Análisis predictivo del riesgo de deserción estudiantil mediante un Sistema de Alerta Temprana al inicio de la carrera

Ignacio Moraga Silva

Proyecto final - Aprendizaje de máquina: Técnicas y aplicaciones en series de tiempo y hardware

Profesor: Rodrigo Cádiz

Pontificia Universidad Católica de Chile

12 de diciembre de 2025

Introducción

La deserción estudiantil en la educación superior es un gran problema tanto para los estudiantes y las Instituciones de Educación Superior [1]

En Chile, cerca del **22.7%** de los estudiantes abandona su carrera tras cursar su primer año [1,2,3] e incluso en carreras como Ingeniería Informática esta cifra ha llegado al 50% para el tercer año [3]

Problema: **Instituciones suelen actuar de manera reactiva**, interviniendo cuando el fracaso académico ya ocurrió. Esto implica perder un estudiante dado que no se empleó la oportunidad de intervención temprana

Introducción

Se propone un Sistema de Alerta Temprana que **perfile el riesgo de deserción** utilizando exclusivamente información disponible al momento de la matrícula, para detectar quiénes necesitan apoyo antes de que sea demasiado tarde

Metodología e Implementación

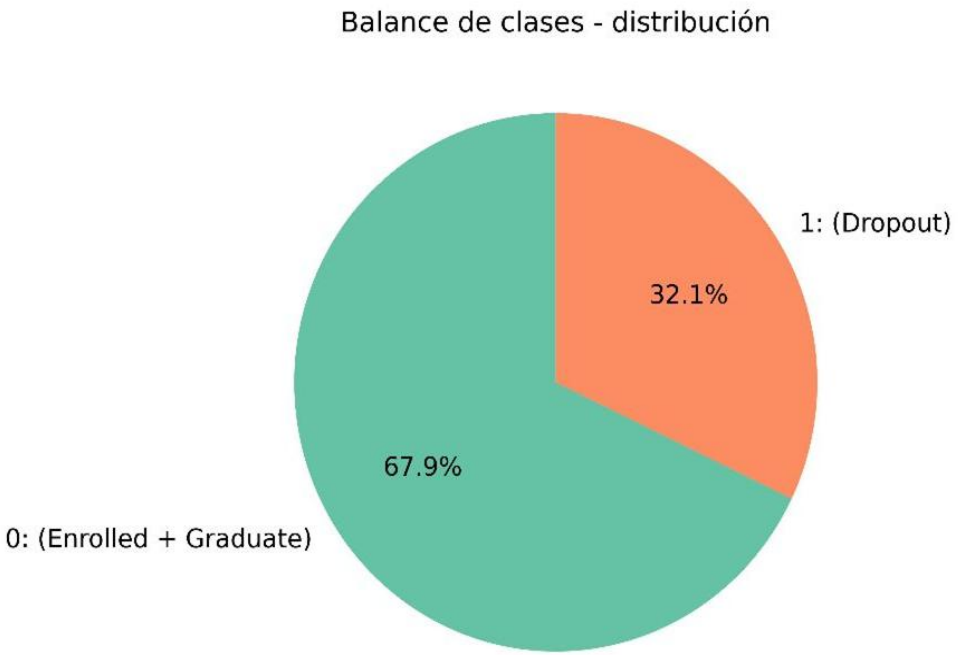
Clasificador binario con el objetivo de identificar si un alumno es **desertor (Clase 1)** o **no desertor (Clase 0)** según características demográficas, socioeconómicas y académicas previas

Dataset empleado: **Predict students' dropout and academic success - Investigating the Impact of Social and Economic Factors** [4]

- ➔ Datos de 4424 estudiantes matriculados en cursos de pregrado en el Instituto Politécnico de Portalegre, Portugal (Kaggle)
- ➔ *Dataset* "proxy" / "sustituto" para efectos del proyecto
- ➔ Se descartaron del dataset variables que no se conocen al momento de la matrícula

Metodología e implementación

Problemas: Desbalance en el *dataset* y alta dimensionalidad de algunas variables categóricas



Fuente: Elaboración propia.

