off points are defined.

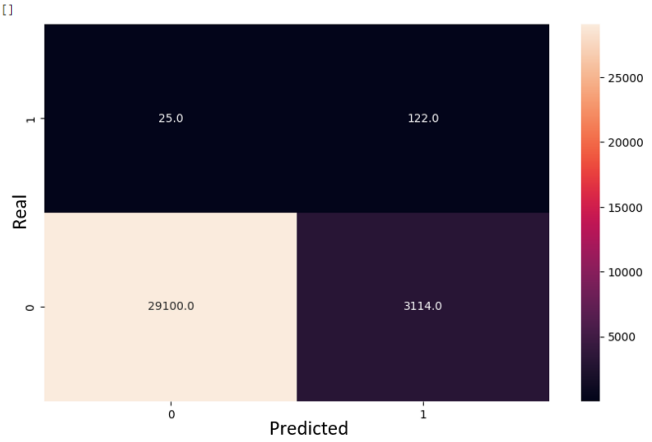

Fig. 4. The graph shows the confusion matrix of the Logistic Regression model, which, of the 147 cases of attrition obtained, the logistic regression has captured 122 cases as predictors of desertion.

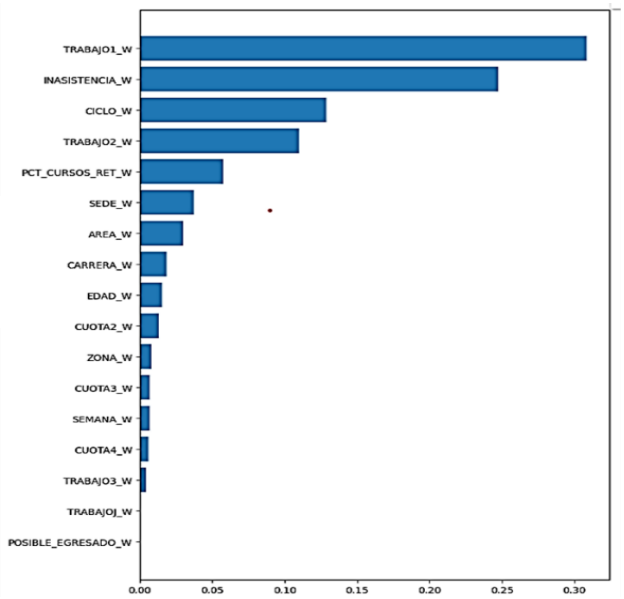

Fig. 5. Variable according to the importance level.

Subsequently, the confusion matrix was run to determine the accuracy of the Random Forest model, which is shown in Fig. 6, where it indicates that, out of a total of 147 dropouts, only the model can correctly predict 116 dropouts.

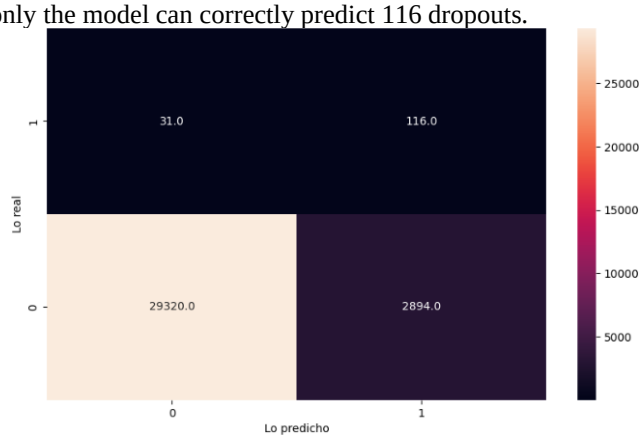

Fig. 6. Confusion matrix of the Random Forest model, which shows that of the 147 cases of attrition obtained, the Random Forest model has captured 116 cases as predictors of attrition.

An analysis of the three models obtained was carried out by performing the ROC (Receiver Operating Characteristic) curve analysis shown in Fig. 7, which indicates that the model furthest from the curve will be the optimal one to apply. In this sense, the most efficient model in choice is the logistic regression model with an accuracy of 0.9332, thus discarding the decision tree and Random Forest [21].

The values obtained were exported to be multiplied by the betas obtained previously in the logistic regression model by the WOES, then the summation was used to obtain the calculation of the probability where the equation (4 and 5) of logistic regression was applied in the column of the results obtained in the exported Excel.

$$logit(p) = \log\left(\log\left(\frac{p}{1-p}\right)\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (4)$$

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}} \quad (5)$$

It is important to mention that it was determined that during the weeks that the initial dropout behavior is very low and therefore the margin of error is greater; as the weeks go by, the margin of error will decrease.

