Clasificación y Regresión

Objetivos:

Afianzar los conceptos introductorios respecto los clasificadores y regresores supervisados, mediante su implementación práctica. Identificar el desempeño de los modelos de acuerdo al tipo de *dataset*, interpretar las métricas y resultados obtenidos. Aplicar los conceptos de descriptores y definirlos.

Promover el interés respecto a la utilidad de los modelos de clasificación y regresión, paramétricos y no paramétricos, mediante aplicaciones prácticas reales.

Recomendaciones para la resolución del trabajo:

Evitar copiar y/o modificar soluciones de pares (compañeros, sitios de Internet, etc.), en lugar de ello, esforzarse por elaborar una producción original propia a partir del análisis y reflexión de cada una de las consignas, teniendo a mano la teoría provista en clases, la bibliografía ofrecida y todo otro material complementario que juzgue necesario para enriquecer su producción.

Reflexionar sobre los conceptos o justificaciones que se ofrecen como solución a la consigna presentada. Es decir, pueden intercambiarse opiniones, debates o puestas en común respecto a un determinado punto, pero la producción entregada debe basarse en su concepción personal del marco teórico, experiencias generales e interpretación de las consignas.

Producción esperada para acreditar la actividad:

Presentar un informe de estilo monográfico con formato libre en la tarea designada en el aula virtual del curso, incluyendo el contenido solicitado en cada punto de la guía. Realizarlo en tiempo y forma, dentro del plazo máximo de una semana desde la disponibilidad del presente documento.

Además del informe, adjuntar los códigos utilizados para llevar a cabo las experiencias. Los resultados presentados, deben poder ser replicables.

Priorizar la calidad por sobre la cantidad, cuidando la prolijidad general en la confección, incluyendo una portada debidamente identificatoria del trabajo.

Consignas:

Ejercicio 1

Se tiene un *dataset* con información referente a la lectura de gases realizada por una matriz de 6 sensores de bajo costo en instantes sucesivos de tiempo. A estas lecturas, se asocia la información de la actividad realizada en el recinto sensado.

El conjunto de sensores utilizados se puede agrupar en dos categorías principales:

- Sensores MQ (MQ2, MQ9, MQ135, MQ137, MQ138) que tienen gran sensibilidad, baja latencia y bajo costo; cada sensor puede responder a diferentes gases;
- Sensor analógico de gas CO2 (MG-811) que tiene una excelente sensibilidad al dióxido de carbono y apenas se ve afectado por la temperatura y la humedad del aire.

El conjunto de datos contiene 1845 muestras recolectadas que describen 4 situaciones objetivo:

1. **Situación normal** - Actividad: aire limpio, una persona que duerme, estudia o descansa. Muestras disponibles: 595.
2. **Preparación de comidas** - Actividades: cocinar carne o pasta, verduras fritas. Una o dos personas en la habitación, circulación de aire forzado. Muestras disponibles: 515.
3. **Presencia de humo** - Actividad: quemar papel y madera por un corto período de tiempo en una habitación con ventanas y puertas cerradas – Muestras disponibles: 195.
4. **Limpieza** - Actividad: uso de detergentes en aerosol y líquidos con amoníaco y/o alcohol. La circulación de aire forzado se puede activar o desactivar - Muestras disponibles: 540.

Cada muestra está compuesta por 7 valores; los primeros seis valores son las salidas de los sensores, mientras que el último es el índice de la acción que generó los valores adquiridos por los mismos. Las cuatro situaciones diferentes están asociadas con una composición del aire distinta, teniendo en cuenta que cualquier actividad produce sustancias químicas (respiración humana, exhalaciones de procesos metabólicos, liberación de volátiles por combustión y/o oxidación, evaporación de detergentes domésticos, etc.).

Los datos se encuentran en el archivo adjunto denominado "*dataset_ADL_clasificacion.csv*". A continuación, se presenta una porción de este *dataset*:

MQ2	MQ9	MQ135	MQ137	MQ138	MG-811	Situación
670	696	1252	1720	1321	2431	4
641	674	1156	1652	1410	2433	1
642	646	1159	1643	1455	2361	3
640	590	1105	1608	1459	2427	4
616	627	1192	1637	1466	2447	2

Tabla 1. Fragmento del dataset, 1, calidad del aire.

En la columna "*Situación*" se encuentra codificada la clase que indica si se trata de una **actividad normal** (**1**), **preparación de comidas** (**2**), **presencia de humo** (**3**) o de **limpieza** (**4**).

El objetivo es determinar si con estas observaciones es posible obtener un modelo que permita ***clasificar*** cada una de las 4 situaciones objetivo, a partir de los valores arrojados por los sensores.

