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]

   La conversión es una etapa dentro de la limpieza de datos con el objetivo de transformar los datos para realizar un análisis más eficiente y con resultados más fácilmente entendibles. Las técnicas de conversión más utilizadas son:

   - Normalización, su objetivo es la reducción o estandarización de los valores en una escala previamente establecida, como por ejemplo (0,1) o (-1,1). Las técnicas más utilizadas son min-max o z-score.
   - Transformación de *Box-Cox*, se utiliza cuando nos encontramos con un conjunto de datos donde existen problemas de normalidad y homogeneidad de varianzas. Su objetivo es transformar los datos para que se parezcan lo más posible a una distribución normal.
   - Discretización, consiste en la sustitución de los valores numéricos por categorías o rangos, pudiendo ser estos conceptuales. Su objetivo es facilitar su uso para la aplicación de algoritmos de aprendizaje automático.

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]

   Los procesos de reducción, integración y selección de datos corresponden a la etapa de preprocesado o limpieza del ciclo de vida de los datos. En cuanto a las alternativas posibles para el proceso de reducción de los datos, podemos citar la reducción de dimensionalidad, y la reducción de la cantidad. Al reducir la dimensionalidad, seleccionamos un subconjunto de atributos del conjunto inicial de los datos, mientras que, al reducir la cantidad, seleccionamos un subconjunto de muestras u observaciones.

Por ejemplo, dada una base de datos que recoja los datos clínicos de un grupo de pacientes con patologías cardio-respiratorias (edad, sexo, altura, peso, enfermedad diagnosticada, síntomas, fumador, medicación, etc), podríamos querer aplicar alguna de las siguientes reducciones:

- Reducción de la dimensionalidad para solo tener en cuenta la incidencia de la enfermedad según el sexo y la edad de los pacientes, por ser la información que queremos comparar en nuestro estudio. En ese caso, solo analizaríamos tres atributos por cada paciente en la base de datos.
- Reducción de la cantidad de pacientes al solo seleccionar un rango de edad específico (mayores de 40 años), por ser el objetivo de nuestro estudio. En este caso, se analizarían todos los atributos, pero solo de aquellos pacientes que correspondan al rango de edad seleccionado.

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

No, es posible que un *outlier* no provenga de datos erróneos. Aunque sean poco probables, existen valores extremos que pueden pertenecer a la población muestreada.

- ***Outliers* causados por errores en la medición o colecta de los datos**. Este valor atípico puede ser originado por un error en la unidad de magnitud introducida al momento de ingresar los datos. Por ejemplo, al analizar la cantidad de un cierto producto almacenado en una fábrica a lo largo del tiempo, nos damos cuenta de que, en un cierto mes, se reportó un valor en gramos cuando generalmente se esperan valores en el orden de miles de kilogramos (toneladas).

- ***Outliers* por sesgo de muestreo**. Son errores surgidos al incluir erróneamente individuos de poblaciones no destinadas a ser muestreadas. Por ejemplo, en un estudio donde se analiza el efecto de un fármaco sobre la frecuencia cardíaca en dos grupos de pacientes, donde uno de ellos recibía placebo, se descubrió que dos de los participantes llevaban un marcapasos.

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.

a. **Falso**. La normalización min-max realiza una transformación afín de los datos: los valores de la distribución resultante v_i' seguirán los de la distribución original, escalados por un factor $\alpha = \frac{(maxA' - minA')}{(maxA - minA)}$ y desplazados $-\alpha \cdot minA + minA'$. Es decir, $v_i' = \alpha(v_i - minA) + minA'$. Por tanto, si la distribución original no era normal, la distribución resultante tampoco será una distribución normal, aunque su media sí estará en el intervalo $[minA', maxA']$.

b. **Falso**. La distribución de la media muestral será normal, y su media se encontrará entre los valores *minA'* y *maxA'*. No obstante, su valor no tiene por qué coincidir con el valor indicado, en general. Nótese que el valor indicado puede estar incluso fuera del intervalo $[minA', maxA']$, cosa que no tiene sentido (p. ej. $minA' = 40, maxA' = 50$). Además, la varianza tampoco tendrá el valor indicado, sino que será $\alpha^2\sigma^2/N$, donde α es el factor del apartado anterior y σ^2 es la varianza de la distribución original.

c. **Falso**. Es cierto que la media tenderá a $(\mu - \mu_A)/\sigma_A \to 0$ y la desviación estándar tenderá a $\sigma/\sigma_A \to 1$. Pero como ocurre en el primer caso, simplemente se está escalando y trasladando los valores de la distribución, por lo que la distribución resultante no será normal, ya que la distribución original tampoco lo es.

d. **Falso**. La distribución de la media muestral, en este caso, seguirá una distribución normal de media 0 y varianza $1/N$, es decir, desviación estándar $1/\sqrt{N}$.

