Predicción de no egreso en Jóvenes a Programar (JaP)

Avances tesis maestría en Data Science

Mayo 2023

Autores:
Ing. Gonzalo J. Harreguy
MSc. Yanedy Pérez
Tutores:
Dr. Ignacio Álvarez
Dra. Natalia Andrea da Silva

Ceibal

Objetivo general

Estimación del no egreso de los estudiantes de Jóvenes a Programar (JaP), a partir de los factores de mayor impacto.

Ceibal

Jóvenes a Programar (JaP)

Jóvenes a Programar (JaP) surge en el 2016 con el fin de capacitar a jóvenes entre 18 y 30 años (de todo el país) en el área de TI.

JaP ofrece cursos de programación, testing y otras tecnologías (a través de la plataforma Mumuki y por videoconferencias), así como talleres de habilidades blandas y formación en idioma inglés.

Una vez los jóvenes alcanzan la certificación inicial de Programador o Tester, se les brinda la oportunidad de formarse en cursos más específicos.

Cuenta con un Servicio de Intermediación Laboral (SIL) que facilita el contacto de los egresados/as del programa con las empresas del sector de TI.

Ceibal

Estado del arte

Estimación de abandono y rendimiento académico

	Técnicas
Fernández, et al. (2021)	**XGBoost**, **Random Forest** y SVM
Amare & Simonova (2021)	**Random Forest**, Naive Bayesian, **Logistic Regression** y Decision Tree
Park & Yoo (2021)	Decision Tree, **Random Forest**, SVM y **Deep Neural Network**
Ghosh & Janan (2021)	**Random Forest**
Li & Liu (2021)	**Deep Neural Network**
Niyogisubizo et al. (2022)	**Random Forest**, **XGBoost** y **Feed-forward Neural Networks**
Moreira et al. (2022)	**Random Forest**, **XGBoost** y **Artificial Neural Network**
Al-Zawqari et al. (2022)	**Random Forest** y **Artificial Neural Network**
Song et al. (2023)	Decision Tree, **Random Forest**, **XGBoost**, **Logistic Regression** y SVM
Mduma (2023)	**Random Forest**, **Logistic Regression** y Perceptrón Multicapa

Ceibal

Dataset

- Información de los estudiantes: edad, género, estudios alcanzados, información laboral, resultado de la prueba de ingreso, etc.
- Información de la interacción de los estudiantes con la plataforma Mumuki Fase 1: número de ejercicios resueltos, fallidos y con advertencia, cantidad de envíos (con periodicidad semanal).
- Total de estudiantes que entraron a la Fase 1.
- Los años considerados fueron 2020, 2021 y 2022.

Ceibal

Dataset
Transformaciones

- Missing (ej. completar con la media en las variables numéricas, eliminar observaciones con muchos NaN).
- Se crearon nuevas variables a partir de los datos de Mumuki.
- Se reagruparon categorías en algunas variables (ej: la variable salud).
- Transformación de variables utilizando label encoding (ej. la variable educación alcanzada).
- Transformación de variables utilizando one hot encoding (ej. la variable salud).
- Se omitieron las variables con una alta correlación.

Ceibal

Dataset
Final

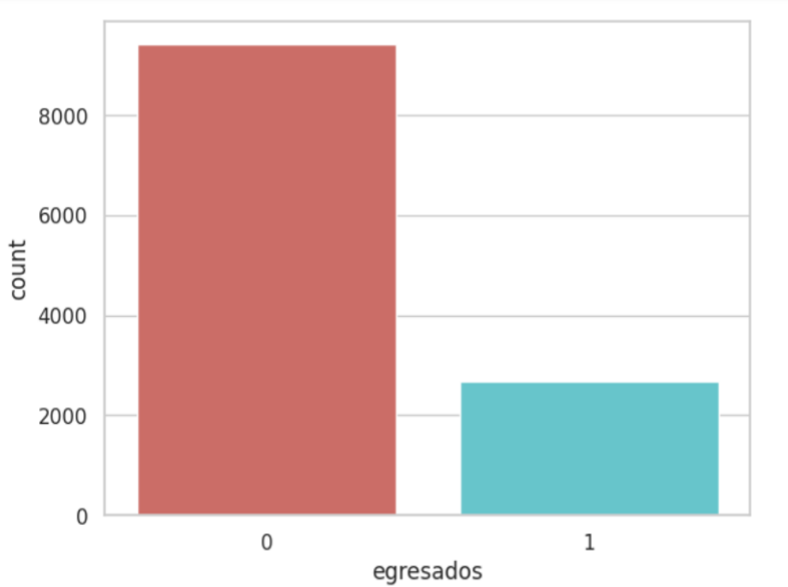
- Observaciones: 12091 estudiantes
- Años: 2020, 2021 y 2022
- Variable explicada: egreso (0 o 1)
- Variables explicativas: 76 variables (género, edad, inscripciones previas a JaP, educación alcanzada, situación laboral, ocupación, trabajo en empresa de TI, trabajo en tares de TI, trabajo publico o privado, puntaje del test de ingreso a JaP, tipo de atención de salud, hijos, región, cantidad de ejercicios correctos, con advertencia y fallidos (semanales) y cantidad de envíos semanales.

