ISSN 2477-9105
DOI: https://doi.org/10.47187/perf.v1i27.142
Perfiles
Número 27 Vol.1 (2022)

PREDICCIÓN DE LA SITUACIÓN ACADÉMICA EN ALUMNOS DE PREGRADO USANDO ALGORITMOS DE MACHINE LEARNING.

Prediction of academic status in undergraduate students using machine learning algorithms.

Jesús Eduardo Gamboa Unsihuay, Jesús Walter Salinas Flores *

Universidad Nacional Agraria la Molina, Facultad de Economía y Planificación, Departamento de Estadística e Informática, Lima, Perú.

*jsalinas@lamolina.edu.pe

Resumen

El rendimiento académico de un estudiante universitario generalmente se mide a través de sus calificaciones, las cuales derivan en una situación académica normal o deficiente, que a su vez depende de diversos factores. El objetivo de esta investigación fue encontrar los principales predictores de la situación académica de un estudiante universitario luego de que transcurrieron seis semestres desde su ingreso. Para el análisis de datos, se hizo uso del algoritmo Boruta para seleccionar variables predictoras y se aplicaron doce algoritmos de clasificación, previa partición de los datos en conjuntos de entrenamiento y evaluación. Luego, se eligieron aquellos modelos con mejores valores de sensibilidad, especificidad y balanced accuracy. Finalmente, se empleó un ensamble y un punto de corte óptimo para mejorar las predicciones. Los modelos con mejor desempeño fueron el de regresión logística, Naive Bayes y máquinas de soporte vectorial con kernel lineal. Al aplicar el ensamble con punto de corte óptimo se obtuvo especificidad de 0.695 y sensibilidad de 0.947. La nota obtenida en el curso de Matemáticas fue una de las más importantes para predecir la situación académica luego de seis semestres de estudios, mientras que las variables sociodemográficas no fueron relevantes.

Palabras clave: Ensamble, minería de datos, Boruta, corte óptimo.

Abstract

The academic performance of a university student is generally measured through grades, which derive in a normal or deficient academic situation, depending in turn on several factors. The objective of this research was to find the main predictors of a university student's academic status after six semesters have elapsed since admission. For data analysis, the Boruta algorithm was used to select predictor variables and twelve classification algorithms were applied, after partitioning the data into training and evaluation sets. Then, those models with the best sensitivity, specificity and balanced accuracy values were chosen. Finally, an optimal assembly and cut-off point were used to improve predictions. The models with the best performance were logistic regression, Naive Bayes and vector support machines with linear kernel. When applying the optimal cut-off assembly, the specificity was 0.695 and sensitivity 0.947. The grade obtained in the mathematics course was one of the most important predictors of academic status after six semesters of study, while sociodemographic variables were not relevant.

Keywords: Ensemble, data mining, Boruta, optimal cut-off.

Fecha de recepción: 28-08-2021 **Fecha de aceptación:** 14-09-2021 **Fecha de publicación:** 31-01-2022

I. INTRODUCCIÓN

En Perú, un estudiante universitario de pregrado es aquel que ha concluido sus estudios de educación secundaria, ha aprobado el proceso de admisión a una universidad, ha alcanzado vacante y se encuentra matriculado en ella. Así, la Universidad Nacional Agraria La Molina de Lima, Perú (UNALM), es una institución educativa que brinda formación en 12 carreras universitarias de pregrado relacionadas al uso y gestión de recursos agropecuarios y la conservación del medio ambiente, organiza procesos de admisión semestrales, mediante los cuales los postulantes buscan alcanzar una vacante para acceder a los estudios universitarios, a través de la resolución de un examen que mide sus conocimientos.

El proceso de admisión se puede dividir en tres etapas: inscripción, examen y asignación de vacantes. En la etapa de inscripción, el postulante, además de brindar sus datos entre los que se encuentra la modalidad de ingreso, la cual se define según los requisitos que el postulante cumple, siendo las más comunes la de concurso ordinario, centro preuniversitario (CEPRE), primer y segundo puesto de colegio, y quinto de secundaria.

