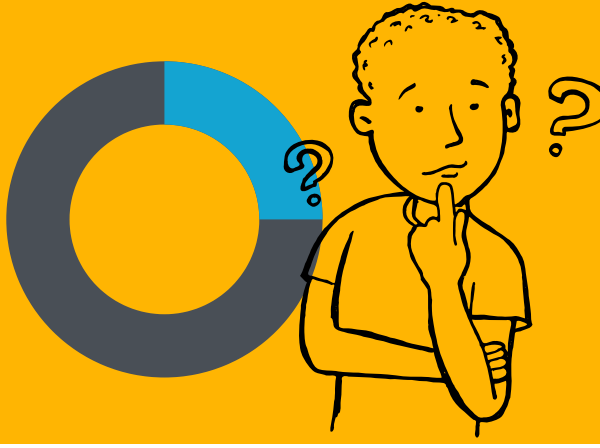
MACHINE LEARNING

Chekin

CRISTINA CEJAS SÁNCHEZ

¿QUÉ ES CHEKIN?



Chekin es un software que permite automatizar todo el proceso de registro de huéspedes, desde la confirmación de la reserva hasta el check-out.

Adaptado a todo tipo de alojamientos turísticos. Cuenta con funcionalidades avanzadas que le permiten ahorrar tiempo al cliente.

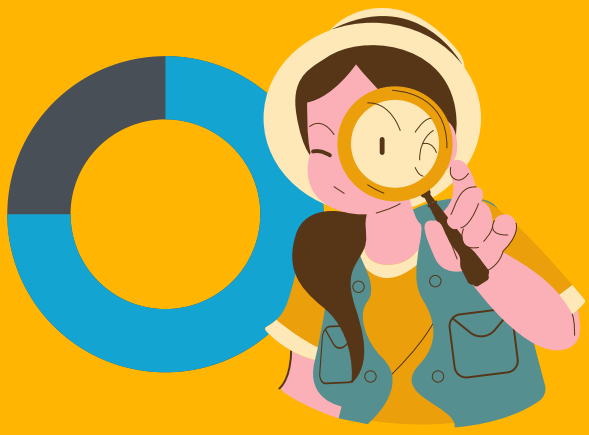
¿QUÉ VAMOS A PREDECIR?

El objetivo de este proyecto será estudiar la tasa de abandono de los clientes de Chekin para intentar predecir qué clientes son los que se han **dado de baja** durante el tiempo de vida de la empresa.

Para ello, vamos a prestar atención a diferentes variables como:

- **Antiguedad** del cliente en la empresa
- **Gasto** total a lo largo de su suscripción
- **Tiempo** transcurrido desde la **última vez** que utilizaron la aplicación
- Número de **incidencias** enviadas a soporte
- Número de **propiedades** que posee el cliente
- Número total de **check-ins realizado**s, es decir, cuántas veces ha usado la aplicación

DATOS UTILIZADOS



Los datos se han obtenido de diferentes fuentes de datos:

- **Back office** de CheKin en Django
- **Hubspot** (plataforma de CRM)
- **ProfitWell** (proporciona métricas financieras)

He trabajado con 3 archivos csv, en los cuales me he centrado en los datos relativos al cliente.
El trabajo de limpieza consistió en:

