Modelo de predicción del rendimiento académico para el curso de nivelación de la Escuela Politécnica Nacional a partir de un modelo de aprendizaje supervisado

Academic performance prediction model for the propedeutic course of the Escuela Politécnica Nacional and the implementation of an automated supervised learning model

ARTICLE HISTORY

Received 04 April 2021
Accepted 17 May 2021

Karen Calva
EcuAnalytics
Quito, Ecuador
karenpris5792@hotmail.com
ORCID: 0000-0002-7280-4724

Miguel Flores
Grupo MODES, SIGTI, Departamento Matemática, Facultad de Ciencias
Escuela Politécnica Nacional
Quito, Ecuador
miguel.flores@epn.edu.ec
ORCID: 0000-0002-7742-1247

Hugo Porras
EcuAnalytics - INsight
Quito, Ecuador
hugo-sxe@hotmail.com
ORCID: 0000-0002-6278-7940

Ana Cabezas-Martínez
Departamento de Estudios Políticos
Facultad Latinoamericana de Ciencias Sociales
Quito, Ecuador
ana.cabezas90@gmail.com
ORCID: 0000-0001-5062-4530

Modelo de predicción del rendimiento académico para el curso de nivelación de la Escuela Politécnica Nacional a partir de un modelo de aprendizaje supervisado

Academic performance prediction model for the propedeutic course of the Escuela Politécnica Nacional and the implementation of an automated supervised learning model

Karen Calva
EcuAnalytics
Quito, Ecuador
karenpris5792@hotmail.com
ORCID: 0000-0002-7280-4724

Miguel Flores
Grupo MODES, SIGTI, Departamento Matemática, Facultad de Ciencias
Escuela Politécnica Nacional
Quito, Ecuador
miguel.flores@epn.edu.ec
ORCID: 0000-0002-7742-1247

Hugo Porras
EcuAnalytics - INsight
Quito, Ecuador
hugo-sxe@hotmail.com
ORCID: 0000-0002-6278-7940

Ana Cabezas-Martínez
Departamento de Estudios Políticos
Facultad Latinoamericana de Ciencias Sociales
Quito, Ecuador
ana.cabezas90@gmail.com
ORCID: 0000-0001-5062-4530

***Resumen*—** En el presente artículo se aplica un modelo de aprendizaje automático supervisado que predice la probabilidad de que un estudiante de la Escuela Politécnica Nacional apruebe el curso de nivelación. Para llevar a cabo esta tarea se describe una metodología estadística basada en gradient boosting y regresión logística donde el problema de aprendizaje se formula en términos de la minimización de la función de error mediante el método del descenso del gradiente. Para explicar la probabilidad de aprobación se toman en consideración dimensiones sugeridas por la literatura relacionadas a variables socioeconómicas, demográficas, familiares, institucionales y de desempeño académico en la postulación y en el curso de nivelación que tiene el estudiante. Los resultados del modelo de árbol de decisión muestran un nivel de precisión del 96% en el conjunto de datos de prueba, con un área bajo la curva ROC de 89.1, siendo estos niveles generalmente aceptados. Por otro lado, los resultados de la regresión logística sugieren que factores como la calificación ponderada del primer bimestre, la calificación con la que postuló, su jornada de estudios, su ubicación geográfica de origen, entre otras, afectan de una u otra manera a la probabilidad del estudiante, de aprobar el curso de nivelación.

Palabras clave— ***r****endimiento académico, regresión logística, árboles de decisión, GBM, método del descenso del gradiente.*

***Abstract*—** In this article, a supervised machine learning model is applied that predicts the probability that a student of the National Polytechnic School will pass the leveling course. To carry out this task, a statistical methodology based on gradient boosting and logistic regression is described where the learning problem is formulated in terms of the minimization of the error function through the gradient descent method. To explain the probability of approval, dimensions suggested by the literature related to socioeconomic, demographic, family, institutional and academic performance variables are taken into consideration in the application and in the leveling course that the student has. The results of the decision tree model show a precision level of 96% in the test data set, with an area under the ROC curve of 89.1, these levels being generally accepted. On the other hand, the results of the logistic regression suggest that factors such as the weighted qualification of the first two months, the qualification with which they applied, their study schedule, their geographical location of origin, among others, affect in one way or another the probability of the student to pass the leveling course.

***Keywords*—** *academic performance, logistic regression, decision trees, GBM, gradient descent method*

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

K. Clava, M. Flores, H. Porras and A. Cabezas-Martínez, “Modelo de predicción del rendimiento académico para el curso de nivelación de la escuela politécnica nacional a partir de un modelo de aprendizaje supervisado”, Latin-American Journal of Computing (LAJC), vol. 8, no. 2, 2021.

I. Introducción

En la actualidad, el rendimiento académico de los estudiantes es uno de los indicadores de calidad más representativos del quehacer académico de las universidades [1], y de este, principalmente en los Cursos de Nivelación (CN) o Propedéutico, depende la oferta académica de los centros de estudio.

El rendimiento académico se considera como el resultado cuantitativo de la comprensión del contenido de los programas de estudio, representado como notas dentro de una escala convencional; mismo que es generado del proceso de enseñanza aprendizaje. Este es obtenido de las tareas, evaluaciones y otras actividades planificadas por los docentes. El rendimiento académico es uno de los factores que muestran el nivel de conocimiento adquirido por el estudiante. Además, se lo considera como uno de los criterios para determinar el éxito o fracaso de los estudiantes a través de un sistema de evaluación [2]. Actualmente, uno de los problemas a los cuales hacen frente las Instituciones de Educación Superior, para el efecto universidades, es el bajo rendimiento académico en los primeros años de la universidad. Esto conlleva a que durante los primeros semestres de formación profesional de los estudiantes se evidencie una “barrera académica”, misma que obstaculiza la oferta de cupos esperada por las universidades, ya que a corto plazo es imposible para estas instituciones incrementar los espacios físicos, donde se imparte la cátedra, así como, la planta docente [3].

El presente artículo busca entender el rendimiento de los estudiantes en el CN de la Escuela Politécnica Nacional; a la vez que se ajusta un modelo que pueda predecir la probabilidad de reprobación de estos. Con este modelo se busca tomar acciones tempranas en su beneficio y planificar de mejor manera la oferta de cupos del próximo periodo en el CN y en cada carrera.

