Trabajo Fin de Grado
Universidad Miguel Hernández

APRENDIZAJE AUTOMÁTICO PARA PREDECIR CANCELACIONES HOTELERAS

Facultad de Ciencias Sociales y Jurídicas de Elche
Grado en Estadística Empresarial
Curso Académico: 2021-2022

Alumna: Inés Teresa Hernández Pastor
Tutora: María Asunción Martínez Mayoral

Índice de contenidos

Resumen

Con la finalidad de orientar a empresas hoteleras en la toma de decisiones, llevamos a cabo un análisis predictivo con técnicas automáticas de clasificación. Se construyen cinco modelos basados en árboles de clasificación para identificar distintos perfiles de clientes respecto a la cancelación de reservas hoteleras *online*, y también para descubrir qué variables están relacionadas con dicha cancelación. Se comentan las ventajas y desventajas de cada una de las modelizaciones, proporcionando criterios de selección y medición de la bondad del ajuste.

Palabras clave

Cancelación hotelera, segmentación, árboles de clasificación, identificación de perfiles.

Introducción

El proceso de globalización ha provocado una serie de cambios en la sociedad, de entre los que destaca la desvinculación del individuo a su comunidad local. Esto ha motivado un aumento del flujo internacional, llevando al turismo a convertirse en uno de los sectores estratégicos a nivel mundial. En Portugal, *"el turismo"* se posiciona como una de las principales actividades económicas que contribuye al Producto Interior Bruto (PIB), que en 2021 representó alrededor del 15% y, dentro de esta, cabe destacar la hostelería. Entendemos por hostelería la "actividad económica consistente en la prestación de servicios ligados al alojamiento y/o la alimentación durante un periodo determinado de tiempo" (Marrero Hernández 2016).

Asimismo, en los últimos años la digitalización empresarial en el ámbito de la hostelería ha generado avances significativos en lo que tiene que ver con la facilitación de la gestión de reservas y la generación de información concerniente. No obstante, dichas facilidades para realizar reservas han venido acompañadas de las mismas facilidades para gestionar cancelaciones, no siempre con la antelación debida para no repercutir negativamente en costes importantes.

Disponer de mucha información, relativa a todas las gestiones de reserva hoteleras realizadas a lo largo de un periodo amplio de tiempo, puede ser aprovechado, con técnicas analíticas basadas en inteligencia artificial, para identificar el tipo de clientes que habiendo realizado reservas, son más proclives a una cancelación más o menos tardía, y también con ello predecir y ajustar los riesgos de cancelación para minorar los costes asociados a cancelaciones.

Con todo, la información disponible no suele estar depurada y sí contener numerosos errores que es preciso tratar antes de llevar a cabo cualquier tipo de análisis estadístico fiable. Las técnicas de tratamiento de información basadas en inteligencia artificial, especialmente cuando son aplicadas sin fundamentación estadística, pueden producir análisis sesgados debido a la falta de tratamiento y depuración previa de la información, así como a una elección poco fundamentada de las técnicas más eficaces para realizar el análisis en función del tipo de información disponible y los objetivos planteados.

Partimos, para desarrollar este trabajo, de un banco de datos considerablemente grande (119.390 registros sobre 31 variables), relativo a reservas hoteleras entre el 1 de julio de 2015 y el 31 de agosto de 2017 en dos hoteles (uno de ciudad y otro de tipo *resort*), en Portugal. Estos datos ya han sido trabajados en Antonio et al.(2019), y están publicados en la plataforma de competición *Kaggle*, (https://www.kaggle.com/datasets/datacertlaboratoria/proyecto-4-anlisis-de-cancelaciones-hoteleras). En la publicación de Antonio et al. (2019) se presenta un mero descriptivo numérico y gráfico de los datos.

En base a la información publicada, las empresas hoteleras involucradas tienen contratos con una agencia de marketing que les ayuda a atraer clientes. Por el servicio pagan un coste fijo mensual a la agencia, pero la mayor parte del gasto proviene de los costes variables ligados a las reservas que se generan. Sin embargo, muchas de las reservas realizadas se cancelan finalmente, con lo que el hotel no sólo pierde el dinero desembolsado a la agencia (1.5€ por reserva), sino también el generado por la propia cancelación, especialmente si se realiza con menos de 3 días de antelación (lo que supone unos costes aproximados de 120€).

Antonio et al.(2019) comentan al describir este banco de datos, cierta dependencia entre la probabilidad de cancelación y las características propias del cliente y la reserva, recogidas durante el proceso de reserva. Resultará pues, especialmente relevante, poder predecir analíticamente el riesgo de cancelación en función de las características registradas sobre los clientes y la propia reserva.
Es más, en una sociedad en la que la inteligencia artificial se involucra cada vez más en las actividades cotidianas, cobra sentido poder incorporarla también en este caso y diseñar el análisis predictivo con técnicas automáticas de clasificación.

Objetivos

En este trabajo se plantea como objetivo principal investigar las cancelaciones de reservas hoteleras que se producen en la base de datos referida anteriormente. Como objetivos secundarios se tratará de:

- Identificar con cuáles de las variables registradas están más relacionadas las cancelaciones.
- Diferenciar perfiles de clientes para encontrar patrones que permitan predecir una cancelación más o menos tardía de su reserva hotelera, y con ello diseñar estrategias de actuación que permitan minorar los costes que suponen especialmente las cancelaciones tardías.

Las repercusiones que puede tener el conseguir respuestas a estos objetivos están directamente relacionadas con la reducción de costes en los hoteles, contribuyendo al diseño de estrategias alternativas de comercialización.

Además, estos objetivos serán abordados desde la implementación de algoritmos y estrategias automatizadas que eviten, en la medida de lo posible, la supervisión humana en la modelización. Se propondrán distintas estrategias alternativas, se compararán y se comentarán sus ventajas e inconvenientes para una predicción eficaz de la probabilidad o el riesgo de cancelación hotelera.

Información disponible

La base de datos que se analiza en este documento se ha obtenido de la plataforma web *Kaggle* y es de dominio público. Los datos han sido tomados del repositorio *Kaggle* (https://www.kaggle.com/datasets/datacertlaboratoria/proyecto-4-anlisis-de-cancelaciones-hoteleras).

Contiene dos bases de datos reales colapsadas, recogidas entre el 1 de julio de 2015 y el 31 de agosto de 2017 a través del sistema PLS de dos hoteles de Portugal, uno situado en el Algarve (tipo *resort*) y otro en la ciudad de Lisboa (tipo ciudad). Ambas tienen la misma estructura, por lo que se presenta una sola base de datos compuesta por datos relativos a gestiones de reserva en los hoteles. Cada observación (fila) representa una reserva hotelera, incluyendo las que efectivamente se efectuaron y las que se cancelaron. Todos los elementos que pertenecen a la identificación del hotel o del cliente han sido eliminados.

El banco de datos original de *Kaggle* está compuesto por 31 variables, de las cuales 13 son cuantitativas y 18 son categóricas. Consta de 119.390 registros, repartidos entre los años 2015 (21.996), 2016 (56.707) y 2017 (40.687).

Las variables disponibles en la base de datos, diferenciadas en categóricas, numéricas discretas y numéricas continuas, las mostramos en la *Tabla 1.*

Tabla 1: Variables disponibles en la base de datos original

Variable	Descripción	Tipo variable	Categorías
"Hotel"	Tipo de hotel	Categórica	Ciudad Resort
"Adultos"	Número de adultos en la reserva.	Numérica discreta	x
"Niños"	Número de niños en la reserva.	Numérica discreta	x
"Bebés"	Número de bebés en la reserva.	Numérica discreta	x
"Asignada"	Tipo de habitación asignada	Categórica	A B C D E F G H I K L P
"Reservada"	Tipo de habitación reservada	Categórica	A B C D E F G H L P
"Entre_semana"	Número de noches pernoctadas entre semana.	Numérica discreta	x
"Fin_semana"	Número de noches pernoctadas en fin de semana.	Numérica discreta	x

“Comida”	Tipo de reserva relativa al régimen de comidas.	Categórica	Cama y desayuno Pensión completa Media pensión Sin comidas Indefinido
“Aparcamiento”	Número de plazas de aparcamiento reservadas.	Categórica	0 1 2 3 8
“Peticiones”	Peticiones especiales realizadas en la reserva.	Categórica	0 1 2 3 4 5
“Cambios”	Cambios en la reserva	Numérica discreta	x
“Agente”	Identificador de la agencia de viajes que hizo la reserva	Categórica	334 categorías
“Compañía”	Identificador de la empresa que hizo la reserva	Categórica	354 categorías
“Canal”	Canal de reserva, disponible también a través de agentes de viajes (AV) y operadores turísticos (OT)	Categórica	Empresa Directo GDS Agentes de viajes/Operadores turísticos Indefinido
“Mercado”	Segmento del mercado	Categórica	Aviación Complementario empresa directo grupos Agentes de viajes/Operadores turísticos Agentes de viajes
“País”	País de origen del cliente.	Categórica	178 categorías
“Cliente”	Tipo de cliente.	Categórica	Contrato Grupo Transitorio Transitorio-fiesta
“Día”	Día del mes en que se proyecta la llegada.	Numérica discreta	x

"Semana"	Semana del año en que se proyecta la llegada.	Numérica discreta	x
"Mes"	Mes en que se proyecta la llegada.	Categórica	Enero Febrero Marzo Abril Mayo Junio Julio Agosto Septiembre Octubre Noviembre Diciembre
"Año"	Año en que se proyecta la llegada.	Categórica	2015 2016 2017
"Estado"	Estado de la reserva.	Categórica	Cancelada Chequeada No disponible
"Fecha_estado"	Fecha en que se chequeó el estado de la reserva.	Categórica	926 categorías
"Repite"	Si el cliente repite reserva o no.	Categórica	0 (No) 1 (Sí)
"Cancelación"	Indica si se ha realizado una cancelación o no.	Categórica	0 (No) 1 (Sí)
"Cancelaciones_previas"	Número de cancelaciones previas realizadas por el cliente.	Numérica discreta	x
"Espera"	Número de días en lista de espera.	Numérica discreta	x
"Previas_no_canceladas"	Número de reservas previas no canceladas.	Numérica discreta	x
"Cadencia"	Tiempo desde la reserva hasta la llegada proyectada.	Numérica discreta	x
"Gasto_medio"	Tarifa media diaria	Numérica continua	x

Procesado de información

Como en todo análisis estadístico, siempre es preciso llevar a cabo en primer lugar un procesado y depuración del banco de datos, a través del cual se excluyan errores, variables irrelevantes para el objetivo del estudio, y se creen variables más eficientes para proporcionar información en relación a los objetivos de análisis. Además, dado el volumen de datos disponible (ya comentado previamente), es fácil intuir que la carga computacional será grande cuando planteemos algoritmos de clasificación con los que resolver los objetivos; intentaremos pues, limpiar y reducir en la medida de lo posible dicha base de datos, para disminuir la carga computacional y conseguir mayor eficiencia en los cálculos.

El procesado de información se ha llevado a cabo pues, según tres tipos de operaciones: eliminación de variables, eliminación de registros y creación de variables.

Eliminación de variables

Eliminamos cinco variables que consideramos irrelevantes para el objetivo de nuestro estudio, que se muestran en la *Tabla 2* identificadas por sus nombres y el motivo de su eliminación.

Tabla 2: Variables originales eliminadas

Variable	Motivo eliminación
"Agente"	No consta información sobre su significado
"Compañía"	No consta información sobre su significado y contiene registros con valor *NULL*
"Estado"	Contiene información duplicada de la variable *"Cancelación"*
"Fecha_estado"	Contiene información duplicada de las variables *"Día"*, *"Mes"* y *"Año"*
"Mercado"	Contiene información duplicada de la variable *"Canal"*

Eliminación de registros

Del total de registros eliminamos aquellos en los que falta información o contienen información errónea.

Analizamos la existencia de datos faltantes e incongruencias en los valores de las diversas variables consideradas. Así pues, se eliminan un total de 409 registros u observaciones, por los motivos que se muestran en la *Tabla 3.*

Tabla 3: Registros eliminados

Motivo eliminación	Número registros eliminados
Datos faltantes en la variable “Niños”	4
“Gasto_medio” realizado por el cliente inferior a 0€ (sin sentido)	1
Reservas con ningún adulto (sin sentido)	403
Registro en la variable "Canal" con valor *"Indefinido" (no está identificada la categoría)*	1

Transformación de variables

Una vez hemos descartado todo lo que consideramos prescindible, y tras examinar con detenimiento la base de datos, decidimos transformar algunas variables en otras nuevas, de modo que aporten información eficiente reduciendo la dimensionalidad de la base de datos. Se muestra la propuesta de nuevas variables en la *Tabla 4* .

Tabla 4: Nuevas variables definidas.

Variable	Descripción	Variables origen y que sustituyen	Tipo/Categorías
“Familiar”	Tipo de cliente en función del número de adultos y menores en la reserva	*“Adultos”*, *“Niños”* y *“Bebés”*	Categórica / Solo Pareja Familia
“Habitación”	Cambio en la asignación del tipo de habitación solicitada al reservar	*“Reservada”* y *“Asignada” (sólo sustituye a “Asignada”)*	Categórica / Igual Peor Mejor
“Estancia”	Días de la semana en que se va a hospedar	*“Fin_semana”* y *“Entre_semana”*	Categórica / Fin semana Entre semana Combinado
“Tasa_cancelación”	Tasa de cancelación del cliente	*“Cancelaciones_previas” (CP)* y *“Previas_no_canceladas” (PNC)* Tasa=CP/max(1,(CP+PNC))	V. numérica

Reducimos así sustancialmente el número de variables “útiles” con estas nuevas variables (de tres a una en *“Familiar”*, de dos a una en *“Estancia”* y de dos a una en *“Tasa_cancelación”*), y enriquecemos la información que contienen: incorporando si se ha producido o no una mejora o un empeoramiento a la hora de asignar una habitación, respecto de la que se asignó inicialmente (en *“Habitación”*) y calculando un peso asociado al comportamiento histórico del

cliente respecto a cancelaciones previas que ha realizado (en *"Tasa_cancelación"*).

Una vez acabado el procesado de información, disponemos de una base de datos que cuenta con 118360 filas o registros y 19 columnas o variables, que se describen posteriormente en el apartado Análisis.

