Taller 1 - Ciencia de Datos Aplicada

Erich Giusseppe Soto Parada y Nicolás Klopstock Triana

Septiembre 2025

1 Entendimiento de datos

El dataset utilizado consta de datos sobre información de reservas para dos modalidades de hoteles: Ciudad y Resort. Sobre estos datos crudos, pudimos hacer los siguientes análisis luego de un reporte de *profile*:

El conjunto está fuertemente dominado por Resort Hotel, así que cualquier patrón agregado tenderá a reflejar el comportamiento de ese hotel. Por eso, cuando se analice algo sensible a mezcla de poblaciones, conviene estratificar por hotel. La variable objetivo is_canceled presenta más 0 que 1 (aprox. 0.70 a 1), lo que implica tener presente el desbalance a la hora de interpretar tasas y correlaciones. lead_time exhibe un sesgo pronunciado; para estabilizar su escala y acercarla a una forma más normal, se aplica una transformación logarítmica (log1p), en caso que se requiera normalizar, dejando abierta la puerta a probar Box-Cox si se requiere. En arrival_date_year hay 614 registros con el valor "20016" que se corrige a 2016 por tratarse de un error tipográfico. A nivel estacional, agosto concentra más reservas (arrival_date_month), mientras que arrival_date_week_number se ve más uniforme con algunos atípicos y arrival_date_day_of_month luce prácticamente uniforme; todas se conservan para capturar estacionalidad y calendarios.

En duración de la estancia, se concentra el análisis en los casos típicos: para stays_in_weekend_nights se filtra a 0–2 noches y para stays_in_week_nights se eliminan valores $\geq$ 5 noches (se pierde $\sim 6.9\%$ de los datos), privilegiando representatividad sobre rarezas operativas (las estancias largas existen, pero distorsionan promedios y no son el foco). children y babies se eliminan porque están extremadamente desbalanceadas (casi todo 0) y, sobre todo, porque no resultan variables accionables para "pegarle" a mayor ocupación en el corto plazo: saber que casi nadie viaja con bebés o niños no ofrece una palanca clara para reducir cancelaciones o incrementar ocupación; en cambio, segmentaciones y canales sí habilitan acciones comerciales. Similar razonamiento aplica a meal: aunque pudiera haber señales débiles con cancelación, no se considera clave ahora y se descarta para simplificar. country se conserva para entender procedencias, útil como capa contextual.

market_segment y distribution_channel se mantienen porque permiten identificar segmentos y canales con mayor o menor propensión a cancelar, abriendo

la posibilidad de incentivar los de menor cancelación o revisar políticas en los de mayor cancelación. is_repeated_guest se descarta por desbalance extremo (poca variabilidad útil). previous_cancellations y previous_bookings_not _canceled se eliminan por su gran proporción de nulos/cero, que los vuelve poco informativos a nivel agregado. Se analizan conjuntamente reserved_room_type y assigned_room_type introduciendo un indicador de desajuste (cuando lo asignado difiere de lo reservado), con la hipótesis de que un mismatch podría relacionarse con cancelaciones, sin imponer ordinalidad inexistente. booking_changes se conserva (pese a muchos nulos) para evaluar si un mayor número de cambios se asocia a cancelar.

deposit_type es de las variables más relevantes desde el negocio: pese al desbalance (predomina "No Deposit"), vale la pena estudiar sus tasas de cancelación, aunque habría cautela al usarla en modelos. agent no se ve accionable ahora y se elimina; lo mismo company y days_in_waiting_list por su masividad de ceros. customer_type se mantiene para explorar si ciertas categorías muestran mayor propensión a cancelar. En adr, se eliminan outliers altos (para evitar que casos extremos gobiernen los resultados) y se estudia su relación con is_canceled; además se revisa la frecuencia de adr == 0 por estado de cancelación y se construye un ingreso esperado simple como adr * (stays_in_week_nights + stays_in_weekend_nights) para entender la pérdida potencial. required_car_parking_spaces y total_of_special_requests se descartan: ambas están muy desbalanceadas y no ofrecen una palanca clara para reducir cancelaciones. Finalmente, reservation_status se elimina por redundancia con is_canceled y por la categoría "No-Show" que introduce ambigüedad temporal; antes de soltar la columna se remueven las filas "No-Show". reservation_status_date también se elimina para evitar inconsistencias asociadas a ese estatus. El resto de variables relevantes se conserva, incluyendo kids (no se especificó acción sobre ella).

