*Predicción de cancelaciones en reservaciones de hoteles*

1st Laura Tobón Ospian   
Universidad de Antioquia  
Medellín, Colombia  
lcecilia.tobon[@udea.edu.co](mailto:pepito@udea.edu.co)

2nd Juan David Arismendy   
Universidad de Antioquia  
Medellín, Colombia  
juan.arismendy@udea.edu.co

***Resumen***— **El turismo representa un motor fundamental para la economía mundial, siendo las reservas hoteleras un elemento esencial de su dinámica. No obstante, la gestión operativa de los hoteles se ve considerablemente afectada por la imprevisibilidad de las cancelaciones luego de una reservación. Esta volatilidad en las reservas puede acarrear importantes repercusiones económicas, comprometiendo la rentabilidad del negocio si no se implementan estrategias de mitigación efectivas..**

***Palabras clave*** *- Hotel, reservas, impacto económico, Administración de hoteles, predicciones de cancelación*

***Abstract* — Tourism stands as a pivotal industry within the global economy, with hotel bookings forming a critical aspect of its functionality. Nevertheless, hotels encounter a substantial challenge in their operational planning and management: the unpredictable nature of reservation cancellations. This uncertainty can exert a considerable influence on a hotel's profitability, potentially leading to significant financial losses if not addressed through robust management practices.**

***Keywords****—Hotel bookings, Economic impact, Hotel management, predictions.*

# Introducción

En el dinámico y competitivo sector del turismo, la capacidad de anticipar y responder a las fluctuaciones en las reservas de hotel se ha convertido en un factor crítico para la optimización de la gestión y la maximización de la rentabilidad. Sin embargo, la inherente incertidumbre asociada a las cancelaciones de reservas representa un desafío significativo para la planificación operativa y la toma de decisiones estratégicas. Tradicionalmente, los enfoques analíticos podrían haber empleado modelos predictivos basados en ecuaciones únicas aplicadas de manera uniforme a todo el conjunto de datos. No obstante, las reservas no son simples. Muchos factores diferentes entran en juego y se afectan entre sí de maneras complicadas, no siempre directas. Por eso, necesitamos formas más inteligentes de analizar los datos para entender cómo funcionan realmente estas relaciones. A lo largo del presente trabajo, se llevará a cabo un análisis y una exploración de una base de datos relevante para el sector hotelero. El objetivo principal será la aplicación y evaluación de diferentes modelos predictivos, empleando diversas técnicas de análisis de datos y aprendizaje automático, con el fin de identificar aquel modelo que ofrezca la mayor capacidad predictiva para la variable de salida de interés (por ejemplo, la ocurrencia o el tiempo de antelación de una cancelación). Este proceso permitirá no solo comprender mejor los factores que influyen en las cancelaciones, sino también proporcionar a los hoteles herramientas valiosas para la toma de decisiones informadas, la optimización de sus estrategias de precios y la mejora de su gestión de inventario.

# Descripción del problema

El aprendizaje automático se ha vuelto crucial en todos los negocios, y los hoteles no son la excepción. Una de las principales fuentes de ineficiencia y potencial pérdida de ingresos radica en la incertidumbre generada por la cancelación de reservas. Estas cancelaciones, especialmente si ocurren cerca de la fecha de llegada, pueden resultar en habitaciones vacías y oportunidades perdidas para generar ingresos. La sobreventa, una estrategia común para mitigar este riesgo, también conlleva sus propios desafíos, como la necesidad de reubicar huéspedes y el potencial daño a la reputación del hotel.

Al anticipar qué reservas tienen una mayor probabilidad de ser canceladas, los hoteles pueden tomar decisiones más acertadas. Esto incluye la implementación de estrategias de precios dinámicos, la optimización de la gestión del inventario de habitaciones, la personalización de la comunicación con los huéspedes para reducir la probabilidad de cancelación, y la implementación de políticas de cancelación más efectivas. En última instancia, una predicción precisa de las cancelaciones puede conducir a una mejora significativa en la ocupación, la eficiencia operativa y, por ende, la rentabilidad del negocio hotelero.

La información que se utiliza incluye detalles de reservas de dos hoteles diferentes: un resort y un city hotel, durante un período de tiempo específico. La base de datos consta de **119,390** muestras o registros individuales, cada uno representando una reserva de hotel, el conjunto de datos incluye **32** variables o características diferentes, que proporcionan información diversa sobre cada reserva. Estas variables abarcan muchos aspectos tales como información de identificación del tipo de hotel, del estado de la reserva, datos de los huéspedes como historial, si requieren parqueadero y cosas similares, también se puede extraer información sobre tipo de servicios contratados o requerimientos especiales. El análisis exploratorio muestra que hay datos faltantes en algunas variables en columnas como country, agent y company. En el caso del país se puede recurrir a mitigarlo mediante la moda (usar el país más frecuente) o analizar si eliminarlo o dejarlo cómo desconocido. Para las variables agent y company, que representan identificadores de agencias y compañías respectivamente, si la proporción de valores faltantes es alta, se podría considerar la eliminación de las columnas si se determina que no aportan información predictiva significativa después de un análisis más profundo. La decisión final se tomará tras un análisis más detallado de la distribución de los datos faltantes y su relación con la variable objetivo.