Metodología

Para abordar el análisis descriptivo inicial de todas las variables involucradas en el análisis consideramos:

- Para las variables de tipo numérico se muestran la media y la desviación típica globales, así como las parciales para los dos tipos de clientes, los que cancelan y los que no (dado que el objetivo es investigar la relación con la cancelación de reservas). También se aportan en las tablas los p-valores obtenidos con el test t de *Student* con el objetivo de investigar si dichas variables están asociadas de forma significativa a la cancelación de reservas.
- Para las variables de tipo categórico se muestran los conteos y porcentajes por categoría. También se realizan los test de asociación para investigar la relación con la cancelación de reservas y se muestran los p-valores obtenidos. Los tests de asociación con la variable *"Cancelación"* que se han utilizado han sido el *Chi-cuadrado* (para las variables dicotómicas) y el *Cramer's V* (para el resto).

Para abordar el análisis inferencial y dar respuesta a los objetivos planteados relativos a predecir la cancelación de reservas, se utilizan varias versiones de los árboles de clasificación, que proporcionan un procedimiento automatizado para la selección de variables y la identificación de perfiles de clientes. En las propuestas se consideran dos objetivos: predecir la probabilidad de cancelación, y predecir el riesgo de cancelación en función del tiempo que transcurre desde la reserva. Esto nos lleva a proponer modelos logísticos y también de supervivencia. En concreto, los modelos propuestos están basados en:

- Árboles de regresión logística para predecir la probabilidad de cancelación, a los que nos referiremos como ARL, incluyendo todos los predictores disponibles en el banco de datos procesado. Los perfiles obtenidos serán descritos en términos del grupo en que son clasificados (cancelan/no cancelan).
- Árboles de regresión logística para predecir la probabilidad de cancelación, incluyendo todos los predictores en el banco de datos salvo la variable que registra los tiempos hasta llegada al hotel (desde la reserva) y que están censurados por el evento de cancelación. Nos referiremos a ellos como ARLR y los perfiles de clientes que generen

serán descritos en términos de curvas *Kaplan-Meier* (tiempo hasta cancelación) para predecir el riesgo de cancelación.
- Árboles de supervivencia para predecir el riesgo de cancelación, utilizando como variable respuesta la conjunción de las variables *"Cadencia"*, que contiene los tiempos hasta evento (cancelación/llegada al hotel) y *"Cancelación"*, que da información sobre la censura por cancelación. Los perfiles generados serán descritos en términos de curvas *Kaplan-Meier*. Nos referiremos a ellos como ARR.

Identificamos a continuación los modelos propuestos bajo esta diferenciación descrita anteriormente:

- MODELO A. Partición recursiva y árboles de regresión basados en regresión logística (ARL)
- MODELO B. Árboles de inferencia condicional basados en regresión logística (ARL)
- MODELO C. Árboles de inferencia condicional basados en regresión logística con los que se describe el riesgo de cancelación (ARLR).
- MODELO D. Árboles de inferencia condicional para predecir la supervivencia o riesgo de cancelación (ARR).
- MODELO E. Por último, se incluyen los modelos de partición recursiva basada en el modelo logístico, en sus dos versiones ARL y ARLR.

La primera implementación de árboles de clasificación (A) la realizamos utilizando la metodología CART, con la librería *rpart* de R (*Recursive Partitioning and Regression Trees,* Therneau, 2022). El algoritmo implementado usa la medida de *Gini* (Gini, 1912) para la división de los nodos, que está basada en la varianza total de las K clases que se generan y da información sobre la pureza de cada uno de los nodos que surgen. Estos árboles son fáciles de usar e interpretar, pero no son competitivos con los mejores métodos de aprendizaje automático. De manera general, lo que hace este algoritmo es encontrar la variable independiente (de entre los predictores disponibles) que mejor separa los datos en dos grupos. Esta mejor separación es expresada con una regla y a cada regla corresponde un nodo.

Según Mendoza (2018), la principal ventaja de este método es su interpretabilidad, ya que nos da un conjunto de reglas a partir de las cuales se pueden tomar decisiones. Es un algoritmo que no es demandante en capacidad de cómputo, comparado con otros y, a pesar de ello, tiende a dar buenos resultados. Por otro lado, su principal desventaja es que da una clasificación "débil", y sus resultados pueden variar mucho dependiendo de la muestra de datos usados para entrenar un modelo. Además, presenta un fuerte sesgo de selección, lo que puede afectar en la preferencia de seleccionar predictores con elevado número de categorías.

Para elegir el árbol óptimo utilizando la librería *rpart* se sigue el criterio del error estándar de Breiman et al.(1984), para comparar modelos en función de su complejidad y el error que generan en la clasificación. Sugieren que un árbol óptimo
debe ser aquel que tenga la menor cantidad de nodos terminales, con un error estándar mínimo y con el menor de los costes bajo el punto de vista de la información que debe aportar. Por lo tanto, se selecciona el modelo en el que el error no varíe sustancialmente con respecto a un modelo con mayor complejidad (o profundidad). Con este algoritmo integramos la selección automática basada en valores óptimos para el parámetro de complejidad (cp) y la profundidad del árbol.

Las aproximaciones B, C y D las efectuamos utilizando la función *ctree* de R, que construye árboles de clasificación condicionales. Los árboles de clasificación condicionales (Hothorn et al, 2006) son un tipo de árboles de regresión no-paramétricos, aplicables a todos los tipos de problemas de regresión, incluyendo respuestas nominales, ordinales, numéricas, censuradas, así como a variables respuesta multivariantes y covariables medidas en escalas arbitrarias. Estos árboles plantean una estrategia de provocar particiones binarias en las variables explicativas (covariables y factores) de forma recursiva, de modo que se van seleccionando como covariables/factores que segmentan, aquellas que manifiestan mayor asociación con la respuesta (en términos de ciertos estadísticos condicionales de asociación). Las variables que aparecen antes en el árbol son aquellas que manifiestan una mayor asociación con la respuesta (p-valores significativos más pequeños), y la partición binaria que se propone sobre ella es aquella que genera las máximas diferencias en las respuestas entre los dos grupos resultantes.
Por otro lado, la ventaja de utilizar estos árboles es la flexibilidad que da el hecho de no exigir el compromiso con ningún modelo paramétrico, y por lo tanto con una distribución concreta para la respuesta.

El modelo B incluye todos los predictores entre las posibles variables predictoras, y origina como resultados clasificaciones binarias (cancelaciones/no cancelaciones).

El modelo C, también sobre la respuesta binaria *"Cancelación"*, se ajusta sin incluir entre los predictores el tiempo hasta la llegada al hotel (censurada por las cancelaciones), y este tiempo se utiliza finalmente para caracterizar los perfiles de clientes a través de curvas *Kaplan-Meier.* Estas curvas representan la probabilidad acumulada del tiempo hasta el evento de interés, que en nuestro caso es la cancelación de la reserva.

La desventaja que presentan los árboles condicionales es que el ajuste no está basado en ningún modelo paramétrico, esto es, no asume ninguna distribución paramétrica para la respuesta, por lo que no es posible comparar modelos

(particiones) alternativos en términos de verosimilitud u otros criterios basados en ella.

A diferencia de los demás métodos de clasificación, *ctree* utiliza un procedimiento de prueba de significación para seleccionar variables en lugar de seleccionar la variable que maximiza una medida de información (como el coeficiente de *Gini* con *rpart*). Este procedimiento se considera imparcial ya que selecciona los predictores a través de una hipótesis nula global de independencia entre cualquiera de los predictores y la respuesta. Es por esto por lo que se considera este algoritmo mejor que el involucrado en *rpart*.

Para el árbol de regresión logística en los modelos B y C, la profundidad óptima la decidimos a través de diversas métricas habituales para problemas de clasificación que presentamos a continuación, y que se definen en términos de bondad de la clasificación. La profundidad para el árbol de supervivencia del modelo D viene dada por las limitaciones computacionales de nuestros equipos, si bien se podrían utilizar criterios similares a los basados en la bondad de la clasificación.

Como regla general en estos modelos, B, C y D, exigimos siempre un número mínimo de observaciones en los nodos terminales: al menos del 10% de los registros, con el fin de garantizar perfiles consecuentes de clientes.

Las principales métricas utilizadas en *Machine Learning* para juzgar la eficiencia de los algoritmos de clasificación, o lo que es lo mismo, concluir sobre la bondad del ajuste, son: exactitud, precisión, recuerdo y puntuación F1. Partimos de datos binarios con respuestas posibles 0 (negativo) y 1 (positivo), y que a través del algoritmo va a proporcionar estimaciones también de 0 (negativo) y 1 (positivo), como se muestra en la *Tabla 5*. Las métricas de bondad de la clasificación habituales son:

- La **exactitud** es, de todas las métricas, la más utilizada y muestra la capacidad del modelo para hacer predicciones correctas. Esta es simplemente un indicador de la bondad del modelo a la hora de hacer predicciones correctas y sólo será útil si el conjunto de datos está equilibrado. Cuando tenemos un conjunto de datos sesgado o hay desequilibrios, necesitamos una perspectiva diferente sobre cómo evaluar el modelo. Por eso introducimos otras métricas que pueden ser mucho más útiles que la exactitud, como la precisión (*precision*), el recuerdo (*recall*) y la puntuación F1 (f1 *score*).
- La **precisión** muestra el porcentaje de las predicciones positivas que son realmente positivas (o correctas).
- Por otro lado, el **recuerdo** muestra el porcentaje de las muestras positivas reales que se clasifican correctamente. Existe un equilibrio entre la

precisión y la recuperación de modo que, generalmente, el aumento de la precisión disminuye la recuperación y viceversa. Un clasificador que no tiene falsos positivos tiene una precisión de 1, y un clasificador que no tiene falsos negativos tiene un recuerdo de 1. Idealmente, un clasificador perfecto tendrá la precisión y el recuerdo de 1.

- Para aunar la información contenida en los índices de precisión y recuerdo, surge una métrica llamada **puntuación F1**. La puntuación F1 es la media armónica de la precisión y el recuerdo, y muestra lo bueno que es el modelo a la hora de clasificar todas las clases sin tener que equilibrar la precisión y la recuperación. Si la precisión o la recuperación son muy bajas, la puntuación F1 también lo será.

Tanto la exactitud como la precisión y el recuerdo pueden calcularse fácilmente utilizando una matriz de confusión, presentada en la *Tabla 5*, que muestra el número de predicciones correctas e incorrectas realizadas por un clasificador en todas las clases disponibles. De forma más intuitiva, una matriz de confusión se compone de 4 elementos principales:

- Verdaderos positivos (TP): Número de muestras que se clasifican correctamente como positivas, y su etiqueta real es positiva.
- Falsos positivos (FP): Número de muestras que se clasifican incorrectamente como positivas, cuando en realidad su etiqueta real es negativa.
- Verdaderos negativos (TN): Número de muestras clasificadas correctamente como negativas, cuando su etiqueta real es negativa.
- Falsos negativos (FN): Número de muestras clasificadas incorrectamente como negativas, cuando en realidad su etiqueta real es positiva.

La matriz de confusión se puede expresar como se muestra en la *Tabla 5*, tomando como posibles valores de la respuesta los valores 0 (negativo) y 1 (positivo):

Tabla 5: Matriz de confusión tras la clasificación de una variable binaria (0/1).

		Valor Predicho	
		0	**1**
Valor Observado	**0**	TN	FP
	1	FN	TP

Utilizando la notación relativa a los conteos de clasificaciones en cada una de las celdas de la matriz de confusión, las métricas definidas se calculan matemáticamente como:

$$Exactitud\ (A) = (TP + TN) / (TP + TN + FP + FN)$$

$$Precisión\ (P) = TP / (TP + FP)$$

$$Recuerdo\ (R) = TP / (TP + FN)$$

$$F1 = (2 * P * R) / (P + R)$$

Por último, para ajustar el modelo E propuesto, utilizamos la función *glmtree* de la librería *partykit*, con la que generamos modelos de partición recursiva basada en el modelo logístico para la variable respuesta binaria *"Cancelación"*.

La partición recursiva basada en modelos (Zeileis and Hothorn, 2015), en lugar de ajustar un modelo global a un conjunto de datos, estima modelos locales en subconjuntos de los datos que se generan aprendiendo entre sí a través de particiones recursivas. La partición finalmente proporcionará una estimación de la probabilidad de cancelación en cada uno de los nodos terminales, garantizando que dichas estimaciones son significativamente diferentes entre los distintos nodos terminales o perfiles de clientes que se generan.

El objetivo de una estimación basada en modelos paramétricos parte de un planteamiento genérico basado en modelizar de modo paramétrico la respuesta a predecir, para minimizar cierta función con la que se cuantifica el error entre la estimación y el valor observado, y que depende de un parámetro incluido en dicha modelización paramétrica de la respuesta (como puede ser la verosimilitud cambiada de signo u otro criterio útil para la comparación de modelos).

El proceso de generación del árbol sigue los siguientes pasos:

1. Se ajusta un modelo paramétrico a un conjunto de datos (en nuestro caso un modelo logístico con respuesta binomial).
2. Se testa la "inestabilidad paramétrica" de θ (variaciones en la estimación de los parámetros del modelo) sobre un conjunto de variables (explicativas), cuya respuesta es susceptible de ser "dividida" en dos trozos.
3. De entre las variables que muestran inestabilidad paramétrica, se secciona el modelo con respecto a la variable con la inestabilidad más

grande (o lo que es lo mismo, la asociación más fuerte o el p-valor más pequeño). Se elige el punto de corte que proporciona la mejora más grande en el ajuste del modelo respecto de la función de error.

4. Se repite el procedimiento en cada una de las submuestras resultantes hasta que no es posible particionar más.

Para calcular la estabilidad de todos los parámetros del modelo en cierta variable de partición se usan los tests propuestos por Zeileis and Hornik (2007), así como una corrección por *Bonferroni* al testar conjuntamente múltiples variables de partición.

Con este algoritmo, es posible también detener el crecimiento del árbol especificando el mínimo tamaño de los nodos terminales (*minsize*) y la máxima profundidad del árbol (*maxdepth*), o incluso especificando criterios concretos de selección como la verosimilitud, el AIC o el BIC.

