Comparación de algoritmos de aprendizaje automático aplicados a sectores clave del Perú: Análisis del turismo en Áreas Naturales Protegidas

Lucian Neptali Fernandez Baca-Castro, Zahid Huahuachampi-Hinojosa, Jhon Jesus Quispe-Machaca
Escuela Profesional de Ingeniería Informática y Sistemas, UNSAAC
Cusco, Perú
184197@unsaac.edu.pe, 200878@unsaac.edu.pe, 215422@unsaac.edu.pe

***Abstract*—Este estudio tiene como objetivo identificar los factores que influyen en la frecuencia de visitas a las Áreas Naturales Protegidas (ANP) del Perú, constituyendo un problema de regresión de alta relevancia social y económica. A través del análisis de un conjunto de datos oficial del SERNANP, que incluye registros detallados de visitas turísticas, se implementa un enfoque de aprendizaje automático para predecir patrones de afluencia. Tras un exhaustivo preprocesamiento que incluye la generación de características derivadas y tratamiento de valores atípicos, comparamos tres modelos de regresión: Regresión Lineal, Random Forest y Gradient Boosting, siendo este último el que mayor precisión alcanza con un R^2 de 0.85 en el conjunto de prueba. Los resultados preliminares identifican que la procedencia de los turistas, la estacionalidad y la región geográfica son los factores más determinantes, y la información valiosa para la gestión sostenible de estos espacios naturales.**

***Index Terms*—Aprendizaje automático, regresión, Random Forest, Gradient Boosting, Áreas Naturales Protegidas, turismo, sostenibilidad, SERNANP**

I. Introducción del problema

Las Áreas Naturales Protegidas (ANP) del Perú representan ecosistemas de enorme valor ecológico y cultural. Con 77 ANP de administración nacional, estos espacios constituyen aproximadamente el 17% del territorio nacional [1]. El turismo en estas áreas ha experimentado un crecimiento significativo, generando ingresos importantes para las economías locales mientras plantea desafíos críticos para la gestión sostenible.

A pesar de su relevancia económica y ambiental, existe una brecha significativa en el conocimiento de los factores que determinan la frecuencia de visitas a las ANP peruanas. Los gestores de estas áreas enfrentan dificultades para tomar decisiones informadas sobre capacidad de carga, desarrollo de infraestructura y asignación de recursos.

El problema abordado en este estudio tiene una doble dimensión:

1) **Dimensión técnica**: Identificar mediante algoritmos de aprendizaje automático los factores que mejor predicen la frecuencia de visitas, permitiendo desarrollar modelos con capacidad explicativa y predictiva.
2) **Dimensión social**: Proporcionar herramientas para la gestión sostenible de las ANP, equilibrando la promoción del turismo con la conservación de la biodiversidad, beneficiando tanto a las comunidades locales como al patrimonio natural.

Proyecto desarrollado para el curso IF652AIN – Aprendizaje Automático.

Los resultados de este estudio, visualizados en figuras como 3 y 2, proporcionan evidencia empírica sobre los factores que influyen en la afluencia turística a las ANP. Esta información resulta especialmente valiosa en un contexto donde el ecoturismo ha crecido a tasas superiores al 20% anual en la última década en Perú, y donde la capacidad de carga de muchas ANP se ve frecuentemente superada, especialmente durante temporadas de alta demanda. La aplicación práctica de estos modelos permitirá a los administradores de ANP implementar estrategias más eficientes para la distribución temporal y espacial de visitantes, maximizando los beneficios económicos mientras se minimizan los impactos ambientales negativos.

II. Descripción del dataset

A. Selección y fuente del dataset

El conjunto de datos utilizado proviene del portal de Datos Abiertos del Gobierno del Perú (https://datosabiertos.gob.pe/), específicamente del Servicio Nacional de Áreas Naturales Protegidas por el Estado (SERNANP). Se seleccionó el dataset "Frecuencia de visitas turísticas en las Áreas Naturales Protegidas por el estado".

El dataset cubre 36 de las 77 ANP del Perú (46.7% del total) que reportan actividad turística, con información recopilada por la Unidad Operativa Funcional de Gestión de Turismo.