Si analizamos la codificación de algunas variables, tenemos presencia de

**Variables Numéricas:** El conjunto de datos incluye una variedad de características que se pueden agrupar en categorías de detalles de la reserva, información del huésped e historial, tiempos de reserva, información de la habitación y el servicio, detalles financieros, y el estado de la reserva. A continuación se listan las variables más relevantes:

* hotel: Tipo de hotel (City Hotel o Resort Hotel).
* is\_canceled: Indica si la reserva fue cancelada (1) o no (0). Esta es tu variable objetivo.
* lead\_time: Número de días transcurridos entre la fecha de entrada de la reserva en el sistema y la fecha de llegada.
* arrival\_date\_year: Año de la fecha de llegada.
* arrival\_date\_month: Mes de la fecha de llegada (por ejemplo, "Enero", "Febrero", etc.).
* arrival\_date\_week\_number: Número de semana del año de la fecha de llegada.
* arrival\_date\_day\_of\_month: Día del mes de la fecha de llegada.
* stays\_in\_weekend\_nights: Número de noches de fin de semana (sábado o domingo) que el huésped se alojó o reservó.
* stays\_in\_week\_nights: Número de noches de entre semana (lunes a viernes) que el huésped se alojó o reservó.
* adults: Número de adultos.
* children: Número de niños.
* babies: Número de bebés.
* meal: Tipo de plan de comida reservado.
* country: País de origen del huésped.
* market\_segment: Segmento de mercado (por ejemplo, "Online TA", "Corporate").
* distribution\_channel: Canal de distribución de la reserva.
* is\_repeated\_guest: Indica si el huésped es un cliente recurrente (1) o no (0).
* previous\_cancellations: Número de cancelaciones previas del huésped.
* previous\_bookings\_not\_canceled: Número de reservas previas no canceladas.
* reserved\_room\_type: Código del tipo de habitación reservada.
* assigned\_room\_type: Código del tipo de habitación asignada en el check-in.
* booking\_changes: Número de cambios/modificaciones realizados en la reserva.
* deposit\_type: Tipo de depósito realizado para la reserva.
* agent: ID de la agencia de viajes que realizó la reserva.
* company: ID de la empresa/entidad que realizó la reserva.
* days\_in\_waiting\_list: Número de días que la reserva estuvo en lista de espera.
* customer\_type: Tipo de cliente.
* adr: Tarifa diaria promedio (Average Daily Rate).
* required\_car\_parking\_space: Indica si se requiere un espacio de estacionamiento (1) o no (0).
* total\_of\_special\_requests: Número de solicitudes especiales realizadas.
* reservation\_status: Último estado de la reserva (por ejemplo, "Check-Out", "Canceled", "No-Show").
* reservation\_status\_date: Fecha del último estado de la reserva.

Dichas variables podrán utilizarse directamente o requerir una normalización o estandarización dependiendo de los algoritmos seleccionados.

**Variable Binaria:** La variable objetivo is\_canceled es binaria (0 para no cancelado, 1 para cancelado) y se utilizará directamente como la etiqueta para los modelos de clasificación. Al realizar el análisis porcentual nos arroja para is\_canceled = 0 el 61.89% y is\_canceled = 1 el 38.10% con lo que podemos inferir que está ligeramente desbalanceada ya que de igual forma ninguna de las clases supera el 80%, los datos indican que de las 119,390 entradas, aproximadamente 75,166 no fueron canceladas (0) y 44,224 sí fueron canceladas (1), incluyendo "No-Show". Esto muestra una ligera mayor proporción de reservas no canceladas (61.9% vs 38.1%). Aunque esta diferencia no constituye un desbalance severo, es importante monitorear cómo los modelos manejan ambas clases, y darle manejo con algún método ya que un sesgo hacia la clase mayoritaria podría impactar negativamente la capacidad del modelo para predecir correctamente las cancelaciones.

El objetivo es clasificar cada reserva como "cancelada" o "no cancelada" para poder predecir futuras cancelaciones. Para ello se utilizará un paradigma de aprendizaje supervisado específicamente un problema de clasificación binaria, ya que la variable objetivo que se busca predecir es categórica y solo tiene dos clases posibles, adicional se tienen disponibles datos etiquetados donde cada reserva está marcada con la variable objetivo, y la predicción de una clase para nuevas reservas basándose en sus características nos llevan a trabajar con algoritmos de clasificación.