En este sentido, en la sección dos se presentan algunos factores influyentes sobre el rendimiento académico sugeridos por la literatura, así como, la preponderancia de la regresión logística como metodología de análisis de este problema. En la sección tres se explica las metodologías empleadas para predecir la reprobación del curso de nivelación (gradient boosting machine) y para realizar inferencia sobre los factores influyentes sobre ella (regresión logística). En la cuarta sección se explica la construcción de las variables dependiente e independientes, así como la construcción de los modelos estadísticos aplicados. Finalmente, en la quinta sección, se presentan los resultados de ambos modelos y en la sexta sección se describe las conclusiones y futuras investigaciones.

II. Marco Teórico

A. *Factores influyentes en el rendimiento académico*

En la actualidad, el rendimiento académico de los estudiantes es uno de los indicadores de calidad y eficiencia más importantes del quehacer académico de las universidades [1], y de este, principalmente en los Cursos de Nivelación (CN) o Propedéutico, depende la oferta académica de los centros de estudio en el contexto ecuatoriano. El bajo rendimiento académico durante los primeros años de estudios universitarios representa un grave problema ya que, el alto porcentaje de reprobados en las asignaturas impartidas en los semestres iniciales genera una “barrera académica”, que imposibilita que las universidades cuenten con la oferta de cupos deseada, ya que no se dispone de los recursos físicos y académicos necesarios [3].

En este sentido, se conoce que el rendimiento académico es considerado como el producto de la asimilación del contenido de los programas de estudio, expresado como un resultado cuantitativo, de la ejecución del proceso de enseñanza aprendizaje, denominado como calificaciones. Estos son obtenidos como resultado de las evaluaciones y otras actividades realizadas durante el quehacer docente. En este contexto, se entiende como rendimiento académico al nivel de conocimiento obtenido por el estudiante durante la jornada académica. Así mismo, es considerado como uno de los criterios para determinar el éxito o fracaso de los estudiantes [2]. Es así como, la predicción y el entendimiento de los factores que ayudan a explicar el rendimiento académico se ha vuelto relevante en el contexto de la educación superior, motivo por el cual son varios los autores que han buscado predecirlo y a la vez de explicarlo.

La referencia [1] menciona que el rendimiento académico es el resultado de la combinación de distintos factores multicausales que incluyen aspectos sociodemográficos, psicosociales, pedagógicos, institucionales y socioeconómico. Este mismo autor menciona que ejemplos de estos factores yacen sobre temáticas como la motivación, ansiedad, entusiasmo y autoestima del estudiante, características del docente, percepción del clima académico, y más situaciones que se ven directamente influenciadas por la toma de decisiones institucionales.

Por otro lado, en un caso de estudio internacional (Italia), otros autores como [4] han estudiado que las variables explicativas del rendimiento académico de los estudiantes de educación superior parten de determinantes personales, su bagaje y redes familiares, e intrínsecos a su institución de estudio. En principio, los determinantes personales incluyen factores como: las aptitudes académicas, habilidades y comportamientos; mientras que el bagaje y las redes familiares hablan de la atención de los padres, el contexto socioeconómico, etc. Las variables sociodemográficas en cambio incluyen: género, edad, estado civil, etc. Finalmente, los factores intrínsecos a la institución se relacionan a la estructura del semestre, el pénsum de estudios, etc.

Para el caso ecuatoriano, se han analizado las distintas causas de deserción en el curso de nivelación previo al ingreso al primer año de las facultades del área técnica de la Universidad de Cuenca, donde se ha argumentado que la deserción es un problema que ocurre en esta etapa en particular, causándole pérdidas de tiempo y recursos al estudiante y a la institución universitaria. Sus resultados sugieren que algunas de las causas de deserción yacen en la carencia de bases en temas de matemáticas, entrar a una carrera que no es de su preferencia, la existencia de un sistema de evaluación complejo y falta de recursos económicos.

Al final, el determinar analíticamente los factores que influyen en el rendimiento académico de los estudiantes permitirá implementar medidas adecuadas para combatir la alta tasa de reprobación, y ayudará a predecir con antelación el número de estudiantes que aprobarán el CN y los que lo harán en segunda matrícula. Al conocer esta predicción se busca tomar acciones tempranas en su beneficio y planificar

de mejor manera los cupos del próximo periodo en el CN y en cada carrera.

Hasta donde llega el conocimiento de este trabajo, pese al extenso estudio del rendimiento académico, la mayor parte de investigaciones realizadas se basan en el uso de estadística descriptiva y modelos de regresión logística [5], [3]. En esta línea, se busca aportar a la literatura con una metodología de predicción con mayor exactitud y robustez, que permita capturar efectos no lineales en los datos [6], a la par del uso de metodologías clásicas para realizar inferencia sobre los factores que influyen en el rendimiento académico.

III. Metodología Analítica

En esta sección se describen las dos metodologías utilizadas en este trabajo: regresión logística y gradient boosting machine, así como algunas nociones y definiciones teóricas necesarias para comprender la construcción de los modelos de predicción e inferencia.

A. Problemas de clasificación y su modelamiento

Para comenzar la explicación de las metodologías es preciso definir el aprendizaje estadístico y los problemas de clasificación.

Acorde a la referencia [7], el concepto de aprendizaje estadístico o "statistical learning" se refiere a un amplio conjunto de herramientas utilizadas para entender datos. Dentro de él existen dos tipos de problemas supervisados (i.e. que poseen una variable dependiente que puede ser estimada): regresión y clasificación. En el problema de regresión, se estiman modelos que puedan predecir una variable de respuesta cuantitativa, mientras que en el problema de clasificación se busca predecir una variable dependiente categórica.

Así, para modelar una variable dependiente categórica existen varias técnicas, de las cuales se explican dos en específico, útiles para las finalidades de este estudio: gradient boosting machine y modelos de regresión logística.

Al final, se explican las metodologías de selección y evaluación de modelos.

B. Gradient Boosting Machine

Para comprender estos modelos se revisa primero la definición de los métodos aditivos generalizados y de los métodos basados en árboles, para luego explicar su combinación en los modelos gradient boosting machine (GBM).

