PROYECTO DE INVESTIGACIÓN INTERNOS SIN FINANCIAMIENTO O AUTOGESTIONADOS

ANEXO 1 - DATOS INFORMATIVOS

Fecha de presentación (24/12/2019):

Título del proyecto: **MODELO DE PREDICCIÓN DEL RENDIMIENTO ACADÉMICO PARA EL CURSO DE NIVELACIÓN DE LA ESCUELA POLITÉCNICA NACIONAL A PARTIR DE UN MODELO DE APRENDIZAJE SUPERVISADO AUTOMATIZADO EN R.**

TIPOS DE INVESTIGACIÓN

Investigación básica ❑ Investigación aplicada □

DEPARTAMENTO(S) Y/O INSTITUTO(S):
1. Departamento de Matemática

LÍNEA(S) DE INVESTIGACIÓN (verificable en el SAEW):
1. Modelos Estadísticos

RESUMEN DE INFORMACIÓN DEL DIRECTOR Y COLABORADORES

Director

Apellidos y nombres	No. de Cédula	HSS	Departamento	Título de mayor nivel y mención.
Flores Sánchez Miguel Alfonso	0918863218	4	Matemática	Doctores en Estadística e Investigación Operativa

Colaborador(es)

Apellidos y nombres	No. de Cédula	HSS	Departamento	Título de mayor nivel y mención.

Colaboradores Externos

Apellidos y nombres	No. de identificación	HSS	Institución	Título de mayor nivel y mención.
Calva Yaguana Karen Priscilla	1724448384	2	Escuela Politécnica Nacional	Egresada Ingeniería Matemática

* HSS = Horas Semana Semestre

PROYECTO DE INVESTIGACIÓN INTERNOS SIN FINANCIAMIENTO O AUTOGESTIONADOS

ANEXO 2 – DETALLES DE LA PROPUESTA

Investigación Básica ❑ Investigación Aplicada □

DEPARTAMENTO(S) Y/O INSTITUTO(S):
1. Departamento de Matemática

LINEA(S) DE INVESTIGACIÓN:
1. Modelos Estadísticos

DISCIPLINA CIENTÍFICA ***(Marque X, solamente una opción)***	
Ciencias Naturales y Exactas;	X
Ingeniería y Tecnologías;	
Ciencias Médicas;	
Ciencias Agrícolas;	
Ciencias Sociales;	
Humanidades	

OBJETIVO SOCIOECONÓMICO ***(Marque X, solamente una opción)***	
Exploración y explotación del medio terrestre;	
Ambiente;	
Exploración y Explotación del espacio;	
Transporte, telecomunicaciones y otras infraestructuras;	
Energía;	
Producción y tecnología industrial;	
Salud;	
Agricultura;	
Educación;	X
Cultura, ocio, religión y medios de comunicación;	
Sistemas políticos y sociales, estructuras y procesos;	
Defensa;	
Avance general del conocimiento: I+D financiada con los Fondos Generales de Universidades (FGU);	
Avance general del conocimiento: I+D financiados con otras fuentes.	

1 Proyecto de Investigación

Título:
MODELO DE PREDICCIÓN DEL RENDIMIENTO ACADÉMICO PARA EL CURSO DE NIVELACIÓN DE LA ESCUELA POLITÉCNICA NACIONAL A PARTIR DE UN MODELO DE APRENDIZAJE SUPERVISADO AUTOMATIZADO EN R.

Resumen del proyecto

El bajo rendimiento académico de los estudiantes en los primeros semestres universitarios es un problema que deben enfrentar las universidades, en particular el Curso de Nivelación (CN) de la Escuela Politécnica Nacional (EPN). Los protagonistas de esta problemática son los estudiantes; mientras no adquieran los conocimientos y habilidades básicas para iniciar sus estudios universitarios, el problema de la reprobación se seguirá manifestando igual o con tendencia a aumentar en semestres posteriores.

El modelo resultante se usará para predecir el estado de aprobación o reprobación de los estudiantes del CN y para inferir características descriptivas del rendimiento académico y de los factores que lo explican.
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; también ayudará a predecir con antelación el número de estudiantes que aprobarán en CN y los que no, para que con esta información se pueda planificar de mejor manera los cupos del próximo periodo en el CN y en cada carrera.