2 Estrategia de análisis

Este análisis va a estar compuesto de dos frentes. El primero es descriptivo–exploratorio y se apoya en estadísticos básicos y correlaciones de Spearman entre is_canceled y las variables que se quieran observar. Elegimos Spearman porque capta asociaciones monótonas sin exigir linealidad, lo que es adecuado para relaciones potencialmente no lineales con cancelaciones. En paralelo, se calculará el rate de cancelación por grupo: para cada categoría de interés (por ejemplo, segmentos, canales, tipos de cliente, tipos de depósito, hotel o el desajuste entre habitación reservada y asignada), se estima la proporción de reservas canceladas dentro de cada categoría, con el fin de identificar de manera directa dónde se concentran las mayores tasas y dónde hay oportunidades de intervención. Para facilitar la lectura y la comunicación, se utilizarán visualizaciones sencillas pero efectivas: diagramas de barras comparando tasas, heatmaps para resumir relaciones entre variables y cancelaciones, y gráficas temporales de cancelaciones para detectar de forma preliminar la presencia de patrones esta-

cionales o tendencias. El segundo, por su lado, también está dividido en dos enfoques diferentes. El primero es construir un modelo de árboles con el objetivo de obtener, de forma discriminativa y con sustento estadístoc, una lectura rápida de las cinco variables potencialmente más influyentes y siendo una propuesta para identificar posibles clientes que van a cancelar una reserva hechos y tomar acción temprana. De la misma, forma el segundo enfoque también busca identificar perfiles de alto riesgo de cancelación y proponer estrategias para reducir esas pérdidas; esta vez con un análisis de clustering considerando características y tipo de depósito. Esto nos ayudará a segmentar a los clientes en grupos con comportamientos similares y calcular un acercamiento a un valor de tasa de cancelación por cluster.

3 Desarrollo de la estrategia

3.1 Descripción y Exploración

Para esta parte se realizará una exploración *variable por variable* para identificar patrones, detectar comportamientos anómalos y proponer focos de revisión a partir de las gráficas.

Importante En relación con las variables seleccionadas para el análisis —que inicialmente deberían ser cinco—, se optó por incluir todas aquellas consideradas relevantes, las cuales se presentan en el análisis. Como se observará, se utilizaron todas excepto room_mistmatch, dado que mostró un comportamiento poco diferenciador y un fuerte desbalance.

Figure 1: Distribución de la variable *is canceled* (balance general).

Balance general. En la Fig. 1 se observa que la proporción de cancelaciones es cercana a ≈ 0.41 del total. Con este punto de partida, conviene priorizar los casos donde se combinan **tasa alta** y **volumen** significativo.

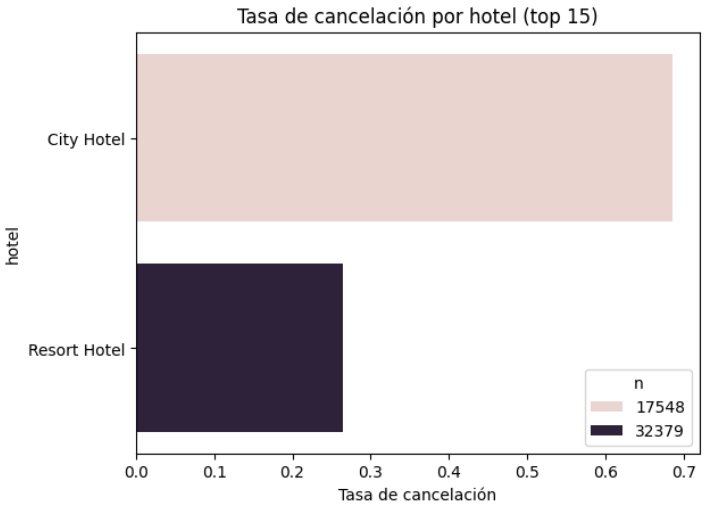
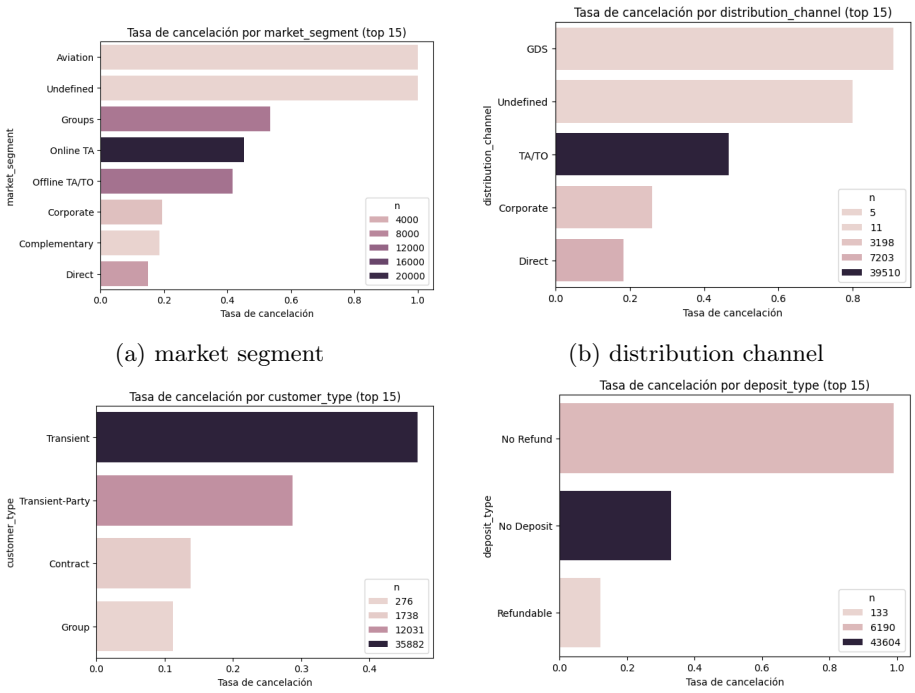