Ceibal

Dataset

Output

Variable a predecir: **egresados**. Toma valores 0 o 1 dependiendo de si el estudiante no tuvo o tuvo éxito en su egreso.



No egresados (9417): 78%

Egresados (2674): 22%

Técnicas de aprendizaje estadístico aplicadas

- ❑ Logistic Regression
- ❑ Random Forest
- ❑ Extreme Gradient Boosting (XGBoost)
- ❑ Deep Neural Network

En Logistic Regression, Random Forest y XGBoost se aplicó **GridSearchCV** y **RandomizedSearchCV** para mejorar los hiperparámetros , ambos métodos incluyen **validación cruzada**.

Ceibal

Logit Regression

Salida

Variable	Coef.	Sign.
inscripciones_previas	-0.0479	***
trabajo_horas	-0.0207	*
trab_emp_TI	-0.0247	*
puntaje_sobre64_20_21_22	0.2388	***
salud_Mutualista	-0.0246	**
salud_No sabe / No contesta	-0.0576	**
salud_Salud Pública / ASSE	-0.0351	***
region_METROPOLITANA	-0.0579	***
region_NOROESTE	-0.0412	**
genero_F	-0.0260	***

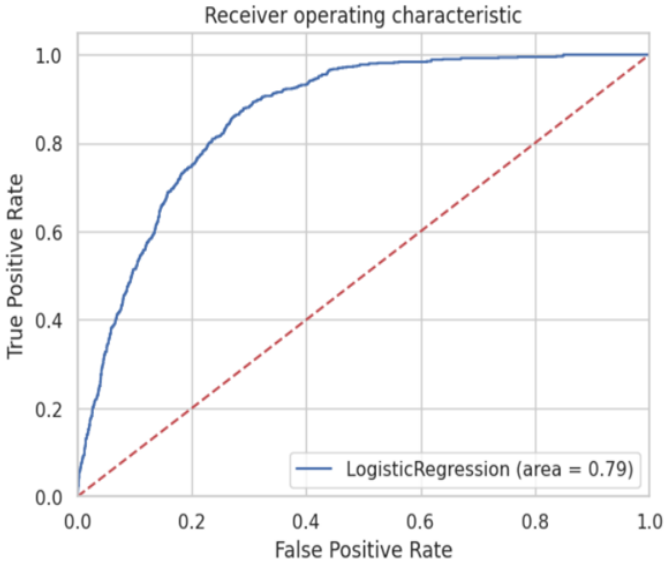
Nota: *, ** y *, denotan significancia al 10%, 5% y 1%, respectivamente.**

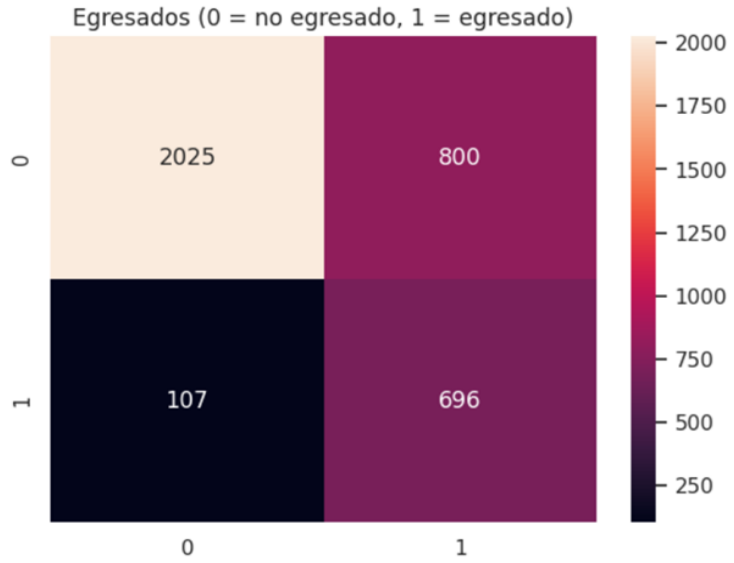
Variable	Coef.	Sign.
semana_3_ejercicios_fallidos_totales	-0.0158	**
semana_4_ejercicios_fallidos_totales	0.0223	**
semana_8_ejercicios_fallidos_totales	-0.0170	**
semana_2_ejercicios_resueltos_totales	0.0008	***
semana_8_ejercicios_resueltos_totales	-0.0010	**
semana_9_ejercicios_resueltos_totales	0.0008	**
semana_11_ejercicios_resueltos_totales	0.0013	***
semana_12_ejercicios_resueltos_totales	-0.0018	***
semana_13_ejercicios_resueltos_totales	0.0019	***
semana_9_envios	0.0014	**
semana_10_envios	0.0047	***
semana_11_envios	0.0051	***
semana_12_envios	0.0133	***
semana_13_envios	-0.0096	***