C. Métodos aditivos generalizados

Según la referencia [7], los modelos de regresión lineal juegan un papel importante en muchos análisis de datos proporcionando reglas de predicción y clasificación. Y aunque estos son atractivos por su simpleza, a menudo fallan en situaciones de la vida real porque los efectos de las covariables sobre la variable dependiente no suelen ser lineales. Es por ello que a continuación, se describen métodos estadísticos flexibles y automáticos que pueden usarse para identificar y caracterizar los efectos de regresión que no son lineales. Estos métodos usualmente se denominan como "modelos aditivos generalizados", los cuales están descritos por la ecuación 1.

$$E(Y|X_1, X_2, \dots, X_p) = \alpha + f_1(X_1) + f_2(X_2) + \dots + f_p(X_p) \quad (1)$$

Donde $X_1, X_2, \dots, X_p$ representan a las variables predictoras o independientes, Y es la variable de salida o dependiente y las f_j's son funciones suaves no especificadas (no paramétricas).

Estos modelos aditivos proporcionan una extensión útil de los modelos lineales, haciéndolos más flexibles y conservando gran parte de su capacidad de interpretación.

Métodos basados en árboles: La referencia [7] proporciona también un resumen completo de los métodos basados en árboles. Este tipo de metodologías busca dividir el espacio de variables en un conjunto de rectángulos y luego ajustan un modelo simple (como una constante) en cada uno. Estos son conceptualmente simples pero potentes, y una ventaja clave del árbol binario recursivo es su interpretabilidad. Este tipo de modelos puede ser utilizado tanto para regresión como para clasificación.

Los árboles de clasificación en específico buscan decidir automáticamente el cómo dividir las variables y en qué puntos, además de la topología (o forma) que el árbol debería tener, de tal manera que este sea capaz de predecir una variable de respuesta categórica. Para ello, se usa como función de pérdida una medida de impureza conocida como el error de clasificación, la cual mide la proporción de observaciones mal clasificadas en el árbol estimado.

D. Modelos boosting y árboles aditivos

Acorde a la referencia [7], el boosting es uno de los modelos de aprendizaje para clasificación más poderosos introducidos en los últimos veinte años. Su motivación nace de usar un procedimiento que combine los resultados de muchos clasificadores "débiles" para producir una clasificación "fuerte". Su éxito realmente no es tan misterioso ya que es una manera de estimar una expansión aditiva en un conjunto de funciones "base" elementales, donde las funciones bases son los clasificadores individuales $G_m(x) \in \{-1,1\}$. Típicamente estos modelos se estiman minimizando una función de pérdida promedio sobre los datos de entrenamiento.

E. Algoritmo Gradient Boosting

Acorde a la referencia [7], el algoritmo gradient boosting es una técnica de aprendizaje automático utilizado para problemas de regresión y clasificación, donde este produce un modelo predictivo en forma de un conjunto de modelos de predicción débiles, los cuales son típicamente árboles de decisión. Este modelo es construido de forma escalonada como lo hacen otros métodos de boosting, y los generaliza permitiendo la optimización arbitraria de una función de pérdida diferenciable. La referencia [6] especifica que este tipo de modelos produce resultados competitivos y altamente robustos para problemas de regresión y clasificación.

De la implementación de este algoritmo se pueden destacar las siguientes ventajas y desventajas:

Ventajas:

- El algoritmo proporciona una buena precisión predictiva.
- Mucha flexibilidad: se puede optimizar las diferentes funciones de pérdida y los hiperparámetros para que la función se ajuste de mejor forma.
- No se requiere procesamiento previo de datos: a menudo funciona muy bien con valores categóricos y numéricos tal como están.
- Maneja datos faltantes: no se requiere imputación. Es recomendado en bases de datos con datos atípicos.

Desventajas:

- El algoritmo GB continúa mejorando para minimizar todos los errores. Esto puede enfatizar a los valores atípicos y causar un sobreajuste. Para solucionar esta problemática se incorpora validación cruzada.
- Computacionalmente caro: el algoritmo a menudo requiere muchos árboles (> 1000) que pueden ser exhaustivos en tiempo y memoria.
- Su alta flexibilidad da como resultado muchos parámetros que interactúan e influyen fuertemente en el comportamiento del enfoque (número de iteraciones, profundidad del árbol, parámetros de regularización, etc.). Esto requiere una búsqueda exhaustiva durante el ajuste.
- Al ser la combinación de cientos e incluso miles de modelos, no es intuitivamente interpretable, para lo cual se ajustará otro modelo.

F. Modelos de regresión logística

Además, de los modelos basados en árboles, existen otras metodologías cuando se da el caso en que la variable dependiente Y cae dentro de una de dos (o más) categorías. Este es el caso del modelo de regresión logística, donde se busca estimar la probabilidad de que Y pertenezca a una categoría en particular, asegurando que los valores de esta probabilidad se encuentren entre cero y uno. Los modelos de regresión logística son usados en su mayoría para el análisis de datos y la inferencia, cuyo objetivo es entender el rol de variables de entrada importantes al explicar la variable respuesta Y [8].

G. Evaluación y selección de modelos

Luego de haber modelado los datos, se debe escoger el mejor modelo a ser utilizado, lo cual se consigue a través de varias técnicas de evaluación, descritas en esta sección y consultadas en [7].

H. Estimación del error de predicción en la muestra

La forma general para la estimación del error en la muestra se describe en la ecuación 2.

$$\widehat{\mathrm{Err}}_{\mathrm{in}} = \overline{\mathrm{err}} + \hat{\omega} \quad (2)$$

Donde $\hat{\omega}$ es la estimación del optimismo promedio. Usando este criterio se ajusta el error de entrenamiento por un factor proporcional al número de funciones base usadas.

En el mismo sentido, el criterio de información de Akaike es un estimador similar pero más generalmente aplicable de Err_{in} cuando se usa una función de pérdida del logaritmo de la verosimilitud.

Para el caso de la regresión logística, usando el logaritmo de verosimilitud binomial se tiene que el criterio de información de Akaike viene definido por la ecuación 3.

$$AIC = -2\frac{2}{N}\mathrm{loglik} + 2\frac{d}{N} \quad (3)$$

Para usar este criterio en la selección de un modelo se escoge aquel modelo que dé el menor AIC sobre el conjunto de modelos considerados.

I. Validación cruzada

Probablemente la forma más simple y ampliamente utilizada para estimar el error de predicción (en el caso de esta investigación, para el modelo GBM) es la validación cruzada. Este método estima directamente el error en varias muestras, estimando así un error de generalización promedio cuando el modelo $\hat{f}(X)$ es aplicado a una muestra de prueba independiente de la distribución conjunta de X e Y. Así, se espera que la validación cruzada estime el error condicional con el conjunto de datos de entrenamiento fijo; sin embargo, las estimaciones de la validación cruzada estiman bien únicamente la esperanza del error de predicción.