La segunda etapa consiste en el examen de admisión, el cual está compuesto preguntas concernientes a las áreas de Razonamiento matemático, Razonamiento Verbal, Matemática, Física, Química y Biología. Sin embargo, existen excepciones en cuanto a los postulantes de algunas modalidades quienes rinden un examen distinto: Traslados Externos y Graduados y Titulados y CEPRE. Finalmente, el proceso de asignación de vacantes se realiza en estricto orden de mérito para cada una de las modalidades del proceso de admisión, es decir aquellos que optaron por distintas modalidades no compiten por una misma vacante.

Luego de haber conseguido una vacante, el ingresante confirma su incorporación a la UNALM realizando su matrícula del primer semestre. En su condición de alumno matriculado y en función a su rendimiento académico, se le adjudica una de las siguientes situaciones académicas: Normal, Observado, Suspendido, Prueba o Separado. La situación Normal es asignada automáticamente a los estudiantes de primer año (dos primeros semestres) y a aquellos que mantienen su promedio semestral en un valor mayor o igual a 11. Las demás situaciones académicas comienzan a regir a partir del tercer semestre. Así, un estudiante es Observado si su último promedio semestral es inferior a 11, y es Suspendido si sus 2 últimos promedios semestrales son inferiores a 11, lo cual le imposibilita la matrícula en el semestre académico siguiente. Luego de subsanar una observación, la situación académica futura puede volver a ser Normal, sin embargo, esto sucede después de subsanar una suspensión, su condición pasaría a ser Normal con antecedente. Finalmente, la situación de Prueba es aquella que presenta el estudiante que se matricula luego de una suspensión. Si en esta situación, vuelve a reportar un promedio semestral inferior de 11, pasa automáticamente a la situación de Separación académica, con la cual pierde la condición de estudiante de la UNALM.

En la literatura se pueden encontrar diversos estudios acerca del rendimiento académico en estudiantes universitarios. En el estudio llevado a cabo por Gómez-Sánchez, Martínez-López, Oviedo-Marín (10) se encontró que el sexo del estudiante y el semestre de estudios, así como su promedio y satisfacción con la carrera influyen en su desempeño académico. Por otro lado, Ocaña (22), en su investigación, lista un conjunto de potenciales variables académicas que tienen repercusión en el rendimiento académico, entre las que menciona las características del colegio de procedencia, el rendimiento en las pruebas de admisión, el desempeño universitario en el año previo al del estudio, la vocación, entre otras.

Por su parte, en la investigación realizada por Jiménez (14) se mencionan tres factores que inciden en el rendimiento académico: el sexo del estudiante, el acceso a becas y el nivel de uso de tecnologías de la información y la comunicación. Es así que diversos estudios señalan distintos factores que repercuten en el desempeño académico, a lo cual cabe mencionar, tal como lo hace Mora (18), que existe una gran cantidad de factores a los cuales no se suele tener completo acceso, tales

como el entorno familiar, laboral o de salud, pero que, a pesar de ello, la consideración de variables principalmente académicas es de utilidad para la toma de decisiones de los gestores universitarios. También se puede mencionar los trabajos sobre deserción universitaria realizados por Barragán (2), Calvache (3), Montserrat (17), Moreira (19) y Munizaga (20).

En la UNALM, Huertas y Bullón (13) desarrollaron un trabajo en el que evaluaron el rendimiento de los ingresantes del año 2000 luego de cinco años, es decir once semestres académicos después de haber ingresado, llegando a la conclusión de que el 11% logró culminar sus estudios y aproximadamente la mitad se encontraba en situación académica normal, además que la modalidad de ingreso no fue un factor diferenciador en el rendimiento académico.

