Revista Ciencia & Tecnología

No. 36, 31 de octubre de 2022
ISSN impreso: 1390 - 6321
ISSN online: 2661 - 6734

Identificación de variables significativas en la deserción estudiantil, mediante un modelo matemático de regresión lineal KDD

Identification of significant variables in student dropout, using a KDD linear regression mathematical model

Cristina Vinueza López[1]
cvinueza31@gmail.com
https://orcid.org/0000-0002-4125-9700

Eduardo Toscano Guerrero[2]
fe.toscano@uta.edu.ec
https://orcid.org/0000-0002-3951-7774

Christian Salazar Noroña[3]
chrisalazarn@gmail.com
https://orcid.org/0000-0001-6449-8862

Cristian Flores Cadena[4]
cflores@uagraria.edu.ec
https://orcid.org/0000-0003-4071-7228

Recibido: 10/05/2022; Aceptado: 14/09/2022

Resumen
En el presente proyecto se desarrolló un modelo de regresión logística para estimar la deserción estudiantil del Instituto Superior Tecnológico Luis A. Martínez Agronómico. Se analizaron los datos de 849 estudiantes, evaluándose las variables: género, estado civil, edad, carrera, repitencia, ocupación e ingresos económicos. Para desarrollar el modelo matemático se utilizó la metodología KDD, que permite generar información a partir de una base de datos con los registros a estudiarse. Dentro el período evaluado el 82,45 % de los estudiantes no desertaron y el 17,55% sí. Para el estudio se establecieron cuatro modelos de regresión logística, sin embargo, se escogió el modelo de regresión logística 4, el cual incluye las variables 'carrera' y 'repitencia', que fueron las únicas significativas. El modelo de regresión logística 4 clasificó correctamente el 83 % de los datos de entrenamiento y el 79 % de los datos de testeo. Adicionalmente, se determinó un modelo de predicción con árboles de decisión, que estableció

[1] Magister en Matemática Aplicada, Universidad Técnica de Ambato, Ecuador
[2] Magister en Matemática Aplicada, Universidad Técnica de Ambato, Ecuador
[3] Ingeniero en Electrónica y Control, Escuela Politécnica Nacional, Ecuador
[4] Magister en Dirección de Operaciones y Seguridad Industrial, Universidad de las Américas, Ecuador

No. 36, 31 de octubre de 2022
ISSN impreso: 1390 - 6321
ISSN online: 2661 - 6734

como variable explicativa 'carrera'. El valor F1_Score del modelo de regresión logística 4 fue mayor que el valor del F1_Score del modelo con árbol de decisión.

Palabras clave: deserción estudiantil, modelo matemático, regresión logística, predicción con árbol de decisión, educación superior, metodología KDD

Abstract
In this research, a logistic regression model was used to estimate student dropout from the IST Luis A. Martínez Agronómico. The data of 849 students registered was used to build the model. The independent variables considered for the model were: gender, marital status, age, career, repetition, occupation and economic status. We used the KDD methodology to estimate the mathematical model, which allows generating information from a database with the records to be studied. In the evaluated period, 82.45% of the students did not dropout but 17.55% did it. In the study, four logistic regression models were established, finally, it was chosen the logistic regression model 4, which only includes the career and repetition variables as the only significant ones. The null hypothesis was rejected because the coefficients $\beta 1$ and $\beta 2$ of the variables 'career' and 'repetition' aren´t zero. The logistic regression model 4 correctly classified 83% of the training data and 79% of the test data. Additionally, we build a prediction model based on decision trees, which established 'career' as a unique explanatory variable. The F1_Score value of the logistic regression model 4 was higher than the F1_Score value of the decision tree model.

Keywords: student dropout, mathematical model, logistic regression, decision tree prediction, higher education, KDD methodology.