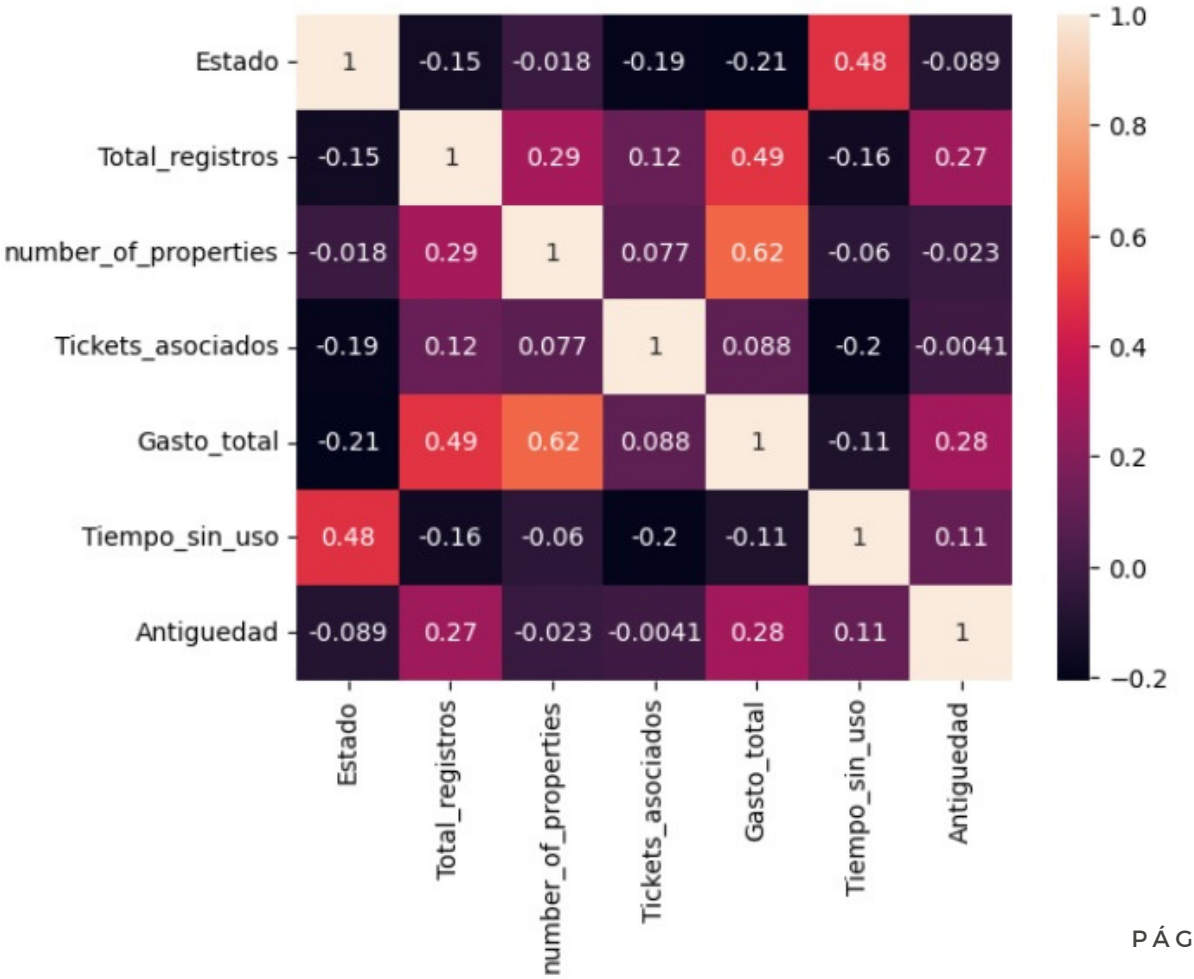
- Eliminar valores nulos
- Trabajar con columnas tipo 'datetime' para obtener rangos de tiempo
- Filtrar aquellos valores que representaban datos incorrectos (mal actualizados en la plataforma)

Después de limpiar y unir los diferentes datasets, obtuve un único dataframe con 2550 filas y 7 columnas, es decir, **2550 clientes**, de los cuales 1662 estaban activos en la plataforma y 888 se habían dado de baja.

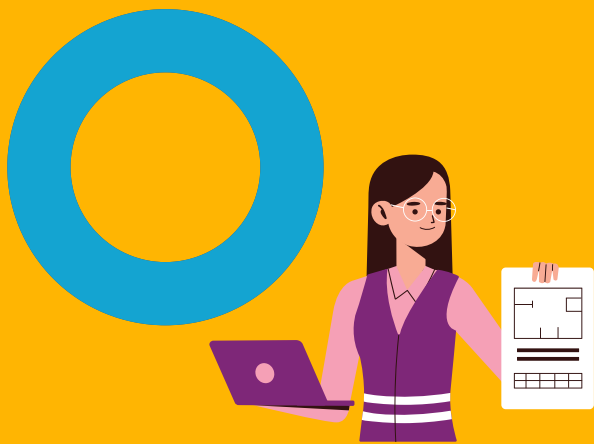
	Estado	Total_registros	number_of_properties	Tickets_asociados	Gasto_total	Tiempo_sin_uso	Antiguedad
3	0	317	5	2	95.64	2	253
4	1	380	2	2	68.71	65	1152
5	0	441	1	0	118.05	5	1433
6	0	18	1	0	49.79	6	182
8	0	340	5	0	242.00	4	199
...	...	...	...	...	...	...	...
2794	1	9	1	0	219.14	140	934
2796	0	4364	1	4	304.05	4	253
2797	1	19	1	0	77.38	587	745
2798	0	110	12	0	205.10	2	44
2799	1	21	1	0	23.40	196	379

2550 rows × 7 columns

Así mismo, estudié la **correlación** entre mi target y el resto de variables para comprobar la relevancia de mis datos para la predicción.



MODELOS ESCOGIDOS



Para realizar la predicción, entrené diferentes modelos de clasificación para evaluar y comparar sus scores.

El foco a la hora de elegir el modelo fue puesto en el **'recall score'**, ya que lo que me interesaba era predecir bien aquellos clientes que se habían dado de baja.