Para llevar a cabo de manera correcta la validación cruzada se divide aleatoriamente la muestra en $K = 10$ grupos de validación cruzada (aunque se podrían utilizar más o menos grupos). Para cada grupo se encuentra un conjunto de "buenos" predictores que muestren una correlación invariada considerable con respecto a la variable dependiente usando todas las observaciones, excepto las que están en el grupo de análisis. Luego, usando solo el subconjunto de predictores, se construye un clasificador multivariado, usando todas las observaciones excepto las que están en el grupo de análisis. Al final se usa el clasificador para predecir la variable dependiente en las observaciones del grupo k.

J. Algoritmo genético para selección del mejor modelo lineal

La referencia [8] plantea, por el lado de la regresión logística, cuando se modelan este tipo de problemas, se estiman varios modelos en búsqueda de un modelo parsimonioso que envuelva un subconjunto de variables de entrada adecuado. Es por ello que para los objetivos de este estudio y para realizar inferencia sobre los determinantes de la probabilidad de reprobación del curso de nivelación de la Escuela Politécnica Nacional, se ha decidido utilizar una aproximación a través de un algoritmo genético.

En este contexto, cuando se piensa en la selección del mejor modelo, la primera idea en mente sería una aproximación a "fuerza bruta", es decir, estimar todos los modelos posibles y luego seleccionar de entre ellos, el mejor, a partir de algún criterio de información. Lamentablemente, este ejercicio puede ser computacionalmente inviable. Para ello, se usa un algoritmo genético que explora solo un subconjunto de todos los posibles modelos, con sesgo hacia los mejores, que gracias a un criterio de selección se vuelve mucho más rápido. Este algoritmo genético es eficiente explorando espacios discretos y puede converger aún con problemas muy complicados.

K. Evaluación del desempeño de modelos de clasificación

Finalmente, en esta sección se analizan las medidas existentes para evaluar los resultados de un proceso de modelización para un problema de clasificación. El objetivo es cuantificar de alguna manera la calidad del ajuste de la solución que se haya encontrado y hacer posible la

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

comparación entre varios modelos, sean de la misma metodología o no.

Cuando se evalúan modelos de clasificación, las medidas de desempeño se calculan comparando las predicciones generadas por este para la muestra de prueba o validación, contra las clases verdaderas del mismo conjunto. Algunas medidas comunes para llevar a cabo esta tarea se describen a continuación, obtenidas de [7].

L. Matriz de confusión

La matriz de confusión es una tabla de doble entrada que permite observar los errores cometidos por el modelo de clasificación entrenado. Esta matriz es conocida además como la matriz de errores, la cual muestra el número de observaciones correcta e incorrectamente clasificadas, donde a cada celda se le asigna una etiqueta distinta. De estas celdas se derivan algunas métricas de desempeño que permiten cuantificar la bondad de ajuste del modelo. El uso de estas medidas dependerá del problema que se esté analizando.

M. Curvas ROC

Una curva Receiver Operating Characteristic o ROC mide el rendimiento respecto a los falsos positivos FP y verdaderos positivos TP. Su diagonal se interpreta como un modelo generado aleatoriamente, mientras que valores inferiores a ella se consideran peores que una estimación aleatoria de nuevos datos.

IV. DATOS, VARIABLES Y CONSTRUCCIÓN DE LOS MODELOS

La Dirección de Gestión de la Información y Procesos (DGIP) almacena, gestiona y administra los recursos informáticos y tecnológicos de la Escuela Politécnica Nacional a través del Sistema de Administración e Información Estudiantil (SAEW); ésta es la fuente subyacente que sirve como insumo de las variables que se consideran en el presente estudio.

Para llevar a cabo el modelamiento se divide aleatoriamente el conjunto de datos en tres partes: para el conjunto de entrenamiento y validación se toma aleatoriamente el 70% y 30% de los periodos 2017-A, 2017-B, 2018-A y 2018-B respectivamente; y para el conjunto de prueba se toma el periodo 2019-A.

A. Evaluación del desempeño de modelos de clasificación

La variable dependiente Y representa el rendimiento académico del estudiante en el Curso de Nivelación (CN), por tanto, es una variable binaria que toma el valor de 1 para los individuos etiquetados como Reprueba y 0 para los estudiantes etiquetados como Aprueba:

$$Reprueba = \begin{cases} 1 & si\ el\ estudiante\ reprueba, \\ 0 & si\ no. \end{cases} \quad (4)$$

En la Tabla I, se puede observar la distribución de los estudiantes que aprueban/reprueban el curso de nivelación por semestre. En esta se puede notar que cada semestre es mayor el porcentaje de estudiantes que reprueban con respecto al total de inscritos.

TABLA I. TASA DE APROBACIÓN POR SEMESTRE

Muestra	Periodo	Número de estudiantes	Rendimiento académico	Número de estudiantes	Porcentaje de estudiantes
Entrenamiento y validación	2017-A	1505	Aprueba	228	15.1%
			Reprueba	1277	84.9%
	2017-B	2364	Aprueba	567	24.0%
			Reprueba	1797	76.0%
	2018-A	2389	Aprueba	493	20.6%
			Reprueba	1896	79.4%
	2018-B	2367	Aprueba	313	13.2%
			Reprueba	2054	86.8%
Prueba	2019-A	2387	Aprueba	582	24.4%
			Reprueba	1805	75.6%

B. Variables independientes

Para el presente análisis se consideran como variables explicativas o independientes a las siguientes, acorde a lo sugerido por estudios previos.

- *Sociodemográficas*: Sexo, Estado civil, Etnia y Edad.
- *Bagaje y familiares:* Número de miembros en el núcleo familiar, Ingreso mensual, Tipo colegio y Residencia.
- *Académicas:* Promedio ponderado del primer bimestre, Calificación de postulación, Calificación del primer bimestre, Número de materias tomadas, Número de matrícula, Segmento poblacional y Número de créditos' por materia.
- *Institucionales:* Curso de nivelación, Semestre del año, Jornada, Materia y Carrera a la que aspira.

C. Construcción del modelo predictivo

La construcción del modelo se realiza mediante aprendizaje supervisado, el cual parte de casos particulares (experiencias) y obtiene casos generales (modelos o reglas) [10]. El aprendizaje supervisado no depende de un experto para "deducir" una regla (modelo o hipótesis) que sirva para describir el conocimiento; por tanto, la ventaja del aprendizaje supervisado es que puede automatizarse [11]. Ya que la variable a predecir Reprueba es binaria, se emplea un modelo de aprendizaje supervisado para clasificación.

