



ESCUELA DE ORGANIZACIÓN INDUSTRIAL
MÁSTER EN BIG DATA Y BUSINESS INTELLIGENCE

Trabajo de Fin de Máster

Smart Yield

Autores:

Marcos Luis García Díaz

Iván Yauci González Hernández

Julio Daniel Pérez Cabrera

Ana María(Miriam) Quintero Padrón

Tutor:

Abel Pacheco Bernabéu

Índice

1	Prefacio.....	4
1.1	Definición del problema.....	4
1.2	¿Dónde está?	4
1.3	¿Cuándo ocurre?.....	4
1.4	Objetivo.....	4
2	Smart Yield	5
2.1	Contextualización y definición del problema	5
2.2	Investigación y toma de datos	6
2.3	Hipótesis de partida.....	8
2.4	Validación de las hipótesis.....	8
2.5	Resultados obtenidos	9
2.6	Conclusiones	10
2.7	Esquema del modelo de Negocio.....	12
2.7.1	Propuesta de valor.	13
2.7.2	Socios claves.	13
2.7.3	Recursos claves:.....	13
2.7.4	Segmento de clientes.	14
2.7.5	Actividades claves	14
2.8	Plan de acción.	15
2.8.1	Alcance.....	15
2.8.2	Objetivos.	15
2.8.3	Mapa de procesos.	16
2.8.4	Procesos estratégicos.	16
2.8.5	Procesos de apoyo.	17
2.8.6	Procesos claves.	18
3	Solución tecnológica.....	20
3.1	Desarrollo del Data Warehouse.....	21
3.1.1	Análisis fuentes de datos.	21
3.1.2	Extracción y transformación de los datos.....	22
3.1.3	Modelo OLAP.	22

3.1.4	Automatización.	24
3.1.5	Visualización.	25
3.2	Inteligencia de negocio para revenue management	25
3.2.1	Entorno de inteligencia de negocio y datos	25
3.2.2	Conectando el Entorno de visualización.....	26
3.2.3	Revenue management e inteligencia de negocio	28
3.2.4	Incertidumbres & Certidumbres en el revenue.....	29
3.2.5	Funcionalidad para Revenue Management.....	29
3.2.6	Principales indicadores del Revenue management.....	30
3.2.7	Módulos/Paneles de Inteligencia para revenue management	30
3.2.8	Estado del la ocupación y ADR	31
3.2.9	Evolución de la ocupación	31
3.2.10	Previsiones vs año anterior de cualquier periodo de tiempo marcado en los filtros. 31	
3.2.11	Diferencias temporales.....	31
3.2.12	Análisis de mercado.....	32
3.3	Desarrollando el modelo de cancelaciones.....	33
3.3.1	Data cleansing.....	33
3.3.2	Desarrollamos y estudiamos los modelos de ML.....	35
3.3.3	Analizando los resultados	38
3.4	Desarrollando el modelo de análisis de sentimientos	39
3.4.1	Encuesta a los clientes.....	39
3.4.2	Pre-procesado de datos.....	40
3.4.3	Obtención de los datos o Data loading.....	40
3.4.4	Procesado de datos.....	40
3.4.5	Dashborad en PowerBI.....	43
4	Recursos físicos y gestión del tiempo.	44
4.1	Análisis de talento humano.	44
4.2	Análisis de los recursos.	45
4.3	Gestión del tiempo.....	45
4.4	Indicadores claves	46

4.5	Beneficios tangibles:	46
4.6	Reducción de costes:	47
4.7	Beneficios intangibles:	47
4.8	Beneficios estratégicos:	47
5	Cash Flow:	48
5.1	Presupuestos:	48
6	Anexos	52
6.1	Desarrollo de modelo de NLP para detectar reseñas positivas y negativas	52
6.2	Desarrollo de modelo de clasificación para determinar cancelaciones	52

1 PREFACIO

1.1 DEFINICIÓN DEL PROBLEMA

El problema radica en que actualmente el mundo del turismo presenta unos grandes cambios a nivel operativo y táctico de los grandes actores del mercado, incluso para un modelo clásico de gestión han aparecido nuevos canales y herramientas como son las OTAs, metabuscadores, etc.

Por otro lado, hay que competir con grandes departamentos de revenue de otros hoteles que si son soportados por el Dato. A su vez, casi todos los alojamientos turísticos están en búsqueda de mejorar la experiencia del cliente.

Por tanto, es necesario disponer de tácticas y estrategias para la obtención, almacenamiento y explotación de los datos de los clientes. Desde sus datos personales a cualquier interacción con nuestra empresa: helpdesk, incidencias, sugerencias, historiales...

1.2 ¿DÓNDE ESTÁ?

Se carece de una estructura de datos donde empezar una estrategia, inteligencia de negocio, revenue y/o Experiencia de Cliente.

El problema se ha generado en que en la evolución de muchos hoteles se van adquiriendo softwares y aplicaciones independientes que solucionen únicamente los problemas operativos del hotel, ya que en general estas empresas tienen una buena salud económica.

1.3 ¿CUÁNDO OCURRE?

El problema ocurre cuando un Hotelero o un equipo de gestión decide implantar un sistema de inteligencia de negocio, cambiar el modo en que se gestiona los canales de venta, la relación con el cliente y/o la experiencia de cliente. Se carece de una estructura de datos donde empezar una estrategia, inteligencia de negocio, revenue y/o Experiencia de Cliente.

1.4 OBJETIVO

- Datos históricos de reservas para gestionar los departamentos del hotel y optimizar los recursos
- Capacidad de reacción y adaptabilidad frente a imprevistos
- Establecer precios adaptados según demanda. (Revenue Management)
- Nos adaptamos a los hábitos de consumo de mis clientes si un cliente tipo pide un servicio, otro cliente del mismo "tipo" se lo ofrezco, personalización de los productos.

- Acciones de marketing dirigidas
- Fidelización del cliente
- Actividades de ocio adaptadas según mi tipo de cliente, mejorando la experiencia del cliente y su fidelización.

2 SMART YIELD

2.1 CONTEXTUALIZACIÓN Y DEFINICIÓN DEL PROBLEMA

Como hemos comentado anteriormente, en los hoteles de mediano y pequeño tamaño no cuentan con los recursos de los grandes actores del sector, que van a la cabeza en cuanto al uso de estrategias basadas en los datos que ellos mismos extraen, así como de los bancos de datos existentes.

Ellos utilizan las últimas técnicas y herramientas del mercado, como machine learning, deep learning, datawarehousing, indicadores mediante software específico de visualización de datos, etc. Esto les permite estar a la vanguardia del mercado y conseguir una mayor clientela, así como una mayor tasa de retorno del mismo cliente y una mayor flexibilidad para adaptarse a cualquier imprevisto.

Por otro lado, la cultura del dato no está arraigada en la estructura de estas empresas medianas o pequeñas, por lo que no se explotan los datos que ya se tienen, ni se define una estrategia clara de recolección del dato. Este hecho tiene como consecuencia más inmediata la poca adaptación y poca flexibilidad de nuestros potenciales clientes a la hora de adaptarse a entorno cambiantes o extraordinarios, como pueden ser los recientes casos de Thomas Cook o la crisis del coronavirus.

A lo anteriormente descrito podemos añadir la baja optimización de los procesos que pueden sufrir estos pequeños actores, que por lo general suelen llevar un control bastante rudimentario o manual, en el que no hacen uso de todas las tecnologías y herramientas que ofrece el mercado. Con todos estos argumentos podemos formular las siguientes hipótesis sobre nuestros potenciales clientes:

- Nuestros clientes no sacan partidos a los datos que utilizan
- Nuestros clientes quieren mejorar la experiencia individual de sus clientes
- Nuestros clientes requieren soluciones que optimicen las tarifas de ventas
- Nuestros clientes quieren mejorar sus procesos operativos
- Nuestros clientes quieren establecer estrategias basadas en los datos que posee
- Nuestros clientes probablemente quieran mejorar su competitividad frente a sus rivales
- Creemos que nuestros clientes conseguirán más clientes a través de las redes sociales

Planteando estas hipótesis podremos cubrir un amplio espectro de necesidades, desde operativas a comerciales. Nuestra misión por tanto es poder actuar sobre todos los departamentos del hotel para poder mejorarlos, automatizarlos y optimizarlos.

2.2 INVESTIGACIÓN Y TOMA DE DATOS

Business Intelligence y Big Data, la Inteligencia de Negocio o BI es un término genérico que incluye las aplicaciones, la infraestructura y las herramientas, y las mejores prácticas que permiten el acceso y el análisis de la información para mejorar y optimizar las decisiones y rendimiento (<https://www.gartner.com/en/information-technology/glossary/big-data>).

Según datos del gobierno de Canarias (<http://www.gobiernodecanarias.org/istac/jaxi-istac/menu.do?Uripub=urn:uuid:ad002616-6e9f-4e4a-8e7a-7c393a69fce0>), en la actualidad existen un descenso importante y preocupante en el número de habitaciones disponibles en el ámbito Canario:

	CANARIAS				
	2020 Mayo	2020 Abril	2020 Marzo	2020 Febrero	2020 Enero
Habitaciones disponibles					
1 Estrella	12	..	1.611	1.611	1.611
2 Estrellas	95	..	3.673	3.698	3.673
3 Estrellas	782	..	24.024	24.169	24.062
4, 5 Estrellas	1.389	..	87.960	88.245	88.113
4 Estrellas	1.264	..	70.973	71.440	71.308
5 Estrellas	125	..	16.987	16.805	16.805

Entendemos que al coincidir con la crisis de COVID-19 el descenso es inevitable, la pregunta es, ¿podríamos mejorar en la recuperación del sector? Entendemos que es importante contactar con estos establecimientos y analizar su situación particular para crear una solución emergente, innovadora y eficiente.

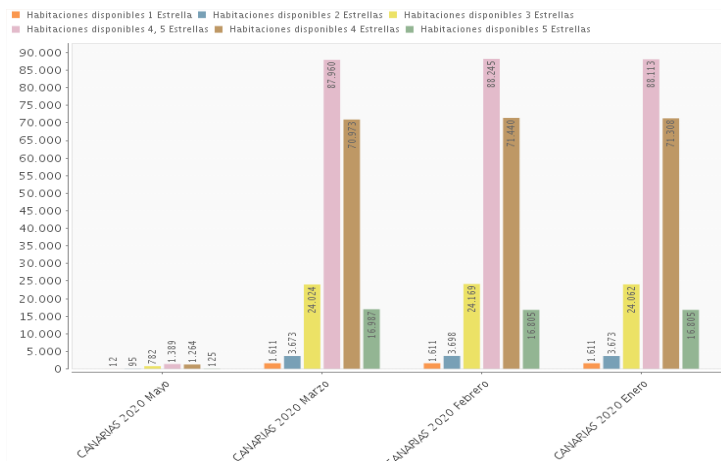
Podemos ver en la gráfica siguiente que la crisis ha afectado a todo tipo de hoteles por igual, pero entendemos que la recuperación será más compleja para estos pequeños y medianos establecimientos.

Con el ánimo y entusiasmo de implantar un sistema que ayude a la recuperación y establezca sistemas en Hoteles que carezcan de dicha inteligencia del dato, se han realizado distintas entrevistas formales e informales con el personal de los hoteles implicados activamente en la propuesta:

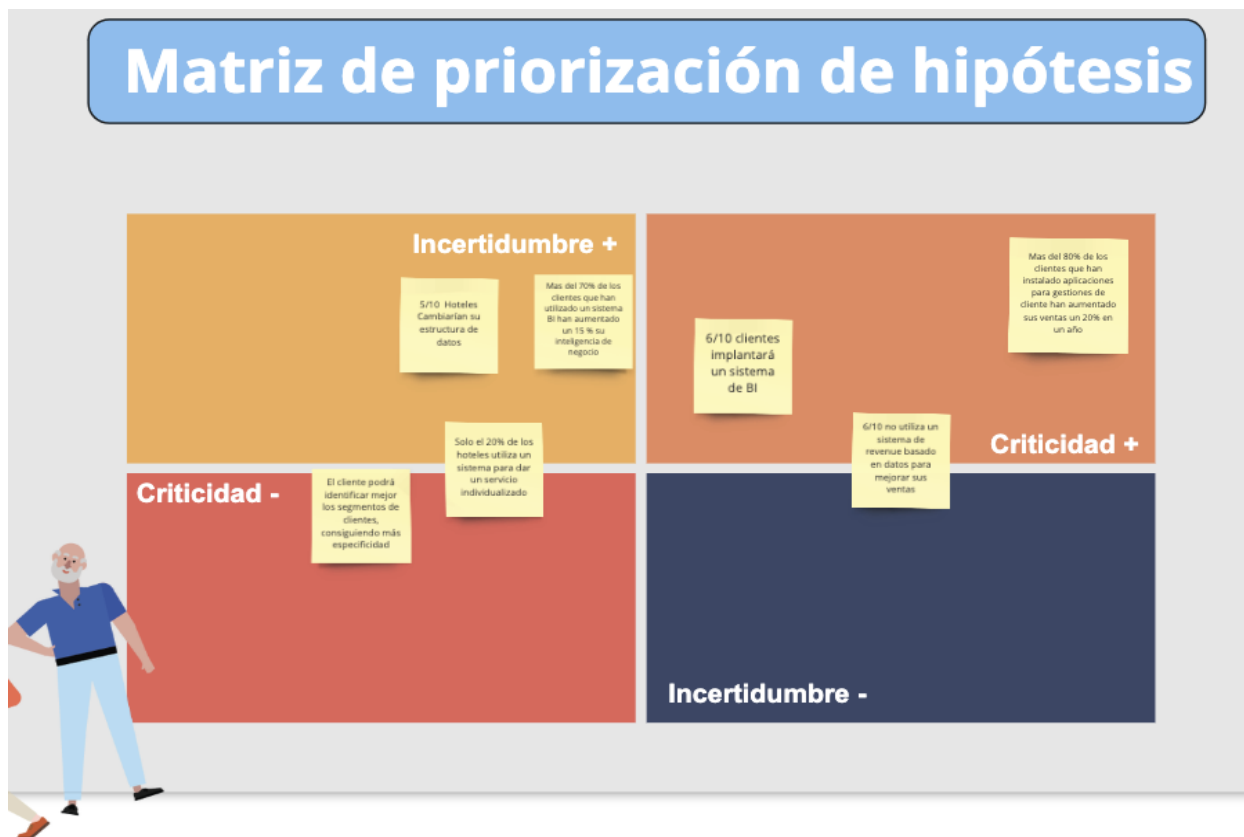
Parque Santiago III: Av. De Las Américas S/N - 38660 Arona – Santa Cruz de Tenerife – España. Un hotel de CATEGORÍA: Apartahotel 3 Llaves (<http://parquesantiagotenerife.es>)

Parque Santiago IV: Av. De Las Américas S/N - 38660 Arona – Santa Cruz de Tenerife – España. Un hotel de CATEGORÍA: Apartahotel 3 Llaves (<http://parquesantiagotenerife.es>)

Las Marismas de Corralejo: Calle Huriamen, s/n, 35660 Corralejo, Las Palmas. Un hotel de CATEGORÍA: Apartahotel 3 Llaves (<http://lasmarismasdecorralejo.com/>)



Dichas entrevistas nos han clarificado y dado un mayor conocimiento del entorno de trabajo actual de nuestros clientes y se han priorizado las siguientes hipótesis:



2.3 HIPÓTESIS DE PARTIDA

De dichas hipótesis, hemos destacado y analizado en profundidad aquellas de mayor criticidad e incertidumbre ya que a su vez son las que más relacionadas están con el éxito de nuestra propuesta, dichas hipótesis son las siguientes:

- 5 de cada 10 hoteles cambiarían su estructura de datos a un entorno BI
- Más de un 70% de los clientes que han utilizado un sistema BI han aumentado sus ventas un 20% anual.

2.4 VALIDACIÓN DE LAS HIPÓTESIS

Se han utilizado dos técnicas parecidas pero diferentes para cada uno de los casos, en nuestro primer grupo de hipótesis, **para validar los sistemas sin un BI integrado**, utilizamos **la entrevista**, ya que necesitábamos más información y más personalizada, así como dirigir o centrar a nuestro interlocutor de modo personalizado. En el segundo grupo de hipótesis, referente a las **ganancias adicionales que aporta un sistema BI**, han sido validadas mediante **la encuesta**, ya que podemos llegar a más gente, son más cortas y poco personalizadas, pero con un amplio target:

5 de cada 10 hoteles cambiarían su estructura de datos a un entorno BI

Nuestros dos clientes iniciales están identificados y han expresado activamente su decisión de cambio, queremos entender qué ocurre con otros hoteles de similares características (Hoteles pequeños o medianos sin un sistema de BI en la actualidad en el territorio Canario), para ello hemos diseñado una encuesta dirigida a los dueños o gerentes y realizado algunas entrevistas.

Más de un 70% de los clientes que han utilizado un sistema BI han aumentado sus ventas un 10% anual

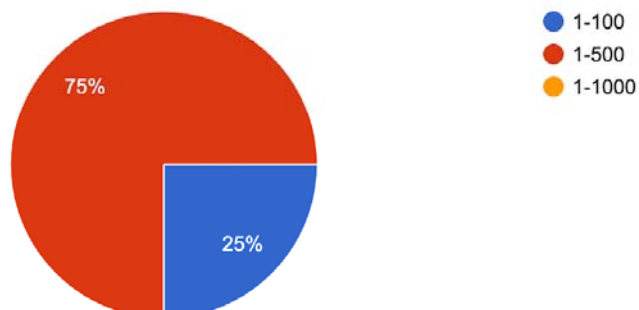
Se ha definido y lanzado una encuesta personalizada, en un grupo de directores, subdirectores, jefes de departamento de distintos hoteles tanto en Canarias como Baleares.

- ¿Cuántas habitaciones tiene su hotel?
- ¿Utiliza un PMS (Property management system) con herramienta principal de gestión?
- ¿Tiene usted un sistema de inteligencia de negocio en su hotel?
- Si la respuesta anterior es sí: ¿En cuánto ha aumentado sus ventas dicho sistema?
- 10% > 100%
- ¿Le facilita tener algún otro sistema para poder obtener los datos que necesite en un momento dado?

2.5 RESULTADOS OBTENIDOS

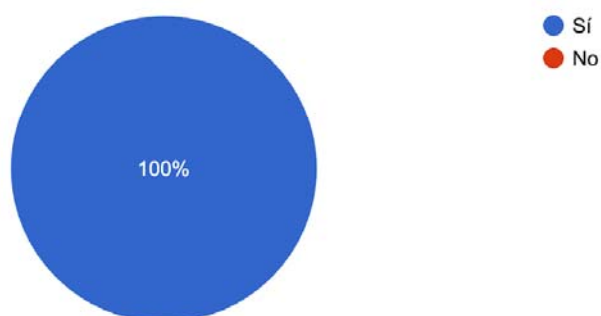
¿Cuántas habitaciones tiene su hotel?

12 respuestas



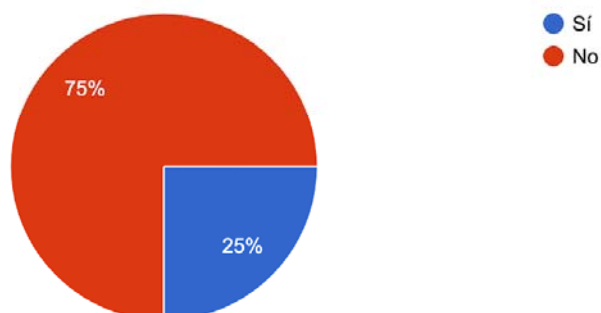
¿ Utiliza un PMS (Property management system) con herramienta principal de gestión ?

12 respuestas



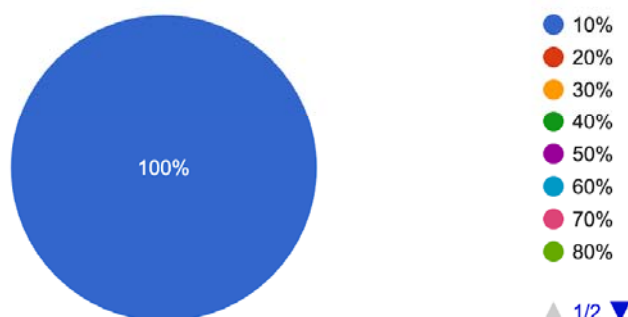
¿Tiene usted un sistema de inteligencia de negocio en su hotel?

12 respuestas



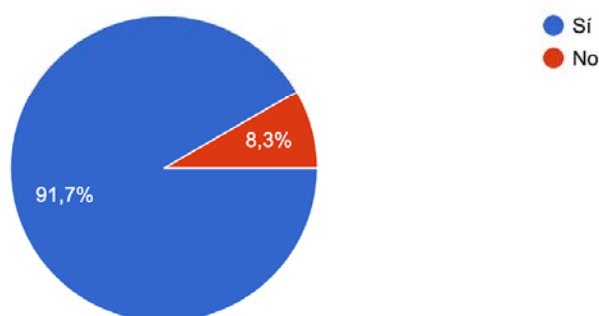
Si la respuesta anterior es sí: ¿En cuánto ha aumentado sus ventas dicho sistema?

2 respuestas



Le facilitaría tener algún otro sistema para poder obtener los datos que necesite en un momento dado ?

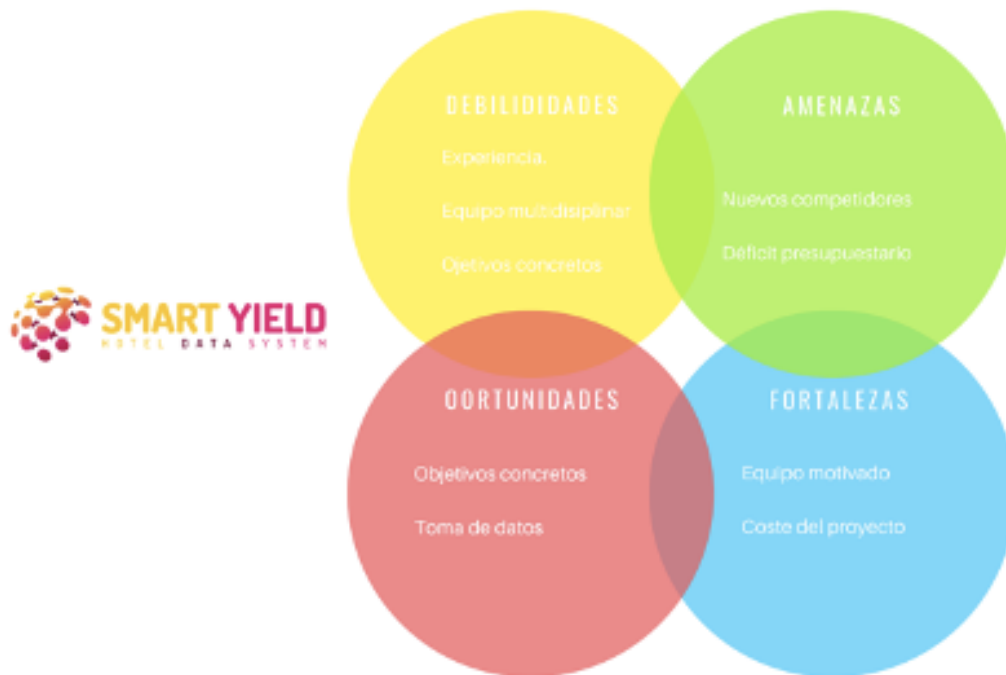
12 respuestas



2.6 CONCLUSIONES

Como se demuestra en la validación de las hipótesis, una vez realizada la encuesta, podemos observar que se cumple los porcentajes que hemos validado. En el grupo que se ha realizado esta encuesta, directores, subdirectores, jefes de departamentos, queda reflejado varias tendencias:

- El porcentaje de hoteles que utilizan un PMS (Property management system) como sistema principal para la organización de todos sus datos es elevado, un 100%.
- Muy pocos hoteles tienen un sistema BI, pero los que lo tienen han incrementado como mínimo un 10% sus ventas.
- La gran mayoría de los encuestados, un 91%, les gustaría poder organizar / visualizar los datos de otra forma, a la que llevan habitualmente.

ANÁLISIS DAFO

1. Debilidades. Aspectos que limitan o reducen la capacidad de desarrollo efectivo de la estrategia de la empresa, que constituyen potenciales obstáculos para cumplir con los objetivos marcados. Entre ellos se han podido identificar los siguientes:

Experiencia en el sector a la hora de ofrecer el servicio será más difícil que los potenciales clientes entiendan la necesidad de tener este servicio. Potenciales clientes con pocos conocimientos del modelo del proyecto a ofrecer.

A pesar de disponer de un equipo multidisciplinar, carecemos de personal con un profundo conocimiento de marketing y de finanzas que pueda guiarnos en el proceso de buscar tendencias en los datos o crear el sistema de BI más eficiente e interesante para el cliente.

La falta de objetivos concretos unido a la gran cantidad de datos de los que disponemos y que presentan una alta variabilidad temporal en un mercado volátil.

2. Amenazas. Toda fuerza del entorno que puede impedir la implantación de una estrategia, o bien reducir su efectividad, o incrementar los riesgos de la misma. Pueden afectar negativamente las posibilidades de conseguir los objetivos generales.

Dado que estamos trabajando en un sector en auge, puede que aparezcan nuevos competidores, o incluso empresa que decidan empezar a tratar los datos como hemos desarrollado nosotros.

Al tener nuestro nicho de mercado en hoteles pequeños / medianos, podemos encontrarnos con un déficit presupuestario para adoptar por este proyecto.

3. Fortalezas. Son capacidades, recursos, posiciones alcanzadas y, consecuentemente, ventajas competitivas que deben y pueden servir para explotar oportunidades.

Equipo motivado, con los conocimientos técnicos y amplia experiencia en el sector turístico hotelero para llevar a cabo de forma eficiente todos los objetivos planteados en el proyecto.

El coste del proyecto será bajo, dado que nos adaptamos a las características de cada cliente, tanto de su infraestructura, trabajando en su mayoría con software Open Source, como de su personal. Pudiendo de esta manera abaratar de forma significativa los costes del proyecto.

4. Oportunidades. Todo aquello que pueda suponer una ventaja competitiva para la empresa, o bien representar una posibilidad para mejorar la rentabilidad de esta o aumentar la cifra de sus negocios.

La toma de datos se podrá realizar sobre informes que podamos sacar de los PMS del cliente, con lo cual no conlleva ningún coste asociado.

Tener un objetivo muy concreto donde pocas empresas han puesto el foco, y que podemos explotar de forma eficiente.

Smart Yield persigue aportar valor al dato de pequeñas medianas empresas hoteleras de la siguiente manera:

- Almacenando todos los datos en un Data Warehouse que se generan por parte de estas empresas, pero que mayoritariamente insertan en software privativos PMS (Property management system) de los cuales no pueden gestionar de un manera eficiente y versátil.
- Aplicando dashboard sencillos y fáciles de entender para una correcta toma de decisiones por parte de la dirección de estas empresas.
- Aplicando algoritmos de Machine Learning IA para:
 - O Sistemas de Revenue
 - O Sistemas de reservas y cancelaciones
 - O Encuestas de satisfacción.
 - O Gestión interna de la organización del trabajo en el hotel.

2.7 ESQUEMA DEL MODELO DE NEGOCIO.

El modelo de negocio que Smart Yield está desarrollando se identifica y describen los aspectos más importantes para tener en cuenta:

2.7.1 Propuesta de valor.

La propuesta de valor para Smart Yield es una solución adaptada a la mejora de la gestión de los datos para pequeñas y medianas empresas hoteleras. Con esto se busca optimizar los recursos que se utilizan para una mejor toma de decisiones.

En primer lugar generando un Data Warehouse, donde poder almacenar todos los datos de la empresa de una forma mucho más efectiva, que la que se tienen actualmente por parte de este tipo de empresas, siendo en un alto porcentaje privativa. Todos estos datos recogidos, transformados y almacenados de una forma eficiente serán puestos a disposición de los departamentos del hotel para una toma de decisiones mucho más eficiente y rápida. De esta manera los departamentos mejorar la distribución y optimización de sus recursos.

En cuanto a los objetivos planteados sobre los análisis de los datos con los algoritmos de ML, podremos decidir, monitorizar y optimizar:

- Los sistemas de yield management.
- Las reservas / cancelaciones.
- Las campañas de marketing.
- Los resultados de las encuestas de satisfacción.
- La gestión interna y distribución de los partes de trabajo.

2.7.2 Socios claves.

Nuestros socios claves son todas aquellas empresas que generan datos que se almacenan en el hotel, como pueden ser:

- Reservas / cancelaciones / modificaciones por parte de tour operadores, agencias de viajes, Agencias Online, web de reservas, etc
- PMS, el sistema de gestión del hotel, ya que es él tiene almacenado la mayor parte de los datos del hotel.
- El propio hotel con sus encuestas de satisfacción y generación de partes de trabajo.

2.7.3 Recursos claves:

El principal recurso clave para desarrollar este proyecto es el dato. Su tratamiento, análisis y optimización es el que nos dará toda la fortaleza del proyecto.

Para el tratamiento de los datos recibidos hemos elegido Pentaho BI Suite, que es un conjunto de programas libres para generar inteligencia empresarial. Incluye herramientas integradas para generar informes, minería de datos, ETL, etc.

En cuanto al almacenamiento elegimos mysql, que es un sistema de gestión de bases de datos relacional desarrollado bajo licencia dual: Licencia pública general/Licencia comercial por Oracle Corporation.

Para la correcta y fácil visualización de los datos usaremos PowerBI, que es un servicio de análisis empresarial de Microsoft, su objetivo es proporcionar visualizaciones

interactivas y capacidades de inteligencia empresarial con una interfaz lo suficientemente simple como para que los usuarios finales creen sus propios informes y paneles.

2.7.4 Segmento de clientes.

El segmento de clientes para este caso, son hoteles independientes o cadenas hoteleras de un tamaño mediano-pequeño, los cuales sus canales de ventas han sido trabajados de la manera tradicional y que dependen fuertemente de la tour-operación.

