PEC 2 (20% nota final)

Presentación

En esta Prueba de Evaluación Continuada (PEC) se trabajan los conceptos generales de integración, validación y análisis de los diferentes tipos de datos.

Es importante tener en cuenta las siguientes consideraciones a la hora de entregar la PEC:

- Es obligatorio y **queda como responsabilidad del estudiante revisar que el archivo entregado en el REC es el correcto.** Un archivo vacío o no pertinente se considerará como no entregado.
- Para que la entrega se considere como realizada, se debe completar al menos el 25% de toda la actividad.

Competencias

En esta PEC se desarrollan las siguientes competencias del Máster de Ciencia de Datos:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

Objetivos

Los objetivos concretos de esta PEC son:

- Conocer los efectos de la utilización de datos de calidad en los procesos analíticos.
- Conocer las principales herramientas de limpieza y análisis de los diferentes tipos de datos.
- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinarios.
- Desarrollar las habilidades de aprendizaje que permitan continuar estudiando de una manera que tendrá que ser en gran medida autodirigida o autónoma.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

Descripción de la PEC a realizar

Ejercicio 1

Después de leer el capítulo 1 del recurso *"Introducción a la limpieza y análisis de los datos"*, responde las siguientes preguntas con tus propias palabras.

1. ¿Qué es la conversión en el proceso de limpieza de datos? Describe brevemente con tus propias palabras las técnicas de conversión más habituales. [Máximo 200 palabras]

2. ¿A qué fase del ciclo de vida de los datos corresponden los procesos de reducción, integración y selección? En el caso de realizar reducción de los datos, ¿cuáles son las dos alternativas posibles y en qué se diferencian? Pon un ejemplo práctico de cada alternativa, indicando el objetivo con el cual se pretende aplicar dichas técnicas. [Máximo 200 palabras]

3. ¿Los *outliers* deben considerarse siempre como medidas no válidas? ¿Puedes indicar algún ejemplo en el que no? [Máximo 200 palabras]

4. Se dispone de un dataset de *N* registros, con un atributo *A* que proviene de una población con distribución **no normal** de media μ y varianza σ^2. Se asume que *N* es un valor *significativo* de registros. Se desea realizar un proceso de normalización sobre este atributo *A*. Indica si las siguientes afirmaciones son verdaderas o falsas y justifica la respuesta: [Máximo 300 palabras]
    a. Después de realizar una normalización min-max, la distribución resultante tenderá a una distribución normal cuya media estará en el intervalo [*minA', maxA'*].

    b. Después de realizar una normalización min-max, la distribución de la media muestral resultante tenderá a una distribución normal de media (*maxA*'-*minA*')/2 y varianza 1/*N*, según el teorema central del límite.

    c. Después de realizar una normalización *z*-score, la distribución resultante tenderá a una distribución normal de media 0 y desviación estándar 1.

    d. Después de realizar una normalización *z*-score, la distribución de la media muestral resultante tenderá a una distribución normal de media 0 y desviación estándar 1/*N*, según el teorema central del límite.

Ejercicio 2

Después de leer el capítulo 2 del recurso *"Introducción a la limpieza y análisis de los datos"*, y el recurso complementario *"Data mining: concepts and techniques"*, contesta las siguientes preguntas con tus propias palabras:

1. ¿Qué son los análisis de normalidad y de homocedasticidad? ¿Por qué se aplica antes del análisis de los datos? [Máximo 150 palabras]

2. Tras finalizar un estudio médico, se dispone de un dataset con los datos de los participantes. El dataset contiene los siguientes atributos:

Atributo	Unidades	Información que se conoce del atributo
Edad	años	-
IMC	kg/m^2	El test de Shapiro-Wilk indica que el p-value es mayor que el nivel de significancia $\alpha = 0.05$.
Sexo	-	Codificado como {M, F}.
País	-	Codificado en una escala numérica con valores de 1 a 195 (cada valor representa un país diferente). El test de Kolmogorov-Smirnov indica que el p-value es mayor que el nivel de significancia $\alpha = 0.05$.
HDL	mg/dL	-
LDL	mg/dL	Sigue una distribución normal.
Tipo_dieta	-	Atributo categórico de más de dos valores.
Cant_medic	mg/día	Cantidad de medicación. Sigue una distribución normal.
Síntoma_1	-	Codificado como {Sí, No}
Síntoma_2	-	Codificado como {imperceptible, muy_débil, débil, medio, alto, muy_alto}.