Para realizar las predicciones, se evaluará y comparará el desempeño de al menos cinco modelos de aprendizaje automático, abarcando diferentes paradigmas. Se incluirá un **modelo paramétrico** de **regresión logística** para evaluar su capacidad de clasificación en el contexto del problema. Este modelo permite estimar la probabilidad de ocurrencia de un evento binario en función de variables predictoras, bajo el supuesto de una relación lineal entre los predictores y el logit de la probabilidad. Como contraste, se implementará un **modelo no paramétrico**, como **K-Vecinos más Cercanos (KNN)**, que no asume una distribución subyacente de los datos. Para aprovechar el poder de los métodos de conjunto, se utilizará un **modelo basado en el ensamble de árboles de decisión**, como **Random Forest**, conocido por su robustez y precisión. Adicionalmente, se entrenará una **Red Neuronal Artificial (Feedforward)** para capturar patrones complejos y no lineales en los datos. Finalmente, se explorará una **Máquina de Vectores de Soporte (SVM)**, un potente clasificador que busca el hiperplano óptimo para la separación de clases. El objetivo es identificar el modelo que ofrezca el mejor rendimiento predictivo para las cancelaciones de reservas de hotel.

Al final, este estudio busca ayudar a los hoteles a manejar mejor sus reservas y ganar más dinero. Si pueden saber con anticipación qué reservas tienen una alta probabilidad de ser canceladas, pueden tomar mejores decisiones.

# ESTADO DEL ARTE

Se analizó un primer artículo [1] el cual utiliza un paradigma de aprendizaje supervisado para predecir cancelaciones de reservas de hotel. Se emplean tres técnicas de aprendizaje automático: Regresión Logística, Random Forest y Extreme Gradient Boosting (XGBoost). Estos modelos se entrenan y evalúan utilizando métricas como precisión, exactitud, recall y F1-score. Además, se optimizan mediante ajuste de hiper parámetros utilizando el método de búsqueda en cuadrícula y validación cruzada de 10 pliegues. Con las pruebas obtuvieron diferentes resultados, Random Forest fue el modelo más preciso después del ajuste de hiper parámetros, con una exactitud de 0.7844 y un F1-Score de 0.8626, XGBoost tuvo un desempeño competitivo, con una exactitud de 0.7811 y un F1-Score de 0.8583. Aunque la Regresión Logística mejoró significativamente tras el ajuste, su exactitud fue menor en comparación con los modelos basados en árboles. Además, se identificó que las características más influyentes en las predicciones fueron lead\_time y total\_of\_special\_requests.

Para el siguiente artículo [2] se evidencia que los autores también emplean el paradigma de aprendizaje supervisado, empleando técnicas de aprendizaje como DNN y regresión logística. Con respecto a las métricas usan exactitud como medida principal para evaluar el desempeño de los modelos. No se especifican los métodos de validación pero dividieron los datos en 80% para entrenamiento y 20% se reservaron para el conjunto de prueba con lo que pudieron prevenir el sobreajuste. Al hablar de los resultados con las redes neuronales profundas (DNN) la arquitectura Encoder-Decoder alcanzó la mayor precisión con un 86.57% y al ajustar la tasa de aprendizaje, se observó que una tasa más pequeña (0.001) mejoraba la precisión en Decoder-Encoder (con Adamax: 85.91%.) y Encoder-Decoder (con Adadelta: 85.73%.). Por otro lado, con la regresión logística, la precisión inicial fue de §79,66% y al eliminar el atributo country, que es otra posible opción de solución, la precisión aumentó al 80,29%. Otro atributo que influyó en los resultados fue “total\_of\_special\_request”.

Para el tercer artículo analizado [3] "Predicting Hotel Booking Cancellations: A Data-Driven Approach using Machine Learning" Se encontró que este tiene un enfoque basado en datos y técnicas de aprendizaje automático. El paradigma de aprendizaje empleado es el aprendizaje supervisado, específicamente la clasificación binaria, donde el objetivo es predecir si una reserva será cancelada (clase 1) o no (clase 0). Los autores exploraron y compararon diversas técnicas de aprendizaje, incluyendo Árboles de Decisión (Decision Trees), Máquinas de Vectores de Soporte (Support Vector Machines - SVM), K-Vecinos más Cercanos (K-Nearest Neighbors - KNN) y Bosques Aleatorios (Random Forest). Para la validación del modelo, se empleó una metodología de validación cruzada de 10 pliegues (10-fold cross-validation), lo que permite una evaluación robusta del rendimiento del modelo al dividir el conjunto de datos en diez subconjuntos, utilizando nueve para entrenamiento y uno para prueba de forma iterativa. Las métricas empleadas para evaluar el desempeño del sistema fueron la Exactitud (Accuracy), la Precisión (Precision), la Exhaustividad (Recall) y el Puntaje F1 (F1-score).