Figure 2: Tasa de cancelación por tipo de hotel.

Hoteles. La Fig. 2 muestra que City Hotel presenta una tasa de cancelación elevada (≈ 0.686) con $n = 17{,}548$, mientras que Resort Hotel es sensiblemente menor (≈ 0.264; $n = 32{,}379$). Dado el peso relativo de City, vale la pena revisar qué está elevando su riesgo (mezcla de canales, política comercial, experiencia del huésped) y priorizar acciones allí.



(a) market segment

(b) distribution channel

(c) customer type

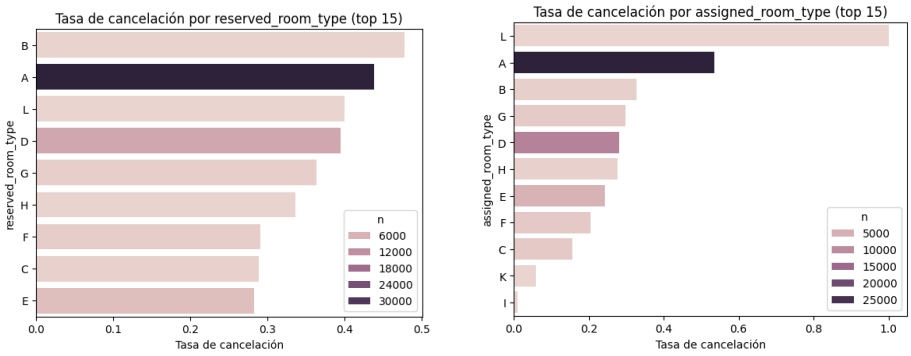
(d) deposit type

Figure 3: Tasa de cancelación por *market segment*, *distribution channel*, *customer type* y *deposit type*.

Canales y segmentos. Las visualizaciones de *market segment* y *distribution channel* (Fig. 3, subfigs. 3a y 3b) indican que Online TA (≈ 0.452; $n = 21{,}339$), TA/TO offline (≈ 0.416; $n = 10{,}410$) y el canal TA/TO (≈ 0.466; $n = 39{,}510$) concentran buena parte del problema. Toca revisar cómo reducir cancelaciones en estos frentes (por ejemplo, mejor reconfirmación e incentivos a mantener la reserva), manteniendo como referencia de menor riesgo a Direct (canal ≈ 0.184; $n = 7{,}203$) y Corporate (segmento ≈ 0.194; $n = 2{,}214$).

Tipo de cliente. En la Fig. 3 (subfig. 3c) se ve que Transient combina tasa alta y mucho volumen (≈ 0.470; $n = 35{,}882$). Conviene analizar qué rasgos de estas reservas explican la deserción (anticipación, canal, precio) y qué prácticas ayudan a reducir su cancelación. Transient-Party queda en riesgo medio (≈ 0.287; $n = 12{,}031$); Contract y Group tienen tasas bajas y menor peso.

Política de depósito. La Fig. 3 (subfig. 3d) muestra que No Refund registra una tasa inusualmente alta (≈ 0.989; $n = 6{,}190$). Antes de extraer conclusiones, es necesario validar su definición y uso en los datos. En paralelo, el grueso del volumen está en No Deposit (≈ 0.331; $n = 43{,}604$), donde vale la pena revisar procesos y comunicaciones para disminuir cancelaciones.



(a) reserved room type (b) assigned room type

Figure 4: Tasa de cancelación por tipo de habitación (reservada y asignada).

Habitaciones. En la Fig. 4 el tipo A concentra volumen y una tasa elevada tanto en *reserved room type* (≈ 0.438; $n = 35{,}081$) como en *assigned room type* (≈ 0.534; $n = 28{,}065$). Toca revisar si hay temas de precio, disponibilidad o expectativas que expliquen estas diferencias. Nota: *assigned room type* se define cerca del check-in; es útil para operación, pero no para predecir al momento de la reserva.