En todo caso el proyecto de Sistema integral para la gestión del dato en empresas hoteleras podría ser implementado en otro tipo de establecimientos hoteleros/turísticos. Pero el segmento inicial se predispone al éxito del proyecto

Estructura de costes.

La estructura de costes contempla los siguientes aspectos:

- Costes de implantación del sistema en servidores propios del hotel.
- Mantenimiento del sistema.
-
-

Relación con los clientes.

La relación que vamos a tener con nuestro cliente se base en estos tres pilares fundamentales:

1. Personalización y trato directo con el cliente. Se elaborará un calendario de reuniones periódicas, en mayor medida por medios online, siendo las primeras y de mayor importancia presenciales, de esta manera obtenemos un feedback directo con nuestro cliente.
2. Como segundo pilar, ofrecemos un dashboard de control para los jefes de departamento del complejo hotelero, de forma que puedan ver los resultados sin necesidad de un perfil más técnico.
3. Una vez implantado nuestro proyecto, se mantendrá un servicio postventa, una asesoría personalizada, para ir comprobando la evolución del proyecto, y si fuese necesario corregir / adaptar a posibles fallos o cambios futuros.

2.7.5 Actividades claves

El flujo de actividades claves contempla los siguientes puntos:

- Recolección de datos del proceso antiguo.
- Definición modelos de datos.
- Procesos automatizados de recolección.
- Definición y generación de procesos de Extracción transformación and carga (ETL)
- Definir modelo IA a utilizar.

- Definir parámetros óptimos para la venta.
- Diseño de cuadros de mando para la gestión interna (operativo y estratégico).



2.8 PLAN DE ACCIÓN.

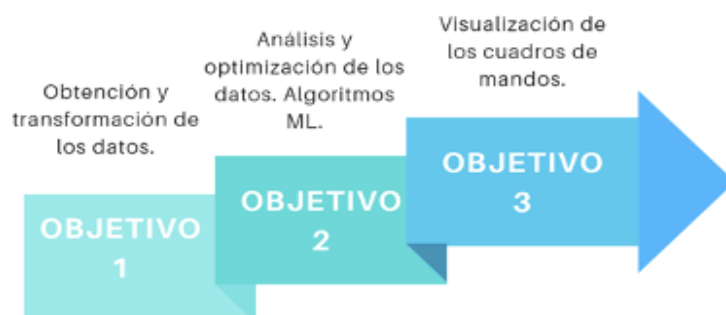
2.8.1 Alcance.

Poder ofrecer un proyecto donde se pueda tratar, almacenar y optimizar los datos generados por una empresa hotelera, de esta manera incrementar la venta directa, conseguir clientes fidelizados, gestionar mejor los procesos internos de gestión, para obtener así un aumento en los ingresos del hotel.

2.8.2 Objetivos.

En una primera fase necesitamos obtener y transformar los datos en nuestro propio Data Warehouse. Estos datos nos servirán para el proceso de análisis y optimización en la toma de decisiones de la empresa. Toda esta información estará disponible desde cualquier dispositivo, en cualquier momento, proporcionando herramientas de visualización que faciliten las gestiones de la toma de decisiones.

Así los objetivos quedan definidos de la siguiente manera:



2.8.3 Mapa de procesos.



2.8.4 Procesos estratégicos.

1. **Dirección estratégica:** Se buscará la especialización en el análisis de los datos, en el sector hotelero, ofreciendo un entorno BI, donde poder tomar mejores decisiones para una mayor rentabilidad del hotel.
2. **Dirección comercial:** Se creará una BD de datos que permita tener todos los hoteles independientes de tamaño pequeño / mediano, para tener localizados todos estos potenciales clientes. De esta forma, nos facilita la labor comercial de búsqueda, y nos permitirá saber aproximadamente el tamaño del mercado al que nos enfrentamos.

El contacto directo con los hoteles será base de nuestro esfuerzo comercial, ya que, al ser un servicio novedoso, requiere un trato más cercano y personalizado.

Nuestra presencia en internet se basará en una web en la que informaremos sobre los servicios que prestamos. Se desarrollará contenido multimedia, muy visual, que facilite la explicación de nuestro producto.

Es muy importante tener presencia en eventos especializados del sector hotelero, donde realizaremos workshop para explicar nuestra solución a potenciales clientes.

3. **Dirección I+D:** Al tratarse de una solución tecnológica, la innovación y mejora continua será una de las principales características principales de nuestro producto. Nuestro modelo de Machine Learning estará sujeto a un proceso de actualizaciones continua, alimentado por los datos que vayamos obteniendo de los distintos canales de entrada.

2.8.5 Procesos de apoyo.

1. Gestión financiera

La gestión financiera de Smart Yield es fundamental y respalda su estructura permitiendo su crecimiento.

Nos gustaría señalar las estrategias definidas para manejar y administrar las finanzas:

Contratación de personal: El personal es un pilar fundamental, desde las fases iniciales y durante la duración del proyecto se contará con 4 expertos. Dichos expertos, están implicados desde las fases iniciales de investigación, conocen en profundidad el sector hotelero y el sector del Business intelligence convirtiéndose en perfiles claves para el éxito de Smart Yield.

Gastos de equipo tecnológico: Se adquirirá lo estrictamente necesario de software, almacenamiento cloud o PC, evitando compras que no sean necesarias. Serán necesarios 4 equipos personales para el pre-procesado y procesado de datos. En un primer momento se propuso la contratación de un servidor para el almacenamiento datos y la gestión de procesos, sin embargo, se ha optado, por motivos económicos, por usar equipos locales y herramientas gratuitas (OneDrive) para el almacenamiento de los datos.

Gastos de software: Si bien se intenta trabajar con herramientas open source tanto como sea posible, es importante saber que para obtener un rendimiento máximo se ha optado por PowerBI con el objetivo de proporcionar visualizaciones interactivas y capacidades de inteligencia empresarial con una interfaz lo suficientemente simple para los usuarios finales.

PowerBI: Se recomienda PowerBI Pro

Cloud storage: Se recomienda que la selección del servidor de Cloud cumpla con la Ley Orgánica de Protección de Datos Personales (LOPD)

Publicidad: Las ventas de nuestro producto será nuestro mayor beneficio, por lo tanto, se hace necesaria la contratación de servicios comerciales. Con este gasto

lograremos el posicionamiento deseado e incrementar las ventas. Por ello se han definido las métricas correspondientes, las cuales ayudarán a conocer el éxito de las campañas. Por otro lado, nunca se pedirán préstamos o créditos con esta finalidad.

Ingresos: Los ingresos proyectados, por cada producto Smart Yield, serían los siguientes:

1. Incremento en las reservas
 2. Optimización de recursos de limpieza
 3. Optimización de recursos de animación
 4. Optimización de servicios
 5. Reducción de gastos
2. **Gestión de proveedores:** Necesitamos establecer relaciones con proveedores de servicio para el correcto funcionamiento del desarrollo del proyecto.
- Proveedores de datos: PMS del hotel, AAVV, TTOO, clientes directos, OTA's
 - Proveedor de BD: Utilizaremos MySQL como sistema de gestión de la BD, y Pentaho como sistema de BI.
 - Proveedor de servidor: Nos podemos adaptar a cada situación, pudiendo utilizar algún servidor que tenga el hotel, o pasando presupuesto valorando cuál es la mejor opción, si la compra de un servidor nuevo, o la utilización de un servicio CLOUD.
3. **Gestión de personas:** Inicialmente Smart Yield empezará su andadura con los 4 fundadores del proyecto y puntualmente con colaboraciones externas. Dentro del equipo, al tener perfiles distintos, podremos hacer frente para cubrir todos los aspectos necesarios para el desarrollo del proyecto.

2.8.6 Procesos claves.

Estos son los procesos claves que se han definido siguiendo las fases habituales del ciclo de vida de una solución tecnológica:

2.8.6.1 Análisis.

Estudio de los requisitos del cliente. Se realiza estudio en función de las características del hotel, para incidir en los aspectos del análisis de datos, que más rentabilidad se le podrá sacar.

Análisis de las fuentes de datos. Se analizarán las posibles fuentes de datos con las que esté trabajando en ese momento el hotel. Todas estas fuentes serán la entrada al sistema DW:

- Recopilación interna:
 - O Histórico de datos del PMS
 - O Gestión interna del hotel. Partes de mantenimiento.
- Recopilación externa:

- O Proveedores de reservas: TTOO, AAVV, OTA's, Web, Email.
- O Encuestas de satisfacción: Wifi, Web, OTA's, portales dedicados.

Análisis del modelo Machine Learning. Se analiza, una vez vista la estructura de datos que se va a utilizar, cual son los modelos de ML que se van a utilizar para predecir el valor de una variable o clasificar diferentes categorías.

2.8.6.2 Diseño.

Definición kpis. Estos indicadores son los que permitirán medir el grado de consecución de los objetivos del cliente y diseñar los Dashboard más adecuados para dar soporte a la toma de decisiones.

Diseño Data Warehouse. Dependiendo de la solución adoptada por el hotel en cuanto al servidor: utilizar su propio servidor, comprar un servidor nuevo o utilizar servicio Cloud. Instalaremos y configuraremos una base de datos MySQL donde crearemos el Data Warehouse, utilizando Pentaho BI Suite para crear los procesos de extracción, transformación y carga de los campos calculados.

Diseño modelos Machine Learning. Para el diseño de los modelos ML se utilizará Python, con el cual accederemos, cruzaremos y analizaremos los datos almacenados en el Data Warehouse. Se aplicarán técnicas de clasificación para poder desarrollar modelos que nos permitan adelantarnos a eventos que a priori atienden a una aleatoriedad, como pueden ser las cancelaciones, o desarrollar modelos de regresión logística para poder predecir comportamientos futuros de las reservas.

El objetivo principal de la utilización y aplicación de NLP es la de conocer la opinión de nuestros clientes sobre los productos del hotel (wifi, habitación, servicios en general, etc.) Y mostrar las opiniones, según positivas o negativas, de un modo sencillo y amigable en un dashboard definido. De este modo se puede conocer rápidamente que podríamos mejorar, dónde y por qué para mejorar la estancia de nuestros clientes en el hotel. Una vez realizada una breve encuesta a nuestros clientes, se reprocesan los datos utilizando librerías de NLP como textblob. Textblob es una biblioteca de Python (2 y 3) para procesar datos textuales, proporciona una API simple con la que podemos realizar etiquetado de parte del discurso, extracción de frases nominales, análisis de sentimientos, clasificación, traducción y más. Una vez analizadas dichas encuestas, y obtenido unos resultados entre -1 y 1 (siendo negativo y positiva respectivamente la opinión de nuestros clientes) se cargarán mediante procesos PowerBI en nuestro dashboard específico.

2.8.6.3 Implementación.

Implementación DW. En este punto se llevará la completa integración del producto ofrecido al cliente. Mediante la combinación de los modelos ML y el DW desarrollaremos distintas analíticas para facilitar la toma de decisiones y optimizar la gestión del hotel.

- Análisis de gestión de ingresos (Yield Management).

- Tasa de futuras cancelaciones.
- Sentimientos y encuestas de satisfacción.
- Hábitos de los huéspedes.
- Gestión de la relación con los clientes.

Desarrollo de cuadros de mandos e informes. Se utilizará PowerBI para la visualización de los indicadores claves, mediante cuadros de mandos e informes, por parte de los directivos hoteleros.

2.8.6.4 Soporte.

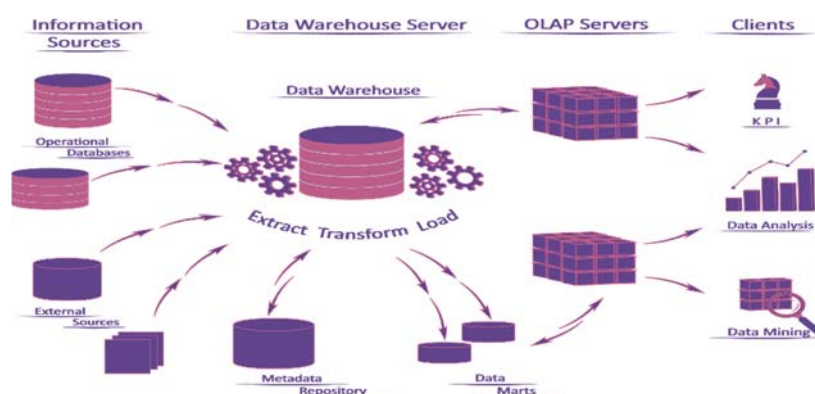
Soporte a clientes. Se dispondrá un servicio de atención al cliente por distintos canales: web, teléfono, email, para resolver cualquier duda o problema sugerido por el cliente.

Mantenimiento. Se definirán un conjunto de pruebas e indicadores que ayuden a mantener la efectividad del sistema.

Actualizaciones. En este apartado se contemplan todas las actualizaciones, tanto tecnológicas en las herramientas seleccionadas como de los propios modelos ML.

3 SOLUCIÓN TECNOLÓGICA.

El proyecto Smart Yield es una solución de optimización de los datos, que facilita la toma de decisiones sobre las necesidades de la gestión hotelera, mediante cuadros de mandos, informes y predicciones, para de esta manera tener un incremento en los beneficios del hotel.



Modelo lógico.

Para llevar a cabo el proyecto Smart Yield, hemos utilizado diferentes herramientas de análisis de datos y distintos lenguajes de programación. Dentro del proyecto inicial vamos a utilizar dos conjuntos de datos de empresas hoteleras distintas (Parque Santiago – Santa Cruz de Tenerife, Las Marismas de Corralejo - Fuerteventura) las cuales se han ofrecido para poder trabajar este proyecto con sus histórico de datos de reservas y cancelaciones.

3.1 DESARROLLO DEL DATA WARHOUSE

El análisis, diseño e implementación de la solución tecnológica la dividimos en los siguientes procesos claves:

3.1.1 Análisis fuentes de datos.

Análisis de las fuentes de datos que se van a utilizar para el DW, una vez analizadas estos dataset se procede a crear un modelo de STG genérico en MySQL, que contemple en su gran mayoría todas las fuentes de datos que se tratarán en un futuro.

El modelo STG es un sistema que permanece entre las fuentes de datos y el Data Warehouse con el objetivo de facilitar la extracción y transformación de los datos, mejorando la calidad de estos.

stg_tfm.reservas							
Info Columns Indexes Triggers Foreign keys Partitions Grants DDL							
Column	Type	Default Value	Nullable	Character Set	Collation	Privileges	
Agencia	varchar(50)		YES	utf8	utf8_general_ci	select,insert,update,references	
Agencia Planning	varchar(50)		YES	utf8	utf8_general_ci	select,insert,update,references	
Agencia Precios	varchar(50)		YES	utf8	utf8_general_ci	select,insert,update,references	
Bloque	varchar(10)		YES	utf8	utf8_general_ci	select,insert,update,references	
Caducidad tarjeta cré...	varchar(10)		YES	utf8	utf8_general_ci	select,insert,update,references	
Camas extra	int	0	YES			select,insert,update,references	
Canal de venta	varchar(10)		YES	utf8	utf8_general_ci	select,insert,update,references	
Cancelación	tinyint(1)		YES			select,insert,update,references	
Cardex (1)	varchar(20)		YES	utf8	utf8_general_ci	select,insert,update,references	
Cardex (2)	varchar(20)		YES	utf8	utf8_general_ci	select,insert,update,references	
Cardex (3)	varchar(20)		YES	utf8	utf8_general_ci	select,insert,update,references	
Cardex (4)	varchar(20)		YES	utf8	utf8_general_ci	select,insert,update,references	
Cardex (5)	varchar(20)		YES	utf8	utf8_general_ci	select,insert,update,references	
Cardex (6)	varchar(20)		YES	utf8	utf8_general_ci	select,insert,update,references	
Cupo	varchar(10)		YES	utf8	utf8_general_ci	select,insert,update,references	
Código postal del die...	varchar(30)		YES	utf8	utf8_general_ci	select,insert,update,references	
Código tarjeta crédito	varchar(10)		YES	utf8	utf8_general_ci	select,insert,update,references	
D.N.I./Pasaporte Ocu...	varchar(20)		YES	utf8	utf8_general_ci	select,insert,update,references	
D.N.I./Pasaporte Ocu...	varchar(20)		YES	utf8	utf8_general_ci	select,insert,update,references	
D.N.I./Pasaporte Ocu...	varchar(20)		YES	utf8	utf8_general_ci	select,insert,update,references	
D.N.I./Pasaporte Ocu...	varchar(20)		YES	utf8	utf8_general_ci	select,insert,update,references	
Departamento de co...	varchar(10)		YES	utf8	utf8_general_ci	select,insert,update,references	
Dirección del cliente	varchar(50)		YES	utf8	utf8_general_ci	select,insert,update,references	
DNI/Pasaporte	varchar(20)		YES	utf8	utf8_general_ci	select,insert,update,references	
Días	int	0	YES			select,insert,update,references	
Edad del cliente	varchar(10)		YES	utf8	utf8_general_ci	select,insert,update,references	
Edad ocupante (1)	varchar(10)		YES	utf8	utf8_general_ci	select,insert,update,references	
Edad ocupante (2)	varchar(10)		YES	utf8	utf8_general_ci	select,insert,update,references	
Edad ocupante (3)	varchar(10)		YES	utf8	utf8_general_ci	select,insert,update,references	
Edad ocupante (4)	varchar(10)		YES	utf8	utf8_general_ci	select,insert,update,references	
Edad ocupante (5)	varchar(10)		YES	utf8	utf8_general_ci	select,insert,update,references	
Email del cliente	varchar(50)		YES	utf8	utf8_general_ci	select,insert,update,references	
Estado	varchar(5)		YES	utf8	utf8_general_ci	select,insert,update,references	
Factura de anticipo	varchar(30)		YES	utf8	utf8_general_ci	select,insert,update,references	
Fax cel cliente	varchar(50)		YES	utf8	utf8_general_ci	select,insert,update,references	
Fecha de anulación	date		YES			select,insert,update,references	
Fecha de check-in	date		YES			select,insert,update,references	
Fecha de documento...	date		YES			select,insert,update,references	
Fecha de documento...	date		YES			select,insert,update,references	
Fecha de documento...	date		YES			select,insert,update,references	
Fecha de documento...	date		YES			select,insert,update,references	
Fecha de documento...	date		YES			select,insert,update,references	
Fecha de la reserva	date		YES			select,insert,update,references	
Fecha de llegada	date		YES			select,insert,update,references	
Fecha de modificació...	date		YES			select,insert,update,references	
Fecha de nacimiento	date		YES			select,insert,update,references	

Count: 144

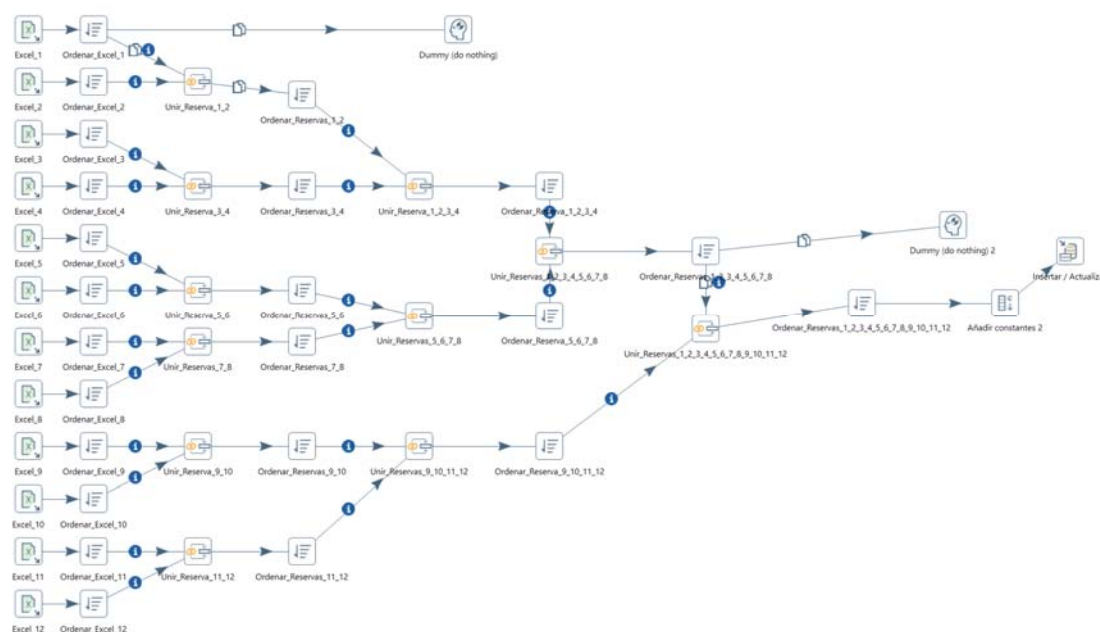
STG Smart Yield

3.1.2 Extracción y transformación de los datos.

Las fuentes de datos que estamos utilizando de los dos hoteles, como en su gran mayoría de este tipo de hoteles, procede del PMS. El PMS es el sistema de gestión de reservas en el cual como funciones principales tiene: generar nuevas reservas, gestionar cancelaciones, bloqueo de habitaciones, etc.

Tanto el Parque Santiago como Las Marismas de Corralejo, exporta el histórico de reservas en archivos Excel. Hemos optado por cargar el STG de MySQL con los dos últimos años completos 2018 - 2019 de reservas y cancelaciones, esto lo hacemos a través de una ETL.

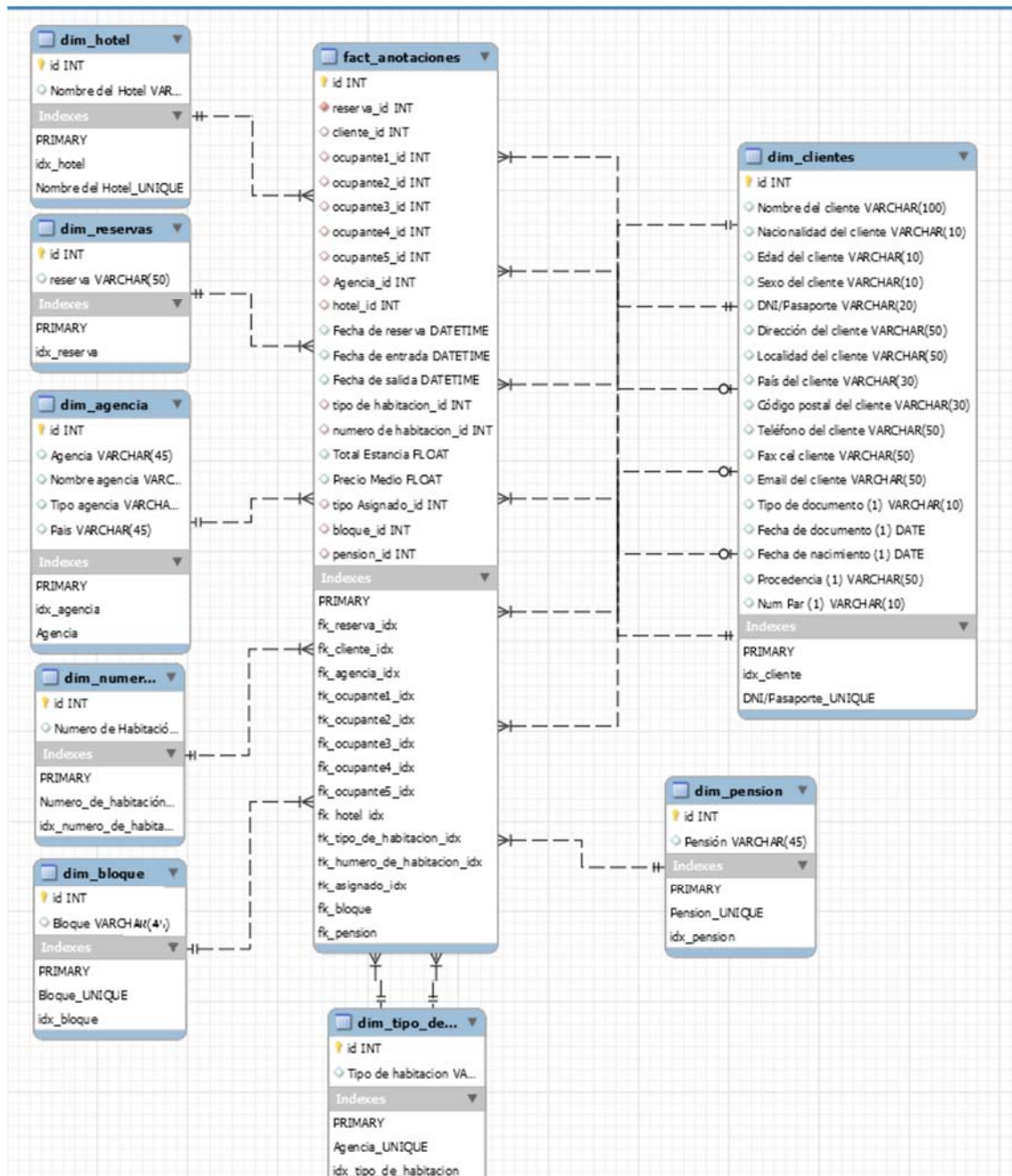
ETL es un proceso que permite mover datos de múltiples fuentes de datos, reformatearlos, limpiarlos y cargarlos en otra base de datos. En nuestro caso la ELT la creamos con Pentaho BI Suite.



ETL Fuente de datos Parque Santiago.

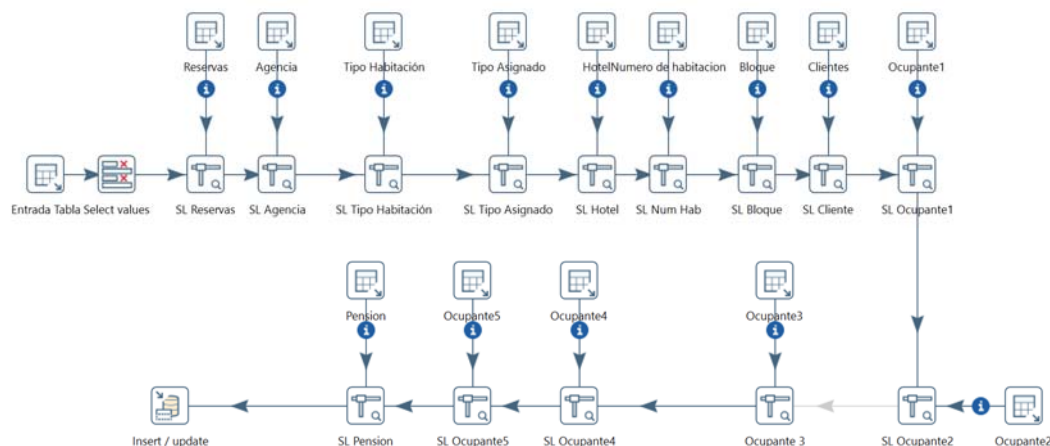
3.1.3 Modelo OLAP.

Una vez cargado los datos en el STG de MySQL, procedemos a crear el modelo OLAP en estrella. Un modelo OLAP, es un sistema utilizado con el objetivo de agilizar las consultas en grandes cantidades de datos. El esquema en estrella consiste en una tabla de hechos en el centro y varias tablas de dimensiones para cada punto de vista del análisis. El modelo OLAP es elaborado pensando en tener un modelo genérico para cualquier establecimiento hotelero, con unos campos comunes a todos.



Modelo OLAP estrella

Ya creado el modelo OLAP estrella en MySQL, procedemos con Pentaho BI Suite a cargar los datos del STG en todas las dimensiones y en la tabla de hechos.



Carga en Pentaho BI de la tabla de hechos

3.1.4 Automatización.

Todos los pasos comentados anteriormente para la carga de datos de los dataset al STG, del STG al OLAP, se automatizan con una serie de Jobs dentro de Pentaho BI Suite.



Jobs Inicial en Pentaho BI



Jobs carga a STG en Pentaho BI



Jobs carga a OLAP en Pentaho BI

3.1.5 Visualización.

Una vez finalizada la carga de datos en el Data Warehouse, se crean una serie de cuadros de mandos para mostrar la información de forma visual e interactiva. La herramienta que utilizamos para esto es PowerBI de Microsoft, la cual nos proporciona todos los servicios necesarios para crear la interfaz con la cual los usuarios finales interactúan con los datos del hotel.



Ejemplo Dashboard de gestión de Smart Yield

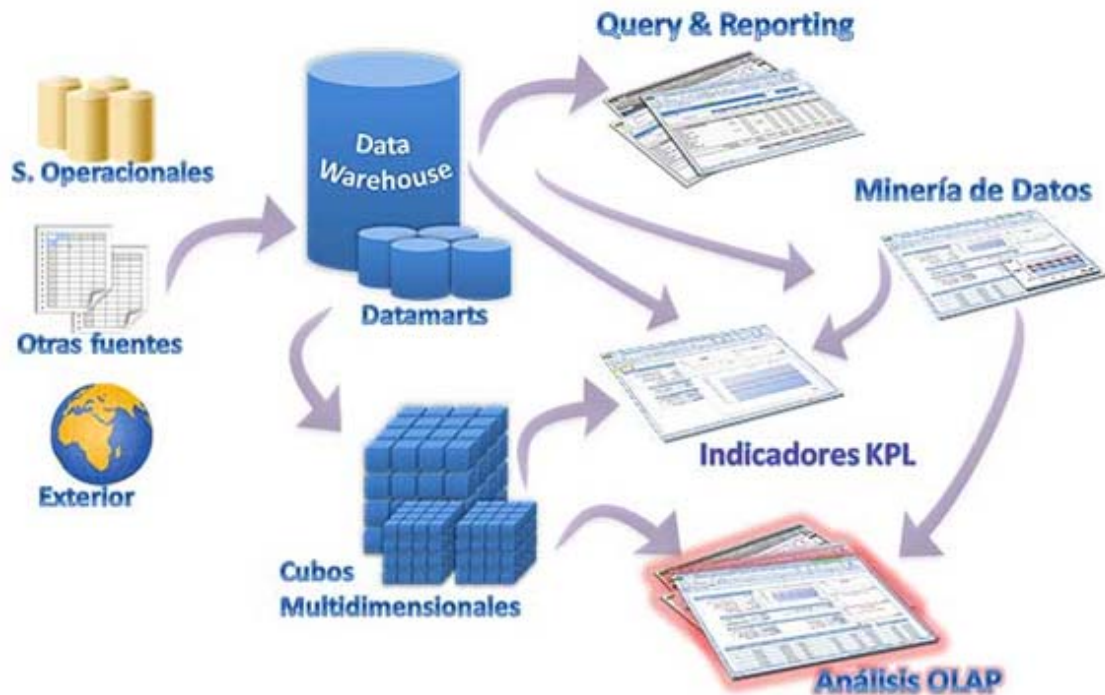
3.2 INTELIGENCIA DE NEGOCIO PARA REVENUE MANAGEMENT

3.2.1 Entorno de inteligencia de negocio y datos

En esta etapa de la solución tecnológica del sistema SamrtYield ya disponemos de una base de datos amplia y perfectamente estructurada de la cual partir hacia una solución de inteligencia de negocio que ayude a los procesos del Revenue del Hotel.