Para decidir la profundidad óptima utilizaremos el criterio BIC, que tiene en cuenta la log-verosimilitud ($\hat{L}$) de los datos y penaliza por la complejidad del modelo, expresada en términos del número de parámetros o nodos terminales generados, *k*, y por el número de datos o registros disponibles *n*.

$$BIC = -2 \cdot ln\left(\hat{L}\right) + k \cdot ln\,(n)$$

En nuestro caso, asumimos una distribución *Bernoulli* para la variable dicotómica *"Cancelación"*, que se relaciona de modo lineal con los predictores a través de una función de enlace o link logit, esto es,

$$logit(p) = X'\beta\ ,$$

donde p es la probabilidad de cancelación, *X* es la matriz que contiene la información de los predictores disponibles en el banco de datos, y β son los coeficientes a estimar en el modelo. La función logit es el logaritmo de p/(1-p).

Resultados del análisis

En este apartado presentamos todos los resultados derivados de aplicar la metodología detallada anteriormente para describir la información disponible y ajustar a continuación distintas alternativas de clasificación con los que predecir la cancelación de reservas hoteleras.

Presentamos en primer lugar el análisis descriptivo del banco de datos y continuamos con el inferencial basado en los árboles de clasificación.

Análisis descriptivo

Realizamos un análisis descriptivo tanto para las variables numéricas (*Tabla 6*) como para las categóricas (*Tabla 7*).
En la *Tabla 6*, como comentamos en el apartado de Metodología, se muestran la media y la desviación típica globales para cada una de las variables de tipo numérico, que son un total de 5. También se dan las medias y desviaciones típicas calculadas para los dos tipos de clientes, los que cancelan y los que no, y los p-valores obtenidos para estudiar si dichas variables están asociadas a la cancelación de reservas, esto es, para concluir si ambos grupos de clientes son distintos en medias.

Tabla 6: Descriptivos variables numéricas

Variable	Cancelación	N	Media	SD	P.valor	N.global	Media.global	SD.global
Cadencia	0	74249	80.44	91.20	< 0.001	118360	104.45	106.93
	1	44111	144.88	118.64				
Cambios	0	74249	0.29	0.72	< 0.001	118360	0.22	0.64
	1	44111	0.10	0.45				
Espera	0	74249	1.60	14.85	< 0.001	118360	2.34	17.66
	1	44111	3.57	21.52				
Gasto_medio	0	74249	101.05	48.41	< 0.001	118360	102.55	50.03
	1	44111	105.07	52.55				
Tasa_cancelación	0	74249	0.00	0.03	< 0.001	118360	0.05	0.22
	1	44111	0.13	0.34				

Los datos indican que todas las variables proporcionan diferencias significativas en media (al 95% de confianza) entre los clientes que cancelaron y los que no, por lo que a priori podemos afirmar que estas variables están asociadas con la cancelación de las reservas cuando las consideramos individualmente, esto es, cuando no incluimos la información/asociación que aportan las restantes variables.

A modo de ejemplo de interpretación de la *Tabla 6*, en los clientes que cancelan, el tiempo que transcurre entre la reserva y la cancelación es de 144.88 días, mientras que en los que no cancelan el tiempo entre reserva y llegada al hotel es de 80.44 días. Esto hablaría a favor de que es más probable una cancelación cuando la reserva se gestiona de modo más anticipado.

En la *Tabla 7* mostramos el análisis descriptivo para las variables categóricas que hemos considerado incluir finalmente en el análisis, esto es, un total de 14 variables, incluida la que contiene la información crítica en nuestro problema, esto es, "*Cancelación"*. La *Tabla 7* presenta el número de registros y el porcentaje que representan en cada una de las categorías que las componen. Así por ejemplo, vemos que el 66.5% de los registros provienen del hotel de ciudad y 33.5% restante corresponden a reservas en el hotel de tipo resort. Es de especial interés que en el banco de datos disponible el 37.2% de los registros corresponden a reservas de clientes que han cancelado, tasa que es considerablemente alta y justifica pues, el análisis que llevamos a cabo para predecir la probabilidad de cancelación.

Tabla 7: Descriptivos variables categóricas

Variable	***Categorías***	***N***	***%***
Hotel	*Ciudad*	*78685*	*66.5*
	Resort	*39675*	*33.5*
Cancelación	*0*	*74249*	*62.7*
	1	*44111*	*37.3*
Año	*2015*	*21802*	*18.4*
	2016	*56108*	*47.4*
	2017	*40450*	*34.2*
Mes	*Abril*	*11027*	*9.3*
	Agosto	*13776*	*11.6*
	Diciembre	*6667*	*5.6*
	Febrero	*7987*	*6.7*
	Enero	*5856*	*4.9*

	Julio	*12554*	*10.6*
	Junio	*10881*	*9.2*
	Marzo	*9700*	*8.2*
	Mayo	*11690*	*9.9*
	Noviembre	*6706*	*5.7*
	Octubre	*11049*	*9.3*
	Septiembre	*10467*	*8.8*
Comida	*Cama y desayuno*	*91520*	*77.3*
	Pensión completa	*797*	*0.7*
	Media pensión	*14380*	*12.1*
	Sin comidas	*10503*	*8.9*
	Indefinido	*1160*	*1.0*
Canal	*Empresa*	*6589*	*5.6*
	Directo	*14422*	*12.2*
	GDS	*190*	*0.2*
	Agentes de viajes/Operadores turísticos	*97158*	*82.1*
	Indefinido	*1*	*0.0*
Repite	*0*	*114862*	*97.0*
	1	*3498*	*3.0*

Reservada	*A*	*85405*	*72.2*
	B	*898*	*0.8*
	C	*924*	*0.8*
	D	*19098*	*16.1*
	E	*6482*	*5.5*
	F	*2877*	*2.4*
	G	*2073*	*1.8*
	H	*597*	*0.5*
	L	*6*	*0.0*
	P	*0*	*0.0*
Cliente	*Contrato*	*4055*	*3.4*
	Grupo	*568*	*0.5*
	Transitorio	*88817*	*75.0*
	Transitorio-fiesta	*24920*	*21.1*
Aparcamiento	*0*	*110969*	*93.8*
	1	*7358*	*6.2*
	2	*28*	*0.0*
	3	*3*	*0.0*
	8	*2*	*0.0*

Peticiones	*0*	*69744*	*58.9*
	1	*32913*	*27.8*
	2	*12853*	*10.9*
	3	*2472*	*2.1*
	4	*338*	*0.3*
	5	*40*	*0.0*
Familiar	*Familia*	*14797*	*12.5*
	Pareja	*81175*	*68.6*
	Solo	*22388*	*18.9*
Habitación	*Igual*	*103929*	*87.8*
	Mejor	*591*	*0.5*
	Peor	*13840*	*11.7*
Estancia	*Combinado*	*60282*	*50.9*
	Entre semana	*51158*	*43.2*
	Fin semana	*6920*	*5.8*

Los resultados de los tests de asociación con la variable *"Cancelación"* han resultado significativos en todas las variables. Esto implica que cada una de ellas está relacionada con la cancelación cuando se la considera individualmente, sin incluir la información que aportan el resto de predictores. Es decir, en principio el reparto por categorías en estas variables es diferente para los clientes que cancelan que para los que no lo hacen.

Análisis inferencial

Posteriormente, llevamos a cabo el análisis inferencial para cada una de las técnicas empleadas. Mostramos los resultados de los árboles construidos siguiendo los criterios descritos anteriormente en la metodología.

MODELO A. Partición recursiva y árboles de regresión

Con el paquete *rpart* de *R* ajustamos el modelo A que ya describimos en la metodología, con el que obtenemos el árbol mostrado en el *Gráfico 1*. Se caracteriza por tener un índice de complejidad (cp) óptimo de 0.01 y profundidad 6.

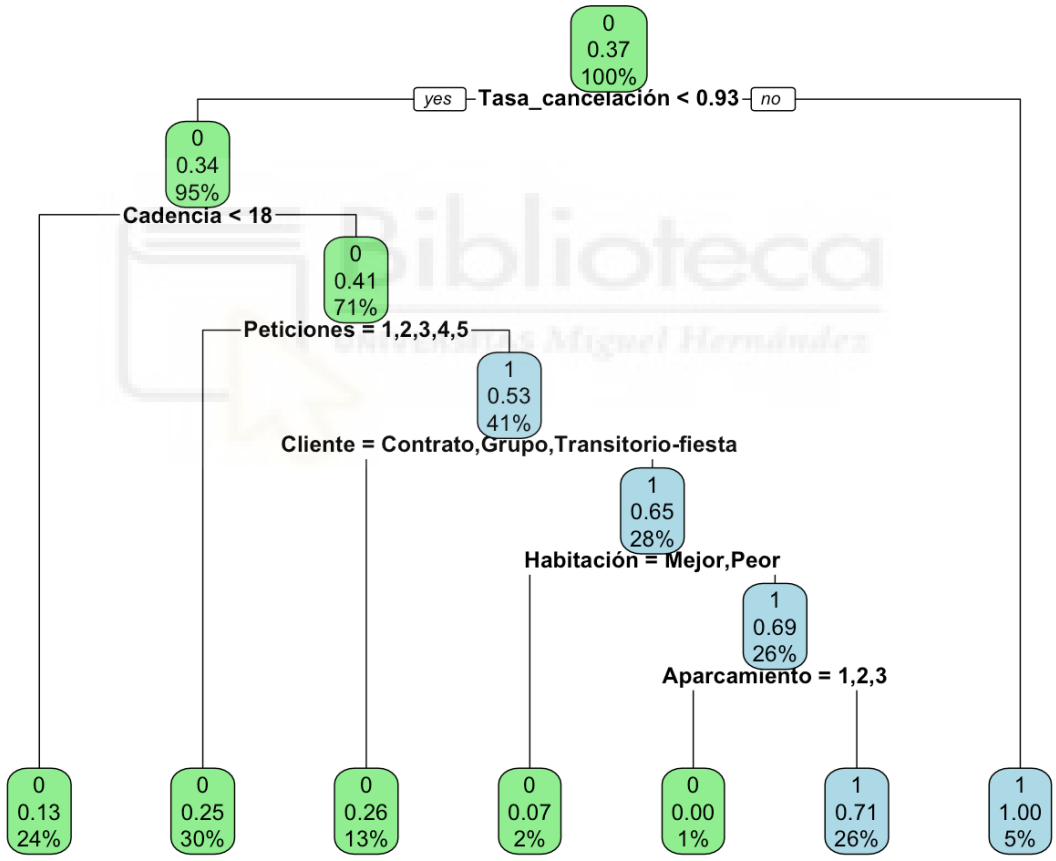

Gráfico 1: Árbol resultante con el modelo A.

Este árbol diferencia inicialmente (nodo raíz) entre sujetos con tasa de cancelación inferior y superior o igual al 93%. Incluye como variables relacionadas con la cancelación, las variables *"Tasa_cancelación" (los de mayor tasa de cancelación en su histórico, superior al 93%, son directamente clasificados como clientes que cancelan), "Cadencia", "Peticiones", "Cliente",*

"Habitación" y *"Aparcamiento",* y finalmente genera un total de 7 perfiles de clientes diferenciados por las particiones generadas en las variables seleccionadas y que se muestran en el árbol.
Además, en este árbol observamos que se generan nodos terminales con tamaño muy pequeño (1%-2%), pues el criterio de selección no impone mínimos.

Como resultado de la implementación del árbol de clasificación con *rpart*, en cada nodo se muestra en el *Gráfico 1*:

- La clase predicha: *Cancela (1) o No cancela (0)*.
- La probabilidad predicha.
- El porcentaje de observaciones en el nodo.

Los perfiles de clientes con el modelo A identificados los mostramos en la *Tabla 8.*

Tabla 8: Perfiles de clientes derivados del modelo A.

	Perfil 1	Perfil 2	Perfil 3	Perfil 4	Perfil 5	Perfil 6	Perfil 7
Tasa_cancelación	< 0.93	< 0.93	< 0.93	< 0.93	< 0.93	< 0.93	≥ 0.93
Cadencia	< 18	≥ 18	≥ 18	≥ 18	≥ 18	≥ 18	
Peticiones		1, 2, 3, 4, 5	0	0	0	0	
Cliente			Contrato, Grupo, Transitorio-fiesta	Transitorio	Transitorio	Transitorio	
Habitación				Mejor, Peor	Igual	Igual	
Aparcamiento					1, 2, 3	0	
Predicho	0	0	0	0	0	1	1
% observaciones	24%	30%	13%	2%	1%	26%	5%
Probabilidad	0.13	0.25	0.26	0.07	0.00	0.71	1

Los clientes clasificados como potencialmente susceptibles de cancelar (probabilidades estimadas próximas a 1) son aquellos que se identifican en los perfiles 6 y 7, con probabilidades estimadas de cancelación de 0.71 y 1 respectivamente. Los que cancelan con probabilidad 1, en el *Perfil 7,* tienen una tasa de cancelación igual o superior al 93%.
El *Perfil 5* es el tipo de cliente deseable para el hotel, con una probabilidad estimada de cancelación de 0. Estos clientes se caracterizan por tener una tasa

de cancelación (en su histórico) inferior al 93%, haber reservado con una antelación superior a 18 días, no haber realizado ninguna petición especial al hotel en su reserva, ser un cliente que transita ocasionalmente por el hotel, no haber variado la asignación del tipo de habitación reservada una vez que recibió la confirmación del hotel, y haber solicitado hasta 3 plazas de *parking*.
Si hubiéramos de establecer un ranking de clientes, de mejores (con menos probabilidad de cancelación p) a peores (con mayor probabilidad de cancelación), sería, según el modelo ajustado:
Perfil 5 (p=0.00) > Perfil 4 (p=0.07) > Perfil 1 (p=0.13) > Perfil 2 (p=0.25) > Perfil 3 (p=0.26) > Perfil 6 (0.71) > Perfil 7 (p=1).

MODELOS B, C Y D. Árboles de inferencia condicional

Con la función *ctree* de la librería *partykit,* construimos los modelos B, C y D, con los que obtenemos los árboles mostrados en el *Gráfico 2, Gráfico 4* y *Gráfico 6* respectivamente.

A la hora de decidir la profundidad óptima, exigimos en primer lugar cierto tamaño mínimo para los nodos terminales (entre el 10 y el 20% del volumen total de datos). Estos dos valores de exigencia para el tamaño de los árboles nos llevan a optar entre un árbol de profundidad 4 (exigiendo nodos terminales con un 20% de registros) y 19 (10% de registros). Para elegir una profundidad entre estas dos, procedemos a marcar otros criterios basados en la bondad de la clasificación; en particular, nos fijamos en la exactitud y la puntuación F1, ya que la precisión y el recuerdo están englobados en esta última. Estas métricas son buenos indicadores de la calidad del árbol para la predicción. En la *Tabla 9* se muestran los valores de bondad de la clasificación para los modelos ajustados con profundidad 4 y hasta profundidad 19 (un total de 16 modelos).

