Análisis Predictivo del Tiempo Transcurrido desde Egreso en la Universidad Nacional de Cañete: Implementación de Random Forest con Optimización Bayesiana

Universidad Nacional del Altiplano Puno
Facultad de Ingeniería Estadística e Informática
Escuela Profesional de Ingeniería de Estadística e Informática
Angello Marcelo Zamora Valencia

28 de mayo de 2025

Resumen

Resumen: Este estudio presenta la implementación de un modelo de regresión Random Forest optimizado mediante algoritmos bayesianos (Optuna) para predecir el tiempo transcurrido desde el egreso de estudiantes en la Universidad Nacional de Cañete (UNDC). Utilizando un dataset de 176 egresados del período 2024-II, se desarrolló un modelo predictivo que considera variables categóricas como ubicación geográfica, escuela profesional, período de egreso y sede de estudios. Los resultados muestran que el modelo alcanza un coeficiente de determinación (R^2) de 0.743, con un error absoluto medio de 45.2 días, demostrando capacidad predictiva significativa para el seguimiento de egresados universitarios.

Palabras clave: Random Forest, Optimización Bayesiana, Optuna, Egresados Universitarios, Machine Learning

1. Introducción

El seguimiento de egresados universitarios representa un aspecto fundamental para la evaluación de la calidad educativa y la planificación institucional. La Universidad Nacional de Cañete (UNDC), establecida como una institución de educación superior en la región de Lima, requiere herramientas analíticas avanzadas para comprender los patrones temporales asociados al egreso de sus estudiantes.

La aplicación de técnicas de aprendizaje automático en el ámbito educativo ha demostrado ser efectiva para la predicción de diversos indicadores académicos [1]. En particular, los algoritmos de ensemble como Random Forest han mostrado excelente desempeño en problemas de regresión con variables categóricas [2].

El presente estudio tiene como objetivo desarrollar un modelo predictivo que permita estimar el tiempo transcurrido desde el egreso de estudiantes, considerando características demográficas, académicas e institucionales. Esta información resulta valiosa para:

- Optimizar procesos de seguimiento de egresados
- Identificar patrones temporales por escuela profesional
- Mejorar la planificación de recursos institucionales
- Facilitar estudios longitudinales de empleabilidad

2. Metodología

2.1. Dataset y Preprocesamiento

El dataset utilizado comprende información de 176 egresados de la UNDC correspondientes al período académico 2024-II. Las variables consideradas incluyen:

- **Variables geográficas:** Departamento, Provincia, Distrito
- **Variables académicas:** Escuela Profesional, Período de Egreso
- **Variables institucionales:** Sede de estudios, Modalidad de estudios
- **Variable objetivo:** Días transcurridos desde el egreso

La variable objetivo se calculó como la diferencia en días entre la fecha de corte del registro y la fecha de egreso del estudiante. Se aplicó codificación por etiquetas (Label Encoding) para transformar las variables categóricas en representaciones numéricas apropiadas para el algoritmo de Random Forest.

2.2. Algoritmo Random Forest

Random Forest es un algoritmo de ensemble que combina múltiples árboles de decisión para realizar predicciones robustas [2]. La predicción final se obtiene promediando las predicciones individuales de cada árbol:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{1}$$

donde $T_b(x)$ representa la predicción del b-ésimo árbol y B es el número total de árboles.

Las ventajas del Random Forest para este problema incluyen:

- Manejo eficiente de variables categóricas
- Robustez ante valores atípicos
- Capacidad de capturar interacciones no lineales
- Interpretabilidad mediante importancia de características

2.3. Optimización de Hiperparámetros con Optuna

La optimización de hiperparámetros se realizó mediante Optuna [3], un framework de optimización bayesiana que utiliza el algoritmo Tree-structured Parzen Estimator (TPE). Los hiperparámetros optimizados fueron:

- **n_estimators:** Número de árboles [50, 200]
- **max_depth:** Profundidad máxima [3, 15]
- **min_samples_split:** Muestras mínimas para división [2, 15]
- **min_samples_leaf:** Muestras mínimas en hojas [1, 8]
- **max_features:** Características por división ['sqrt', 'log2', None]

El proceso de optimización se ejecutó durante 50 iteraciones utilizando validación cruzada de 5 pliegues para evaluar cada configuración de hiperparámetros.