Propón un análisis estadístico para cada una de las siguientes parejas de atributos, y qué condiciones se deberían cumplir para aplicarlo: [Máximo 300 palabras]

   a. IMC y LDL
   b. Sexo y Síntoma_1
   c. HDL y Síntoma_2
   d. LDL y País
   e. Cant_medic y Síntoma_1
   f. País y Tipo_dieta

Ejercicio 3

Después de leer el capítulo 2 del recurso *"Introducción a la limpieza y análisis de los datos"*, y el recurso complementario *"Data mining: concepts and techniques"*, contesta las siguientes preguntas con tus propias palabras:

1. Queremos hacer un algoritmo de clasificación en el ámbito médico, y queremos que nuestro sistema evite no detectar a los enfermos que realmente lo son. ¿Qué métrica crees que se debe maximizar en este caso, la exactitud, la sensibilidad, la especificidad o la precisión? Justifica tu respuesta. [Máximo 100 palabras]

2. ¿En qué se diferencian los modelos de regresión lineal y regresión logística, suponiendo que las variables independientes empleadas fueran las mismas? ¿Cuáles son las métricas que nos permiten evaluar la calidad de estos modelos y de qué forma lo indican? ¿Cómo se podría diseñar un algoritmo que utilice alguno de estos modelos para predecir una variable categórica nominal con cuatro posibles valores? Justifica tus respuestas. [Máximo 300 palabras]

3. Para evaluar el rendimiento de los modelos de clasificación, cuáles son las técnicas más empleadas para la partición de los datos en subconjuntos de entrenamiento y de prueba. Mencione y explique tres de ellas. [Máximo 300 palabras]

4. Tras ejecutar un algoritmo de clasificación de tres grupos, obtenemos como resultado la siguiente matriz de confusión:

		Valor en la realidad		
		Motocicleta	Bicicleta	OTRO
Predicción	Motocicleta	*A* = 97	*B* = 7	*C* = 3
	Bicicleta	*D* = 51	*E* = 120	*F* = 5
	OTRO	*G* = 17	*H* = 13	*I* = 121

Explica con tus propias palabras qué significan las siguientes medidas en este caso. Además, indica su fórmula, utilizando los identificadores A, B, C, etc., y calcula su valor numérico. [Máximo 200 palabras]

   a. La exactitud global del algoritmo.
   b. La sensibilidad para la clase "Bicicleta".
   c. La especificidad para la clase "Bicicleta".
   d. La precisión para la clase "Bicicleta".
   e. El F1-score para la clase "Bicicleta" (puedes dejarlo indicado en función de variables previamente calculadas en apartados anteriores).