Palabras clave (4-6):
Modelo Clasificación, Rendimiento académico, Predicción de reprobación, Software R.

2 Objetivos, relevancia, productos y resultados esperados de esta propuesta de investigación

2.1 Objetivos

2.1.1 Objetivo General

- El objetivo de este proyecto es explicar y predecir el rendimiento académico, en función de variables de rendimiento y factores sociodemográficos de los estudiantes del curso de nivelación en la Escuela Politécnica Nacional mediante modelos de aprendizaje supervisado; cuyos resultados estén disponibles a través de una aplicación web.

2.1.2 Objetivos Específicos

1. Describir el perfil de los estudiantes y su rendimiento académico para el período 2019-A.
2. Desarrollar un modelo de aprendizaje supervisado
3. Aplicar el modelo predictivo, inferir resultados y validarlos.
4. Edificar una aplicación web en Shiny-R que presente las predicciones.

2.2 Detalle de los resultados esperados

a. Los modelos resultantes se usarán para predecir el estado de aprobación o reprobación de los estudiantes del CN y para presentar características descriptivas del rendimiento académico y de los factores que lo explican; siendo ésta una aplicación importante desde el punto de vista de la extracción de conocimiento.

b. Se espera que 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; también ayudará a predecir con antelación el número de estudiantes que aprobarán en CN y los que harán segunda matrícula, para que con esta información se pueda planificar de mejor manera los cupos del próximo periodo en el CN y en cada carrera.

c. La construcción de una aplicación interactiva Shiny, facilitará la visualización de las predicciones y demás resultados obtenidos de los modelos, la aplicación estará desarrollada en un servidor de la Dirección de Gestión de la Información y Procesos (DGIP).

3 Relevancia de la propuesta de investigación y su relación con la(s) líneas de investigación

En el rendimiento académico interactúan elementos multicausales, tanto sociodemográficos, psicosociales, pedagógicos, institucionales y socioeconómicos; la combinación y ponderación de estos factores no siempre son los mismos, razón por la cual este es un tema que amerita constante investigación [1]. El presente trabajo busca mejorar el entendimiento de los factores que influyen en el rendimiento académico (score/nota académica) de los estudiantes del CN de la EPN y predecir el número de estudiantes que aprobarán en CN y los que harán segunda matrícula.

La variable de estudio a predecir y explicar de la presente investigación es la reprobación de los estudiantes de la cual consideraremos los siguientes estados: Si (Reprueba) o No (Aprueba). Con la finalidad de predecir a qué estado de reprobación o aprobación pertenece cada estudiante se aplicará un modelo mediante el aprendizaje inductivo también llamado aprendizaje supervisado, el cual parte de casos particulares (experiencias) y obtiene casos generales (modelos o reglas). El aprendizaje inductivo tiene la ventaja de poder automatizarse [2].

La técnica estadística escogida para modelizar y predecir el estado de la variable reprobación, es la regresión logística, ya que el resultado de la regresión logística es la estimación de la probabilidad de que un estudiante pertenezca a un estado de la variable reprobación (Sí ó No); el ajuste del modelo logístico permite crear un perfil de los estudiantes en base a las variables predictivas y no es necesario que se cumpla el supuesto de normalidad ya que la variable dependiente es dicotómica [3]. El conocimiento de los coeficientes y su ponderación es muy importante para conocer los factores que influyen en la reprobación.