El algoritmo de Gradient Boosting de la sección anterior es el escogido como modelo de predicción de la variable Reprueba. Este es un algoritmo popular de aprendizaje automático que ha demostrado ser exitoso en muchos campos y es uno de los métodos líderes para ganar las competencias en Kaggle.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

Para explicar los factores que influyen sobre la probabilidad de la variable Reprueba se utiliza en cambio un modelo de regresión logística, dada su alta interpretabilidad.

V. RESULTADOS

A. Resultados del modelo predictivo

Para el modelo GBM se observará primero el número de árboles necesarios para que el modelo alcance su óptimo. Específicamente, en la Figura 1, la línea vertical indica el número de árboles que se necesitan para que la función de perdida alcance su punto mínimo.

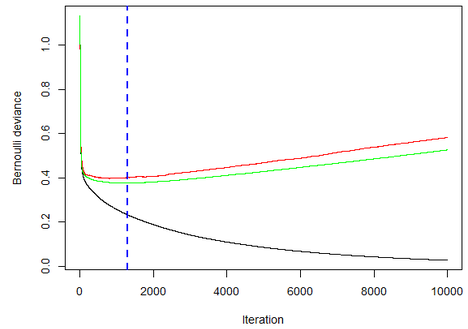


Fig. 1. Error de entrenamiento (negro), error de validación (rojo) y error de prueba (verde) con validación cruzada a medida que se agregan más árboles al algoritmo GBM. El número óptimo de árboles es 1288.

Una característica fundamental en el modelado de GBM es la importancia de las variables. En la Figura 2, se muestran las variables en función de su influencia relativa, que es una medida que indica la importancia relativa de cada variable en el entrenamiento del modelo. Las variables con la mayor disminución promedio en el error de clasificación se consideran las más importantes.

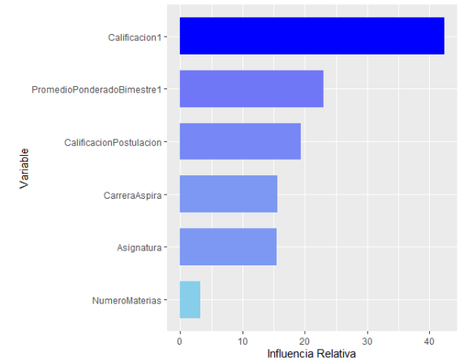


Fig. 2. Las variables con la mayor disminución promedio en el error de clasificación se consideran las más importantes.

Respecto al desempeño del modelo, se evalúa primero los resultados de la matriz de confusión. La Tabla II muestra algunos indicadores de la matriz de confusión para los datos de entrenamiento y de prueba. El punto óptimo de corte sugiere que la tasa de estudiantes que reprueban es mayor a la de los que aprueban. El valor del punto se escogió en función de minimizar el error de clasificación.

TABLA II. ESTADÍSTICOS DE LA MATRIZ DE CONFUSIÓN

Muestra	Optimal Cut Off	MSE	TPR	FPR	Specificity
Entrenamiento	0.69	0.02	0.99	0.04	0.96
Prueba		0.10	0.96	0.09	0.94

Para el error de clasificación (MSE) de la tabla anterior, se tiene que este es ligeramente menor en el conjunto de entrenamiento que en el conjunto de prueba; ya que el conjunto de datos prueba no ha sido utilizado en el proceso de entrenamiento. Así, se puede concluir que el desempeño del modelo es bueno y este no se encuentra sobreajustado al conjunto de datos de entrenamiento.

Los indicadores para los verdaderos positivos (TPR) y los verdaderos negativos (FPR) nos muestran que hay más tendencia a cometer el error de predecir que Reprueba cuando en realidad no es así, es la razón por la que la tasa de los correctamente clasificados como Reprueba (Especificidad) es tan cercana a uno. Esta conclusión se ve reforzada en la Figura 3.

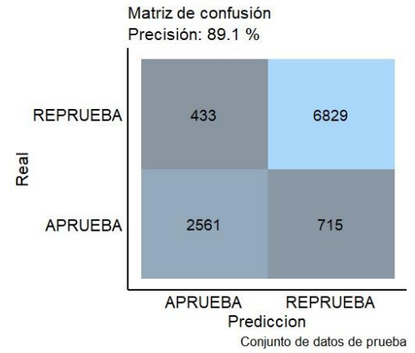


Fig. 3. Matriz de confusión

En la Figura 4, se muestra la curva ROC para los datos de prueba. Dado que el área bajo la curva es cercana a uno, se puede afirmar que el rendimiento del modelo es bueno con respecto a los TPR y los FPR. Así, se ha encontrado un clasificador con un rendimiento muy bueno sin que este se sobreajuste.

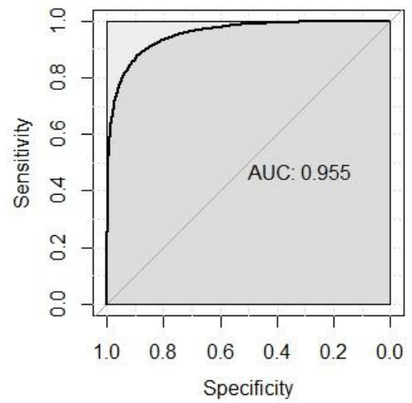


Fig. 4. Curva ROC en la muestra de validación

B. Resultados del modelo inferencial

El modelo inferencial busca mejorar el entendimiento de los factores que influyen en el rendimiento académico de los estudiantes del Curso de Nivelación de la Escuela Politécnica Nacional. La regresión logística es la técnica que se ha seleccionado para cumplir con este objetivo.

Este modelo permite crear un perfil de los estudiantes en base a las variables predictivas; tal que todas en una sola ecuación conjunta explican la probabilidad de aprobar o reprobar de cada estudiante. El conocimiento de los coeficientes y su ponderación es muy importante para conocer los factores que influyen en la reprobación y con ello poder recomendar acciones que ayuden a reducirla. Por ello, en la Tabla II se muestra el mejor modelo inferencial de todos los posibles, estimado a partir de una regresión logística y escogido de un conjunto de estimaciones por tener el mejor criterio de información de Akaike, a través de un algoritmo de selección genético [9].