Finally, after evaluating the performance of the logistic regression, decision tree and random forest models, it was identified that the model with the best performance in predicting possible student dropout was the logistic regression model. Using the estimated parameters obtained in the logistic regression, we proceeded to calculate the probability of dropout, which can be used as a risk score. In this sense, it would be expected that students who did not drop out would have a very low or zero probability of dropping out. Similarly, students who dropped out during the academic period would be expected to have a high level of dropout or a tendency to one.

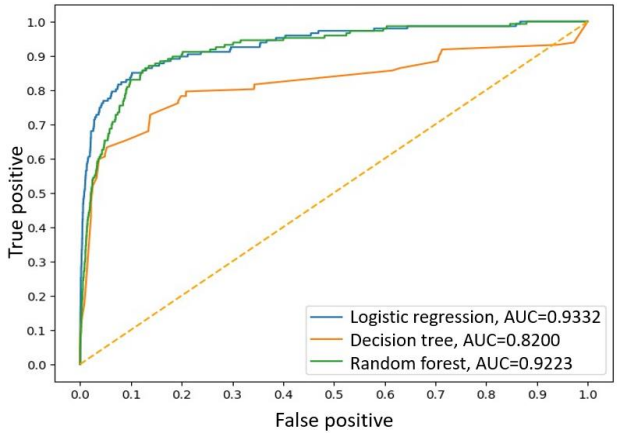

Fig. 7. ROC analysis, in which the performance of each model was validated, being the logistic regression model the one with the highest capacity, since it is farther away from the curve.

In Figure 8, we can see that effectively for the group of students who did not drop out during the academic period, the model generated values mostly very close to zero. However, it can also be seen that a minimum number of students were identified with a high score. These students would be part of the margin of error involved in the probabilistic model. Similarly, it is possible to observe in Figure 10 that in the case of students who dropped out, the model generated a score very close to one. In other words, the model is allowing us to correctly discriminate the groups of students who may or may not drop out.

After a general review of the students' performance, the weekly performance of the model was evaluated, taking into account that little information was available for the first weeks. Since 16 academic weeks were considered and in order to be able to compare the performance of the model in each week, it was proposed to divide the predicted probability of dropout into 10 groups called deciles, as shown in Table 11. As seen in Table 8, in decile 10, most of the students who dropped out are located

within each week. Also, in the first week the model calculated an intermediate dropout probability between 0.4 and 0.7 for the two students who dropped out, which implies a very high margin of error. In the opposite case, it was observed that in week 16 the model identified 32 students who dropped out of a total of 39 with a high probability of dropping out, between 0.9 and 1, which implies a very small margin of error in the case of the 7 students with a probability of less than 0.9.

In the results obtained, it can be observed how the accuracy of the predictive models has an increasing trend with respect to the analysis of the variables identified, due to the fact that more information is added to the models as the academic weeks are completed, and although on some occasions the accuracy is maintained according to the variables, in general it tends to increase. This can be seen in Fig. 9, which indicates that the model captures better the accuracy of the possible students who are going to drop out in the last weeks.

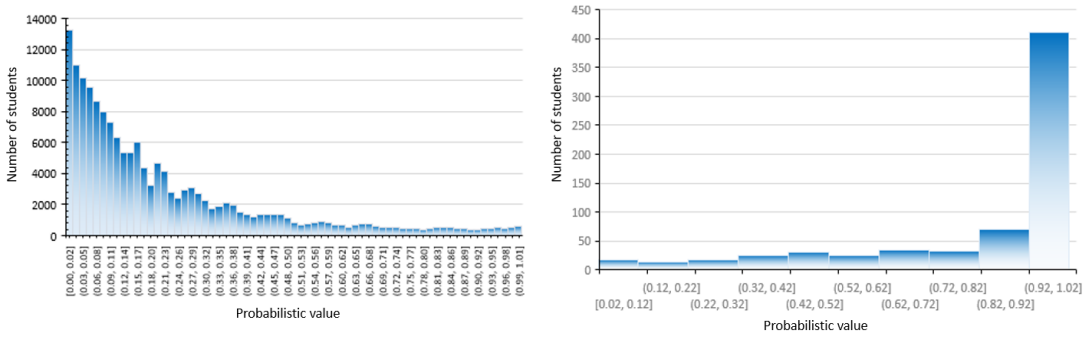

Fig 8. Right: The graph demonstrates the effectiveness of low probability of non-dropout students. Left: The graph demonstrates the effectiveness of high probability of students dropping out.