Dataset considerado (22 características)

→ *One-hot Encoding* para variables categóricas (**221 características**)

Dataset considerado (22 características)

→ Agrupación de categorías con menos de 2% de frecuencia

→ *One-hot Encoding* para variables categóricas (**83 características**)

Metodología e implementación

Modelos empleados y prevención de sobreajuste:

- ➔ **Regresión Logística** (scikit-learn) → *class_weight* = *balanced*
- ➔ Gradient Boosting (scikit-learn + **XGBoost**) → *scale_pos_weight* = *2.11* (instancias de clase 0 sobre instancias de clase 1)
- ➔ Red neuronal **MLP** (scikit-learn) → SMOTE para generar datos sintéticos de la clase minoritaria y asegurar que el modelo aprendiera efectivamente a identificar a los desertores [5,6]

Para cada modelo se realizó validación cruzada estratificada con K=5 (*folds*)

Resultados: Entrenamiento y Validación Cruzada

Métrica	Regresión Logística	XGBoost	MLP
Accuracy	0.6937 (± 0.0339)	**0.7090 (± 0.0281)**	0.6982 (± 0.0251)
ROC-AUC	**0.7617 (± 0.0294)**	0.7481 (± 0.0277)	0.7503 (± 0.0265)
F1-score Macro	**0.6734 (± 0.0339)**	0.6662 (± 0.0287)	0.6700 (± 0.0249)
Recall (Clase 1: `Dropout`)	**0.6905 (± 0.0382)**	0.5462 (± 0.0348)	0.6314 (± 0.0343)
Precision (Clase 1: `Dropout`)	0.5189 (± 0.0409)	**0.5492 (± 0.0444)**	0.5266 (± 0.0359)
Neg Log-Loss	**-0.5798 (± 0.0262)**	-0.6279 (± 0.0489)	-0.5842 (± 0.0246)

Resultados promedios de la validación cruzada ($K = 5$). Se destaca en negrita el mejor desempeño por métrica. Fuente: Elaboración propia.