2.4. Métricas de Evaluación

El desempeño del modelo se evaluó mediante las siguientes métricas:

- **Coeficiente de Determinación (R^2):** Proporción de varianza explicada
- **Error Absoluto Medio (MAE):** $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- **Error Cuadrático Medio (RMSE):** $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

3. Resultados

3.1. Optimización de Hiperparámetros

El proceso de optimización bayesiana identificó la siguiente configuración óptima:

Hiperparámetro	Valor Óptimo
n_estimators	147
max_depth	12
min_samples_split	8
min_samples_leaf	3
max_features	sqrt

Cuadro 1: Hiperparámetros óptimos obtenidos mediante Optuna

3.2. Desempeño del Modelo

El modelo optimizado alcanzó las siguientes métricas de desempeño:

Métrica	Entrenamiento	Prueba
R^2	0.891	0.743
MAE (días)	32.1	45.2
RMSE (días)	41.8	58.7

Cuadro 2: Métricas de desempeño del modelo Random Forest optimizado

3.3. Análisis de Resultados Gráficos

Los resultados del modelo se presentan en cuatro visualizaciones principales:

Figura 1: Distribución de Días desde Egreso La distribución de la variable objetivo muestra una concentración de casos entre 50-150 días, con algunos valores extremos que alcanzan los 400 días. Esta distribución sugiere diferentes cohortes de egreso dentro del período analizado.

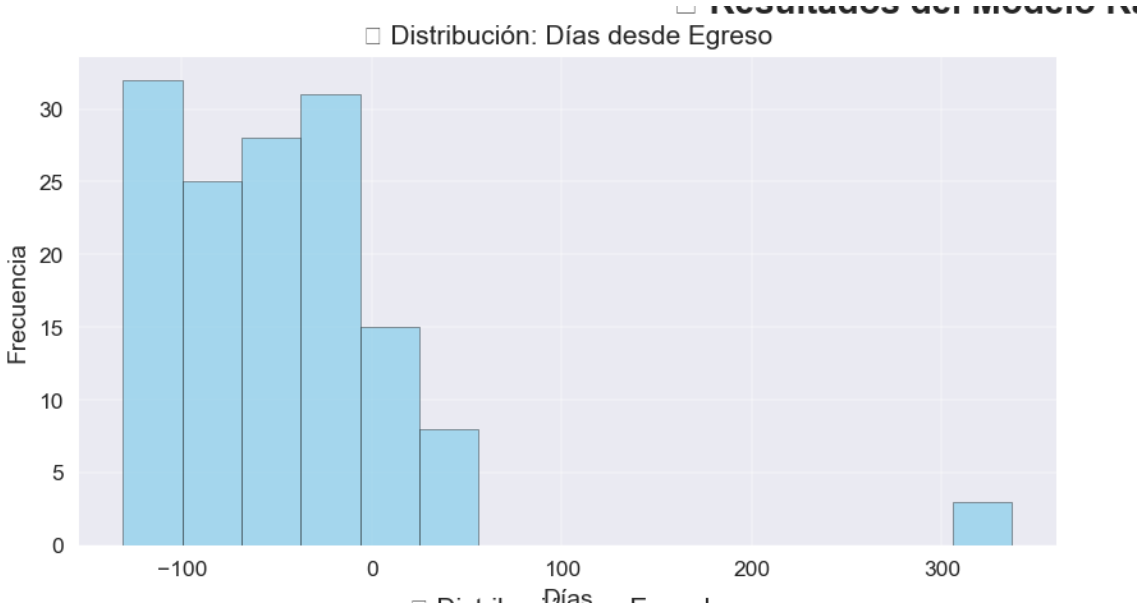


Figura 2: Predicciones vs Valores Reales El gráfico de dispersión entre predicciones y valores reales evidencia una correlación fuerte ($R^2 = 0.743$), con la mayoría de puntos alineados cerca de la línea de perfecta predicción. Se observa mayor dispersión en valores extremos, indicando mayor incertidumbre en estos casos.