Table 8. Comparative Performance of the Regression Linear Regression Model

Probability of desertion	Sem 1		Sem 2		Sem 3		Sem 4		Sem 5		Sem 6		Sem 7		Sem 8		Sem 9		Sem 10		Sem 11		Sem 12		Sem 13		Sem 14		Sem 15		Sem 16	
	ND	D	ND	D	ND	D	ND	D	ND	D	ND	D	ND	D	ND	D	ND	D	ND	D	ND	D	ND	D	ND	D	ND	D	ND	D	ND	D
[0.00 - 0.10>	2546		2336		2333		1019		791		3106		3226	1	3338	1	3411	2	3047		6248	2	6452	2	6546		6590	3	6587	1	7763	1
[0.10 - 0.20>	3278		3260		3230		1836	1	1677		2764	1	2777	1	2794	1	2796	1	2736	1	1519	2	1393	1	1346	1	1316	3	1319		736	
[0.20 - 0.30>	1986		2109		2100	1	2285	2	2428	3	1520	2	1479	1	1462	2	1411	3	1560	2	559	1	545	1	534		537		540		244	
[0.30 - 0.40>	1005		1105	2	1131	1	1789	7	1842	2	801	2	775	2	731		741		816	1	330	1	309		294	1	286	3	283		129	1
[0.40 - 0.50>	614	1	633	2	637	2	1202	8	1250	3	604	2	570	4	536	2	515	1	552	1	253	1	240	2	235		232	1	226	1	93	2
[0.50 - 0.60>	320		365		368	2	840	4	896	2	320	3	298	2	287	1	281	3	332	2	172		167	2	159	4	148	1	141	1	61	
[0.60 - 0.70>	294	1	222	3	217	2	585	4	652	4	319	2	314		295		279		291		163	1	146	3	145		143	2	140		73	1
[0.70 - 0.80>	145		166	1	170	3	348	5	338	4	234	3	220	5	211	2	207		241	2	139	3	137	2	122	2	126	2	119		72	1
[0.80 - 0.90>	139		135	4	132	1	295	14	275	11	235	8	219	11	212	1	205	1	200	1	203	1	190	3	184	5	173	3	170		108	1
[0.90 - 1.00]	72		55	1	52	4	108	18	119	10	295	47	250	43	224	28	209	24	212	58	351	38	306	36	269	38	248	17	236	35	443	32
Students per week	10401		10399		10386		10370		10307		10268		10198		10128		10090		10055		9987		9937		9885		9834		9799		9761	
Desertion	2		13		16		63		39		70		70		38		35		68		50		52		51		35		38		39	
Identifier with probability > 0.5	0.50		0.69		0.75		0.71		0.79		0.90		0.87		0.84		0.80		0.93		0.86		0.88		0.96		0.71		0.95		0.90	

Legend
ND: Non-dropouts are active students in that week.
D: Total number of deserters.

This project proposes the application of 3 predictive models based on machine learning methods using variables from personal information and student academic performance factors. These models yielded good results of accuracy and prediction (more than 90% in the Logistic Regression and Random Forest models), demonstrating the fit of the models with real data. In comparison with similar studies, Acero, Reference [11] developed a model with 80% accuracy using Bayesian algorithms for classification of probabilistic machine learning. Moreover, Logistic Regression was developed by the authors, concluding that the prediction result is related to the data obtained. In our application, the logistic regression shows that the amount of data results improves the indicators, being more accurate and stronger for the early detection of students with possibilities of desertion. In terms of the results, when applying the progressive predictive models from the automatic learning method, it was verified that the logistic regression technique has a better performance greater than 93% in terms of the accuracy of the predictions, compared to the other prediction models proposed. Just as explained by Reference [25], the logistic regression method helps to eliminate numerous variables with little relevance and stays with the variables that present high rates of predicting desertion, which increases its probability of success of the predictive models by selecting the terms for retention variable. Regarding the variables selected as

predictors of the model, Reference [22] argue that the most outstanding antecedents are based on the following criterion or characteristic, the entrance to the different higher institutions, collection of information of the problematic from their entrance and during the whole process of academic permanence. Likewise, it considers the academic performance and variables as characteristics of the variables. For this case it is evidenced that the variables with a higher IV value (Information Value) have correlation with the mentioned, these are age, cycle, quotas, possible graduates and graduated courses, being those that have a high level of indicator of predicting a desertion.