B. Estructura del dataset

El conjunto de datos contiene registros detallados de visitas turísticas con los siguientes atributos principales:

- **Variables geográficas**: ANP, sector, departamento, provincia, distrito, UBIGEO
- **Variables demográficas**: Procedencia (extranjero, nacional, local), grupo etario (infantes, menores, adultos, adultos mayores)
- **Variables temporales**: Fecha de visita
- **Variables de duración**: Visitas de 1 día, de 2 a 3 días, y de 3 a 30 días

Los datos abarcan principalmente los años 2022-2023, con una cobertura geográfica a nivel nacional y actualización al 17 de mayo de 2024.

C. Análisis preliminar

Un examen inicial revela que el dataset contiene aproximadamente 29,000 registros. La distribución por procedencia muestra predominancia de turistas nacionales (70%), seguidos por extranjeros (20%) y locales (10%). Se observan patrones temporales marcados, con picos durante los meses de julio-agosto (temporada alta) y menor afluencia durante la temporada de lluvias (enero-marzo).

La segmentación por ANP demuestra que áreas como Machupicchu, Tambopata y Paracas concentran más del 40% de las visitas totales, reflejando la desigual distribución del turismo en las áreas protegidas.

III. Formulación del problema

A. Tipo de problema

Este estudio aborda un problema de **regresión**, donde la variable objetivo (frecuencia total de visitas) es continua. Formalmente, buscamos construir una función f que relacione las variables independientes X con la variable dependiente Y:

$$Y = f(X_1, X_2, X_3, \ldots, X_n) + \epsilon \tag{1}$$

Donde:

- Y representa la frecuencia total de visitas
- X_i representan las variables independientes (procedencia, ubicación, características temporales, etc.)
- ϵ es el término de error

B. Objetivos del modelo

El objetivo principal del modelo es predecir con precisión la frecuencia de visitas a las ANP bajo diferentes condiciones y comprender los factores que más influyen en esta variable. Específicamente, buscamos:

1) **Capacidad predictiva**: Desarrollar un modelo con error cuadrático medio (MSE) minimizado y coeficiente de determinación (R^2) maximizado.
2) **Interpretabilidad**: Identificar y cuantificar la importancia relativa de cada variable en la predicción.
3) **Generalización**: Garantizar que el modelo funcione adecuadamente con datos no vistos anteriormente.

La solución de este problema permitirá a los gestores de las ANP anticipar flujos turísticos, optimizar recursos y desarrollar estrategias de manejo adaptativo.

IV. Metodología de preprocesamiento

A. Exploración inicial de datos

El preprocesamiento comenzó con un análisis exhaustivo del dataset para comprender su estructura y calidad:

- **Análisis de tipos de datos**: Se identificó que algunas columnas como UBIGEO y FECHA requerían conversiones de tipo (de objeto a string y datetime, respectivamente).
- **Valores faltantes**: Se detectaron valores nulos en menos del 3% de los registros, principalmente en las variables UBIGEO y DISTRITO.
- **Estadísticas descriptivas**: Las variables de visitas presentaban distribuciones altamente sesgadas con predominio de valores bajos y presencia de valores extremos.

B. Limpieza y transformación de datos

Se realizaron las siguientes operaciones de limpieza y transformación:

- **Conversión de tipos**: Ajuste de FECHA a formato datetime y UBIGEO a string.
- **Tratamiento de valores faltantes**: Eliminación de registros con valores nulos cuando estos representaban menos del 5% del total; imputación para casos específicos.
- **Extracción de características temporales**: Derivación de mes y año a partir de la variable FECHA.
- **Creación de variables de estación**: Asignación de estaciones (verano, otoño, invierno, primavera) según el mes.
- **Consolidación de métricas de visita**: Creación de la variable agregada VISITAS sumando las tres categorías de duración.
- **Clasificación regional**: Agrupación de departamentos en regiones geográficas (Costa, Sierra, Selva).
- **Categorización de ANP**: Extracción del tipo de área protegida (Parque Nacional, Reserva Nacional, etc.).