Para la interpretación de los resultados, se partirá de varios puntos de los puntos de vista revisados en la literatura. Al final, el estudio de estas características puede ayudar desde dos enfoques. A nivel micro, sus conclusiones pueden ser utilizadas para mejorar los planes de apoyo y contingencia de las instituciones de educación superior en temas de logística, admisión y cuidado de sus estudiantes; y desde el punto de vista macro, a formular políticas de educación superior que permitan mejorar la calidad de la educación, tener mayores tasas de aprobación y, por ende, mejores resultados macroeconómicos, a la par de cumplimiento de objetivos de desarrollo social.

Así, para entender la problemática desde un espectro más amplio, se cita principalmente a [4], quienes explican en uno de sus más recientes estudios que la decisión de una persona sobre invertir en la educación terciaria, desde el punto de vista económico, es un proceso secuencial que se va haciendo sobre niveles descendientes de incertidumbre sobre los costos de educación y sus retornos futuros, debido a que los estudiantes actualizan su información disponible y a la vez, su decisión, con cada semestre que pasa. Es decir, un estudiante acabará sus estudios sí y solo sí el valor presente neto de la inversión en su educación (tanto de fuentes pecuniarias como no) es superior a cero. Evidentemente, los costos pecuniarios tienen alta relevancia en esta decisión. Acorde a la referencia [12], [13], un estudiante tendrá éxito en la universidad si logra integración académica y social, obedeciendo a características de su pasado escolar y contexto social. Al final, todos estos costos y determinantes podrían en algún momento, causar que el estudiante falle uno u otro curso, y los determinantes de esta situación son los que se tomarán como guía para interpretar los resultados.

TABLA III. RESULTADOS DEL MODELO INFERENCIAL: SE MUESTRAN LOS NIVELES DE SIGNIFICANCIA AL 0.01, 0.05 Y 0.1

Variable	Estimador	Pr(>|z|)	Significancia
(Intercepto)	8.3894	0	***
Sociodemografía			
Edad	0.0239	0.01266	**
Lugar de residencia: Otras provincias			
Extranjero	0.5565	0.0066	***
Quito	-0.1004	0.0249	**
Estado civil: Casado			
Divorciado	-3.2269	0.0007	***
Soltero	-1.7953	0	***
Unión libre	-1.9019	0.0032	***
Bagaje y familia			
Número de miembros en la familia	0.1391	0	***
Tipo de colegio: Extranjero			
Fiscal	0.6011	0.0379	**
Fiscomisional	0.7508	0.0121	**
Municipal	0.6684	0.026	**
Particular	0.7588	0.0092	***
Características académicas			
Promedio ponderado del primer bimestre	-0.9649	0	***
Calificación de postulación	0.0002	0	***
Calificación del primer bimestre	-0.8633	0	***
Número de materias tomadas: Una			
Dos	0.8765	0.0093	***
Tres	1.4357	0	***
Cuatro	2.3101	0	***
Cinco	2.9183	0	***
Número de matrícula: Primera			
Segunda	0.8121	0	***
Tercera	-0.0533	0.7697	
Segmento poblacional: Acción afirmativa			
GAR	-0.6059	0.0073	***
Mérito territorial	-0.4715	0.0071	***
Población general	-0.2332	0.0107	**
Características institucionales			
Curso de nivelación: Ingeniería y Ciencias Nivel Tecnológico Superior	0.1611	0.0022	***
Semestre del año: A B	-0.3436	0	***
Jornada: Matutina, Vesp.	-0.1314	0.0015	***
Materia: Física			
Matemática	-0.635	0	***
Fundamentos de Química	-1.6994	0	***
Geometría y Trigonometría	-0.2912	0	***
Lenguaje y Comunicación	-2.8853	0	***

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

En primer lugar, se analizan las variables sociodemográficas de los estudiantes y su efecto sobre su probabilidad de reprobar.

La edad presenta un signo significativo y positivo, lo cual sugiere que personas mayores en el curso de nivelación tienen una mayor probabilidad de reprobar. En línea con las referencias [14], [4], esto puede deberse a que personas con más edad pueden sentir la obsolescencia de su conocimiento previo, lo que ocasionará que se incremente su dificultad de estudio. Estudiantes más jóvenes en cambio son más conscientes de la actualidad de sus propias habilidades y aptitudes, lo que les permitirá decidir, estando mejor informados al momento de cursar la nivelación. Sin embargo, cabe la posibilidad de que esta variable interactúe con otros factores sociodemográficos.

Cuando se revisa el lugar de residencia del estudiante, se puede observar que el residir en Quito, reduce la probabilidad de reprobar el curso de nivelación, con respecto a personas de otras ciudades. Este resultado va acorde a la literatura económica, ya que según [15] en un estudio para universidades italianas, se sugiere que personas que provienen de lugares lejanos al de la ubicación de la universidad tienen mayor probabilidad de fallar en sus cursos debido a las dificultades que estos atraviesan para ajustarse a nuevos ambientes con más responsabilidades.

Respecto al estado civil, los resultados muestran que estudiantes divorciados, solteros y en unión libre presentan una menor probabilidad de reprobar que estudiantes casados. Esto podría deberse a que el estudiante presenta, al estar casado, mayores responsabilidades y, por ende, les dedica menos tiempo a sus estudios, factor que será explicado más adelante como comportamiento y motivación del estudiante.

En segundo lugar, se analizan las variables relacionadas al bagaje de los padres y las redes familiares.

Empezando por el tipo de colegio donde estudió el individuo previo a ingresar a la universidad, se puede observar que, el estudiar en un colegio fiscal, fiscomisional, municipal o particular, respecto a uno extranjero, presenta en promedio mayor probabilidad de reprobar. Este resultado podría venir de varios factores, siendo los más relacionados, aquellos que acorde a [4], encajan dentro del bagaje educativo y económico de los padres o familia a cargo del estudiante. En efecto, se argumenta que en familias con mayores recursos económicos y cuyos miembros han alcanzado mayores niveles de educación, las nuevas generaciones tienen mayor probabilidad de terminar sus carreras, debido a que estos últimos pueden tener como beneficios, guía y asesoría al momento de la postulación y estudios, además de que, con mayores niveles de ingreso, la carga de costos será también más leve. La referencia [14] sostiene que aquellos estudiantes que tengan el apoyo de sus padres en su elección de carrera tendrán una probabilidad mayor de tener éxito en comparación a aquellos que solo toman una carrera por complacer a sus familias.