Los posibles criterios de decisión a la hora de elegir el árbol óptimo entre los 16 modelos propuestos son los siguientes:

- Si buscamos una exactitud por encima del 75% nos podríamos conformar con una profundidad de 5 ya que con dicho árbol se alcanza el criterio y el valor de los indicadores se estabiliza, aún profundizando un nivel más en el árbol: pasar de una profundidad 5 a 6 no hace variar la calidad del árbol respecto a los cuatro indicadores calculados.
- Si preferimos utilizar como criterio el recuerdo y la precisión por tener datos desequilibrados, como es el caso en nuestro banco de datos, donde es mucho inferior el número de cancelaciones, acudimos al criterio F1, en cuyo caso un posible árbol óptimo lo daría el de profundidad 7, no excesivamente complejo, pero con un valor de F1 superior a los siguientes dos árboles más complejos (profundidad 8 y 9).

- Si buscamos los valores máximos de exactitud y F1 nos quedaríamos con una profundidad máxima entre los límites considerados, esto es, con un árbol de profundidad 19.

Tabla 9: Valores de las métricas para distintas profundidades del modelo B.

Profundidad	Exactitud	Precisión	Recuerdo	F1
4	76.07	70.13	62.35	66.01
5	77.55	73.59	62.00	67.30
6	77.55	73.59	62.00	67.30
7	78.38	73.92	64.86	69.09
8	78.83	77.87	60.36	68.01
9	79.06	76.87	62.67	69.05
10	79.61	77.69	63.55	69.91
11	80.36	78.15	65.67	71.37
12	80.75	77.77	67.71	72.39
13	81.26	80.03	66.25	72.49
14	81.70	79.09	69.18	73.80
15	82.22	80.06	69.65	74.49
16	82.66	81.42	69.30	74.87
17	82.92	82.03	69.36	75.16
18	83.11	82.28	69.70	75.47
19	83.21	82.80	69.36	75.49

Para ejemplificar el modelo generado con el algoritmo de inferencia condicional, mostramos a continuación el árbol de profundidad 5, en el *Gráfico 2.*

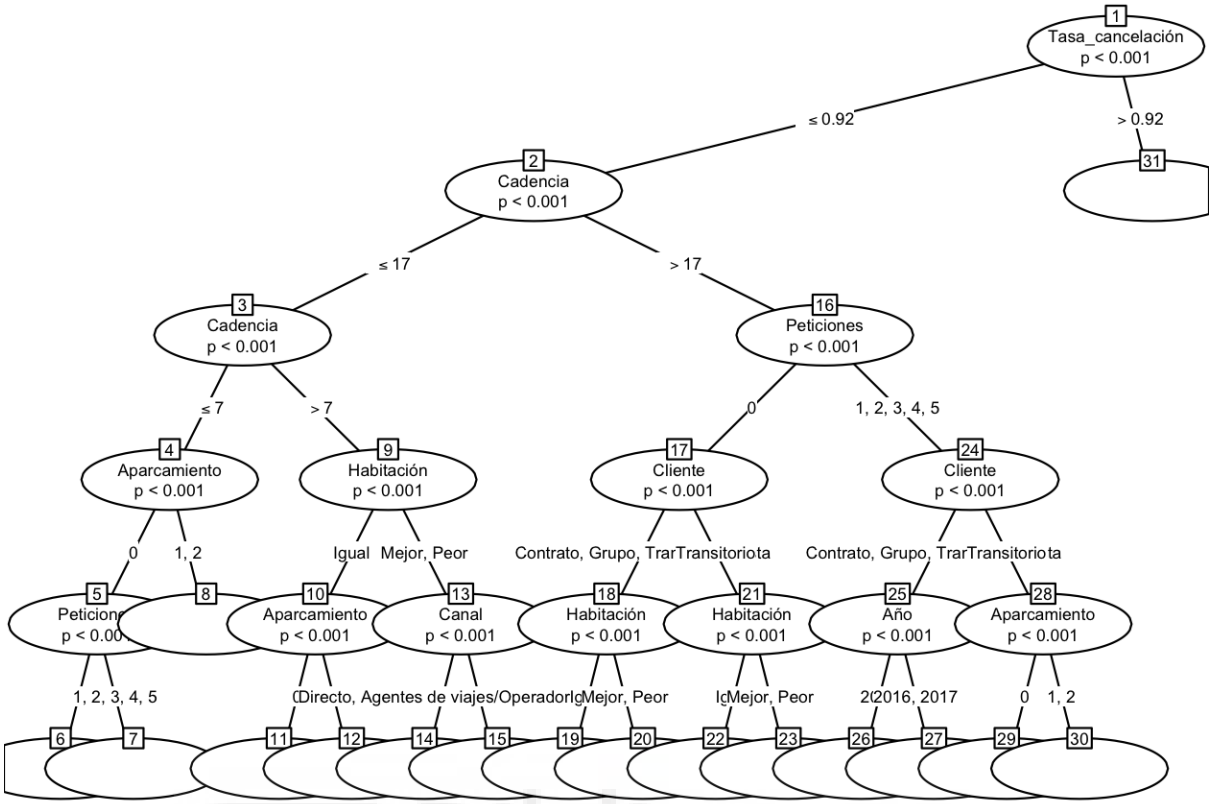

Gráfico 2: Árbol resultante con el modelo B.

En este gráfico podemos observar que las variables que intervienen y por lo tanto están asociadas con la cancelación son: *"Tasa_cancelación", "Cadencia", "Peticiones", "Cliente", "Habitación" , "Canal" , "Año"* y *"Aparcamiento"*.
En los nodos terminales no mostramos la información de manera gráfica ya que se superponen y no se puede distinguir la información. Esta información la mostramos a continuación en la *Tabla 10.1* y en la *Tabla 10.2*.

En cada nodo intermedio se muestra:

1. La variable que segmenta.
2. P-valor (indica el grado de asociación con la respuesta).

Los perfiles de clientes identificados son los que mostramos a continuación en las tablas 10.1 y 10.2.

Tabla 10.1: Perfiles (6-19) de clientes derivados del modelo B.

	Nodo 6	Nodo 7	Nodo 8	Nodo 11	Nodo 12	Nodo 14	Nodo 15	Nodo 19
Tasa_ cancelación	≤ 0.92	≤ 0.92	≤ 0.92	≤ 0.92	≤ 0.92	≤ 0.92	≤ 0.92	≤ 0.92
Cadencia	≤ 7	≤ 7	≤ 7	> 7, ≤ 17	> 7, ≤ 17	> 7, ≤ 17	> 7, ≤ 17	> 17
Aparcamiento	0	0	1, 2	0	1, 2, 3			
Peticiones	0	1, 2, 3, 4, 5						0
Habitación				Igual	Igual	Mejor, Peor	Mejor, Peor	Igual
Cliente								Contrato, Grupo, Transitorio-fiesta
Canal						Empresa	Directo, Agentes de viajes/ Operadores turísticos	
Año								
Predicción	0	0	0	0	0	0	0	0
% error de clasificación	13,00%	7,00%	0,00%	27,80%	0,00%	8,30%	2,90%	29,80%
Observaciones	10003	6490	2618	7194	665	252	1299	12410

Tabla 10.2: Perfiles (20-31) de clientes derivados del modelo B.

	Nodo 20	Nodo 22	Nodo 23	Nodo 26	Nodo 27	Nodo 29	Nodo 30	Nodo 31
Tasa_ cancelación	≤ 0.92	≤ 0.92	≤ 0.92	≤ 0.92	≤ 0.92	≤ 0.92	≤ 0.92	> 0.92
Cadencia	> 17	> 17	> 17	> 17	> 17	> 17	> 17	
Aparcamiento						0	1, 2	
Peticiones	0	0	0	1, 2, 3, 4, 5	1, 2, 3, 4, 5	1, 2, 3, 4, 5	1, 2, 3, 4, 5	
Habitación	Mejor, Peor	Igual	Mejor, Peor					
Cliente	Contrato, Grupo, Transito-rio-fiesta	Transitorio	Transitorio	Contrato , Grupo, Transito-rio-fiesta	Contrato , Grupo, Transito -rio-fiesta	Transitorio	Transitorio	

Canal								
Año				2015	2016, 2017			
Predicción	0	1	0	0	0	0	0	1
% error de clasificación	6,80%	31,30%	7,50%	13,30%	4,90%	31,60%	0,00%	0,40%
Observaciones	2480	31323	2053	1308	5495	26791	2136	5843

Los clientes clasificados por el modelo como clientes que cancelan (Predicción=1) son aquellos que se identifican en los nodos 22 y 31 (*Tabla 10.2*). El *Nodo 31* identifica muy claramente a los clientes que cancelan, puesto que el error de predicción es tan solo del 0.40%. Estos clientes se caracterizan por tener un valor en la variable *"Tasa_cancelación"* superior al 92%.
En este modelo, los nodos 8, 12 y 30 están identificando a clientes deseables para el hotel, esto es, que no reservan, con predicción de 0 y error de clasificación del 0%. Tienen en común el valor en la variable *"Tasa_cancelación"* inferior al 92%. Asimismo, los del *Nodo 8* se caracterizan por haber reservado con una antelación inferior o igual a 7 días, no haber solicitado plaza de aparcamiento y haber realizado entre 1 y 5 peticiones especiales al hotel en su reserva. Por otro lado, los del *Nodo 12*, se identifican por haber reservado entre una semana y 17 días de antelación, solicitar entre 1 y 3 plazas de aparcamiento y alojarse en una habitación del mismo tipo o categoría a la que reservaron. Por último, los clientes identificados en el *Nodo 30* se caracterizan por haber reservado con más de 17 días de antelación, solicitar una o dos plazas de aparcamiento y entre 1 y 5 peticiones especiales, y ser un tipo de cliente que transita ocasionalmente por el hotel.

A continuación, en el *Gráfico 3*, presentamos las proporciones de cancelación/no cancelación estimadas y observadas de cada uno de los nodos terminales con el modelo B. Se reconocen de nuevo los nodos 22 y 31 como aquellos en los que se estiman más cancelaciones (1) que llegadas al hotel (0). En el *Nodo 31* la cantidad de cancelaciones predichas y observadas es bastante similar, siendo esta prácticamente del 100%, mientras que en el *Nodo 22* es bastante superior la cantidad de cancelaciones predichas. Corroboramos lo comentado anteriormente sobre los nodos 8, 12 y 30 pues son la proporción observada de cancelaciones es del 0% (no hay barra para la proporción de cancelaciones).

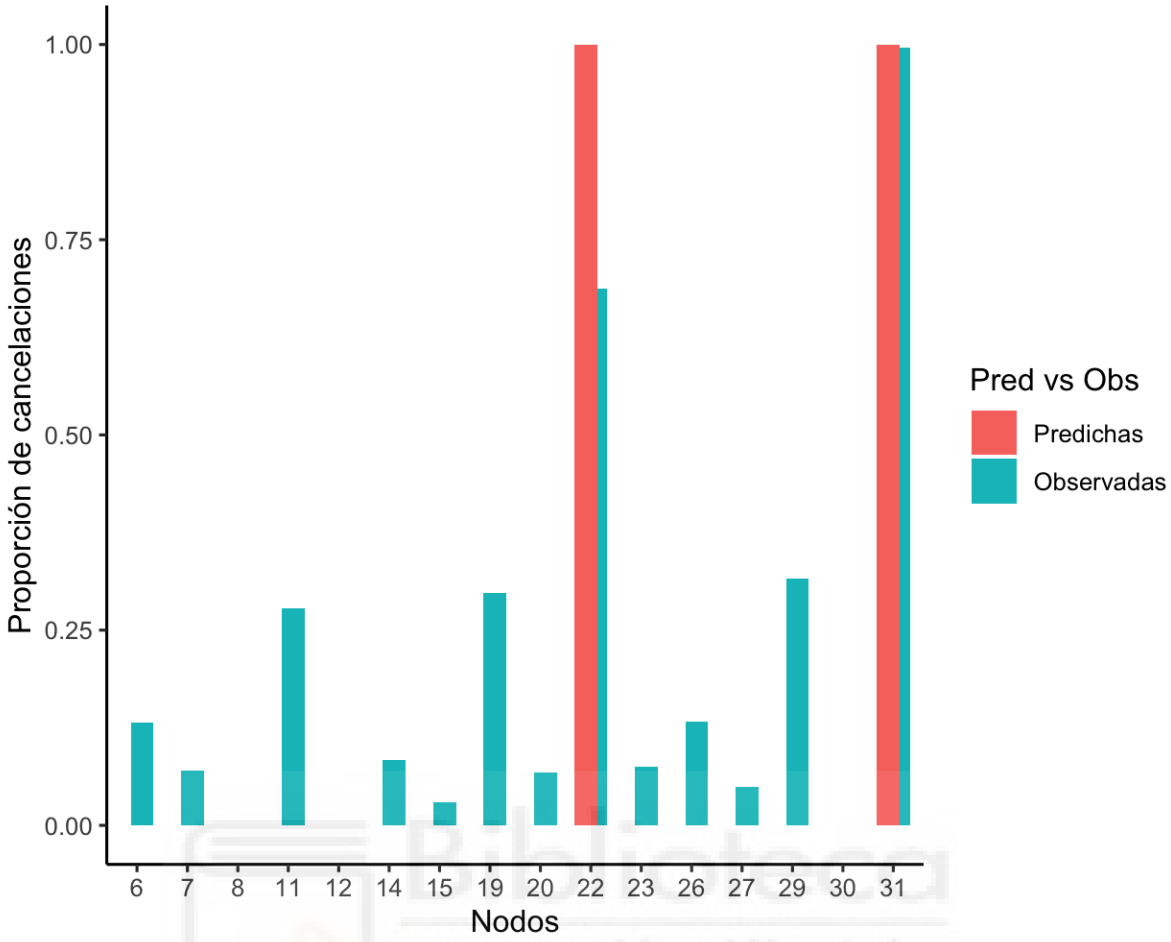

Gráfico 3: Proporciones predichas y observadas de cancelación de los nodos terminales del modelo B.