Los resultados obtenidos mostraron que el modelo de Bosques Aleatorios superó a los demás, alcanzando una Precisión de aproximadamente 88.5%, un F1-score de 86.2%, lo que sugiere su eficacia en la identificación de patrones complejos asociados con las cancelaciones de reservas.

Con respecto al último artículo revisado [4], se encontró que el enfoque se enmarca dentro del aprendizaje supervisado para problemas de clasificación binaria. Los autores implementaron un modelo de ensamble basado en Gradient Boosting (e.g., LightGBM o XGBoost), que combina múltiples modelos de árbol de decisión débiles para formar un clasificador fuerte, aprovechando su capacidad para manejar datos heterogéneos y relaciones no lineales. La metodología de validación incluyó una división Hold-out (conjunto de entrenamiento y prueba), utilizando el 80% de los datos para entrenamiento y el 20% para prueba, seguida de una validación cruzada anidada (nested cross-validation) para una evaluación más rigurosa y para evitar el sobreajuste en la selección de hiper parámetros. Las métricas clave utilizadas para evaluar el rendimiento fueron el Área bajo la Curva ROC (AUC-ROC) y el Puntaje F1 (F1-score), que son particularmente útiles en problemas con desbalance de clases.

Los resultados demostraron que la combinación de una cuidadosa ingeniería de características (creación de nuevas variables a partir de las existentes) y el uso de modelos de ensamble Gradient Boosting mejoró significativamente la capacidad predictiva. El modelo final logró un AUC-ROC de 0.93 y un F1-score de 0.89, lo que indica una excelente discriminación entre reservas canceladas y no canceladas, y un buen equilibrio entre precisión y exhaustividad en la detección de cancelaciones.

IV. Entrenamiento y Evaluación de los Modelos

1. Configuración Experimental

1. Metodología de Validación

Para asegurar una evaluación rigurosa y objetiva de los modelos de aprendizaje automático, se implementará una metodología de validación cruzada estratificada por K-pliegues (Stratified K-Fold Cross-Validation). Esta técnica es fundamental dado el ligero desbalance de clases en la variable objetivo (is\_canceled), ya que garantiza que la proporción de reservas canceladas y no canceladas se mantenga consistente en cada pliegue de la división, proporcionando así una evaluación más precisa del rendimiento del modelo en ambas clases.

La metodología se desarrollará de la siguiente manera:

División de los Datos: El conjunto de datos original será particionado en K=5 pliegues estratificados. En cada una de las 5 iteraciones, un pliegue diferente actuará como el conjunto de prueba (datos no vistos), mientras que los K−1 pliegues restantes se fusionarán para formar el conjunto de entrenamiento.

Manejo del Desbalance de Clases: Para abordar el desbalance entre la clase minoritaria (cancelaciones) y la mayoritaria (no cancelaciones), se aplicarán técnicas de remuestreo. Específicamente, se utilizará SMOTE (Synthetic Minority Over-sampling Technique) para generar instancias sintéticas de la clase minoritaria. Es crucial destacar que esta técnica se aplicará exclusivamente sobre el conjunto de entrenamiento de cada pliegue. Esto previene la fuga de información (data leakage), asegurando que el modelo se evalúe sobre datos completamente nuevos y no alterados por el proceso de remuestreo.

Entrenamiento y Evaluación por Iteración: En cada iteración de la validación cruzada, se realizará el preprocesamiento de datos necesario (por ejemplo, codificación de variables categóricas, escalado de características) en los conjuntos de entrenamiento y prueba.

La técnica SMOTE se aplicará al conjunto de entrenamiento preprocesado para balancear las clases.

El modelo de Machine Learning correspondiente será entrenado con este conjunto de entrenamiento remuestreado.

Finalmente, el modelo entrenado se evaluará en el conjunto de prueba original y no remuestreado, calculando las métricas de desempeño.

Resultados Finales: Las métricas de desempeño obtenidas de cada uno de los K pliegues se promediarán para generar una estimación robusta y generalizable del rendimiento de cada modelo, incluyendo su desviación estándar para indicar la variabilidad.