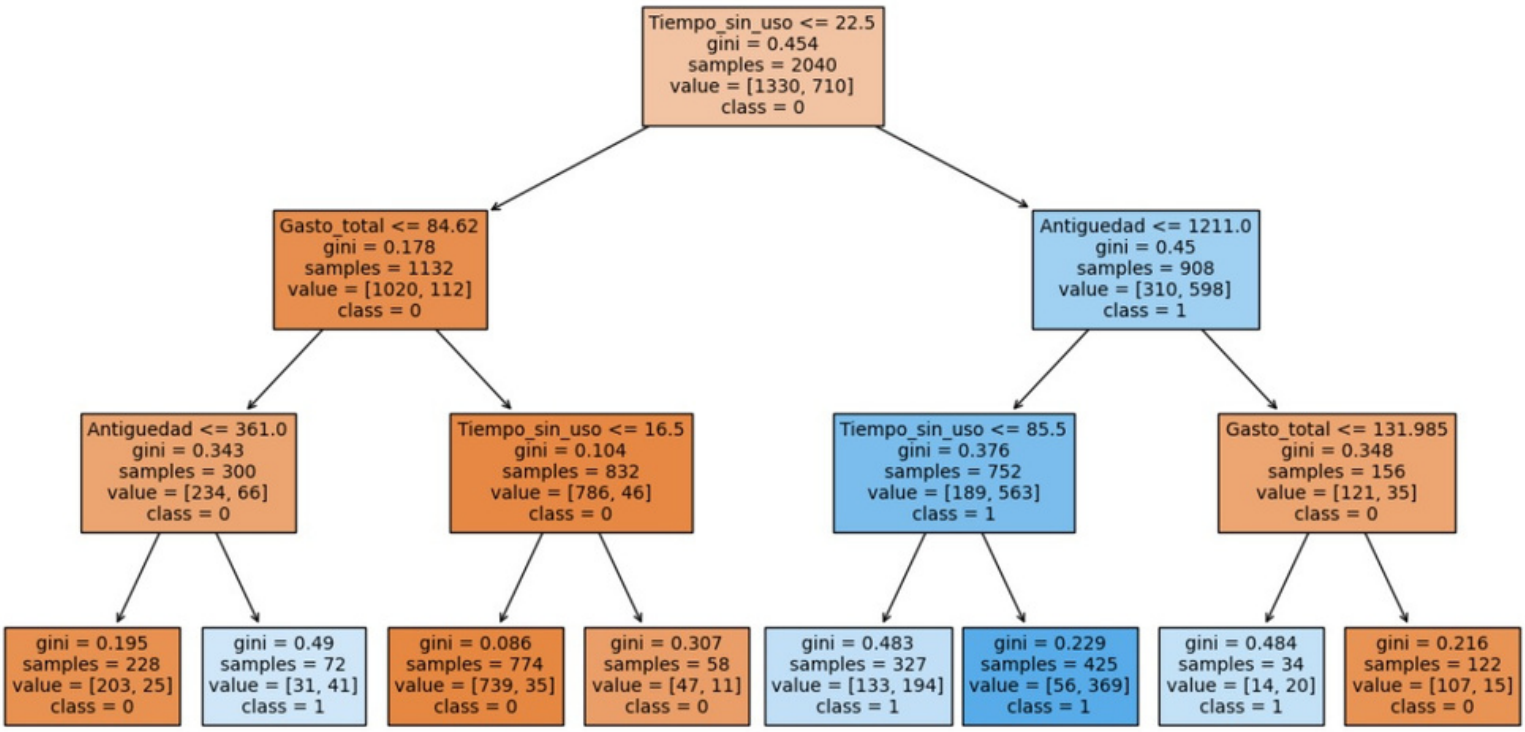
Modelos	Accuracy	Precision	Recall	F1	ROC AUC
Grid Search CV_SVC	0.545098	0.426630	0.882022	0.575092	0.623240
Gradient Boosting	0.876471	0.814208	0.837079	0.825485	0.867335
Decision Tree	0.815686	0.698113	0.831461	0.758974	0.819345
Ada Boost	0.882353	0.835227	0.825843	0.830508	0.869247
Random Forest	0.837255	0.743590	0.814607	0.777480	0.832002
KNeighbors	0.825490	0.737968	0.775281	0.756164	0.813845
SVC	0.849020	0.806061	0.747191	0.775510	0.825403
Logistic Regression	0.801961	0.727811	0.691011	0.708934	0.776229

El modelo con mejor score fue el **Support Vector Classifier.**

Para su entrenamiento se construyó un pipeline donde se balancearon las muestras, se escalaron los valores y posteriormente se entrenó el modelo mediante **Grid Search Cross Validation**.

Los parámetros escogidos fueron los siguientes:
C=1, coef0=10, kernel='sigmoid'

Sin embargo, para facilitar la visualización y poder sacar conclusiones que ayuden en la toma de decisiones, se construyó un **árbol de decisión**:



CONCLUSIONES

Al realizar este estudio pude concluir que mi modelo se enfocaba en 3 únicas variables, de las cuales sobresalía el tiempo que pasaba el cliente sin usar la aplicación.

	Feature importances
Tiempo_sin_uso	0.746156
Antiguedad	0.203189
Gasto_total	0.050655
Total_registros	0.000000
number_of_properties	0.000000
Tickets_asociados	0.000000

Por otro lado, la baja correlación con el número de incidencias puestas por los clientes, me ayuda a pensar que el motivo de baja en los servicios no se debe a problemas técnicos o falta de soporte.

Esto nos puede indicar que, al ser un software que da soporte a clientes con negocios turísticos, el motivo de la baja puede ser causa del cese del servicio post-vacacional.