Logit Regression

Métricas de evaluación

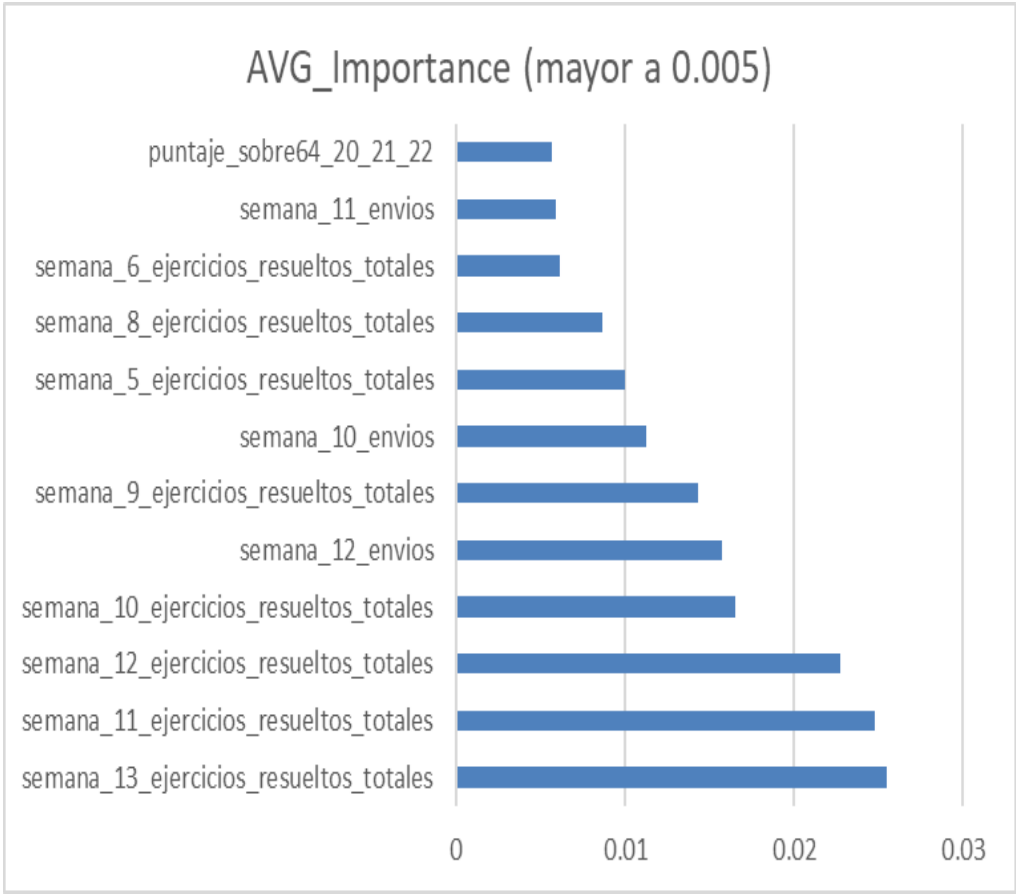
	precision	recall	f1-score	support
0	0.95	0.72	0.82	2825
1	0.47	0.87	0.61	803
accuracy			0.75	3628
macro avg	0.71	0.79	0.71	3628
weighted avg	0.84	0.75	0.77	3628