Introducción

La presente investigación pretende dotar de una herramienta que permita predecir el riesgo de deserción de estudiantes que cursan una carrera de nivel tecnológico superior en una institución del centro del país. Con este fin, se utilizarán modelos predictivos y técnicas de minería de datos para determinar patrones de comportamiento de los estudiantes, que determinen su condición de potencial desertor, asociándole un índice de deserción como probabilidad de abandono del sistema educativo. Dicha información podrá ser utilizada por las instituciones de educación superior del país para tomar medidas que eviten o reduzcan el nivel de deserción al mínimo posible. El modelo predictivo se desarrollará mediante un Modelo Multivariante con Regresión Logarítmica, ya que los modelos de elección discreta son bastantes apropiados para analizar los factores determinantes de la probabilidad de un suceso como el que se pretende estudiar. La presente investigación es de tipo caso de estudio, ya que estudia un sujeto o una realidad de carácter específico. Cabe recalcar que los estudios de casos se utilizan especialmente cuando las preguntas "cómo" o "por qué" se plantean, el investigador tiene poco control sobre los eventos, y cuando se investiga un fenómeno contemporáneo dentro de su contexto de la vida real, considerando que los límites entre el fenómeno y el contexto no son claramente evidentes (Yin, 2003).

No. 36, 31 de octubre de 2022
ISSN impreso: 1390 - 6321
ISSN online: 2661 - 6734

La presente investigación se realiza para diseñar un modelo matemático predictivo, que permita identificar de forma temprana los casos de estudiantes que presenten mayor probabilidad de desertar, para que la institución pueda aplicar medidas preventivas que evitan dichos casos de deserción. El diseño de obtuvo mediante un análisis multivariable con regresión logarítmica. La disminución de la tasa de deserción es de vital importancia en las instituciones de educación superior, ya que representa un indicador de la calidad educativa evaluado por las instancias superiores.

En la educación, la deserción se refiere al abandono prematuro de un programa de estudios antes de alcanzar un título o grado, considerando un tiempo suficientemente largo como para descartar la posibilidad de que el estudiante retome sus estudios (Himmel,2002). La deserción estudiantil constituye un grave problema en la educación de nivel superior, técnica, tecnológica o universitaria, ya que tiene una incidencia negativa sobre los procesos políticos, económicos, sociales y culturales del desarrollo nacional (Sopalo, Guevara y Burbano, 2020). Además, constituye un grave problema para la institución de educación superior, ya que el nivel deserción constituye el indicador "Tasa de retención" del Modelo de evaluación institucional para los institutos superiores técnicos y tecnológicos del proceso de acreditación establecido por el Consejo de Aseguramiento de la Calidad de la Educación Superior (CACES, 2020). De allí la importancia de establecer una alternativa que les permita identificar los casos de estudiantes próximos a retirarse e implementar acciones preventivas para evitarlo.

Metodología

La población de estudio corresponde a todos los estudiantes del Instituto Superior Tecnológico Luis A. Martínez Agronómico, instituto público de educación superior tecnológica de la ciudad de Ambato. Se estudiaron a 849 estudiantes de las carreras de Tecnología Superior en Procesamiento de Alimentos (296), Tecnología Superior en Gastronomía (299) y Tecnología Superior en Producción Pecuaria (254). El procesamiento de datos se realizó siguiendo la metodología KDD, o generación de conocimiento en bases de datos, que es reconocido como un proceso no trivial para identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos (Fayyad, Piatetsky-Shapiro y Smyth, 1996). Las fases de la técnica están en la Figura 1.