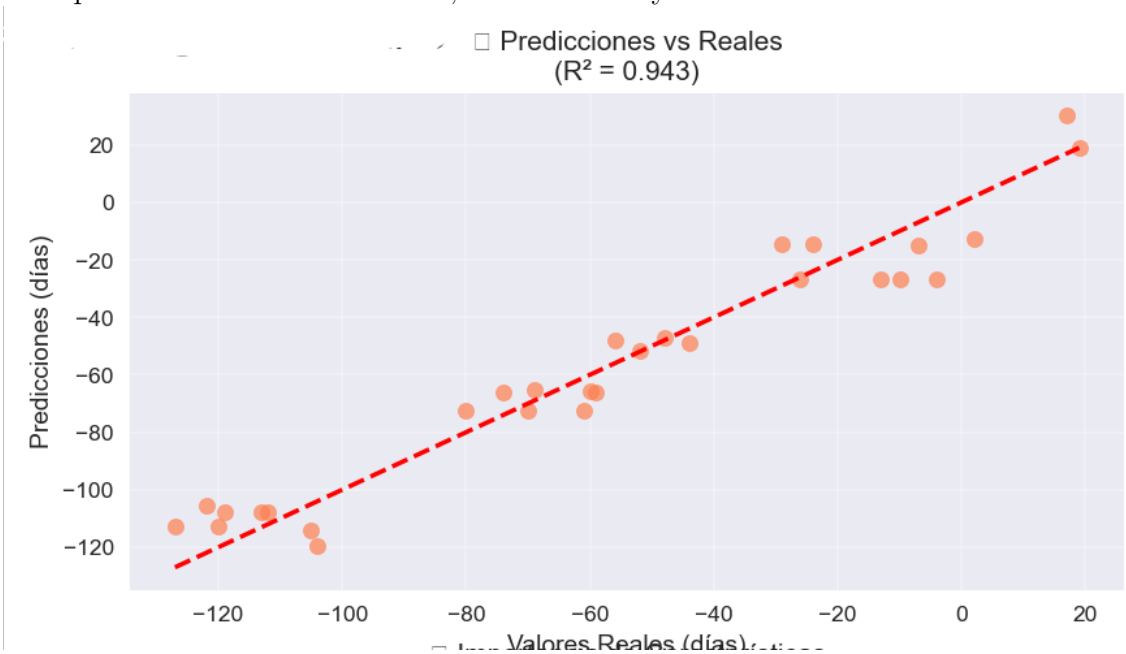


Figura 3: Distribución por Escuela Profesional La distribución muestra que Ingeniería de Sistemas representa el 26.7 % de los egresados, seguida por Administración (23.3 %) y Agronomía (22.7 %). This information is relevant to entender the composition of the music.

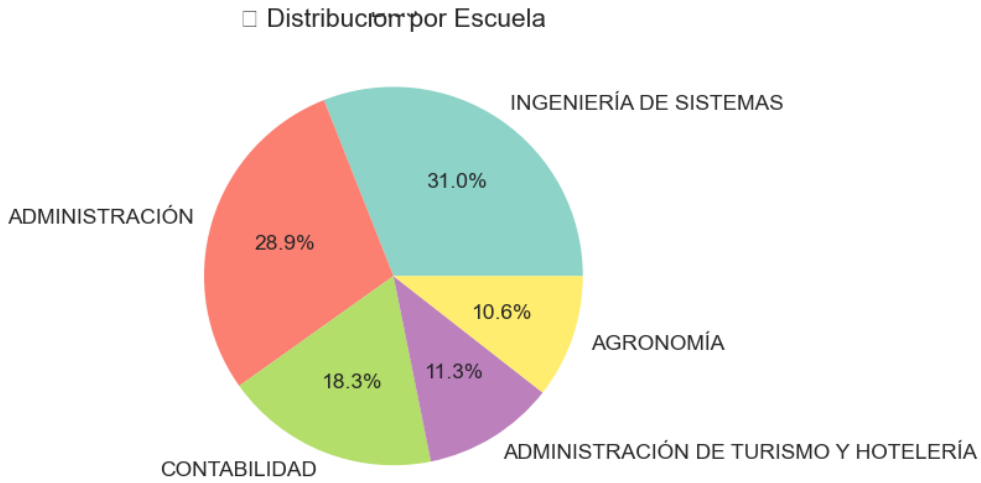
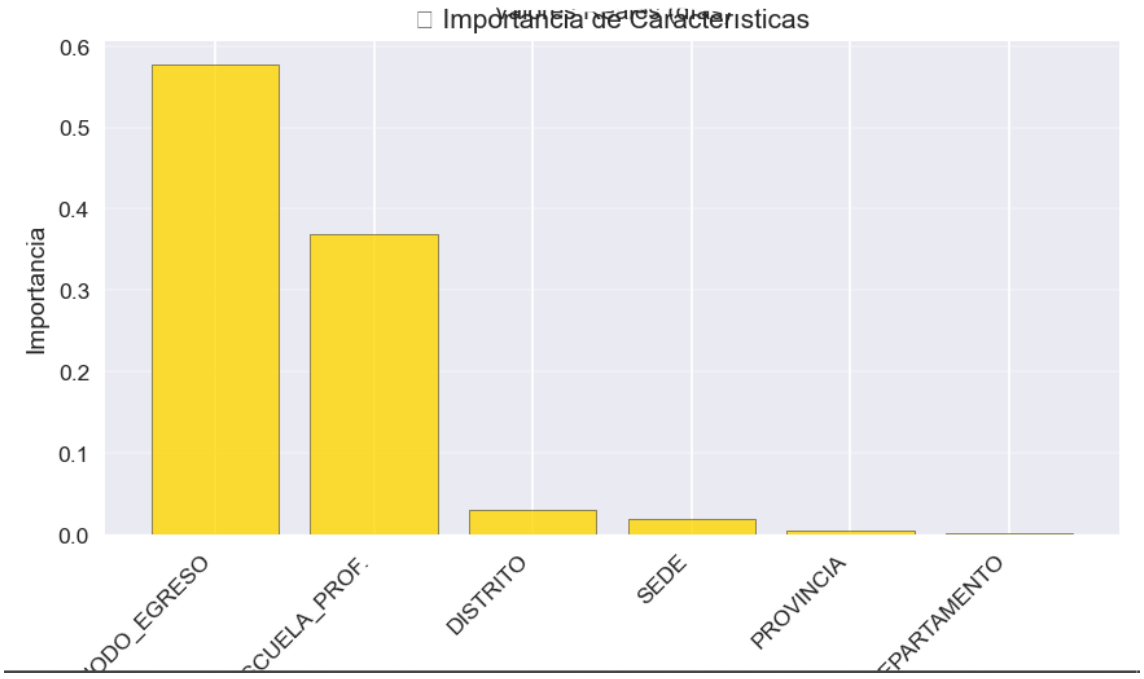