Random Forest

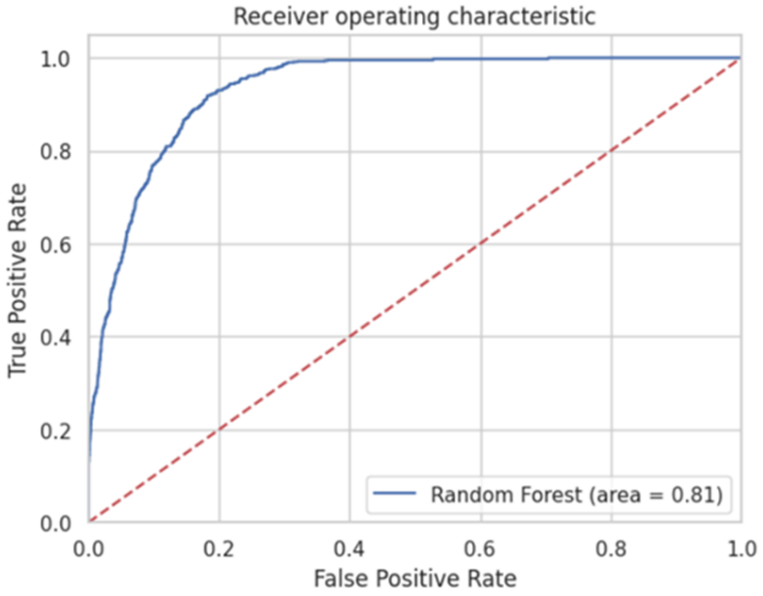
Importancia de las variables

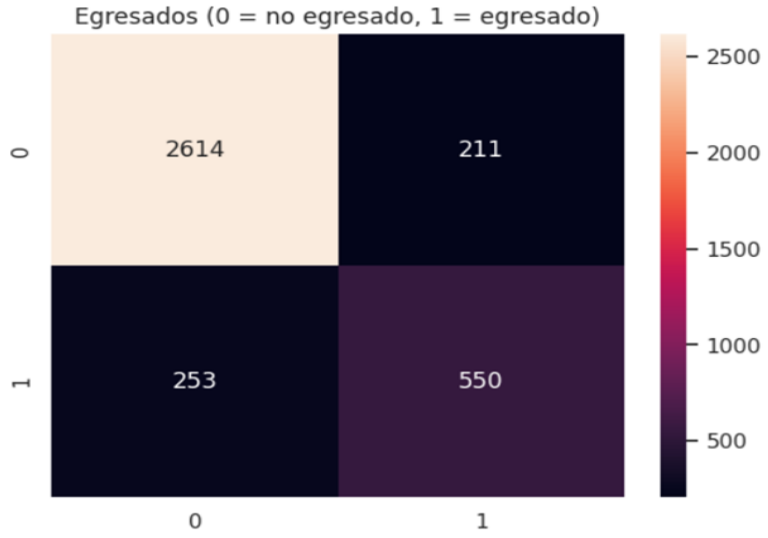


Random Forest

Métricas de evaluación