Nuestra ETL ya ha procesado nuestro esquema OLAP, el cual hemos elegido por ser más rápido a la hora de ejecutar sentencias SQL de tipo **SELECT**, en contraposición con OLTP que es la mejor opción para operaciones de tipo **INSERT, UPDATE Y DELETE**

En los sistemas OLAP se encuentra el concepto de cubo OLAP (también llamado **cubo multidimensional**). Se compone de hechos numéricos o **medidas**, que se clasifican por dimensiones. En este caso se ha creado a partir de un esquema en estrella. Las medidas se obtienen de los registros de una tabla de hechos y las dimensiones se derivan de la dimensión de los cuadros.



Partiendo de la definición de revenue, como objetivo:

Vender las unidades hoteleras de la forma más rentable posible en el canal de distribución adecuada, con la mejor eficiencia de comisiones.

Utilizaremos todo el potencial de esta estructura para crear los paneles de inteligencia en la solución de software de **Power Bi.** es un lenguaje de expresiones de fórmulas **DAX.**

3.2.2 Conectando el Entorno de visualización.

Esta estructura nos servirá para crear unos paneles en PowerBi, para comenzar debemos de conectar nuestro Software a nuestra base de datos procediendo con uno de sus conectores y añadiendo calve y contraseña

Comienza a Cargar nuestra estructura OLAP y los datos en nuestro Archivo de PowerBI

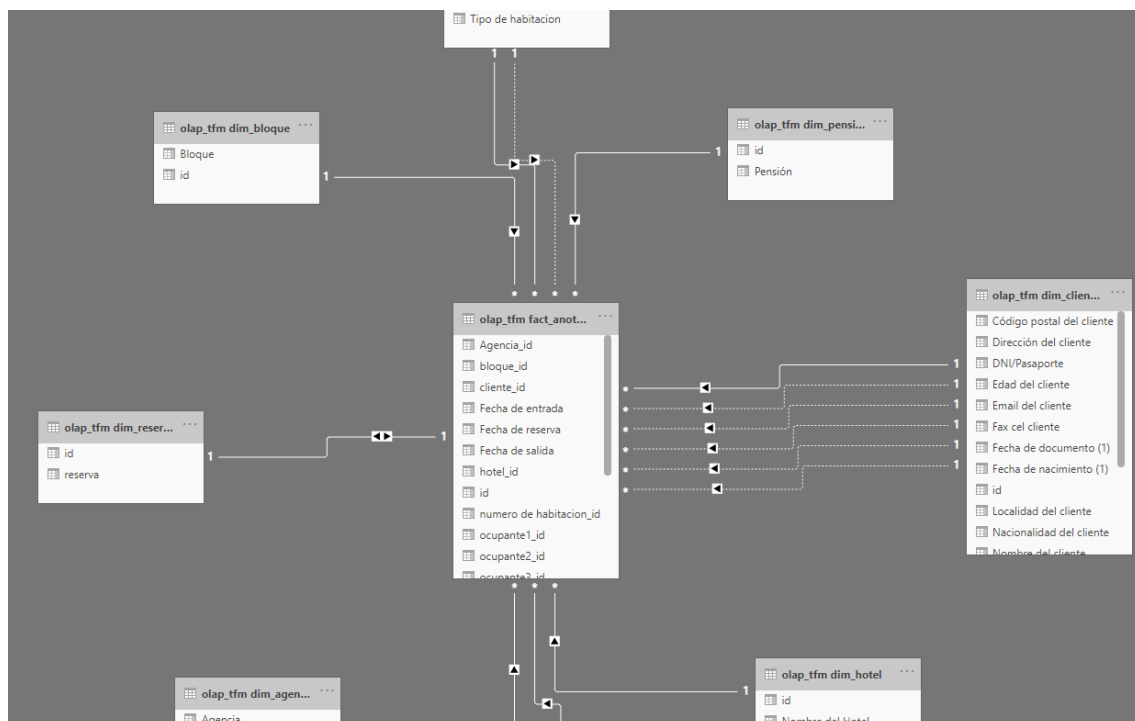
Cargar

- olap_tfm dim_agencia
1 fila de localhost/olap_tfm.
- olap_tfm dim_bloque
21 filas de localhost/olap_tfm.
- olap_tfm dim_clientes
Cargando datos al modelo...
- olap_tfm dim_hotel
4 filas de localhost/olap_tfm.
- olap_tfm dim_numero_de_habitacion
911 filas de localhost/olap_tfm.

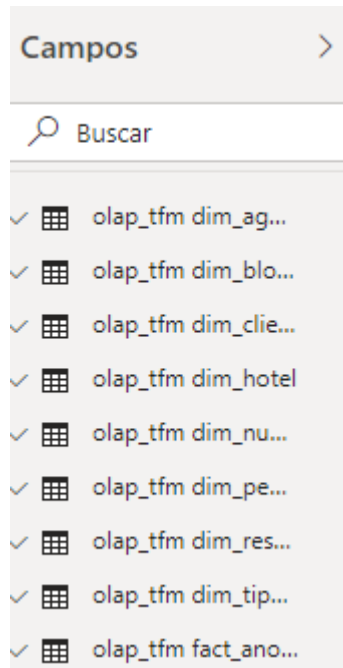


Cancelar

Una vez cargada nuestros datos podemos visualizar todo lo que hemos realizado en nuestra ETL y en MySQL Workbench: índices, índices foráneos... Sin que falte nuestra tabla de Anotaciones, sin la cual la magia de nuestros paneles no sería posible



En el panel de diseño disponemos de todas nuestra dimensiones y tabla de hecho para ir generando nuestros paneles o módulos de visualización.



3.2.3 Revenue management e inteligencia de negocio

Un proceso de revenue management excelente, implica: Información, cultura, herramientas, procesos operativos y alineamiento estratégico. Por tanto, más de la mitad del éxito de un proceso de revenue management podría ser derivado de una correcta gestión y tratamiento de los datos, tanto del hotel, como de datos de competencia y datos en abierto.

Una vez preparada la infraestructura debemos saber para que estamos creando los módulos y paneles de visualización, ahora veremos una definición de revenue management más cercana al Business Intelligence.

Conjunto de técnicas y herramientas de análisis que nos permiten maximizar la rentabilidad hotelera mediante la optimización de: Demanda, Inventario, Canales y Tarifas.

Esta parte de inteligencia de negocio de Smart Yield se podría traducir como un sistema de control de la una palabra mágica en el revenue y es “la certeza”. La certeza es el conocimiento claro y seguro de algo.

Quien tiene una certeza está convencido de que sabe algo sin posibilidad de equivocarse, aunque la certeza no implica veracidad o exactitud. Esto quiere decir un sistema de inteligencia puede afirmar que tiene una certeza y, sin embargo, la información que maneja es falsa o errónea. Por tanto, cobra más sentido la parte de ETL y DW de nuestra solución tecnológica.

Con nuestro sistema Smart Yield nos podemos alejar de la incertidumbre y acercarnos más a la certeza.

3.2.4 Incertidumbres & Certidumbres en el revenue

En este escenario evolutivo queda clara la determinación de las incertidumbres que un sistema de inteligencia de negocio para revenue management debe atender y controlar, fortaleciendo y maximizando el retorno de las certezas de las que se dispone.

Principalmente las expuestas en la siguiente tabla:

Incertidumbres	Certidumbres
Llegada futura de visitantes al destino y demanda de habitaciones del hotel	Histórico de distribución de las reservas por tipo de petición o canal
	Stock del hotel (tipos de habitación)
	Pickup por estatus de reserva
	Histórico porcentaje de confirmaciones sobre peticiones por tipo de reserva
Set competitivo ponderado	Tarifas (restricciones) y penalizaciones
Elasticidad-precio de la demanda	Obligaciones contractuales
Elasticidad-precio de la oferta del set competitivo	Elasticidad-precio de la oferta del hotel
Incidencia de <i>no shows</i> y cancelaciones	Histórico duración media de estancia y personas por habitación
	Histórico coste variable directos por habitación ocupada y márgenes directos por servicios accesorios

3.2.5 Funcionalidad para Revenue Management

Las principales funcionalidades de Smart Yield nos facilitara las siguientes acciones de revenue management, entre otras:

- » Segmentación de mercado.
- » Análisis del pasado: demanda y tendencias.
- » Previsiones/forecast.
- » Estrategias de precios, gestión de la duración y gestión de las disponibilidades.
- » Estrategias de overbooking.
- » Sistemas de información.

Estas funcionalidades no deben ni se pueden considerar estáticas, sino más bien recurrentes y procedimentales, es decir:



Por lo cual, nuestro sistema de obtención, procesamiento y visualización de datos es extremadamente potentes tanto en este proceso hotelero como en cualquier otro donde los datos sean cambiantes.

3.2.6 Principales indicadores del Revenue management

Los KPI más utilizados para el revenue management son:

$$\% \text{ ocupación} = \frac{\text{Habitaciones vendidas}}{\text{Habitaciones disponibles}}$$

$$\text{ADR (Average Daily Rate)} = \frac{\text{Ingresos totales por habitaciones}}{\text{Número de habitaciones vendidas}}$$

$$\text{RevPAR (Revenue per available room)} = \frac{\text{Ingresos totales habitaciones}}{\text{Habitaciones disponibles}}$$

$$\text{RevPAR} = \text{ADR} \times \% \text{ ocupación}$$

$$\text{GOP (Gross Operating Profit)} = \text{Cifra de Negocio} - (\text{Costes Fijos} + \text{C. Variables})$$

$$\text{GOPPAR (GOP per available room)} = \frac{\text{Cifra de Negocio} - (\text{C. Fijos} + \text{C. Variables})}{\text{Habitaciones disponibles}}$$

3.2.7 Módulos/Paneles de Inteligencia para revenue management

Se han desarrollado módulos y paneles de control para monitorizar analizar y realizar predicciones de numerosos parámetros, esta es muestra de alguno de ello:

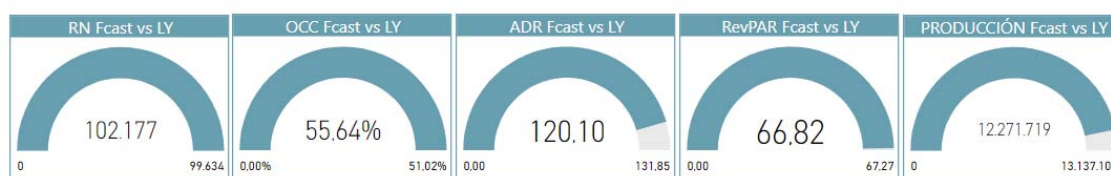
3.2.8 Estado del la ocupación y ADR

Fecha Reserva	OCC	ADR	BC OCC	BC ADR
28/11/2019	22,34 %	86,98	0,04 %	-0,01
27/11/2019	22,31 %	86,99	0,07 %	-0,04
26/11/2019	22,23 %	87,03	0,19 %	-0,06
25/11/2019	22,04 %	87,08	0,26 %	-0,04
24/11/2019	21,79 %	87,12	0,02 %	0,02
23/11/2019	21,76 %	87,10	0,27 %	-0,14
22/11/2019	21,49 %	87,24	0,41 %	-0,08
21/11/2019	21,08 %	87,32	0,26 %	-0,19
20/11/2019	20,83 %	87,54	0,23 %	-0,05
Total	22,34 %	86,98	22,34 %	-0,01

3.2.9 Evolución de la ocupación



3.2.10 Previsiones vs año anterior de cualquier periodo de tiempo marcado en los filtros.

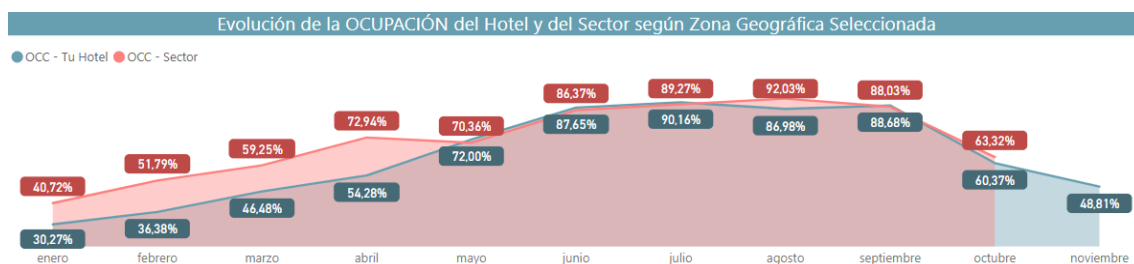


3.2.11 Diferencias temporales.

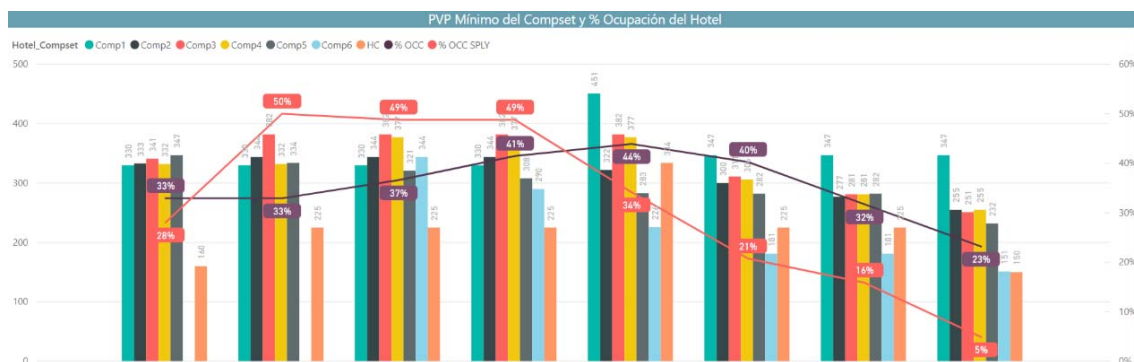
Diferencia Ocupación entre Real y Año Anterior a la fecha						
Mon	Tue	Wed	Thu	Fri	Sat	Sun
			1	2	3	4
				-1.31 %	-3.93 %	2.24 %
5	6	7	8	9	10	11
-1.02 %	0.20 %	5.76 %	9.21 %	2.83 %	-0.48 %	3.17 %
12	13	14	15	16	17	18
-7.02 %	-6.60 %	2.26 %	-1.84 %	-2.30 %	-0.78 %	6.49 %
19	20	21	22	23	24	25
3.68 %	4.87 %	2.63 %	-1.11 %	-6.54 %	-0.94 %	-0.19 %
26	27	28	29	30		
0.73 %	-0.13 %	-1.05 %	2.35 %	7.10 %		

3.2.12 Análisis de mercado.

3.2.12.1 Datos en abierto.



3.2.12.2 Análisis de competidores.



3.3 DESARROLLANDO EL MODELO DE CANCELACIONES

Una vez que hemos cargado y depurado todos los datos de las reservas de Parque Santiago 3, Parque Santiago 4 y Las Marismas del Corralejo, procedemos al desarrollo de un modelo de ML capaz de predecir cuándo se cancelará una reserva en función de varios parámetros que describiremos a continuación.

Por lo tanto, nuestro trabajo se centrará en realizar diversas pruebas con diferentes modelos y parámetros para ver qué modelo predice mejor nuestro target.

3.3.1 Data cleansing

Una vez que hemos cargado los datos en el entorno Anaconda mediante un conector SQL que nos permite conectarnos a la base de datos que hemos desarrollado en los apartados anteriores, utilizamos el lenguaje Python y las librerías NUMPY, MATPLOTLIB y PANDAS como librerías principales para este apartado. Estas librerías nos permiten trabajar específicamente con funciones matemáticas, tratar con dataframes y visualizar datos.

```
import numpy as np
#crear un dataframe con los nulos y mostrarlos ordenados
df_null=pd.DataFrame({'Campos':list(df.columns),'nº de nulos':df.isnull().sum()})
df_null=df_null.sort_values('nº de nulos')
df_null
```

		Campos	nº de nulos
Numero de reserva	Numero de reserva		0
Nombre del Hotel	Nombre del Hotel		0
Importe prepago/depósito	Importe prepago/depósito		0
Tipo de habitación	Tipo de habitación		0
Fecha de la reserva	Fecha de la reserva		0
Cancelación	Cancelación		0
Fecha de llegada	Fecha de llegada		0
Fecha de salida	Fecha de salida		0
Nombre del cliente	Nombre del cliente		7
Numero de personas	Numero de personas		173
Tipo de pension	Tipo de pension		195

Una vez que detectamos los valores nulos, los extraemos y formamos un dataframe nuevo con estos valores. Además, consideramos necesario extraer información de las fechas, de tal forma que usaremos el día y el mes como variables independientes.

```
df_1['mes_de_llegada'] = df_1['fecha_de_llegada'].dt.month
df_1['dia_de_llegada'] = df_1['fecha_de_llegada'].dt.day
df_1['mes_de_salida'] = df_1['fecha_de_salida'].dt.month
df_1['dia_de_salida'] = df_1['fecha_de_salida'].dt.day
df_1['mes_de_la_reserva'] = df_1['fecha_de_la_reserva'].dt.month
df_1['dia_de_la_reserva'] = df_1['fecha_de_la_reserva'].dt.day
```

Nos fijamos en las variables categóricas y revisamos si existen algunos valores que simplemente no deberían estar.

```
#Seleccionamos todas las agencias menos estas 4
df_1=df_1[~df_1.agencia.isin(['PROPI', 'PROPI4', 'OBRAS', 'REFOR'])]
```

Convertimos en variables categóricas aquellas que son de tipo Objeto

```
#convertimos en variables categóricas aquellas que son de tipo objetos
cat_features = df_1.columns[df_1.dtypes=='object']
num_features = df_1.columns[df_1.dtypes!='object']
|
for col in cat_features:
    df_1[col] = pd.Categorical(df_1[col])
```

A continuación, transformamos aquellas variables categóricas que están compuestas de texto en binarias, de cara a implementarlo en los modelos. Lo que hemos hecho aquí es que en las variables que contienen dos posibles valores, le asignamos a uno de ellos el valor 1, y el otro posible valor será 0.

```
#como sólo hay dos opciones donde elegir en cuanto a nombre del hotel,
#usamos esto para asignar '1' al valor Parque Santiago 3 y al tipo de pension SS

df_1.nombre_del_hotel = df_1.nombre_del_hotel.eq('PARQUE SANTIAGO 3').astype('int')
df_1.tipo_de_pension = df_1.tipo_de_pension.eq('SS').astype('int')
df_1.head()
```

Para aquellas variables categóricas donde hayan más de 2 opciones, utilizamos la función `get_dummies()`, que nos convierte estas opciones en columnas donde estarán a 1 si se daba ese valor o 0 si no se daba en esa fila.

agencia__VITA	agencia__VSIERR	agencia__WEB
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1

Por último, observamos el porcentaje de Cancelaciones que se dieron frente a las que no se dieron, y las variables más correlacionadas con la columna de Cancelación.

```
df_1.cancelacion.value_counts().div(df_1.shape[0] / 100)
```

```
0    70.471533
1    29.528467
Name: cancelacion, dtype: float64
```

```
#observamos las correlaciones con la variable cancelacion
corr.loc['cancelacion'].sort_values(ascending=False)
```

```
cancelacion          1.000000
agencia__COSMOS      0.156901
agencia__DIR         0.148171
agencia__SOME        0.143127
nombre_del_hotel     0.103409
agencia__OTB         0.081232
agencia__TLT         0.077396
agencia__ARTURO      0.075534
```

3.3.2 Desarrollamos y estudiamos los modelos de ML

A continuación, nos disponemos a realizar el estudio de varios modelos y variantes de estos con la idea de estudiar cuál es su desempeño a la hora de predecir. Para ellos nos basaremos en las métricas preestablecidas como el Accuracy, Precision o F1-Score, que vienen derivados de la Matriz de Confusión, la cual nos indica cómo de bien está detectando los positivos y los negativos.

VALORES PREDICCIÓN		VALORES REALES
Verdaderos positivos	Falsos positivos	
Falsos negativos	Verdaderos negativos	

1. La Exactitud (accuracy)

La Exactitud (en inglés, “Accuracy”) se refiere a lo cerca que está el resultado de una medición del valor verdadero. En términos estadísticos, la exactitud está relacionada con el sesgo de una estimación. También se conoce como Verdadero Positivo (o “True positive rate”). Se representa por la proporción entre los positivos reales predichos por el algoritmo y todos los casos positivos.

En forma práctica la Exactitud es la cantidad de predicciones positivas que fueron correctas.

2. La Precisión (precision)

La Precisión (en inglés “Precision”) se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión. Se representa por la proporción entre el número de predicciones correctas (tanto positivas como negativas) y el total de predicciones.

En forma práctica es el porcentaje de casos positivos detectados.

3. La Sensibilidad (“Recall” o “Sensitivity”),

La sensibilidad y la especificidad son dos valores que nos indican la capacidad de nuestro estimador para discriminar los casos positivos, de los negativos. La sensibilidad es la fracción de verdaderos positivos, mientras que la especificidad, es la fracción de verdaderos negativos, aunque en nuestro estudio vamos a obviar la especificidad.

También se conoce como Tasa de Verdaderos Positivos (True Positive Rate) ó TP. Es la proporción de casos positivos que fueron correctamente identificadas por el algoritmo.

4. El F1 Score

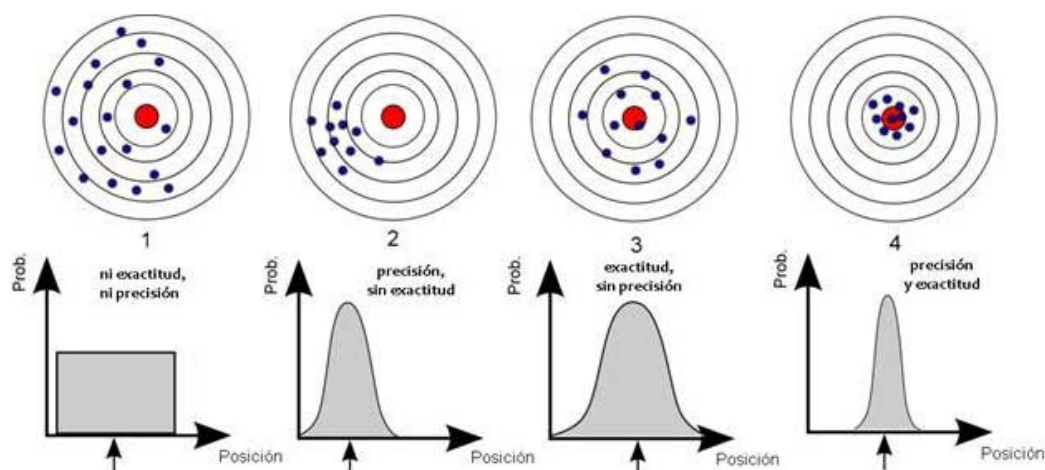
Esta es otra métrica muy empleada porque nos resume la precisión y sensibilidad en una sola métrica. (Precision y Recall en una sola métrica) por ello es de gran utilidad cuando

la distribución de las clases es desigual, por ejemplo, cuando el número de pacientes con una condición es del 15% y el otro es 85%

Conforme a estas nuevas métricas podemos obtener cuatro casos posibles para cada clase:

- Alta precisión y alto recall: el modelo maneja perfectamente esa clase.
- Alta precisión y bajo recall: el modelo no detecta la clase muy bien, pero cuando lo hace es altamente confiable.
- Baja precisión y alto recall: El modelo detecta bien la clase, pero también incluye muestras de la otra clase.
- Baja precisión y bajo recall: El modelo no logra clasificar la clase correctamente.

Cuando tenemos un “dataset” con desequilibrio, suele ocurrir que obtenemos un alto valor de precisión en la clase Mayoritaria y un bajo recall en la clase Minoritaria. Esto es mayormente lo que nos pasa en todas las predicciones pues como hemos visto anteriormente, la proporción de reservas canceladas y no canceladas es de 30/70.



La precisión es un gran estadístico, pero es útil únicamente cuando se tienen “datasets” simétricos (la cantidad de falsos negativos y de falsos positivos es similar).

El indicador F1 de la matriz de confusión es útil si se tiene una distribución de clases desigual.

Elegimos mayor Accuracy para conocer qué tan seguro estamos de los verdaderos positivos, mientras que la sensibilidad o “Recall” nos podría servir para saber si no está perdiendo positivos.

Por ejemplo, si creemos que es mejor en nuestro caso tener falsos positivos que falsos negativos, utilizamos una sensibilidad alta (Recall), cuando la aparición de falsos negativos resulta inaceptable pero no importa tener falsos positivos adicionales (falsas alarmas).

Con lo anteriormente explicado, presentamos los siguientes resultados de estas métricas para cada modelo probado de Decision Tree y Random Forest:

	Decision Tree							
	1(Sí hubo cancelación)				0(No hubo cancelación)			
	Recall	Precisión	Accurac	F1-Score	Recall	Precisión	Accurac	F1-Score
Base	0,4	0,5	0,70467	0,44	0,83	0,77	0,70467	0,8
Reduced features	0,2	0,57	0,71964	0,3	0,94	0,74	0,71964	0,82
Randomizedsearchcv	0,19	0,62	0,72673	0,29	0,95	0,74	0,72673	0,83
Gridsearchcv	0,2	0,68	0,73576	0,31	0,96	0,74	0,73576	0,84

	Random Forest							
	1(Sí hubo cancelación)				0(No hubo cancelación)			
	Recall	Precisión	Accurac	F1-Score	Recall	Precisión	Accurac	F1-Score
Base	0,41	0,53	0,71821	0,46	0,85	0,77	0,71821	0,81
Reduced features	0,22	0,58	0,72295	0,32	0,93	0,74	0,72295	0,83
Randomizedsearchcv	0,17	0,74	0,75113	0,28	0,98	0,74	0,75113	0,84
Gridsearchcv	0,29	0,69	0,75063	0,4	0,95	0,76	0,75063	0,84

3.3.3 Analizando los resultados

En nuestro caso se trata de una distribución no equilibrada, ya que tenemos el 30% de reservas con cancelaciones realizadas y 70% de reservas con cancelaciones no realizadas. Por tanto, nuestro scoring principal sería el **F1 Score**. Por lo tanto, fijándonos en esta métrica observamos que a medida que aplicamos diferentes métodos para buscar el mejor estimador, mejoramos sustancialmente en la detección como en el acierto de que sean 0 (No cancelaciones), pero el rendimiento en los 1(Cancelaciones) decae a la hora de detectar aquellas reservas que sí fueron canceladas (Bajo recall), no así el acierto, que si mejora. Por lo tanto, en este estudio tenemos un modelo más equilibrado, que correspondería con el Random Forest básico, pero elegiremos el Random Forest con gridsearchcv, detecta y acierta muy bien las no cancelaciones, y además tiene un F1 Score de las cancelaciones más alto que el resto de las variantes de randomforest. Aun así, se debería seguir trabajando en los modelos para obtener unos mejores resultados en la detección de las cancelaciones que sí se hicieron efectivas.

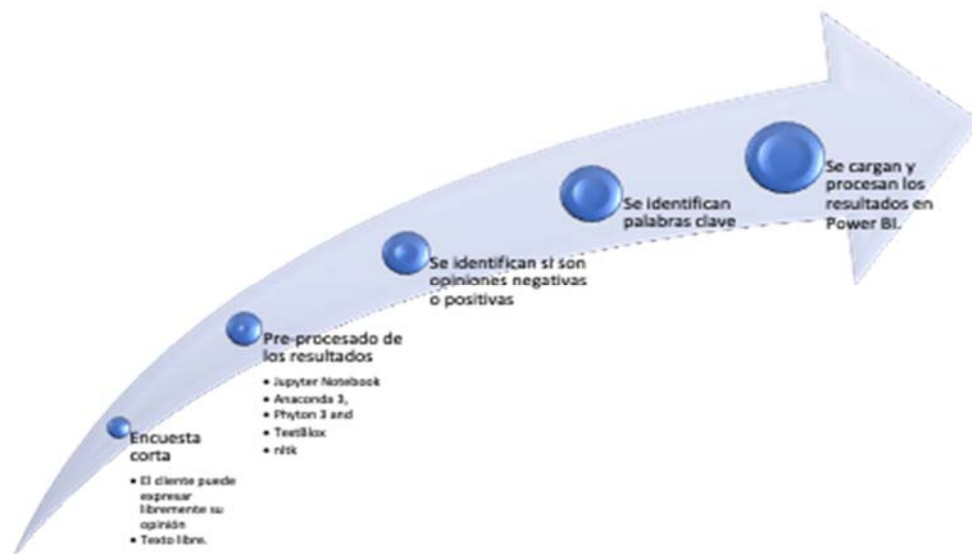
3.4 DESARROLLANDO EL MODELO DE ANÁLISIS DE SENTIMIENTOS

Para nuestro caso de estudio, en Smart Yeild nos concentramos en un subproblema del campo del NLP; la clasificación de textos en positivos o negativos, según se entienda que el encuestado está hablando mal (negativo) o bien (positivo) de un determinado tema (wifi, limpieza, etc.). Conocer la opinión de nuestros clientes es crucial para mejorar nuestros servicios de modo eficiente.