Por otro lado, la variable de número de miembros de la familia muestra que a medida que esta incrementa, lo hace también la probabilidad de reprobar el curso de nivelación, manteniendo todo lo demás constante. La referencia [4] sugiere que familias con un mayor número de miembros genera que sus estudiantes disminuyan su probabilidad de aprobar, debido a que los recursos del hogar (tanto pecuniarios como no, e.g. ingreso disponible y atención de los padres) se diluyen con más miembros.

A continuación, se analizan las variables de características de los estudiantes, sus habilidades y comportamientos.

La variable de número de materias otorga información para interpretar que aquellos individuos que toman un mayor número de ellas son más propensos a fallar. Este efecto está correlacionado con el tiempo dedicado al estudio, ya que, al ser un curso propedéutico presencial, por lo general no permite que sus estudiantes asistan a clases y trabajen al mismo tiempo. Entonces es plausible asumir que quienes tienen menos materias le dedican más tiempo a estudiarlas.

En efecto, la referencia [16] muestra evidencia empírica de universidades estadounidenses que respaldan esta conclusión. Sin embargo, es preciso recalcar que este tiempo dedicado al estudio podría tener interacciones adicionales con variables relacionadas principalmente a la motivación del estudiante, las cuales deberían ser estudiadas más a fondo. Es así como, al topar la temática de la motivación del estudiante, es prudente analizar el efecto de la variable de número de matrícula ya que se puede observar que aquellos individuos que están cursando su segunda o tercera matrícula tienen menor probabilidad de reprobar que aquellos que se encuentran cursando la primera. Este efecto, en línea con [12], [4] podría estar representando la motivación del estudiante, en especial de aquellos que a pesar de haber reprobado ya una vez el curso de nivelación, lo siguen intentando. Quienes ya lo han repetido dos veces, no presentan diferencias estadísticamente significativas con aquellos que van en su primera matrícula.

VI. CONCLUSIONES Y RECOMENDACIONES

En este trabajo se ha mostrado la utilidad de combinar modelos con objetivos de predicción e inferencia utilizando técnicas de aprendizaje supervisado, para entender a fondo un problema de alta relevancia social.

Por un lado, el uso de Gradient Boosting Machine (GBM) tiene buenos resultados en la predicción de si un estudiante aprobará o no el curso de nivelación, potenciado a través de validación cruzada. Por ello también su preferencia de uso en varias ramas de la ciencia. Este algoritmo predice con una tasa de aciertos del 89 % a aquellos estudiantes que reprobarán el curso de nivelación, logrando un área bajo la ROC en el conjunto de datos de validación de 0.95, la cual indica un buen desempeño de la estimación realizada.

Por otro lado, en el modelo de inferencia se muestran algunos puntos. La selección del mejor modelo logit a través del algoritmo de selección genético ha sido útil para

determinar qué variables afectan a la probabilidad de reprobar. Factores como la calificación ponderada del primer bimestre, la calificación con la que postuló, su jornada de estudios, su ubicación geográfica de origen, entre otras, afectan de una u otra manera a la probabilidad del estudiante, de aprobar el curso de nivelación.

En definitiva, el uso de estas técnicas estadísticas permite el análisis de políticas relacionadas tanto a la situación del estudiante (e.g. política de cuotas) como al manejo del curso de nivelación (e.g. jornadas), que permitan obtener mejores resultados en la aprobación de este. Por ejemplo, es importante dar soporte al problema del pasado académico de los nuevos estudiantes universitarios y generar políticas en la educación secundaria que permitan que los bachilleres lleguen con el menor número de vacíos académicos a su primer año de universidad. Por otro lado, se puede también dar mayor aporte socioeconómico y psicológico a los estudiantes para que puedan aprobar sus cursos sin problema, sea con apoyo respecto a la elección de su carrera, ayuda económica para él y su familia (si aplicase), tutorías y guías para alivianar la carga de estudios, mejor planeación de las jornadas académicas para evitar agotamientos, entre otras medidas a nivel institucional. A nivel más general, el solucionar el problema de infraestructura para acoger un mayor número de estudiantes y la potencialización de universidades alejadas de las ciudades más grandes son también posibles recomendaciones para evitar que estudiantes de provincia se queden sin cupo.

Así mismo, se recomienda continuar con mayor profundidad el estudio de los efectos de las acciones afirmativas y políticas de cuotas. Esto debido a que [17] y los resultados de este trabajo muestran que los estudiantes acogidos a estas medidas son más propensos a reprobar el Curso de Nivelación y abandonar la universidad; donde algunos de sus principales factores de reprobación asociados son el ingreso, la nota de postulación, la provincia de procedencia, falta de acompañamiento académico, entre otros.

Con ello en mente, y regresando al aspecto pragmático de este trabajo, se recomienda dar seguimiento a los resultados del modelo GBM para analizar posibles cambios debido a nuevos efectos que puedan surgir con el paso del tiempo, sea debido a fenómenos externos o propios al curso de nivelación. Se sugiere, además, sociabilizar el modelo para permitir que otras facultades, e incluso, otras instituciones puedan hacer de esta herramienta y así aportar a la mejora de las condiciones de aprobación en instituciones de educación superior. Además, este modelo podría ser mejorado al ser puesto en competencia con otros modelos con buena capacidad de predicción (tales como redes neuronales artificiales) o con el uso de técnicas de balanceo para evitar que el porcentaje de reprobados en la muestra supere a los aprobados.

En línea con todo este análisis, se puede llegar a un nivel más alto de la discusión y poner sobre la mesa la coyuntura actual: la inminente intersección entre la ciencia de datos y la política pública. Varios artículos publicados por académicos y profesionales de la industria sugieren que esta es necesaria debido a que el análisis provisto por los algoritmos de la ciencia de datos puede ayudarnos a entender y quizás resolver problemas complejos, siempre y cuando se haya entendido su trasfondo histórico, legal y socioeconómico. De esta manera, es oportuno mantener la idea de [19], quien menciona que el aparecimiento de la revolución tecnológica, el big data y el aprendizaje automático implican tanto el aprovechamiento de los datos para una mejor toma de decisiones, así como un cambio de paradigma dentro de las ciencias sociales, haciendo posible la aceleración de los procesos de investigación y desarrollo, enfocándolos en una aplicación práctica (entre otras cosas) a la política pública.

Referencias

[1] G. Guiselle, “Factores asociados al rendimiento académico en estudiantes universitarios desde el nivel socioeconómico: Un estudio en la Universidad de Costa Rica”, El Salvador: Revista Electrónica Educare, vol. 17, 2013.