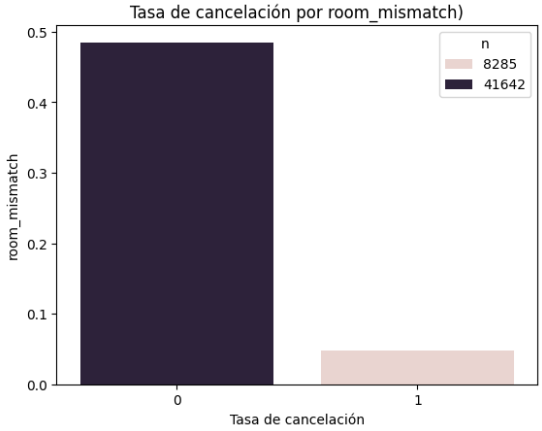


Figure 5: Tasa de cancelación cuando hay desajuste entre tipo de cuarto (*room mismatch*).

Otras señales. En *room mismatch* (Fig. 5) aparece un patrón atípico: menor cancelación con desajuste declarado (≈ 0.048; $n = 8{,}285$). Esto sugiere revisar definiciones y el momento de captura de la variable. Por ingresos, las reservas que cancelan concentran más *expected revenue* (media ≈ 330 vs ≈ 283), por lo que el impacto económico se sesga a reservas de mayor valor (tambien puede

ser por la hipotesis sugerida arriba y es que puede ser el cuarto asignado al momento de llegar al hotel en cuyo caso no serviria mucho esta variable).

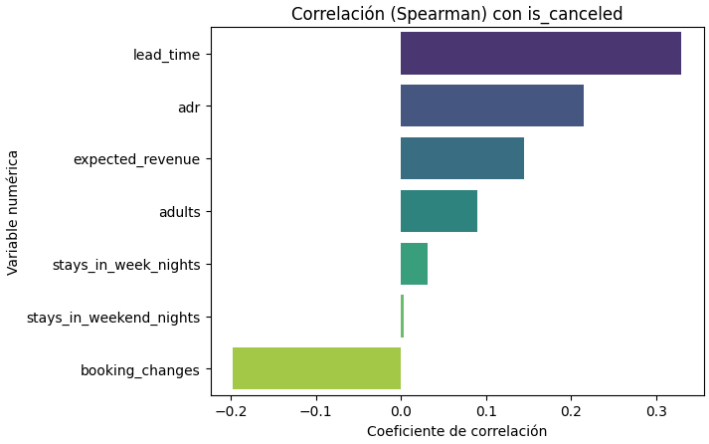


Figure 6: Correlaciones de Spearman: ninguna asociación individual alcanza valores altos.

Correlaciones. La matriz de Spearman (Fig. 6) indica asociaciones *moderadas* con *is canceled* (por ejemplo, $\rho_{\text{lead time}} \approx 0.33$, $\rho_{\text{adr}} \approx 0.21$) y ninguna cercana a 0.8. Esto implica que no hay relaciones monótonas fuertes por variable individual; lo relevante está en combinaciones (p. ej., lead time $\times$ canal $\times$ hotel). Con base en ello, el siguiente paso es profundizar en los focos con alta tasa y alto n (City Hotel, TA/TO y Online TA, cliente Transient, tipos de habitación A) para identificar causas y definir acciones concretas.

3.2 Estrategia - Árboles de Decisión

Para la estrategia basada en árboles de decisión se construyó un modelo orientado a identificar las variables y particiones más relevantes para predecir la cancelación de reservas. La selección de divisiones internas se realizó con la métrica de impureza Gini, habitual en este tipo de modelos, y la evaluación externa se hizo con métricas de desempeño en el conjunto de prueba. El análisis se organiza en dos etapas: primero, la interpretación de las particiones del árbol y de las reglas resultantes; después, la discusión de la relevancia de las variables y de la utilidad predictiva global.

Table 1: Métricas de test

Clase / Métrica	Precisión	Recall	F1-score	Soporte
0	0.8099	0.8671	0.8375	5868
1	0.7894	0.7101	0.7476	4118
Exactitud			0.8023	9986
Promedio macro	0.7997	0.7886	0.7926	9986
Promedio ponderado	0.8015	0.8023	0.8005	9986

En el conjunto de prueba, el modelo alcanza una exactitud de 0.8023 y un ROC AUC de 0.8760, lo que indica buena capacidad discriminativa. La clase 0 (no cancelación) presenta *precision* 0.8099, *recall* 0.8671 y *f1* 0.8375 con 5,868 observaciones; la clase 1 (cancelación) muestra *precision* 0.7894, *recall* 0.7101 y *f1* 0.7476 con 4,118 observaciones. En términos agregados, el rendimiento es sólido, con un leve sesgo hacia la clase mayoritaria, pero con una separación entre clases consistente según el AUC.