Para ello el hotel ha de pasar una encuesta a sus clientes y Smart Yeild procesará los datos, los analizará y mostrará en el Dashboard de forma personalizada, mediante el uso y diseño de interfaz cómoda y funcional.

El reto es trabajar con información no estructurada presentada en texto libre y expresada mediante Lenguaje Natural, se entiende que el cliente se siente más cómodo pudiendo expresar libremente su opinión. La gran dificultad que entraña esta técnica es que no tiene una estructura conocida como por ejemplo los lenguajes de programación, o datos en una base de datos MySQL.

El proceso seguido para realizar las encuestas, analizado y procesado los datos es el siguiente:



A continuación, explicaremos el proceso con detalles

3.4.1 Encuesta a los clientes

La mejor manera de conocer la opinión de nuestros clientes es preguntar directamente en la fuente, dichas encuestas serán definidas junto con el cliente para personalizar y adaptar a sus preferencias, servicios, etc. En la actualidad se ha utilizado datos de los 2 pilotos de Smart Yeild, donde podemos conocer lo que los clientes han escrito

3.4.2 Pre-procesado de datos

Ésta es el pase (junto con el procesado) del problema que toma una mayor importancia. Para el pre-procesado de los datos (carga de las encuestas, limpieza del texto, etc..) Se ha utilizado las principales herramientas y librerías.

Se consideraron varias herramientas para el Procesamiento del Lenguaje Natural, las cuales fueron evaluadas según las ventajas y desventajas para el desarrollo del proyecto teniendo en cuenta los objetivos que se deseaban alcanzar. Tiempo de carga, facilidad de uso, lenguaje a analizar, etc. Las herramientas seleccionadas y con las que se ha desarrollado la solución final son las siguientes:

- Jupyter Notebook Anaconda 3: Open-source Editor que nos permite de realizar ciencia de datos y aprendizaje automático de Python / R
- Python: Python es un lenguaje de programación interpretado, de alto nivel y de propósito general.
- NLTK: Natural Language Toolkit o NLTK es un paquete de Python para el procesamiento del lenguaje natural que nos permite.
- textblob.

3.4.3 Obtención de los datos o Data loading

La obtención del texto a procesar puede realizarse de muchas maneras, existiendo varias opciones para cada fuente de texto que se puede presentar. En nuestro caso de ha optado por almacenar todas las encuestas en un directorio específico y que podrá ser configurable.

```
# importing os module
import os

#ruta de trabajo para evitar rutas absolutas
present_working_directory = %pwd
print(present_working_directory)

with os.scandir(present_working_directory) as entries:
    for entry in entries:
        if entry.name.endswith('.txt'):
            print(entry.name)
            file = open(entry.name)
```

Una vez almacenadas las encuestas en nuestra carpeta específica, procedemos a la lectura de los datos:nada ni cads

```
lines = file.readlines()
file.close()
cadena = ''.join(lines)

cadena = cadena.replace("\n", ".\n");
cadena = cadena.replace("-", "");
cadena = cadena.replace("?", ".");
cadena = cadena.replace("!", ".");
cadena = cadena.replace("...", ".");
cadena = cadena.replace("...", ".");
cadena = cadena.lower()
print(cadena)
```

Una vez leídos todos los ficheros pendientes de procesar limpiamos una serie de caracteres que dificultan el encontrar el inicio y fin de frase correctamente.

3.4.4 Procesado de datos

Analizamos nuestro texto con la librería, previamente importada, textblob:

```
blob = TextBlob(str1)|
```

Esto nos permite visualizar los tags, lista de palabras y frases que tenemos, esto nos permite hacernos una primera idea de lo que tenemos y confirmar que la carga de datos se ha realizado correctamente:

```
blob.tags  
blob.noun_phrases  
for sentence in blob.sentences:  
    print(sentence)
```

Frases a analizar:

```
better wifi .  
would like the pools to be un salty.  
better internet .  
had difficulties with chanel 4 and wifi keeps dropping. otherwise we have stayed at the resort about 7 times and keep coming  
back as we love the resort and atmosphere. no need for any changes.  
don't change your customers for the safe. gives a cheap impression for a nice hotel like yours. upgrade the internet access for  
speed and security.  
none .  
a more prompt service to issues reported. wifi down for 6 days in room, only responded to because electricity went off in the  
room on 6th day. disappointed by psantiagos attitude to this. as access to wifi is of extreme importance to us as a family. o  
ther than that, we love it here, as this is our 6th time to visit. looking forward to seeing it all renovated and modernised  
in the near future hopefully. .  
improve the wifi and also if you could leave washing up liquid so we could wash up in the apartment .  
could have tea and coffee in room also few toiletries.  
slow wifi in room .  
a hook for the towel in the kitchen.  
add a micro wave.  
the glasses for water, milk, drinks etc are to big.  
fix the broken paint in the pool, moore steady wifi .  
heat the big pool at this time of year. improve wifi.
```

Analizamos sentimentalmente nuestro texto teniendo en cuenta varias cosas. Por un lado, los valores de sentimiento y objetividad, donde ambos tienen un rango de -1 a 1:

Entendemos que es más eficiente quedarnos con valores más cercanos a los extremos ya que un valor entre -0.2 y 0.2 podría ser dado como neutro, así mismo es interesante que el valor de objetividad sea cercano a 1, esto nos diría que la clasificación es correcta (en términos estadísticos).

En la siguiente tabla mostramos los resultados obtenidos después del procesado ordenados por su polaridad:

```
blob = TextBlob(str1)
# blob.tags

# Each word in the lexicon has scores for:
# 1) polarity: negative vs. positive (-1.0 => +1.0)
# 2) subjectivity: objective vs. subjective (+0.0 => +1.0)

output = []
keyword = '';
result = 'Not relevant';

for sentence in blob.sentences:
    #output.append({"sentence":sentence,"Polarity":sentence.sentiment.polarity, "subjectivity":s
    doc_aux = nlp(str(sentence))
    keyword = '';
    for ent in doc_aux.ents:
        # print the entity text and its label
        #if (ent.Label_ == 'ORG' or ent.Label_ == 'PERSON'):
        keyword = keyword + ent.text + ent.Label_;

    if (sentence.sentiment.polarity < -0.2):
        result = 'NEGATIVE'
    elif (sentence.sentiment.polarity > 0.2):
        result = 'POSITIVE'
    else:
        result = 'Not Relevant'

    output.append({"sentence":sentence,"Polarity":sentence.sentiment.polarity,
                  "subjectivity":sentence.sentiment.subjectivity, "keyword":keyword,
                  "Result": result},
    )
```

```
output_pd = pd.DataFrame(output)
output_pd.sort_values(by=['Polarity'])
```

	sentence	Polarity	subjectivity	keyword	Result
193	(w,h,y,i,s,t,h,e,r,e,n,o,t,a,n,y,t,e,a,p,o,t,...,t,e,r,r,i,b,l,e,...)	-1.0	1.0		NEGATIVE
374	(e,l,a,g,u,a,c,a,l,i,e,n,t,e,i,n,s,u,f,i,c,i,e,n,t,e,h,a,s,t,a,p,a,r,a,d,o,s,e,l,w,i,f,i,h,o,r,r,i,b,l,e,y,r,u,i,d,o,s,c,añ,e,r,i,a,s,s,e,a,v,i,s,o,p,e,r,o,a,u,n,t,i,e,...)	-1.0	1.0	el el wifi	NEGATIVE
337	(w,h,y,i,s,t,h,e,r,e,n,o,t,a,n,y,t,e,a,p,o,t,...,t,e,r,r,i,b,l,e,...)	-1.0	1.0		NEGATIVE
243	(s,o,r,t,o,u,t,y,o,u,r,t,e,r,r,i,b,l,e,w,i,f,i,c,o,n,n,e,c,t,i,o,n,s,...)	-1.0	1.0		NEGATIVE
230	(e,l,a,g,u,a,c,a,l,i,e,n,t,e,i,n,s,u,f,i,c,i,e,n,t,e,h,a,s,t,a,p,a,r,a,d,o,s,e,l,w,i,f,i,h,o,r,r,i,b,l,e,y,r,u,i,d,o,s,c,añ,e,r,i,a,s,s,e,a,v,i,s,o,p,e,r,o,a,u,n,t,i,e,...)	-1.0	1.0	el el wifi	NEGATIVE
...	...	...	...	...	...
221	(e,v,e,r,y,t,h,i,n,g,w,a,s,w,o,n,d,e,r,f,u,l,a,l,l,t,h,e,s,t,a,f,f,w,e,e,n,c,o,u,n,t,e,r,e,d,w,e,r,e,w,o,n,d,e,r,f,u,l,...)	1.0	1.0		POSITIVE

Una vez obtenidos nuestros resultados los pasamos a un formato válidos (CSV) para poder cargarlos en PowerBI.

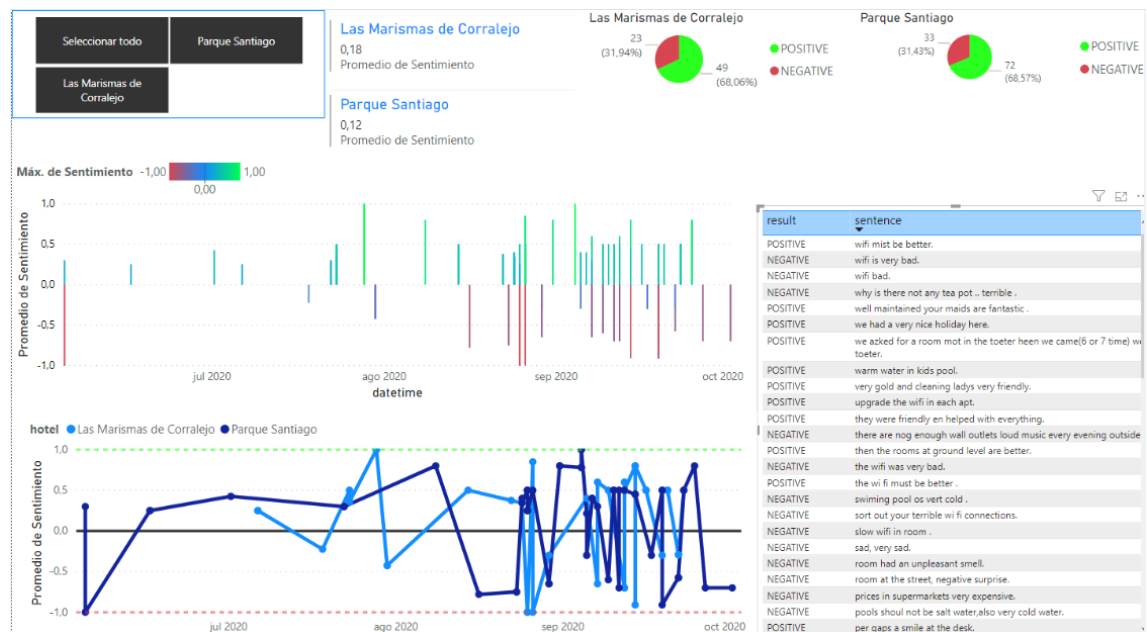
Ejemplo de fichero csv:

```
hotel,date,time,sentence,polarity,subjectivity,keyword,result
Parque Santiago,2020-09-22 15:10:12.717229,Better WiFi ...0.5,0.5,Better WiFi,POSITIVE
Las Marismas de Corralejo,2020-09-20 15:10:12.798229,Would like the pools to be an outdoor,0.0,0.0,unORG,POSITIVE
Parque Santiago,2020-09-18 15:10:12.765229,Better internet ...0.5,0.5,B,POSITIVE
Las Marismas de Corralejo,2020-09-11 15:10:12.794229,Bad difficutlty with Chanel 4 and wifi keeps dropping,0.0,0.0,Chanel 4ORG,POSITIVE
Parque Santiago,2020-08-18 15:10:12.772229,Otherwise we have stayed at the resort about 7 times and keep coming back as we love the resort and atmosphere,0.25,0.3,about 7CARDINAL,POSITIVE
Las Marismas de Corralejo,2020-09-11 15:10:12.783229,No need for any changes,0.0,0.0,,POSITIVE
Parque Santiago,2020-08-25 15:10:12.794229,Don't charge your customers for the wifi,0.0,0.0,,POSITIVE
Las Marismas de Corralejo,2020-08-13 15:10:12.804229,Gives a cheap impression for a nice hotel like yours,0.5,0.05,,POSITIVE
Parque Santiago,2020-09-15 15:10:12.812229,Upgrade the internet access for speed and security,0.0,0.0,,POSITIVE
Las Marismas de Corralejo,2020-09-29 15:10:12.817229,Bone ...0.0,0.0,,POSITIVE
Parque Santiago,2020-08-25 15:10:12.824229,A more prompt service to lessons reported,0.5,0.5,,POSITIVE
Las Marismas de Corralejo,2020-06-27 15:10:12.836229,Wifi Down for 4 days in room, only responded to because electricity went off in the room on 6th day,0.0,0.7777777777777778,0.6666666666666666
Parque Santiago,2020-08-25 15:10:12.844229,Disappointed by Fñantiago attitude to this,0.0,0.75,0.75,NEGATIVE
Las Marismas de Corralejo,2020-08-09 15:10:12.833229,As access to WiFi is of extreme importance to us as a family,0.0,0.125,1.0,WIFIORG,NEGATIVE
Parque Santiago,2020-07-30 15:10:12.841229,Mother than that, we love it here, as this is our 6th time to visit,0.1875,0.4875,eachCARDINAL,POSITIVE
Las Marismas de Corralejo,2020-07-10 15:10:12.870229,Looking forward to seeing it all renovated and modernized in the near future hopefully,0.05,0.2625,,POSITIVE
Parque Santiago,2020-08-28 15:10:12.858229,Improve the WiFi and also if you could leave washing up liquid so we could wash up in the apartment,0.0,0.0,,POSITIVE
Las Marismas de Corralejo,2020-08-12 15:10:12.892229,Could have Tea and coffee in room also few toiletries,0.2,0.1,TeaPERSON,NEGATIVE
```

3.4.5 Dashborad en PowerBI

La siguiente figura nos muestra el resultado de carga de nuestro fichero csv en un dashboard definido para la correcta lectura y transformación de los datos.

La figura nos muestra el menú principal (izquierda), selector de hotel y muestra de datos generales (derecha).



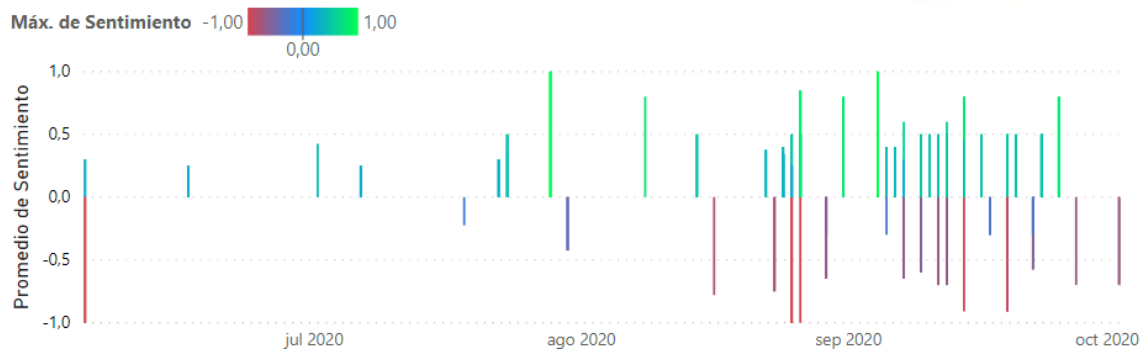
A continuación, procederemos a la descripción de cada una de las partes de nuestro dashboard NLP.

La siguiente gráfica nos ayuda a entender de una vez y rápidamente la opinión en términos generales de los encuestados respecto a nuestros hoteles. Donde si bien podemos ver que en términos generales hay más opiniones positivas, no parecen ser suficientes.



Por lo que tendremos que indagar más en las deficiencias de nuestros servicios, bien por fechas o por palabras clave, etc. Éste panes será total y completamente configurable, el servicio de mantenimiento del producto nos permitirá adaptar los parámetros a cada hotel y circunstancia.

Se han utilizado un código de colores internacional donde los datos de comentarios positivos se muestran en verde y los comentarios negativos en rojo, como se puede ver en la siguiente gráfica:



Así mismo también tendremos una sección, donde podremos listar las opiniones directamente, si se considera.

result	sentence
POSITIVE	wifi must be better.
NEGATIVE	wifi is very bad.
NEGATIVE	wifi bad.
NEGATIVE	why is there not any tea pot .. terrible .
POSITIVE	well maintained your maids are fantastic .
POSITIVE	we had a very nice holiday here.
POSITIVE	we asked for a room mot in the toeter heen we came(6 or 7 time) w toeter.
POSITIVE	warm water in kids pool.
POSITIVE	very gold and cleaning ledys very friendly.
POSITIVE	upgrade the wifi in each apt.
POSITIVE	they were friendly en helped with everything.
NEGATIVE	there are nog enough wall outlets loud music every evening outside
POSITIVE	then the rooms at ground level are better.
NEGATIVE	the wifi was very bad.
POSITIVE	the wi fi must be better .
NEGATIVE	swiming pool os vert cold .
NEGATIVE	sort out your terrible wi fi connections.
NEGATIVE	slow wifi in room .
NEGATIVE	sad, very sad.
NEGATIVE	room had an unpleasant smell.
NEGATIVE	room at the street, negative surprise.
NEGATIVE	prices in supermarkets very expensive.
NEGATIVE	pools shoul not be salt water.also very cold water.
POSITIVE	per gaps a smile at the desk.

4 RECURSOS FÍSICOS Y GESTIÓN DEL TIEMPO.

4.1 ANÁLISIS DE TALENTO HUMANO.

Este proyecto necesita perfiles técnicos, dentro del equipo de fundadores nos hemos repartido las áreas de la siguiente manera:

- Analista/consultor BI. Iván Yauci, experto en diseño de bases de datos y almacenamiento de DW, alto dominio de herramientas de reportes de información, gran conocimiento de ERP en sistemas gestores de bases de datos y diseño ETL.
- Científico de datos. Julio Daniel, experto en análisis de información, dedicado a analizar e interpretar grandes bases de datos.

- Responsable financiera y analista de calidad. Miriam, experta en proyectos de proveedores, se encarga de la búsqueda y contacto con los clientes, utiliza metodologías de aseguramiento de la calidad del software, con conocimiento y experiencia en las normas ISO.
- Desarrollador y soporte técnico. Marcos Díaz, experto en lenguajes de programación y manejo de BD. Proporcionará soporte técnico en la instalación y configuración de todo el entorno.

4.2 ANÁLISIS DE LOS RECURSOS.

Dentro del proyecto Smart Yield contaremos con los recursos hardware de los propios integrantes. A continuación, se describen los recursos hardware y software necesarios para el proyecto:

- Soporte de almacenamiento: Servidor, donde se aloja el DW. Contemplamos varias opciones: utilizar el propio del hotel, compra de un nuevo servidor, utilización de un servicio Cloud. El sistema de BD que se utilizara es mysql, con licencia GPL.
- Soporte de análisis: Python, Pentaho, PowerBI.
- Soporte colaborativo: Google Drive, Bitbucket, Trello.

4.3 GESTIÓN DEL TIEMPO.

Cada proyecto tendrá un contenido diferente según las condiciones y requisitos que determine el cliente. En cualquier caso, a continuación, planteamos un cronograma de trabajo tipo, donde tenemos que completar una serie de fases:

1 Fase, la identificación de los recursos por parte del clientes, sobretodo que opción toma para el servidor: utilizar un servidor propio, comprar un servidor nuevo, utilizar una opción Cloud. Instalar los programas necesarios para el correcto funcionamiento del producto (MySQL, Pentaho, PowerBI, servidor Web).

2 Fase, analizar de donde proviene los datos (reservas) que vamos a utilizar, configuración de una interfaz con el PMS para obtener los datos históricos de años anteriores.

3 Fase, donde cargamos los datos al Data Warehouse con una ETL para controlar todos los datos introducidos en el sistema.

4 Fase, se cargan los modelos Machine Learning y se configura los dashboard para su correcta visualización.

5 Fase, de testeo y puesta de funcionamiento.



4.4 INDICADORES CLAVES

	OBJETIVO	KPI	PERIODICIDAD
1	EFFECTIVIDAD DE LA HERRAMIENTA	% DE RENDIMIENTO DE LOS MODELOS ML	TRIMESTRAL
2	AUMENTO DE BENEFICIOS	% GASTOS DE EXPLOTACIÓN SOBRE EL TOTAL DE GASTOS	TRIMESTRAL
3	AUMENTAR CLIENTES	% AUMENTO DE NUEVOS CLIENTES	ANUAL
4	ANÁLISIS DEL MERCADO	% CUOTA DE MERCADO	TRIMESTRAL
5	SATISFACCIÓN DEL CLIENTE	ENCUESTA NIVEL DE SATISFACCIÓN DE LOS CLIENTES	ANUAL

4.5 BENEFICIOS TANGIBLES:

- Mejorar la adquisición de clientes y su conversión mediante el uso de la segmentación.
- Reducir la tasa de abandono de clientes, incrementar su fidelidad, teniendo en cuenta cuál es su valor.
- Incrementar los ingresos por el crecimiento de las ventas.
- Aumentar los resultados, consiguiendo que nuestros clientes actuales compren más productos o servicios.
- Aumentar la satisfacción de los clientes.
- Conocer mejor cuáles son las características demográficas de nuestra zona de influencia.
- Hacer crecer la participación de mercado.
- Mejorar aquellas actividades relacionadas con la captura de datos.
- Facilitar la adopción de los cambios en la estrategia.

- Medir la efectividad de las campañas rápidamente y ser capaces de hacer los ajustes durante el ciclo de vida de estas.
- Optimizar la atención a los clientes
- Optimizar las acciones de marketing
- Analizar a la competencia cuando establecemos precios.
- Crear nuevas oportunidades.

4.6 REDUCCIÓN DE COSTES:

- Negociar mejores precios con los proveedores, identificar a los proveedores más importantes, gestionar descuentos por cantidades de compra; permitir el análisis del cumplimiento de los proveedores y asignar las compras de acuerdo con ello; medir el nivel de calidad, servicio y precio.
- Reducir o reasignar el personal necesario para llevar a cabo los procesos.
- Incrementar la productividad con información más inmediata y mejor.
- Aumentar el control de costes.
- Disminuir los gastos.
- Eliminar ineficiencias y reducir los costes operativos.
- Eliminarlas sobreproducciones de productos.
- Acortar los tiempos de respuesta a las peticiones de informes. O Analizar los problemas, reparaciones y defectos y proveer la información para hacer un seguimiento y corregir los problemas recurrentes.
- Abaratar los costes de las acciones de marketing.

4.7 BENEFICIOS INTANGIBLES:

- Mejorar el acceso a los datos a través de consultas, análisis o informes.
- Información más actualizada.
- Dotar a la información de mayor precisión.
- Controlar mejor la información.
- Mayor integración de la información.

4.8 BENEFICIOS ESTRATÉGICOS:

- Mayor habilidad para identificar y nutrir a aquellos clientes con mayor potencial.
- Mayor habilidad para analizar estrategias de precios.
- Mejorar la toma de decisiones, realizándose de forma más rápida, informada y basada en hechos.
- Mayor visibilidad de la gestión.
- Dar soporte a las estrategias.
- Aumentar el valor de mercado.

5 CASH FLOW:

Inicialmente partimos de la base de que es un proyecto en empresa, pero dado que se ha comportado bien con los datos de dos hoteles, no descartamos crear un cashflow como producto en el proyecto

5.1 PRESUPUESTOS:

Se han tenido en cuenta, a modo general, los siguientes **Gastos**:

1. Instalaciones: Sería necesario el alquiler de oficinas,
2. Software, cloud storage y equipos técnicos (PC)
3. Personal: 4 especialistas durante los 5 años que dura el plan.

Estos gastos no se prevé que sufran modificaciones durante los cinco primeros años que dura el plan, no siendo necesario ampliaciones de plantilla ni suscripciones a sw de los 4 (Power BI Pro, etc.). Sólo variará los gastos de cloud storages que dependen del nº de proyectos vendidos. Durante el primer año, dichos gastos de cloud son asumidos por los socios e irán incluidos en el precio del producto/servicio. A partir del segundo año serán asumidos por el cliente.

- Sobre los gastos generales del punto anterior, se estima un incremento del 2% anual correspondiente a un IPC estimado.
- El salario medio de cada uno es de 40.000 €, con el correspondiente incremento del 2% anual (IPC estimado)
- Se estima un 30% del salario destinado a Seguridad Social, IRPF, etc.
- Como amortizaciones, sólo se amortizan los equipos informáticos de los 4 especialistas en 3 años.

COSTES					
Concepto	Año 1	Año 2	Año 3	Año 4	Año 5
Instalaciones	12.160,00 €	12.403,20 €	12.651,26 €	12.904,29 €	13.162,38 €
Infraestructura	3.000,00 €	4.080,00 €	4.161,60 €	4.244,83 €	4.329,73 €
Comercial	5.440,00 €	5.548,80 €	5.659,78 €	5.772,97 €	5.888,43 €
Personal	201.448,00 €	205.476,96 €	209.586,50 €	213.778,23 €	218.053,79 €
Otros	3.300,00 €	3.366,00 €	3.433,32 €	3.501,99 €	3.572,03 €
Total	225.348,00 €	230.874,96 €	235.492,46 €	240.202,31 €	245.006,35 €

Respecto a los **Beneficios**, nuestro principal beneficio es la venta del **Smart Yield: Hotel Data System**, al cual se le ha puesto un precio de venta de 87.000 euros por cadena hotelera, donde se incluye modelos de revenues, dashboards, modelos de gº operativa para optimizar recursos internos, modelos NPL sobre reseñas de encuestas de los propios hoteles para mejorar experiencia de cliente, etc. Así mismo, también incluye implantación y adaptaciones a las necesidades de cada cadena hotelera, ya que aunque el 80-90% del núcleo es el producto que estáis construyendo ahora, es muy probable que

cada cadena hotelera necesite realizar adaptaciones para optimizar y adaptar sus propios recursos.

- Se estima vender Smart Yield a 2 clientes, dedicando los 4 primeros meses a construir la aplicación y los otros 8 a implantación, resolución de incidencias, ajustes, etc. En los 2 clientes. Durante los años sucesivos se estima vender el producto en 3 clientes porque donde se incluirán las adaptaciones personalizadas que solicite cada uno.
- Los años siguientes se dará soporte y mantenimiento de la solución, dicho soporte será un 20% del coste del producto/año. Por tanto, a partir del año 2 se prevé el soporte de los productos vendidos en años anteriores + 3 nuevos clientes. En cambio, no es necesario crecer en plantilla para el incremento de soporte porque debido a la curva de aprendizaje que estáis adquiriendo, las implantaciones y adaptaciones del producto son cada vez más rápidas de hacer.
- Para ser competitivos, no estamos proyectando incremento alguno en el precio del servicio, que será el mismo durante los 5 años del plan.
- Realizareis una aportación inicial de 7.500 € por socio en concepto de capital social. No será necesario apalancarse (solicitar préstamo). Por tanto, los intereses en la cuenta de resultados serán siempre de 0 al no tener financiación y el impuesto sobre la renta considerado es del 29,5% .

INGRESOS					
Concepto	Año 1	Año 2	Año 3	Año 4	Año 5
Número de nuevos proyectos	2	3	3	3	3
Precio anual del producto	80.000,00 €	80.000,00 €	80.000,00 €	80.000,00 €	80.000,00 €
Total ingresos por Ventas	160.000,00 €	240.000,00 €	240.000,00 €	240.000,00 €	240.000,00 €
Soporte anual por producto	17.400,00 €	17.400,00 €	17.400,00 €	17.400,00 €	17.400,00 €
Total ingresos Soporte	- €	34.800,00 €	87.000,00 €	139.200,00 €	191.400,00 €
Total ingresos	160.000,00 €	274.800,00 €	327.000,00 €	379.200,00 €	431.400,00 €

Toda esta información podemos verla reflejada en la siguiente tabla:

GASTOS PRIMER AÑO

Concepto	Imp. Concepto	Subconcepto	Imp. Subconc.	Comentarios
Instalaciones	12.160,00 €	Alquiler de oficinas	6.000,00 €	500 € * 12 meses
		Suministros (electricidad, agua, etc.)	2.160,00 €	180 € * 12 meses
		Limpieza	3.600,00 €	300 € * 12 meses
		Material de oficina	400,00 €	
Infraestructura	3.000,00 €	Suscripciones software	1.000,00 €	
		Gastos Cloud proyecto	2.000,00 €	1.000 € * 2 proyectos el primer año
		Publicidad	2.400,00 €	200 € * 12 meses
Comercial	5.440,00 €	Atenciones con Clientes	400,00 €	
		Comunicaciones	2.400,00 €	200 € * 12 meses
		Mensajerías	240,00 €	20 € * 12 meses
		Sueldos y salarios	126.857,14 €	35.000 € de salario medio * 4 personas / 12 pagas * 12 meses
Personal	201.448,00 €	Pagos extras	21.142,86 €	35.000 € de salario medio * 4 personas / 2 pagas * 2 pagas extra
		Seguridad Social	44.400,00 €	30% sobre los sueltos, salarios y pagas extras
		Gastos de formación	2.800,00 €	700 € * 4 persona
		Gastos de desplazamiento	4.400,00 €	11 desplazamientos al año (no se considera agosto) * 100 € de coste * 4 personas
Otros	3.300,00 €	Dietas	1.848,00 €	11 desplazamientos al año * 42 € de dieta * 4 personas
		Asesoría jurídica y contable	2.400,00 €	200 € * 12 meses
		Otros gastos	900,00 €	
Amortizaciones	2.400,00 €	Equipos informáticos	2.400,00 €	1.800 € * 4 equipos. Como inversión, sólo primer año. A amortizar en 3 años
TOTAL GASTOS GENERALES	227.748,00 €		227.748,00 €	

BENEFICIOS ANUALES VENTA PROYECTO

Concepto	Importe anual	Comentarios
Venta de proyecto a cadena hotelera X	80.000,00 €	Adaptado a las necesidades de cada cliente.
Soporte de proyecto a cadena hotelera X tras implantación	17.400,00 €	1.450 € * mes. Incluye reentrenamientos, mejoras en los algoritmos, etc.