Procedemos a la construcción del modelo C, el árbol con todas las explicativas menos *"Cadencia".*

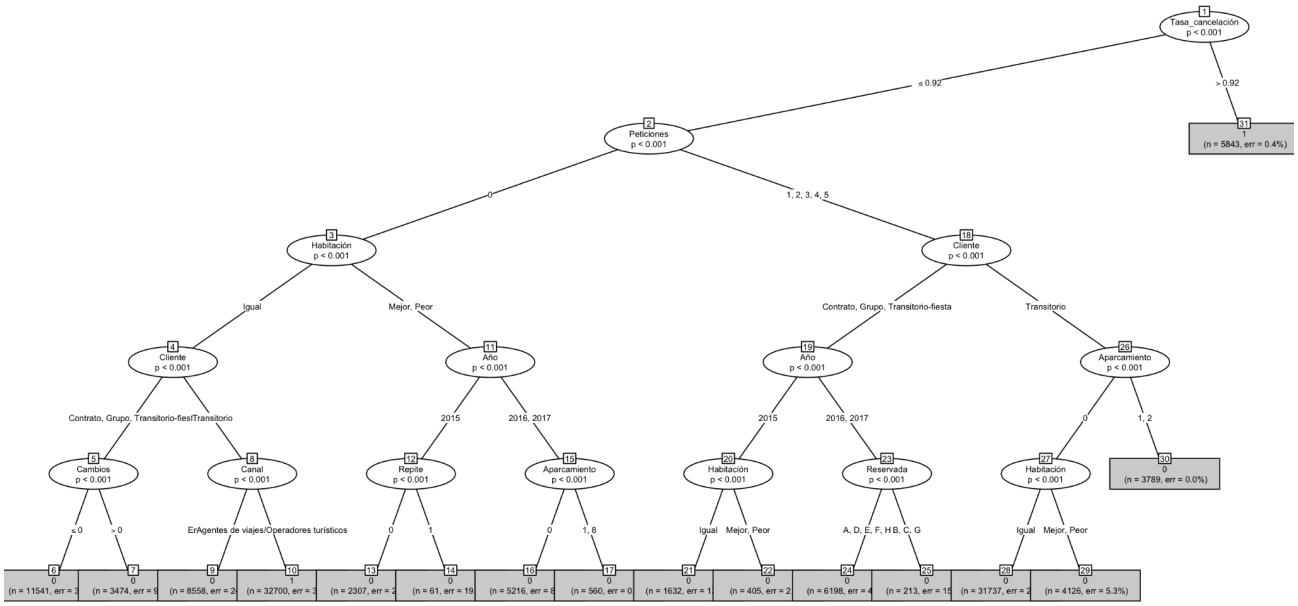

Gráfico 4: Árbol resultante con el modelo C.

En el *Gráfico 4* podemos observar que las variables que intervienen y están asociadas con la cancelación son: *"Tasa_cancelación", "Peticiones", "Habitación"* , *"Cliente", "Cambios", "Canal"* , *"Año", "Repite", "Reservada"* y *"Aparcamiento"*.

Dibujamos las curvas *Kaplan-Meier* para cada uno de los nodos y las probabilidades del tiempo hasta cancelación relativas a cadencia temporal desde la reserva, y con ellas identificamos cada uno de los perfiles. En el *Gráfico 5* observamos que tan solo obtenemos 14 curvas *Kaplan-Meier* de los 16 perfiles obtenidos con el modelo C. En concreto, los perfiles que no se muestran son el *Perfil 17* y *30*, esto se debe a que todos los que clientes que están clasificados en estos nodos son clientes que se han hospedado (no han cancelado) y por lo tanto no hay observaciones con las que ajustar las probabilidades a tiempo de cancelación.

Tabla 11.1: Perfiles (6-16) de clientes derivados del modelo C.

	Nodo 6	Nodo 7	Nodo 9	Nodo 10	Nodo 13	Nodo 14	Nodo 16
Tasa_ cancelación	≤ 0.92	≤ 0.92	≤ 0.92	≤ 0.92	≤ 0.92	≤ 0.92	≤ 0.92
Peticiones	0	0	0	0	0	0	0
Habitación	Igual	Igual	Igual	Igual	Mejor, Peor	Mejor, Peor	Mejor, Peor

Cliente	Contrato, Grupo, Transitorio -fiesta	Contrato, Grupo, Transitorio -fiesta	Transitorio	Transitorio			
Cambios	≤ 0	> 0					
Canal			Empresa, Directo, GDS	Agentes de viajes, Operadores turísticos			
Año					2015	2015	2016, 2017
Repite					0	1	
Aparcamiento							0
Reservada							
Clase	0	0	0	1	0	0	0
% error	32,80%	9,10%	24,40%	34,40%	2,70%	19,70%	8,80%
Observaciones	11541	3474	8558	32700	2307	61	5216

Tabla 11.2: Perfiles (17-31) de clientes derivados del modelo C.

	Nodo 17	Nodo 21	Nodo 22	Nodo 24	Nodo 25	Nodo 28	Nodo 29	Nodo 30	Nodo 31
Tasa_ cancelación	≤ 0.92	≤ 0.92	≤ 0.92	≤ 0.92	≤ 0.92	≤ 0.92	≤ 0.92	≤ 0.92	> 0.92
Peticiones	0	1, 2, 3, 4, 5	1, 2, 3, 4, 5	1, 2, 3, 4, 5	1, 2, 3, 4, 5	1, 2, 3, 4, 5	1, 2, 3, 4, 5	1, 2, 3, 4, 5	
Habitación	Mejor, Peor	Igual	Mejor, Peor			Igual	Mejor, Peor		
Cliente		Contrato , Grupo, Transito -rio- fiesta	Contrato , Grupo, Transito- rio-fiesta	Contrato, Grupo, Transito- rio -fiesta	Contrato , Grupo, Transito- rio-fiesta	Transitori o	Transitori o	Transitori o	
Cambios									
Canal									
Año	2016, 2018	2015	2015	2016, 2017	2016, 2017				
Repite									
Aparcamiento	1, 8					0	0	1, 2	
Reservada				A, D, F,	B, C, G				

				H					
Clase	0	0	0	0	0	0	0	0	1
% error	0,00 %	13,60%	2,20%	4,30%	15,50%	29,50%	5,30%	0,00%	0,40%
Observaciones	560	1632	405	6198	213	31737	4126	3789	5843

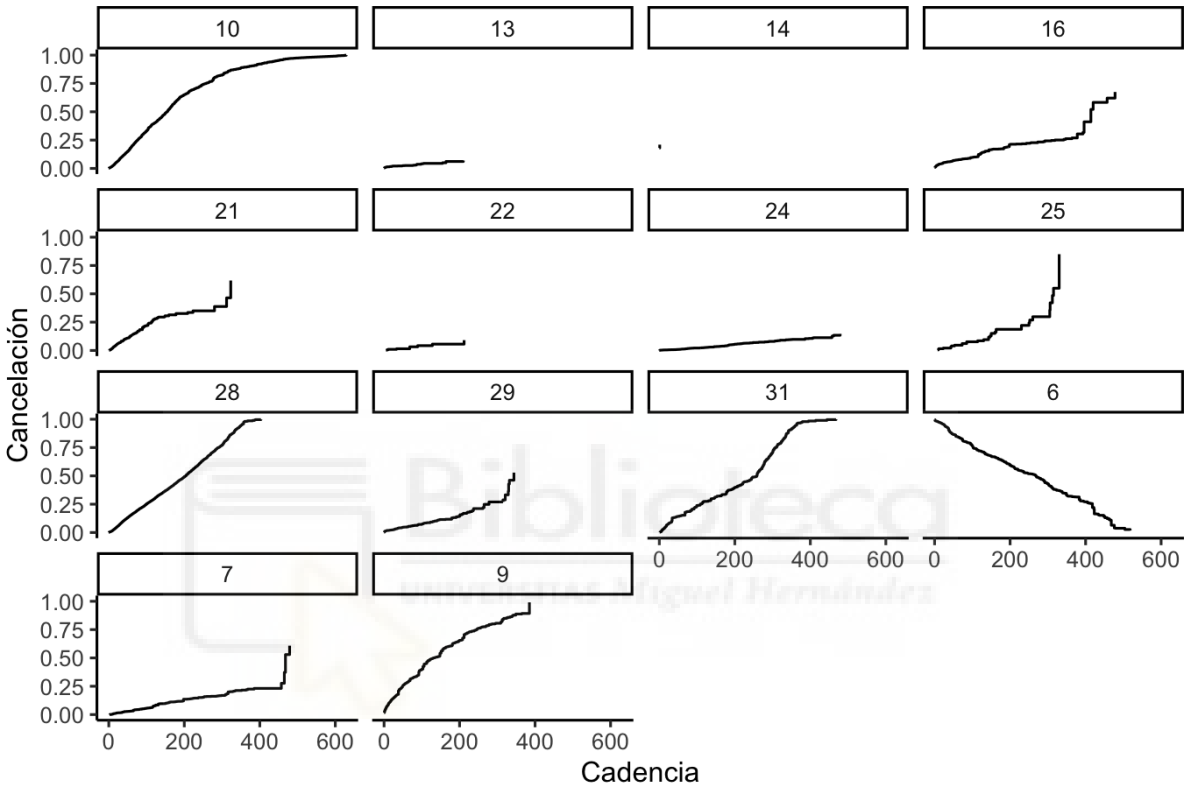

Gráfico 5: Curvas Kaplan-Meier de tiempo a evento (cancelación) del modelo C.

Los perfiles en los que la curva es más empinada, y por lo tanto identifican una mayor probabilidad a cancelar en un tiempo más corto, son los perfiles 9, 10 y 31. El perfil del *Nodo 6* identifica un comportamiento inesperado, con una curva decreciente para el tiempo hasta evento, y que requeriría un estudio más concienzudo para explicarlo.

Procedemos a continuación a la construcción del modelo D, basado en predecir el riesgo de cancelación, planteándolo como un problema de supervivencia (o riesgo de cancelación).

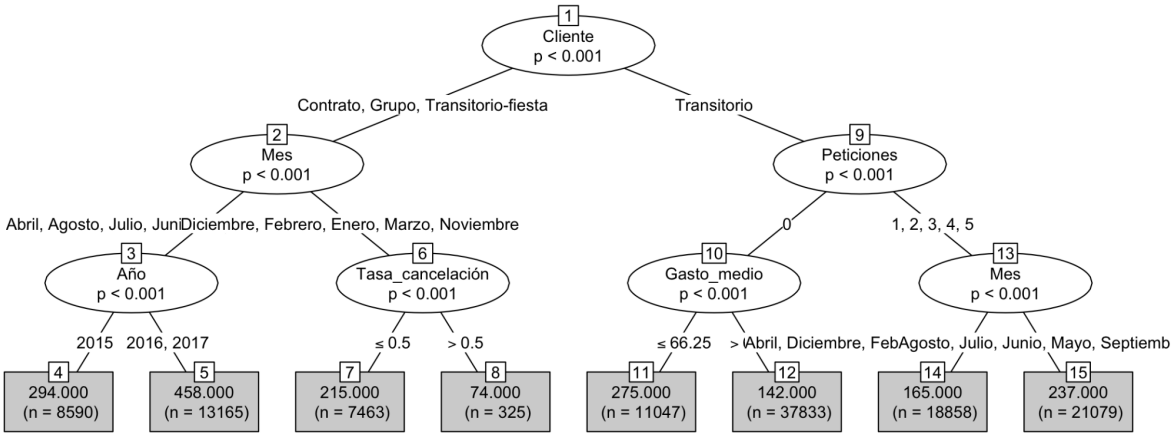

Gráfico 5: Árbol resultante con el modelo D.

En el *Gráfico 5* se muestra para cada nodo terminal:

1. Tiempo medio hasta cancelación.
2. Número de observaciones.

Las variables que intervienen e influyen en los tiempos hasta cancelación son: *"Cliente", "Mes", "Año", "Tasa_cancelación", "Peticiones"* y *"Gasto_medio".*

Los perfiles de clientes identificados son los mostrados en la *Tabla 12*.

Tabla 12: Perfiles de clientes derivados del modelo D.

	Cliente	Mes	Año	Tasa_ cancelación	Peticiones	Gasto_ medio	Tiempo medio (días)	Observaciones
Nodo 4	Contrato, Grupo, Transitorio -fiesta	Abril, Agosto, Julio, Junio, Mayo, Octubre Septiembre	2015				294	8590
Nodo 5	Contrato, Grupo, Transitorio -fiesta	Abril, Agosto, Julio, Junio, Mayo, Octubre Septiembre	2016, 2017				458	13165
Nodo 7	Contrato, Grupo, Transitorio -fiesta	Diciembre, Enero, Febrero, Marzo, Noviembre		≤ 0.5			215	7463

Nodo 8	Contrato, Grupo, Transitorio -fiesta	Diciembre, Enero, Febrero, Marzo, Noviembre		> 0.5			74	325
Nodo 11	Transitorio				0	≤ 66.25	275	11047
Nodo 12	Transitorio				0	> 66.25	142	37833
Nodo 14	Transitorio	Diciembre, Enero, Febrero, Marzo, Noviembre			1, 2, 3, 4, 5		165	18858
Nodo 15	Transitorio	Agosto, Junio, Julio, Mayo, Septiembre			1, 2, 3, 4, 5		237	21079

