GUÍA PARA REALIZAR EL TRABAJO FIN DE MASTER

EL EQUIPO TENDRÁ QUE ELEGIR UNA DE LAS OPCIONES SIGUIENTES

1. ANÁLISIS DE UN DATASET (ORIENTACIÓN DATA SCIENTIST)

a. Objetivos:
   i. Analizar un dataset disponible públicamente (Kaggle, UCI Machine Learning Repository, Gapminder u otra fuente que el alumno considere siempre que el conjunto no tenga ningún derecho de uso).

b. Fases:
   i. El estudio y análisis de este dataset deberá de cumplir de forma general las fases de un proceso de modelización analítica estándar, entre las que se encuentran:
      i. Crear un análisis descriptivo del conjunto (gráfico en lo posible).
      ii. Realizar las transformaciones que se consideren más adecuadas o relevantes para el conjunto.
      iii. Crear modelos de predicción utilizando diferentes técnicas de modelización (machine learning) justificando su uso, determinando el nivel de precisión y detallando las bondades, debilidades de cada técnica utilizada.
         a. En este punto, se busca que el alumno proponga un desarrollo que aporte algo más que lo que se podría conseguir con un AutoML.
      iv. Discusión de los resultados del modelo: explicatividad/interpretabilidad.
      v. Realizar un informe final de conclusiones en el que las diferentes fases queden bien delimitadas y en particular donde las mejoras ofrecidas por el modelo queden bien explicitadas y las mejoras futuras que podrían plantear sobre el trabajo realizado.
         a. Este informe final, tendrá una orientación tal que pueda ser entendida por un equipo de “Negocio”.
         b. Podrá incluir elementos técnicos, pero deberá de incluir en mayor proporción detalles que expliquen los resultados del modelo a una persona sin muchos conocimientos técnicos.

vi. Además de las fases anteriormente descritas (propias de la metodología de modelización), se valorará muy positivamente, el que este modelo pueda productivizarse.
    a. Entendemos por este aspecto el que el modelo pueda ser utilizado en un equivalente a una aplicación empresarial. Que al modelo se le puedan pasar nuevos valores y el modelo devuelva una predicción.

c. Extensión:
    i. La extensión total del trabajo no debe superar 20 caras (tamaño folio) sin contar los anexos, ni el índice de contenidos.
    ii. El código asociado y los estudios preliminares se aportarán como anexos. La extensión de estos anexos no cuentan para el tope de 20 caras comentado anteriormente.
    iii. El trabajo se puede realizar por entero en un notebook tipo Jupyter, exportándose a formato HTML:
        i. por favor tened especial cuidado en no generar listados amplios de datos que no aportan valor.
d. Tecnologías:
    i. Lenguajes de programación: Python.
        i. Se valorará la legibilidad del código, el uso de comentarios y un correcto formateado.
    ii. Se recomienda el uso de un notebook: Jupyter
e. Visibildad del trabajo si el conjunto es de Kaggle:
    i. Si el dataset elegido es de Kaggle, el código desarrollado se compartirá como un “Kernel” en el espacio asociados a los datos para este fin, incluyendo que el análisis forma parte de un proceso de evaluación del “Máster – XXXX”.

2. CREACIÓN DE UN PIPELINE DE PREPARACIÓN/DISPONIBILIZACIÓN DE DATOS (PERFIL DATA ENGINEER)

a. Objetivos:
    i. El objetivo consiste en preparar un pipeline, un conjunto de scripts que permitan realizar una/s ETLs (Extraction Transformation Loading) de diferentes fuentes e integrarlas en una base datos que pudiera ser utilizada para realizar un modelo.
    ii. Estas ETLs deberán ser configurables en cuanto a la periodicidad de su ejecución y deberán contar con las soluciones necesarias para monitorizar su progreso/debugging.
b. Tecnologías:
    i. Cualquiera de las estudiadas en el Máster.
    ii. Se puede optar por preparar el pipeline en una tecnología en particular o una combinación de Tecnologías.
c. Documentación:
    i. Se tendrá que documentar la arquitectura técnica elegida:
        i. Sus componentes, sus inter-relaciones y las tecnologías empleadas en cada uno de estos elementos.
    ii. En cuanto al código:
        i. O bien se podrá incluir un repositorio GitHub o referir algún otro repositorio en la nube (Google, Amazon, Azure, etc).
    iii. Además de la solución técnica, la documentación deberá incluir detalles del caso de uso de negocio que solucionaría. Incluyendo referencias a alternativas existentes, diferenciando las mejoras que la propuesta introduce.
d. Extensión:
    i. En cuanto a la extensión de la solución, tampoco se espera que se presente una solución perfectamente disponible para un entorno empresarial, pero sí demostrar que la solución es perfectamente funcional de extremo a extremo.
    ii. Que cumple el objetivo de la captura de diferentes fuentes de datos
    iii. Y que éstos se disponibilizan en una/s tablas listas para ser explotadas: por procesos de modelización, de BI, etc.

3. EL ALUMNO PUEDE PROPONER UN TRABAJO QUE NO ENCAJE EN LAS PROPUESTAS ANTERIORES

a. Objetivos:
   i. El trabajo ha de estar relacionado alguno de los temas impartidos en el curso, pero siempre con una orientación de corte técnico.
   ii. Que el trabajo implique el desarrollo de una solución software y que se pueda encuadrar en el ámbito de la analítica avanzada.

b. Extensión:
   i. La extensión total del trabajo no debe superar 20 caras (tamaño folio), con las mismas consideraciones comentadas en el punto 1 (también en el epígrafe de Extensión).

c. Tecnologías:
   i. Cualquiera de las impartidas en el Máster.

Notas Generales:

- No se admitirán cambios de tema del TFM a menos de quince días para la fecha de entrega.
- El TFM se realizará ver detalles adjuntos (epígrafe de "Realización de los trabajos").
- En el nombre del fichero se incluirá el nombre del alumno o del grupo (Nombre y dos apellidos), separando el nombre y los apellidos con un guión bajo ("_"):
  - Ejemplo: Maria_Garcia_Perez_Estudio_pajaros.zip
  - Grupo_A_Estudio_mariposas.zip
- Los tutores a cargo de mentorizar y corregir los trabajos serán **Carlos Ortega** y **Santiago Mota**.
  - Los tutores pueden ayudar en sugerir una orientación adecuada a una propuesta de trabajo, pero se evitará el enviar diferentes versiones del trabajo para confirmar si el enfoque o el nivel de avance, es el correcto.

Realización de los trabajos:

- Los trabajos se realizarán en modalidad:
  - Grupal (grupos de 5 o 6 personas).

Sobre el informe del TFM, a modo de resumen la estructura del entregable sería:

- Documento (el de las 20 caras) que contiene:
  i. el detalle del trabajo expuesto de una forma (a poder ser no muy técnica). Que incluye tablas resumen, uso de bullets para enumerar ideas, etc.

ii. En el texto se incluyen referencias a diferentes partes del Anexo donde se dan detalles más profundos de la idea expuesta.

- Como Anexo se puede incluir:

iii. el código desarrollado

iv. Estudio más detallados de por ejemplo el EDA (si se escogió la opción 1), o de la ejecución de diferentes modelos (si se escogió la opción 1).