Revista Ciencia & Tecnología

No. 36, 31 de octubre de 2022
ISSN impreso: 1390 - 6321
ISSN online: 2661 - 6734

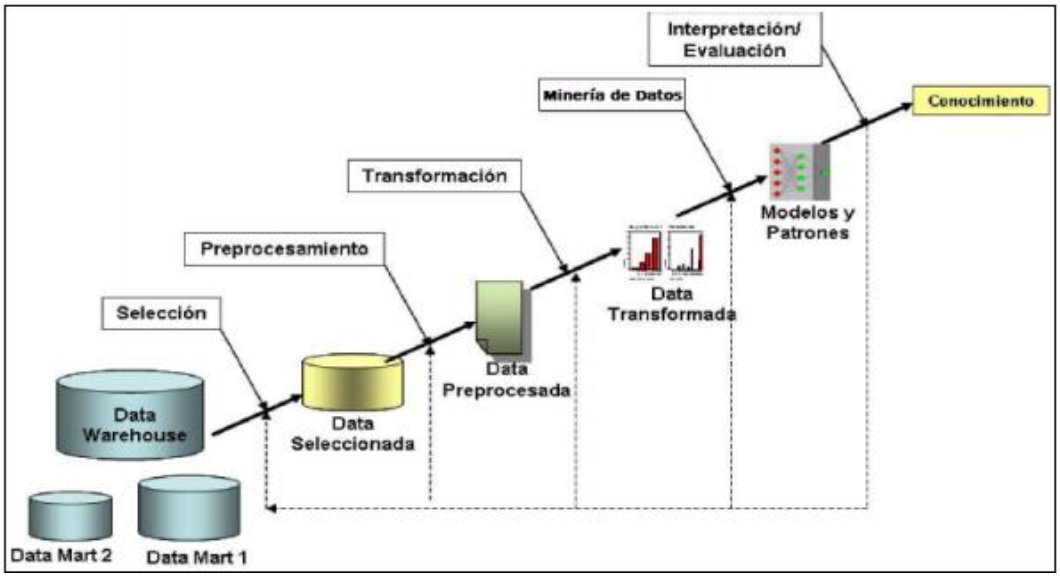

Figura 1. – Metodología KDD
Fuente: Fayyad, Piatetsky-Shapiro y Smyth, 1996)

Para el presente estudio se consideraron las siguientes variables independientes: género, estado civil, edad, carrera, repitencia, ocupación e ingresos económicos, como se indica en la Tabla 1.

Tabla 1.- Variables dependientes del proyecto de investigación

Variables independientes	**Descripción**
Género	Femenino Masculino
Estado civil	Soltero Casado Unión libre Divorciado
Edad	Valor en años
Carrera	PA Procesamiento de Alimentos G Gastronomía PP Producción Pecuaria
Repitencia	Estudiante a reprobado al menos una asignatura Estudiante no ha reprobado ninguna asignatura
Ocupación	Estudiante estudia y trabaja Estudiante solamente estudia

No. 36, 31 de octubre de 2022
ISSN impreso: 1390 - 6321
ISSN online: 2661 - 6734

Ingresos económicos	Valor promedio en USD de ingresos económicos percibidos mensualmente

Elaborado por: Vinueza, 2021

El 70% de los datos fue utilizado para generar el modelo predictivo requerido y el 30% restante para verificar su eficacia, como se indica en el trabajo de Zamorano y Martín (2018). El modelo predictivo se desarrolló mediante un Modelo Multivariante con Regresión Logarítmica, ya que los modelos de elección discreta son bastantes apropiados para analizar los factores determinantes de la probabilidad de un suceso.

El modelo de regresión logística se realizó en el Programa R Studio Versión 1.3.959 2019:

```
library(dplyr)
library(lmtest)
library(ResourceSelection)
library(MLmetrics)
library(InformationValue)
library(stargazer)
library(sjPlot)
library(DescTools)

 mod1<-glm(formula = Deserción~Género+Estadocivil+Repitencia+
           Ocupación+Ingresos+Edad+Carrera,
         records, family=binomial(link = "logit"))

 summary(mod1)
 #Retirar variables no significativas de dos en dos, para ver como cambia modelo
 #las variables a retirarse son las que tienen mayor valor p
#Se elimina género y edad
 mod2<-glm(formula = Deserción~Estadocivil+Repitencia+
           Ocupación+Ingresos+Carrera,
         records, family=binomial(link = "logit"))
  summary(mod2)
  #Se elimina ingresos y edad
 mod3<-glm(formula = Deserción~Repitencia+
           Ocupación+Carrera,
         records, family=binomial(link = "logit"))
  summary(mod3)
  #Se elimina ocupación
 mod4<-glm(formula = Deserción~Repitencia+Carrera,
         records, family=binomial(link = "logit"))
  summary(mod4)

anova(mod4, test = "Chisq")
 coef(mod4)
 exp(coef(mod4))
```

Revista Ciencia & Tecnología

No. 36, 31 de octubre de 2022
ISSN impreso: 1390 - 6321
ISSN online: 2661 - 6734

Resultados y discusión

En la Tabla 2 se muestra la distribución porcentual de la población de las variables independientes cualitativas del estudio.