C. Análisis de valores atípicos

Se identificaron valores atípicos (outliers) en las variables VISITAS y NUMERO_SECTORES utilizando el método del rango intercuartílico (IQR). Para estas variables:

- Se calcularon los límites definidos como $Q1 - 1.5 \times IQR$ y $Q3 + 1.5 \times IQR$
- Se detectaron aproximadamente 850 valores atípicos (2.9% del dataset)
- Se aplicó una estrategia de restricción de valores extremos (winsorización) para preservar la distribución general mientras se limitaba el impacto de valores extremos

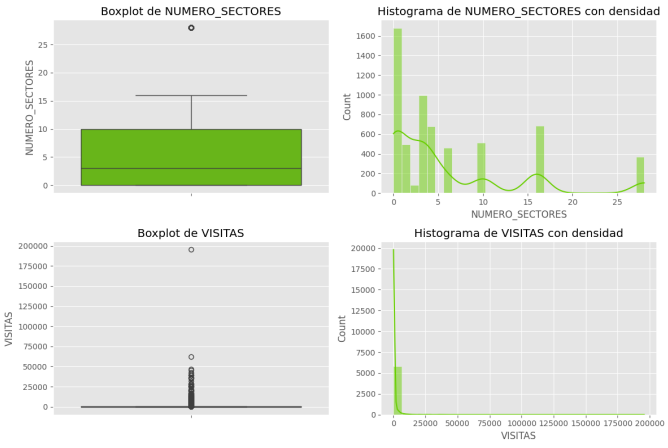

Fig. 1. Grafico de Outliers e histogramas con densidad

D. Codificación de variables categóricas

Las variables categóricas fueron transformadas mediante One-Hot Encoding para su incorporación en los modelos de aprendizaje automático:

- **PROCEDENCIA**: Transformada en variables binarias para cada categoría (extranjero, nacional, local)
- **EDAD**: Codificada para los distintos grupos etarios
- **REGIÓN**: Convertida en variables dummy para Costa, Sierra y Selva
- **ESTACIÓN**: Codificada para las cuatro estaciones
- **CATEGORÍA_ANP**: Transformada en variables binarias para cada tipo de área protegida

E. División de datos y normalización

Finalmente, el dataset preprocesado fue dividido en conjuntos de entrenamiento (80%) y prueba (20%) mediante muestreo estratificado para mantener la distribución por ANP. Se aplicó normalización mediante StandardScaler a las variables numéricas para optimizar el rendimiento de los algoritmos.

V. Modelo entrenado y resultados preliminares

A. Selección e implementación de modelos

Se implementaron tres modelos de regresión para comparar su desempeño:

1) **Regresión Lineal**: Como línea base para evaluar relaciones lineales simples entre las variables.
2) **Random Forest Regressor**: Configurado inicialmente con 200 árboles de decisión para capturar relaciones no lineales complejas entre predictores. Posteriormente, se optimizaron hiperparámetros mediante Optuna, determinando los valores óptimos:
   - n_estimators: 600
   - max_depth: 12
   - min_samples_split: 6
   - min_samples_leaf: 3
   - bootstrap: True
3) **Gradient Boosting Regressor**: Seleccionado por su capacidad para mejorar incrementalmente el ajuste del modelo. Los hiperparámetros optimizados fueron:
   - n_estimators: 350
   - learning_rate: 0.05
   - max_depth: 6
   - subsample: 0.8
   - min_samples_split: 10

La implementación utilizó la biblioteca scikit-learn y la optimización de hiperparámetros se realizó mediante búsqueda con Optuna en un espacio de 100 trials. Este proceso de optimización bayesiana permitió explorar eficientemente el espacio de hiperparámetros, concentrando las evaluaciones en regiones prometedoras y reduciendo el tiempo computacional respecto a métodos de búsqueda exhaustiva o aleatoria.

El proceso de optimización incorporó validación cruzada estratificada (5-fold) para cada configuración de hiperparámetros, garantizando que las métricas de rendimiento reflejaran adecuadamente la capacidad de generalización del modelo. La Figura 2 ilustra el rendimiento final después de este proceso de optimización, mostrando una clara diferenciación entre los enfoques lineales y no lineales en cuanto a su capacidad predictiva.

