Máster Universitario en Big Data – Minería de datos

Práctica final: Clasificación de embarcaciones según su sonido

Resumen

El proyecto DeuteroNoise tiene como objetivo principal abordar los desafíos del ruido submarino generado por actividades humanas, especialmente las relacionadas con el tráfico marítimo. Este proyecto busca desarrollar herramientas innovadoras para la monitorización, clasificación y mitigación del ruido submarino, con el propósito de preservar la biodiversidad marina y facilitar estudios sobre su impacto ambiental.

En este contexto, una de las tareas fundamentales que quiere abordar el proyecto es la clasificación de los ruidos generados por diferentes tipos de embarcaciones. Cada tipo de embarcación, como barcos de carga, petroleros o barcos de pasajeros, emite sonidos diferentes debido a factores como el tamaño, el diseño del casco y las características del motor. Identificar estos sonidos permite evaluar zonas de alto impacto acústico y proponer estrategias para minimizar la contaminación sonora en hábitats sensibles.

Para esta práctica, se cuenta con un conjunto de datos que incluye grabaciones submarinas obtenidas mediante hidrófonos instalados en ubicaciones estratégicas.

Objetivos

Analizar un problema del mundo real y proponer una solución de Machine Learning *end-to-end*. En este sentido, debéis diseñar un pipeline para resolver el problema y entender el *workflow* tradicional de los proyectos de Machine Learning. Específicamente, tendréis que realizar los siguientes puntos:

1. Definir vuestro pipeline y los diferentes módulos que lo van a componer.
2. Explorar los datos y, si es necesario, incrementar el corpus.
3. Aplicar técnicas de *feature engineering.*
4. Aplicar distintos modelos y reportar los resultados.
5. Aplicar técnicas de validación apropiadas.

Entregable y grupos

Para poder aprobar la práctica, tendréis que entregarla en el eStudy. El entregable debe constar de un documento con (como mínimo) el siguiente contenido:

1) Introducción al problema.
2) Explicación del pipeline utilizado.
3) Análisis exploratorio y *features* escogidas.

4) Explicación de los distintos modelos de Machine Learning que habéis utilizado.
5) Explicación de los resultados obtenidos.
6) Conclusiones.

Además del documento PDF, deberéis entregar un script de Python o un notebook con vuestro código. Podéis hacer la práctica en grupos de un máximo de **dos personas**. Los grupos deben notificarse por correo a la profesora en la dirección ester.vidana@salle.url.edu antes del día 16 de Febrero. Si no mandáis el correo a tiempo, se entenderá que realizáis el trabajo de manera individual y, por lo tanto, no lo podréis entregar en pareja.

En cuanto hayáis hecho vuestra entrega, tendréis que realizar una entrevista en pareja con la profesora para explicar vuestro trabajo. En dicha entrevista, se evaluarán los modelos de Machine Learning entrenados con un conjunto de datos de Test que se os proporcionará durante la misma entrevista. El formato de los datos de Test será igual que el formato proporcionado en el conjunto de datos de entrenamiento. El pipeline entregado, por lo tanto, debe ser capaz de (1) leer audios nuevos, (2) sacar sus características y (3) obtener resultados de clasificación.

Consideraciones

Como modelos de Machine Learning, como mínimo debéis utilizar:

- Un algoritmo de Machine Learning.
- Un *ensemble*.
- Un algoritmo de Deep Learning.

Y en cuanto a librerías, podéis utilizar las vistas en clase u otras bajo la aprobación de la profesora. Para el algoritmo de Deep Learning, debéis utilizar Tensorflow o Pytorch.

Dataset

Podréis partir de dos conjuntos de datos: o bien de un dataset completo de unos 20 GB, o bien de una versión reducida del mismo de menos de 1 GB. El segundo es un subconjunto de datos del primero, para que la práctica se pueda manejar mejor con infraestructuras de computación más modestas.

Además de este dataset, si queréis, podéis utilizar datos externos bajo la aprobación de la profesora, y tendréis que incluir en el informe de dónde los habéis obtenido.

Métricas de evaluación

Podéis utilizar las métricas que creáis convenientes, pero como mínimo tenéis que mostrar una matriz de confusión y la F1-Score de vuestro sistema.

Formato y fecha de entrega

Para considerar la práctica como entregada, es necesario que cumpláis con dos requerimientos:

1- Subir la práctica al pozo del eStudy. Ambos miembros del grupo tienen que subir la entrega en el eStudy.
2- Superar la entrevista con la profesora. Solamente se podrá peticionar un horario para hacer la entrevista cuando se haya subido el contenido de la práctica en el pozo. La hora de entrevista se acordará entre la profesora y el grupo de prácticas por correo electrónico. Si un grupo no se presenta a la hora acordada de entrevista o llega tarde, el grupo dispondrá de una segunda oportunidad para presentar la práctica en convocatoria ordinaria. Si en la segunda oportunidad el grupo vuelve a no presentarse, entonces la calificación del grupo será de No Presentado (NP) en ordinaria y deberá presentar la práctica en convocatoria extraordinaria, donde dispondrá de una última oportunidad.

La fecha máxima para depositar la práctica al pozo del eStudy para poder tener nota en convocatoria ordinaria es el 27 de Abril. El grupo dispondrá entonces hasta el 30 de Abril para acordar una fecha de entrevista con la profesora, que deberá realizarse como máximo el día 16 de Mayo.

Si un grupo termina la práctica antes de tiempo, puede subir la práctica al eStudy y peticionar la entrevista antes de las fechas establecidas.

Recordad que todas las actividades de evaluación de esta asignatura se consideran actividades altamente significativas según la normativa académica (https://www.salleurl.edu/es/normativa-de-copias). Por lo tanto, las copias totales o parciales en cualquier actividad evaluable, se penalizarán con el que está establecido en la normativa académica, tanto en la fuente como la copia sin excepción.