Tabla 2.- Distribución porcentual en la muestra de las variables independientes

Variable	Descripción	Total	Porcentaje
Género	Femenino	511	60,18 %
	Masculino	338	39,82 %
Estado civil	Soltero	762	89,75 %
	Divorciado	54	6,36 %
	Casado	11	1,30 %
	Unión libre	22	2,68 %
Carrera	Procesamiento de Alimentos	296	34,86 %
	Gastronomía	299	35,22 %
	Producción Pecuaria	254	29,92 %
Repitencia	Estudiante a reprobado al menos una asignatura	126	14,84 %
	Estudiante no ha reprobado ninguna asignatura	723	85,16 %
Ocupación	Estudiante estudia y trabaja	276	32,50 %
	Estudiante solamente estudia	573	67,50 %

Elaborado por: Vinueza, 2021 (Fuente: IST Luis A. Martínez Agronómico)

No. 36, 31 de octubre de 2022
ISSN impreso: 1390 - 6321
ISSN online: 2661 - 6734

Las variables estudiadas se escogieron, debido a que la deserción por problemas personales puede verse afectada por el sexo del estudiante, afectando más a las mujeres que a los hombres, los conflictos de carácter económico y entorno familiar y social, se incluyen en esta causa. Se constató que los estudiantes de sexo masculino, vinculados al mercado laboral y provenientes de otras regiones del país, presentan mayor probabilidad de desertar (Baquerizo, Tam y López, 2014). Adicionalmente, según Sánchez (2016), en el Ecuador una de las principales razones que causan la deserción es la falta de acompañamiento a los jóvenes secundarios a la hora de escoger su carrera superior. Sánchez afirma que aproximadamente el 40% de los jóvenes no saben qué estudiar en la universidad, por lo cual los estudiantes optan por una carrera incorrecta, que conlleva finalmente a repitencia y deserción.

La importancia del estudio de los ingresos económicos de los estudiantes se debe a que el nivel superior la deserción estudiantil se produce con mayor frecuencia en los primeros semestres académicos y el factor económico es el que mayor relevancia justifica estas deserciones (77,9 %); seguido de los problemas familiares (7,5%) y los relacionados con las insuficiencias en el desempeño del personal-académico (6,9%) (Yaselga y Yépez, 2010).

Con relación a la deserción de los estudiantes, se observó que el 82,45% de estudiantes no desertaron y se matricularon en el último período académico de cada carrera, a diferencia del 17,55% que abandonó la institución por varios motivos. En el estudio de Matallana, Gonzalez y Fonseca (2020) se realizó el seguimiento a 510 estudiantes en un periodo comprendido entre el 26 de marzo de 2019 y 26 de marzo de 2020, de los cuales 348 eran mujeres y 162 hombres, de los cuales 96 desertaron equivalente al 19%.
Posteriormente, para realizar el análisis predictivo de los datos, se desarrolló un primer modelo de regresión logística con todas las variables independientes disponibles, obteniéndose los resultados de la Figura 2.

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -1.123e+00  9.692e-01  -1.159   0.2464
GéneroMasculino            -4.215e-02  2.312e-01  -0.182   0.8553
EstadocivilDivorciado       4.049e-01  8.175e-01   0.495   0.6204
EstadocivilSoltero         -3.167e-01  4.584e-01  -0.691   0.4896
Estadocivilunión libre     -1.101e+00  1.140e+00  -0.966   0.3342
RepitenciaSi                5.979e-01  2.731e-01   2.189   0.0286 *
OcupaciónTrabaja y estudia  3.974e-01  2.470e-01   1.609   0.1077
Ingresos                   -8.573e-05  2.742e-04  -0.313   0.7545
Edad                       -2.132e-03  2.977e-02  -0.072   0.9429
CarreraPA                  -6.728e-01  2.742e-01  -2.454   0.0141 *
CarreraPP                  -2.779e-01  2.707e-01  -1.027   0.3046
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 547.57  on 592  degrees of freedom
Residual deviance: 528.61  on 582  degrees of freedom
  (3 observations deleted due to missingness)
AIC: 550.61
```

Figura 2. – Resumen de modelo de regresión logística 1
Fuente: Programa R Studio,2021

Revista Ciencia & Tecnología

No. 36, 31 de octubre de 2022
ISSN impreso: 1390 - 6321
ISSN online: 2661 - 6734

De los resultados obtenidos, se observó que solamente las variables 'carrera' ('PA-Procesamiento de Alimentos' y 'PP-Producción Pecuaria') y 'repitencia' resultaron significativas, con 95 % de confianza, ya que el valor de p es menor a 0,05. La evaluación del criterio de información de Akaike (AIC) del modelo arrojó un resultado de 550,61. Para mejorar el modelo 1, se decidió proponer un segundo modelo, en el cual se eliminen las variables que presenten mayor valor de p, que en nuestro caso corresponden a las variables 'Edad' (p = 0,9429) y 'Género' (p = 0,8553). Los resultados del este nuevo modelo se presenta en la Figura 3.

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -1.191e+00  4.623e-01  -2.575   0.0100 *
EstadocivilDivorciado       4.185e-01  8.141e-01   0.514   0.6072
EstadocivilSoltero         -3.085e-01  4.009e-01  -0.769   0.4416
EstadocivilUnión libre     -1.082e+00  1.124e+00  -0.963   0.3357
RepitenciaSi                5.982e-01  2.730e-01   2.191   0.0284 *
OcupaciónTrabaja y estudia  3.847e-01  2.350e-01   1.637   0.1016
Ingresos                   -9.285e-05  2.720e-04  -0.341   0.7329
CarreraPA                  -6.718e-01  2.740e-01  -2.452   0.0142 *
CarreraPP                  -2.772e-01  2.706e-01  -1.025   0.3056
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 547.57  on 592  degrees of freedom
Residual deviance: 528.65  on 584  degrees of freedom
  (3 observations deleted due to missingness)
AIC: 546.65
```

