ACTIVIDAD 2.3 – Ejercicio Teórico
Anteproyecto de minería de datos web

Imagina que decides montar un sistema de minería de datos web basado en un motor predictivo de Machine Learning.

Este es un ejercicio teórico. Explica en una **extensión máxima de 1 página** (aproximado) tu idea de proyecto, intentando contestar en la medida de lo posible a cuantas puedas de las siguientes preguntas. No es necesario que escribas nada de código, pero si lo ves necesario puedes comentar algunas de las tecnologías que usarías también.

Nombre Alumno: Pablo Vázquez Fernández

Otros Alumnos Grupo (si hay): Andrea del Vado Puell , Joaquín Ángel Tejero Cañero

- ¿Cuál es el objetivo de tu proyecto?
- ¿Qué tipo de minería web usarás: de contenido, de utilización, de estructura?
- ¿Cuáles serían tus fuentes de datos? ¿Qué web o conjunto de webs usarías para extraer datos?
    - ¿Usarías una web de un dominio propio o de tu empresa, o usarías webs externas?
    - ¿Qué tipo de datos quieres minar: numéricos, textos, imágenes?
    - Señala la web concreta y pon ejemplos concretos de datos en esa web querrías minar
    - ¿Prevés algún problema en la adquisición de los datos?
- ¿Con qué frecuencia tendrías que hacer el minado de datos?
    - ¿Qué cantidad de datos esperas manejar para crear tu dataset de entrenamiento?
    - ¿Y una vez esté el sistema en funcionamiento?
- ¿Qué tipo de sistema predictivo (de Machine Learning) propones?
    - ¿Qué cosa te interesa predecir?
    - ¿Vas a usar sistemas de clustering, clasificación, regresión, análisis de textos?
    - ¿Cuáles crees que son las métricas adecuadas para validar tu sistema de minería web?
- Más allá de este ejercicio teórico, ¿cómo de viable crees que sería tu propuesta de sistema en la vida real?

Propuesta de Proyecto

El **objetivo** principal de este proyecto es desarrollar un sistema de minería de datos web basado en *Machine Learning* para predecir tendencias de precios de productos electrónicos en diferentes plataformas de comercio electrónico. Se busca proporcionar a los usuarios información predictiva sobre posibles cambios en los precios de productos específicos para ayudar en la toma de decisiones de compra.

En cuando a la **minería web**, se opta por contenido y estructura. La extracción de datos se centrará en la información de precios, características de productos y reseñas de usuarios.

Hablando de **fuentes de datos**, se utilizarán diversas plataformas de comercio electrónico como Amazon o eBay para extraer datos sobre precios, características de productos, reseñas y tendencias de ventas.

Relacionado con los **tipos de datos a minar**, principalmente serán datos numéricos como los precios, datos de texto como las descripciones de productos o las reseñas y es posible que imágenes, en el caso de ser relevante para la predicción de precios.

Algunos **problemas o desafíos** podrían aparecer debido a restricciones de acceso, políticas de *scraping* de las plataformas o cambios en la estructura de datos de las páginas web, lo que requeriría un monitoreo constante y ajustes en los algoritmos de extracción.

Para obtener datos lo más actualizados posibles, la **frecuencia de minado** será diaria. Como se esperaría manejar un gran volumen de datos, se creará un conjunto de entrenamiento inicial que se seguirá actualizando una vez que el sistema esté en funcionamiento.

En cuando al **tipo de sistema predictivo**, se propone un modelo de regresión para predecir los posibles cambios en los precios de los productos en función de diferentes variables como: la demanda, la competencia y las características del producto. También se podría considerar un análisis de texto para comprender las reseñas de los usuarios y su impacto en los precios.

Para **validar el sistema**, se considerarían métricas como el error cuadrático medio (RMSE) que evaluará la precisión de las predicciones de precios. Además, métricas de clasificación podrían utilizarse para ver la relevancia de las características de los productos en las predicciones.

La **viabilidad** en la vida real dependerá, en gran medida, de la capacidad para acceder a los datos de las plataformas de comercio electrónico de manera ética y legal, así como de la precisión y capacidad predictiva del modelo desarrollado. Sería muy importante tener en cuenta los aspectos éticos y legales de la extracción de datos web. Además, serán necesarios gran cantidad de recursos para mantener el sistema actualizado y adaptado a cambios en las plataformas web.