Máster de FP en Inteligencia Artifical y Big Data

Ejercicio de clasificación: detección de spam

Realiza un proyecto de machine learning enfocado en la detección de mensajes de spam siguiendo las siguientes pautas:

- El clasificador se puede realizar para correos electrónicos, para sms o para mensajes directos de cualquier plataforma. También se pueden considerar spam los posts falsos, es decir, los creados por bots, por tanto, realizar un clasificador para detectar este tipo de texto también será válido.
- El dataset debe contener suficientes muestras de mensajes como para que los resultados sean fiables, a ser posible, más de 500 registros. Si son varios miles de registros, mejor.
- Para poder utilizar algoritmos enfocados en aprendizaje automático supervisado, el dataset debe estar etiquetado, es decir, uno de los atributos debe indicar si el mensaje correspondiente es spam o no.
- Se darán los pasos oportunos, al igual que con otros proyectos realizados en clase: obtención de datos, descripción de los datos, exploración y visualización de los datos, exposición del objetivo, preparación de los datos para los algoritmos de machine learning, entrenamiento del modelo, comprobación del rendimiento y exposición de conclusiones.
- Se pueden seguir guías o tutoriales siempre que se haga de forma crítica y consciente de lo que se hace en cada momento; modificando, resumiendo o ampliando el código en base a decisiones justificadas.
- Es muy importante indicar los enlaces de las fuentes, tanto del dataset como de la documentación consultada.
- Hay datasets muy “masticados” que están casi listos para entrenar el modelo. Son válidos si el trabajo cumple todos los puntos que se piden. Pero es aconsejable, si se quiere sacar buena nota, partir de los mensajes originales y

realizar todo el proceso de tokenización, eliminación de signos de puntuación, eliminación de preposiciones y otras partículas, etc.

- El proyecto se debe realizar en grupos de dos alumnos/as y se expondrá en clase en las fechas indicadas.
- Durante la exposición, el profesor puede hacer preguntas sobre líneas concretas de código o sobre aspectos generales del ejercicio.
- No se permite la utilización de redes neuronales (se verán en el segundo trimestre).