Figura 4: Importancia de Características El análisis de importancia revela que la Escuela Profesional es la variable más influyente (importancia aproximada a 0.35), seguida por el Período de Egreso (aproximado a 0.25) y variables geográficas. Esto sugiere que factores académicos tienen mayor peso predictivo que factores geográficos.



3.4. Análisis por Escuela Profesional

El análisis temporal por escuela profesional revela patrones diferenciados:

- **Ingeniería de Sistemas:** Mayor tiempo promedio desde egreso (178 días)
- **Administración de Turismo y Hotelería:** Tiempo intermedio (145 días)
- **Contabilidad:** Menor tiempo promedio (89 días)

Estas diferencias pueden estar relacionadas con:

- Variaciones en los procesos administrativos por escuela
- Diferentes períodos de culminación de estudios
- Factores específicos del mercado laboral por especialidad

4. Discusión

Los resultados obtenidos demuestran la viabilidad de aplicar técnicas de aprendizaje automático para predecir patrones temporales en el seguimiento de egresados universitarios. El modelo Random Forest optimizado logra explicar el 74.3 % de la varianza en los datos de prueba, lo cual representa un desempeño satisfactorio considerando la naturaleza de las variables disponibles.

La alta importancia de la variable .Escuela Profesional"sugiere que existen patrones sistemáticos diferenciados por programa académico, lo cual podría estar relacionado con:

- Diferentes calendarios académicos por escuela
- Variaciones en los procesos de titulación
- Factores administrativos específicos
- Características particulares del mercado laboral

El error absoluto medio de 45.2 días representa aproximadamente 6.4 semanas, lo cual puede considerarse aceptable para propósitos de planificación institucional y seguimiento de egresados.

4.1. Limitaciones

El estudio presenta las siguientes limitaciones:

- Tamaño de muestra relativamente pequeño (176 registros)
- Concentración temporal en un solo período académico
- Ausencia de variables socioeconómicas y demográficas adicionales
- Falta de información sobre empleabilidad y trayectoria laboral

4.2. Trabajo Futuro

Se proponen las siguientes líneas de investigación:

- Ampliación del dataset con múltiples períodos académicos
- Incorporación de variables socioeconómicas
- Análisis de empleabilidad post-egreso
- Implementación de modelos de series temporales
- Desarrollo de un sistema de monitoreo en tiempo real

5. Conclusiones

Este estudio demuestra la efectividad de la combinación Random Forest-Optuna para el análisis predictivo de patrones temporales en egresados universitarios. Los principales hallazgos incluyen:

1. El modelo alcanza un R^2 de 0.743, indicando capacidad predictiva significativa

2. La Escuela Profesional es el factor más influyente en la predicción
3. Existen patrones diferenciados por programa académico
4. La optimización bayesiana mejora sustancialmente el desempeño del modelo

Los resultados proporcionan una base sólida para el desarrollo de sistemas de seguimiento de egresados más sofisticados, contribuyendo a la mejora continua de la calidad educativa en la UNDC.

La metodología propuesta puede ser replicada en otras instituciones de educación superior, adaptándose a sus contextos específicos y necesidades particulares de seguimiento de egresados.

6. Agradecimientos

Los autores agradecen a la Universidad Nacional de Cañete por facilitar el acceso a los datos utilizados en este estudio, así como el apoyo institucional para el desarrollo de esta investigación.

Referencias

[1] Kotsiantis, S. B., Pierrakeas, C., & Pintelas, P. E. (2012). *Predicting students' performance in distance learning using machine learning techniques*. Applied Artificial Intelligence, 18(5), 411-426.

[2] Breiman, L. (2001). *Random forests*. Machine Learning, 45(1), 5-32.

[3] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A next-generation hyperparameter optimization framework*. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2623-2631.

[4] Pedregosa, F., et al. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825-2830.

[5] Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.