4 Productos esperados (marcar con una "X" al menos uno de los productos no señalados)

Tipo de Producto:	Marcar con una "X"
a. Disertación a la Comunidad Politécnica (obligatorio);	X
b. Presentación de un artículo en formato de la Revista Politécnica (obligatorio)	X
c. Proyecto de Titulación;	X
d. Aplicación tecnológica construida o implementada;	X
e. Patente presentada;	
f. Perfil de proyecto de mayor impacto científico, técnico, pedagógico o de innovación.	
g. Publicaciones científicas indexada en SCIMAGO-SCOPUS/WoS/SCIELO/Latindex Catálogo o un artículo en congreso indexado en SCOPUS.	X

5 Descripción y metodología y diseño del proyecto

5.1 Descripción, metodología y diseño del proyecto

La variable de estudio a predecir y explicar de la presente investigación es la reprobación de los estudiantes que está en función del rendimiento del estudiante, de la cual consideraremos los siguientes estados: Si (Reprueba) o No (Aprueba); éste se conoce como un problema de clasificación ya que el objetivo es predecir a qué estado de reprobación o aprobación pertenece cada estudiante. Para determinar las variables

independientes del modelo de regresión logístico se cuenta con los registros académicos de los estudiantes desde el año 2009 que ha sido proporcionado por la Dirección de Gestión de la Información y Procesos (DGIP) administra los recursos informáticos y tecnológicos de la EPN, almacena y gestiona entre otros. Las variables que se considerarán como posibles factores que intervengan en el rendimiento académico, se muestra en el Cuadro 1 una breve descripción de éstas variables:

Cuadro 1: Descripción de las variables

Variable	Tipo	Descripción
numeroMaterias	numérica	Número de materias en la que está matriculado el estudiante (aplica a los de segunda matrícula)
calificacion1	numérica	Calificación del primer bimestre
notaPostulacion	numérica	Calificación del examen Ser Bachiller
edad	numérica	Edad en años
miembrosFamilia	numérica	Número de miembros del núcleo familiar
ingreso	numérica	Información socioeconómica, ingreso mensual familiar neto
sexo	categórica	F: Femenino, M: Masculino
estadoCivil	categórica	C: Casado/a, D: Divorciado/a, S: Soltero/a, U: Unión libre
etnia	categórica	Información personal
jornada	categórica	matutina: 7am - 1pm, vespertina: 2pm - 8pm
tipoColegio	categórica	Tipo de colegio: Fiscal, Particular o privado, Fiscomisional, Municipal.
segmentoPoblacional	categórica	Segmento poblacional: Población general, Política de cuotas, etc.
ciudadResidencia	categórica	Ciudad de residencia
numeroAsignacion	categórica	Número de asignación de cupo por parte del SNNA
carreraAspiraEstudiante	categórica	Carrera a la que aspira el estudiante
carreraSSNA	categórica	Si el estudiante aspira a una Carrera bajo el Reglamento de Régimen Académico (RRA)

La variable Reprueba es cualitativa, y se la codifica de la siguiente manera: Reprueba = 1 (si), si el estudiante reprueba la asignatura o Reprueba = 0, caso contrario. Ya que la variable Reprueba es binaria (toma únicamente dos valores 0 o 1), entonces se requiere ajustar un modelo para el cual los valores estimados para la respuesta sea 0 o 1. Se utilizará la función logística, por las siguientes razones:

- Es adecuada en la mayoría de los casos en los cuales la respuesta es binaria; es decir, toma valores 0 o 1.
- Las variables independientes para la regresión logística pueden ser continuas o categóricas.

Para el modelo de regresión logística, sea $X = X_1, X_2, \ldots, X_n$ el conjunto de variables independientes o explicativas y sean $\beta_1, \beta_2, \ldots \beta_n$ los parámetros, entonces la formulación es la siguiente:

$$P_i = E(Y_i = 1|X) = \frac{1}{1 + e^{(\beta_1 + \beta_2 X_i + \ldots + \beta_n X_n)}} \quad (1)$$

Por facilidad, podemos expresar la ecuación anterior como

$$P_i = \frac{1}{1+e^{-Z_i}} = \frac{e^Z}{1+e^z} \quad (2)$$

Donde $Z_i = \beta_1+\beta_2X_i+...+\beta_nX_n$. La ecuación (2) descrita anteriormente se conoce como la función de distribución logística (acumulativa). Z_i se encuentra dentro de un rango de $-\infty$ a ∞, por tanto P_i se encuentra dentro de un rango de 0 a 1; P_i no está linealmente relacionado con Z_i; es decir, P_i no está linealmente relacionado con X_i, lo que significa que no se puede estimar los parámetros con el procedimiento habitual de MCO sino por estimadores de máxima verosimilitud. Tenemos que si P_i es la probabilidad de que un evento ocurra, entonces $(1 - P_i)$, la probabilidad de que el evento no ocurra es:

$$1 - P_i = \frac{1}{1+e^{Z_i}} \quad (3)$$

Entonces:

$$\frac{P_i}{1-P_i} = \frac{1-e^{Z_i}}{1+e^{-Z_i}} = e^{Z_i}$$

Así, $P_i/(1-P_i)$ es la razón de las probabilidades, la razón de la probabilidad de que un evento ocurra respecto de la probabilidad de que no ocurra. Tomando el logaritmo natural de la última ecuación se obtiene el siguiente resultado:

$$L_i = \ln(\frac{P_i}{1-P_i}) = Z_i = \beta_1 + \beta_2X_i + ... + \beta_nX_n; \quad (4)$$

es decir, L, es el logaritmo de la razón de las probabilidades [4]. Algunas características importantes que se puede mencionar del modelo de regresión logística son las siguientes:

- A medida que P toma los valores de 0 a 1, es decir, a medida que Z varía de $-\infty$ a ∞, tenemos que L toma valores de $-\infty$ a ∞.
- Aunque L es lineal en X, las probabilidades en sí mismas no lo son.
- En el modelo de regresión logística los coeficientes expresan el cambio en el logaritmo de las probabilidades, cuando una de las variables explicativas cambia en una unidad, permaneciendo constantes las demás.
- Si L, es positivo, significa que cuando se incrementa el valor de la(s) explicativa(s), aumenta la posibilidad de que la variable dependiente sea igual a 1. Si L es negativo, la posibilidad de que la variable dependiente sea igual a 1 disminuye conforme se incrementa el valor de X.
- βi, la pendiente del modelo, mide el cambio en L ocasionado por un cambio unitario en Xi.
- El modelo de regresión logística supone que el logaritmo de la razón de probabilidades está relacionado linealmente con Xi [5].

La calidad del modelo, medida como el equilibrio entre un ajuste razonable de los datos y un número mínimo de parámetros, se evaluará usando índices como el Criterio de Información de Akaike (AIC) o el Criterio de Información Bayesiano (BIC o SBC). Al comparar varios modelos paramétricos entre sí, el modelo con el índice más bajo es el que presenta la mejor calidad en el conjunto de modelos evaluados [6].

5.2 Bibliografía

[1] Montero, E., Villalobos, J., y Valverde, A. (2007). Factores institucionales, pedagógicos, psicosociales y sociodemográficos asociados al rendimiento académico en la universidad de costa rica: Un análisis multinivel. *Revista Electrónica de Investigación y Evaluación Educativa,* volumen 13, n. 2, p. 215-234.

[2] K.J. Hunt. (1993). Classification by induction: Aplications to modelling and control of non linear dynamic systems. *Intelligent Systems Engineering*, volumen 2, n. 4, p. 231 - 245.

[3] Tanış, Caner. (2014). Analysis of Influence on Academic Achievement of Students' Attitudes and

Some Habits with Logistic Regression. *Journal of Selçuk University Natural and Applied Science,* volumen 3, n. 2, p.61-72.
[4] Hastie, T., Tibshirani, R., Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (Segunda edición). Stanford, California: Editorial Springer.
[5] Kononenko, I., Bratko, I., y Kukar, M. (2017). *An Introduction to Machine Learning: Methods and aplications.* (Segunda edición). Stanford, California: Editorial Springer.
[6] Calcagno, V., y de Mazancourt, C. (2010). glmulti: An r package for easy automated model selection with (generalized) linear models. *Journal of statistical software.* volumen 34, n. 12, p. 1 - 29.

6 Infraestructura, equipos y fondos adicionales.

6.1 Infraestructura y equipos

- *Indicar la infraestructura y equipos **disponibles** para la ejecución del proyecto, con la ubicación actual de los mismos*

Infraestructura	Equipos	
	Nombre del Equipo	**Ubicación del Equipo**
Computadora de escritorio		Departamento Matemática
Servidor de la DGIP		DGIP

6.2 Breve justificación del equipo requerido

- El servidor contiene los datos que servirán de insumo para los modelos y se llamará a estas estructuras de datos a través del computador de escritorio.