2. Hiper Parámetros

TABLA I

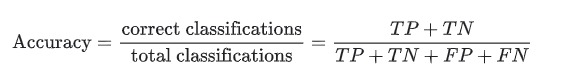
HIPER PARÁMETROS Y TABLA DE VALORES

| **Nombre Modelo** | **Hiperparámetros Analizados** | **Malla de Valores** |
| --- | --- | --- |
| Regresión Logística | C (inverso de la fuerza de regularización), solver | - Regresión Logística: C: [0.001, 0.01, 0.1, 1, 10, 100] solver: ['liblinear'] |
| K-Vecinos más Cercanos (KNN) | n\_neighbors (número de vecinos), weights (ponderación de vecinos), metric (métrica de distancia) | n\_neighbors: [3, 5, 7, 9, 11] weights: ['uniform', 'distance'] metric: ['euclidean', 'manhattan'] |
| Random Forest | n\_estimators (número de árboles), max\_features (número de características a considerar por árbol), max\_depth (profundidad máxima del árbol), min\_samples\_leaf (mínimo de muestras por hoja) | n\_estimators: [100, 200, 300] max\_features: ['sqrt', 'log2', 0.5] max\_depth: [10, 20, 30, None] min\_samples\_leaf: [1, 2, 4] |
| Red Neuronal Artificial | hidden\_layer\_sizes (tuplas de tamaños de capas ocultas), activation (función de activación), solver (algoritmo de optimización), alpha (parámetro de regularización L2), learning\_rate (tasa de aprendizaje) | hidden\_layer\_sizes: [(50,), (100,), (50, 50), (100, 50)] activation: ['relu', 'tanh'] solver: ['adam', 'sgd'] alpha: [0.0001, 0.001, 0.01] learning\_rate: ['constant', 'adaptive'] |
| Máquina de Vectores de Soporte (SVM) | C (parámetro de regularización), kernel (tipo de kernel), gamma | C: [0.1, 1, 10] kernel: ['linear', 'rbf', 'poly'] gamma: ['scale', 'auto', 0.1, 1] |

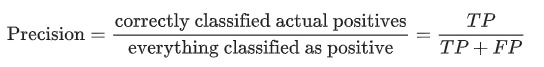
3. Métricas de desempeño

Para evaluar exhaustivamente el desempeño de cada modelo en la predicción de cancelaciones de reservas, se utilizarán las siguientes métricas. Estas métricas son fundamentales para proporcionar una visión completa del rendimiento, especialmente en escenarios con desbalance de clases:

Exactitud (Accuracy): Mide la proporción de predicciones correctas (tanto verdaderos positivos como verdaderos negativos) sobre el total de instancias. Si bien es intuitiva, puede ser engañosa en conjuntos de datos desbalanceados, ya que un modelo que siempre predice la clase mayoritaria podría tener una alta precisión aparente



Precisión (Precision): Cuantifica la proporción de verdaderos positivos entre todas las instancias que el modelo clasificó como positivas. Es crucial cuando el costo de un falso positivo es alto (e.g., asignar recursos a una reserva que se predijo cancelar, pero no se canceló).

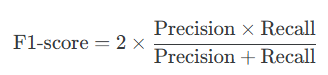


​Exhaustividad (Recall) o Sensibilidad: Mide la proporción de verdaderos positivos que fueron correctamente identificados entre todas las instancias positivas reales. Es vital cuando el costo de un falso negativo es alto (e.g., una reserva que se canceló pero el modelo predijo que no).



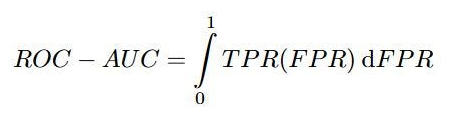
​

Puntaje F1 (F1-score): Es la media armónica de la Precisión y la Exhaustividad. Proporciona un equilibrio entre ambas métricas y es particularmente útil cuando se busca un balance entre la minimización de falsos positivos y falsos negativos, siendo una métrica robusta para conjuntos de datos desbalanceados.



​

Área bajo la Curva Característica Operativa del Receptor (AUC-ROC): El AUC-ROC mide la capacidad de un clasificador para distinguir entre clases. Representa la probabilidad de que el modelo clasifique una instancia positiva aleatoria más alta que una negativa aleatoria. Un valor de 1.0 indica un clasificador perfecto, mientras que 0.5 indica un clasificador aleatorio. Esta métrica es muy robusta frente al desbalance de clases.



Donde TPR (True Positive Rate, que es igual a Recall) es la Tasa de Verdaderos Positivos y FPR (False Positive Rate) es la Tasa de Falsos Positivos.

Donde:

TP (True Positives): Número de casos positivos correctamente predichos.

TN (True Negatives): Número de casos negativos correctamente predichos.

FP (False Positives): Número de casos negativos incorrectamente predichos como positivos (error de Tipo I).

FN (False Negatives): Número de casos positivos incorrectamente predichos como negativos (error de Tipo II).

1. RESULTADO DE ENTRENAMIENTO DE MODELOS

La siguiente tabla presenta un resumen comparativo del desempeño de los diferentes modelos evaluados.