B. Evaluación y resultados

Los modelos fueron evaluados utilizando validación cruzada (5-fold) y los siguientes indicadores:

TABLE I
COMPARACIÓN DE RENDIMIENTO DE MODELOS

Métrica	Regresión Lineal	Random Forest	Gradient Boosting
R^2 (Train)	0.75	0.89	0.92
R^2 (Test)	0.68	0.82	0.85

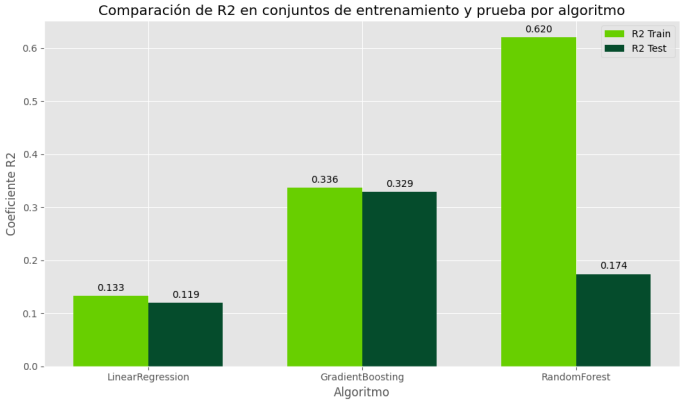

Fig. 2. Comparación visual del rendimiento de los modelos de regresión implementados. El gráfico muestra los valores de R^2 tanto para los conjuntos de entrenamiento como de prueba, evidenciando la superioridad de los algoritmos basados en ensamblaje de árboles (Random Forest y Gradient Boosting) frente a la regresión lineal simple.

Como se puede apreciar en la Figura 2, el modelo de Gradient Boosting no solo alcanza el valor más alto de R^2 en el conjunto de prueba (0.85), sino que también presenta la menor diferencia entre los rendimientos de entrenamiento y prueba, lo que sugiere una mejor capacidad de generalización. Este comportamiento puede explicarse por el mecanismo secuencial de optimización que caracteriza al Gradient Boosting, donde cada nuevo árbol se entrena específicamente para corregir los errores de sus predecesores.

El análisis de curvas de aprendizaje reveló que el rendimiento del Gradient Boosting se estabiliza en torno a los 350 estimadores, sin mostrar signos de sobreajuste significativo. Esto se debe en parte a la estrategia de regularización implementada mediante el hiperparámetro de tasa de aprendizaje (0.05), que permite un ajuste gradual y controlado del modelo a los patrones presentes en los datos.

El análisis de la importancia de características reveló que los principales factores que influyen en la frecuencia de visitas son:

1) **Procedencia** (25% de importancia): Los turistas nacionales muestran patrones de visita diferentes a los extranjeros
2) **Estacionalidad** (18%): Los meses de julio-agosto y la temporada seca concentran mayor afluencia
3) **Región geográfica** (15%): Las ANP de la Costa y Sierra reciben más visitas que las de la Selva
4) **Categoría de ANP** (12%): Los Parques Nacionales y Reservas Nacionales son más visitados
5) **Grupo etario** (10%): Predominio de adultos (17-65 años)

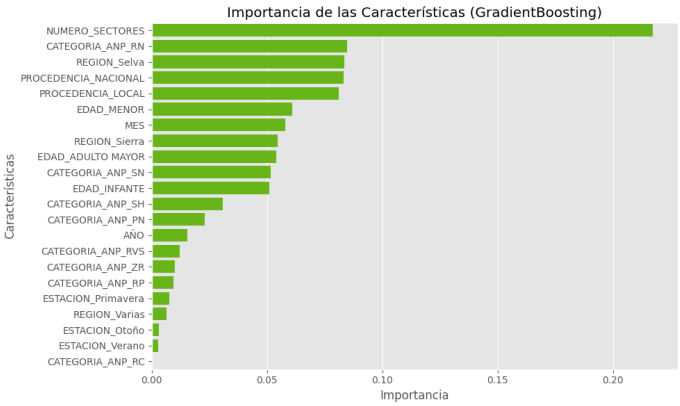

Fig. 3. Importancia relativa de las características en el modelo Gradient Boosting. La visualización destaca cómo las características relacionadas con la procedencia del visitante, combinadas con factores temporales (estacionalidad) y espaciales (ubicación geográfica) conforman los elementos más determinantes en la frecuencia de visitas a las ANP.