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]

   Los análisis de normalidad son pruebas que tienen como objetivo analizar cuánto difiere la distribución de datos observados con respecto a lo esperado si procediesen de una distribución normal con la misma media y desviación típica. Por otro lado, los análisis de homocedasticidad buscan comprobar si la varianza del error es constante a lo largo de las observaciones.

   Estos análisis se aplican antes de los análisis de los datos ya que ciertos análisis estadísticos asumen ciertas características o suposiciones sobre los datos que deben cumplirse para que dichas técnicas, junto con las conclusiones extraídas, sean válidas. En el caso de los test paramétricos, entre las suposiciones más habituales, están el hecho de que los datos se encuentren distribuidos normalmente, así como que los grupos de datos presenten varianzas similares.

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

a. **IMC y LDL:** Puesto que IMC y LDL son variables numéricas continuas y normales (p-valor mayor que el nivel de significancia $\alpha = 0.05$ en el test de Shapiro-Wilk para IMC), podría realizarse un análisis de correlación de Pearson. Para ello, debería verificarse, además, la ausencia de outliers, la relación lineal entre las dos variables y la condición de homocedasticidad.
b. **Sexo y Síntoma_1:** Al tratarse de dos variables categóricas podría realizarse un test de χ^2 para comprobar si existen diferencias significativas entre los grupos definidos, p.ej., por el atributo Sexo.
c. **HDL y Síntoma_2:** En este caso nos encontramos con una variable numérica (asumimos no-normal) y una variable categórica ordinal, por lo que podría realizarse un análisis de correlación de Spearman, tras mapear el atributo Síntoma_2 a una escala del 0 al 5.
d. **LDL y País:** El test de Kolmogorov-Smirnov no tiene sentido aplicarlo en este caso sobre el atributo País, ya que se trata de una variable categórica no ordinal, aunque esté expresada de forma numérica. Puesto que el atributo LDL sigue una distribución normal, debería comprobarse la homocedasticidad con respecto al atributo País (p.ej., test de Levene) para realizar un test de ANOVA. En caso contrario, debería aplicarse la alternativa no paramétrica de Kruskal-Wallis.
e. **Cant_medic y Síntoma_1:** Al tratarse de un atributo que cumple la condición de normalidad, si se verifica la condición de homocedasticidad de Cant_medic respecto Síntoma_1 (p.ej., test de

Levene), podría aplicarse la prueba paramétrica *t* de Student. Si esto no ocurriese, sería conveniente aplicar la prueba de Mann-Whitney.

f. **País y Tipo_dieta:** De nuevo, son dos variables categóricas nominales, por lo que el análisis conveniente en este caso sería un test de χ^2.

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]

   En este caso queremos maximizar los verdaderos positivos (VP) sobre el total de positivos reales (P), por lo tanto, de las propuestas métricas queremos maximizar la sensibilidad.

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]

   Los modelos de **regresión lineal** nos permiten establecer la relación de dependencia lineal que existe entre una variable numérica dependiente (variable a predecir) y las variables independientes o predictoras. En el caso de la **regresión logística**, la variable dependiente es dicotómica (Enfermo/Sano, 0/1), y el modelo se basa en estimar las probabilidades de ocurrencia de una clase u otra, empleando una escala transformada basada en una función logística.

   Para evaluar la **calidad** de los modelos de regresión lineal, se emplea el parámetro R^2 o *R-squared*, donde valores de R^2 cercanos a 1, indican una mayor bondad en el ajuste del modelo obtenido. Para los modelos de regresión logística, se emplea el parámetro AIC (Akaike Information Criterion), el cual considera simultáneamente la bondad del ajuste y la complejidad del modelo, siendo el mejor de entre varios modelos, aquel con menor AIC.

Para realizar una clasificación de una variable categórica con **cuatro posibles valores**, un método habitual consiste en utilizar un modelo de regresión logística dicotómico para cada una de las clases. Por ejemplo, para clasificar entre [coche, avión, barco, tren], dispondríamos de una instancia de regresión logística que indicaría si el objeto a clasificar es un coche (0/1), otra instancia para avión (0/1), etc. y tomaríamos como resultado aquella clase cuyo valor sea más próximo a 1 (*one-hot encoding*). No sería conveniente utilizar un modelo de regresión lineal, entre otras cosas, porque se trata de una variable no ordinal.

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]

Los métodos más comunes empleados en la partición de los datos son el método de exclusión, el método de submuestreo aleatorio, y el método de validación cruzada.