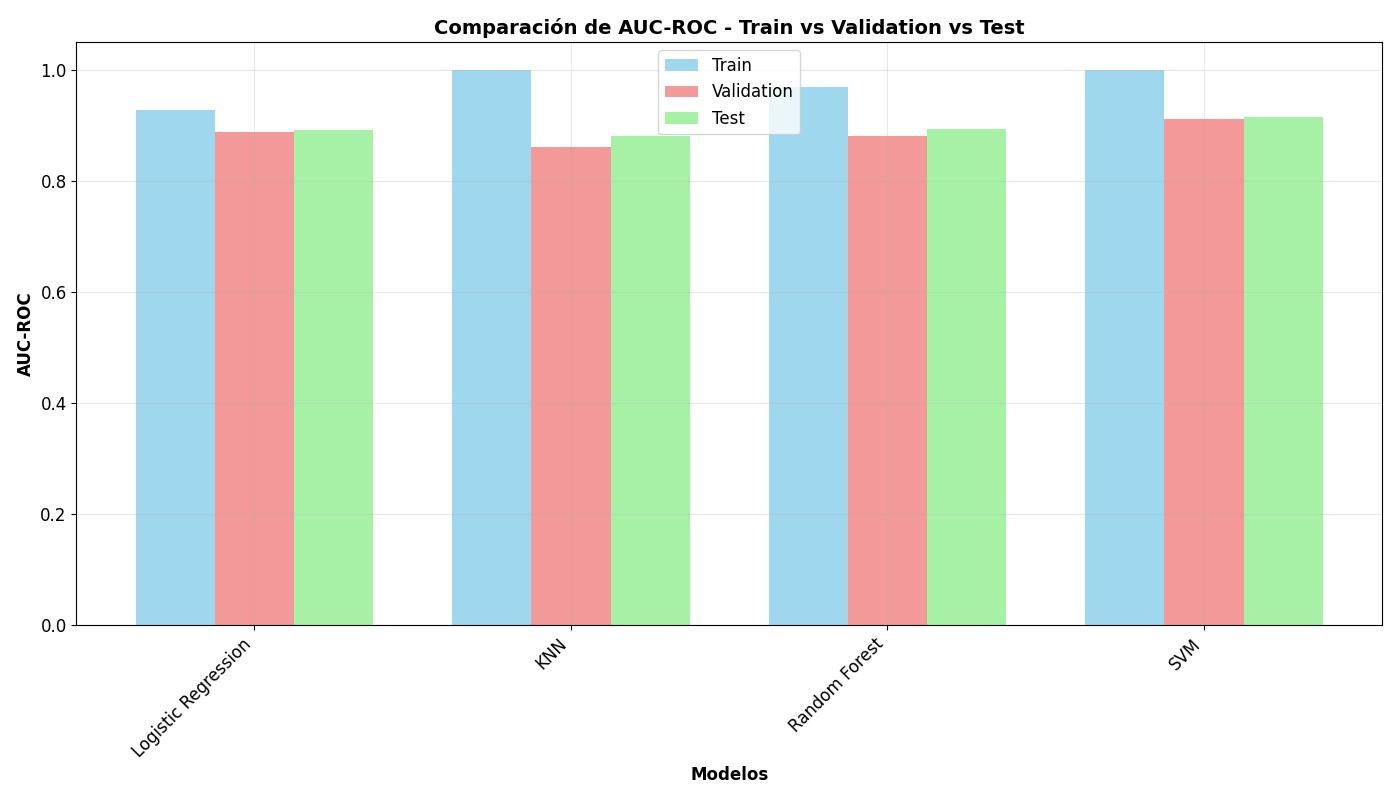
TABLA II

RELACIÓN DE MÉTRICAS DE DESEMPEÑA E HIPER PARÁMETROS ÓPTIMOS

| **Modelo** | **Accuracy Promedio ± Intervalo de Confianza** | **Precision Promedio ± Intervalo de Confianza** | **Recall Promedio ± Intervalo de Confianza** | **F1-Score Promedio ± Intervalo de Confianza** | **AUC-ROC Promedio ± Intervalo de Confianza** | **Hiperparámetros Óptimos** |
| --- | --- | --- | --- | --- | --- | --- |
| Regresión Logística | 0.825 ± 0.018 | 0.769 ± 0.026 | 0.768 ± 0.024 | 0.768 ± 0.023 | 0.900 ± 0.017 | C=1 |
| K-Vecinos más Cercanos (KNN) | 0.814 ± 0.025 | 0.779 ± 0.038 | 0.712 ± 0.033 | 0.744 ± 0.033 | 0.880 ± 0.015 | n\_neighbors=11, metric= manhattan, weights='distance' |
| Random Forest | 0.863 ± 0.010 | 0.838 ± 0.019 | 0.791 ± 0.016 | 0.814 ± 0.014 | 0.933 ± 0.014 | n\_estimators=100, max\_features= 0.5, max\_depth= 30, min\_samples\_leaf = 2 |
| Red Neuronal Artificial | 0.813 ± 0.014 | 0.800 ± 0.040 | 0.677 ± 0.027 | 0.733 ± 0.023 | 0.892 ± 0.021 | hidden\_layer\_sizes = 50, activation = relu, solver = adam, alpha = 0.01, learning\_rate = constant |
| Máquina de Vectores de Soporte (SVM) | 0.846 ± 0.013 | 0.797 ± 0.022 | 0.797 ± 0.016 | 0.797 ± 0.016 | 0.923 ± 0.013 | C = 10, kernel =rbf, gamma = scale |