6.3 Fondos Adicionales

ESCUELA POLITÉCNICA NACIONAL

Proyecto de Investigación Interno Sin Financiamiento o Autogestionado

ANEXO 3 - CRONOGRAMA DE ACTIVIDADES DEL PROYECTO

Título del Proyecto: **MODELO DE PREDICCIÓN DEL RENDIMIENTO ACADÉMICO PARA EL CURSO DE NIVELACIÓN DE LA ESCUELA POLITÉCNICA NACIONAL A PARTIR DE UN MODELO DE APRENDIZAJE SUPERVISADO AUTOMATIZADO EN R**

AÑO 1

Nº	Actividad	Mes 1				Mes 2				Mes 3				Mes 4				Mes 5				Mes 6				Mes 7				Mes 8				Mes 9				Mes 10				Mes 11				Mes 12			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	**Descripción del perfil de los estudiantes**																																																
1.1	Actividad 1: Describir el perfil actual de los estudiantes y su rendimiento académico																																																
1.2	Actividad 2: Desarrollar los procesos y técnicas con las que se va a implementar la predicción del rendimiento académico																																																
2	**Desarrollar los modelos de aprendizaje supervisado**																																																
2.1	Actividad 1: Automatizar los algoritmos en R																																																
2.2	Actividad 2: Aplicar el modelo predictivo, inferir resultados y validarlos.																																																
2.3	Edificar una aplicación web en Shiny que presente las predicciones																																																
3	**Desarrollo de productos**																																																
3.1	Elaboración de un artículo y envío SCOPUS																																																
3.2	Conferencia Nacional																																																
3.3	Cierre del proyecto																																																

PROYECTO DE INVESTIGACIÓN INTERNOS SIN FINANCIAMIENTO O AUTOGESTIONADOS

ANEXO 4 - DECLARACIÓN

TIPO DE INVESTIGACIÓN

Investigación básica ❑	Investigación aplicada ☒

TÍTULO DEL PROYECTO

MODELO DE PREDICCIÓN DEL RENDIMIENTO ACADÉMICO PARA EL CURSO DE NIVELACIÓN DE LA ESCUELA POLITÉCNICA NACIONAL A PARTIR DE UN MODELO DE APRENDIZAJE SUPERVISADO AUTOMATIZADO EN R.

DECLARACIÓN DEL DIRECTOR DEL PROYECTO

El equipo de investigadores, representado por el Director del Proyecto declara lo siguiente:

- Que el presente proyecto es una creación original de mi autoría y del equipo de investigadores, y por tanto asumimos la completa responsabilidad legal en caso de que un tercero alegue la titularidad de los derechos intelectuales del proyecto, exonerando a la EPN de cualquier acción legal que se derive por esta causa.
- Que el presente proyecto no ha sido presentado en ninguna convocatoria de otra institución pública o privada. El incumplimiento será causal para que el proyecto no sea tomado en consideración.
- Que todos los bienes adquiridos en proyecto permanecerán bajo la custodia y responsabilidad del director de proyecto durante la ejecución del mismo.
- Que si el proyecto genera algún producto o procedimiento susceptible de obtener derechos de propiedad intelectual, de los cuales se deriven beneficios, aceptamos que éstos serán compartidos entre los investigadores y la institución o las instituciones participantes en el proyecto, conforme a lo establecido en el COESC.
- Que el equipo de investigadores y/o instituciones participantes se comprometen a mantener la confidencialidad de la información si ésta podría ser susceptible de protección por patentes, y solicitar la valoración de propiedad intelectual respectiva previa a cualquier publicación o difusión.
- Que para el caso de derechos de autor otorgamos una licencia de uso exclusivo con fines académicos para la o las instituciones participantes en el proyecto.

Firma del Director del Proyecto
Nombre: Miguel Flores
C.I.:0918863218

DECLARACIÓN DEL JEFE DE DEPARTAMENTO