Proyecto Júpiter – Reconocimiento de emociones

El proyecto Júpiter tiene como objetivo principal terminar de unir los conceptos para el alumno de forma práctica mediante un proyecto.

Introducción

La empresa de parques temáticos Pontia World quiere mejorar la experiencia de los visitantes a sus distintos parques de atracciones y evaluar la experiencia de los usuarios que disfrutan de cada atracción. Esta compañía recibe miles de visitantes diarios. Todos los parques están constantemente videovigilados y las distintas atracciones captan en muchos momentos fotos de los visitantes (en muchas ocasiones para su posterior venta). En muchas ocasiones estos vídeos los utilizan los analistas de Pontia World para identificar patrones de elección de atracciones. En esta ocasión, se desean usar imágenes extraídas de los vídeos y las fotos para determinar las emociones de los visitantes, de manera que puedan ver la experiencia de estos durante su estancia en el parque e identificar, por ejemplo, emociones preponderantes en cada una de las atracciones. Sin embargo, resulta muy costoso e ineficiente llevar a cabo un etiquetado manual de las imágenes para luego obtener estadísticas (se necesitan muchas imágenes para alcanzar resultados concluyentes).

La empresa carece de un sistema adecuado para almacenar y gestionar sus datos, por lo que no solo necesita ser capaz de automatizar el etiquetado, sino que es necesario llevar a cabo una transformación digital completa alrededor de estos datos: empezando por su almacenamiento, pasando por su procesamiento y finalizando con la generación de resultados y cálculo de KPIs útiles para el negocio.

¿Qué aprenderás en el proyecto?

- Gestión de tablas con SQL:
    - Creación de tablas
    - Asignación de claves primarias y claves foráneas
    - Asignación de tipos de datos
    - Normalización de tablas
    - Inserción de datos
    - Modificar o Alterar tablas
- Conexión a una BBDD con Python
- Análisis exploratorio de datos:
    - Cálculo de estadísticos descriptivos
    - Cálculo de correlación
    - PCA

    - Visualización
    - Detección de *outliers*
    - Análisis de la distribución de los datos
    - Análisis de la asimetría
    - Análisis de curtosis
    - Detección de *missing values*
    - Identificación de datos erróneos
- Limpieza de datos con SQL:
    - Convertir a decimales campos numéricos
    - Creación de vistas
    - Unión con otras tablas para corroborar información JOIN
    - Funciones de limpieza como: CAST, RIGHT, REPLACE
    - Filtrar tablas con WHERE
    - Tratamiento de NULL
    - Categorizar con CASE
    - Redondear campos con la función CEILING
    - Detectar duplicados en una tabla con CTE'S & HAVING
    - Eliminar registros de una tabla DELETE
    - Parsear un JSON con SQL
- Limpieza de datos con Python:
    - Crear un pipeline de limpieza de datos
- Machine Learning:
    - Definición del problema y los objetivos
    - Selección de datos
    - Data Augmentation y técnicas de corrección del balanceo
    - Selección del algoritmo de aprendizaje
    - Entrenamiento y parametrización del algoritmo
    - Evaluación del modelo
    - Explotación del modelo
    - Comparativa de modelos
    - Metodología iterativa del ML
- Presentación de resultados:
    - Visualización de resultados
    - Conexión de herramienta de visualización con BBDD
    - Cálculo de KPIs
    - Diseño e implementación de dashboards
    - Storytelling
    - Informe de proyecto

Herramientas a usar en el proyecto

- **MySQL:** se utilizará una base de datos que se importará a MySQL para trabajar desde allí.

- **Python:** utilizaremos python para limpiar y analizar los datos, así como el desarrollo de los modelos de ML.
- **Power BI / Tableau:** haremos uso de una herramienta de visualización para presentar los resultados.