Detallar y fundamentar cada aspecto de la solución propuesta, incluyendo los aspectos que crea conveniente y respondiendo como mínimo las siguientes premisas:

- ¿Qué clasificador o meta clasificador se ajusta mejor a la solución buscada? ¿Por qué?
- En base a los resultados obtenidos, ¿Sería posible utilizar el modelo para predecir el comportamiento de nuevas mediciones?
- Evaluar aspectos relacionados al rendimiento en el proceso de entrenamiento y vincularlos con la relación costo/beneficio.
- Ídem para los tiempos de inferencia o consulta, ¿Cuál es fundamental si pretendo utilizar un sistema en tiempo real con el modelo desarrollado?
- De los resultados visualizados en la matriz de confusión, ¿Para todas las clases el modelo seleccionado presenta el mismo comportamiento?
- ¿Existen problemas evidentes en este *dataset*? Si es así, ¿Cuáles son y cómo los solucionaría?
- ¿Son necesarios todos los sensores para que el modelo pueda identificar de manera eficiente las clases de interés?

Ejercicio 2

La resistencia a la compresión simple es la característica mecánica principal del concreto. Se define como la capacidad para soportar una carga por unidad de área, y se expresa en términos de esfuerzo [*MPa*]. El cemento es el material más activo de la mezcla de concreto, por tanto, sus características y sobre todo su contenido (proporción) dentro de la mezcla tienen una gran influencia en la resistencia del concreto, a cualquier edad. A mayor contenido de cemento se puede obtener una mayor resistencia y a menor contenido, la resistencia del concreto va a ser menor. Además, existen varios "agregados" que afectan de manera no lineal las proporciones de los materiales para una resistencia determinada.

Se tiene un *dataset* con información referente a factores que, según los especialistas, se vinculan con la resistencia a la compresión del hormigón. La información de estas variables, su tipo y unidades, se especifican de la siguiente manera:

- ***Cement*** *(component 1) -- quantitative -- kg in a m3 mixture -- Input Variable*
- ***Blast Furnace Slag*** *(component 2) -- quantitative -- kg in a m3 mixture -- Input Variable*
- ***Fly Ash*** *(component 3) -- quantitative -- kg in a m3 mixture -- Input Variable*
- ***Water*** *(component 4) -- quantitative -- kg in a m3 mixture -- Input Variable*
- ***Superplasticizer*** *(component 5) -- quantitative -- kg in a m3 mixture -- Input Variable*
- ***Coarse Aggregate*** *(component 6) -- quantitative -- kg in a m3 mixture -- Input Variable*
- ***Fine Aggregate*** *(component 7) -- quantitative -- kg in a m3 mixture -- Input Variable*
- ***Age*** *-- quantitative -- Day (1~365) -- Input Variable*
- ***Concrete compressive strength*** *-- quantitative -- MPa -- Output Variable*

Los datos se encuentran en el archivo adjunto denominado "*dataset_hormigon_regresion.csv*". A continuación, se presenta una porción de este *dataset*:

cement	slag	flyash	water	superplasticizer	coarseaggregate	fineaggregate	age	csMPa
540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	79.99
540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	61.89
332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	40.27
332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	41.05
198.6	132.4	0.0	192.0	0.0	978.4	825.5	360	44.30

Tabla 2. Fragmento del dataset, 2, resistencia del concreto.

En la columna "*csMPA*" se encuentran los valores numéricos correspondientes a la resistencia a la compresión del hormigón, obtenida para la combinación de los demás materiales en las cantidades indicadas en la fila correspondiente.

El objetivo es determinar si con estas observaciones es posible obtener un modelo que permita ***predecir*** el valor de resistencia a partir de los componentes y proporciones utilizados en la elaboración del hormigón.

Detallar y fundamentar cada aspecto de la solución propuesta, incluyendo los aspectos que crea conveniente y respondiendo como mínimo las siguientes premisas:

- ¿Qué regresor se ajusta mejor a la solución buscada? ¿Por qué?
- En base a los resultados obtenidos, ¿Sería posible utilizar su modelo para predecir la resistencia que tendrá el concreto con una "receta" diferente a las utilizadas?
- ¿Qué métricas utiliza para evaluar el desempeño de un modelo frente a un problema de este tipo? ¿Podría utilizar las mismas métricas empleadas para el modelo del ejercicio 1?
- ¿Existen problemas evidentes en este *dataset*? Si es así, ¿Cuáles son y cómo los solucionaría?
- El preprocesamiento de los datos, ¿Afecta de alguna manera el desempeño del modelo?
- Un preprocesamiento incorrecto de los datos, ¿Qué inconvenientes cree que podría generar?

Anexos:

Saberes vinculados a la actividad:

Saberes conocer	Saberes hacer	Saberes ser
IA en Ingeniería (intra curso): - *Data shapes* - Validación cruzada - *Leave-out* - Clasificadores paramétricos - Clasificadores no paramétricos - Regresores paramétricos - Regresores no paramétricos - Matriz de confusión - Extracción de *features* Estadística (conceptos previos): - Ajuste de curva - Probabilidad - Incertidumbre Programación (conceptos previos): - Manipulación de archivos - Procesamiento de datos Sistemas (conceptos previos): - Función de transferencia	IA en Ingeniería (intra curso): - Identificar clasificador óptimo - Identificar regresor óptimo - Interpretar matriz de confusión - Calcular probabilidades - Obtener información a partir de datos Otras habilidades transversales: - Vincular nuevos conceptos a los consolidados - Tomar decisiones respaldadas por información	- Autosuficiencia y proactividad - Razonamiento basado en sentido común - Inferencia