Método de exclusión o *holdout:* El conjunto de datos se divide aleatoriamente en dos subconjuntos, el de entrenamiento y el de prueba, siendo el primero, por ejemplo, dos tercios del total de los datos originales, y el resto se emplea como conjunto de prueba. Este método se considera más apropiado para grandes conjuntos de datos.

Método de submuestreo aleatorio: Es similar al método *holdout*, pero realizado *k* veces de forma aleatoria, para posteriormente obtener una medida promedio del rendimiento global del modelo a partir de las *k* estimaciones. Suele proporcionar una medida más realista del rendimiento del modelo.

Método de validación cruzada: Este método divide el conjunto de los datos originales en *k* subconjuntos (*folds*) mutuamente exclusivos, de forma aleatoria y con tamaños similares. El entrenamiento se realiza con *k-1* subconjuntos dejando el subconjunto restante para testear el modelo. Este proceso se repite *k* veces, para finalmente calcular la exactitud como el número total de clasificaciones correctas obtenidas en las *k* iteraciones, dividido por el número total de muestras en el conjunto de datos original. La validación cruzada puede ser también estratificada, donde cada subconjunto mantiene aproximadamente, la misma distribución de las clases que el conjunto de datos original. El método *leave-one-out* es un caso particular de validación cruzada, donde *k* representa el total de muestras del conjunto original.

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).

a. Proporción de registros correctamente clasificados de todas las clases, sobre el total de registros:

$$Exactitud = \frac{A+E+I}{A+B+C+D+E+F+G+H+I} = 0,779.$$

b. Proporción de registros clasificados como Bicicletas y que efectivamente lo son, sobre el total de registros reales de Bicicletas:

$$Sensibilidad_{Bicicleta} = \frac{E}{B+E+H} = 0,857.$$

c. Proporción de registros no clasificados como Bicicletas y que efectivamente no lo son, sobre el total de registros que no son realmente Bicicletas:

$$Especificidad_{Bicicleta} = \frac{A+C+G+I}{A+C+D+F+G+I} = 0,810.$$

d. Proporción de registros clasificados como Bicicletas y que efectivamente lo son, sobre el total de registros clasificados como Bicicletas:

$$Precisión_{Bicicleta} = \frac{E}{D+E+F} = 0,682.$$

e. Media armónica de la precisión y la sensibilidad de la clase "Bicicleta":

$$F1_{Bicicleta} = 2\,\frac{Precisión_{Bicicleta} \cdot Sensibilidad_{Bicicleta}}{Precisión_{Bicicleta} + Sensibilidad_{Bicicleta}} = \frac{2E}{B+D+2E+F+H} = 0,759.$$

f. Media aritmética de los F1-score de todas las clases:

$$F1_{MACRO} = \frac{1}{3}F1_{Motocicleta} + \frac{1}{3}F1_{Bicicleta} + \frac{1}{3}F1_{OTRO}.$$

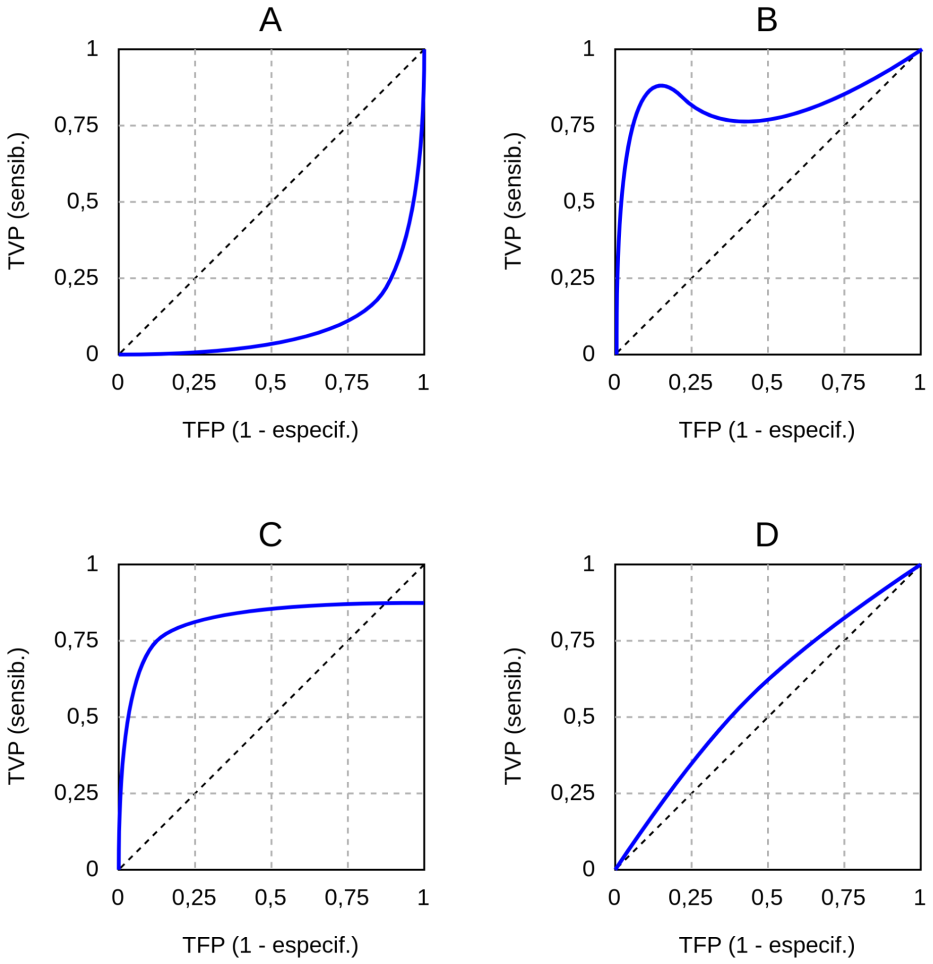
g. Media ponderada (por el total de registros reales) de los F1-score de cada clase:

$$F1_{WEIGHTED} = \frac{1}{A+B+C+D+E+F+G+H+I}\,((A + D + G)\cdot F1_{Motocicleta} + (B + E + H)\cdot F1_{Bicicleta} + (C + F + I)\cdot F1_{OTRO}),$$

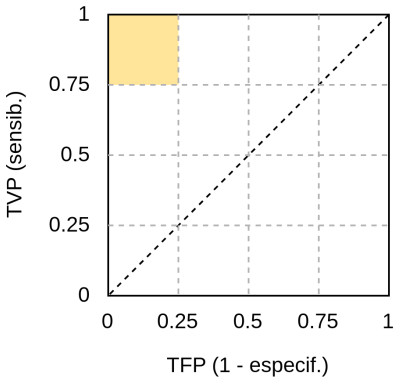
$$F1_{WEIGHTED} = \frac{165}{434}F1_{Motocicleta} + \frac{140}{434}F1_{Bicicleta} + \frac{129}{434}F1_{OTRO}.$$

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]

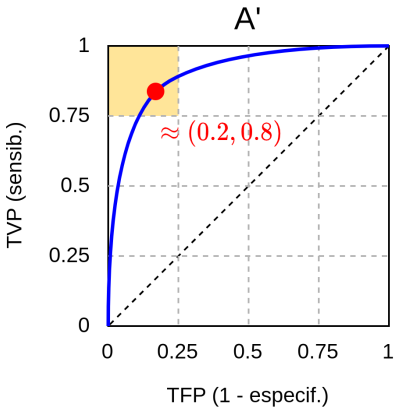


En primer lugar, identificamos que la región en la que la curva ROC satisface los requerimientos del problema es el recuadro de color amarillo en la siguiente figura:



Una vez identificada esta región, veamos caso por caso cada curva ROC:

A. Este caso corresponde a un clasificador inverso: confunde los casos positivos con los negativos. Es decir, “casi siempre” que recibe como entrada “Z” indica un valor próximo a 0 (cuando debería ser próximo a 1), y viceversa. En base a esto, podría construirse un nuevo algoritmo A’ cuya salida consista en invertir la salida del clasificador A (es decir, si A retorna *x*, el algoritmo A’ retornará 1-*x*). La curva ROC de este nuevo algoritmo sería un reflejo en la diagonal de la gráfica A, pudiéndose tomar un umbral de clasificación que corresponda a un punto de la curva ROC dentro de la región de interés, aproximadamente (0,2, 0,8), según la curva:



B. Esta curva ROC es incorrecta. Una curva ROC siempre es no-decreciente, ya que ningún caso que es clasificado como positivo será clasificado como negativo para cualquier valor umbral de clasificación menor. Por tanto, no puede extraerse ninguna conclusión de este caso.
C. De nuevo, una curva ROC incorrecta, ya que debería tratarse de una curva no-decreciente que empieza en la esquina (0, 0) y termina en (1,1). Tampoco es posible extraer ninguna conclusión, más que el hecho de que el análisis realizado en este clasificador ha sido incorrecto.
D. Esta curva corresponde a un clasificador cuya curva es muy próxima a la diagonal, y que por tanto sería levemente mejor que un clasificador aleatorio (“lanzar una moneda al aire”). Al estar la curva tan próxima a la diagonal, queda lejos de satisfacer los requerimientos del problema.

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.