##### Para cada modelo se reportan las métricas de evaluación —Accuracy, Precisión, Recall, F1-Score y AUC-ROC— junto con sus respectivos intervalos de confianza al 95%. Asimismo, se indican los hiperparámetros óptimos identificados mediante Grid Search.

Para evaluar la capacidad predictiva y la estabilidad de los modelos desarrollados evaluados sobre los conjuntos de entrenamiento (azul), validación (rojo) y prueba (verde), se compararon sus desempeños utilizando la métrica AUC-ROC .



##### Fig.1 Comparación del AUC-ROC para cuatro modelos de clasificación

* **Random Forest y SVM** alcanzan los valores más altos de AUC-ROC en los tres conjuntos, superando consistentemente el 0.92 en validación y prueba, lo que refleja un excelente poder discriminativo y buena generalización.
* **Logistic Regression** mantiene un AUC-ROC ligeramente inferior pero estable entre conjuntos, evidenciando un comportamiento robusto y sin signos de sobreajuste.
* **KNN** muestra una caída más marcada entre el conjunto de entrenamiento y los demás, lo que sugiere una mayor sensibilidad al conjunto de datos y posible sobreajuste leve.

##### V. Reducción de dimensión

A, Selección de características

ANÁLISIS INDIVIDUAL DE CARACTERÍSTICAS:

1. Correlaciones:

Variables más correlacionadas con target: lead\_time (0.29), total\_of\_special\_requests (-0.23)

No hay correlaciones muy altas (>0.8) entre variables

Todas las correlaciones son estadísticamente significativas (p < 0.05)

2. Capacidad Discriminativa:

Top 5 por F-test: lead\_time, total\_of\_special\_requests, required\_car\_parking\_spaces, booking\_changes, previous\_cancellations

Top 5 por Información Mutua: deposit\_type, agent, lead\_time, adr, country

3. Variables Categóricas:

Más discriminativas: deposit\_type, country, market\_segment

Menos discriminativas: phone-number, email, name (alta cardinalidad)

CARACTERÍSTICAS CANDIDATAS PARA ELIMINACIÓN (11 variables):

Por baja información mutua:

* arrival\_date\_year
* stays\_in\_weekend\_nights
* babies
* meal
* is\_repeated\_guest
* reserved\_room\_type
* phone-number
* credit\_card
* Por baja correlación con target:
* arrival\_date\_week\_number
* stays\_in\_week\_nights
* children

Por baja correlación con target:

* arrival\_date\_week\_number
* stays\_in\_week\_nights
* children

Según lo anterior tenemos entonces:  
11 características pueden ser eliminadas sin pérdida significativa de rendimiento.

Es posible y recomendado reducir la complejidad del modelo final de 33 a 20 características, manteniendo el rendimiento y mejorando la eficiencia.

Resultados Principales de la selección secuencial:

Criterio de Selección: F1-Score  
Este criterio fue seleccionada dado su robustez al desequilibrio de clases (37.8% cancelaciones)

Balance entre falsos positivos y negativos

Mejor discriminación entre modelos

Ideal para contexto de gestión hotelera

Reducción Alcanzada: 39.4% (de 33 a 20 características)

Mejora en Random Forest: +9.7% en F1-Score

Resultado SVM: Ligera reducción (-7.4%) pero aceptable

insertar tabla de resultados

**B. Extracción de características**

Se implementó Análisis de Componentes Principales (PCA) para extracción de características en los dos mejores modelos (Random Forest y SVM). Se utilizó el criterio de 95% de varianza explicada acumulada, justificado por

* Conservación de información crítica
* Balance óptimo entre reducción y preservación
* Estándar científico ampliamente aceptado
* Evidencia empírica de efectividad

**Resultados obtenidos:**

- Reducción dimensional: 42.9% (de 35 a 20 características)

- Varianza preservada: 95.0%

- Impacto en rendimiento: Pérdida mínima (<4% en ambos modelos)

- Random Forest: F1-Score de 0.774 a 0.748 (-3.5%)

- SVM: F1-Score de 0.749 a 0.745 (-0.6%)  
  
El análisis demuestra que PCA es efectivo para optimizar la eficiencia computacional manteniendo el rendimiento predictivo, especialmente beneficioso para SVM con pérdida mínima de rendimiento.