CAPITAL SOCIAL (SÓLO PRIMER AÑO)

Concepto	Importe anual	Comentarios
Aportación de socios	30.000,00 €	7.500 € * 4 socios

AMORTIZACION

Concepto	Inversión	% Anual	Valor Anual
Hardware	7.200,00 €	33,33%	2.400,00 €

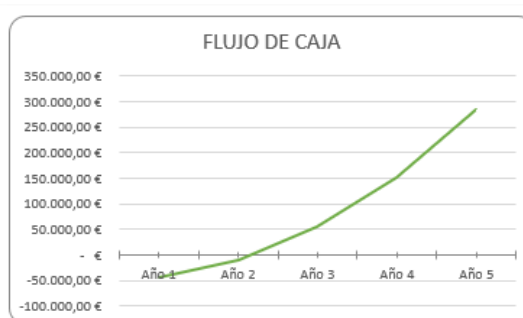
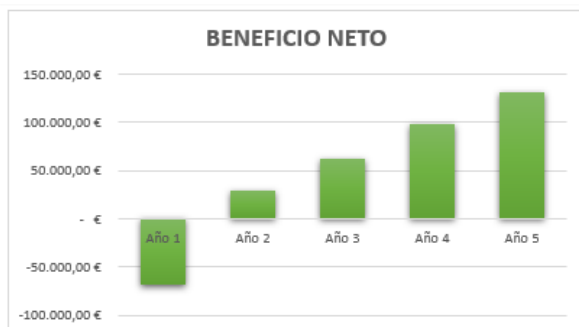
CUENTA DE RESULTADOS

Concepto	Año 1	Año 2	Año 3	Año 4	Año 5
Total ingresos	160.000,00 €	274.800,00 €	327.000,00 €	379.200,00 €	431.400,00 €
Gastos por Estructura	225.348,00 €	230.874,96 €	235.492,46 €	240.202,31 €	245.006,35 €
Margen Operativo (EBITDA)	- 65.348,00 €	43.925,04 €	91.507,54 €	138.997,69 €	186.393,65 €
Amortización	2.400,00 €	2.400,00 €	2.400,00 €	- €	- €
EBIT	- 67.748,00 €	41.525,04 €	89.107,54 €	138.997,69 €	186.393,65 €
Intereses	- €	- €	- €	- €	- €
EBT	- 67.748,00 €	41.525,04 €	89.107,54 €	138.997,69 €	186.393,65 €
Impuestos	- €	12.249,89 €	26.286,72 €	41.004,32 €	54.986,13 €
BENEFICIO NETO	- 67.748,00 €	29.275,15 €	62.820,82 €	97.993,37 €	131.407,52 €
MARGEN DE BENEFICIO	-42,34%	10,65%	19,21%	25,84%	30,46%

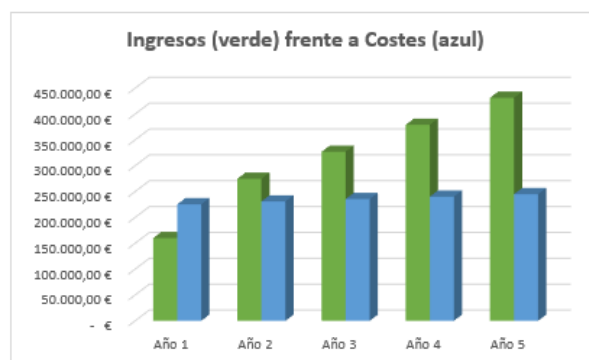
FLUJO DE CAJA

Concepto	Año 1	Año 2	Año 3	Año 4	Año 5
Aporte de los socios	30.000,00 €	- €	- €	- €	- €
Saldo Inicial	30.000,00 €	- 42.548,00 €	- 10.872,85 €	54.347,97 €	152.341,34 €
Total ingresos por Ventas	160.000,00 €	274.800,00 €	327.000,00 €	379.200,00 €	431.400,00 €
Total Ingresos	160.000,00 €	274.800,00 €	327.000,00 €	379.200,00 €	431.400,00 €
Hardware	- 7.200,00 €	- €	- €	- €	- €
Total Inversiones	- 7.200,00 €	- €	- €	- €	- €
Gastos por Estructura	- 225.348,00 €	- 230.874,96 €	- 235.492,46 €	- 240.202,31 €	- 245.006,35 €
Impuestos	- €	- 12.249,89 €	- 26.286,72 €	- 41.004,32 €	- 54.986,13 €
Total Gastos	- 225.348,00 €	- 243.124,85 €	- 261.779,18 €	- 281.206,63 €	- 299.992,48 €
SALDO FINAL	- 42.548,00 €	- 10.872,85 €	54.347,97 €	152.341,34 €	283.748,86 €

El proyecto arrancará con una aportación inicial de los socios de 30.000 euros que según nuestras previsiones se recuperarán a partir del segundo año.



TIR	103%	PayBack	2,34
			Al comienzo del segundo año
VAN	273.403,52 €	TIRM	76%



6 ANEXOS

6.1 DESARROLLO DE MODELO DE NLP PARA DETECTAR RESEÑAS POSITIVAS Y NEGATIVAS

6.2 DESARROLLO DE MODELO DE CLASIFICACIÓN PARA DETERMINAR CANCELACIONES

6.1 DESARROLLO DE MODELO DE NLP PARA DETECTAR RESEÑAS POSITIVAS Y NEGATIVAS

Import needed libraries to work with NLP. Ensure preconditions, spacy correctly installed: pip install -U spacy python -m spacy download en_core_web_md

In [2]:

```
import pandas as pd
pd.set_option('display.max_colwidth', 100000)
import unicodedata
import json
from nltk.corpus import stopwords
import statistics
import re
import spacy
from spacy.matcher import PhraseMatcher
from spacy.matcher import Matcher
import numpy as np
```

In [3]:

```
from textblob import TextBlob
import nltk
nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]      C:\Users\A177889\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]      C:\Users\A177889\AppData\Roaming\nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

Out[3]:

True

In [4]:

```
import en_core_web_sm
import csv
import datetime
```

In [5]:

```
nlp = spacy.load('en_core_web_sm')
```

In []:

```
stop_words = set(stopwords.words('english'))
print(stop_words)
```

In [9]:

```
# importing os module
import os

#ruta de trabajo para evitar rutas absolutas
present_working_directory = %pwd
print(present_working_directory)
cadena = '';

with os.scandir(present_working_directory) as entries:
    for entry in entries:
        if entry.name.endswith('.txt'):
            print(entry.name)
            file = open(entry.name)
            lines = file.readlines()
```

```

file.close()
cadena = ''.join(lines)

cadena = cadena.replace("\n", ".\n");
cadena = cadena.replace("-", "");
cadena = cadena.replace("?", ".");
cadena = cadena.replace("!", ".");
cadena = cadena.replace("..", ".");
cadena = cadena.replace("..", ".");
cadena = cadena.lower()
print(cadena)

```

C:\Users\A177889\OneDrive - Atos\04_Cursos\01 Master EOI\PROYECTO\NLP

sugerencias.txt

better wifi .

would like the pools to be un salty.

better internet .

had difficulties with chanel 4 and wifi keeps dropping. otherwise we have stayed at the resort about 7 times and keep coming back as we love the resort and atmosphere. no need for any changes.

don't charge your customers for the safe. gives a cheap impression for a nice hotel like yours. upgrade the interest access for speed and security.

none .

a more prompt service to issues reported. wifi down for 6 days in room, only responded to because electricity went off in the room on 6th day. disappointed by psantiagos attitude to this. as access to wifi is of extreme importance to us as a family. other than that, we love it here, as this is our 6th time to visit. looking forward to seeing it all renovated and modernised in the near future hopefully. .

improve the wifi and also if you could leave washing up liquid so we could wash up in the apartment .

could have tea and coffee in room also few toiletries.

slow wifi in room .

a hook for the towel in the kitchen.

add a micro wave.

the glasses for water, milk, drinks etc are to big.

fix the broken paint in the pool, moore steady wifi .

heat the big pool at this time of year. improve wifi.

norsk tv.

no.

badly needs refurbished .

to much noice from the restaurants.

no.

a big thank you to the cleaning staff and the reception staff. they were friendly en helped with everything. we had a very nice holiday here.

i come often .please check gregor.brucel@btinternet.com.. this never confers favour when getting a room south/west facing overlooking the pool.

better wifi connection .

less building work .

3020.

too much noise/music from the restaurants in th evening.

time to put in new beds and a new sofa in our room (1029).

had a fantastic time will definatly be back . fans ok but miss air conditioning would not return in the hottest months . well maintained your maids are fantastic .

if traveling with small kids the upstairs rooms are a bit dangerous. then the rooms at ground level are better. also we didn't know about the heated pool until yesterday. maybe a sign or map of the water temperature at the big pool would be nice .

make the rooms as modern as the rest of the area. exterier i great and a 4 star hotel. but abit disapointed tjat the standard in the rooms where so old. more like a 3 star hotel.

asked for sea view didn't get it and we are here for 3 weeks very dissatisfied.

i would recommand to make the bathroom a bit bigger so the shower would be more comfortable and maybe even remove the bidet. .

need hairdryer .

great stay.

upgrade wifi.

too much salt in the pool .

a good map showing best way to apartment.

ban large inflatable pool toys.

better wifi .

wifi needs to be attended to. it is a major problem all the time.

microwaves walkin showers refurbished as apartments are looking dated.

suggest one pool should be heated in santiago 3. would suggest not putting guests into rooms next door to rooms that are being renovated. wifi stopped working after 2 days in our

r room . otherwise a lovely resort .
very little vegetables available. bad quality om mondstuk.
the wi fi must be better .
wifi mist be better. bad in lanzarote.
old and bad beds. old and bad furnitures in all roms, especially in the kitchen.
why is there not any tea pot .. terrible . bad.
we asked for a room mot in the toeter heen we came(6 or 7 time) we hot a room in the toet
er. the reception had change or. but moe we have a room without sunshine.
nine.
can we book the same room, in your homesite, direct from home.
i want to use my hairdryer in the bathroom, not the kitchen. .
our rom have no sun in the morning and no sin in the evenong. sad, very sad. something we
will consortes next time we order. .
big broom for cleanup (sand , balcony) is missing.
better wifi connection .
swiming pool os vert cold .
very gold and cleaning ladys very friendly.
very gold and cleaning ladys very friendly.
very gold and cleaning ladys very friendly.
more selection on the tv and double beds .
have wifi available throughout.
staff at reception didn't do anything to try and resolve our issue. wonth ever stay here
again and will warn anyone else of coming to stay here.
ceiling fans not working .
ceiling fans not working.
microwave would be very welcome.
i only missed watercooker and micro wave. .
wifi is very bad.
you need showers in the bathroom instead of baths modern kitchens need to be fitted your
price is now getting too expensive paid 90 euròs for 1 bed apartment in 2016 now pay 140
euros a night in november but you haven't upgraded the rooms or wifi.
you need showers in the bathroom instead of baths modern kitchens need to be fitted your
price is now getting too expensive paid 90 euròs for 1 bed apartment in 2016 now pay 140
euros a night in november but you haven't upgraded the rooms or wifi.
everything was wonderful all the staff we encountered were wonderful. the housekeeping s
taff even set the kids teddy bears neatly on the beds. our only issue the whole week was
that the pools are colder than the ocean. .
be thee light in the bathroom.
poolrest.
prices in supermarkets very expensive.
none.
warmer pools.
beds too hard.
el agua caliente insuficiente hasta para dos,el wifi horrible y ruidos cañerías,se aviso
pero auntie vinieron no se arregló.
think you should have mircowave and a electric kettle in the rooms .
better wifi .
great like every year.
per gaps a smile at the desk.
more wifi power.
we were given a room next to 3028 which was being refurbished to a private client . very
noisy all day.
more comfortable beds. room had an unpleasant smell.
this list should not come day 1.
tv noruega falta.
room at the street, negative surprise.
wifi is terribel.
wifi is terribel.
fix the wifi it is really slow.
warm water in kids pool.
i hawed staid at parque santiago 24 times last 13 yrs so i hop better apart next time.
i should like "dynor" on the sittplats on the balkong and towels with the pools. .
better wifi .
would like the pools to be un salty.
better internet .
had difficulties with chanel 4 and wifi keeps dropping. otherwise we have stayed at the
resort about 7 times and keep coming back as we love the resort and atmosphere. no need
for any changes.
don't charge your customers for the safe. gives a cheap impression for a nice hotel like
yours. upgrade the interest access for speed and security.
none .
a more prompt service to issues reported. wifi down for 6 days in room, only responded to

because electricity went off in the room on 6th day. disappointed by psantiagos attitude to this. as access to wifi is of extreme importance to us as a family. other than that, we love it here, as this is our 6th time to visit. looking forward to seeing it all renovated and modernised in the near future hopefully. .
improve the wifi and also if you could leave washing up liquid so we could wash up in the apartment .
could have tea and coffee in room also few toiletries.
slow wifi in room .
a hook for the towel in the kitchen.
add a micro wave.
the glasses for water, milk, drinks etc are too big.
fix the broken paint in the pool, more steady wifi .
heat the big pool at this time of year. improve wifi.
norsk tv.
no.
badly needs refurbished .
too much noise from the restaurants.
no.
a big thank you to the cleaning staff and the reception staff. they were friendly and helped with everything. we had a very nice holiday here.
i come often .please check gregor.brucel@btinternet.com.. this never confers favour when getting a room south/west facing overlooking the pool.
better wifi connection .
less building work .
3020.
too much noise/music from the restaurants in the evening.
time to put in new beds and a new sofa in our room (1029).
had a fantastic time will definitely be back . fans ok but miss air conditioning would not return in the hottest months . well maintained your maids are fantastic .
if traveling with small kids the upstairs rooms are a bit dangerous. then the rooms at ground level are better. also we didn't know about the heated pool until yesterday. maybe a sign or map of the water temperature at the big pool would be nice .
make the rooms as modern as the rest of the area. exterior is great and a 4 star hotel. but a bit disappointed that the standard in the rooms were so old. more like a 3 star hotel.
asked for sea view didn't get it and we are here for 3 weeks very dissatisfied.
i would recommend to make the bathroom a bit bigger so the shower would be more comfortable and maybe even remove the bidet. .
need hairdryer .
great stay.
upgrade wifi.
too much salt in the pool .
a good map showing best way to apartment.
ban large inflatable pool toys.
better wifi .
wifi needs to be attended to. it is a major problem all the time.
microwaves walk-in showers refurbished as apartments are looking dated.
suggest one pool should be heated in santiago 3. would suggest not putting guests into rooms next door to rooms that are being renovated. wifi stopped working after 2 days in our room . otherwise a lovely resort .
very little vegetables available. bad quality food and drink.
the wifi must be better .
wifi must be better. bad in lanzarote.
old and bad beds. old and bad furniture in all rooms, especially in the kitchen.
why is there not any tea pot .. terrible . bad.
we asked for a room not in the tower when we came (6 or 7 times) we got a room in the tower. the reception had changed. but now we have a room without sunshine.
nine.
can we book the same room, in your website, direct from home.
i want to use my hairdryer in the bathroom, not the kitchen. .
our room has no sun in the morning and no sun in the evening. sad, very sad. something we will consider next time we order. .
big broom for cleanup (sand , balcony) is missing.
better wifi connection .
swimming pool is very cold .
very good and cleaning ladies very friendly.
very good and cleaning ladies very friendly.
very good and cleaning ladies very friendly.
more selection on the tv and double beds .
have wifi available throughout.
staff at reception didn't do anything to try and resolve our issue. won't ever stay here again and will warn anyone else of coming to stay here.
ceiling fans not working .

ceiling fans not working .
microwave would be very welcome.
i only missed watercooker and micro wave. .
wifi is very bad.
you need showers in the bathroom instead of baths modern kitchens need to be fitted your
price is now getting too expensive paid 90 euròs for 1 bed apartment in 2016 now pay 140
euros a night in november but you haven't upgraded the rooms or wifi.
you need showers in the bathroom instead of baths modern kitchens need to be fitted your
price is now getting too expensive paid 90 euròs for 1 bed apartment in 2016 now pay 140
euros a night in november but you haven't upgraded the rooms or wifi.
everything was wonderful all the staff we encountered were wonderful. the housekeeping s
taff even set the kids teddy bears neatly on the beds. our only issue the whole week was
that the pools are colder than the ocean. .
be thee light in the bathroom.
poolrest.
prices in supermarkets very expensive.
none.
warmer pools.
beds too hard.
el agua caliente insuficiente hasta para dos,el wifi horrible y ruidos cañerías,se aviso
pero auntie vinieron no se arregló.
think you should have mircowave and a electric kettle in the rooms .
better wifi .
great like every year.
per gaps a smile at the desk.
more wifi power.
we were given a room next to 3028 which was being refurbished to a private client . very
noisy all day.
more comfortable beds. room had an unpleasant smell.
this list should not come day 1.
tv noruega falta.
room at the street, negative surprise.
make sure disabled people are on ground floor if possible .
sort out your terrible wi fi connections.
pools shoul not be salt water,also very cold water.
upgrade the wifi in each apt.
please in german.
a ramp would be a good thing as our son and many other people are invalids.
you must get wifi which is working in the hotel rooms.
the tv should be bigger.
facility to buy news paper on the reception would be great.
the wifi was very bad. sometimes it was not possible to login. sometimes when you were co
nnected it was not efficiant enough to open simple websites. .
nu tv asked evert time connectie your tv to your smartdevice. when i click denyed it come
s back evert time. introef close aswell. i van not switch it of in the menu's .
loved my stay will defo come back lovely staff and lovely manners would recommend thank
you all.
we need 2 rooms on same level,no stairs.
wifi bad.
need better wifi. also it shouldn'task for you to login all the time. hope you can take
my recommendations into account.
more uk channels.
improve wifi rooms not ready or what we had booked.
wifi not working all the time.
there are nog enough wall outlets loud music every evening outside is terrible.
better wifi .
would like the pools to be un salty.
better internet .
had difficulties with chanel 4 and wifi keeps dropping. otherwise we have stayed at the
resort about 7 times and keep coming back as we love the resort and atmosphere. no need
for any changes.
don't charge your customers for the safe. gives a cheap impression for a nice hotel like
yours. upgrade the interest access for speed and security.
none .
a more prompt service to issues reported. wifi down for 6 days in room, only responded to
because electricity went off in the room on 6th day. disappointed by psantiagos attitude
to this. as access to wifi is of extreme importance to us as a family. other than that, w
e love it here, as this is our 6th time to visit. looking forward to seeing it all renova
ted and modernised in the near future hopefully. .
improve the wifi and also if you could leave washing up liquid so we could wash up in the
apartment .
could have tea and coffee in room also few toiletries.

slow wifi in room .
a hook for the towel in the kitchen.
add a micro wave.
the glasses for water, milk, drinks etc are to big.
fix the broken paint in the pool, moore steady wifi .
heat the big pool at this time of year. improve wifi.
norsk tv.
no.
badly needs refurbished .
to much noice from the restaurants.
no.
a big thank you to the cleaning staff and the reception staff. they were friendly en help
ed with everything. we had a very nice holiday here.
i come often .please check gregor.bruce1@btinternet.com.. this never confers favour when
getting a room south/west facing overlooking the pool.
better wifi connection .
less building work .
3020.
too much noise/music from the restaurants in th evening.
time to put in new beds and a new sofa in our room (1029).
had a fantastic time will definatly be back . fans ok but miss air conditioning would n
ot return in the hottest months . well maintained your maids are fantastic .
if traveling with small kids the upstairs rooms are a bit dangerous. then the rooms at gr
ound level are better. also we didn't know about the heated pool until yesterday. maybe a
sign or map of the water temperature at the big pool would be nice .
make the rooms as modern as the rest of the area. exterier i great and a 4 star hotel. bu
t abit disapointed tjat the standard in the rooms where so old. more like a 3 star hotel.
asked for sea view didn't get it and we are here for 3 weeks very dissatisfied.
i would recommand to make the bathroom a bit bigger so the shower would be more comfortab
le and maybe even remove the bidet. .
need hairdryer .
great stay.
upgrade wifi.
too much salt in the pool .
a good map showing best way to apartment.
ban large inflatable pool toys.
better wifi .
wifi needs to be attended to. it is a major problem all the time.
microwaves walkin showers refurbished as apartments are looking dated.
suggest one pool should be heated in santiago 3. would suggest not putting guests into r
ooms next door to rooms that are being renovated. wifi stopped working after 2 days in ou
r room . otherwise a lovely resort .
very little vegetables available. bad quality om mondstuk.
the wi fi must be better .
wifi mist be better. bad in lanzarote.
old and bad beds. old and bad furnitures in all roms, especially in the kitchen.
why is there not any tea pot .. terrible . bad.
we asked for a room mot in the toeter heen we came(6 or 7 time) we hot a room in the toet
er. the reception had change or. but moe we have a room without sunshine.
nine.
can we book the same room, in your homesite, direct from home.
i want to use my hairdryer in the bathroom, not the kitchen. .
our rom have no sun in the morning and no sin in the evenong. sad, very sad. something we
will consortes next time we order. .
big broom for cleanup (sand , balcony) is missing.
better wifi connection .
swiming pool os vert cold .
very gold and cleaning ladys very friendly.
very gold and cleaning ladys very friendly.
very gold and cleaning ladys very friendly.
more selection on the tv and double beds .
have wifi available throughout.
staff at reception didn't do anything to try and resolve our issue. wonth ever stay here
again and will warn anyone else of coming to stay here.
ceiling fans not working .
ceiling fans not working .
microwave would be very welcome.
i only missed watercooker and micro wave. .
wifi is very bad.
you need showers in the bathroom instead of baths modern kitchens need to be fitted your
price is nowgetting too expensive paid 90 euròs for 1 bed apartment in 2016 now pay 140
euros a night in november but you haven't upgraded the rooms or wifi.

you need showers in the bathroom instead of baths modern kitchens need to be fitted your price is now getting too expensive paid 90 euròs for 1 bed apartment in 2016 now pay 140 euros a night in november but you haven't upgraded the rooms or wifi. everything was wonderful all the staff we encountered were wonderful. the housekeeping staff even set the kids teddy bears neatly on the beds. our only issue the whole week was that the pools are colder than the ocean. . be thee light in the bathroom. poolrest. prices in supermarkets very expensive. none. warmer pools. beds too hard. el agua caliente insuficiente hasta para dos,el wifi horrible y ruidos cañerías,se aviso pero auntie vinieron no se arregló. think you should have mircowave and a electric kettle in the rooms . better wifi . great like every year. per gaps a smile at the desk. more wifi power. we were given a room next to 3028 which was being refurbished to a private client . very noisy all day. more comfortable beds. room had an unpleasant smell. this list should not come day 1. i come often .please check gregor.bruce1@btinternet.com.. this never confers favour when getting a room south/west facing overlooking the pool

In [10]:

```
blob = TextBlob(cadena)
# blob.tags
```

Each word in the lexicon has scores for: # 1) polarity: negative vs. positive (-1.0 => +1.0) # 2) subjectivity: objective vs. subjective (+0.0 => +1.0)

In [11]:

```
output = []
keyword = '';
result = 'Not relevant';

for sentence in blob.sentences:
    doc_aux = nlp(str(sentence))
    keyword = '';
    for ent in doc_aux.ents:
        keyword = keyword + ent.text + ent.label_;
    if (sentence.sentiment.polarity < -0.2):
        result = 'NEGATIVE'
    elif (sentence.sentiment.polarity > 0.2):
        result = 'POSITIVE'
    else:
        result = 'Not Relevant'

    output.append({"sentence":sentence,"Polarity":sentence.sentiment.polarity,
                  "subjectivity":sentence.sentiment.subjectivity, "keyword":keyword,
                  "Result": result},
    )
```

In [12]:

```
output_pd = pd.DataFrame(output)
output_pd.head(50)
```

Out[12]:

	sentence	Polarity	subjectivity	keyword	Result
0	(b, e, t, t, e, r, , w, i, f, i, , .)	0.500000	0.500000		POSITIVE
1	(w, o, u, l, d, , l, i, k, e, , t, h, e, , p, o, o, l, s, , t, o, , b, e, , u, n, , s, a, l, t, y, .)	0.000000	0.000000	unORG	Not Relevant
2	(b, e, t, t, e, r, , i, n, t, e, r, n, e, t, , .)	0.500000	0.500000		POSITIVE

3	(n, a, d, , d, i, t, i, l, c, u, i, t, i, e, s, , w, i, t, n, , c, n, a, n, e, l, , 4, , a, n, d, , w, i, f, i, , k, e, e, p, s, , d, i, o, p, p, i, n, g, .)	0.000000	0.000000	character	keyword	Result Not Relevant
4	(o, t, h, e, r, w, i, s, e, , w, e, , h, a, v, e, , s, t, a, y, e, d, , , a, t, , t, h, e, , r, e, s, o, r, t, , a, b, o, u, t, , 7, , t, i, m, e, s, , a, n, d, , k, e, e, p, , c, o, m, i, n, g, , b, a, c, k, , a, s, , w, e, , l, o, v, e, , t, h, e, , r, e, s, o, r, t, , a, n, d, ...)	0.250000	0.300000	about 7	CARDINAL	POSITIVE
5	(n, o, , n, e, e, d, , f, o, r, , a, n, y, , c, h, a, n, g, e, s, .)	0.000000	0.000000			Not Relevant
6	(d, o, n, ' , t, , c, h, a, r, g, e, , y, o, u, r, , c, u, s, t, o, m, e, r, s, , f, o, r, , t, h, e, , s, a, f, e, .)	0.500000	0.500000			POSITIVE
7	(g, i, v, e, s, , a, , c, h, e, a, p, , i, m, p, r, e, s, s, i, o, n, , f, o, r, , a, , n, i, c, e, , h, o, t, t, e, l, , l, i, k, e, , y, o, u, r, s, .)	0.500000	0.850000			POSITIVE
8	(u, p, g, r, a, d, e, , t, h, e, , i, n, t, e, r, e, s, t, , a, c, c, e, s, s, , f, o, r, , s, p, e, e, d, , a, n, d, , s, e, c, u, r, i, t, y, .)	0.000000	0.000000			Not Relevant
9	(n, o, n, e, , .)	0.000000	0.000000			Not Relevant
10	(a, , m, o, r, e, , p, r, o, m, p, t, , s, e, r, v, i, c, e, , t, o, , i, s, s, u, e, s, , r, e, p, o, r, t, e, d, .)	0.500000	0.500000			POSITIVE
11	(w, i, f, i, , d, o, w, n, , f, o, r, , 6, , d, a, y, s, , i, n, , r, o, o, m, , , o, n, l, y, , r, e, s, p, o, n, d, e, d, , t, o, , b, e, c, a, u, s, e, , e, l, e, c, t, r, i, c, i, t, y, , w, e, n, t, , o, f, f, , i, n, , t, h, e, , r, o, o, m, , o, n, , 6, t, h, , d, a, y, .)	- 0.077778	0.644444	6 days	DATE6th dayDATE	Not Relevant
12	(d, i, s, a, p, p, o, i, n, t, e, d, , b, y, , p, s, a, n, t, i, a, g, o, s, , a, t, t, i, t, u, d, e, , t, o, , t, h, i, s, .)	- 0.750000	0.750000			NEGATIVE
13	(a, s, , a, c, c, e, s, s, , t, o, , w, i, f, i, , i, s, , o, f, , e, x, t, r, e, m, e, , i, m, p, o, r, t, a, n, c, e, , t, o, , u, s, , a, s, , a, , f, a, m, i, l, y, .)	- 0.125000	1.000000			Not Relevant
14	(o, t, h, e, r, , t, h, a, n, , t, h, a, t, , , w, e, , l, o, v, e, , i, t, , h, e, r, e, , , a, s, , t, h, i, s, , i, s, , o, u, r, , 6, t, h, , t, i, m, e, , t, o, , v, i, s, i, t, .)	0.187500	0.487500	6th	ORDINAL	Not Relevant
15	(l, o, o, k, i, n, g, , f, o, r, w, a, r, d, , t, o, , s, e, e, i, n, g, , i, t, , a, l, l, , r, e, n, o, v, a, t, e, d, , a, n, d, , m, o, d, e, r, n, i, s, e, d, , i, n, , t, h, e, , n, e, a, r, , f, u, t, u, r, e, , h, o, p, e, f, u, l, l, y, .)	0.050000	0.262500			Not Relevant
16	(.)	0.000000	0.000000			Not Relevant
17	(i, m, p, r, o, v, e, , t, h, e, , w, i, f, i, , a, n, d, , a, l, s, o, , i, f, , y, o, u, , c, o, u, l, d, , l, e, a, v, e, , w, a, s, h, i, n, g, , u, p, , l, i, q, u, i, d, , s, o, , w, e, , c, o, u, l, d, , w, a, s, h, , u, p, , i, n, , t, h, e, , , a, p, a, r, t, m, e, n, t, ...)	0.000000	0.000000			Not Relevant
18	(c, o, u, l, d, , h, a, v, e, , t, e, a, , a, n, d, , c, o, f, f, e, e, , i, n, , r, o, o, m, , a, l, s, o, , f, e, w, , t, o, i, l, e, t, r, i, e, s, .)	- 0.200000	0.100000			Not Relevant
19	(s, l, o, w, , w, i, f, i, , i, n, , r, o, o, m, , .)	- 0.300000	0.400000			NEGATIVE
20	(a, , h, o, o, k, , f, o, r, , t, h, e, , t, o, w, e, l, , i, n, , t, h, e, , k, i, t, c, h, e, n, .)	0.000000	0.000000			Not Relevant
21	(a, d, d, , a, , m, i, c, r, o, , w, a, v, e, .)	0.000000	0.000000			Not Relevant
22	(t, h, e, , g, l, a, s, s, e, s, , f, o, r, , w, a, t, e, r, , , m, i, l, k, , , d, r, i, n, k, s, , e, t, c, , a, r, e, , t, o, , b, i, g, .)	0.000000	0.100000			Not Relevant
23	(f, i, x, , t, h, e, , b, r, o, o, k, e, n, , p, a, i, n, t, , i, n, , t, h, e, , p, o, o, l, , , m, o, o, r, e, , s, t, e, a, d, y, , w, i, f, i, , .)	- 0.116667	0.450000			Not Relevant
24	(h, e, a, t, , t, h, e, , b, i, g, , p, o, o, l, , a, t, , t, h, i, s, , t, i, m, e, , o, f, , y, e, a, r, .)	0.000000	0.100000	this time of year	DATE	Not Relevant
25	(i, m, p, r, o, v, e, , w, i, f, i, .)	0.000000	0.000000			Not Relevant
26	(n o r s k t v)	0.000000	0.000000	norsk tv	ORG	Not Relevant

	sentence	Polarity	subjectivity	keyword	Relevant Result
27	(n, o, .)	0.000000	0.000000		Not Relevant
28	(b, a, d, l, y, , n, e, e, d, s, , r, e, f, u, r, b, i, s, h, e, d, , .)	0.700000	0.666667		NEGATIVE
29	(t, o, , m, u, c, h, , n, o, i, c, e, , f, r, o, m, , t, h, e, , r, e, s, t, a, u, r, a, n, t, s, .)	0.200000	0.200000		Not Relevant
30	(n, o, .)	0.000000	0.000000		Not Relevant
31	(a, , b, i, g, , t, h, a, n, k, , y, o, u, , t, o, , t, h, e, , c, l, e, a, n, i, n, g, , s, t, a, f, f, , a, n, d, , t, h, e, , r, e, c, e, p, t, i, o, n, , s, t, a, f, f, .)	0.000000	0.100000		Not Relevant
32	(t, h, e, y, , w, e, r, e, , f, r, i, e, n, d, l, y, , e, n, , h, e, l, p, e, d, , w, i, t, h, , e, v, e, r, y, t, h, i, n, g, .)	0.375000	0.500000		POSITIVE
33	(w, e, , h, a, d, , a, , v, e, r, y, , n, i, c, e, , h, o, l, i, d, a, y, , h, e, r, e, .)	0.780000	1.000000		POSITIVE
34	(i, , c, o, m, e, , o, f, t, e, n, , ., p, l, e, a, s, e, , c, h, e, c, k, , g, r, e, g, o, r, ., b, r, u, c, e, 1, @, b, t, i, n, t, e, r, n, e, t, ., c, o, m, ., ., , t, h, i, s, , n, e, v, e, r, , c, o, n, f, e, r, s, , f, a, v, o, u, r, , w, h, e, n, , g, e, t, t, i, n, g, , a, , r, ...)	0.000000	0.000000	gregor.bruce1@btinternet.comORG	Not Relevant
35	(b, e, t, t, e, r, , w, i, f, i, , c, o, n, n, e, c, t, i, o, n, , .)	0.500000	0.500000		POSITIVE
36	(l, e, s, s, , b, u, i, l, d, i, n, g, , w, o, r, k, , .)	0.166667	0.066667		Not Relevant
37	(3, 0, 2, 0, ., \n, t, o, o, , m, u, c, h, , n, o, i, s, e, /, m, u, s, i, c, , f, r, o, m, , t, h, e, , r, e, s, t, a, u, r, a, n, t, s, , i, n, , t, h, , e, v, e, n, i, n, g, .)	0.200000	0.200000	3020CARDINALeveningTIME	Not Relevant
38	(t, i, m, e, , t, o, , p, u, t, , i, n, , n, e, w, , b, e, d, s, , a, n, d, , a, , n, e, w, , s, o, f, a, , i, n, , o, u, r, , r, o, o, m, , (, 1, 0, 2, 9,), .)	0.136364	0.454545	1029DATE	Not Relevant
39	(h, a, d, , a, , f, a, n, t, a, s, t, i, c, , t, i, m, e, , , w, i, l, l, , d, e, f, i, n, a, t, i, y, , b, e, , b, a, c, k, , .)	0.200000	0.450000		Not Relevant
40	(f, a, n, s, , o, k, , b, u, t, , m, i, s, s, , a, i, r, , c, o, n, d, i, t, i, o, n, i, n, g, , , w, o, u, l, d, , n, o, t, , r, e, t, u, r, n, , i, n, , t, h, e, , h, o, t, t, e, s, t, , m, o, n, t, h, s, , .)	0.500000	0.500000	the hottest monthsDATE	POSITIVE
41	(w, e, l, l, , m, a, i, n, t, a, i, n, e, d, , , y, o, u, r, , m, a, i, d, s, , a, r, e, , f, a, n, t, a, s, t, i, c, , .)	0.400000	0.900000		POSITIVE
42	(i, f, , t, r, a, v, e, l, i, n, g, , w, i, t, h, , s, m, a, l, l, , k, i, d, s, , t, h, e, , u, p, s, t, a, i, r, s, , r, o, o, m, s, , a, r, e, , a, , b, i, t, , d, a, n, g, e, r, o, u, s, .)	0.425000	0.650000		NEGATIVE
43	(t, h, e, n, , t, h, e, , r, o, o, m, s, , a, t, , g, r, o, u, n, d, , l, e, v, e, l, , a, r, e, , b, e, t, t, e, r, .)	0.500000	0.500000		POSITIVE
44	(a, l, s, o, , w, e, , d, i, d, n, ' , t, , k, n, o, w, , a, b, o, u, t, , t, h, e, , h, e, a, t, e, d, , p, o, o, l, , u, n, t, i, l, , y, e, s, t, e, r, d, a, y, .)	0.000000	0.000000	yesterdayDATE	Not Relevant
45	(m, a, y, b, e, , a, , s, i, g, n, , o, r, , m, a, p, , o, f, , t, h, e, , w, a, t, e, r, , t, e, m, p, e, r, a, t, u, r, e, , a, t, , t, h, e, , b, i, g, , p, o, o, l, , w, o, u, l, d, , b, e, , n, i, c, e, , .)	0.300000	0.550000		POSITIVE
46	(m, a, k, e, , t, h, e, , r, o, o, m, s, , a, s, , m, o, d, e, r, n, , a, s, , t, h, e, , r, e, s, t, , o, f, , t, h, e, , a, r, e, a, .)	0.200000	0.300000		Not Relevant
47	(e, x, t, e, r, i, e, r, , i, , g, r, e, a, t, , a, n, d, , a, , 4, , s, t, a, r, , h, o, t, e, l, .)	0.800000	0.750000	4CARDINAL	POSITIVE
48	(b, u, t, , a, b, i, t, , d, i, s, a, p, o, i, n, t, e, d, , t, j, a, t, , t, h, e, , s, t, a, n, d, a, r, d, , i, n, , t, h, e, , r, o, o, m, s, , w, h, e, r, e, , s, o, , o, l, d, .)	0.050000	0.100000	abitPERSON	Not Relevant
49	(m, o, r, e, , l, i, k, e, , a, , 3, , s, t, a, r, , h, o, t, e, l, .)	0.500000	0.500000	3CARDINAL	POSITIVE

In [13]:

```
#myData = [{"sentence", "polairy", "sujetivity"}]
output = []
```

```

keyword = '';
result = 'Not relevant';

myFile = open(present_working_directory+'\\sentimientos.csv', 'w')
keyword = '';
result = 'Not relevant';
index = 0;
hotelname = '';
with myFile:
    writer = csv.writer(myFile, delimiter=',', quotechar='"', quoting=csv.QUOTE_MINIMAL)
    writer.writerow(['hotel', 'datetime', 'sentence', 'polarity', 'subjectivity', 'keyword',
                    'result'])

    for sentence in blob.sentences:
        doc_aux = nlp(str(sentence))
        keyword = '';
        for ent in doc_aux.ents:
            # print the entity text and its label
            #if (ent.label_ == 'ORG' or ent.label_ == 'PERSON'):
            keyword = keyword + ent.text + " ";

        if (len (str(sentence)) > 3):
            if (sentence.sentiment.polarity < 0):
                result = 'NEGATIVE'
            else:
                result = 'POSITIVE'

        output.append({"sentence":sentence, "Polarity":sentence.sentiment.polarity,
                      "subjectivity":sentence.sentiment.subjectivity, "keyword":key
word,
                      "Result": result},
                      )
        if (index%2 == 0):
            hotelname = 'Parque Santiago'
        else:
            hotelname = 'Las Marismas de Corralejo'
        index = index + 1;
        writer.writerow([hotelname,datetime.datetime.now() - datetime.timedelta(days
=len(sentence)) ,
                        sentence,sentence.sentiment.polarity, sentence.sentiment.subjec
tivity,
                        keyword, result])

print("Writing complete")

```

Writing complete

In [14]:

```

output_pd = pd.DataFrame(output)
output_pd.sort_values(by=['Polarity'])

```

Out[14]:

	sentence	Polarity	subjectivity	keyword	Result
193	(w, h, y, , i, s, , t, h, e, r, e, , n, o, t, , a, n, y, , t, e, a, , p, o, t, , ., ., , t, e, r, r, i, b, l, e, ,)	-1.0	1.0		NEGATIVE
374	(e, l, , a, g, u, a, , c, a, l, i, e, n, t, e, , i, n, s, u, f, i, c, i, e, n, t, e, , h, a, s, t, a, , p, a, r, a, , d, o, s, , , e, l, , w, i, f, i, , h, o, r, r, i, b, l, e, , y, , r, u, i, d, o, s, , c, a, ñ, e, r, i, a, s, , , s, e, , a, v, i, s, o, , p, e, r, o, , a, u, n, t, i, e, ...)	-1.0	1.0	el el wifi	NEGATIVE
337	(w, h, y, , i, s, , t, h, e, r, e, , n, o, t, , a, n, y, , t, e, a, , p, o, t, , ., ., , t, e, r, r, i, b, l, e, ,)	-1.0	1.0		NEGATIVE
243	(s, o, r, t, , o, u, t, , y, o, u, r, , t, e, r, r, i, b, l, e, , w, i, , f, i, , c, o, n, n, e, c, t, i, o, n, s, ,)	-1.0	1.0		NEGATIVE
230	(e, l, , a, g, u, a, , c, a, l, i, e, n, t, e, , i, n, s, u, f, i, c, i, e, n, t, e, , h, a, s, t, a, , p, a, r, a, , d, o, s, , , e, l, , w, i, f, i, , h, o, r, r, i, b, l, e, , y, , r, u, i, d, o, s, , c, a, ñ, e, r, i, a, s, , , s, e, , a, v, i, s, o, , p, e, r, o, , a, u, n, t, i, e, ...)	-1.0	1.0	el el wifi	NEGATIVE
...	...	...	...	...	...

	Sentence	Polarity	subjectivity	keyword	Result
221	(e, v, e, r, y, t, h, i, n, g, , w, a, s, , w, o, n, d, e, r, f, u, l, , , a, l, l, , t, h, e, s, e, , f, f, , w, e, , e, n, c, o, u, n, t, e, r, e, d, , w, e, r, e, , w, o, n, d, e, r, f, u, l, .)	1.0	1.0		POSITIVE
360	(m, i, c, r, o, w, a, v, e, , w, o, u, l, d, , b, e, , v, e, r, y, , w, e, l, c, o, m, e, .)	1.0	1.0	microwave	POSITIVE
97	(e, v, e, r, y, t, h, i, n, g, , w, a, s, , w, o, n, d, e, r, f, u, l, , , a, l, l, , t, h, e, , s, t, a, f, f, , w, e, , e, n, c, o, u, n, t, e, r, e, d, , w, e, r, e, , w, o, n, d, e, r, f, u, l, .)	1.0	1.0		POSITIVE
92	(m, i, c, r, o, w, a, v, e, , w, o, u, l, d, , b, e, , v, e, r, y, , w, e, l, c, o, m, e, .)	1.0	1.0	microwave	POSITIVE
365	(e, v, e, r, y, t, h, i, n, g, , w, a, s, , w, o, n, d, e, r, f, u, l, , , a, l, l, , t, h, e, , s, t, a, f, f, , w, e, , e, n, c, o, u, n, t, e, r, e, d, , w, e, r, e, , w, o, n, d, e, r, f, u, l, .)	1.0	1.0		POSITIVE

385 rows × 5 columns

In []:

6.1 DESARROLLO DE MODELO DE CLASIFICACIÓN PARA DETERMINAR CANCELACIONES

Cargamos los datos de la base de datos

```
In [101]: import mysql.connector as mysql
```

```
In [102]: ## connecting to the database using 'connect()' method  
## it takes 3 required parameters 'host', 'user', 'passwd'  
db = mysql.connect(  
    host = "localhost",  
    user = "root",  
    passwd = "pentaho",  
    database = "stg_tfm"  
)
```

```
In [103]: cursor = db.cursor()
```

```
In [104]: query = "select * from reservas"
```

```
In [105]: cursor.execute(query)
```

```
In [106]: records = cursor.fetchall()  
  
## Showing the data  
#for record in records:  
#    print(record)
```

```
In [107]: import pandas as pd  
#indico las dimensiones de la ventana de salida  
pd.set_option('display.max_rows', 500)  
pd.set_option('display.max_columns', 500)  
pd.set_option('display.width', 1000)
```

```
In [108]: import numpy as np
```

```
In [109]: df = pd.read_sql(("SELECT * FROM reservas"), db) #Lanzar la query para seleccionar todo de los datos reservas
```

```
In [11]: #print(df)
```

```
In [22]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 93324 entries, 0 to 93323  
Columns: 144 entries, Numero de reserva to Cancelación  
dtypes: float64(13), int64(1), object(130)  
memory usage: 102.5+ MB
```



```
In [110]: df.describe(include = 'all');#realizamos un describe para analizar tanto las v  
         ariables categóricas como las variables numéricas.  
         #la coletilla all implica que se incluyan todas las variables
```

```
In [29]: import numpy as np
          #crear un dataframe con los nulos y mostrarlos ordenados
          df_null=pd.DataFrame({'Campos':list(df.columns),'nº de nulos':df.isnull().sum
          ()})#
          df_null=df_null.sort_values('nº de nulos')
          df_null
```

Out[29]:

	Campos	nº de nulos
Numero de reserva	Numero de reserva	0
Nombre del Hotel	Nombre del Hotel	0
Importe prepago/depósito	Importe prepago/depósito	0
Tipo de habitación	Tipo de habitación	0
Fecha de la reserva	Fecha de la reserva	0
Cancelación	Cancelación	0
Fecha de llegada	Fecha de llegada	0
Fecha de salida	Fecha de salida	0
Nombre del cliente	Nombre del cliente	7
Numero de personas	Numero de personas	173
Tipo de pension	Tipo de pension	195
Agencia	Agencia	794
Total estancia	Total estancia	1614
Días	Días	4326
Importe de la estancia sin oferta	Importe de la estancia sin oferta	4490
Número última Reserva	Número última Reserva	17507
Fecha de modificación de la reserva	Fecha de modificación de la reserva	21960
Tipo de habitación asignado	Tipo de habitación asignado	21960
Nacionalidad del cliente	Nacionalidad del cliente	21967
Nombre agencia	Nombre agencia	23020
Cupo	Cupo	23202
Precio Medio	Precio Medio	23612
Numero de habitación	Numero de habitación	23755
Cardex (1)	Cardex (1)	33214
Número de niños	Número de niños	39408
Habitación compartida	Habitación compartida	41013
Hora de salida	Hora de salida	41013
Hora de llegada	Hora de llegada	41013
Reserva online	Reserva online	41013
Estado	Estado	41013
Ofertas aplicadas	Ofertas aplicadas	41013
Procedencia (1)	Procedencia (1)	41014
Hora modificación	Hora modificación	41271
Sexo del cliente	Sexo del cliente	42065
Tipo de documento (1)	Tipo de documento (1)	42133

		Campos	nº de nulos
Edad del cliente	Edad del cliente		42202
Fecha de nacimiento (1)	Fecha de nacimiento (1)		42202
DNI/Pasaporte	DNI/Pasaporte		42238
Fecha de documento (1)	Fecha de documento (1)		42298
Importe total de las ofertas aplicadas	Importe total de las ofertas aplicadas		42763
Nombre ocupante (1)	Nombre ocupante (1)		46510
Procedencia (2)	Procedencia (2)		46529
Nación ocupante (1)	Nación ocupante (1)		46530
Sexo ocupante (1)	Sexo ocupante (1)		46531
Fecha de nacimiento (2)	Fecha de nacimiento (2)		46552
Edad ocupante (1)	Edad ocupante (1)		46555
Tipo de documento (2)	Tipo de documento (2)		46560
D.N.I./Pasaporte Ocupante (1)	D.N.I./Pasaporte Ocupante (1)		46564
Fecha de documento (2)	Fecha de documento (2)		46602
Número de bebes	Número de bebes		49550
Número de Bono	Número de Bono		50744
Número de niños (2)	Número de niños (2)		51458
Camas extra	Camas extra		52299
Límite de crédito	Límite de crédito		52311
Canal de venta	Canal de venta		52969
Última Salida (1)	Última Salida (1)		54237
Cardex (2)	Cardex (2)		56162
Última Salida (2)	Última Salida (2)		58464
Observaciones llegada	Observaciones llegada		61054
Hora reserva	Hora reserva		63657
Nombre canal de venta	Nombre canal de venta		63921
Forma de pago	Forma de pago		64814
Observaciones llegada (2)	Observaciones llegada (2)		69795
Operador de la anulación	Operador de la anulación		71215
Motivo de la anulación	Motivo de la anulación		71364
Fecha de anulación	Fecha de anulación		71364
Sexo ocupante (2)	Sexo ocupante (2)		73596
Nombre ocupante (2)	Nombre ocupante (2)		73603
Fecha de nacimiento (3)	Fecha de nacimiento (3)		73607
Edad ocupante (2)	Edad ocupante (2)		73614
Nación ocupante (2)	Nación ocupante (2)		73624

Campos		nº de nulos
Procedencia (3)	Procedencia (3)	73624
Tipo de documento (3)	Tipo de documento (3)	73637
D.N.I./Pasaporte Ocupante (2)	D.N.I./Pasaporte Ocupante (2)	73644
Fecha de documento (3)	Fecha de documento (3)	73662
Email del cliente	Email del cliente	74883
Observaciones llegada (3)	Observaciones llegada (3)	75817
Cardex (3)	Cardex (3)	77940
Precio fijo	Precio fijo	79223
Última Salida (3)	Última Salida (3)	79277
Sexo ocupante (3)	Sexo ocupante (3)	83294
Fecha de nacimiento (4)	Fecha de nacimiento (4)	83296
Nombre ocupante (3)	Nombre ocupante (3)	83299
Edad ocupante (3)	Edad ocupante (3)	83300
Procedencia (4)	Procedencia (4)	83307
Nación ocupante (3)	Nación ocupante (3)	83307
Tipo de documento (4)	Tipo de documento (4)	83312
D.N.I./Pasaporte Ocupante (3)	D.N.I./Pasaporte Ocupante (3)	83313
Fecha de documento (4)	Fecha de documento (4)	83319
Observaciones Salida	Observaciones Salida	84017
Localizador/Referencia	Localizador/Referencia	84765
Cardex (4)	Cardex (4)	85491
Última Salida (4)	Última Salida (4)	86310
Observaciones llegada (4)	Observaciones llegada (4)	86872
Caducidad tarjeta crédito	Caducidad tarjeta crédito	87056
Password consulta del cliente	Password consulta del cliente	87057
Titular de la tarjeta	Titular de la tarjeta	87101
Tarjeta de crédito	Tarjeta de crédito	87463
Agencia Planning	Agencia Planning	87837
Bloque	Bloque	88624
Agencia Precios	Agencia Precios	88969
Nombre ocupante (4)	Nombre ocupante (4)	90696
Sexo ocupante (4)	Sexo ocupante (4)	90699
Fecha de nacimiento (5)	Fecha de nacimiento (5)	90703
Edad ocupante (4)	Edad ocupante (4)	90704
Nación ocupante (4)	Nación ocupante (4)	90705
D.N.I./Pasaporte Ocupante (4)	D.N.I./Pasaporte Ocupante (4)	90705

	Campos	nº de nulos
Tipo de documento (5)	Tipo de documento (5)	90705
Fecha de documento (5)	Fecha de documento (5)	90706
Código tarjeta crédito	Código tarjeta crédito	91097
Cardex (5)	Cardex (5)	91324
Última Salida (5)	Última Salida (5)	91709
Observaciones factura	Observaciones factura	92442
Edad ocupante (5)	Edad ocupante (5)	92535
Nación ocupante (5)	Nación ocupante (5)	92537
D.N.I./Pasaporte Ocupante (5)	D.N.I./Pasaporte Ocupante (5)	92537
Sexo ocupante (5)	Sexo ocupante (5)	92538
Nombre ocupante (5)	Nombre ocupante (5)	92538
Cardex (6)	Cardex (6)	92732
Última Salida (6)	Última Salida (6)	92845
Ocupación discapacitados	Ocupación discapacitados	92879
Turno de comedor	Turno de comedor	92894
Fecha de check-in	Fecha de check-in	92955
Hora de anulación	Hora de anulación	93190
Tipo de depósito en la reserva	Tipo de depósito en la reserva	93264
Departamento de cobro depósitos	Departamento de cobro depósitos	93266
Permite cargos extras	Permite cargos extras	93295
Permite crédito TV	Permite crédito TV	93295
Permite crédito TLF	Permite crédito TLF	93295
Num Par (1)	Num Par (1)	93316
Teléfono del cliente	Teléfono del cliente	93317
Num Par (2)	Num Par (2)	93318
Num Par (3)	Num Par (3)	93319
Fecha prepago/depósito	Fecha prepago/depósito	93322
Num Par (4)	Num Par (4)	93323
Prepago/Deposito	Prepago/Deposito	93324
Dirección del cliente	Dirección del cliente	93324
Nombre subagencia	Nombre subagencia	93324
Subagencia	Subagencia	93324
Factura de anticipo	Factura de anticipo	93324
Código postal del cliente	Código postal del cliente	93324
Localidad del cliente	Localidad del cliente	93324
Fax cel cliente	Fax cel cliente	93324

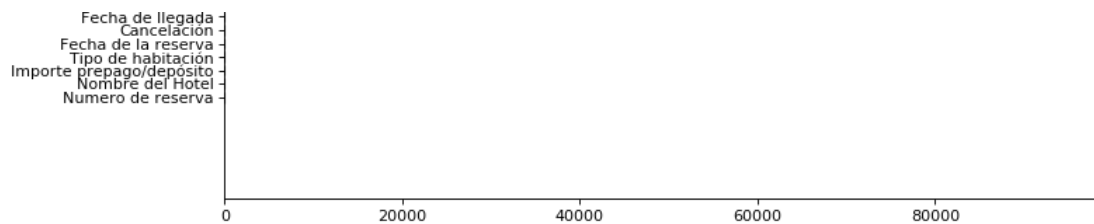
Campos		nº de nulos
País del cliente		93324

```
In [32]: import matplotlib.pyplot as plt
```

```
In [33]: plt.figure(figsize=(10,25), dpi=80)
plt.barh(df_null['Campos'], df_null['nº de nulos'], edgecolor='black')
```


Out[33]: <BarContainer object of 144 artists>

País del cliente	
Fax cel cliente	
Localidad del cliente	
Código postal del cliente	
Factura de anticipo	
Subagencia	
Nombre subagencia	
Dirección del cliente	
Prepago/Deposito	
Num Par (4)	
Fecha prepago/deposito	
Num Par (3)	
Num Par (2)	
Teléfono del cliente	
Num Par (1)	
Permite crédito TLF	
Permite crédito TV	
Permite cargos extras	
Departamento de cobro depósitos	
Tipo de depósito en la reserva	
Hora de anulación	
Fecha de check-in	
Turno de comedor	
Ocupación discapacitados	
Ultima Salida (6)	
Cardex (6)	
Nombre ocupante (5)	
Sexo ocupante (5)	
D.N.I./Pasaporte Ocupante (5)	
Nación ocupante (5)	
Edad ocupante (5)	
Observaciones factura	
Ultima Salida (5)	
Cardex (5)	
Código tarjeta crédito	
Fecha de documento (5)	
Tipo de documento (5)	
D.N.I./Pasaporte Ocupante (4)	
Nación ocupante (4)	
Edad ocupante (4)	
Fecha de nacimiento (5)	
Sexo ocupante (4)	
Nombre ocupante (4)	
Agencia Precios	
Bloque	
Agencia Planning	
Tarjeta de crédito	
Titular de la tarjeta	
Password consulta del cliente	
Caducidad tarjeta crédito	
Observaciones llegada (4)	
Ultima Salida (4)	
Cardex (4)	
Localizador/Referencia	
Observaciones Salida	
Fecha de documento (4)	
D.N.I./Pasaporte Ocupante (3)	
Tipo de documento (4)	
Nación ocupante (3)	
Procedencia (4)	
Edad ocupante (3)	
Nombre ocupante (3)	
Fecha de nacimiento (4)	
Sexo ocupante (3)	
Ultima Salida (3)	
Precio fijo	
Cardex (3)	
Observaciones llegada (3)	
Email del cliente	
Fecha de documento (3)	
D.N.I./Pasaporte Ocupante (2)	
Tipo de documento (3)	
Procedencia (3)	
Nación ocupante (2)	
Edad ocupante (2)	
Fecha de nacimiento (3)	
Nombre ocupante (2)	
Sexo ocupante (2)	
Fecha de anulación	
Motivo de la anulación	
Operador de la anulación	
Observaciones llegada (2)	
Forma de pago	
Nombre canal de venta	
Hora reserva	
Observaciones llegada	
Ultima Salida (2)	
Cardex (2)	
Ultima Salida (1)	
Canal de venta	
Límite de crédito	
Camas extra	
Número de niños (2)	
Número de Bono	
Número de bebés	
Fecha de documento (2)	
D.N.I./Pasaporte Ocupante (1)	
Tipo de documento (2)	
Edad ocupante (2)	
Fecha de nacimiento (2)	
Sexo ocupante (1)	
Nación ocupante (1)	
Procedencia (2)	
Nombre ocupante (1)	
Importe total de las ofertas aplicadas	
Fecha de documento (1)	
DNI/Pasaporte	
Fecha de nacimiento (1)	
Edad del cliente	
Tipo de documento (1)	
Sexo del cliente	
Hora modificación	
Procedencia (1)	
Ofertas aplicadas	
Estado	
Reserva online	
Hora de llegada	
Hora de salida	
Habitación compartida	
Número de niños	
Cardex (1)	
Número de habitación	
Precio Medio	
Cupo	
Nombre agencia	
Nacionalidad del cliente	
Tipo de habitación asignado	
Fecha de modificación de la reserva	
Número ultima Reserva	
Importe de la estancia sin oferta	
Días	
Total estancia	
Agencia	
Tipo de pension	
Número de personas	
Nombre del cliente	
Fecha de salida	



```
In [34]: df['Nombre del Hotel'].unique()
```

```
Out[34]: array(['PARQUE SANTIAGO 3', 'PARQUE SANTIAGO 4', 'Marimas'], dtype=object)
```

```
In [35]: #Extraer los campos de los que no tengo nulos
```

```
In [111]: #df_ps.describe(include='all') siguen saliendo 3 nombres de hotel
not_Marismas=df['Nombre del Hotel'] != 'Marimas'
df_1=df[not_Marismas]
df_1.describe(include='all')
```

```
Out[111]:
```

	Numero de reserva	Fecha de llegada	Fecha de salida	Fecha de check- in	Nombre del cliente	Nacionalidad del cliente	Edad del cliente	Sexo del cliente	DNI/Pasajero
count	74271	74271	74271	369	74271	52309	51122	51259	5
unique	74271	730	767	179	57478	64	95	2	4
top	3135626,0	2018-10-21	2018-11-04	2018-08-04	REFORMAS	GB	56	M	43387
freq	1	292	265	41	390	24152	1127	26535	
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

```
In [112]: df_1['Agencia'].unique(); #confirmamos que los datos de precio de Marismas no  
están mal incluidos
```

```
In [39]: #crear un data frame con columnas específicas a partir de df
df_1=df_1[['Numero de personas','Fecha de llegada','Fecha de salida','Fecha de  
la reserva','Tipo de habitación','Tipo de pensión','Agencia', 'Importe prepago  
o depósito', 'Nombre del Hotel','Cancelación']]
```

```
In [41]: #cambiar el tipo de dato de los campos de fechas para el tipo de dato datetime
df_1['Fecha de la reserva'] = pd.to_datetime(df['Fecha de la reserva'])
df_1['Fecha de llegada'] = pd.to_datetime(df['Fecha de la reserva'])
df_1['Fecha de salida'] = pd.to_datetime(df['Fecha de la reserva'])
df_1.dtypes
```

```
Out[41]: Numero de personas          float64
Fecha de llegada          datetime64[ns]
Fecha de salida           datetime64[ns]
Fecha de la reserva       datetime64[ns]
Tipo de habitación        object
Tipo de pension           object
Agencia                   object
Importe prepago/depósito  float64
Nombre del Hotel          object
Cancelación               int64
dtype: object
```

```
In [42]: #cambiar los nombres de las columnas a minúsculas y remplazar los espacios por _
df_1.columns = df_1.columns.str.lower().str.replace(' ', '_')
```