[2] F. Carlos. “Sistemas de evaluación académica”, El Salvador: Editorial Universitaria, 2014.

[3] V. Jorge y col., “Una explicación del rendimiento estudiantil universitario mediante modelos de regresión logística”. Venezuela: Visión Gerencial, 2009.

[4] A. Carmen y col., “DISCUSSION PAPER SERIES The Economics of University Dropouts and Delayed Graduation : A Survey The Economics of University Dropouts and Delayed Graduation : A Survey". En: 11421, 2018.

[5] Rodríguez Ayán, M. N., & Coello García, M. T. (2008), Prediction of university students’ academic achievement by linear and logistic models. Spanish Journal of Psychology, 11(1), 275–288. https://doi.org/10.1017/s1138741600004315

[6] Friedman, Jerome H. “Greedy Function Approximation: A Gradient Boosting Machine.” The Annals of Statistics 29, 33,. 5, pp. 1189–1232, 2001 JSTOR, www.jstor.org/stable/2699986.

[7] T. Hastie, T. Robert y F. Jerome, “The elements of statistical learning: data mining, inference, and prediction”, New York: Springer, 2017.

[8] Jordi Gironés Roig y col. Minería de datos: modelos y algoritmos. pp. 274, 2017 isbn: 9788491169048.

[9] C. Vincent, glmulti: “Model Selection and Multimodel Inference Made Easy”. R package version 1.0.7.1. [Online]. Available: https://CRAN.R-project.org/package=glmulti. [Accessed: 2019].

[10] J. Hunt, “Classification by induction: Aplications to modelling and control of non linear dynamic systems. Intelligent Systems Engineering”, 1993.

[11] I. Kononenko, I. Bratko and M. Kukar. Machine, “Learning and Data MIning: Methods and Aplications”. John Wiley & Sons Ltd, 1998.

[12] S. Larose y col. “Nonintellectual learning factors as determinants for success in college". En: Research in Higher Education 39.3, pp. 275-297, 1998.

[13] T. Ernest, P. Patrick, T. Terenzini y Lee M. “Wole. Orientation to College and Freshman Year Persistence/Withdrawal Decisions". En: The Journal of Higher Education 57.2, pp. 155, 1986.

[14] N. Alexander y W. Ruth. “Determinants of College Success". En: The Journal of Higher Education 11.9, pp. 479-485, 1940.

[15] Carmen Aina. Success and failure of Italian university students. Evidence from administrative data". pp 1-51, (2010).

[16] P. Babcock y M. Mindy. “The falling time cost of college: Evidence from half a century of time use data". En: Review of Economics and Statistics, 2011

[17] S. Iván y col.”Factores Asociados Al Abandono En Estudiantes De Grupos Vulnerables. Caso Escuela Politécnica Nacional". En: Congresos CLABES, pp. 132-141. [Online]. Available: https : / / revistas. [Accessed: 2018].

[18] S. Walter, Escudero. “Big data y aprendizaje autom_atico: Ideas y desafíos para economistas". En: Una nueva econometría. isbn: 978-987-655-201-1, 2018.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

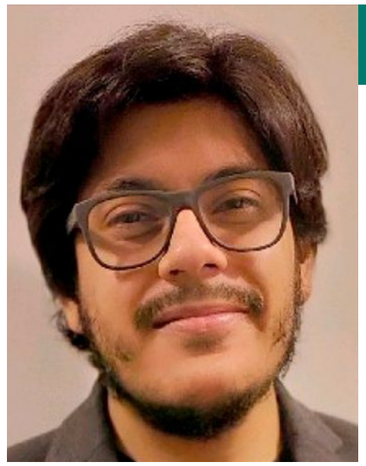
AUTHORS

Karen Calva

Ingeniera Matemática especializada en estadística y ciencia de datos, graduada en la Escuela Politécnica Nacional. Especialista de analítica avanzada en una de las instituciones financieras más grandes de Ecuador. Consultora independiente en temas de machine learning, geo-estadística, gestión de procesos y logística para el levantamiento de información, desarrollo de aplicativos webs con motores analíticos. Docente para estudiantes de pregrado o profesionales, con experiencia dictando cursos y conferencias relacionadas a matemáticas, técnicas estadísticas y manejo de software libre como R, Spark y Python, en instituciones como la Sociedad Ecuatoriana de Estadística.

Miguel Flores

Ph.D. en Estadística e Investigación de Operaciones, Máster en Técnicas Estadísticas (Universidad de La Coruña). Tiene experiencia en Educación y Formación profesional superior, universitaria y empresarial en el campo de la Statistics & Machine Learning. Profesor Titular de la cátedra Probabilidad y Estadística, en la Escuela Politécnica Nacional. Miembro del Grupo de Investigación Multidisciplinar en Sistemas de Información, Gestión de la Tecnología e Innovación (SIGTI) de la Escuela Politécnica Nacional y del Grupo de Modelización, Optimización e Inferencia Estadística (MODES) de la Universidad de La Coruña.



Hugo Porras

Ingeniero en Ciencias Económicas y Financieras graduado con mención Summa Cum Laude en la Escuela Politécnica Nacional, y estudiante de la maestría en inteligencia artificial en la Universidad Internacional de la Rioja. Soy especialista científico de datos en Banco del Pacífico e investigador independiente en temas relacionados a finanzas de real state, riesgo de crédito, geo-analítica, procesamiento del lenguaje natural, economía del bienestar, economía del desarrollo y organización industrial. Además, he sido profesor en cursos de programación en R y procesamiento del lenguaje natural con la Sociedad Ecuatoriana de Estadística.

Ana Cabezas-Martínez

Ingeniera Comercial con mención en Administración de Empresas por la Universidad de las Américas (UDLA). Actualmente, me encuentro finalizando la maestría de Política Comparada en la Facultad Latinoamericana de Ciencias Sociales FLACSO Ecuador.

Soy especialista en educación superior con experiencia en el diseño, implementación, monitoreo, análisis y evaluación de política pública en proyectos enfocados en el sector sociales; la generación e implementación de metodologías enfocadas en la articulación y cooperación de los sectores sociales y productivos (públicos, privados y organizaciones no gubernamentales).

K. Calva, M. Flores, H. Porras, A. Cabezas-Martínez, "Modelo de predicción del rendimiento académico para el curso de nivelación de la Escuela Politécnica Nacional a partir de un modelo de aprendizaje supervisado", Latin-American Journal of Computing (LAJC), vol. 8, no. 2, 2021.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021