Figura 3. – Resumen de modelo de regresión logística 2
Fuente: Programa R Studio,2021

Así como en el modelo anterior, se observa que solamente las variables 'carrera' y 'repitencia' resultaron significativas, así como el intercepto, con 95 % de confianza, ya que el valor de p es menor a 0,05. El valor de AIC disminuye de 550,61 a 546,65, lo que significa que el modelo 2 es mejor que modelo 1. En el estudio de Segura y Loza (2017) se reportó que las variables 'beca académica', 'edad', 'provincia' y 'título de escuela secundaria' influyen en el rendimiento académico de los estudiantes. Adicionalmente, González y Arismendi (2018) en su trabajo mostraron que los factores con mayor significancia estadística corresponden al 'género', 'año de egreso de enseñanza media' y 'jornada de estudio'.

Una vez más se eliminan nuevamente las variables que presentan mayor valor de p, es decir, 'Estado civil' (p = 0,6072) e 'Ingresos económicos' (p = 0,7329) y se obtiene el modelo 3, cuyo resultado se muestra en la Figura 4. El estudio de Albarrán (2019) revela que 65% de los estudiantes podrían dejar la institución por la falta de oportunidades laborales futuras, aumento de sus gastos personales y académicos, interrupción de las labores académicas por continuas protestas sociales, carencia de recursos económicos, baja formación escolar secundaria y desmotivación, principalmente.

No. 36, 31 de octubre de 2022
ISSN impreso: 1390 - 6321
ISSN online: 2661 - 6734

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -1.5338     0.2098  -7.312 2.64e-13 ***
RepitenciaSi               0.5881     0.2710   2.170   0.0300 *
OcupaciónTrabaja y estudia 0.4270     0.2280   1.873   0.0611 .
CarreraPA                 -0.6596     0.2715  -2.430   0.0151 *
CarreraPP                 -0.3040     0.2683  -1.133   0.2572
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 547.57  on 592  degrees of freedom
Residual deviance: 530.82  on 588  degrees of freedom
  (3 observations deleted due to missingness)
AIC: 540.82
```

Figura 4. – Resumen de modelo de regresión logística 3
Fuente: Programa R Studio,2021

En el modelo 3, se mantienen las variables ‘carrera’ y ‘repitencia’ como variables significativas, así como el intercepto, con 95 % de confianza, ya que el valor de p es menor a 0,05. El valor de AIC disminuye de 546,65 a 540,82, lo que indica que el modelo está mejorando. Por último, se elimina la variable ‘ocupación’ (p = 0,611), obteniéndose el modelo 4, mostrado en la Figura 5.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.3305     0.1745  -7.622 2.49e-14 ***
RepitenciaSi   0.6068     0.2698   2.249  0.02451 *
CarreraPA     -0.7494     0.2664  -2.813  0.00492 **
CarreraPP     -0.3752     0.2649  -1.417  0.15656
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 547.57  on 592  degrees of freedom
Residual deviance: 534.27  on 589  degrees of freedom
  (3 observations deleted due to missingness)
AIC: 542.27
```

Figura 5. – Resumen de modelo de regresión logística 4
Fuente: Programa R Studio,2021