En un estudio más reciente, llevado a cabo por Delgado (6), se analizó el rendimiento de los ingresantes de los semestres 2017-I y 2017-II mediante su nota obtenida en el curso de Matemática y la cantidad de créditos aprobados en su primer semestre de estudios universitarios. Luego de su análisis, concluyó que la nota de matemática del examen de admisión de la universidad fue la más importante para la clasificación del desempeño académico.

El objetivo de esta investigación consiste en encontrar las variables que permitan predecir la situación académica de un estudiante universitario (normal o deficiente) luego de que transcurrieron seis semestres desde su ingreso, usando algoritmos de Machine Learning.

II. MATERIALES Y MÉTODOS

Se utilizó la metodología CRISP (Cross-Industry Standard Process for Data Mining), la cual es una metodología probada para trabajos de minería de datos e incluye seis fases que pueden apreciarse en la Figura 1 y que comprende: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelación, evaluación y despliegue de resultados. Estas fases son mencionadas por Cichosz (4) y Witten, Frank, Hall, Pall (25).

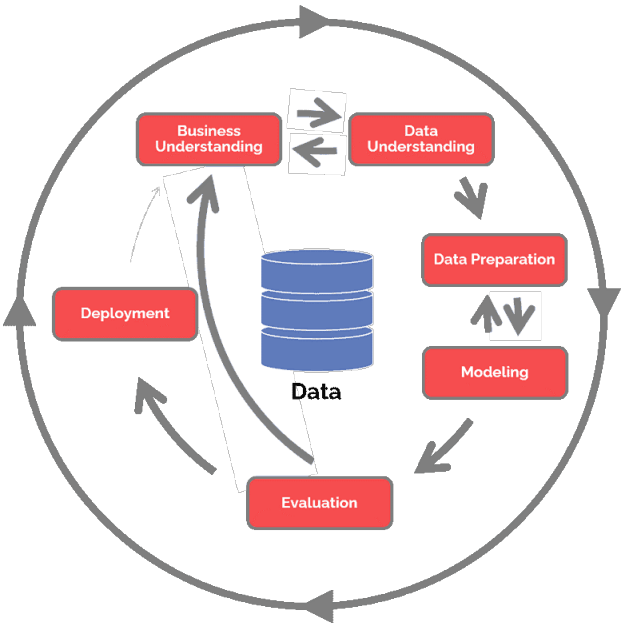

Figura 1. Fases de la Metodología CRISP (Data Science Process Alliance, 2020).

Unidad elemental y variables

Un estudiante que ingresó a la UNALM en los semestres 2017-I o 2017-II por cualquier modalidad excepto Graduados y titulados, Traslado externo y CEPRE, y que además cursó por lo menos un semestre de estudios. El conjunto de datos está compuesto por 622 unidades elementales.

La variable objetivo considerada para el modelo es la situación académica del estudiante 6 semestres después de haber iniciado sus estudios universitarios, la cual es una variable dicotómica que toma el valor 0 si la situación académica del estudiante es Normal u Observado, o 1 para todas las demás situaciones académicas. Las variables predictoras pueden ser agrupadas en tres categorías: sociodemográficas, relacionadas al examen de admisión y referidas al primer semestre de estudios.

Metodología Estadística

Se realizó un análisis descriptivo univariado y bivariado con las variables predictoras y la variable dependiente a predecir (situación académica) y se aplicó el algoritmo BORUTA para la selección de las principales variables predictoras. Este algoritmo duplica el conjunto de datos y mezcla los valores en cada columna. Lantz (16) denomina a estos valores como variables de sombra. Luego, entrena un clasificador usando el algoritmo Random Forest en el conjunto de datos y calcula

el Mean Decrease Accuracy o el Mean Decrease Impurity para cada una de las variables del conjunto de datos. Cuanto mayor sea el puntaje, mejor o más importante es la variable.