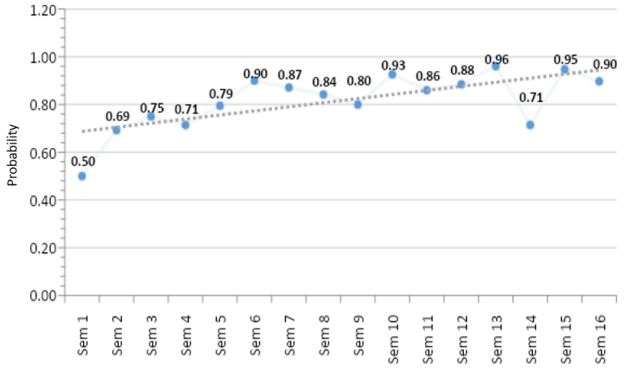

Fig. 9. Trend of student identification, which shows the identification of students with a probability greater than 0.5 and is used to identify dropouts.

IV. Conclusion

This project proposes the application of three predictive models based on machine learning methods using variables from personal information and student academic performance factors. These models yielded good results in terms of accuracy and prediction (more than 90% in the Logistic Regression and Random Forest models) based on real data from the student database from a Peruvian institution.

For the extraction of information, it is important to perform a previous analysis of each item for a proper treatment. An adequate process allows us to stablish different limits, academic standards and dates for each report. This analysis determines the sources of data collection that will contain the variables to develop the predictive models. The analysis of the probability level of each variable was decisive in order to categorize them. In this way, "WOES" and "Information value" functions were used with an automatic learning method and, through the use of libraries, listed according to the probability levels. The implementation of statistical models allows us to establish values from highest to lowest dropout probability according to the reports in order to sclerifying the performance of the model.

The statistical models analyzed were Logistic Regression, Decision Tree and Random Forest, using the confusion matrix to capture the cases of desertion predicted by the model compared to the real rate of desertion and dropout. Consequently, it was concluded by means of the analysis of the ROE curve that the model with the closest approximation to the real dropout rate was the Logistic Regression model with an accuracy of 93.3%. From comparing the probabilistic values for the number of real dropouts, it was demonstrated that the results obtained by the Logistic Regression model have a high level of precision, indicating that the more information on the variables presented week after week, the better the performance in terms of the level of accuracy, precision and sensitivity of the model, achieving up to 95% reliability in prediction. the variables with a higher IV value (Information Value) are related to student age, cycle, quotas, possible graduates and graduated courses.

Acknowledgment

The authors would like to thank the Research Directorate of Universidad Tecnológica del Perú for financial support.

References

[1] F. E. Bailon, A. A. Pantigozo, B. P. Barriga, and ..., "Machine learning as a key element in the prospective of academic performance in Peruvian Universities," *Int. J. Aquat. Sci.*, vol. 12, no. 2, pp. 4626–4636, 2021.

[2] S. Bunoti, "The Quality of Higher Education in Developing Countries Needs Professional Support," *22nd Int. Conf. High. Educ.*, vol. 4, no. 2, pp. 1–10, 2011.

[3] M. Alban and D. Mauricio, "Predicting University Dropout trough Data Mining: A systematic Literature," *Indian J. Sci. Technol.*, vol. 12, no. 4, pp. 1–12, 2019, doi: 10.17485/ijst/2019/v12i4/139729.

[4] P. J. Maquera-Luque, J. L. Morales-Rocha, and C. M. Apaza-Panca, "Socio-economic and cultural factors that influence the labor insertion of University Graduates, Peru," *Heliyon*, vol. 7, no. 7, 2021, doi: 10.1016/j.heliyon.2021.e07420.

[5] H. Vega, E. Sanez, P. De La Cruz, S. Moquillaza, and J. Pretell, "Intelligent System to Predict University Students Dropout," *Int. J. online Biomed. Eng.*, vol. 18, no. 7, pp. 27–43, 2022, doi: 10.3991/ijoe.v18i07.30195.

[6] M. Alban and D. Mauricio, "Neural networks to predict dropout at the universities," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 2, pp. 149–153, 2019, doi: 10.18178/ijmlc.2019.9.2.779.

[7] G. Ortiz-Martínez, P. Vázquez-Villegas, M. I. Ruiz-Cantisani, M. Delgado-Fabián, D. A. Conejo-Márquez, and J. Membrillo-Hernández, "Analysis of the retention of women in higher education STEM programs," *Humanit. Soc. Sci. Commun.*, vol. 10, no. 1, 2023, doi: 10.1057/s41599-023-01588-z.

[8] M. Albán, D. Mauricio, and .., "Decision Trees for the Early Identification of University Students at Risk of Desertion," *Int. J. Eng. Technol.*, vol. 7, no. 4.44, p. 51, 2018, doi: 10.14419/ijet.v7i4.44.26862.