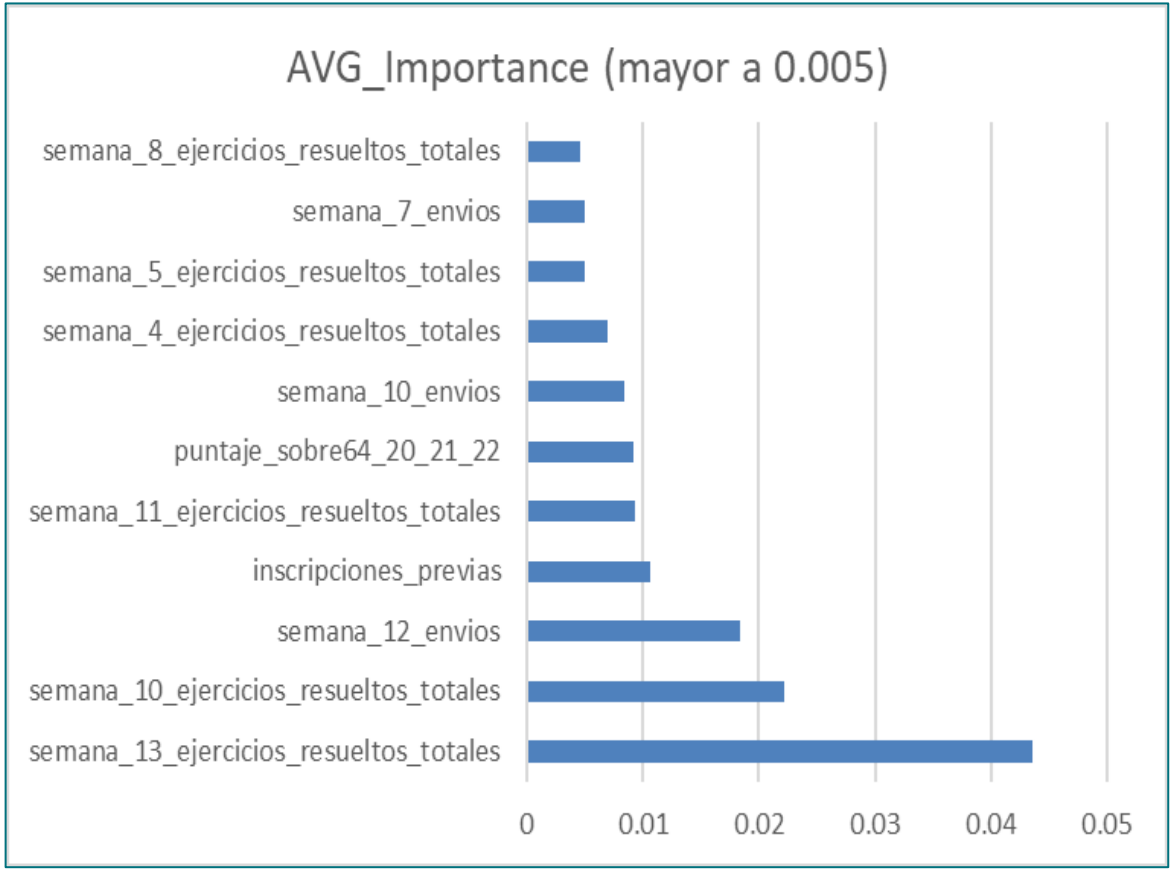
	precision	recall	f1-score	support
0	0.91	0.93	0.92	2825
1	0.73	0.69	0.71	803
accuracy			0.87	3628
macro avg	0.82	0.81	0.81	3628
weighted avg	0.87	0.87	0.87	3628