El conjunto de datos fue dividido asignando de manera aleatoria el 80% del total de registros para los datos de entrenamiento, y el 20% restante para los datos de evaluación, verificando que en ambas particiones la proporción de estudiantes por situación académica sea similar. Las variables numéricas fueron estandarizadas. Al tener un 84,41% de alumnos en situación académica normal y un 15,59% en situación académica No Normal, se realizó un balanceo de datos utilizando el algoritmo SMOTE, el cual está basado en el principio de oversampling que genera datos artificiales o sintéticos basados en las similitudes del conjunto de variables de la clase minoritaria usando el algoritmo de los vecinos más cercanos o k-nn. Estos algoritmos son descritos por Fernández, García, Galar, Prati, Krawczyk, Herrera (8) y Haibo y Yunqian (11).

Para la etapa de modelamiento se usó la validación cruzada 10-folds para la estimación y selección de hiperparámetros de los modelos. Se usaron los siguientes algoritmos, descritos por Gareth (9) y Hastie (12):

-Regresión logística
-K-NN
-Naive Bayes
-Árbol C5.0
-Árbol CART
-Bagging
-Random Forest
-Gradient Boosting Machine (GBM)
-XGBoosting
-Red Neuronal Perceptrón Multicapa
-Máquina de Soporte Vectorial con kernel lineal (SVL)
-Máquina de Soporte Vectorial con kernel radial (SVM)

Posteriormente, con los tres algoritmos que proporcionaron los mejores indicadores en el entrenamiento y que no estén correlacionados, se realizó un algoritmo de ensamble basado en el promedio de las probabilidades obtenidas y se mejoraron los indicadores usando el punto de corte óptimo sugerido por la curva ROC. Los métodos de ensamble son técnicas para combinar varios algoritmos de aprendizaje con la finalidad de poder construir un algoritmo de aprendizaje más fuerte. Existen ensambles basados en promedio, promedio ponderado y voto mayoritario, descritos por Alfaro (1), Dixit (7), Kumar (15), Narayanachar (21), Rokach (23) y Zhou (26). Al mantener la muestra de evaluación sin balancear se usaron indicadores robustos a esta desproporción, tales como la sensibilidad, la especificidad y el accuracy balanceado.

III. RESULTADOS

Selección de variables predictoras

Como resultado de la comprensión de los datos y aplicando el algoritmo BORUTA se seleccionaron las siguientes variables predictoras numéricas de la situación académica de un alumno:

PUNTAJE.MATEMÁTICAS: Puntaje obtenido en el área de Matemática en el examen de admisión

- PUNTAJE.RM: Puntaje obtenido en el área de Razonamiento matemático en el examen de admisión.
- PUNTAJE.FÍSICA: Puntaje obtenido en el área de Física en el examen de admisión
- PUNTAJE.FINAL: Puntaje obtenido en el examen de admisión
- LENGUA: Nota en el curso de Lengua
- QUIM: Nota en el curso de Química
- MATE: Nota en el curso de Matemáticas
- CREDAP: Número de créditos aprobados en el primer semestre de estudios.
- PROMSEM: Promedio ponderado del primer semestre de estudios.

Análisis descriptivo de las variables predictoras

- En el primer semestre de estudios, la nota promedio de Matemática fue de 11.1 puntos, para Química su media fue de 10.5, mientras que la nota media de Lengua fue igual a 13.3.
- Los estudiantes obtuvieron un primer promedio semestral con media de 12.2 puntos y aprobaron 14.7 créditos en promedio.
- El 23.1% de los estudiantes que obtuvieron hasta 13.4 de nota en el área de Matemáticas (mediana de la variable) en el examen de admisión presentaron una situación académica de riesgo, mientras que en el grupo restante

(más de 13.4 de nota) esta cifra se redujo a casi la mitad (12.9%).

• El 23.8% de los estudiantes que obtuvieron 11 o menos nota en Matemática (mediana de la variable) en el primer semestre presentaron una situación académica de riesgo, mientras que en el grupo restante (más de 11) este indicador alcanzó solo el 6.7%.

• El 24% de los estudiantes que aprobaron hasta 15 créditos (mediana de la variable) en el primer semestre presentaron una situación académica de riesgo, mientras que en el grupo restante (16 a más créditos aprobados) este indicador alcanzó solo el 7.1%.