En el modelo 4 se observa que las dos variables ‘repitencia’ y ‘carrera’ (Procesamiento de alimentos) son significativas, con un valor p de 0,0245 y 0,0049 respectivamente, presentando ambas variables valores de p menores a 0,05. El intercepto también se muestra como un valor significativo con un valor de p igual a 2,49x10-14. Cuando se compara el AIC del modelo 3 (540,82) con el valor AIC del modelo 4 (542,27), se observa un aumento del valor, es decir, según el criterio Akaike, el modelo 3 es mejor que el modelo 4. Sin embargo, el criterio AIC si disminuyó si se compara el modelo 1 (550,60) con el modelo 4 (542,27), por lo tanto, se escoge el modelo 4 que puede explicarse con solamente dos variables, ‘repitencia’ y’ carrera del estudiante’. El modelo 4 posteriormente se validó con otros métodos.

En el programa se determinó el valor de p es 0,000121 el cual es menor a 0, 05; por lo tanto, el modelo es significativo con 95 % de confianza, como se muestra en la Figura 6.

Revista Ciencia & Tecnología

No. 36, 31 de octubre de 2022
ISSN impreso: 1390 - 6321
ISSN online: 2661 - 6734

```
>   #Estadistico de prueba
>   with(mod4,null.deviance-deviance)
[1] 13.29436
>
>   #Valor de P del estadistico de prueba
>   with(mod4,pchisq(null.deviance-deviance,df.null-df.residual,lower.tail = FALSE))
[1] 0.004041405
>   #El valor de p es 0.000121 menor a 0.05 por lo tanto el modelo es significativo con
>   #95 % de confianza
>
```

Figura 6. – Análisis de significancia del modelo de regresión logística 4
Fuente: Programa R Studio,2021

Además, se realizó el análisis ANOVA del modelo 4, el cual se muestra en la Figura 7.

```
>   anova(mod4, test = "Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Deserción

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                         592     547.57
Repitencia  1   5.0204       591     542.55  0.02505 *
Carrera     2   8.2740       589     534.27  0.01597 *
```

Figura 7. – Análisis ANOVA del modelo de regresión logística 4
Fuente: Programa R Studio,2021

Del análisis ANOVA se observa que el valor p de 'repitencia' y 'carrera' es menor a 0,05 y por tanto significativo con un nivel de confianza de 95 %.
Posteriormente se determinan los coeficientes y razón Odd del modelo 4, como se muestra en la Tabla 3.

Tabla 3.- Coeficientes y valores de razón Odd del modelo de regresión logística

Variables significativas	Coeficientes	Razón Odd
Repitencia (Si)	0,5539975	1,7401955
Carrera (Procesamiento de Alimentos)	-0,8537485	0,4258158
Carrera (Producción Pecuaria)	-1,0143231	0,3626478

Fuente: Programa R Studio, 2021

Revista Ciencia & Tecnología
No. 36, 31 de octubre de 2022
ISSN impreso: 1390 - 6321
ISSN online: 2661 - 6734

Se observa que para la variable 'repitencia', la razón Odd es 1,74. Por lo tanto, se puede afirmar que el hecho de que un estudiante repita alguna asignatura aumenta en un 74 % la probabilidad de que abandone la institución. Por otro lado, con respecto a la variable 'carrera', el valor de la razón Odd de la carrera 'Procesamiento de Alimentos' es 0,42, lo que muestra que los estudiantes de esta carrera tienen 58% menor probabilidad de desertar.
Por su lado, la razón Odd de la carrera de 'Producción Pecuaria' es 0,36; lo que permite concluir que los estudiantes de esta carrera tienen 64% menor probabilidad de desertar. Esto se puede observar en la Figura 8.

