

Pronóstico de Graduación de Estudiantes Subgraduados Universidad de Puerto Rico Recinto Mayagüez

Jesús D. Hernández Londoño¹ & Roberto Rivera Santiago²

¹Universidad de Puerto Rico. Departamento de Matemáticas.
Recinto Mayagüez.

4 de mayo de 2022



Contenido

- 1 Introducción
- 2 Objetivo
- 3 Metodología
- 4 Resultados
- 5 Conclusiones
- 6 Referencias



Contenido

- 1 Introducción
- 2 Objetivo
- 3 Metodología
- 4 Resultados
- 5 Conclusiones
- 6 Referencias



Introducción

¿Qué beneficios brinda una población educada?

Una sociedad educada brinda a la nación seres humanos

- Capaces de luchar (comunicativos), generar cambio, crear y decidir.
- Capaces de respetar las diferencias.
- Un medio efectivo para poder acceder y ascender a los diferentes puestos de trabajo.
- La mejor garantía del progreso de una Nación.



Introducción

¿Cómo está la educación universitaria en Puerto Rico?

En promedio:

- Solo el 45.75 % de estudiantes subgraduados obtienen su grado a 150 % del tiempo.
- Algunas universidades ofrecen talleres y consejería a estudiantes para facilitar que se gradúen.
- No existe una manera objetiva de predecir si una estudiante en particular completará su grado.



Contenido

- 1 Introducción
- 2 **Objetivo**
- 3 Metodología
- 4 Resultados
- 5 Conclusiones
- 6 Referencias



Objetivo

Objetivo

Predecir si un estudiante subgraduado se gradúa al 150 % del tiempo de la Universidad de Puerto Rico Recinto Mayagüez, usando métodos de aprendizaje automático según varias variables del estudiante.



Contenido

- 1 Introducción
- 2 Objetivo
- 3 Metodología**
- 4 Resultados
- 5 Conclusiones
- 6 Referencias



Metodología

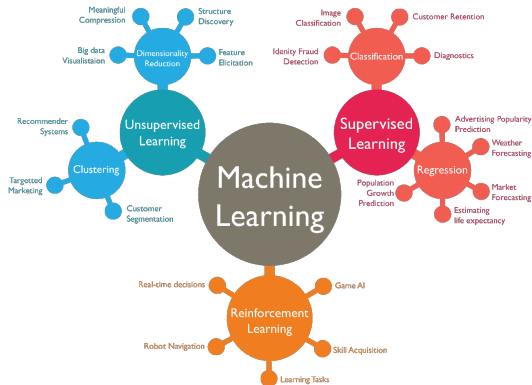


Figura 1: Tipos de aprendizaje (Extraída de towards data science [2])



Metodología

Supongamos que se tiene una respuesta cualitativa Y_i que representa si un estudiante i se graduó ($Y_i = 0$) o no ($Y_i = 1$) y p diferentes predictoras $X = (X_{i,1}, \dots, X_{i,p})$, existe una función f tal que

$$Y_i = f(X_{i,1}, \dots, X_{i,p})$$

Para indicar la predicción de Y se usa \hat{f} la estimación de f donde \hat{Y} representa el resultado de predicción de Y

$$\hat{Y}_i = \hat{f}(X_{i,1}, \dots, X_{i,p})$$

Para cuantificar la precisión del modelo \hat{f} se usa la tasa de error

$$\frac{1}{n} \sum_{i=1}^n I(Y_i \neq \hat{Y}_i)$$



Metodología

- Se preprocesan los datos. Extracción de predictoras y eliminación de datos duplicados.
- Se imputan datos usando la imputación multivariante por ecuaciones encadenadas (MICE).
- Exploración de los datos.
- Dos versiones del modelo: uno antes de comenzar estudios subgraduados, y otro luego del 1er año en el RUM.
- Comparar Métodos para validar predicciones.
 - Árboles de Clasificación.
 - Bosque Aleatorio.
 - Boosting.
 - Naïve Bayes.
 - Regresión Logística.
 - TabNet. No se han confirmado sus componentes teóricos. En consecuencia no se ha publicado en algún jornal.