[9] X. Palacios-Pacheco, W. Villegas-Ch, and S. Luján-Mora, "Application of data mining for the detection of variables that cause university desertion," *Commun. Comput. Inf. Sci.*, vol. 895, pp. 510–520, 2019, doi: 10.1007/978-3-030-05532-5_38.

[10] P. I. Giovagnoli, "Determinants in university desertion and graduation: an application using duration models," *Económica, La Plata*, vol. 51, no. 1–2, pp. 59–90, 2005, [Online]. Available: http://sedici.unlp.edu.ar/bitstream/handle/10915/9215/20081127035216PM_Economica_543.pdf.

[11] A. Acero, J. C. Achury, and J. C. Morales, "University dropout: A prediction model for an engineering program in bogotá, Colombia," *Proc. 8th Res. Eng. Educ. Symp. REES 2019 - Mak. Connect.*, no. July, pp. 483–490, 2019.

[12] C. A. Garrido Silva and J. Pajuelo Diaz, "Dropout among students in higher education: a case study," *Univ. Cienc. y Tecnol.*, vol. 27, no. 119, pp. 18–28, 2023, doi: 10.47460/uct.v27i119.703.

[13] El peruano, "Minedu: Tasa de deserción en educación universitaria se redujo a 11.5%," *03 De Abril*, 2023. .

[14] K. T. Joshy, F. J. Peterkumar, and S. Vakayil, "The impact of service quality on customer satisfaction; an empirical study," *Int. J. Manag.*, vol. 11, no. 3, pp. 76–88, 2020, doi: 10.34218/IJM.11.3.2020.009.

[15] A. Valencia-Arias, S. Chalela, M. Cadavid-Orrego, A. Gallegos, M. Benjumea-Arias, and D. Y. Rodríguez-Salazar, "University Dropout Model for Developing Countries: A Colombian Context Approach," *Behav. Sci. (Basel).*, vol. 13, no. 5, 2023, doi: 10.3390/bs13050382.

[16] C. Lavalle and V. L. De Nicolas, "Peru and its new challenge in higher education: Towards a research university," *PLoS One*, vol. 12, no. 8, pp. 1–12, 2017, doi: 10.1371/journal.pone.0182631.

[17] V. C. J. Elvis, N. R. Á. Fernando, L. F. A. César, and J. F. J. Deisy, "Deserción universitaria: Evaluación de diferentes algoritmos de Machine Learning para su predicción," *Rev. Ciencias Soc.*, vol. 28, no. 3, pp. 362–375, 2022, doi: 10.31876/rcs.v28i3.38480.

[18] S. I. Logroño Naranjo, N. A. Estrada Brito, V. A. Vásconez Núñez, and E. M. Rosero Ordóñez, "Analysis of the use of the Python programming language for statistical calculations.," *Espirales Rev. Multidiscip. Investig.*, vol. 6, no. 41, 2022, doi: 10.31876/er.v6i41.813.

[19] Saldaña Romero, "La Prueba Chi cuadrado," *Enfermería del Trab.*, vol. 63, pp. 31–38, 2011.

[20] I. Challenger Pérez, Y. Díaz Ricardo, and R. Becerra García, "El lenguaje de programación Python/The programming language Python," *Rev. Ciencias Holguín*, vol. 20, pp. 1–13, 2014, [Online]. Available: http://www.linuxjournal.com/article/2959.

[21] P. Martínez-Camblor, "Comparación de pruebas diagnósticas desde la curva ROC," *Rev. Colomb. Estad.*, vol. 30, no. 2, pp. 163–176, 2007.

[22] G. Fonseca and F. García, "Permanencia y abandono de estudios en estudiantes universitarios: un análisis desde la teoría organizacional," *Rev. la Educ. Super.*, vol. 45, no. 179, pp. 25–39, 2016, doi: 10.1016/j.resu.2016.06.004.

[23] J. Zárate, N. Bedregal, and V. Cornejo, "Modelos de clasificación para reconocer patrones de deserción en estudiantes universitarios," *Ingeniare. Rev. Chil. Ing.*, vol. 29, no. 1, pp. 168–177, 2021.

[24] Y. Zhang, J. Huang, J. Zhang, S. Liu, and S. Shorman, "Analysis and prediction of second-hand house price based on random forest," *Appl. Math. Nonlinear Sci.*, vol. 7, no. 1, pp. 27–42, 2022, doi: 10.2478/amns.2022.1.00052.

[25] 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.*, pp. 1–6, 2018, doi: 10.1109/ColCACI.2018.8484847.