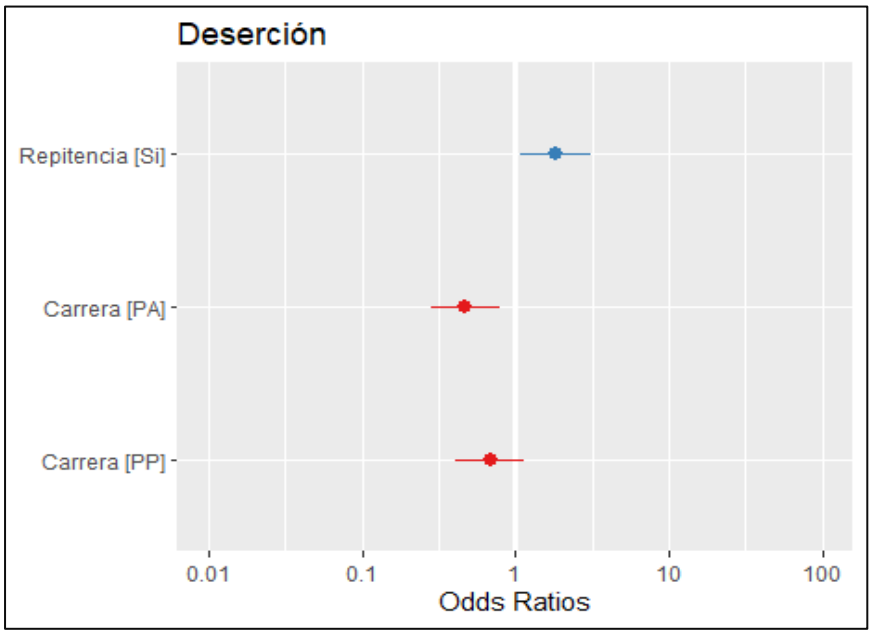

Figura 8 – Valores de razón Odd de las variables repitencia y carrera del modelo de regresión logística 4
Fuente: Programa R Studio,2021

En el estudio de Barahona, Veres y Aliaga (2016) quienes, utilizaron un modelo de regresión logística para determinar aquellos factores que han incidido en la tasa de deserción, el valor de la razón Odd, indica que los créditos (asignaturas) inscritos reducen en un 9 % la probabilidad de desertar. Es decir, los estudiantes que no desertaron son aquellos con la mayor cantidad de créditos inscritos y que han realizado un mayor esfuerzo académico. De igual forma sucede con la variable rendimiento académico que disminuye en 14 % veces la probabilidad de desertar
A continuación, se realizó la validación del modelo utilizando una Prueba Razón de verosimilitud, comparando el modelo 1, que implicaba todas las variables independientes, con el modelo 4 que considera solamente las variables 'carrera' y 'repitencia'. Realizada la prueba, se obtiene un valor de p igual a 0,5793, mayor a 0,05. Por lo tanto, el modelo 1 es estadísticamente igual a modelo 4. El resultado del test se muestra en la Figura 9.

No. 36, 31 de octubre de 2022
ISSN impreso: 1390 - 6321
ISSN online: 2661 - 6734

```
Likelihood ratio test

Model 1: Deserción ~ Género + Estadocivil + Repitencia + Ocupación + Ingresos +
    Edad + Carrera
Model 2: Deserción ~ Repitencia + Carrera
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  11 -264.30
2   4 -267.14 -7 5.6659     0.5793
```

Figura 9. – Prueba de razón de similitud entre modelo 1 y modelo 4
Fuente: Programa R Studio,2021

Posteriormente, se estableció la matriz de confusión de los datos de entrenamiento y datos de testo, utilizando el modelo de regresión logística 4.

En la Tabla 5, se puede observar que 497 de los datos fueron clasificados correctamente, mientras 97 fueron clasificados incorrectamente. Con respecto a los datos de entrenamiento (70%), se obtuvo que el porcentaje de datos que se clasifican correctamente es del 83 %.

1. Tabla 5.- Matriz de confusión de datos de entrenamiento

		PREDICCIÓN	
		No desertó	Desertó
REAL	No desertó	497	1
	Desertó	96	0

Fuente: Programa R Studio,2021

Para la implementación del modelo predictivo se consideró el 30 % de datos que no fueron utilizados en el entrenamiento. En las pruebas, se encontró que el 79 % de los datos de testeo fueron clasificados correctamente. La matriz de confusión de los datos de testeo se muestra en la Tabla 6.

2. Tabla 6.- Matriz de confusión de datos de testeo

		PREDICCIÓN	
		No desertó	Desertó
REAL	No desertó	202	0
	Desertó	53	0

Fuente: Programa R Studio,2021

No. 36, 31 de octubre de 2022
ISSN impreso: 1390 - 6321
ISSN online: 2661 - 6734

También se calcularon los valores de precisión, exhaustividad (recall) y F1-score de la matriz de confusión de los datos de testeo. Los resultados obtenidos se muestran en la Tabla 7.