Metodología

	Actual Positive	Actual Negative
Predicted Positive	True Positive (TP)	False Positive (FP) Type II Error
Predicted Negative	False Negative (FN) Type I Error	True Negative (TN)

Cuadro 1: Matriz de Confusión para métodos de aprendizaje automatizado.

- Sensibilidad = Recall = $\frac{TP}{TP+FN}$.
- $P1 = \frac{FN}{FN+TP} = 1 - \text{Recall}$.
- $P2 = \frac{FN}{FN+TP+FP+TN}$.
- AUC es la proporción de casos positivos y negativos predichos correctamente por el modelo.



Contenido

- 1 Introducción
- 2 Objetivo
- 3 Metodología
- 4 Resultados**
- 5 Conclusiones
- 6 Referencias



Datos

Datos

- Cohorte hasta el año de comienzo 2010.
- Observaciones 24,432. Predictoras 17.
- De los 24,432 resultan 19,546 para entrenamiento (80 %) y 4,886 (20 %) de prueba.

Predictoras

Año, Facultad, Programa de admisión, Apt Verbal, Aprov Matemáticas, Apt Matemáticas, Aprov Español, Aprov Inglés, Ingreso Familiar, Educación Padre, Educación Madre, Genero, Tipo de Escuela, GPA Primer año, Rel Estudiante GPA, Rel Escuela GPA y Graduación.



Análisis Exploratorio

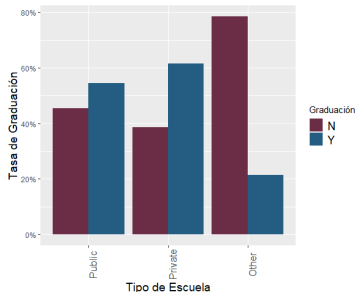


Figura 4: Tasa de Graduación por Tipo de Escuela.

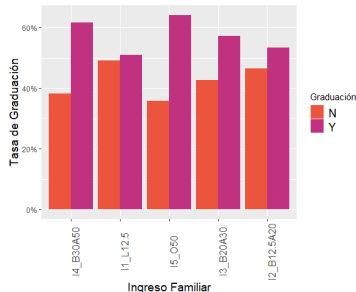


Figura 5: Tasa de Graduación por Ingreso Familiar.



Análisis

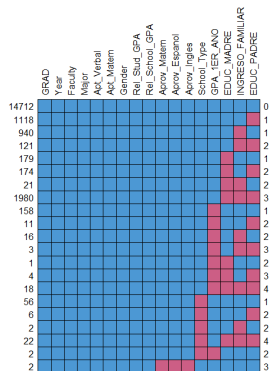


Figura 6: Patrones
Valores Faltantes

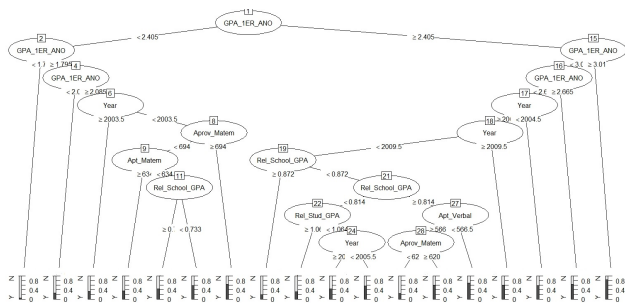


Figura 7: Ejemplo de Árbol decisional.