```
In [44]: df_1['mes_de_llegada'] = df_1['fecha_de_llegada'].dt.month
df_1['dia_de_llegada'] = df_1['fecha_de_llegada'].dt.day
df_1['mes_de_salida'] = df_1['fecha_de_salida'].dt.month
df_1['dia_de_salida'] = df_1['fecha_de_salida'].dt.day
df_1['mes_de_la_reserva'] = df_1['fecha_de_la_reserva'].dt.month
df_1['dia_de_la_reserva'] = df_1['fecha_de_la_reserva'].dt.day
```

```
In [45]: df_1.drop(['fecha_de_llegada', 'fecha_de_salida', 'fecha_de_la_reserva'], axis
= 'columns', inplace=True)
```

```
In [46]: # Rellena los NA con la moda de esta variable
sum(df_1.numero_de_personas.mode())
df_1.fillna({'numero_de_personas':df_1.numero_de_personas.mode().sum()},inplace=True);
df_1.describe(include='all')
```

Out[46]:

	numero_de_personas	tipo_de_habitación	tipo_de_pension	agencia	importe_prepago/dep
count	74271.000000	74271	74271	74271	74271
unique	NaN	17	2	81	81
top	NaN	STD	SS	JET2	JET2
freq	NaN	36407	73368	13856	13856
mean	2.668861	NaN	NaN	NaN	NaN
std	1.080258	NaN	NaN	NaN	NaN
min	1.000000	NaN	NaN	NaN	NaN
25%	2.000000	NaN	NaN	NaN	NaN
50%	2.000000	NaN	NaN	NaN	NaN
75%	3.000000	NaN	NaN	NaN	NaN
max	9.000000	NaN	NaN	NaN	NaN

```
In [47]: df_1.numero_de_personas=df_1.numero_de_personas.astype('int')
df_1.drop(['importe_prepago/depósito'], axis = 'columns', inplace=True)
```

```
In [48]: df_1.cancelación=df_1.cancelación.astype('int')
```

```
In [49]: #df_1.describe(include='all')
```

```
In [50]: #Le quitamos las tildes a ambas columnas
df_1['cancelacion']=df_1['cancelación']
df_1['tipo_de_habitacion']=df_1['tipo_de_habitación']
df_1.drop(['cancelación'], axis = 'columns', inplace=True)
df_1.drop(['tipo_de_habitación'], axis = 'columns', inplace=True)
```

```
In [51]: df_1['nombre_del_hotel'].unique()
```

Out[51]: array(['PARQUE SANTIAGO 3', 'PARQUE SANTIAGO 4'], dtype=object)

```
In [52]: #Seleccionamos todas las agencias menos estas 4
df_1=df_1[~df_1.agencia.isin(['PROPI', 'PROPI4', 'OBRAS', 'REFOR'])]
```

```
In [53]: #Generamos otro dataframe con los datos de las marismas
#df_2=df_1[df_1.nombre_del_hotel.isin(['Marimas'])]
```

```
In [54]: #convertimos en variables categóricas aquellas que son de tipo objetos
cat_features = df_1.columns[df_1.dtypes=='object']
num_features = df_1.columns[df_1.dtypes!='object']

for col in cat_features:
    df_1[col] = pd.Categorical(df_1[col])
```

```
In [55]: list_=df_1['tipo_de_habitacion'].unique()
for a in list_:
    print(a)
```

```
STD
1BA
2BA
VLA
VLB
STDI
2BAI
1BAI
1BAE
2BAE
STDE
3BA
1BAF
STDS
1BAS
2BAS
STDB
```

```
In [56]: list_2=df_1['agencia'].unique()
         for a in list_2:
             print(a)
```

WEB
UNI
ITS
ARTURO
ALTURA
TUI
DERFRA
HEIM
ARK
SCHAU
TUIPO
VITA
COLAB
SUMA
MCCA
BARCE
SSUR
ITSAU
ITSSU
DIR
KUO
GURIA
DTOS
HCH
TUISU
LUNA
HALCON
EROSKI
CLICK
OTB
TLT
SAGA
TUIAU
VING
TJAE
NVN
JET2
NHL
FIR
NUR
HOL
TENMA
VIACO
BUC
COVA
CORTE
NOMA
ZADY
ECU
ACONFI
VSIERR
BIVES
ARROBA
UNITV
ANUR
ERIOJA
AIRBUS

NOVA
 ALLTOU
 AMAIA
 TCNE
 SATT
 GLOBA
 BOOK
 BADA
 CANAT
 BOOK1
 ALISIO
 HERNA
 SOME
 COSMOS
 PTWEB
 PTAGE
 SOCEAN
 PAN
 NIEVE
 BEST

```
In [57]: list_3=df_1['tipo_de_pension'].unique()
for a in list_3:
    print(a)
```

SS
 MP

```
In [59]: df_1.head()
```

Out[59]:

	numero_de_personas	tipo_de_pension	agencia	nombre_del_hotel	mes_de_llegada	dia_de_l
8	2	SS	WEB	PARQUE SANTIAGO 3		2
9	2	SS	WEB	PARQUE SANTIAGO 3		2
10	2	SS	WEB	PARQUE SANTIAGO 3		2
11	2	SS	WEB	PARQUE SANTIAGO 3		2
12	4	SS	WEB	PARQUE SANTIAGO 3		2

```
In [60]: #aplicar get_dummies a aquellos campos que son categóricos para convertirlos en 1 y 0
#Lo que hace el get_dummies es categorizar las variables en 0 y 1 de cara a usar el árbol de decisión, que sólo usa
#variables numéricas
df_1 = pd.get_dummies(df_1, columns=['tipo_de_habitacion', 'agencia'], prefix_sep='__')
```

```
In [61]: #como sólo hay dos opciones donde elegir en cuanto a nombre del hotel,
#usamos esto para asignar '1' al valor Paque santiago 3 y al tipo de pension S

df_1.nombre_del_hotel = df_1.nombre_del_hotel.eq('PARQUE SANTIAGO 3').astype(
'int')
df_1.tipo_de_pension = df_1.tipo_de_pension.eq('SS').astype('int')
df_1.head()
```

Out[61]:

	numero_de_personas	tipo_de_pension	nombre_del_hotel	mes_de_llegada	dia_de_llegada	n
8	2	1	1	2	22	
9	2	1	1	2	22	
10	2	1	1	2	22	
11	2	1	1	2	22	
12	4	1	1	2	22	

```
In [62]: #df_1.dtypes
```

```
In [63]: df_1 = df_1.sample(frac=1).reset_index(drop=True)
X=df_1.drop('cancelacion',axis=1)
y=df_1.cancelacion
```

```
In [64]: #df_1.head(100)
list(X);
```

```
In [65]: df_1.cancelacion.value_counts().div(df_1.shape[0] / 100)
```

```
Out[65]: 0    70.471533
1     29.528467
Name: cancelacion, dtype: float64
```

```
In [66]: #df_1.describe(include='all')
```

```
In [67]: df_1.corr();
```

```
In [99]: #observamos la correlación lineal
corr = df_1.corr().stack();
corr.loc[corr < 1.0].sort_values(ascending=False);
```

```
In [98]: #observamos las correlaciones con la variable cancelacion  
corr.loc['cancelacion'].sort_values(ascending=False)
```

```

Out[98]: cancelacion                1.000000
          agencia__COSMOS            0.156901
          agencia__DIR               0.148171
          agencia__SOME              0.143127
          nombre_del_hotel           0.103409
          agencia__OTB               0.081232
          agencia__TLT               0.077396
          agencia__ARTURO            0.075534
          tipo_de_habitacion__STD    0.063236
          agencia__ALTURA           0.061946
          tipo_de_habitacion__2BA    0.057963
          numero_de_personas         0.049263
          agencia__PTWEB             0.048408
          agencia__VITA              0.041710
          agencia__HEIM              0.038240
          agencia__SUMA              0.033952
          agencia__PTAGE             0.032994
          agencia__SCHAU             0.022604
          agencia__BOOK              0.022461
          agencia__BOOK1             0.016639
          tipo_de_pension            0.016001
          agencia__NOVA              0.015804
          agencia__COLAB             0.011343
          agencia__5OCEAN            0.009946
          agencia__NOMA              0.009843
          agencia__TUIAU             0.009405
          agencia__WEB               0.008704
          agencia__PAN               0.008121
          agencia__DERFRA            0.007119
          agencia__BEST              0.005742
          agencia__NIEVE             0.005742
          agencia__NVN               0.004454
          tipo_de_habitacion__1BAF   0.003816
          agencia__TUISU             0.003120
          tipo_de_habitacion__3BA    0.002625
          agencia__BARCE             0.002613
          agencia__ALISIO            0.002359
          agencia__CORTE             0.002359
          agencia__TUIPO             0.002316
          agencia__GURIA             0.001074
          tipo_de_habitacion__STDB   0.000708
          agencia__LUNA              0.000537
          agencia__ITSSU             -0.001381
          agencia__ITSAU             -0.001901
          agencia__VIACO             -0.002016
          agencia__UNITV             -0.002406
          agencia__ARROBA            -0.002406
          agencia__ZADY              -0.002406
          agencia__AMAIA             -0.002406
          agencia__CANAT             -0.002406
          agencia__BIVES             -0.002406
          agencia__GLOBA             -0.002406
          agencia__ACONFI            -0.002406
          agencia__ALLTOU            -0.002406
          agencia__COVA              -0.002406
          agencia__BADA              -0.002406
          agencia__ECU               -0.002406

```

agencia__TUI	-0.002824
tipo_de_habitacion__1BA	-0.002881
agencia__VSIERR	-0.003403
agencia__SATT	-0.003403
agencia__HERNA	-0.003403
agencia__ANUR	-0.003403
agencia__CLICK	-0.003403
agencia__ERIOJA	-0.003403
agencia__AIRBUS	-0.004167
agencia__HALCON	-0.004648
dia_de_salida	-0.004698
dia_de_la_reserva	-0.004698
dia_de_llegada	-0.004698
agencia__HCH	-0.004750
agencia__EROSKI	-0.004812
agencia__TENMA	-0.005894
agencia__KUO	-0.009363
agencia__ITS	-0.011149
agencia__NUR	-0.012868
agencia__SSUR	-0.013349
tipo_de_habitacion__2BAI	-0.013395
agencia__HOL	-0.013540
agencia__BUC	-0.014307
tipo_de_habitacion__VLA	-0.017877
tipo_de_habitacion__2BAS	-0.020403
mes_de_llegada	-0.021016
mes_de_salida	-0.021016
mes_de_la_reserva	-0.021016
agencia__ARK	-0.021825
tipo_de_habitacion__VLB	-0.023670
tipo_de_habitacion__1BAS	-0.026663
tipo_de_habitacion__2BAE	-0.027208
agencia__NHL	-0.028250
agencia__MCCA	-0.035693
agencia__TJAE	-0.039683
tipo_de_habitacion__STDS	-0.045739
agencia__SAGA	-0.048063
agencia__JET2	-0.048457
tipo_de_habitacion__1BAE	-0.051704
tipo_de_habitacion__STDE	-0.057450
tipo_de_habitacion__1BAI	-0.057769
agencia__FIR	-0.062172
tipo_de_habitacion__STDI	-0.065041
agencia__DTOS	-0.078682
agencia__UNI	-0.093077
agencia__VING	-0.098412
agencia__TCNE	-0.099669
dtype: float64	

Aplicamos árbol de decisión

```
In [70]: from sklearn import tree

from sklearn.tree import DecisionTreeClassifier

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, stratify=y, random_state=0)
```

```
In [71]: X_train.shape, y_train.shape, X_test.shape, y_test.shape
```

```
Out[71]: ((50666, 103), (50666,), (21715, 103), (21715,))
```

```
In [72]: dtree = DecisionTreeClassifier()
dtree = dtree.fit(X_train, y_train)
```

```
In [73]: print(dtree.predict(X_test))
```

```
[0 0 0 ... 0 1 0]
```

```
In [74]: from sklearn import metrics
print("Accuracy:", metrics.accuracy_score(y_test, dtree.predict(X_test)))
```

```
Accuracy: 0.7081740732212756
```

```
In [75]: accuracy_basic_dtree=metrics.accuracy_score(y_test, dtree.predict(X_test))
accuracy_basic_dtree
```

```
Out[75]: 0.7081740732212756
```

```
In [77]: from sklearn.metrics import classification_report, confusion_matrix
```

```
print(" == Clasifcation report Decision Tree - Base ==")
print(classification_report(y_test,dtree.predict(X_test)))
print(" == Confusion Matrix - Base ==")
print(confusion_matrix(y_test,dtree.predict(X_test)))
```

```
== Clasification report Decision Tree - Base ==
              precision    recall  f1-score   support

     0       0.77       0.84       0.80      15303
     1       0.51       0.40       0.45       6412

 micro avg       0.71       0.71       0.71      21715
 macro avg       0.64       0.62       0.62      21715
 weighted avg    0.69       0.71       0.70      21715

== Confusion Matrix - Base ==
[[12834  2469]
 [ 3868  2544]]
```

Estudiamos las variables más importantes en el modelo de árbol de decisión

```
In [78]: X.describe(include='all');
```

```
In [79]: feature_imp_basic_dtree = pd.Series(dtree.feature_importances_, index=list(X)).  
         sort_values(ascending=False)  
         feature_imp_basic_dtree
```


Out[79]:

dia_de_la_reserva	0.108264
dia_de_llegada	0.107116
dia_de_salida	0.103722
numero_de_personas	0.091943
mes_de_la_reserva	0.068831
mes_de_llegada	0.064042
mes_de_salida	0.060296
nombre_del_hotel	0.050700
agencia__COSMOS	0.040036
agencia__DIR	0.034798
agencia__SOME	0.033833
tipo_de_habitacion__1BA	0.019229
agencia__OTB	0.015880
tipo_de_habitacion__STD	0.015421
agencia__TLT	0.014155
tipo_de_habitacion__2BA	0.012584
agencia__ARTURO	0.012345
agencia__WEB	0.011215
agencia__ALTURA	0.009894
agencia__TCNE	0.008568
agencia__UNI	0.008074
agencia__VING	0.007659
agencia__ARK	0.006941
agencia__DTOS	0.005895
agencia__TUI	0.005741
agencia__NVN	0.005565
agencia__HOL	0.005430
agencia__FIR	0.004022
agencia__ITS	0.003809
agencia__SAGA	0.003697
agencia__PTWEB	0.003511
agencia__VITA	0.003499
tipo_de_pension	0.003312
agencia__NUR	0.002817
tipo_de_habitacion__3BA	0.002513
tipo_de_habitacion__STDE	0.002409
agencia__TJAE	0.002276
agencia__JET2	0.002221
tipo_de_habitacion__VLA	0.002171
agencia__MCCA	0.002127
agencia__HCH	0.001989
agencia__TUIPO	0.001900
tipo_de_habitacion__STDI	0.001884
agencia__TUISU	0.001854
tipo_de_habitacion__1BAI	0.001750
agencia__HEIM	0.001724
agencia__COLAB	0.001672
agencia__PTAGE	0.001631
agencia__DERFRA	0.001558
agencia__SUMA	0.001538
agencia__SSUR	0.001369
tipo_de_habitacion__1BAE	0.001324
agencia__NHL	0.001165
agencia__SCHAU	0.000904
agencia__NOVA	0.000854
agencia__BUC	0.000821
agencia__ITSAU	0.000806

tipo_de_habitacion__2BAI	0.000797
tipo_de_habitacion__2BAE	0.000767
agencia__ITSSU	0.000675
agencia__KUO	0.000632
tipo_de_habitacion__STDB	0.000628
tipo_de_habitacion__STDS	0.000626
agencia__BOOK1	0.000602
agencia__BARCE	0.000601
agencia__TUIAU	0.000582
agencia__BOOK	0.000470
tipo_de_habitacion__2BAS	0.000397
agencia__NOMA	0.000350
tipo_de_habitacion__1BAS	0.000332
tipo_de_habitacion__1BAF	0.000190
agencia__VIACO	0.000164
agencia__SOCEAN	0.000151
agencia__PAN	0.000150
agencia__HALCON	0.000106
agencia__LUNA	0.000106
tipo_de_habitacion__VLB	0.000088
agencia__CORTE	0.000084
agencia__BEST	0.000071
agencia__CLICK	0.000040
agencia__CANAT	0.000034
agencia__AIRBUS	0.000033
agencia__TENMA	0.000017
agencia__UNITV	0.000000
agencia__SATT	0.000000
agencia__NIEVE	0.000000
agencia__VSIERR	0.000000
agencia__HERNA	0.000000
agencia__BADA	0.000000
agencia__ERIOJA	0.000000
agencia__GURIA	0.000000
agencia__GLOBA	0.000000
agencia__EROSKI	0.000000
agencia__ARROBA	0.000000
agencia__ECU	0.000000
agencia__COVA	0.000000
agencia__ACONFI	0.000000
agencia__ALISIO	0.000000
agencia__ALLTOU	0.000000
agencia__BIVES	0.000000
agencia__AMAIA	0.000000
agencia__ANUR	0.000000
agencia__ZADY	0.000000
dtype: float64	

```
In [80]: X_imp_features=X[['dia_de_llegada', 'dia_de_la_reserva', 'dia_de_salida', 'numero_de_personas', 'mes_de_salida', 'mes_de_la_reserva', 'nombre_del_hotel', 'agencia__COSMOS', 'agencia__SOME', 'agencia__DIR']]
X_imp_features
```

Out[80]:

	dia_de_llegada	dia_de_la_reserva	dia_de_salida	numero_de_personas	mes_de_salida	m
0	23	23	23	2	4	
1	29	29	29	2	8	
2	14	14	14	2	6	
3	4	4	4	4	1	
4	25	25	25	2	7	
...	...	...	...	...	...	...
72376	30	30	30	2	12	
72377	17	17	17	3	9	
72378	22	22	22	2	10	
72379	24	24	24	2	6	
72380	30	30	30	2	10	

72381 rows × 10 columns



```
In [84]: X_train_imp_feat, X_test_imp_feat, y_train_imp_feat, y_test_imp_feat = train_test_split(X_imp_features, y, test_size=0.3, stratify=y, random_state=0)
```

```
In [100]: X_train_imp_feat.shape, y_train_imp_feat.shape, X_test_imp_feat.shape, y_test_imp_feat.shape
```

Out[100]: ((50666, 10), (50666,), (21715, 10), (21715,))

```
In [101]: dtree_imp_feat = DecisionTreeClassifier()
dtree_imp_feat = dtree_imp_feat.fit(X_train_imp_feat, y_train_imp_feat)
```

```
In [102]: print(dtree_imp_feat.predict(X_test_imp_feat))

[0 1 0 ... 0 0 0]
```

```
In [103]: from sklearn import metrics
print("Accuracy:", metrics.accuracy_score(y_test_imp_feat, dtree_imp_feat.predict(X_test_imp_feat)))
```

Accuracy: 0.7196408012894313

```
In [122]: print(" === Clasification report Decision Tree - Reduced Features ===")
print(classification_report(y_test_imp_feat,dtree_imp_feat.predict(X_test_imp_
feat)))
print(" === Confusion Matrix - Reduced Features ===")
print(confusion_matrix(y_test_imp_feat,dtree_imp_feat.predict(X_test_imp_feat
)))
```

```
=== Clasification report Decision Tree - Reduced Features ===
              precision    recall  f1-score   support

     0       0.74      0.94      0.82      15303
     1       0.57      0.20      0.30       6412

 micro avg       0.72      0.72      0.72      21715
 macro avg       0.65      0.57      0.56      21715
weighted avg       0.69      0.72      0.67      21715

=== Confusion Matrix - Reduced Features ===
[[14344   959]
 [ 5129 1283]]
```

Utilizamos Decision Tree con RandomizedSearch CV

```
In [113]: from scipy.stats import randint

param_RandSearch = {"max_depth": [3, None],
                    "max_features": randint(1, 9),
                    "min_samples_leaf": randint(1, 9),
                    "criterion": ["gini", "entropy"]}
```

```
In [114]: tree_cv = RandomizedSearchCV(dtree_imp_feat, param_RandSearch, cv=5)
```

```
In [115]: tree_cv.fit(X_train_imp_feat,y_train_imp_feat)
```

```
Out[115]: RandomizedSearchCV(cv=5, error_score='raise-deprecating',
                             estimator=DecisionTreeClassifier(class_weight=None, criterion='gini',
                             max_depth=None,
                             max_features=None, max_leaf_nodes=None,
                             min_impurity_decrease=0.0, min_impurity_split=None,
                             min_samples_leaf=1, min_samples_split=2,
                             min_weight_fraction_leaf=0.0, presort=False, random_state=None,
                             splitter='best'),
                             fit_params=None, iid='warn', n_iter=10, n_jobs=None,
                             param_distributions={'max_depth': [3, None], 'max_features': <scipy
                             y.stats._distn_infrastructure.rv_frozen object at 0x0000022F0FD3BC08>, 'min_s
                             amples_leaf': <scipy.stats._distn_infrastructure.rv_frozen object at 0x000002
                             2F0FA14048>, 'criterion': ['gini', 'entropy']},
                             pre_dispatch='2*n_jobs', random_state=None, refit=True,
                             return_train_score='warn', scoring=None, verbose=0)
```

```
In [118]: print("Tuned Decision Tree Parameters: {}".format(tree_cv.best_params_))
print("Best score is {}".format(tree_cv.best_score_))
print("Accuracy:", metrics.accuracy_score(y_test_imp_feat, tree_cv.predict(X_test_imp_feat)))
```

Tuned Decision Tree Parameters: {'criterion': 'gini', 'max_depth': 3, 'max_features': 5, 'min_samples_leaf': 4}
 Best score is 0.7265029803023724
 Accuracy: 0.7267326732673267

```
In [123]: print(" == Clasification report Decision Tree - Randomized Search - Reduced features ==")
print(classification_report(y_test_imp_feat, tree_cv.predict(X_test_imp_feat)))
print(" == Confusion Matrix - Randomized Search - Reduced features ==")
print(confusion_matrix(y_test_imp_feat, tree_cv.predict(X_test_imp_feat)))
```

== Clasification report Decision Tree - Randomized Search - Reduced features ==

	precision	recall	f1-score	support
0	0.74	0.95	0.83	15303
1	0.62	0.19	0.29	6412
micro avg	0.73	0.73	0.73	21715
macro avg	0.68	0.57	0.56	21715
weighted avg	0.70	0.73	0.67	21715

== Confusion Matrix - Randomized Search - Reduced features ==
 [[14575 728]
 [5206 1206]]

Aplicamos Decision Tree con GridSearchCV

```
In [145]: param_dtree_gridSearch={
    'criterion': ['gini', 'entropy'],
    'max_depth': [3,4,5,6,7,8],
    'max_features': [4,5,6,7,8,9,10],
    'min_samples_leaf': [3,4,5]
}
```

```
In [146]: dtree_gridSearchCV= GridSearchCV(estimator=DecisionTreeClassifier(),
    param_grid=param_dtree_gridSearch,
    cv=5,
    refit=True,
    n_jobs=-1)
```

```
In [147]: dtree_gridSearchCV.fit(X_train_imp_feat,y_train_imp_feat)
```

```
Out[147]: GridSearchCV(cv=5, error_score='raise-deprecating',
      estimator=DecisionTreeClassifier(class_weight=None, criterion='gini',
      max_depth=None,
      max_features=None, max_leaf_nodes=None,
      min_impurity_decrease=0.0, min_impurity_split=None,
      min_samples_leaf=1, min_samples_split=2,
      min_weight_fraction_leaf=0.0, presort=False, random_state=None,
      splitter='best'),
      fit_params=None, iid='warn', n_jobs=-1,
      param_grid={'criterion': ['gini', 'entropy'], 'max_depth': [3, 4, 5,
      6, 7, 8], 'max_features': [4, 5, 6, 7, 8, 9, 10], 'min_samples_leaf': [3, 4,
      5]},
      pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
      scoring=None, verbose=0)
```

```
In [148]: dtree_gridSearchCV.predict(X_test_imp_feat)
```

```
Out[148]: array([0, 0, 0, ..., 0, 0, 0])
```

```
In [149]: print("Accuracy:",metrics.accuracy_score(y_test_imp_feat, dtree_gridSearchCV.p
      predict(X_test_imp_feat)))
      print(" == Clasifcation report  Decision Tree - GridSearch - Reduced feature
      s == \n")
      print(classification_report(y_test_imp_feat,dtree_gridSearchCV.predict(X_test_
      imp_feat)))
      print(" == Confusion Matrix - Randomized Search - Reduced features ==")
      print(confusion_matrix(y_test_imp_feat,dtree_gridSearchCV.predict(X_test_imp_f
      eat)))
```

Accuracy: 0.7357586921482846

== Clasifcation report Decision Tree - GridSearch - Reduced features ==

	precision	recall	f1-score	support
0	0.74	0.96	0.84	15303
1	0.68	0.20	0.31	6412
micro avg	0.74	0.74	0.74	21715
macro avg	0.71	0.58	0.57	21715
weighted avg	0.72	0.74	0.68	21715

== Confusion Matrix - Randomized Search - Reduced features ==

```
[[14698  605]
 [ 5133 1279]]
```

```
In [ ]:
```

Aplicamos random forest

```
In [82]: from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import cross_val_score
```

```
rfc=RandomForestClassifier(
    n_estimators=100,
    criterion='gini',
    max_depth=None,
    min_samples_split=2,
    min_samples_leaf=1,
    min_weight_fraction_leaf=0.0,
    max_features='auto',
    max_leaf_nodes=None,
    bootstrap=True,
    oob_score=False,
    n_jobs=-1,
    random_state=123,
    verbose=0,
    warm_start=False,
    class_weight=None
)

rfc.fit(X_train, y_train)
```

```
Out[82]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
    max_depth=None, max_features='auto', max_leaf_nodes=None,
    min_impurity_decrease=0.0, min_impurity_split=None,
    min_samples_leaf=1, min_samples_split=2,
    min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=-1,
    oob_score=False, random_state=123, verbose=0, warm_start=False)
```

```
In [68]: rfc_predict = rfc.predict(X_test)
```

```
In [ ]:
```

```
In [69]: print(" == Clasifcation report ==")
print(classification_report(y_test,rfc_predict))
```

```
== Clasifcation report ==
```

	precision	recall	f1-score	support
0	0.77	0.85	0.81	15303
1	0.53	0.41	0.46	6412
micro avg	0.72	0.72	0.72	21715
macro avg	0.65	0.63	0.64	21715
weighted avg	0.70	0.72	0.71	21715

```
In [70]: print(" == Confusion Matrix ==")  
print(confusion_matrix(y_test, rfc_predict))
```

```
== Confusion Matrix ==  
[[12960  2343]  
 [ 3776  2636]]
```

```
In [71]: from sklearn.metrics import accuracy_score  
accuracy_score(y_test, rfc_predict)
```

```
Out[71]: 0.7182132166705043
```

Medimos la importancia de las variables


```
In [72]: #Medimos la importancia de las variables utilizando la lista de features, es decir, variables independientes  
feature_imp = pd.Series(rfc.feature_importances_, index=list(X)).sort_values(ascending=False)  
feature_imp
```

```

Out[72]: dia_de_llegada      0.126239
          dia_de_salida      0.126072
          dia_de_la_reserva  0.125319
          numero_de_personas 0.123146
          mes_de_salida      0.049379
          mes_de_llegada     0.049226
          mes_de_la_reserva  0.048724
          nombre_del_hotel   0.037746
          agencia__COSMOS    0.031258
          agencia__DIR       0.026900
          agencia__SOME      0.026176
          agencia__UNI       0.012945
          agencia__OTB       0.012168
          agencia__WEB       0.011414
          tipo_de_habitacion__STD 0.011258
          tipo_de_habitacion__1BA 0.010803
          agencia__JET2      0.010727
          agencia__DTOS      0.010320
          agencia__TLT       0.009741
          agencia__VING      0.009043
          agencia__ARTURO    0.008065
          tipo_de_habitacion__2BA 0.007792
          agencia__TCNE      0.007722
          agencia__ARK       0.007496
          agencia__ALTURA   0.006982
          agencia__TUI       0.005445
          agencia__NVN       0.005071
          agencia__VITA      0.004407
          agencia__FIR       0.004375
          agencia__SUMA      0.004058
          tipo_de_pension    0.003936
          agencia__HOL       0.003933
          agencia__HEIM      0.003840
          agencia__ITS       0.003807
          agencia__NUR       0.003708
          agencia__SCHAU     0.003187
          agencia__PTWEB     0.003139
          tipo_de_habitacion__STDE 0.002912
          tipo_de_habitacion__STDI 0.002871
          agencia__COLAB     0.002630
          tipo_de_habitacion__1BAI 0.002465
          agencia__SAGA      0.002248
          agencia__BOOK1     0.002145
          agencia__SSUR      0.002037
          tipo_de_habitacion__3BA 0.001864
          agencia__NHL       0.001689
          agencia__TJAE      0.001681
          agencia__HCH       0.001641
          agencia__TUIPO     0.001571
          tipo_de_habitacion__1BAE 0.001545
          agencia__DERFRA    0.001391
          agencia__MCCA      0.001341
          agencia__NOVA      0.001280
          tipo_de_habitacion__VLA 0.001171
          agencia__BOOK      0.001076
          agencia__TUISU     0.001041
          agencia__KUO       0.000949

```