Recursos complementarios para el proyecto

- Documentación de Scikit-Learn: https://scikit-learn.org
- Documentación de MySQL: https://www.mysql.com/
- Repositorio de datos: TBD
- Documentación de Tensorflow: https://www.tensorflow.org
- Documentación de Keras: https://keras.io/

Problema a resolver

La empresa Pontia World carece de un sistema de datos capaz de almacenar y procesar toda la información relativa a sus procesos en el negocio de gestión de parques temáticos. Actualmente utilizan un sistema NoSQL que almacena toda la información en distintos archivos con formato JSON. Sin embargo, hoy por hoy, no llevan a cabo ningún tipo de procesamiento ni análisis sobre ellos, lo que les priva de la posibilidad de tomar decisiones estratégicas basadas en los datos, detectar errores o incidencias que se produzcan en sus sistemas y extraer información y conocimiento que les permita monitorizar y analizar las imágenes.

Ante esta situación, Ponta World ha diseñado un plan de transformación de cómo la empresa debería gestionar sus datos. Aunque la intención es mejorar toda su infraestructura de datos, desean comenzar con las relacionadas con experiencia del usuario en sus atracciones para, posteriormente, evaluar su eficiencia y si realmente representa una mejora notable para el negocio. En caso de obtener resultados positivos, procederían con la transformación del resto de su infraestructura. El plan relacionado con la experiencia de usuario cuenta con los siguientes pasos:

1. Diseñar e implementar un modelo de datos relacional que le permita llevar a cabo análisis fácilmente y extraer métricas y KPIs útiles para el negocio.
2. Identificar patrones de errores e incidencias dentro de sus sistemas gracias a los datos proporcionados.
3. Automatizar la tarea de detección de emociones.
4. Proponer soluciones de IA Generativa para optimizar sus procesos.

Con la primera tarea lo que se pretende es facilitar la tarea a sus analistas, contestar ciertas preguntas de negocio y monitorizar sus sistemas adecuadamente gracias a la extracción de métricas. Esto les permita tomar decisiones basadas en sus datos, así como descubrir áreas de mejora dentro del negocio.

A pesar de que sus sistemas funcionan razonablemente bien y no presenta incidencias críticas, es posible que tanto en el procesamiento como en el almacenamiento de los datos de las

imágenes se cometan errores o se produzcan *bugs* que es conveniente solucionar. Por eso, parte de este plan de transformación pone su foco en identificar y comunicar las incidencias.

Por último, Pontia World quiere competir tecnológicamente con otras empresas de su sector desarrollando sistemas que le permitan automatizar tareas complejas. Por este motivo, desea comenzar con el diseño e implementación de tecnologías capaces de identificar con técnicas de visión artificial las distintas emociones.

Para lograr estos cuatro objetivos, la empresa ha facilitado especificaciones de cada una de las tareas, así como los datos necesarios para desempeñarlas. Aunque, por seguridad, Pontia World prefiera no dar acceso completo a sus datos, ha llevado a cabo una extracción de datos para poder trabajar adecuadamente.

Los datos

La empresa ha extraído de sus sistemas un conjunto de datos del mes de septiembre de 2022. Una misma fotografía del visitante genera varios archivos JSON distintos (cada uno con una información concreta de la fotografía) que, juntos, recogen toda la información relativa a dicha experiencia del cliente. Los campos que aparecen en los JSON son los siguientes:

- **t_id**: identificador de la fotografía
- **tiempo_recogida**: unidad de tiempo (número entero) que representa el momento en que se toma la fotografía contando el número de horas que han pasado desde las 07:00 del 1 de septiembre de 2022. Por ejemplo, si este campo indica un 8 significa que la fotografía se tomó a las 15:00 (07:00 más 8 horas) del 1 de septiembre; mientras que si indica un 25 significa que se tomó a las 08:00 del 2 de septiembre de 2022 (25 horas después del momento de referencia).
- **comienzo_atraccion**: unidad de tiempo (igual que en el campo anterior) del momento en que visitante se sube a la atracción de la que se tomó la foto.
- **atraccion**: atracción en la que el visitante se subió.
- **emocion**: emoción etiquetada en la fotografía
- **id_visitante**: identificador del visitante.
- **tipo_entrada**: tipo de la entrada comprada (familiar, individual, fast-pass, …).
- **coste**: coste de la entrada adquirida.
- **procedencia**: procedencia del visitante.
- **duracion**: tiempo en minutos de su estancia en el parque temático.
- **valoracion**: puntuación que el visitante le ha dado a la atracción.
- **tiempo_de_espera**: tiempo que el visitante ha estado esperando para subirse a la atracción.
- **antelacion_de_compra**: número de días de antelación con la que el visitante compró las entradas (si el valor es 0 significa que la adquirió en taquilla).

Modelo de datos relacional y KPIs

El objetivo de esta tarea es diseñar e implementar un modelo de datos relacional que permita almacenar los datos de las visitas al parque, monitorizarlos y calcular ciertas métricas de negocio que permitan contestar a preguntas de la compañía. Para desarrollar este modelo, se exige utilizar tecnologías SQL. Además, Pontia World quiere auditar el trabajo realizado en este apartado, para lo cual solicita que se le entregue:

- **Esquema relacional**: diagrama de relación de las distintas entidades del modelo junto con sus campos y tipos de datos.
- **Script SQL**: uno o varios archivos que incluyan todas las sentencias SQL utilizadas a lo largo de todo el proceso: creación, modificación y actualización de tablas y vistas, consultas, inserciones en tablas, procesamiento de archivo JSON, …

Además, también le gustaría calcular los siguientes KPIs y contestar a las siguientes preguntas (las sentencias SQL utilizadas para resolverlas debe incluirse en el script mencionado):

- Calcular la media diaria de visitantes.
- Calcular la cuantía total de visitantes.
- ¿Qué días del mes ha habido más visitas y cuántas?
- ¿A qué horas del día sube más gente en la atracción más visitada?
- ¿Cuáles son los 5 visitantes que se han subido en más atracciones y en cuántas?
- ¿Cuáles son los 5 visitantes que se han subido en menos atracciones y en cuántas?
- ¿Cuál ha sido la recaudación total del parque de atracciones?
- Por cada atracción, ¿cuál ha sido la emoción más frecuente?
- ¿Cuál es la media de valoración de cada atracción?
- ¿De dónde son los 3 visitantes que peores valoraciones de media han puesto?
- ¿Cuál es la antelación máxima con la que se adquiere cada tipo de entrada?
- ¿Qué día y hora del mes se producen los tiempos de espera máximos en cada atracción?
- Para cada cliente, calcular el tiempo que no ha estado esperando durante su estancia en el parque.
- El tiempo total de espera de las 3 atracciones mejor valoradas y las 3 peor valoradas.
- De los visitantes que compraron la entrada en taquilla, ¿cuál fue la atracción a la que más se subieron?
- ¿Cuál es la atracción que tiene más número de visitantes con entrada de tipo fast-pass?

Pontia World aceptará y valorará positivamente el planteamiento y resolución de otras preguntas y métricas.

Identificación de errores e incidencias

Para este objetivo, Pontia World quiere identificar los datos erróneos que se encuentren entre los proporcionados (un ejemplo es que el tiempo que un visitante espera en total sea superior a la duración de su estancia en el parque). Con el fin de lograrlo, será necesario aplicar el

conocimiento que se tiene del negocio, así como las reglas de negocio que la empresa nos ha dado:

- La estancia de un visitante no puede superar las 9 horas.
- Las entradas fast-pass se venden como máximo con 3 días de antelación.
- No hay ninguna atracción a la que se puedan subir más de 500 visitantes en una misma hora.
- La valoración es un valor comprendido entre 0 y 10.

Además de estos errores e incidencias que se pueden detectar, resulta necesario identificar aquellos valores nulos que aparezcan y detectar si existen características comunes de los errores, incidencias y valores nulos del mismo tipo.