• El promedio semestral del primer semestre presentó una alta correlación (mayor a 0.8) con al menos una de las demás variables predictoras, por lo que fue retirado del análisis.

Evaluación de los modelos

En la figura 2 se puede observar que los algoritmos que son más estables son la Regresión Logística, Naive Bayes y un SVM con kernel lineal, puesto que con estos se obtuvieron las menores diferencias entre el Balanced Accuracy en los datos de entrenamiento y evaluación, asimismo, alcanzaron el mayor valor en este indicador al utilizar los datos de evaluación. Estos tres algoritmos fueron ensamblados.

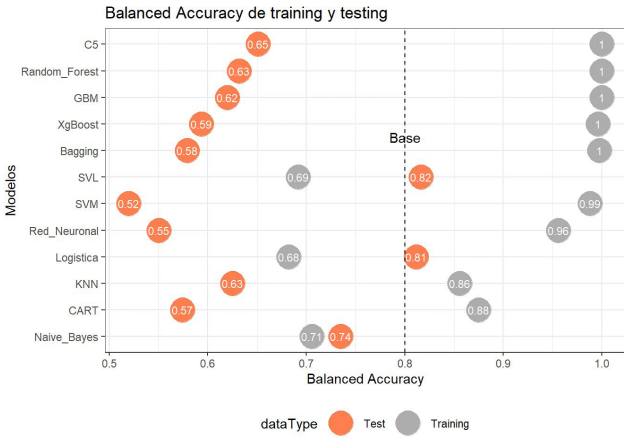

Figura 2. Comparación del Balanced Accuracy para los doce modelos.

Ensamble de modelos

Los resultados de la sensibilidad, especificidad y accuracy balanceado para cada uno de los tres modelos elegidos se muestran en la tabla 1, así también para el ensamble de éstos, usando el punto de corte tradicional (0.5) y el óptimo sugerido por la curva ROC que se aprecia en la figura 3. Con este valor óptimo (0.439), la probabilidad de detectar correctamente a los alumnos en situación académica normal es de 0.695, mientras que la detección de estudiantes cuya situación es no normal se realiza con una probabilidad de 0.947.

Algoritmo	Sensibilidad	Especificidad	Accuracy Balanceado
Ensamble con umbral óptimo	0.9473684	0.6952381	0.8213033
SVM con kernel lineal	0.8421053	0.7904762	0.8162907
Regresión Logística	0.8421053	0.7809524	0.8115288
Ensamble con umbral de 0.5	0.7894737	0.7333333	0.7614035
Naive-Bayes	0.7368421	0.7333333	0.7350877

Tabla 1. Comparación de indicadores para los tres algoritmos y el ensamble con y sin punto de corte óptimo.

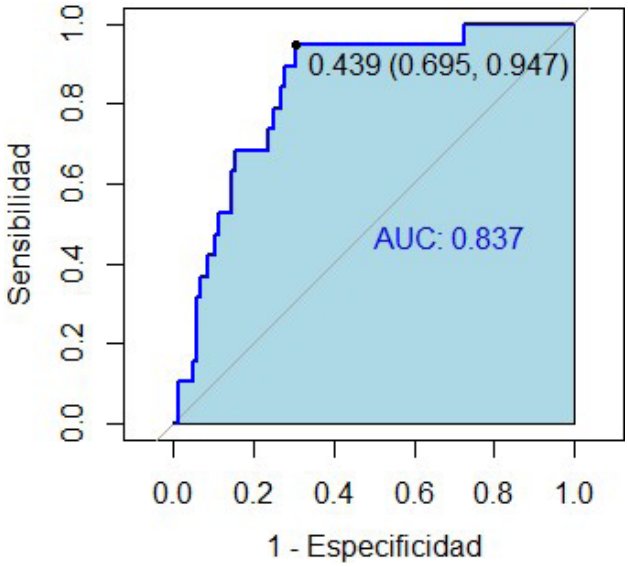

Figura 3. Curva ROC para el ensamble con los tres modelos.