Table 2: Top 10 variables más importantes

Feature	Variable	Importancia
32	deposit_type_No Refund	0.388456
0	hotel_City Hotel	0.222098
53	lead_time	0.123974
29	customer_type_Transient	0.084630
20	market_segment_Online TA	0.057623
57	adr	0.045801
43	assigned_room_type_A	0.036006
19	market_segment_Offline TA/TO	0.013480
58	booking_changes	0.011233
34	reserved_room_type_A	0.007323

Las importancias del modelo confirman que la política de depósito "No Refund" es el principal determinante del riesgo (0.388), seguida por el tipo de hotel "City Hotel" (0.222) y la antelación de la reserva, o lead_time (0.124). En un segundo nivel aparecen el tipo de cliente "Transient" (0.085), el segmento de mercado "Online TA" (0.058) y el precio promedio diario ADR (0.046), junto con variables de configuración como la habitación asignada tipo A (0.036), el segmento "Offline TA/TO" (0.013), los cambios de reserva (0.011) y la habitación reservada tipo A (0.007). En conjunto, estos resultados sugieren que la combinación de política de depósito, contexto del hotel (urbano) y antelación, modulada por el segmento y el precio, explica una fracción sustantiva de la propensión a cancelar.

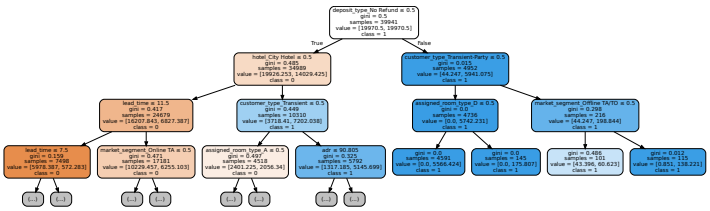


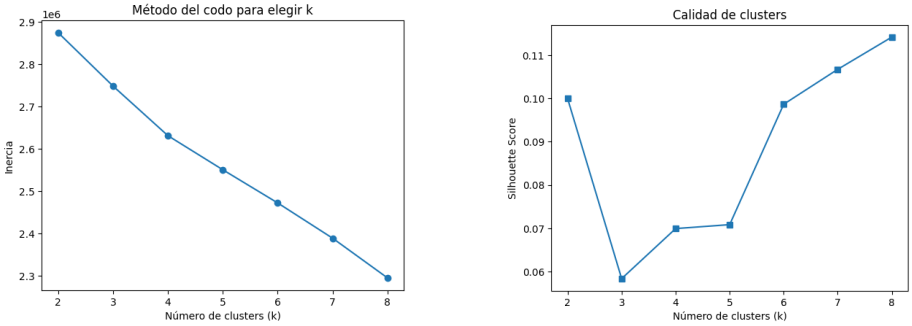
Figure 7: Estructura y particiones del arbol.

La estructura del árbol refuerza esta lectura. En las primeras capas, cuando la variable deposit_type_No Refund es 0 (es decir, no aplica un depósito no reembolsable), la siguiente partición clave es hotel_City Hotel; dentro de ese ámbito, el lead_time y el segmento market_segment_Online TA organizan ramas que diferencian con claridad comportamientos de cancelación. En cambio, cuando deposit_type_No Refund es 1, varias ramas convergen rápidamente en la clase de cancelación, con matices aportados por customer_type_Transient-Party, assigned_room_type_D y market_segment_Offline TA/TO. En términos prácticos, el nodo raíz y sus primeras divisiones ya entregan reglas útiles para priorizar casos de riesgo.

3.3 Estrategia - Clústers

Operativamente, las reglas permiten orientar decisiones. Para reservas sin "No Refund", conviene vigilar especialmente las realizadas en hoteles urbanos con determinada antelación y canal online, donde se observan incrementos de riesgo. Bajo política "No Refund", aunque el sesgo hacia la cancelación es alto, la combinación de tipo de cliente y asignación de habitación introduce gradientes que pueden explotarse para confirmaciones reforzadas, comunicación segmentada o estrategias de *overbooking* prudentes. Dado el AUC de 0.8760 y el *recall* de la clase 1 de 0.7101, el modelo es apto para cribado temprano; si el objetivo fuese captar un mayor número de cancelaciones, podrían ajustarse los umbrales de decisión, relajar la poda o explorar ensamblados (p.,ej., Random Forest o Gradient Boosting) manteniendo la interpretabilidad mediante extracción de reglas.