Tabla 7.- Valores de evaluación de la matriz de confusión de datos de testeo

Parámetros	Valores
Precisión	1,00
Exhaustividad (Recall)	0,79
F1-Score	0,88

Fuente: Programa R Studio,2021

Los valores del F1-score muestran que el modelo matemático número 4, presenta un rendimiento del 88% en la clasificación de estudiantes que desertan de la institución.

Conclusiones

- Se realizó el análisis descriptivo de las variables de estudio, los resultados más representativos indican que 60, 18 % de los estudiantes son de género femenino, el 89,75 % son solteros, el 32,50 % trabajan y estudian y se encontró un nivel de repitencia del 14,84 %. La edad promedio de los estudiantes es 23,19 años y el valor promedio de ingresos económicos es 560, 21 USD.
- Según la razón Odd de la variable 'repitencia' si un estudiante repite alguna asignatura aumenta en un 74 % la probabilidad de que abandone la institución. Por ello, se concluye que se deben realizar otros estudios para mejorar la calidad de la educación y disminuir el porcentaje de repitencia en la institución.
- Se encontró que la variable significativa 'carrera' indica que los estudiantes de la Carrera de 'Gastronomía' tiene mayor probabilidad de desertar que de las otras dos carreras ('Procesamiento de Alimentos' y 'Producción Pecuaria'). Por lo tanto, se propone formular nuevos modelos de regresión logística, solamente en la carrera de 'Gastronomía' para determinar otras variables significativas.
- Se implementó el modelo predictivo para estimar deserción estudiantil, y se observó que el modelo de regresión logística 4 clasificó correctamente el 83 % de los datos de entrenamiento y el 79 % de los datos de testeo.

Referencias bibliográficas

Albarrán-Peña, J. (2019). La deserción estudiantil en la Universidad de Los Andes (Venezuela). *Educación y Humanismo*, *21*(36), 60-92.

Barahona, P., Veres, E y Aliaga, V. (2016). Deserción académica de la Universidad de Atacama, Chile. Comunicación, 7(2), 27-37.

Baquerizo, P., Tam, A. y López, J. (2014). La deserción y la repitencia en las instituciones de Educación Superior: algunas experiencias investigativas en el Ecuador. Universidad y Sociedad, 6(1).

CEDIA. (2020). *Indicadores para la gestión de la calidad en la educación Superior Ecuatoriana.* Recuperado de https://www.cedia.edu.ec/dmdocuments/publicaciones/Libros/indicadores2020.pdf

Fayyad, U., Piatetsky-Shapiro, G. y Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27-34.

González, F y Arismendi, K. (2018). Deserción estudiantil en la educación superior técnico-profesional: Explorando los factores que indicen en alumnos del primer año. *Revista de la Educación Superior, 47*(188), 109-137.

Matallana, A., González, J. y Fonseca, L. (2020). Modelo sobrevida para la deserción estudiantil de los programas de nivel técnico en una IES de formación para el trabajo en la ciudad de Bogotá.

Sánchez, T. (2016). *La deserción universitaria*. Recuperado de https://www.eltelegrafo.com.ec/noticias/sociedad/4/la-desercion-universitaria-bordea-el-40.

Segura-Morales, M., & Loza-Aguirre, E. (2017). Using Decision Trees for Predicting Academic Performance Based on Socio-Economic Factors. In *2017 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 1132-1136). IEEE.

Sopalo, S., Guevara, G y Burbano, R. (2020). *Análisis de los factores que inciden en la deserción estudiantil de los niños, niñas y adolescentes ecuatorianos en el periodo 2009 – 2019*. (Tesis de grado). Escuela Politécnica Nacional, Quito.

Yaselga, B., y Yépez, P. (2010). *Factores que intervienen en la deserción de los estudiantes de segundo y cuarto semestres de la Carrera de Enfermería, Facultad Ciencias de la Salud.* Universidad Técnica del Norte (Tesis de Licenciatura). Ibarra- Ecuador.

Yin, R.K. *(2003). Case study research: Design and Methods.* Recuperado de *https://iwansuharyanto.files.wordpress.com/2013/04/robert_k-_yin_case_study_research_design_and_mebookfi-org.pdf*

Zamorano, J. (2018). Comparativa y análisis de algoritmos de aprendizaje automático para la predicción del tipo predominante de cubierta arbórea.