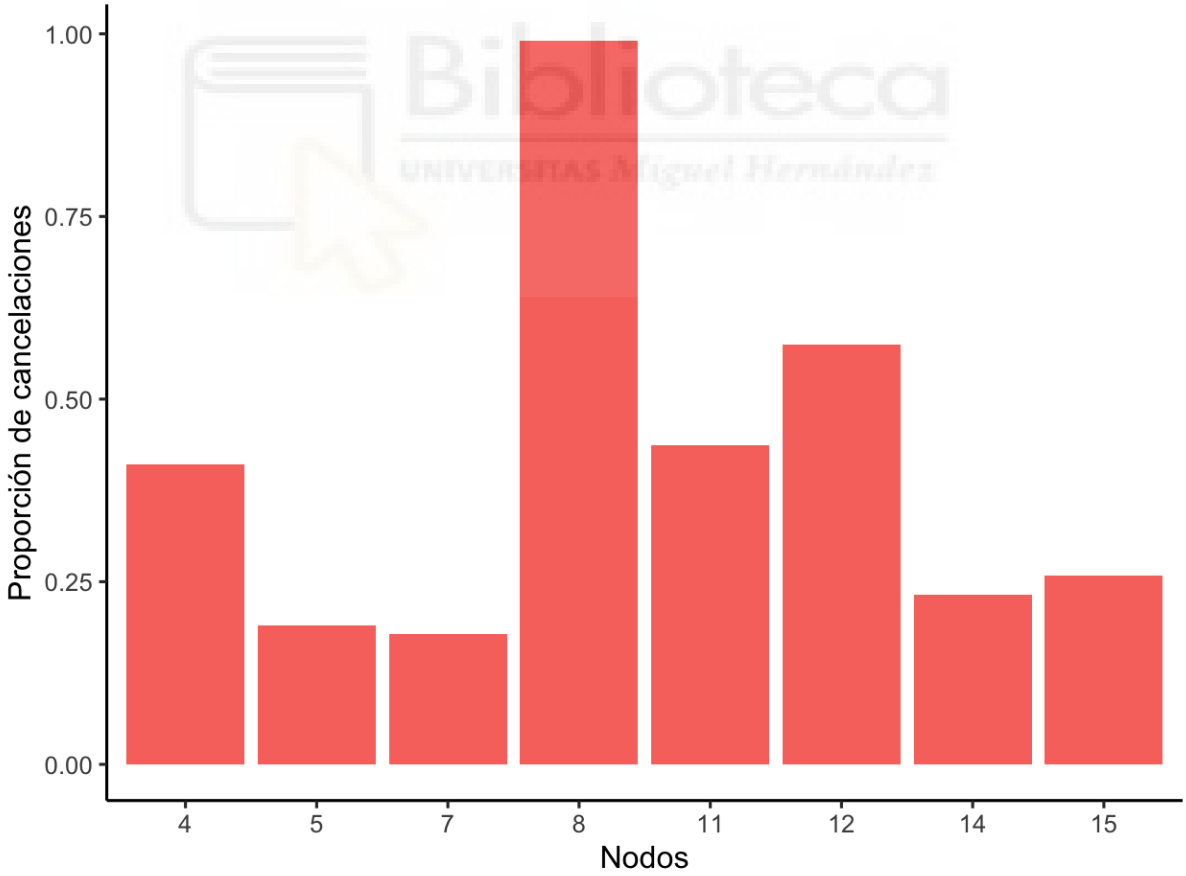

Gráfico 6: Proporciones observadas de cancelación de los nodos terminales del modelo D.

En el *Gráfico 6*, el nodo en el que más sujetos cancelan es el *Nodo 8,* teniendo un porcentaje de cancelación observado del 100%. Por el contrario, el nodo en el que menos cancelaciones se efectúan es el *Nodo 7,* con un porcentaje de cancelación inferior al 25%.

Dibujamos las curvas *Kaplan-Meier:*

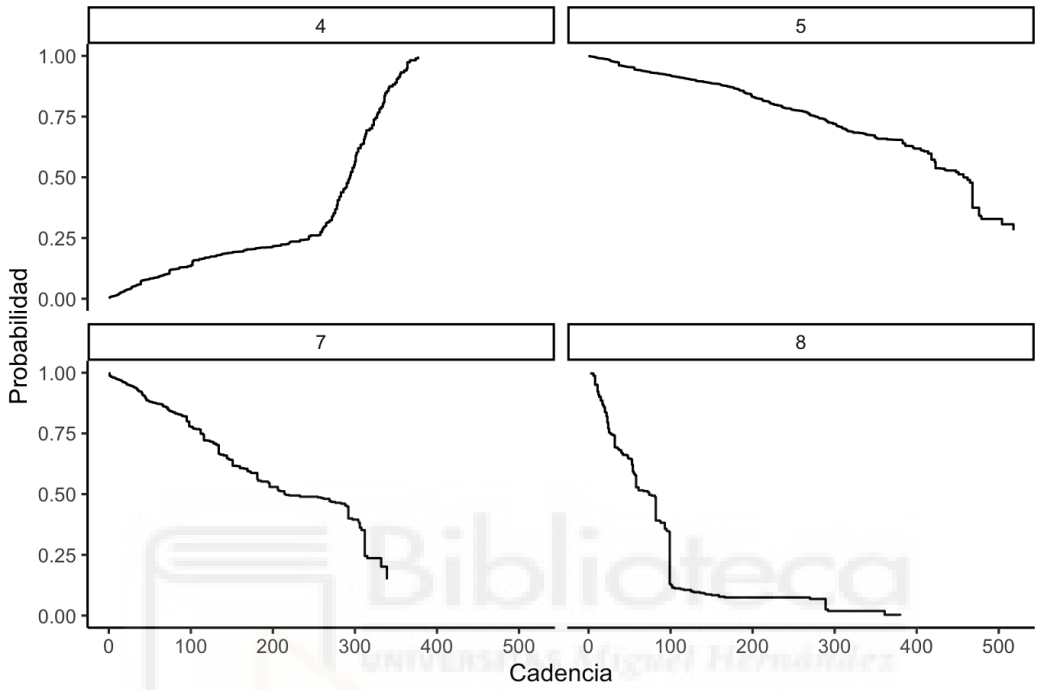

Gráfico 7: Curvas Kaplan-Meier de tiempo a evento (cancelación) del modelo D.

En el *Gráfico 7*, la probabilidad de cancelación en el *Nodo 4* aumenta de manera directa con el tiempo de cadencia. Por el contrario, en los nodos 5, 7 y 8, observamos una relación indirecta entre la probabilidad de cancelación y el tiempo de cadencia.

MODELO E. Modelos de partición recursiva basada en el modelo logístico

Construimos el modelo de partición recursiva basada en el modelo logístico para la variable binaria *"Cancelación"*. Primero lo haremos con todas las variables y, a continuación, con todas menos *"Cadencia"* (modelo E).

Construimos el modelo E siguiendo el criterio BIC de selección de variables, y por tanto de elección de la profundidad del árbol, desarrollado en la metodología. Obtenemos un árbol con 27 nodos terminales. No lo mostramos de manera gráfica ya que al tener tantas bifurcaciones y nodos terminales, es muy complicado mostrar con claridad la información en un gráfico. Las variables que

tienen relación con la cancelación según este modelo son: *"Cambios", "Espera", "Cliente", "Tasa_cancelación", "Año", "Peticiones", "Habitación", "Hotel", "Gasto_medio", "Mes", "Canal", "Estancia"* y *"Familiar"*.

Puesto que la variable *"Cadencia"* no aparece como predictor en el árbol, ni siquiera proponemos reajustarlo excluyéndola del conjunto de predictores, por lo que no se plantea el ajuste ARLR propuesto en la Metodología. El análisis de perfiles se resolvería de modo similar a como se presentó previamente para el modelo C a través de las curvas KM. Puesto que no es objetivo de este trabajo concluir con una solución óptima, sino proponer alternativas de análisis, no se presentan los resultados de este modelo de modo exhaustivo. De manera gráfica, presentamos el *Gráfico 7*, donde observamos que los nodos en los que más sujetos cancelan son el 5, 17 y 38, teniendo un porcentaje de cancelación observado del 100%. Por el contrario, los nodos en el que menos cancelaciones se efectúan son el 26, 28, 32, 52 y 53, con un porcentaje de cancelación inferior del 0%.

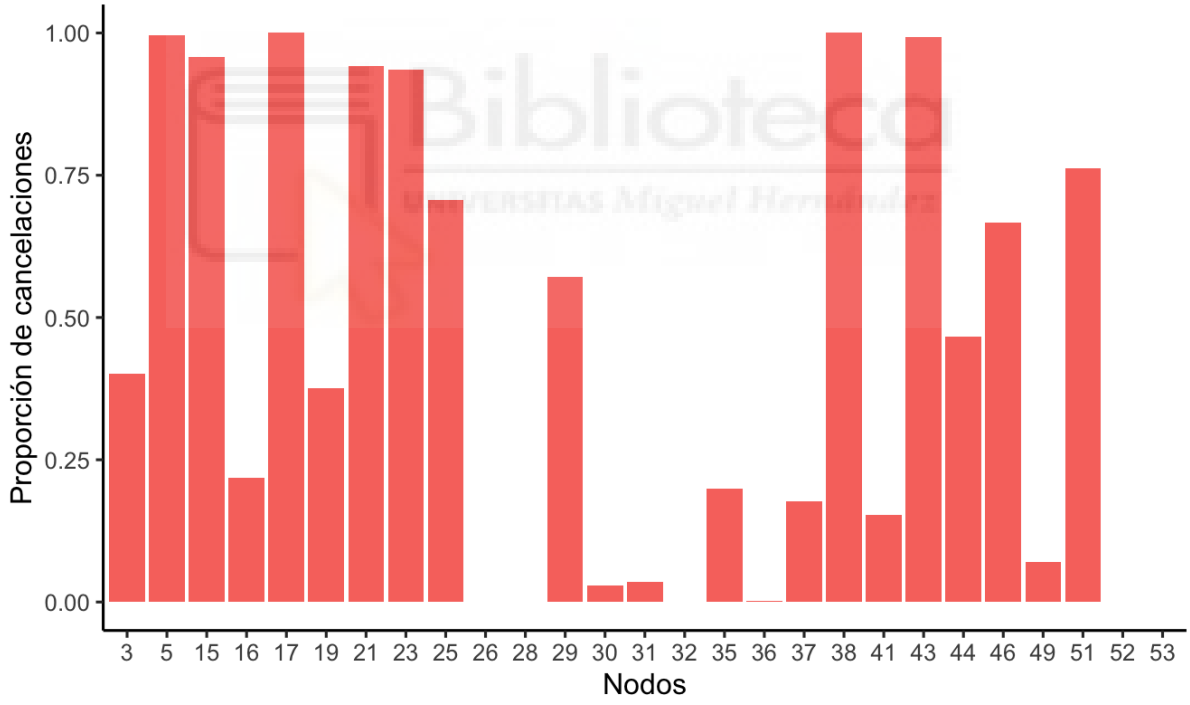

Gráfico 7: Proporciones observadas de cancelación de los nodos terminales del modelo E.

Conclusiones y líneas futuras

Empresas hoteleras portuguesas desean recibir una orientación en la toma de decisiones. Nuestro objetivo principal es investigar las cancelaciones de reservas hoteleras que se producen. Como objetivos secundarios tratamos de identificar con cuáles de las variables registradas están más relacionadas las cancelaciones y diferenciar perfiles de clientes para encontrar patrones que permitan predecir una cancelación más o menos tardía de su reserva hotelera.

Nos aportan una base de datos que contiene registros de reservas hoteleras realizadas *online* en dos hoteles de Portugal. Se detecta un problema de interés, relativo a una tasa de cancelaciones considerablemente alta, concretamente del 37.3%, y las pérdidas que ello ocasiona por el tipo de política que tienen los hoteles a la hora de remunerar a las agencias que les proporcionan las reservas, y especialmente por los gastos que generan las reservas tardías a menos de tres días de la fecha programada de llegada al hotel. Con el fin de dar respuesta a los hoteles al respecto de predecir la probabilidad o riesgo de cancelación en función de las características de las reservas y los clientes que las realizan, se plantea el análisis que se ha desarrollado en este trabajo.

Dado el volumen de información y la transversalidad del problema planteado, migrable prácticamente a cualquier centro hotelero, se propone resolverlo a través de técnicas automatizadas de *Machine Learning* que identifiquen perfiles de clientes diversos, para cada uno de los cuales sea posible estimar la probabilidad o el riesgo de cancelación transcurrido cierto tiempo desde la reserva. Dichas técnicas servirán así mismo para seleccionar qué variables están relacionadas con la cancelación de reservas, y cuáles no, y en consecuencia concretar las estrategias de recogida de información complementaria a la reserva, o diseñar igualmente, estrategias comerciales para retener el máximo de reservas posibles.

Una vez conocidos los objetivos, las técnicas de análisis automatizado elegidas en este trabajo son los árboles de clasificación. Puesto que la algoritmia desarrollada para ajustar estos árboles es muy variada, se ha optado por ajustar y comparar los algoritmos más comunes y utilizados en la práctica y basados en programación con el lenguaje estadísticos R. En concreto, se han utilizado algoritmos basados en particiones recursivas (modelo A con *rpart*), inferencia condicional (modelos B, C y D, con *ctree*) y partición basada en modelos (en particular, en el modelo logístico, modelo E, con *glmtree*). Los dos primeros tipos de algoritmos son no paramétricos y no asumen distribución alguna sobre la respuesta dicotómica que contiene la información sobre cancelaciones/no cancelaciones. En el modelo condicional se propone una primera modelización (B) para predecir la variable respuesta utilizando a priori todos los predictores disponibles (un total de 18: 13 variables categóricas y 5 numéricas), un segundo

modelo (C) prescindiendo en los predictores la variable que contabiliza el tiempo entre la reserva y la llegada al hotel o a cancelación (y que convierte a la variable *"Cancelación"* en una variable de censura), y un tercer modelo (D) que utiliza directamente la información combinada de ambas para predecir la probabilidad del tiempo hasta cancelación en términos de riesgo (a través del análisis de supervivencia). La última partición propuesta (E) basada en modelos, es de tipo paramétrico y asume una distribución *Bernoulli* sobre la variable *"Cancelación"*.

Cada uno de los algoritmos, por su modo propio de proceder en la estimación, requiere de unos criterios específicos de selección o elección de la profundidad del árbol de clasificación ajustado. En este trabajo se aplica el criterio de selección habitual para los modelos de partición recursiva, basada en complejidad óptima y error; para los árboles condicionales se proponen criterios populares en problemas de clasificación y basados en métricas de bondad de la clasificación (observados versus predichos); y para los árboles basados en modelos se usa el criterio BIC, implementado en el algoritmo y que da peso a la verosimilitud de los datos tratando de minimizar a la vez la complejidad del árbol, cuantificada con el número de nodos terminales que genera. El trabajo no trata de concluir con un modelo óptimo de ajuste, sino que muestra diferentes alternativas y criterios de parada, ilustrando los resultados que se obtienen. Proporciona además diversos modos de descripción de los perfiles de clientes que surgen, en términos de la probabilidad de cancelación que se estima (que se muestra gráficamente a través de gráficos de barras), y también en términos del riesgo de cancelar un tiempo dado después de la reserva (como se ilustra con las curvas *Kaplan-Meier* para los modelos que no incluyen la variable temporal como predictor).