XGBoost

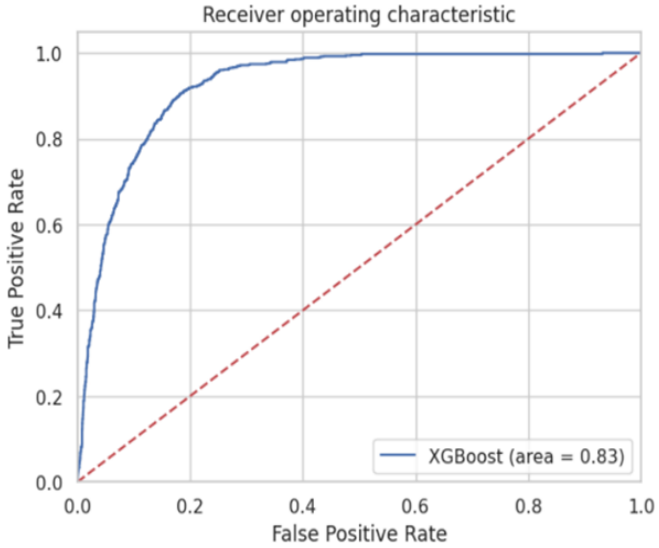
Importancia de las variables

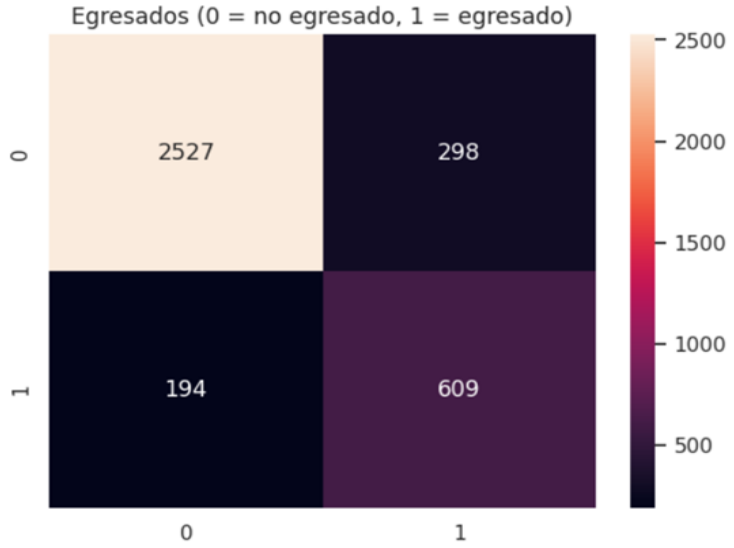


XGBoost

Métricas de evaluación

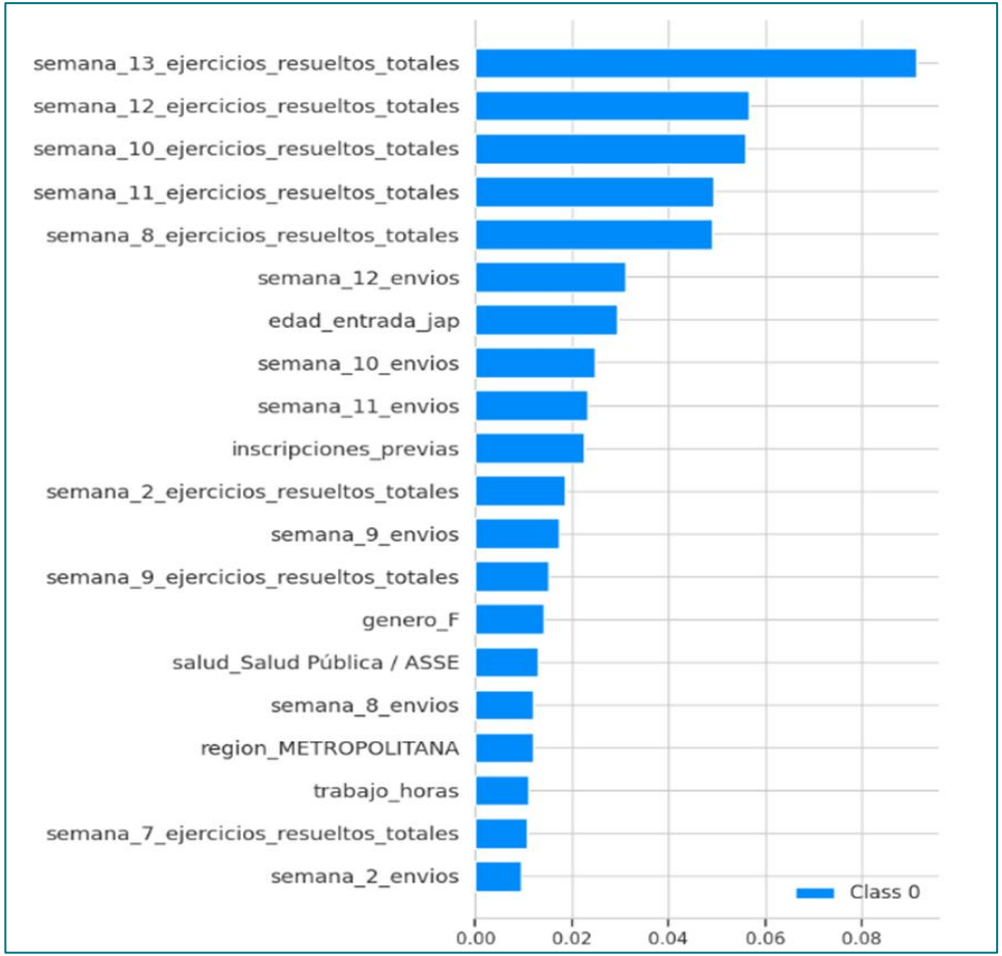
	precision	recall	f1-score	support
0	0.93	0.89	0.91	2825
1	0.67	0.76	0.71	803
accuracy			0.86	3628
macro avg	0.80	0.83	0.81	3628
weighted avg	0.87	0.86	0.87	3628