Rendimiento Predictivo

Métrica	Árbol de Clasificación	Bosque Aleatorio	Boosting	Naive Bayes	Regresión Logística	Tabnet
Sensibilidad	0.4896	0.4469	0.7260	0.5217	0.4500	0.7017
AUC	0.5953	0.6672	0.6868	0.6679	0.7023	0.6707
P(Modelo predice si se gradúa Realmente no se gradúa)	0.5104	0.5531	0.2740	0.4783	0.5500	0.2983
P(Modelo predice incorrectamente que el estudiante si se gradúa)	0.2205	0.2389	0.1184	0.2066	0.2376	0.1265

Cuadro 2: Modelos antes de entrar a la universidad.

Métrica	Árbol de Clasificación	Bosque Aleatorio	Boosting	Naive Bayes	Regresión Logística	Tabnet
Sensibilidad	0.6166	0.6223	0.7756	0.6298	0.6675	0.8046
AUC	0.8155	0.8286	0.8373	0.7831	0.8392	0.8168
P(Modelo predice si se gradúa Realmente no se gradúa)	0.3834	0.3777	0.2244	0.3702	0.3325	0.1954
P(Modelo predice incorrectamente que el estudiante si se gradúa)	0.1656	0.1632	0.0969	0.1599	0.1436	0.0829

Cuadro 3: Modelos luego del primer año en el RUM.



Rendimiento Predictivo

Métrica	Árbol de Clasificación	Bosque Aleatorio	Boosting	Naive Bayes	Regresión Logística	Tabnet
Sensibilidad	0.5135	0.4944	0.7591	0.5168	0.4516	0.341
AUC	0.6154	0.6639	0.6936	0.6558	0.6896	0.5304
P(Modelo predice si se gradúa Realmente no se gradúa)	0.4865	0.5056	0.2409	0.4832	0.5484	0.6589
P(Modelo predice incorrectamente que el estudiante si se gradúa)	0.1996	0.2075	0.0988	0.1983	0.2250	0.2860

Cuadro 4: Modelos antes de entrar a la universidad - datos imputados.

Métrica	Árbol de Clasificación	Bosque Aleatorio	Boosting	Naive Bayes	Regresión Logística	Tabnet
Sensibilidad	0.6485	0.6372	0.7650	0.6149	0.6702	0.7462
AUC	0.7611	0.8218	0.8307	0.7705	0.8312	0.8195
P(Modelo predice si se gradúa Realmente no se gradúa)	0.3515	0.3621	0.235	0.3851	0.3298	0.2537
P(Modelo predice incorrectamente que el estudiante si se gradúa)	0.1442	0.1486	0.0964	0.1580	0.1353	0.1099

Cuadro 5: Modelos luego del primer año en el RUM - datos imputados.



Boosting

Hiperparámetros

Finalmente para Boosting se eligen los hiperparámetros de un espacio de búsqueda de valores, obteniendo N trees= 99, Interaction depth= 3, Shrinkage= 0,10, N min obs in node= 15 y Bag fraction= 1.

	Actual N	Actual Y
Predicted N	1153	1099
Predicted Y	366	1083

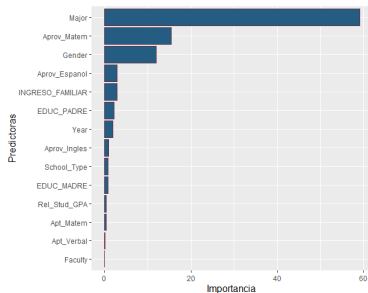
Cuadro 6: Matriz de confusión Boosting.



Boosting

Métricas de Evaluación

- Sensibilidad = $\frac{1153}{1153+366} = 0,759$
- $P(\text{Modelo predice si se gradúa} \mid \text{Realmente no se gradúa}) = \frac{366}{1153+366} = 0,2409$
- $P(\text{Modelo predice si se gradúa pero realmente no se gradúa}) = \frac{366}{1153+366+1099+1083} = 0,0988$



Tabnet

Hiperparámetros

Finalmente para TabNet se eligen los hiperparámetros de un espacio de búsqueda de valores, obteniendo Max epochs= 22, Batch size= 1029, Virtual batch size=131 y Patience= 22.

