Proyecto Júpiter – Visión artificial

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

Introducción

La empresa distribuidora de alimentos Pontia Logista quiere automatizar sus procesos mecánicos. Esta compañía distribuye toneladas de alimentos diarios entre distintos supermercados. Estos alimentos deben ser correctamente etiquetados. Esta tarea, actualmente la llevan a cabo empleados ya que, en ocasiones, en el proceso de manipulación, puede ocurrir que haya distintos tipos de alimentos mezclados (puede pasar que durante la recolección, por ejemplo, una fruta de un tipo se mezcle con las de otro tipo). Sin embargo, resulta muy costoso e ineficiente necesitar un etiquetado manual de los alimentos existiendo maquinaria capaz de llevar a cabo esta tarea si no existiera la posibilidad de error en el etiquetado que pueden suponer multas muy costosas.

La empresa carece de un sistema adecuado para almacenar y gestionar los datos de sus operaciones, 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 Ponta Logista carece de un sistema de datos capaz de almacenar y procesar toda la información relativa a sus procesos en el negocio de distribución de alimentos. 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 ellas, 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 operaciones.

Ante esta situación, Ponta Logista 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 la distribución de frutas 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 distribución de frutas 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 etiquetado de frutas.

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 permite 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 operaciones 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 Logista 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 los distintos alimentos.

Para lograr estos tres objetivos, la empresa ha facilitado especificaciones de cada una de las tareas, así como los datos necesarios para desempeñarlas. Aunque, por seguridad, Pontia Logista 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 operación de distribución de alimentos genera varios archivos JSON distintos (cada uno con una información concreta de la operación) que, juntos, recogen toda la información relativa a dicha manipulación del alimento. Los campos que aparecen en los JSON son los siguientes:

- **t_id**: identificador de la etiqueta del producto
- **tiempo_recogida**: unidad de tiempo (número entero) que representa el momento en que se produce la recoge el alimento del proveedor agrícola 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 recogida se produjo a las 15:00 (07:00 más 8 horas) del 1 de septiembre; mientras que si indica un 25 significa que se produjo a las 08:00 del 2 de septiembre de 2022 (25 horas después del momento de referencia).
- **tiempo_venta**: unidad de tiempo (igual que en el campo anterior) del momento en que el producto se pone a la venta a disposición del cliente final.
- **coste_inicial**: coste pagado al proveedor agrícola.
- **precio_venta**: precio de venta al supermercado o distribuidor final.
- **tipo**: tipo de fruta.
- **proveedor**: proveedor agrícola.
- **cliente**: supermercado o distribuidora final a la que se vende el producto.
- **peso**: peso en gramos del producto.
- **marca**: marca del producto.
- **lote**: código alfanumérico que identifica el lote en el que se ha distribuido el producto.

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 operaciones de distribución, 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 Logista 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 la cuantía de las distribuciones
- Calcular la cuantía total de las distribuciones
- ¿Qué días del mes se han producido más distribuciones y cuántas?
- ¿A qué horas del día se producen más recogidas de alimentos y cuántas?
- ¿Cuáles son los 5 clientes que más dinero han gastado comprando la fruta y cuánto?
- ¿Cuáles son los 5 clientes que menos dinero han gastado comprando la fruta y cuánto?
- ¿Cuáles son los 10 proveedores que han recibido más dinero y cuánto?
- ¿Cuáles son los 3 productos con mayor beneficio a lo largo del mes (aquellos que al restarle al coste de venta el precio de compra se quedan con un mejor resultado) y cuál ha sido su balance?
- ¿Cuáles son los 3 productos con peor beneficio a lo largo de todo el mes y cuál ha sido?
- ¿Cuál es el precio de venta medio de cada fruta?
- Suponiendo que si no se dispone de información de venta se trata de una fruta que no ha podido venderse por haber sido dañada durante la distribución, ¿cuánta fruta de cada tipo ha sido dañada?
- ¿Cuál ha sido la pérdida total de la fruta dañada?
- ¿Cuál es la cuantía total de cada tipo de fruta que han comprado los 5 clientes que más dinero han gastado?
- Para cada producto, calcular el porcentaje de beneficio.

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

Identificación de errores e incidencias

Para este objetivo, Pontia Logista quiere identificar los datos erróneos que se encuentren entre los proporcionados (un ejemplo es si la fecha de venta a cliente es posterior a la fecha de recogida del proveedor agrícola). 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:

- Como mínimo se tarda un día desde que la fruta se recoge hasta que llega al cliente final.
- A cada proveedor no se le pueden vender en un mismo día más de 100kg de una misma fruta.
- No hay marcas que produzcan más de un tipo de fruta.
- Un mismo lote no puede tener marcas o frutas distintas.

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 Logista exige uno o varios archivos donde se registren las sentencias SQL realizadas.

Etiquetado automático

Por último, se quiere desarrollar un sistema capaz de automatizar la tarea de etiquetado de la fruta. 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 (el tipo de fruta que es).
11. Repetir las fases 5-10 con al menos dos.

Pontia Logista quiere auditar también estos desarrollos por lo que será necesario entregar uno o varios archivos en los que se registren todas las operaciones realizadas y sus salidas (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.

Entrega y Evaluación

Se deberá entregar para su evaluación lo siguiente:

- **Un archivo ejecutable**: Este archivo deberá ser un Jupyter Notebook o Google Colab que se usará para comprobar los resultados. En él deberá aparecer las respuestas a las preguntas de negocio, el código SQL, un esquema relaciones, todo el código de Python con su salida y la predicción obtenida de los datos no etiquetados.

- **Informe ejecutivo del proyecto**: Deberás entregar un resumen del proyecto (en él no se debe plasmar todos los detalles ni el código ejecutado) que se puede realizar en diapositivas que expondrán brevemente este informe no debe superar las 5 páginas y debe tener 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.

De cara a la presentación del proyecto, se puede hacer uso de las herramientas que se prefiera, aunque se recomienda la utilización de una herramienta de BI para presentar los resultados con los principales drivers del problema, su análisis y la propuesta de su mejora o solución.