A modo de conclusión práctica, tras mostrar todos los modelos ajustados, observamos cómo la variable más determinante a la hora de predecir si un cliente cancelará o no es la variable que hemos creado inicialmente a partir de la información sobre cancelaciones previas de los clientes, esto es, la *"Tasa_cancelación".* De hecho, en todos los árboles ajustados, a excepción del modelo D, es la primera variable que aparece en la partición. Todos los clientes cuya tasa de cancelación es superior al 93% tienen probabilidad 1 de cancelar su reserva.

Para los modelos A, B (con profundidad 5) y C, los perfiles de clientes más proclives a una cancelación (con probabilidades de cancelación próximas a 1) se muestran en la Tabla 13.

Tabla 13: Perfiles de clientes que cancelan en los modelos A, B y C.

	Tasa_ cancelación	Cadencia	Peticiones	Cliente	Habitación	Canal
Modelo A	≥ 0.93					
	< 0.93	≥ 18	0	Transitorio	Igual	
Modelo B	> 0.92					
	≤ 0.92	> 17	0	Transitorio	Igual	
Modelo C	> 0.92					
	≤ 0.92		0	Transitorio		Agentes de viajes/ Operadores turísticos

En la *Tabla 13* obtenemos seis perfiles de clientes que cancelan. Tres de ellos se caracterizan por estar definidos exclusivamente por la variable *"Tasa_cancelación"*, siendo en el modelo A mayor o igual al 93% y en los modelos B y C superior al 92%. El cuarto perfil de cliente se caracteriza por tener una tasa de cancelación inferior al 93%, reservar con 18 o más días de antelación, no realizar peticiones especiales, ser un cliente transitorio y tener asignada la misma habitación que reservó. El quinto perfil se distingue por una tasa de cancelación no superior al 92%, reservar con más de 17 días de antelación, no realizar peticiones especiales, ser un cliente transitorio y tener asignada la misma habitación que reservó. Por último, el sexto perfil de cliente se caracteriza por tener una tasa de cancelación inferior o igual al 92%, no realizar peticiones especiales, ser un cliente que está de paso y haber realizado la reserva a través de agentes de viajes (AV) y operadores turísticos (OT).

No es viable la comparación de estos modelos con el modelo de supervivencia (D), al variar la respuesta y la interpretación: en A, B y C se predice la probabilidad de cancelación y en el D se predice la probabilidad a un tiempo dado de cancelación.

Por otra parte, dado que el modelo E óptimo tiene mucha profundidad, y en consecuencia muchos nodos terminales (27), además de que incluye prácticamente todas las variables predictoras (13 de las 18 consideradas), no mostramos el "mejor perfil" del cliente que cancela con probabilidad alta, pues como se aprecia en el *Gráfico 7*, son bastantes los nodos (perfiles) en los que la proporción observada de cancelaciones es muy próxima al 1. Incluir esta descripción no aporta valor añadido a este trabajo por cuanto el mero hecho de haberlo ajustado y seleccionado ya ha cubierto los objetivos iniciales de mostrar

alternativas para la identificación de perfiles de clientes y selección de variables asociadas a la cancelación.

Como futuras líneas de trabajo, podríamos plantear seguir investigando a partir de los modelos propuestos, para conseguir un modelo óptimo de predicción de la probabilidad de cancelación, y en consecuencia la concreción de aquellas variables vinculadas a la misma, con las que guiar la toma de decisiones en los hoteles, para diseñar nuevas políticas comerciales.

Bibliografía

Arribalzaga, E. B. (2007). Interpretación de las curvas de supervivencia. *Revista Chilena de Cirugía*, *59*(1), 75-83. http://dx.doi.org/10.4067/S0718-40262007000100013

Assis Gomes, C. M., Lemos, G. C., & Jelihovschi, E. G. (2020). Comparing the Predictive Power of the CART and CTREE algorithms. *Avaliação Psicológica*, *19*(1), 87-96. http://dx.doi.org/10.15689/ap.2020.1901.17737.10

Borrás, F., Martínez Mayoral, M. A., & Departamento de Estadística, Matemáticas e Informática. Universidad Miguel Hernández de Elche. (2022). *Introducción a los algoritmos de aprendizaje automático*.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification And Regression Trees*. https://doi.org/10.1201/9781315139470

Corporación Sanitaria Parc Taulí. (2005). Conceptos básicos del análisis de supervivencia. *Cirugía Española*, *78*(4), 222-230. 10.1016/S0009-739X(05)70923-4

Chicco, D. & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary

classification evaluation. BMC genomics, 21(1), 1-13. https://doi.org/10.1186/s12864-019-6413-7

Fernández Casal, R., Costa Bouzas, J., & Oviedo de la Fuente, M. (2021, 10 13). *Aprendizaje Estadístico*. Aprendizaje Estadístico. Retrieved May 27, 2022, from https://rubenfcasal.github.io/aprendizaje_estadistico/

Gómez, G., & Departament d'Estadística i Investigació Operativa. (n.d.). Hablemos de… Análisis de supervivencia. *3*(4), 51-58. http://aeeh.es/wp-content/uploads/2012/05/v3n4a203pdf001.pdf

Hothorn, A., Hornik, K & Zeileis, A. (2006). Ctree: Conditional Inference Trees. Retrieved May, 31, 2022 from Hothorn, A., Hornik, K & Zeileis, A. (2006). Ctree: Conditional Inference Trees. Retrieved May, 31, 2022 from https://cran.r-project.org/web/packages/partykit/vignettes/ctree.pdf.

Marrero Hernández, F. J. (2016). *glosario de términos hoteleros, turísticos y relacionados*. Francisco José Marrero Hernández. https://www.hosteltur.com/files/web/templates/term/wikitur.pdf

Martínez Mayoral, M. A. (2001). *Modelos lineales generalizados*. Universidad Miguel Hernández.

Martinez Mayoral, M. A. (2021, 11 25). *Árboles de Supervivencia*. Wikipedia, the free encyclopedia. Retrieved May 27, 2022, from https://bookdown.org/asun_mayoral/arboles/arboles.html

Mendoza, J. B. (2018, April 23). *Arboles de decisión con R - Clasificación*. RPubs. Retrieved May 26, 2022, from https://rpubs.com/jboscomendoza/arboles_decision_clasificacion

Obregón N, N., García C, Ó., & Vivas A, L. (2005). *Modelamiento estadístico*. Universidad Nacional de Colombia.

Therneau, T. (2022, January 24). CRAN - Package rpart. Retrieved May 26, 2022, from https://cran.r-project.org/web/packages/rpart/index.html

Therneau, T. M., & Mayo Foundation. (2019). *An Introduction to Recursive Partitioning Using the RPART Routines*.

Zeileis, A., & Hothorn, T. (2015). Parties, Models, Mobsters: A New Implementation of Model-Based Recursive Partitioning in R. The comprehensive R archive network. Retrieved May, 31, 2022 from https://cran.r-project.org/web/packages/partykit/vignettes/mob.pdf.

Anexos (código fuente)

```
#Leemos las librerías que vamos a utilizar
library(readr)
library(ggplot2)
library(forecast)
library(knitr)
library(tseries)
library(seastests)
library(gridExtra)
library(tidyverse)
library(partykit)
library(MASS)
library(reporttools)
library(papeR)
library(skimr)
library(kableExtra)
library(sjPlot)
library(rpart)
library(rpart.plot)
library(ggplot2)
library(survival)

# Leemos la base de datos
datos <- read_csv("~/Desktop/TFG/datos/hotel_bookings.csv")
datos <-data.frame(datos)

nrow(datos) # 119390 registros
```

```
ncol(datos) # 31 variables

any(is.na(datos)) # vemos si hay NA
which(is.na(datos))

datos <- na.omit(datos) # eliminamos los registros con NA
any(is.na(datos)) # ya no hay NA
nrow(datos) # 119386 registros

# Resumen de los datos
summary(datos)

which(datos$adr<0) # valor de adr negativo
datos_finales <- datos[-14970,] # eliminamos la fila donde adr<0
nrow(datos_finales) # 119385 registros


# Eliminamos variables
datos_finales <- datos_finales[,-c(23,24,30,31)]

# Convertimos variables categóricas a factor
datos_finales$hotel <- factor(datos_finales$hotel)
datos_finales$is_canceled <- as.factor(datos_finales$is_canceled)
datos_finales$arrival_date_year <- factor(datos_finales$arrival_date_year)
datos_finales$arrival_date_month <- factor(datos_finales$arrival_date_month)
datos_finales$meal <- factor(datos_finales$meal)
datos_finales$country <- factor(datos_finales$country)
datos_finales$market_segment <- factor(datos_finales$market_segment)
datos_finales$distribution_channel <- factor(datos_finales$distribution_channel)
datos_finales$is_repeated_guest <- factor(datos_finales$is_repeated_guest)
datos_finales$reserved_room_type <-
factor(datos_finales$reserved_room_type)
datos_finales$assigned_room_type <-
factor(datos_finales$assigned_room_type)
datos_finales$customer_type <- factor(datos_finales$customer_type)
datos_finales$required_car_parking_spaces <-
factor(datos_finales$required_car_parking_spaces)
datos_finales$total_of_special_requests <-
factor(datos_finales$total_of_special_requests)

str(datos_finales) # comprobamos que cada variable tiene su tipo
correspondiente
```

```
# Recodificación (traducción de nombres de variables y categorías)
datos_finales <- rename(datos_finales, Hotel=hotel, Cancelación=is_canceled,
Cadencia=lead_time, Año=arrival_date_year,
                       Mes=arrival_date_month, Semana=arrival_date_week_number,
Día=arrival_date_day_of_month,
                       Fin_semana=stays_in_weekend_nights,
Entre_semana=stays_in_week_nights, Adultos=adults,
                       Niños=children, Bebés=babies, Comida=meal, País=country,
Mercado=market_segment, Canal=distribution_channel,
                       Repite=is_repeated_guest,
Cancelaciones_previas=previous_cancellations,
                       Previas_no_canceladas=previous_bookings_not_canceled,
Reservada=reserved_room_type, Asignada=assigned_room_type,
                       Cambios=booking_changes, Espera=days_in_waiting_list,
Cliente=customer_type, Gasto_medio=adr,
                       Aparcamiento=required_car_parking_spaces,
Peticiones=total_of_special_requests)

levels(datos_finales$Hotel) <- c("Ciudad", "Resort")
levels(datos_finales$Mes) <- c("Abril", "Agosto","Diciembre","Febrero",
"Enero","Julio","Junio", "Marzo","Mayo",
                               "Noviembre", "Octubre","Septiembre")
levels(datos_finales$Comida) <- c("Cama y desayuno", "Pensión
completa","Media pensión","Sin comidas","Indefinido")
levels(datos_finales$Mercado) <- c("Aviación",
"Complementario","Empresa","Directo","Grupos", "Agentes de
viajes/Operadores turísticos",
                                   "Agentes de viajes")
levels(datos_finales$Canal) <- c("Empresa","Directo","GDS", "Agentes de
viajes/Operadores turísticos", "Indefinido")
levels(datos_finales$Cliente) <- c("Contrato","Grupo","Transitorio", "Transitorio-
fiesta")


## Descriptivos numéricos por grupos (cancelación: SÍ/ NO)

### Variables numéricas
tab1.g=papeR::summarize(datos_finales, type="numeric", group="Cancelación")
kbl(tab1.g[c(1,2,4,6,7)],caption="Descriptivos de las variables numéricas por
grupos", digits = 2,
    col.names=c("Variable","Cancelación","N","Media",
"SD"),row.names=FALSE) %>%
  kable_styling(bootstrap_options = "condensed",full_width = F)
```

```
# Combinamos todas las tablas para tener toda la información en una única tabla de descriptivos numéricos
tab=tab1[,c(1,2,4,5)]
colnames(tab)=c("Variable","N","Media","SD")
tab.g=tab1.g[c(1,2,4,6,7,15)];rownames(tab.g)=NULL;colnames(tab.g)=c("Variable","Cancelación","N","Media","SD", "P.valor")
opts <- options(knitr.kable.NA = "")
kbl(caption="Descriptivos de las variables numéricas",left_join(tab.g,tab,by="Variable",suffix=c("",".global"))) %>%
  kable_styling(bootstrap_options = "condensed",full_width = F)

### Variables categóricas
# Identificación de las categóricas
des_skim=skim(datos_finales)
tipo=des_skim$skim_type # identifica el tipo de cada variable
var=des_skim$skim_variable
# identificación variables categóricas
categoricas=var[tipo=="factor"]


## PREPROCESADO


### Variable Familiar


datos_finales<-datos_finales[-which(datos_finales$Adultos==0),]

datos_finales$Familiar <- ifelse(datos_finales$Adultos==1 & datos_finales$Niños==0 & datos_finales$Bebés==0,"Solo",
                                 ifelse(datos_finales$Adultos == 2 & datos_finales$Niños==0 & datos_finales$Bebés==0,"Pareja","Familia"))

datos_finales$Familiar <- as.factor(datos_finales$Familiar)


###  Variable Habitación

#Recodificacion de asignada y reservada
levels(datos_finales$Reservada) <- c(1,2,3,4,5,6,7,8,11,12)
levels(datos_finales$Asignada) <- c(1,2,3,4,5,6,7,8,9,10,11,12)

datos_finales$Reservada <- as.numeric(datos_finales$Reservada)
datos_finales$Asignada <- as.numeric(datos_finales$Asignada)
```

```
datos_finales$Habitación <-
ifelse(datos_finales$Asignada==datos_finales$Reservada,"Igual",

ifelse(datos_finales$Asignada<datos_finales$Reservada,"Mejor","Peor"))

datos_finales$Reservada <- as.factor(datos_finales$Reservada)
datos_finales$Asignada <- as.factor(datos_finales$Asignada)
levels(datos_finales$Reservada) <- c("A", "B", "C", "D", "E", "F", "G", "H", "L", "P")
levels(datos_finales$Asignada) <- c("A", "B", "C", "D", "E", "F", "G", "H", "I", "K",
"L", "P")

datos_finales$Habitación <- as.factor(datos_finales$Habitación)

###  Variable Estancia

datos_finales<-datos_finales[-which(datos_finales$Cancelación==0 &
datos_finales$Fin_semana==0
                    & datos_finales$Entre_semana==0),] # Eliminamos los
que se hospedan y se quedan 0 noches

datos_finales$Estancia <- ifelse(datos_finales$Fin_semana!=0 &
datos_finales$Entre_semana==0, "Fin semana",
                  ifelse(datos_finales$Fin_semana==0 &
datos_finales$Entre_semana!=0, "Entre semana","Combinado"))

datos_finales$Estancia <- as.factor(datos_finales$Estancia)

### Variable Tasa cancelación

for (i in 1:nrow(datos_finales)){
  datos_finales$Tasa_cancelación[i] <-
round(sum(as.numeric(datos_finales$Cancelaciones_previas[i]))/
                        max(1,
as.numeric(datos_finales$Cancelaciones_previas[i]) +
as.numeric(datos_finales$Previas_no_canceladas[i])),2)
}


datos_finales <- datos_finales[,-c(6,7,8,9,10,11,12,14,15,18,19,21)]


## DESCRIPTIVOS VARIABLES NUMÉRICAS
```

```
tab1=papeR::summarize(datos_finales, type="numeric")
tab1.g=papeR::summarize(datos_finales, type="numeric", group="Cancelación")

tab=tab1[,c(1,2,4,5)]
colnames(tab)=c("Variable","N","Media","SD")
tab.g=tab1.g[c(1,2,4,6,7,15)];rownames(tab.g)=NULL;colnames(tab.g)=c("Variable","Cancelación","N","Media","SD", "P.valor")
opts <- options(knitr.kable.NA = "")
kbl(caption="Descriptivos de las variables numéricas",left_join(tab.g,tab,by="Variable",suffix=c("",".global"))) %>%
  kable_styling(bootstrap_options = "condensed",full_width = F)



## DESCRIPTIVOS VARIABLES CATEGÓRICAS

des_skim=skim(datos_finales)
tipo=des_skim$skim_type # identifica el tipo de cada variable
var=des_skim$skim_variable
# identificación tipos de variables
categoricas=var[tipo=="factor"]

tab2=papeR::summarize(datos_finales, type="factor")
colnames(tab2)=c("Variable","Categorías","","N","%")

kbl(caption="Descriptivos de las variables categóricas",tab2) %>%
  kable_styling(bootstrap_options = "condensed",full_width = F)



## RPART

# Modelo completo con rpart
arbol_rpart=rpart(Cancelación~Hotel+Cadencia+Año+Mes+Comida+Canal+Repite+Reservada+

Cambios+Espera+Cliente+Gasto_medio+Aparcamiento+Peticiones+

Familiar+Habitación+Reservada+Tasa_cancelación,data=datos_finales)


## decisión profundidad óptima

plotcp(arbol_rpart)
```

```
printcp(arbol_rpart)


xerror <- arbol_rpart$cptable[,"xerror"]
imin.xerror <- which.min(xerror)
# Valor óptimo
fila <- arbol_rpart$cptable[imin.xerror, ]

upper.xerror <- xerror[imin.xerror] + arbol_rpart$cptable[imin.xerror, "xstd"]
icp <- min(which(xerror <= upper.xerror))
cp <- arbol_rpart$cptable[icp, "CP"]
nsplit <- fila[2]



# Modelo óptimo con rpart (MODELO A)

arbol_rpart3=rpart(Cancelación~Hotel+Cadencia+Año+Mes+Comida+Canal+R
epite+Reservada+

Cambios+Espera+Cliente+Gasto_medio+Aparcamiento+Peticiones+

Familiar+Habitación+Reservada+Tasa_cancelación,data=datos_finales,
                   method = "class", cp=cp)

colores <- ifelse(arbol_rpart3$frame$yval==1, "lightgreen", "lightblue")
rpart.plot(arbol_rpart3, box.col = colores, cex = 0.70)

## CTREE

## 1) CTREE CON RESPUESTA CANCELACIÓN Y TODOS LOS
PREDICTORES (MODELO B)

profundidades=4:19
exitos=matrix(0,nrow=length(profundidades),ncol=4,dimnames=list(profundidad
es,c("e","p","r","f1")))
for(i in profundidades){

arbol_ctree2=ctree(Cancelación~Hotel+Cadencia+Año+Mes+Comida+Canal+R
epite+Reservada+

Cambios+Espera+Cliente+Gasto_medio+Aparcamiento+Peticiones+

Familiar+Habitación+Reservada+Tasa_cancelación,data=datos_finales,
```

```
                  control=ctree_control(maxdepth = i))
  # exactitud
  clasificacion=table(datos_finales$Cancelación,predict(arbol_ctree2))
  e=sum(diag(clasificacion))/nrow(datos_finales)
  # precision
  p=clasificacion[2,2]/sum(clasificacion[,2])
  # recuerdo
  r=clasificacion[2,2]/sum(clasificacion[2,])
  # f1
  f1=2*p*r/(p+r)
  exitos[i-3,]=round(c(e,p,r,f1)*100,2)
}
exitos
rownames(exitos)=profundidades


### Construimos el árbol con ctree y profundidad 5

arbol_ctree5=ctree(Cancelación~Hotel+Cadencia+Año+Mes+Comida+Canal+R
epite+Reservada+

Cambios+Espera+Cliente+Gasto_medio+Aparcamiento+Peticiones+

Familiar+Habitación+Reservada+Tasa_cancelación,data=datos_finales,
                control=ctree_control(maxdepth = 5))
print(arbol_ctree5)
plot(arbol_ctree5,gp = gpar(fontsize = 6), type="simple",
    terminal_panel=node_inner(arbol_ctree5, id = T, pval = F, abbreviate = F,fill =
"white", gp = gpar()))



### dibujamos la proporción para las cancelaciones observadas y las predichas

arbol_ctree5=ctree(Cancelación~Hotel+Cadencia+Año+Mes+Comida+Canal+R
epite+Reservada+

Cambios+Espera+Cliente+Gasto_medio+Aparcamiento+Peticiones+

Familiar+Habitación+Reservada+Tasa_cancelación,data=datos_finales,
                control=ctree_control(maxdepth = 5))
#print(arbol_ctree5)
plot(arbol_ctree5)
```

```
nodos=predict(arbol_ctree5,type="node")
fit=predict(arbol_ctree5,type="response")

datoss=datos_finales%>%
  mutate(nodos=as.factor(nodos),fit=fit)%>%
  rename(obs=Cancelación)%>%
  group_by(nodos)%>%
  dplyr::summarise(n=n(),obss=sum(obs=="1")/n,fits=sum(fit=="1")/n)

datoss[,c(1,3,4)] %>%
  pivot_longer(cols=c(2,3),names_to = "tipo",values_to="prop")%>%
  ggplot(aes(x=nodos,y=prop,fill=tipo))+
  geom_col(position=position_dodge(width=0.4))+
  theme_classic()+
  labs(x="Nodos",y="Proporción de cancelaciones",fill="Pred vs Obs")+
  scale_fill_hue(labels = c ('Predichas','Observadas'))
```