IV. DISCUSIÓN

En los hallazgos reportados de esta investigación, las variables sociodemográficas como el sexo, la edad de ingreso y el tipo de colegio de procedencia del estudiante no contribuyeron en la predicción de la situación académica, a diferencia de lo señalado en las investigaciones de Gómez-Sánchez, Martínez-López, Oviedo-Marín (10) y Cortez, Tutiven, Villavicencio (5). Por otro lado, el estudio de Tapasco, Ruiz y García (24) indica que el puntaje conseguido en el examen de admisión no fue una variable significativa en el promedio del estudiante de carreras de ingeniería y ciencias agropecuarias, lo cual concuerda con lo detectado a través de los algoritmos de Machine Learning para alumnos de la UNALM, puesto

que solo la nota de Razonamiento Matemático estuvo entre las predictoras más significativas y solo en uno de los algoritmos. Esto se debería a que el examen de admisión mide los conocimientos de los postulantes mas no su aptitud hacia la carrera, un factor importante en su desarrollo académico. Finalmente, las notas de Matemática y Química en el primer semestre siempre aparecen en el grupo de mejores predictoras en los tres modelos ensamblados. Es importante mencionar también que el trabajo de Ocaña (22) señala al rendimiento académico previo como una variable que repercute a futuro. Así, los hallazgos reportados por el modelo en estudio podrán ser de utilidad para los tutores, quienes están a cargo del seguimiento de estudiantes universitarios en cuanto a su desempeño académico.

V. CONCLUSIONES

- El modelo de regresión logística, y los algoritmos Naive Bayes y Máquinas de Soporte Vectorial con kernel lineal fueron elegidos por tener un Balanced Accuracy altos y estabilidad de resultados en las muestras de entrenamiento y evaluación.
- En los modelos elegidos para el ensamble, las variables relacionadas al primer semestre presentaron mayor importancia que las del examen de admisión. La nota obtenida en el curso de Matemáticas fue una de las más importantes para predecir la situación académica.
- Las variables sociodemográficas no fueron relevantes en la predicción del rendimiento académico luego de seis semestres de estudios.
- El modelo encontrado debe actualizarse rutinariamente dado que los contenidos y/o formas de enseñanza - aprendizaje van variando a lo largo del tiempo.

VI. AGRADECIMIENTOS

Los autores desean agradecer al personal de la UNALM por las facilidades brindadas en la recopilación de la información de las bases de datos de cada oficina.

Referencias

1. Alfaro E, Gámez M, García N. Ensemble Classification Methods with Applications in R. New Jersey: John Wiley & Sons, Ltd.; 2019.
2. Barragán S, González L. Un modelo para explicar la retención en la universidad de Bogotá Jorge Tadeo Lozano: árboles de decisión. Congresos CLABES. 2016. Disponible en: https://revistas.utp.ac.pa/index.php/clabes/article/view/1090
3. Calvache L, Alvarez V, Triviño J, Quiceno C, Pulgarin R. Aplicación de técnicas de minería de datos para la identificación de patrones de deserción estudiantil como apoyo a las estrategias de SARA (sistema de acompañamiento para el rendimiento académico). Congresos CLABES. 2018. Disponible en: https://revistas.utp.ac.pa/index.php/clabes/article/view/2021
4. Cichosz P. Data Mining Algorithms: Explained Using R. New Jersey: John Wiley & Sons, Ltd.; 2015.
5. Cortez F, Tutiven J, Villavicencio M. Determinantes del rendimiento académico universitario. Revista Publicando. 2017; 4(10): 284 - 296
6. Delgado R. Uso de los métodos multivariante para el análisis del desempeño académico de los estudiantes de la educación superior (Caso: Estudiantes ingresantes en el primer curso de Matemática de la UNALM) [tesis de maestría]. Perú: UNMSM; 2020
7. Dixit A. Ensemble Machine Learning. A beginner's guide that combines powerful machine learning algorithms to build optimized models. United Kingdom: Packt Publishing Ltd.; 2017.
8. Fernández A, García S, Galar M, Prati R, Krawczyk B, Herrera, F. Learning from Imbalanced Data Sets. New York: Springer; 2018.
9. Gareth J, Witten D, Hastie T, Tibshirani R, 2013. An Introduction to Statistical Learning: with Applications in R. New York: Springer Texts in Statistics; 2013.
10. Gómez-Sánchez D, Martínez-López E, Oviedo-Marín R. Factores que influyen en el rendimiento académico del estudiante universitario. Tecnociencia. 2011; 5(2): 90 – 97.
11. Haibo H, Yunqian M. Imbalanced Learning: Foundations, Algorithms, and Applications.