El análisis combinado del método del codo y del coeficiente de silueta sugiere que usar $k = 4$ clústers es una elección razonable cuando buscamos mayor segmentación sin perder demasiada calidad. En este caso, lo que nos interesa no es tanto la máxima compactación o separación absoluta, sino obtener un nivel de detalle que permita diferenciar subgrupos de clientes con patrones diferenciados.



(a) Método del codo (b) Coeficiente de silueta

Figure 8: Criterios para elegir k en *K-Means*.

La proyección bidimensional (sacada usando PCA) muestra claramente cuatro clústers definidos, entre los cuales uno sobresale por encontrarse significativamente separado de los demás. Este grupo aislado puede entenderse como un clúster con comportamiento atípico, mientras que los otros tres parecen compartir un "clúster padre" más amplio, dentro del cual se subdividen en patrones específicos.

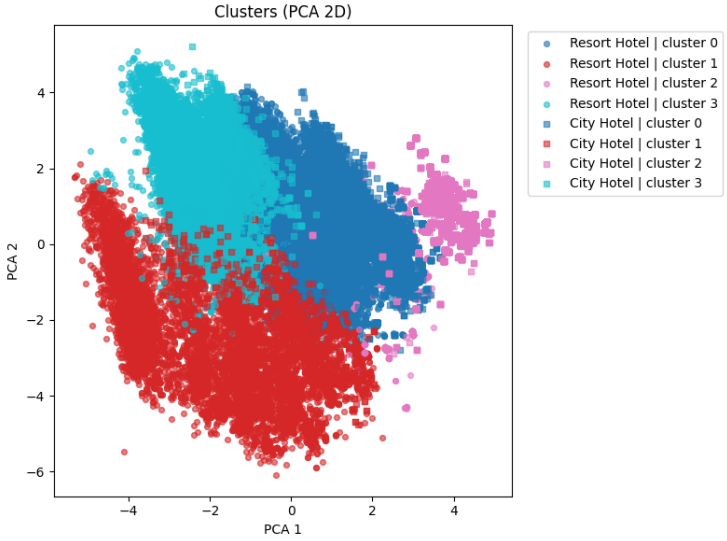


Figure 9: Resultados Clústers con K-Means

El clúster separado (clúster 2 en la gráfica) resulta particularmente crítico porque muestra una tasa promedio de cancelación de 0.99, es decir, casi todas sus reservas terminan canceladas. Este tiene un $n = 6137$.

Este clúster se caracteriza por un lead time extremadamente alto (207.39

días en promedio antes de la fecha de llegada), combinado con un ADR inferior al promedio (86.61), lo que muestra que son reservas hechas con muchísima anticipación y a tarifas relativamente bajas. El número de cambios de reserva es prácticamente nulo (0.01 en promedio), lo que refuerza la idea de que no existe un proceso de ajuste o replanificación en este grupo, sino cancelaciones directas. Aproximadamente un 75% de estas reservas corresponden a hoteles de ciudad y un 25% a resorts. Este clúster también se distingue por su uso predominante del segmento de mercado "Offline TA/TO", con un canal de distribución TA/TO en el 92% de los casos. El tipo de cliente registrado es casi siempre Transient. En cuanto a la política de depósitos, el patrón es curioso: el 100% de estas reservas están bajo la categoría "No Refund", lo cual contrasta con su altísima tasa de cancelación (*¿por qué cancelaría si no se le devuelve ningún dinero?*). Finalmente, el comportamiento respecto a las habitaciones es muy homogéneo: el 97% de los casos se reservó y fue asignada la habitación tipo A, lo que sugiere que la oferta de este tipo de habitación, en combinación con las condiciones de depósito, podría estar relacionada con la propensión extrema a la cancelación.

En contraste, los otros tres clústers presentan tasas de cancelación más bajas y comportamientos más heterogéneos ($n_{c_0} = 22927, n_{c_1} = 9837, n_{c_3} = 11026$). Entre ellos, destaca el clúster 1, cuya tasa de cancelación es apenas de 0.15, mostrando así un grupo de huéspedes con alta probabilidad de llevar a cabo su reserva. Los clientes en este clúster reservan con un lead time promedio de alrededor de mes y medio antes de la llegada; se concentran fuertemente en resorts (90%), y muestra un peso importante de Online TA (56%) con canal de distribución casi exclusivo en TA/TO (0.99). El tipo de cliente vuelve a ser Transient, y en todos los casos el depósito es No Deposit (1.0). A diferencia del clúster 2, aquí no existe una preferencia marcada por un tipo de habitación, ya que la asignación se distribuye de forma casi equitativa entre las distintas categorías.