2) CTREE CON RESPUESTA CANCELACIÓN Y TODOS LOS PREDICTORES MENOS CADENCIA (MODELO C)

Modelo con ctree y profundidad óptima

```
arbol_ctree_sin=ctree(Cancelación~Hotel+Año+Mes+Comida+Canal+Repite+R
eservada+

Cambios+Espera+Cliente+Gasto_medio+Aparcamiento+Peticiones+

Familiar+Habitación+Reservada+Tasa_cancelación,data=datos_finales,
                  control=ctree_control(maxdepth=5))

plot(arbol_ctree_sin,gp = gpar(fontsize = 6),
    inner_panel=node_inner, type = "simple"
)
print(arbol_ctree_sin)
```

Dibujamos las curvas Kaplan-Meier

```
datos_finales$cancelacion01<-as.numeric(datos_finales$Cancelación)-1
datos_finales$nodos=as.factor(predict(arbol_ctree_sin,type="node"))
```

```
nnodos=length(levels(datos_finales$nodos))
nodos=levels(datos_finales$nodos)

sel=which(datos_finales$nodos==nodos[1])
skm=summary(survfit(Surv(Cadencia, cancelacion01) ~
1,data=datos_finales[sel,]))
datosfin=tibble(time=skm$time,surv=skm$surv,nodo=nodos[1])

for(i in 2:nnodos){
  sel=which(datos_finales$nodos==nodos[i])
  skm=summary(survfit(Surv(Cadencia, cancelacion01) ~
1,data=datos_finales[sel,]))
  pred=tibble(time=skm$time,surv=1-(skm$surv),nodo=nodos[i])

  datosfin=bind_rows(datosfin,pred)
}

ggplot(datosfin,aes(x=time,y=surv))+
  geom_step()+
  facet_wrap(vars(nodo))+
  theme_classic()+
  ylab("Cancelación")+
  xlab("Cadencia")


## 3) CTREE CON RESPUESTA SURV(CADENCIA CANCELACIÓN)
(MODELO D)

#Creamos una nueva variable que es la variable Cancelación pero de tipo
numérico

is.factor(datos_finales$Cancelación)
head(datos_finales$Cancelación)
as.numeric(datos_finales$Cancelación)-1

datos_finales$Cancelación_num <- as.numeric(datos_finales$Cancelación)-1



### Supervivencia con ctree profundidad 3

survival <- ctree(Surv(Cadencia,Cancelación_num) ~
Hotel+Año+Mes+Comida+Canal+Repite+Reservada+
```

```
Cambios+Espera+Cliente+Gasto_medio+Aparcamiento+Peticiones+

Familiar+Habitación+Estancia+Reservada+Tasa_cancelación,data=
                datos_finales,control = ctree_control(maxdepth =3))

print(survival)
plot(survival,gp = gpar(fontsize = 10), type= "simple")


### dibujamos la proporción para las cancelaciones observadas

nodos3=predict(survival,type="node")

datoss=datos_finales%>%
  mutate(nodos3=as.factor(nodos3))%>%
  rename(obs=Cancelación)%>%
  group_by(nodos3)%>%
  dplyr::summarise(n=n(),obss=sum(obs=="1")/n)

datoss[,c(1,3)] %>%
  pivot_longer(cols=c(2),names_to = "tipo",values_to="prop")%>%
  ggplot(aes(x=nodos3,y=prop,fill=tipo))+
  geom_col(position=position_dodge(width=0.4))+
  labs(x="Nodos",y="Proporción de cancelaciones")+
  theme_classic()+
  theme(legend.position='none')



### Dibujamos las curvas Kaplan-Meier

datos_finales$nodos=as.factor(predict(survival,type="node"))

nnodos=length(levels(datos_finales$nodos))
nodos=levels(datos_finales$nodos)

sel=which(datos_finales$nodos==nodos[1])
skm=summary(survfit(Surv(Cadencia,                cancelacion01)            ~
1,data=datos_finales[sel,]))
datosfin=tibble(time=skm$time,surv=1-skm$surv,nodo=nodos[1])
for(i in 2:nodos){
  sel=which(datos_finales$nodos==nodos[i])
```

```
  skm=summary(survfit(Surv(Cadencia, cancelacion01) ~
1,data=datos_finales[sel,]))
  pred=tibble(time=skm$time,surv=skm$surv,nodo=nodos[i])

  datosfin=bind_rows(datosfin,pred)
}


## 4) GLMTREE CON RESPUESTA CANCELACIÓN Y TODOS
PREDICTORES(MODELO E)

### glmtree con profundidad óptima

survival_mob <-
glmtree(Cancelación~Hotel+Año+Cadencia+Mes+Comida+Canal+Repite+Rese
rvada+

Cambios+Espera+Cliente+Gasto_medio+Aparcamiento+Peticiones+Familiar+
                Habitación+Estancia+Reservada+Tasa_cancelación, data
=datos_finales,
                family = binomial, maxdepth = 5)

plot(survival_mob,gp = gpar(fontsize = 10), type = "simple")
print(survival_mob)



### dibujamos la proporción para las cancelaciones observadas

nodos2=predict(survival_mob,type="node")

datoss=datos_finales%>%
  mutate(nodos2=as.factor(nodos2))%>%
  rename(obs=Cancelación)%>%
  group_by(nodos2)%>%
  dplyr::summarise(n=n(),obss=sum(obs=="1")/n)

datoss[,c(1,3)] %>%
  pivot_longer(cols=c(2),names_to = "tipo",values_to="prop")%>%
  ggplot(aes(x=nodos2,y=prop,fill=tipo))+
  geom_col(position=position_dodge(width=0.4))+
  labs(x="Nodos",y="Proporción de cancelaciones")+
  theme_classic()+
  theme(legend.position='none')
```

5) GLMTREE CON RESPUESTA CANCELACIÓN Y TODOS PREDICTORES MENOS CADENCIA (MODELO E)

glmtree con profundidad óptima

```
survival_mob_sin <- glmtree(Cancelación~Hotel+Año+Mes+Comida+Canal+Repite+Reservada+Cambios+Espera+Cliente+Gasto_medio+Aparcamiento+Peticiones+Familiar+Habitación+Estancia+Reservada+Tasa_cancelación, data =datos_finales, family = binomial, prune = "BIC")

plot(survival_mob_sin, gp = gpar(fontsize = 3), type= "simple")
print(survival_mob_sin)
```

dibujamos la proporción para las cancelaciones observadas

```
nodos4=predict(survival_mob_sin,type="node")

datoss=datos_finales%>%
  mutate(nodos4=as.factor(nodos4))%>%
  rename(obs=Cancelación)%>%
  group_by(nodos4)%>%
  dplyr::summarise(n=n(),obss=sum(obs=="1")/n)

datoss[,c(1,3)] %>%
  pivot_longer(cols=c(2),names_to = "tipo",values_to="prop")%>%
  ggplot(aes(x=nodos4,y=prop,fill=tipo))+
  geom_col(position=position_dodge(width=0.4))+
  labs(x="Nodos",y="Proporción de cancelaciones")+
  theme_classic()+
  theme(legend.position='none')
```

Dibujamos las curvas Kaplan-Meier

```
datos_finales$nodos=as.factor(predict(survival_mob_sin,type="node"))

nnodos=length(levels(datos_finales$nodos))
nodos=levels(datos_finales$nodos)

sel=which(datos_finales$nodos==nodos[1])
```

```
skm=summary(survfit(Surv(Cadencia, cancelacion01) ~
1,data=datos_finales[sel,]))
datosfin=tibble(time=skm$time,surv=1-skm$surv,nodo=nodos[1])
for(i in 2:nodos){
  sel=which(datos_finales$nodos==nodos[i])
  skm=summary(survfit(Surv(Cadencia, cancelacion01) ~
1,data=datos_finales[sel,]))
  pred=tibble(time=skm$time,surv=skm$surv,nodo=nodos[i])

  datosfin=bind_rows(datosfin,pred)
}
```