New Jersey: John Wiley & Sons, Ltd. The Institute of Electrical and Electronics Engineers, Inc.; 2013.
12. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd. ed. New York: Springer; 2001.
13. Huertas C, Bullón C. Evaluación del desempeño de los alumnos de la UNALM según su modalidad de ingreso. Anales Científicos. 2009. 70(3): 58-70.
14. Jiménez M. Análisis cuantitativo de las variables que influyen en el rendimiento universitario. RIDE Revista Iberoamericana para la Investigación y el Desarrollo Educativo. 2018. 9(17): 623 – 638.
15. Kumar I, Jaim M. Ensemble Learning for AI Developers. Learn Bagging, Stacking, and Boosting. Methods with Use Cases. New York: Springer Science+Business Media; 2020.
16. Lantz B. Machine Learning with R. United Kingdom: Packt Publishing Ltd.; 2019.
17. Montserrat V, González J, Patricio J. Modelo predictivo para la permanencia en la Educación Superior. Congresos CLABES. 2017. Disponible en: https://revistas.utp.ac.pa/index.php/clabes/article/view/1588
18. Mora R. Factores que intervienen en el rendimiento académico universitario: Un estudio de caso. Repositorio Institucional de la Universidad de Alicante [Internet]. 2021 [citado el 15 Agosto 2021]. 6: 1041 – 1063. Disponible en http://rua.ua.es/dspace/handle/10045/52320#vpreview
19. Moreira T, Hernández M, Solís M, Fernández T. Estudio descriptivo del perfil desertor en tres cohortes de estudiantes universitarios de primer ingreso. Congresos CLABES, 38-49. 2020. Disponible en: https://revistas.utp.ac.pa/index.php/clabes/article/view/2622
20. Munizaga F, Rojas-Murphy A, Leal R. Variables que Influyen en la retención de estudiantes de primer año en un programa de bachillerato chileno. Congresos CLABES. 2018. Disponible en: https://revistas.utp.ac.pa/index.php/clabes/article/view/1892
21. Narayanachar P. Hands-On Ensemble Learning with R. United Kingdom: Packt Publishing Ltd.; 2018.
22. Ocaña Y. Variables académicas que influyen en el rendimiento académico de los estudiantes universitarios. Investigación Educativa. 2011. 15(27): 165 – 180.
23. Rokach L. Pattern Classification using Ensemble Methods. Series in Machine Perception and Artificial Intelligence – Vol. 75. Singapur: World Scientific Publishing Co. Pte. Ltd.; 2010.
24. Tapasco O, Ruiz F, Osorio D. Estudio del poder predictivo del puntaje de admisión sobre el desempeño académico Universitario. Revista Latinoamericana de Estudios Educativos (Colombia). 2016. vol. 12, núm. 2, pp. 148-165.
25. Witten I, Frank E, Hall M, Pal C. Data Mining: Practical Machine Learning Tools and Techniques. 4er. ed. Massachusetts: Morgan Kauffman; 2019.
26. Zhou Z. Ensemble Methods. Foundations and Algorithms. Florida: Chapman & Hall/CRC. Machine Learning & Pattern Recognition Series; 2017.