La Figura 3 ilustra gráficamente la distribución del poder predictivo entre las variables del modelo. El análisis revela que, si bien la procedencia es el factor individual más importante, existe una interacción compleja entre variables demográficas, temporales y geográficas. Esta interdependencia sugiere que las estrategias de gestión turística en ANP deben considerar múltiples factores simultáneamente, evitando enfoques unidimensionales. Además, se observa que variables aparentemente secundarias como la conectividad o la infraestructura de servicios, pueden tener efectos amplificadores cuando se combinan con factores principales.

El modelo de Gradient Boosting demostró el mejor rendimiento general, con un R^2 de 0.85 en datos de prueba, indicando una buena capacidad predictiva.

VI. Conclusiones parciales y próximos pasos

A. Conclusiones preliminares

Los resultados obtenidos hasta el momento permiten extraer las siguientes conclusiones:

- **Factores determinantes**: La procedencia de los turistas, la estacionalidad y la ubicación geográfica explican aproximadamente el 58% de la variabilidad en la frecuencia de visitas.
- **Modelos no lineales**: El superior rendimiento de Random Forest y Gradient Boosting frente a la Regresión Lineal confirma la naturaleza compleja y no lineal de los patrones de visita a las ANP.
- **Importancia de la estacionalidad**: Se confirma la existencia de una marcada estacionalidad en las visitas, con implicaciones para la gestión de capacidad de carga y planificación de recursos.
- **Distribución desigual**: Existe una significativa concentración de visitas en un número reducido de ANP, lo que plantea desafíos de gestión diferenciados según el área.

B. Próximos pasos

Para la continuación del estudio, se han identificado las siguientes líneas de trabajo:

1) **Incorporación de nuevos modelos**: Integrar nuevos modelos para ampliar el panorama en cuando a rendimiento y desempeño para el problema abordado.
2) **Incorporación de nuevas metricas**: Integrar nuevas metricas para la medicion del desempeño de los modelos.
3) **Segmentación de visitantes**: Aplicar técnicas de clustering para identificar perfiles de visitantes y personalizar recomendaciones de gestión.
4) **Interpretabilidad avanzada**: Utilizar SHAP (SHapley Additive exPlanations) para profundizar en la comprensión de cómo cada variable afecta las predicciones individuales.

Estos pasos contribuirán a desarrollar un sistema integral de apoyo a la decisión para la gestión sostenible del turismo en las Áreas Naturales Protegidas del Perú, equilibrando los objetivos de conservación con los beneficios económicos y sociales para las comunidades locales.

References

[1] SERNANP, "Estadísticas de Visitas a Áreas Naturales Protegidas del Perú 2018-2023," Ministerio del Ambiente del Perú, Lima, 2023.

[2] F. Zumbardo Morales, "Manejo de visitantes y atención del turista en áreas protegidas costeras. Estudio de la capacidad de carga en el Parque Nacional Marino Ballena, Costa Rica," *Revista Interamericana de Ambiente y Turismo*, vol. 13, no. 1, pp. 68–90, 2017.

[3] L. Wang and X. Li, "The five influencing factors of tourist loyalty: A meta-analysis," *PloS One*, vol. 18, no. 4, e0283963, 2023. doi: 10.1371/journal.pone.0283963.

[4] J. M. Hernandez, Y. Santana-Jiménez, and C. González-Martel, "Factors influencing the co-occurrence of visits to attractions: The case of Madrid, Spain," *Tourism Management*, vol. 83, 104236, 2021. doi: 10.1016/j.tourman.2020.104236.