```

agencia__PTAGE      0.000936
tipo_de_habitacion__STDS  0.000843
agencia__ITSAU      0.000767
agencia__BUC        0.000689
tipo_de_habitacion__2BAE  0.000652
tipo_de_habitacion__1BAS  0.000485
agencia__ITSSU      0.000460
tipo_de_habitacion__VLB  0.000436
tipo_de_habitacion__2BAS  0.000425
tipo_de_habitacion__2BAI  0.000384
agencia__NOMA       0.000364
agencia__TUIAU      0.000353
agencia__BARCE      0.000328
tipo_de_habitacion__STDB  0.000309
tipo_de_habitacion__1BAF  0.000219
agencia__GURIA      0.000189
agencia__HALCON     0.000178
agencia__VIACO      0.000166
agencia__LUNA       0.000089
agencia__SOCEAN     0.000085
agencia__CORTE      0.000055
agencia__EROSKI     0.000046
agencia__BEST       0.000045
agencia__TENMA      0.000040
agencia__NIEVE      0.000036
agencia__PAN        0.000034
agencia__ALISIO     0.000032
agencia__ANUR       0.000028
agencia__VSIERR     0.000019
agencia__AIRBUS     0.000019
agencia__BIVES      0.000017
agencia__CANAT      0.000016
agencia__UNITV      0.000016
agencia__ZADY       0.000015
agencia__ACONFI     0.000014
agencia__HERNA      0.000013
agencia__GLOBA      0.000013
agencia__BADA       0.000012
agencia__CLICK      0.000012
agencia__COVA       0.000011
agencia__ERIOJA     0.000010
agencia__ARROBA     0.000010
agencia__ALLTOU     0.000007
agencia__SATT       0.000006
agencia__AMAIA      0.000000
agencia__ECU        0.000000
dtype: float64

```

```
In [73]: importance=rfc.feature_importances_
```

Como las variables más importantes son las mismas que en el árbol de decisión, utilizamos la misma variable `X_train_imp_feat`

```
In [85]: rfc_imp_feat=RandomForestClassifier(
    n_estimators=100,
    criterion='gini',
    max_depth=None,
    min_samples_split=2,
    min_samples_leaf=1,
    min_weight_fraction_leaf=0.0,
    max_features='auto',
    max_leaf_nodes=None,
    bootstrap=True,
    oob_score=False,
    n_jobs=-1,
    random_state=123,
    verbose=0,
    warm_start=False,
    class_weight=None
)

rfc_imp_feat.fit(X_train_imp_feat, y_train_imp_feat)
```

```
Out[85]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
    max_depth=None, max_features='auto', max_leaf_nodes=None,
    min_impurity_decrease=0.0, min_impurity_split=None,
    min_samples_leaf=1, min_samples_split=2,
    min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=-1,
    oob_score=False, random_state=123, verbose=0, warm_start=False)
```

```
In [86]: print("Accuracy:",metrics.accuracy_score(y_test_imp_feat, rfc_imp_feat.predict(
    (X_test_imp_feat)))
    print(" == Clasifcation report  RandomForest - GridSearch - Reduced features
    == \n")
    print(classification_report(y_test_imp_feat,rfc_imp_feat.predict(X_test_imp_fe
    at)))
    print(" == Confusion Matrix - RandomForest - GridSearch - Reduced features ==
    ==")
    print(confusion_matrix(y_test_imp_feat,rfc_imp_feat.predict(X_test_imp_feat)))
```

Accuracy: 0.7229564816946811

== Clasifcation report RandomForest - GridSearch - Reduced features ==

	precision	recall	f1-score	support
0	0.74	0.93	0.83	15303
1	0.58	0.22	0.32	6412
micro avg	0.72	0.72	0.72	21715
macro avg	0.66	0.58	0.57	21715
weighted avg	0.69	0.72	0.68	21715

== Confusion Matrix - RandomForest - GridSearch - Reduced features ==

```
[[14289 1014]
 [ 5002 1410]]
```

Aplicamos randomized search

```
In [87]: from pprint import pprint
# Look at parameters used by our current forest
print('Parameters currently in use:\n')
pprint(rfc.get_params())
```

Parameters currently in use:

```
{'bootstrap': True,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': None,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 100,
 'n_jobs': -1,
 'oob_score': False,
 'random_state': 123,
 'verbose': 0,
 'warm_start': False}
```

```

In [88]: from sklearn.model_selection import RandomizedSearchCV

# Numero de arboles que componen el ramdon forest
n_estimators = [int(x) for x in np.linspace(start = 200, stop = 2000, num = 10
)]
# Numero de funciones a implementar
max_features = ['auto', 'sqrt']
# Máximo nº de niveles en el arbol
max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
max_depth.append(None)
# Mínimo numero de muestras requerido para separar un nodo
min_samples_split = [2, 5, 10]
# Número mínimo de muestras necesarias en cada nodo de la hoja en el random fo
rest
min_samples_leaf = [1, 2, 4]
# Metodo de seleccion de muestras para entrar cada árbol
bootstrap = [True, False]

# Creamos la red aleatoria de parámetros
random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap}

pprint(random_grid)

{'bootstrap': [True, False],
 'max_depth': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None],
 'max_features': ['auto', 'sqrt'],
 'min_samples_leaf': [1, 2, 4],
 'min_samples_split': [2, 5, 10],
 'n_estimators': [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000]}

```

```
In [89]: # Búsqueda aleatoria de parámetros, suando tres validaciones cruzadas,
# buscar a través de 100 combiaciones diferentes usando todos los núcleos
rf_random = RandomizedSearchCV(estimator=rfc, param_distributions=random_grid,
                               n_iter = 100, scoring='accuracy',
                               cv = 3, verbose=2, random_state=42, n_jobs=-1,
                               return_train_score=True)

# Entrenar el modelo
rf_random.fit(X_train_imp_feat, y_train_imp_feat);
```

Fitting 3 folds for each of 100 candidates, totalling 300 fits

```
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 4 concurrent workers.
[Parallel(n_jobs=-1)]: Done 33 tasks      | elapsed: 6.0min
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\externals\joblib\externals\
loky\process_executor.py:706: UserWarning: A worker stopped while some jobs
were given to the executor. This can be caused by a too short worker timeout
or by a memory leak.
  "timeout or by a memory leak.", UserWarning
[Parallel(n_jobs=-1)]: Done 154 tasks      | elapsed: 30.2min
[Parallel(n_jobs=-1)]: Done 300 out of 300 | elapsed: 57.5min finished
```

```
In [91]: rf_random.best_params_
```

```
Out[91]: {'n_estimators': 200,
          'min_samples_split': 5,
          'min_samples_leaf': 4,
          'max_features': 'auto',
          'max_depth': 10,
          'bootstrap': True}
```

```
In [93]: best_random = rf_random.best_estimator_
```

```
In [84]: best_random
```

```
Out[84]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                                max_depth=100, max_features='auto', max_leaf_nodes=None,
                                min_impurity_decrease=0.0, min_impurity_split=None,
                                min_samples_leaf=4, min_samples_split=5,
                                min_weight_fraction_leaf=0.0, n_estimators=1000, n_jobs=-1,
                                oob_score=False, random_state=123, verbose=0, warm_start=False)
```

```
In [94]: best_random.fit(X_train_imp_feat, y_train_imp_feat)
```

```
Out[94]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                                max_depth=10, max_features='auto', max_leaf_nodes=None,
                                min_impurity_decrease=0.0, min_impurity_split=None,
                                min_samples_leaf=4, min_samples_split=5,
                                min_weight_fraction_leaf=0.0, n_estimators=200, n_jobs=-1,
                                oob_score=False, random_state=123, verbose=0, warm_start=False)
```

```
In [81]: best_random_predict = best_random.predict(X_test)
```

In [85]: `accuracy_score(y_test, best_random_predict)`

Out[85]: 0.7511397651393046

In [95]: `print("Accuracy:", metrics.accuracy_score(y_test_imp_feat, best_random.predict(X_test_imp_feat)))`
`print(" == Clasification report RandomForest - RandomizedSearch - Reduced features == \n")`
`print(classification_report(y_test_imp_feat, best_random.predict(X_test_imp_feat)))`
`print(" == Confusion Matrix - RandomForest - RandomizedSearch - Reduced features ==")`
`print(confusion_matrix(y_test_imp_feat, best_random.predict(X_test_imp_feat)))`

Accuracy: 0.7384757080359199

== Clasification report RandomForest - RandomizedSearch - Reduced features ==

	precision	recall	f1-score	support
0	0.74	0.98	0.84	15303
1	0.74	0.17	0.28	6412
micro avg	0.74	0.74	0.74	21715
macro avg	0.74	0.57	0.56	21715
weighted avg	0.74	0.74	0.68	21715

== Confusion Matrix - RandomForest - RandomizedSearch - Reduced features ==

```

=
[[14921  382]
 [ 5297 1115]]

```

Aplicamos Grid search

In [154]: *#En este caso aplicamos la búsqueda en malla combinada con la validación cruzada, basándonos en los parámetros obtenidos del best_random*

```

gs_cv_grid = {
    'bootstrap': [True],
    'max_depth': [80, 90, 100, 110],
    'criterion': ['gini'],
    'max_features': ['auto'],
    'min_samples_leaf': [3, 4, 5],
    'min_samples_split': [4, 5, 6],
    'n_estimators': [800, 900, 1100]
}

```

In [155]: `rfc_gs_cv=RandomForestClassifier()`


```
In [156]: from sklearn.model_selection import GridSearchCV

grid_search=GridSearchCV(estimator=rfc_gs_cv, param_grid=gs_cv_grid,cv=3,n_jobs=-1, verbose=2)
```

```
In [157]: grid_search.fit(X_train, y_train)
```

Fitting 3 folds for each of 108 candidates, totalling 324 fits

```
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 4 concurrent workers.
[Parallel(n_jobs=-1)]: Done 33 tasks      | elapsed: 8.8min
[Parallel(n_jobs=-1)]: Done 154 tasks    | elapsed: 38.3min
[Parallel(n_jobs=-1)]: Done 324 out of 324 | elapsed: 80.9min finished
```

```
Out[157]: GridSearchCV(cv=3, error_score='raise-deprecating',
      estimator=RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
      max_depth=None, max_features='auto', max_leaf_nodes=None,
      min_impurity_decrease=0.0, min_impurity_split=None,
      min_samples_leaf=1, min_samples_split=2,
      min_weight_fraction_leaf=0.0, n_estimators='warn', n_jobs=None,
      oob_score=False, random_state=None, verbose=0,
      warm_start=False),
      fit_params=None, iid='warn', n_jobs=-1,
      param_grid={'bootstrap': [True], 'max_depth': [80, 90, 100, 110], 'criterion': ['gini'], 'max_features': ['auto'], 'min_samples_leaf': [3, 4, 5], 'min_samples_split': [4, 5, 6], 'n_estimators': [800, 900, 1100]},
      pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
      scoring=None, verbose=2)
```

```
In [158]: print("Accuracy:",metrics.accuracy_score(y_test, grid_search.predict(X_test)))
print(" == Clasifcation report RandomForest - GridSearch - Reduced features == \n")
print(classification_report(y_test,grid_search.predict(X_test)))
print(" == Confusion Matrix - RandomForest - GridSearch - Reduced features == ")
print(confusion_matrix(y_test,grid_search.predict(X_test)))
```

Accuracy: 0.7506332028551692

== Clasifcation report RandomForest - GridSearch - Reduced features ==

	precision	recall	f1-score	support
0	0.76	0.95	0.84	15303
1	0.69	0.29	0.40	6412
micro avg	0.75	0.75	0.75	21715
macro avg	0.72	0.62	0.62	21715
weighted avg	0.74	0.75	0.71	21715

== Confusion Matrix - RandomForest - GridSearch - Reduced features ==

```
[[14467  836]
 [ 4579 1833]]
```



SMART YIELD

HOTEL DATA SYSTEM

Resumen Ejecutivo

Sistema integral para la gestión del dato en empresas hoteleras y de inteligencia de negocio para su explotación.

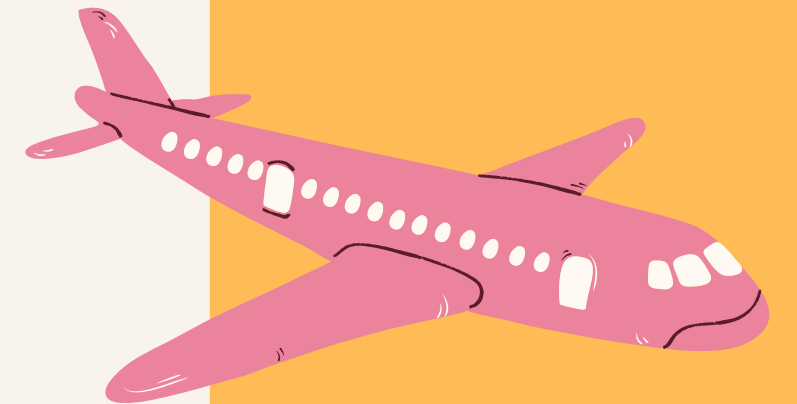


PROBLEMA DE NEGOCIO

Actualmente el mundo del turismo presenta unos grandes cambios a nivel operativo y táctico de los grandes actores del mercado. Incluso para un modelo clásico de gestión han aparecido nuevos canales y herramientas como son las OTAs, metabuscadores, etc.

Por otro lado, hay que competir con grandes departamentos de revenue de otros hoteles que sí son soportados por el dato. A su vez, casi todos los alojamientos turísticos están en búsqueda de mejorar la experiencia del cliente.

Por tanto, es necesario disponer de tácticas y estrategias para la obtención, almacenamiento y explotación de los datos de los clientes. Desde sus datos personales a cualquier interacción con nuestra empresa: helpdesk, incidencias, sugerencias, historiales...



Optimización de procesos para el sector hotelero

Generamos una oferta de unos precios adaptados a las demandas reales de servicios de los clientes, con unas acciones de marketing perfectamente segmentadas. Esto nos permitirá optimizar todos los procesos de gestión del hotel, eliminando ofertas y servicios innecesario. No solo optimizamos la venta y la experiencia de cliente, sino que disminuimos los costes.

PROPUESTA

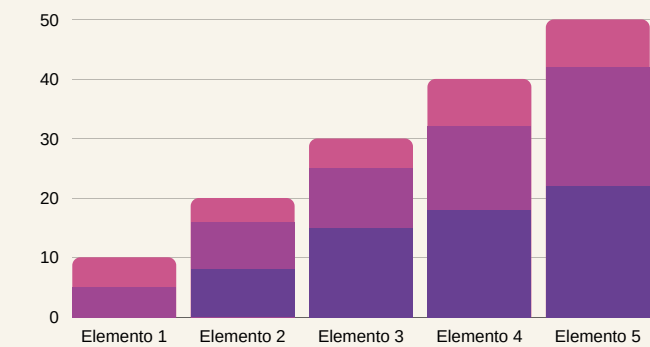


DATA WAREHOUSE

Almacenando todos los datos que se generan por parte de estas empresas, pero que mayoritariamente insertan en software privativos PMS (Property management system) de los cuales no pueden gestionar de un manera eficiente y versátil.

DASHBOARD

Aplicando dashboard sencillos y fáciles de entender para una correcta toma de decisiones por parte de la dirección de estas empresas.



MACHINE LEARNING

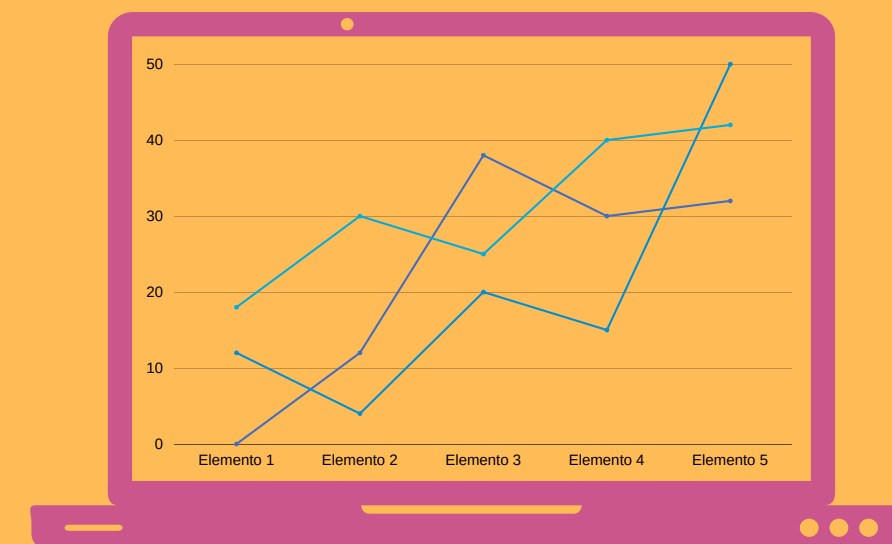
Aplicando algoritmos de Machine Learning e Inteligencia Artificial para:

Sistemas de Revenue.

Encuestas de satisfacción.

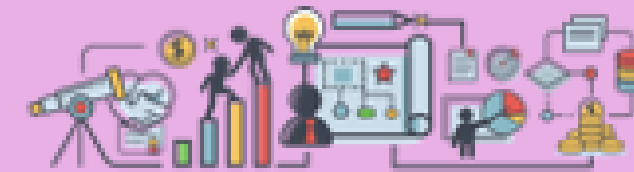
Sistemas de reservas y cancelaciones.

Gestión interna de la organización del trabajo en el hotel.



OPORTUNIDAD

VENTAS + PRODUCCIÓN + GESTIÓN => DATOS



Optimización de procesos para el sector hotelero



Un sistema de gestión basado en el dato permitiría la toma de decisiones tácticas y técnicas basada en hechos empíricos, redundando directamente en beneficios y rapidez para gerencia.





BI

El sistema genera una inteligencia de negocio basada en datos históricos, modelos de optimización, minimiza tiempos de reacción en cualquier circunstancia adversa e intervenciones urgentes, proporcionando beneficio directo.

VISUALIZACIÓN

El trabajo diario se verá beneficiado por la implantación de un sistema de cuadros de mando con datos centralizados, agilizando la comunicación, la distribución de la información y de trabajo de manera más efectiva.

GESTIÓN INTERNA

Los partes de trabajos realizados por todos los departamentos del hotel, se gestionan de manera más óptima por parte de la dirección del hotel, pudiendo estos acceder a todo el trabajo efectuado por los empleados en cualquier momento.



REVENUE

Consiguiendo una gestión orientada al dato, y utilizando las nuevas herramientas de BD & BI, se consiguen extraer indicadores y tomar decisiones estratégicas que mejoren los ingresos “Revenue”.



EXPERIENCIA

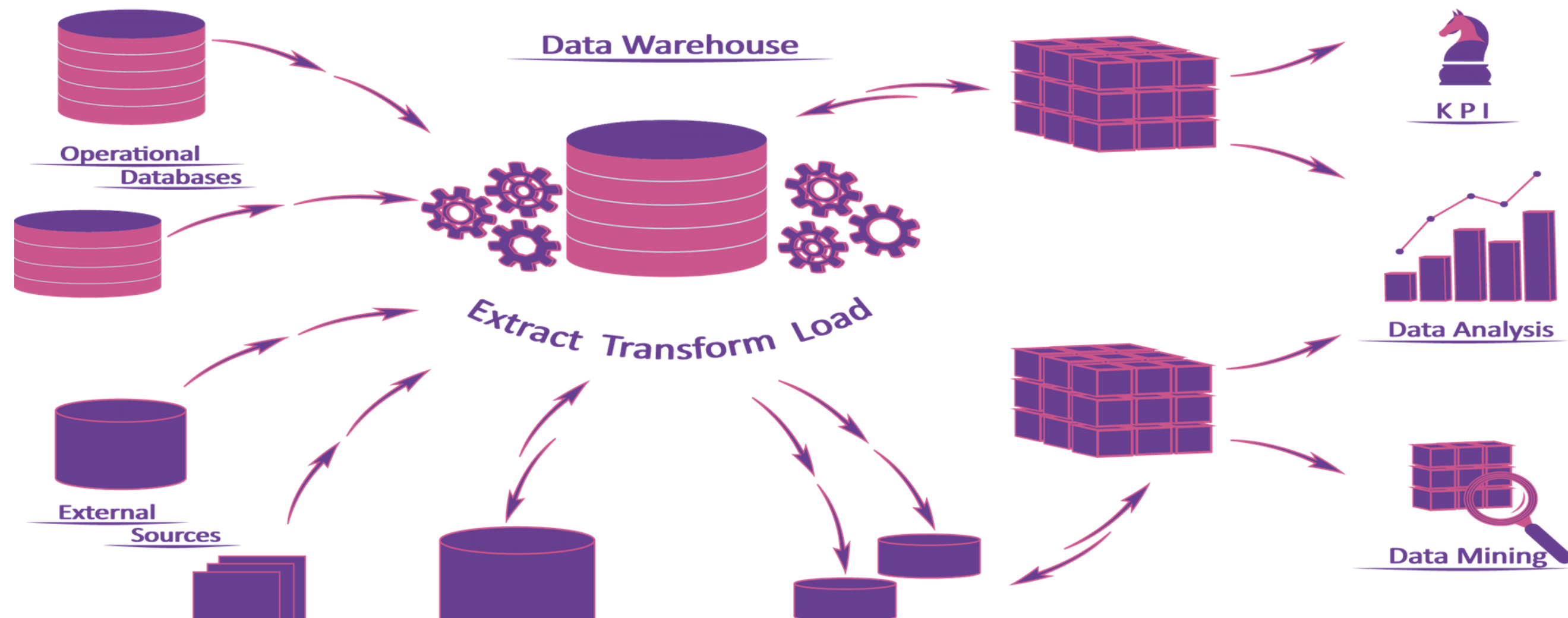
Basándose en el dato como herramienta, mejora la gestión interna del hotel, y podremos desarrollar una experiencia de cliente más adaptada, que nos otorgará competitividad.



COSTE

El coste del proyecto no es elevado, dado que realizamos proyectos utilizando tecnología Open Source y adaptable a cada cliente.

DISEÑO TECNOLÓGICO



Recursos disponibles buscando alcanzar la mayor eficiencia



DATOS

- **MySql**
- **Pentaho BI Suite**
- **Excel**
- **Json**

Procesos de obtención y transformación de los datos en Excel, Json, XML, a una Base de datos MySql, con una solución de inteligencia Negocio Pentaho BI Suite.

MACHINE LEARNING

- **Classification**
- **Forecasting**
- **Análisis de sentimiento**

Modelos de clasificación y análisis de sentimiento para clasificar los clientes y ofrecer mejores servicios y más personalizados.
Modelos predictivos “forecast” para definir los precios de venta

DASHBOARD

- **Pentaho BI Suite**
- **PowerBI**

Visualizaciones de cuadros de mandos interactivos

INDICADORES CLAVES



**EFFECTIVIDAD DE LA
HERRAMIENTA**



AUMENTO DE BENEFICIOS



**AUMENTAR CARTERA DE
CLIENTES**



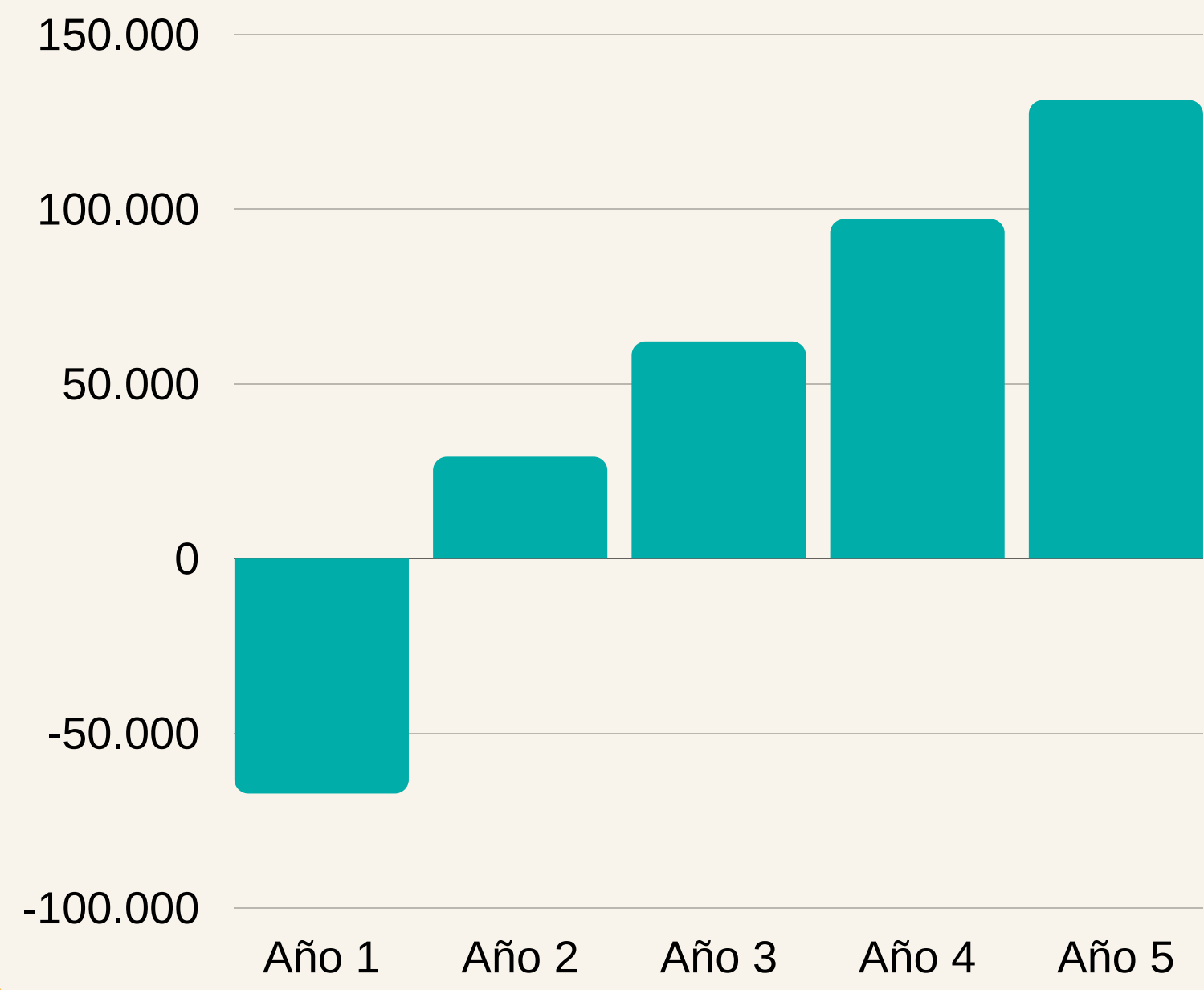
ANÁLISIS DEL MERCADO



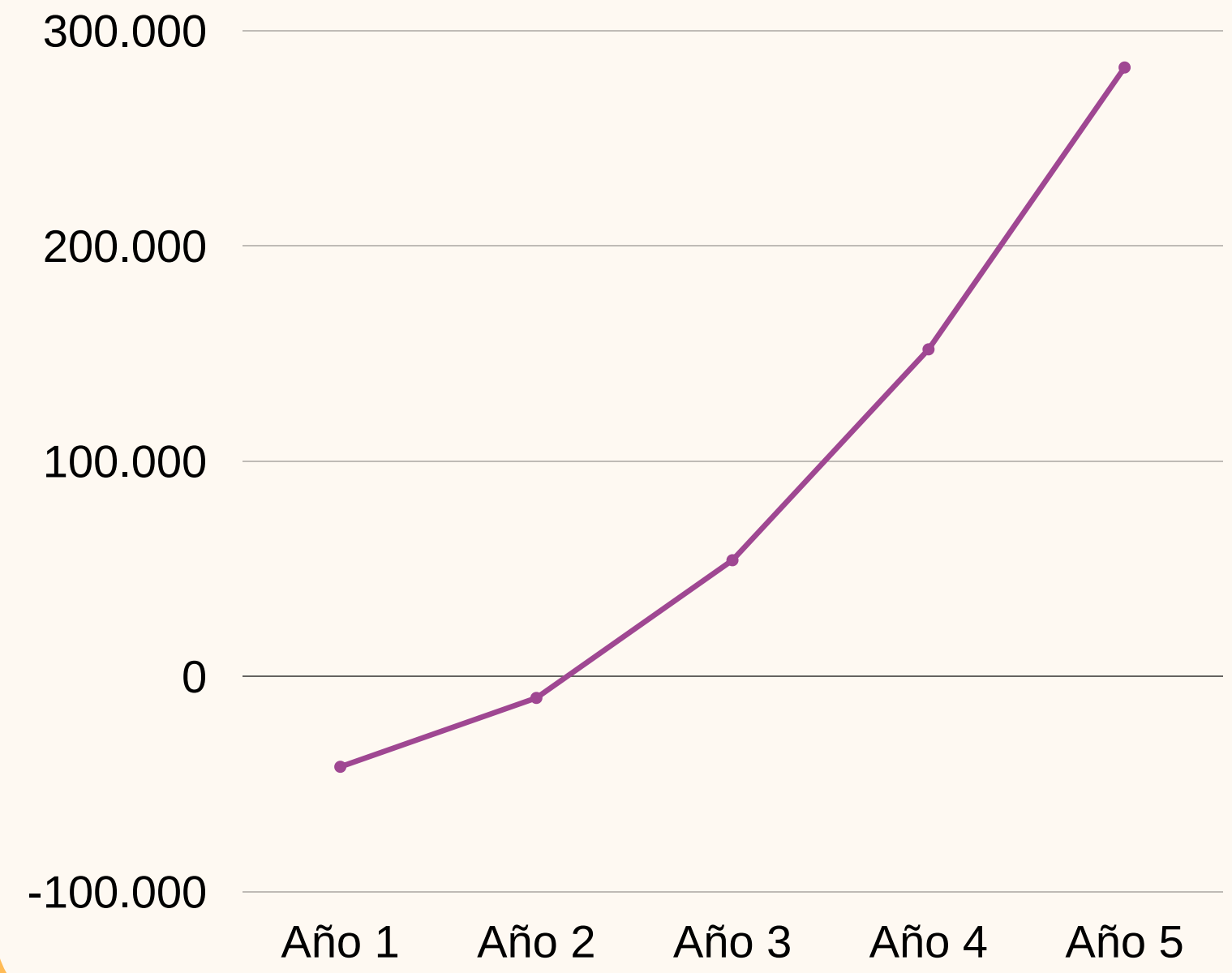
**SATISFACCIÓN DE
CLIENTES**

CUENTA DE RESULTADOS

BENEFICIO NETO



FLUJO DE CAJA



RENTABILIDAD

TIR

103%

VAN

273.403€

PAYBACK

**Al comienzo
del 2º año**

TIRM

76%



SMART YIELD

HOTEL DATA SYSTEM