Proyecto de Datos 1

Entrega 1: Definición del proyecto

El objetivo de esta entrega es definir el proyecto en el que vais a trabajar durante todo el curso. En el proyecto vais a desarrollar un sistema de aprendizaje automático que deberá ser útil para una empresa u organización específica.

El sistema debe abordar un problema claramente definido que ayude a la empresa a cumplir sus objetivos. El sistema realizará una tarea de aprendizaje automático usando datos reales obtenidos de fuentes públicas o datos web de empresas o redes sociales accesibles mediante *web scraping*. En esta entrega debéis aterrizar esta idea de forma realista.

Para ello debéis abordar los siguientes puntos.

1. Contorno del proyecto

1.1 Organización destinataria

Describe la empresa u organización (real o ficticia) que vaya a usar el sistema de aprendizaje automático que vais a desarrollar. Especifica en qué sector opera y a qué se dedica.

1.2 Objetivo de negocio de la organización

Define el objetivo u objetivos de negocio que la empresa u organización busca lograr con la implementación del sistema. Indica por qué son importantes para la empresa.

1.3 Objetivo del sistema de aprendizaje automático

Define el objetivo u objetivos del sistema de aprendizaje automático y su relación con los objetivos del apartado anterior.

1.4 Uso del sistema de aprendizaje automático

Indica cómo se va a usar el sistema dentro de la organización, quién va a usarlo y para qué.

1.5 Requisitos de rendimiento

Identifica un volumen aproximado de uso del sistema (predicciones por unidad de tiempo) e identifica si existen momentos del año donde puede haber picos y valles o eventos que puedan ocasionar cambios abruptos en el volumen.

Identifica también si el tiempo que se tarda en generar la predicción puede ser un requisito a tener en cuenta.

1.6 Requisitos legales

Identifica posibles restricciones con las que el proyecto puede contar, ya sean legales (por ejemplo, relacionado con privacidad o datos sensibles) o problemas de sesgos sociales que el sistema pueda ayudar a consolidar.

2. Tarea de aprendizaje automático

2.1 Definición de la tarea de aprendizaje automático

Para la tarea de aprendizaje automático debes elegir entre regresión, clasificación binaria, clasificación multiclase o clasificación multietiqueta.

Si el problema puede resolverse de varias maneras, menciona las ventajas e inconvenientes que ves a cada una e indica si hay alguna que ves más o menos adecuada que las otras.

Explica por qué la tarea seleccionada es útil para el logro del objetivo del sistema y para los objetivos de la compañía.

2.2 Definir la variable objetivo

Identifica la variable objetivo de tu sistema de aprendizaje automático, es decir, la variable que vas a predecir. Indica qué valores puede tomar, incluyendo sus unidades si las tuviera, y reflexiona sobre su distribución (p.ej. ¿Hay valores extremos? ¿Hay clases dominantes o infrarrepresentadas?).

Además, debes definir cómo vas a obtener los datos de dicha variable.

2.3 Identificar los datos de entrada

En un sistema de aprendizaje automático los datos de entrada son aquellos que usa el sistema para generar la salida. Debes identificar los datos de entrada que ves factible que use tu sistema, es decir, qué variables usará o con qué datos contará.

2.4 Fuente de datos principal

Indica cómo planeas obtener datos reales para tu proyecto. Puedes utilizar APIs de redes sociales, webs públicas, aplicaciones diversas o realizar *web scraping* de sitios web específicos.

Debes indicar si la fuente de datos te permite obtener la información necesaria que indicaste en los dos apartados anteriores y si no es así, qué alternativas tienes.

También debes indicar las posibles restricciones o problemas que podáis tener para acceder a los datos. Por ejemplo, algunas APIs son de pago o tienen una limitación de uso.

Por ello, en este apartado debes presentar al menos una **demo de código** que muestre que habéis probado a extraer los datos y un **fichero de datos** que muestre que habéis logrado extraer datos. El objetivo es asegurar la viabilidad del proyecto en lo referente a los datos.

2.5 Fuentes de datos secundarias

Puede darse el caso de que existan otras fuentes de datos que complementen los datos anteriores y que puedan servir para enriquecer vuestro sistema de aprendizaje automático.

En este punto del proyecto, debes identificar dichas fuentes de datos e indicar por qué consideres que pueden contribuir a mejorar el sistema y cómo se integrarían los datos en el sistema. Opcionalmente, puedes presentar también la demo y el fichero de datos con datos extraídos de la fuente o fuentes secundarias.

2.6 Factores que afecten al rendimiento

Identifica qué factores pueden obligar a reentrenar el sistema porque se degrade su rendimiento o porque tenga que adaptarse a cambios en el entorno.

3. Evaluación del Sistema

3.1 Evaluación de éxito del sistema

Indica la medida cuantitativa que va a usar la organización para validar si el sistema cumple su objetivo o no. Podría haber más de una y algunas pueden ser indirectas, es decir, no derivadas directamente del éxito o el fracaso de la predicción que da el sistema.

Determina también si existe una alternativa en la organización contra la que comparar el sistema y si se os ocurre alguna heurística que pueda realizar la tarea de forma aproximada y que también sirva para comparar.

Explica cómo se llevaría a cabo la evaluación y las comparaciones con las posibles alternativas.

3.2 Evaluación de éxito de la tarea de aprendizaje automático

Indica con qué medida o medidas tiene sentido evaluar el rendimiento de la tarea de aprendizaje automático, tanto durante el entrenamiento como durante el uso del mismo.

Indica cómo se relacionan las medidas de éxito de la tarea de aprendizaje automático y la del sistema, y el impacto de ésta en los objetivos de negocio de la organización.

Entrega

Prepara un informe de 5 páginas máximo (fuente a tamaño 11 e interlineado sencillo) en la que respondas a lo que se indica en las secciones anteriores. Debes adjuntar también el código para recuperar los datos y una pequeña muestra con los datos recuperados.

La entrega se realiza a través del campus virtual y entregará únicamente un miembro.

Fecha límite de entrega: 5 de febrero 23:55