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 y justifiquen 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, ni por su puesto la portada o la contraportada.
        i. El tamaño de letra y el interlineado se deja a decisión del alumno, pero primando el sentido común y la legibilidad del documento (documentos a tamaño de letra 8 ó 9 o de 20 no tienen mucho sentido, el preferido sería de tamaño 10 u 11).
        ii. Sobre el tipo de letra, recomendamos Verdana o Arial.
    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. Tampoco contarán ni la portada ni el índice de contenidos.
    iii. El trabajo se puede realizar por entero en un notebook tipo Jupyter exportándose a formato HTML.
    En estos casos:
        i. por favor tened especial cuidado en no generar listados amplios de datos que no aportan valor.
    iv. Si el trabajo se realiza en el espacio de Colab, igualmente se ha de exportar el resultado a un .html para su correcta lectura.
        i. En este caso, se puede adjuntar un link dentro del informe de conclusiones con la url utilizada de Colab.

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. Visibilidad del trabajo si el conjunto es de Kaggle:
    i. Si el dataset elegido es de Kaggle, se recomienda (por mejorar la marca persona) compartir el código desarrollado como un “notebook” 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 une 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. LOS ALUMNOS PUEDEN PROPONER UN TRABAJO QUE NO ENCAJE EN LAS PROPUESTAS ANTERIORES

a. Objetivos:
    i. El objetivo del trabajo ha de ser primeramente comentado con los tutores para su discusión/aprobación.
    ii. El trabajo ha de estar relacionado alguno de los temas impartidos en el curso, pero siempre con una orientación de corte técnico.
    iii. 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 grupo:
    - 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.

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

i. 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.
ii. Como Anexo se puede incluir:
    i. El código desarrollado
    ii. Estudio más detallados de por ejemplo el EDA, o de la ejecución de diferentes modelos.

¿QUÉ HACER SI VOY A USAR DATOS DE UNA EMPRESA?

En este caso, lo que sugerimos en lo siguiente:

- Asegurarse con mucho margen de antelación que contáis con el visto bueno de vuestra empresa. Hemos visto que en el último momento, poco antes de la entrega las empresas se echan atrás.
- Muy probablemente vuestra empresa os pida que se firme un NDA (Non Disclosure Agreement) para evitar que la propiedad intelectual se filtre a terceros.
- A la hora de crear este NDA, por favor comunicad a vuestra empresa (departamento legal) las siguientes limitaciones.

    1. Es responsabilidad del alumno cerciorarse de que tiene licencia para utilizar los datos del proyecto. En caso de ser necesario, porque así lo demanden los propietarios legales de los datos, el máster proporcionará una plantilla de Acuerdo de Confidencialidad (NDA) que firmarán, a título personal, Carlos Ortega y Santiago Mota.
    2. La autoría y propiedad de los TFM será de los alumnos, las únicas personas que accederán al material entregado serán los citados profesores, con objeto de calificar los trabajos y, una vez cerradas las actas, se procederá a la eliminación del material.
    3. No se firmarán acuerdos en los 30 días previos a la entrega del TFM, por lo que el alumno(s) debe(n) gestionar el proceso antes de llegar a esa fecha. En caso de que no se llegará a un acuerdo con los propietarios de los datos, será responsabilidad del alumno(s) presentar un proyecto alternativo.

CHECKLIST:

A modo de lista de comprobación de elementos importantes, se recomienda considerar lo siguiente:

- ¿Has mirado los derechos de uso de los datos?
- ¿Tienes el código compartido en un Github o en un Drive?
    - ¿Es accesible desde el link?
    - ¿Santiago Mota y Carlos Ortega tienen permisos de acceso?
- ¿La memoria ocupa 20 hojas?
- ¿Tienes el código en los Anexos?
- ¿El proyecto es reproducible?
- ¿Has incluido un apartado de conclusiones?
- ¿Has incluido una breve lista (media cara) con la bibliografía y/o referencias?.

PREGUNTAS FRECUENTES:

- ¿Si mi conjunto de datos es de 1000 - 2000 filas, es suficiente?
    - Valoramos el uso de conjuntos grandes. Los conjuntos grandes suponen retos de procesamiento muy próximos a los que nos enfrentamos en entornos empresariales.
    - Si el conjunto es limitado, no impide hacer un TFM, pero se valorará menos que el uso de un conjunto grande.
    - De conjuntos de datos de 300-400 filas, es muy complicado poder realizar un TFM que no difiera de un trabajo de fin de Módulo.
- ¿Puedo hacerlo en inglés?
    - Sí, el TFM se puede hacer en inglés.
- ¿Puedo presentarlo en PowerPoint?
    - No. Pensamos que el TFM ha de presentarse en forma de memoria técnica con su redacción de forma equivalente a un informe. Este enfoque es mucho más complicado al usarse un PowerPoint donde se prima los mensajes más escueto.
- ¿Se cuenta la portada en la extensión?
    - No, ni la contraportada, ni el índice de contenidos.
- ¿Tengo que poner bibliografía?
    - Sí, pero de forma escueta. Que no ocupe más allá de media página.
- ¿En un HTML cómo veo que sean 20 páginas?
    - Puedes exportar el HTML a pdf y contar las páginas.,
    - Otra alternativa es contar el número de pantallas consecutivas que ocupa tu HTML (sobre un monitor de 13-14 pulgadas).
- Mi TFM es de un conjunto de datos de Kaggle que tiene ya mucho código desarrollado por otras personas, ¿cómo se califica el TFM en este caso?
    - En estos casos, sugerimos cambiar de conjunto de datos.
    - Kaggle contiene conjuntos de datos muy orientados a la educación, práctica. Hacer un TFM de estos casos, no difieren de hacer un trabajo de fin de módulo.
- La empresa con la que estoy haciendo el TFM, solo me pide que haga un cuadro de mando, ¿es esto suficiente para el TFM?.
    - No.
    - Se sugiere ofrecer a la empresa llegar a presentar un modelo relacionado con el caso propuesto y aunque esta vía no se acepte, en el TFM todos los elementos que complementen lo solicitado por la empresa se valorarán.
    - No incluir aspectos de modelización o de otros aspectos desarrollados durante el Máster (por ejemplo productivización) hace que el TFM sea limitado.
- Tengo diferentes dudas, ¿se puede mantener una reunión/call con los tutores para resolverlas?

    - Por experiencias previas, preferimos que las dudas se trasladen o bien el foro de la plataforma o en los correos personales (siempre con copia a los Gestores) de forma escrita.
    - Hemos visto que el hecho de expresar las dudas por escrito sirve mucho para aclarar el alcance de la duda y precisar mucho más lo que se necesita.
    - Otro punto a tener en cuenta es que las dudas no pueden ser de tipo técnico sobre aspectos particulares del enfoque del TFM, salvo que exista un bloqueo que impida avanzar.
    - Tampoco pueden ser sobre errores que impiden el avance. Los errores en la instalación de librerías, o en la ejecución forma parte del día a día de alguien que analice datos. El alumno tiene que ser capaz de poder gestionar estas situaciones de forma autónoma consultando foros, Google, etc.
- ¿Se puede disponer de TFMs pasados para ver la estructura seguida?.
    - Lo hicimos en el pasado y no fue bien. Cada TFM tiene su enfoque previo y la presentación válida para un TFM en particular puede no ser válida para otro.
    - Por tanto, no se proporcionarán TFMs de referencia.