	Actual N	Actual Y
Predicted N	626	364
Predicted Y	152	691

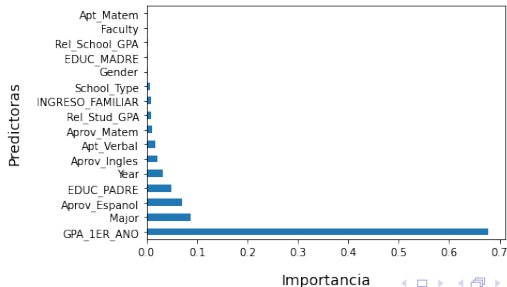
Cuadro 7: Matriz de confusión TabNet.



Tabnet

Métricas de Evaluación

- Sensibilidad = $\frac{626}{626+152} = 0,8046$
- $P(\text{Modelo predice si se gradúa} \mid \text{Realmente no se gradúa}) = \frac{152}{626+152} = 0,1953$
- $P(\text{Modelo predice si se gradúa pero realmente no se gradúa}) = \frac{152}{626+364+152+691} = 0,0829$



Contenido

- 1 Introducción
- 2 Objetivo
- 3 Metodología
- 4 Resultados
- 5 Conclusiones**
- 6 Referencias



Conclusiones

Conclusiones

- La imputación de los datos mejoró las predicciones solo para Boosting en el primer caso (un modelo para antes de entrar a la universidad).
- Boosting identifica correctamente el 75,91 % de los estudiantes que no se graduaron y la probabilidad de predecir incorrectamente que el estudiante si se gradúa es del 9,88 %.
- Tabnet identifica correctamente el 80,46 % de los estudiantes que no se graduaron y la probabilidad de predecir incorrectamente que el estudiante si se gradúa es del 8,29 %.



Conclusiones

Conclusiones

- Se pueden desarrollar estrategias de intervención basados en la detección de estudiantes con bajas probabilidades de graduación.
 - Consejería académica.
 - Asistencia en manejo de tiempo y estrés.
 - Orientación de programas en el RUM que ayudan al estudiante.



Conclusiones

Trabajo Futuro

- Es importante resaltar que como los métodos tienen que ser eficientes la intervención de los oficiales también lo debe ser.
- Considerar XGBoost un método de ensamble de aprendizaje automático que ha superado a otros métodos en diferentes competiciones de Kaggle.
- Planificar nuevos modelos ML a implementar una vez intervención rinda frutos.



Contenido

- 1 Introducción
- 2 Objetivo
- 3 Metodología
- 4 Resultados
- 5 Conclusiones
- 6 Referencias**



Referencias



Barshay, J. y S. Aslanian. 2019.

Colleges are using big data to track students in an effort to boost graduation rates, but it comes at a cost.

Enlace del Hechinger Report



El costo creciente de no ir a la universidad. Pew Research Center. 2014.

<https://www.pewresearch.org/social-trends/2014/02/11/the-rising-cost-of-not-going-to-college/>






¿Vale la pena una educación universitaria?. ProCon.org. 2020.

<https://college-education.procon.org/>



Referencias

-  U.S. Bureau of Labor Statistics. 2019.
<https://www.bls.gov/emp/chart-unemployment-earnings-education.htm>
-  Lorberfeld A. 2019.
Machine Learning Algorithms In Layman's Terms, Part 1 [En Línea].
Disponible:
Towards data science
-  Rodriguez, A. 2010.
Retención de Estudiantes



Referencias



Karamouzis, S.T. y A. Vrettos. 2008.

An artificial neural network for predicting student graduation outcomes.

in Proceedings of the World Congress on Engineering and Computer Science, in San Francisco, USA.



Ploutz, E. C. 2018.

Machine Learning Applications in Graduation Prediction at the University of Nevada, Las Vegas.

Tesis B.S., Universidad de Nevada, Las Vegas, N.V., USA.

<http://dx.doi.org/10.34917/13568668>