Al igual que con la tarea anterior, Pontia World exige uno o varios archivos donde se registren las sentencias SQL realizadas.

Reconocimiento de emociones

Por último, se quiere desarrollar un sistema capaz de automatizar la tarea de reconocimiento automático de emociones. Para ello hará falta seguir los siguientes pasos:

1. Acceder con Python a las imágenes y extraer los datos.
2. Llevar a cabo un análisis exploratorio de los datos.
3. Efectuar las tareas de limpieza del dataset necesarias.
4. Si se requiere, utilizar técnicas de data augmentation.
5. Identificar el problema y los objetivos.
6. Fase de selección de datos.
7. Selección del algoritmo.
8. Entrenamiento y parametrización del algoritmo.
9. Evaluación del modelo.
10. Utilizar el modelo de visión artificial a los datos en producción no etiquetados y generar un archivo con las predicciones. Este archivo debe poseer la siguiente información: identificador de la imagen y resultado de la predicción (la emoción detectada).
11. Repetir las fases 5-10 con al menos dos.

Pontia World quiere auditar también estos desarrollos por lo que será necesario entregar uno o varios archivos en los que se registren todas las imágenes analizadas (puede ser un Jupyter Notebook, Google Colab u otro tipo de presentación en el que se pueda ver tanto el código como el resultado de su ejecución). Es preferible, que para poder obtener los mismos resultados si se vuelve a ejecutar, se fije una semilla aleatoria en los distintos algoritmos.

Propuesta de valor con IA Generativa

Además de los problemas que buscan resolver con analítica y ciencia de datos, la empresa Pontia World quiere destinar el proyecto a buscar cómo optimizar y mejorar la eficiencia de

sus procesos internos y su forma de trabajar utilizando tecnologías de IA Generativa. Para ello, se deben identificar posibles casos de uso sobre los que poner en funcionamiento esta tecnología y proponer una arquitectura que aborde estos casos de uso. Para este hito, deben justificarse los casos de uso elegidos con métricas y detallar lo máximo posible la solución técnica propuesta.

Entrega y Evaluación

El formato de entrega de todo el desarrollo del proyecto es libre, mientras cumpla con las condiciones expuestas anteriormente (se incluya todo el código SQL, un esquema relacional, todo el código Python con su salida, la predicción obtenida de los datos no etiquetados y se conteste a las preguntas de negocio). No obstante, para su corrección, también será necesario entregar:

- **Archivo ejecutable**: independientemente del tipo (Jupyter Notebook, Google Colab u otros) debe entregarse un archivo que pueda ejecutarse para comprobar los resultados.
- **Propuesta de IA Generativa**: documento que recoja los casos de uso y la solución técnica planteadas con herramientas de IA Generativa.
- **Informe ejecutivo del proyecto**: documento resumen del proyecto (en él no se debe plasmar todos los detalles ni el código ejecutado) que exponga brevemente (el informe no debe superar las 5 páginas) los siguientes apartados:
    - **Equipo del proyecto y objetivos**: exponer brevemente los integrantes del equipo, las tareas desempeñadas por cada uno y los objetivos planteados para el trabajo.
    - **Modelo relacional**: en él se debe exponer el modelo relacional diseñado e implementado (puede hacerse uso del esquema solicitado).
    - **Limpieza de datos**: breve exposición de los errores encontrados y los pasos a seguir en la limpieza de los datos.
    - **Metodología de ML**: explicar brevemente cada una de las fases de la metodología seguidas.
    - **Comparación de modelos**: comparativa de los modelos de ML desarrollados.
- **Dashboard**: se deben diseñar cuadros de mandos que reflejen los KPIs estudiados y que presenten los resultados con los principales drivers del problema, su análisis y la propuesta de su mejora o solución.

De cara a la presentación del proyecto, se puede hacer uso de las herramientas que se prefiera.