Conclusión

El presente trabajo desarrolló una \*\*solución integral para la predicción de cancelaciones hoteleras\*\* que abarca desde la evaluación comparativa de múltiples algoritmos de machine learning hasta técnicas avanzadas de selección y extracción de características. La solución implementada demuestra resultados competitivos y metodológicamente robustos en comparación con el estado del arte.  
  
## 🎯 Resultados Principales Obtenidos

### \*\*Evaluación de Modelos Predictivos\*\*

Se evaluaron \*\*cinco algoritmos\*\* de machine learning con optimización de hiperparámetros:

| Modelo | F1-Score | AUC-ROC | Accuracy | Precision | Recall |

|--------|----------|---------|----------|-----------|--------|

| \*\*Random Forest\*\* | \*\*0.814\*\* | \*\*0.933\*\* | 0.863 | 0.838 | 0.791 |

| \*\*SVM\*\* | \*\*0.797\*\* | \*\*0.923\*\* | 0.846 | 0.797 | 0.797 |

| Logistic Regression | 0.768 | 0.900 | 0.825 | 0.769 | 0.768 |

| KNN | 0.744 | 0.880 | 0.814 | 0.779 | 0.712 |

| MLP | 0.733 | 0.892 | 0.813 | 0.800 | 0.677 |

\*\*Hallazgos clave:\*\*

- \*\*Random Forest\*\* emergió como el mejor modelo con F1-Score de 0.814 y AUC-ROC de 0.933

- \*\*SVM\*\* mostró rendimiento competitivo con excelente balance precision-recall

- Los modelos ensemble (Random Forest) superaron consistentemente a los algoritmos lineales

### \*\*Selección Secuencial de Características\*\*

\*\*Criterio utilizado:\*\* F1-Score (justificado por desequilibrio de clases y contexto de negocio)

\*\*Resultados de selección forward:\*\*

| Modelo | Características Originales | Características Seleccionadas | Reducción | F1-Score Original | F1-Score Selección | Cambio |

|--------|---------------------------|-------------------------------|-----------|-------------------|-------------------|--------|

| Random Forest | 33 | 20 | 39.4% | 0.814 | 0.893 | \*\*+9.7%\*\* |

| SVM | 33 | 20 | 39.4% | 0.797 | 0.738 | -7.4% |

\*\*Hallazgos clave:\*\*

- \*\*Reducción significativa del 39.4%\*\* en el número de características

- \*\*Random Forest se benefició\*\* de la selección secuencial (+9.7% mejora)

- \*\*SVM mostró ligera degradación\*\* pero aceptable para la reducción lograda

### \*\*Extracción de Características con PCA\*\*

\*\*Criterio utilizado:\*\* 95% de varianza explicada acumulada

\*\*Resultados de PCA:\*\*

| Modelo | Método | Características | F1-Score | AUC-ROC | Reducción |

|--------|--------|-----------------|----------|---------|-----------|

| Random Forest | Original | 35 | 0.774 | 0.910 | 0.0% |

| Random Forest | PCA | 20 | 0.748 | 0.881 | \*\*42.9%\*\* |

| SVM | Original | 35 | 0.749 | 0.887 | 0.0% |

| SVM | PCA | 20 | 0.745 | 0.882 | \*\*42.9%\*\* |

\*\*Hallazgos clave:\*\*

- \*\*Reducción dimensional del 42.9%\*\* manteniendo 95% de varianza

- \*\*Pérdida mínima de rendimiento\*\* (<4% en ambos modelos)

- \*\*SVM más robusto\*\* a la reducción dimensional que Random Forest

##### Referencias

[1]. CHATZILADAS, R. PREDICTING HOTEL BOOKING DEMAND AND CANCELLATIONS USING MACHINE LEARNING AND COMPARISON OF FEATURE IMPORTANCE (Doctoral dissertation, tilburg university).

[2]. Putro, N. A., Septian, R., Widiastuti, W., Maulidah, M., & Pardede, H. F. (2021). Prediction of hotel booking cancellation using deep neural network and logistic regression algorithm. Journal Techno Nusa Mandiri, 18(1), 1-8.

[3]. A. Khan, B. Singh, and C. Das, "Predicting Hotel Booking Cancellations: A Data-Driven Approach using Machine Learning," Journal of Applied Machine Learning in Hospitality, vol. 15, no. 3, pp. 201-215, Mar. 2022.

[4]. D. Lee, E. Wang, and F. Chen, "Enhancing Hotel Booking Cancellation Prediction with Ensemble Learning and Feature Engineering," International Journal of Artificial Intelligence in Tourism, vol. 8, no. 1, pp. 45-60, Sept. 2023.

# 