Realmente se pueden ver las condiciones de ventas opuestas entre ambos gran clústers. Por un lado se puede que que el hecho de "No Deposit" o "No Refunds" no evita cancelaciones. Un cambio signficativo puede ser empujar más el uso de Online TA en lugar del offline. En este último suele haber menos control directo con el cliente, menor comunicación clara. El hecho de reservar con más o menos 7 meses de antelación implica "reservas especulativas". Se podrían introducir condiciones específicas para reservas con *lead time* muy alto; algún tipo de precios cambiantes, penalidades suevas o beneficios adicionales para los que efectivamente mantengan la reserva.

4 Propuestas sobre los modelos en producción

El análisis basado en clústers y el modelo de árboles de decisión se complementan de manera natural, porque tienen un *approach* al problema de la cancelación desde perspectivas distintas pero convergentes. Clustering permite descubrir perfiles de clientes y reservas que se identifican de manera no supervisada, destacando segmentos extremos y contrastando patrones de comportamiento dentro

de la población. A su vez, los árboles de decisión ofrecen un marco supervisado y jerárquico, capaz de identificar las variables más influyentes y traducirlas en reglas claras de clasificación y predicción. Mientras el enfoque de clústers ayuda a caracterizar poblaciones y visualizar heterogeneidades que pueden orientar la segmentación estratégica, el modelo de árboles aporta interpretabilidad y priorización de factores de riesgo, facilitando la toma de decisiones operativas. En conjunto, la estrategia combinada no solo valida hallazgos desde ángulos distintos, sino que también permite diseñar intervenciones más sólidas: por un lado, dirigiendo acciones específicas hacia segmentos de alto riesgo identificados en los clústers, y por otro, utilizando las reglas derivadas del árbol para implementar políticas preventivas y criterios de gestión automatizada. Por otro lado, ambos modelos pueden contribuir a clasificar las reservas y gestionarlas en función de su probabilidad de cancelación, lo cual es un enfoque común en la mayoría de los modelos de clasificación. Esto permitiría ofrecer incentivos específicos a los usuarios con mayor riesgo de cancelar, con el fin de reducir dicha probabilidad. En el caso del modelo de clustering, es posible aplicar un enfoque semisupervisado, estimando probabilidades de cancelación a partir de la proporción de cancelaciones observadas dentro de cada grupo. Sin embargo, dado que en este escenario se dispone de etiquetas, resulta más recomendable emplear modelos supervisados, como los basados en árboles de decisión, que permiten definir con mayor precisión las probabilidades de cancelación.

Cuando se identifique un cliente con alta probabilidad de cancelar o que es perteneciente a un grupo de alto riesgo, la empresa puede intervenir con medidas específicas: introducir tarifas dinámicas o escalonadas que incentiven la permanencia de la reserva y/o diseñar beneficios asociados a mantener la compra después de cierto tiempo específico (para reservas hechas con muchos días de antelación). Al mismo tiempo, se recomienda empujar el uso de canales en línea frente al offline, ya que los primeros permiten comunicación directa, personalización de ofertas y seguimiento más efectivo. El árbol de decisión, al dar reglas claras, puede incorporarse a sistemas de alerta temprana: cuando se cumplan combinaciones de variables de alto riesgo (ej. depósito “No Refund” + City Hotel + lead time gigante), el sistema podría activar automáticamente campañas de retención, ajustes en la política de habitación o incentivos comerciales. De esta forma, los clústers aportan la segmentación de los grupos a intervenir y el árbol entrega las reglas operativas para actuar en tiempo real, configurando una estrategia integral que busca reducir sustancialmente la tasa de cancelación en segmentos críticos.

Para implementar las recomendaciones se sugiere emplear A/B testing, dividiendo la muestra en dos grupos: uno de control y otro experimental (con la "cura", es decir, promociones, incentivos u otros proyectos). Este enfoque permitirá evaluar la efectividad de las acciones propuestas mediante un análisis estadístico riguroso.

4.1 Resumen de hallazgos y acciones prioritarias

Qué vimos El descriptivo muestra concentración de riesgo en *City Hotel* (Fig. 2) y en canales/segmentos *Online TA* y *TA/TO* (Fig. 3), con señales numéricas donde *lead_ time* y *adr* se asocian con mayor probabilidad de cancelación (Fig. 6). El *clustering* identifica un grupo extremo (tasa cercana a 1.0 y n relevante) que combina *lead_ time* muy alto, *TA/TO* y *No Refund*. El árbol confirma reglas simples y accionables: las particiones por *deposit_ type_ No Refund*, *hotel_ City Hotel*, *lead_ time* y *market_ segment_ Online TA* capturan buena parte del riesgo.

Acciones inmediatas (dónde intervenir).