DNN

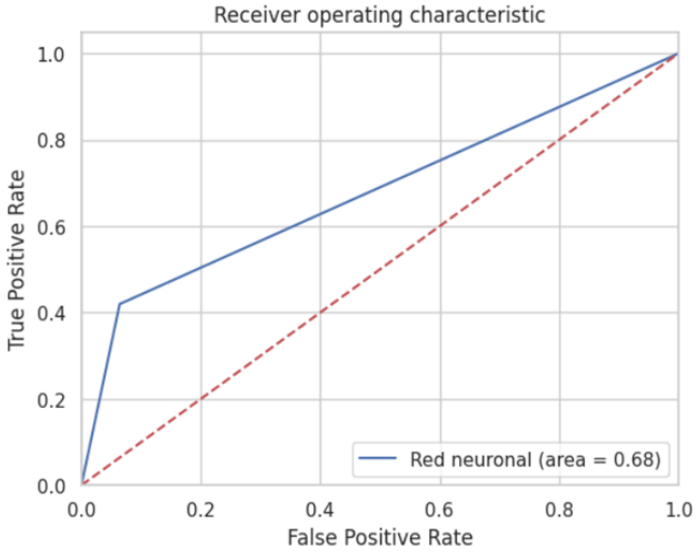
Importancia de las variables

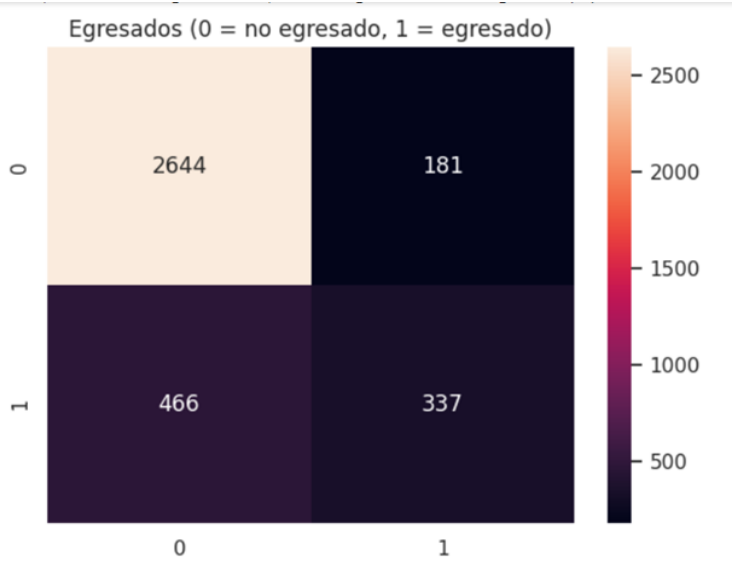


DNN

Métricas de evaluación

	precision	recall	f1-score	support
0	0.85	0.94	0.89	2825
1	0.65	0.42	0.51	803
accuracy			0.82	3628
macro avg	0.75	0.68	0.70	3628
weighted avg	0.81	0.82	0.81	3628





Métricas de evaluación

Tabla comparativa

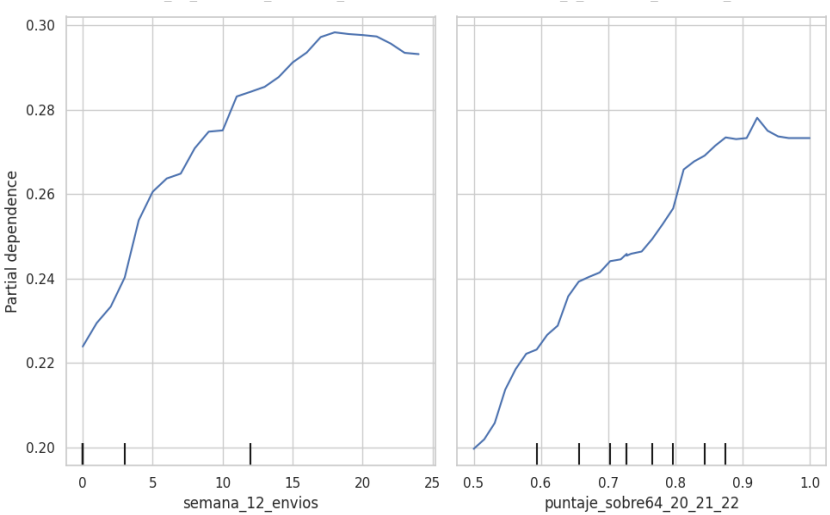
Algoritmo	accuracy	AUC	precision (0)	recall (0)	f1-score (0)
Logistic regression	0.75	0.79	0.95	0.72	0.82
Random forest	0.87	0.81	0.91	0.93	0.92
XGBoost	0.86	0.83	0.93	0.89	0.91
DNN	0.82	0.68	0.85	0.94	0.89