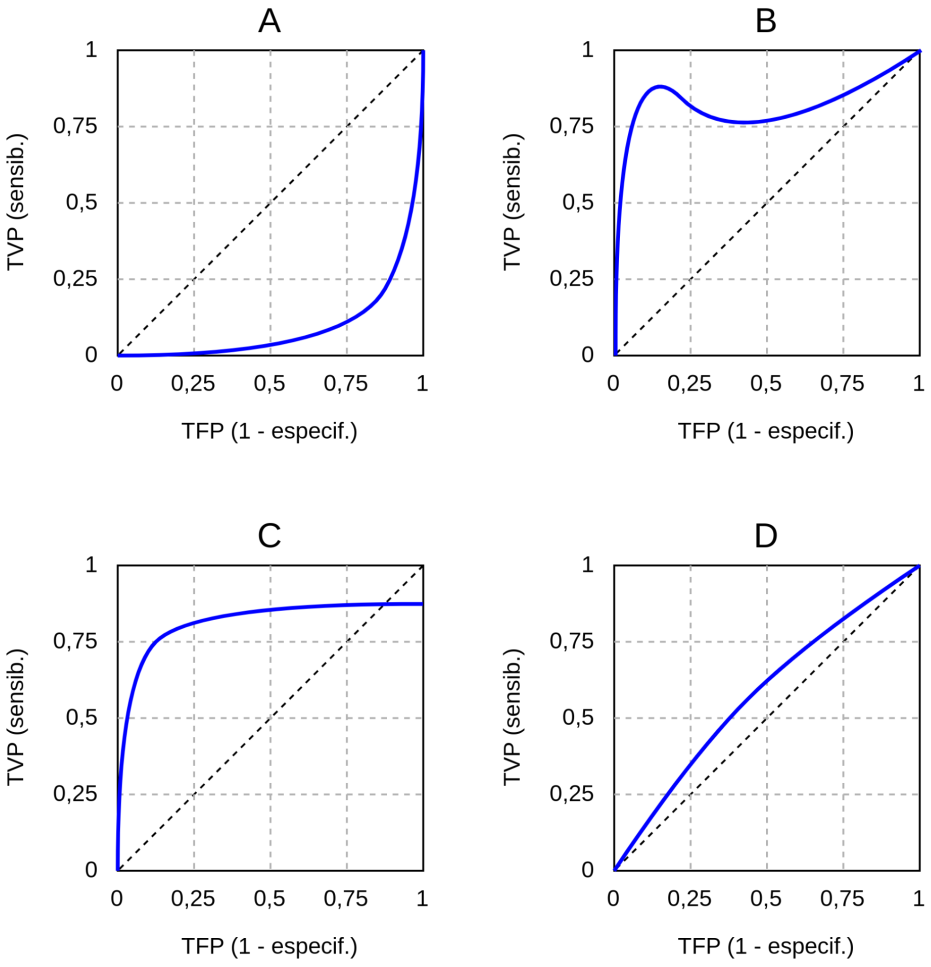
Define, de manera teórica, la fórmula de:

f. El macro F1-score (puedes dejarlo indicado en función del F1-score de cada categoría).

g. El weighted F1-score (puedes dejarlo indicado en función del F1-score de cada categoría).

5. En un sistema de reconocimiento de texto, se desea identificar la letra "Z" con los siguientes requerimientos:
   - Tasa de verdaderos positivos (TVP) superior al 75%.
   - Tasa de falsos positivos (TFP) inferior al 25%.

   Para ello, se dispone de cuatro algoritmos de clasificación (A, B, C y D) sobre los que previamente se ha realizado un análisis para evaluar su rendimiento, obteniéndose las curvas ROC que se muestran en la figura. Para cada caso, justifica cuáles son las conclusiones más relevantes, y si se podría utilizar alguno de estos algoritmos para cumplir con los requerimientos del problema. En caso afirmativo, indica de forma aproximada el punto de la curva ROC asociado al umbral de decisión elegido. [Máximo 300 palabras]



Recursos

Los siguientes recursos son de utilidad para la realización de la PEC:

Básicos

- Calvo M., Pérez D., Subirats L (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.

Complementarios

- Megan Squire (2015). *Clean Data*. Packt Publishing Ltd. Capítulos 1 y 2.
- Jiawei Han, Micheine Kamber, Jian Pei (2012). *Data mining: concepts and techniques*. Morgan Kaufmann. Capítulo 3.
- Jason W. Osborne (2010). *Data Cleaning Basics: Best Practices in Dealing with Extreme Scores.* Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.

Criterios de valoración

La ponderación de los ejercicios es la siguiente:

Ejercicio	1.1	1.2	1.3	1.4	2.1	2.2	3.1	3.2	3.3	3.4	3.5
Puntos	0,75	0,75	0,75	1	1	1,5	0,5	1	0,75	1	1

Se valorará la idoneidad de las respuestas, que deberán ser claras y completas. Cuando sea necesario, deberán acompañarse de ejemplos representativos y bien justificados.

Formato y fecha de entrega

Se debe entregar un único documento en **formato PDF** (no se aceptarán otros formatos) con las respuestas a las preguntas.

Este documento se debe subir al espacio de Entrega y Registro de EC del aula antes de las **23:59h CET del día 13 de diciembre**. No se aceptarán entregas fuera de plazo.