- **City Hotel** (Fig. 2): investigar causas específicas del mayor porcentaje de cancelación (mix de canal, pricing, experiencia); priorizar acciones allí por su combinación de tasa alta y volumen.
- **Canales/segmentos de alto aporte al problema** (Fig. 3): enfocar revisión en *Online TA* y *TA/TO* (proporción y volumen altos). Objetivo: entender por qué cancelan más y qué ajustes de proceso/condiciones reducen la deserción.
- **Clientes *Transient*** (Fig. 3): analizar por qué concentran mayor cancelación (anticipación, canal, precio) y documentar prácticas que funcionan en subgrupos de menor riesgo.
- **Reservas con *lead_ time* largo y *adr* alto** (Fig. 6): priorizar su seguimiento por mayor propensión e impacto económico; definir umbrales operativos de atención temprana basados en estas dos señales.
- ***Room types* con mayor peso en el riesgo** (Fig. 4): revisar expectativas y pricing del tipo A (tanto reservado como asignado); verificar si hay desalineación entre promesa y experiencia.

Acciones de calidad de datos y gobernanza.

- **Auditar *deposit_ type_ No Refund***: la tasa observada es contraintuitiva; validar significado, momento de registro y consistencia histórica antes de usarla para decisiones.
- **Revisar variables cercanas a operación tardía** (p. ej., *assigned_ room_ type*, *room_ mismatch*; Fig. 4 y Fig. 5): confirmar su semántica y evitar *leakage* en modelos que predicen al momento de la reserva.
- **Etiquetas y categorías** (p. ej., unificación de "No Refund/Non Refund", "Online TA/TA"): normalizar nomenclaturas para análisis y despliegues consistentes.

Hallazgos de clústers (grupos críticos y rasgos operativos). El análisis del codo y de la silueta (Figs. 8) sugiere $k = 4$; la proyección bidimensional (Fig. 9) hace visible un *grupo crítico* claramente separado del resto. Este clúster concentra una **tasa de cancelación cercana a 1.0** con n relevante e integra un patrón operativo muy definido: *lead_ time* **muy alto** (reservas realizadas con mucha anticipación), *adr* por debajo del promedio, predominio de *TA/TO* (offline) como canal, cliente *Transient* y presencia total de *No Refund*. Esta combinación sugiere **reservas especulativas de largo plazo** en intermediación offline con condiciones de depósito que, al menos en los datos, no están disuadiendo la cancelación. Recomendación: (i) trazar el flujo end-to-end de ese segmento (origen de demanda, acuerdos comerciales, mensajes y tiempos de reconfirmación); (ii) revisar la *semántica* y consistencia de *No Refund*; (iii) definir reglas de seguimiento y comunicación para *lead_ time* muy largos en TA/TO (reconfirmaciones programadas, cambios de condiciones con anticipación suficiente); y (iv) evaluar ajustes de *pricing* y beneficios condicionados al *no show/cancelación*. Los otros tres clústers muestran **tasas moderadas/bajas** con perfiles más heterogéneos (por ejemplo, uno fuertemente asociado a *Resort*, *Online TA* y *No Deposit*, con cancelación baja ≈ 0.15): estos sirven como **contraste operativo** para identificar prácticas que *sí* reducen cancelación (anticipación más corta, mejor comunicación y condiciones más claras) y llevarlas, cuando sea pertinente, a los grupos de mayor riesgo.

Hallazgos del árbol (variables clave, umbrales y relevancia). El modelo supervisado (Fig. 1 y Fig. 2) confirma pocas palancas, pero muy influyentes, y las organiza en reglas operables. El primer corte es *deposit_ type_ No Refund*; cuando está presente, múltiples ramas convergen a cancelación, lo que obliga a auditar su definición y a tratar esa señal con cautela hasta validar su semántica. Despues de *No Refund*, el árbol usa *hotel_ City Hotel*, *lead_ time* y *market_ segment_ Online TA* para separar riesgo: aparecen umbrales prácticos (p. ej., *lead_ time* en torno a 10–12 días en primeras capas) que diferencian claramente la propensión. Además, la relevancia de *customer_ type_ Transient* y *adr* indica que anticipación + canal + precio forman un triángulo crítico: reservas de *lead_ time* largo, en *City Hotel* y *Online TA*, con *adr* relativamente alto, deben entrar a listas de seguimiento temprano. En síntesis, el árbol entrega: (i) variables clave (*deposit_ type_ No Refund*, *hotel_ City Hotel*, *lead_ time*, *market_ segment_ Online TA*, *adr*); (ii) umbrales de decisión para *lead_ time* y combinaciones canal×hotel; y (iii) reglas que pueden traducirse en alertas y priorización operativa (por ejemplo: "si *City Hotel* & *Online TA* & *lead_ time* > umbral, entonces reconfirmar y monitorear precio/beneficios").