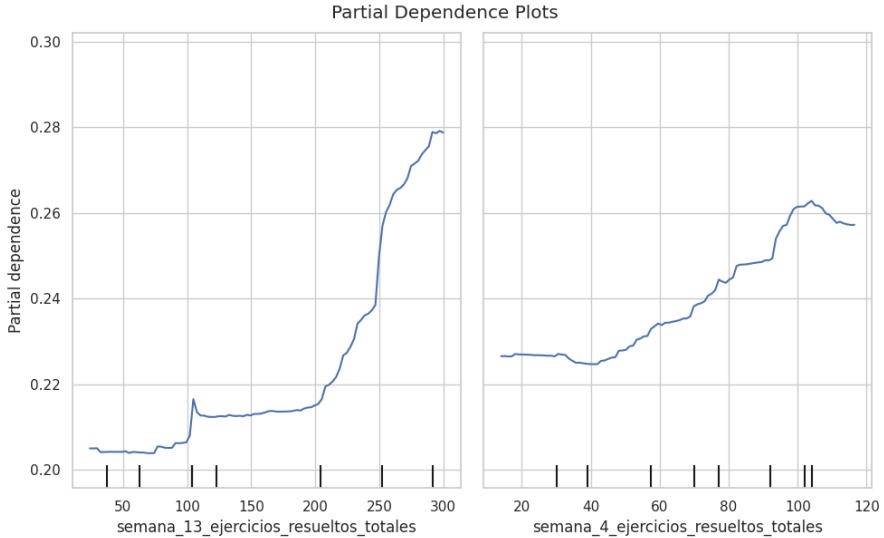
Nota: 0 = No egreso

Ceibal

Random Forest

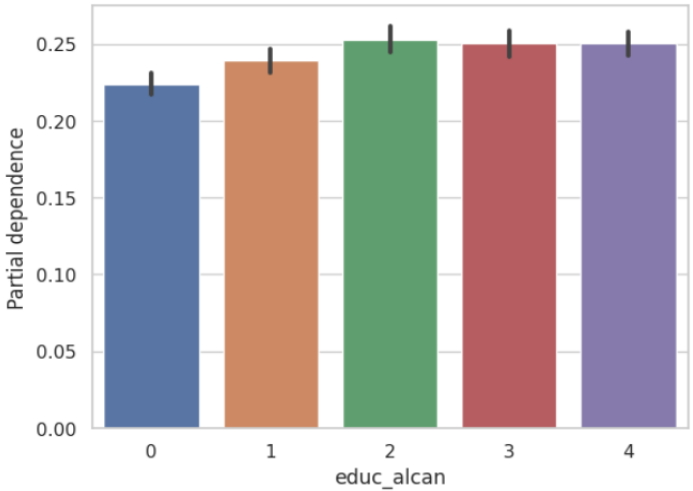
Partial dependence plots (PDP) I

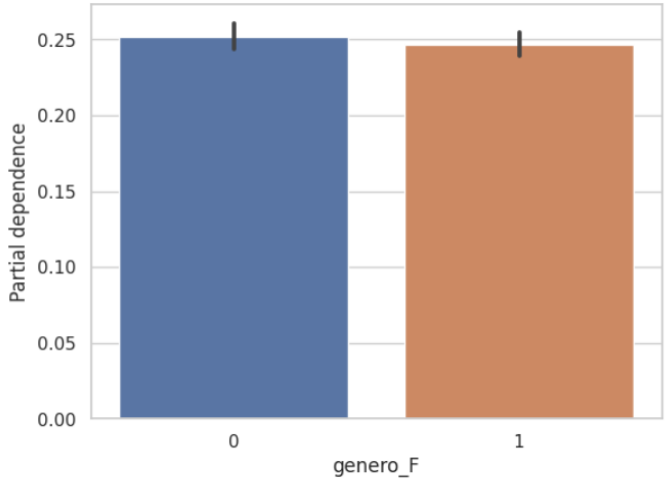




Random Forest

Partial dependence plots (PDP) II

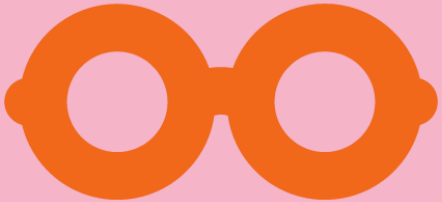




educ_alcan: educación alcanzada ("Ciclo básico/Media Básica": 0, "Otros estudios": 1, "Bachillerato/Educación Media Superior": 2, "Terciaria": 3, "Universitaria": 4).
genero_F: género (“Femenino”: 1, “Masculino”: 0)

Notas finales

- La ingeniería de datos y la aplicación de mecanismos de validación cruzada mejoraron significativamente el ajuste de los modelos.
- Los modelos testeados arrojan métricas de evaluación con valores muy favorables, sobre todo para la categoría: no egreso.
- Random Forest y XGBoost presentan métricas de evaluación con valores muy cercanos.
- Las variables creadas a partir de los datos de la plataforma Mumuki son las que más aportan a la disminución del error.



¡Gracias!

ceibal.edu.uy

Ceibal