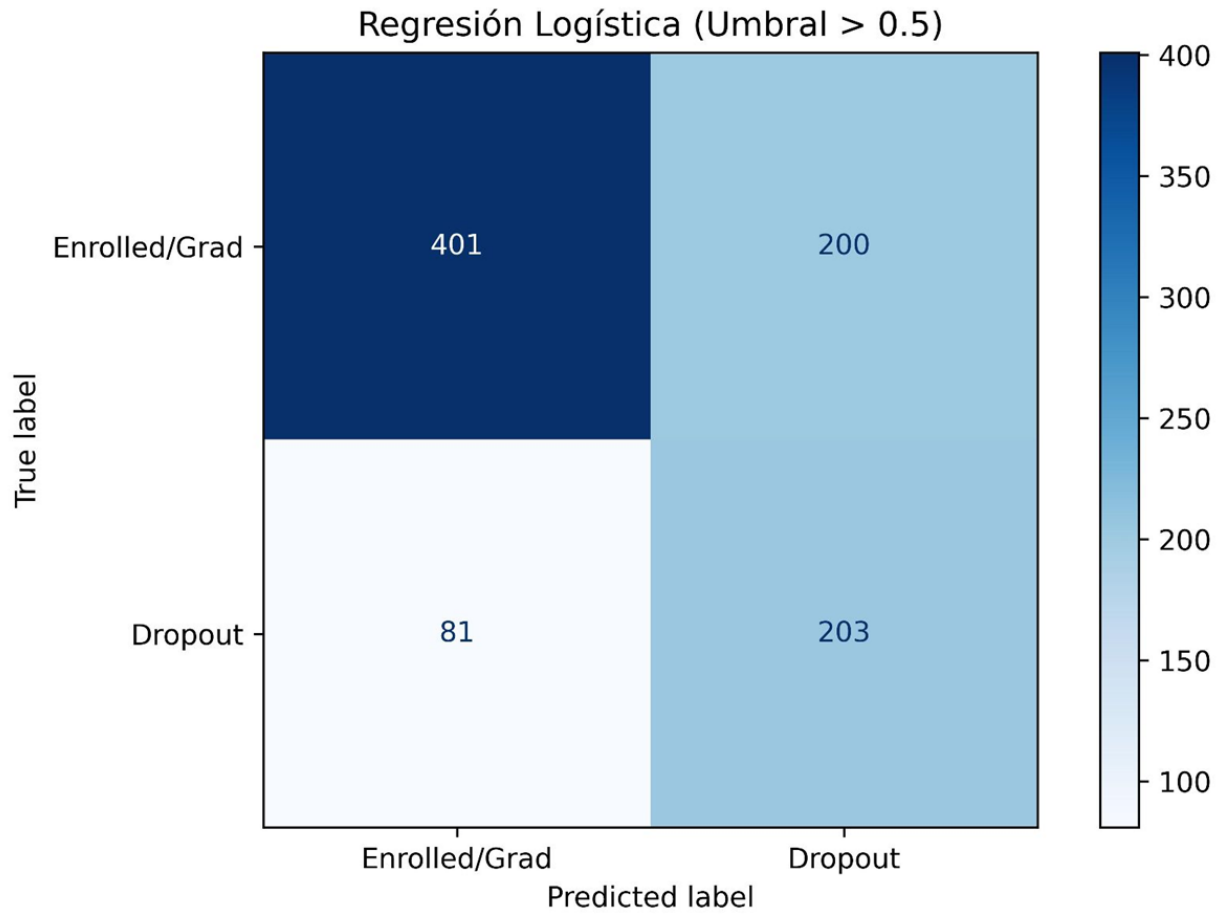
Resultados: Inferencia

Métrica	Valor
Accuracy	0.68
ROC-AUC	0.7754
F1-score Macro	0.67
Recall (Clase 1: `Dropout`)	**0.71**
Recall (Clase 0: `Enrolled/Graduate`)	0.67
Precision (Clase 1: `Dropout`)	0.50
Precision (Clase 0: `Enrolled/Graduate`)	0.83

Métricas para inferencia realizada con modelo de Regresión Logística sobre el set de prueba. Fuente: Elaboración propia.

Recall del 71% → **7 de cada 10 estudiantes son detectados correctamente como desertores en el futuro**

Resultados: Inferencia



Matriz de confusión para inferencia en conjunto de prueba con modelo de Regresión Logística. Fuente: Elaboración propia.

Interpretabilidad y características determinantes

Tipo de Factor	Feature	Coeficiente/parámetro	Odds Ratio
Factores que aumentan riesgo de deserción	Mother's occupation_1	0.9960	2.7073
	Mother's qualification_19	0.8760	2.4014
	Course_8	0.8078	2.2430
	Course_16	0.6511	1.9177
	Application mode_4	0.6484	1.9124
Factores que reducen riesgo de deserción	Scholarship holder_1	-1.0997	0.3330
	Course_12	-0.6839	0.5046
	Course_3	-0.5773	0.5614
	Course_10	-0.5518	0.5759
	Application mode_1	-0.5400	0.5827

Los 5 factores más importantes asociados a la deserción y a la permanencia. El signo del coeficiente y el Odds Ratio [7] indican la dirección y magnitud del impacto de la variable asociada en la probabilidad de deserción (Clase 1), respectivamente. Fuente: Elaboración propia.

Conclusiones y Trabajo Futuro

- ➔ Se logró validar la efectividad de un Sistema de Alerta Temprana basada en un modelo predictivo que clasifica a alumnos en posibles desertores/no desertores en base a sus características al momento de la matrícula

- ➔ El modelo ganador es Regresión Logística, con un Recall alto y lograr generalizar en el conjunto de prueba

- ➔ Se identifican posibles políticas de acción a realizar por parte de las instituciones en base a los resultados

- ➔ A futuro se plantea segmentar inicialmente con el modelo actual para luego actualizar las probabilidades ocurridas las primeras pruebas/evaluaciones. También, variar el umbral de decisión

Referencias

[1] D. Opazo, S. Moreno, E. Álvarez Miranda, and J. Pereira, "Analysis of first-year university student dropout through machine learning models: A comparison between universities," *Mathematics*, vol. 9, no. 20, 2021. Available: https://www.mdpi.com/2227-7390/9/20/2599

[2] Servicio de Información en Educación Superior [SIES], "Informe 2025, Retención de 1er año de pregrado, Cohortes 2020-2024," Subsecretaría de Educación Superior Ministerio de Educación, Tech. Rep., 2025.

[3] F. A. Bello, J. Kohler, K. Hinrechsen, V. Araya, L. Hidalgo, and J. L. Jara, "Using machine learning methods to identify significant variables for the prediction of first-year informatics engineering students dropout," in *2020 39th International Conference of the Chilean Computer Science Society (SCCC)*, 2020, pp. 1-5. Available: 10.1109/SCCC51225.2020.9281280

[4] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting student dropout and academic success," *Data*, vol. 7, no. 11, 2022. Available: https://www.mdpi.com/2306-5729/7/11/146

[5] Imbalanced Learn, "SMOTE." [Online]. Available: https://imbalanced-learn.org/stable/references/generated/imblearn.oversampling.SMOTE.html

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, p. 321-357, 2002. Available: http://dx.doi.org/10.1613/jair.953

[7] U. (https://stats.stackexchange.com/users/182590/user1607), "logit interpreting coefficients as probabilities," *Cross Validated*. [Online]. Available: https://stats.stackexchange.com/q/363806