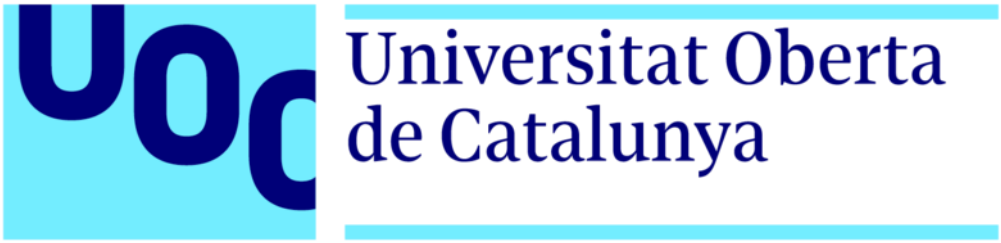

Máster universitario de Ciencia de Datos

Prueba de Evaluación Continua 2 – PEC2

Trabajo sobre los conceptos generales de integración, validación y análisis de los datos

Autor:

Mario Ubierna San Mamés

Índice de Contenido

1. Introducción

1.1. Presentación

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

1.2. Objetivos

Los objetivos concretos de esta Prueba de Evaluación Continuada 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.

2. Enunciado

2.1. Ejercicio 1

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

1. *¿A qué fase del ciclo de vida de los datos corresponden los procesos de reducción, integración y selección? En el caso de realizar la reducción de los datos, ¿cuáles son las dos alternativas posibles y en que se diferencian? Ponga 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 se corresponden a la fase del ciclo de vida llamada limpieza de datos o preprocesado, en ella se busca eliminar o corregir registros en los cuales faltan datos, o son incorrectos, inexactos…

   Por otro lado, para reducir la dimensionalidad de un conjunto podemos hacer uso de dos alternativas [1]:

   - Reducción de la dimensionalidad: su principal objetivo es reducir el tamaño del conjunto de datos a partir de sus atributos.

     Un ejemplo sería, si tenemos un *dataset* que contiene información sobre estadísticas de jugadores de baloncesto, y en ese *dataset* tenemos 50 atributos (altura, peso, % de tiro de 2, % de tiro de 3, puntos anotados, rebotes, robos, tapones…).

   - Reducción de la cantidad: su principal objetivo es reducir el tamaño del conjunto de datos a partir de la cantidad de registros.

     Un ejemplo sería, las transacciones de compra que se hacen en *Amazon*, cada segundo se hacen muchas compras, si analizamos el *dataset* que tenemos al cabo de un año tendríamos millones y

millones de registros, por lo tanto, podría ser interesante tener un *dataset* con un conjunto de datos menor que represente de la misma o máxima forma posible el *dataset* original.

2. *Describe con tus propias palabras y mediante un ejemplo los cuatro principales métodos de submuestreo aleatorio que permiten la reducción de la cantidad (máximo 300 palabras).*

   Los cuatro principales métodos de submuestreo aleatorio son [1]:

   - Muestra aleatoria simple sin sustitución: básicamente lo que busca este método es elegir de forma aleatoria diferentes registros cuya probabilidad para ser seleccionado sea la misma. De tal forma que vamos a tener un *dataset* más pequeño que el original.

     Siguiendo con el apartado anterior, imaginémonos que tenemos estadísticas de jugadores de baloncesto de todas las ligas del mundo, tendríamos muchísimos registros, por lo tanto podemos elegir de forma aleatoria diferentes jugadores cuya probabilidad de ser elegido para formar parte del nuevo conjunto de datos sea 1/tamaño del *dataset* original, de esta forma conseguimos una muestra de la población.

   - Muestra aleatoria simple con sustitución: es similar que en el caso anterior, con la excepción de que antes una vez que se elegía un jugador ya no se podía volver a elegir, sin embargo, ahora un mismo jugador puede escogerse varias veces.

     Por lo tanto, puede darse el caso de que en la muestra poblacional que tenemos un mismo jugador pueda aparecer más de una vez en la misma muestra.

   - Muestra de clústeres: este método tiene un enfoque diferente a los dos anteriores, ahora lo que hacemos es agrupar los registros, en nuestro caso los jugadores de características similares (altura, peso…), y una vez que tenemos todos los grupos identificados se elige de forma aleatoria x grupos, dónde x es menor que el número de clústeres identificados originalmente.

   - Muestra estratificada: es una fusión de los métodos anteriores, ya que dividimos el *dataset* original en diferentes muestras disociadas llamadas estratos, y elegimos de forma aleatoria x registros de cada estrato.

Por lo tanto, podríamos tener diferentes estratos a partir de la posición de cada jugador (en baloncesto hay 5 posiciones, 5 estratos), y elegiríamos x jugadores de forma aleatoria dentro de cada estrato.

2.2. Ejercicio 2

Después de leer el capítulo 1.5 y 1.6 del recurso "Introducción a la limpieza y análisis de los datos", contesta las siguientes preguntas con tus propias palabras.

1. *¿Qué se considera un outlier?¿Cuáles son los posibles efectos de su presencia en los resultados finales de los análisis estadísticos? (máximo 150 palabras)*

   Podríamos definir *outlier* como aquellos valores que se alejan de los valores "normales" de una variable o población que se les consideran como observaciones anómalas y extremas [2], es decir, son valores que están alejados de la media de la muestra o de la población.

   Los efectos que tiene los *outliers* sobre el análisis estadístico son adversos, es decir, al fin y al cabo tener valores extremo implica que va a variar la varianza de los datos, ya que al tener valores más alejados a la media más diferentes son los datos entre sí. Si los datos son muy diferentes al realizar un análisis estadístico inferencial no podríamos generalizar de forma correcta, ya que estaríamos cometiendo un error muy grande, es decir, los datos están sesgados.

2. *¿Los outliers pueden considerarse cómo medidas válidas de los datos? Explica con tus propias palabras dos posibles causas que pueden dar lugar a la aparición de outliers, poniendo un ejemplo práctica en cada una. (máximo 200 palabras)*

   Los valores extremo lógicamente pueden considerar medidas válidas de los datos, es decir, que un valor por muy alejado que esté de la media no significa que no sea correcto, por ejemplo, si tomamos la altura de los jugadores de baloncesto en cm y la media nos da 180 cm, si tenemos una observación cuya altura es de 216 cm, por muy alejada que esté de la media este valor puede ser correcto, otra cosa sería si ese valor fuera 400 cm.

   Los *outliers* pueden aparecer por diferentes factores, uno de ellos un error humano, si alguien al introducir la altura de un jugador en vez de poner 190 cm pone 290 porque el número dos está cerca del uno tendremos un valor extremo.

Otro caso por el que pueden aparecer valores extremos es porque medimos una misma variable en diferentes escalas, es decir, si la altura está en cm pero hay valores en la que está en metros, puede darse el caso de que tengamos un valor 190 cm y otro valor 2 m, realmente no hay mucha diferencia entre ellos pero al estar en diferentes escalas la diferencia es muy grande, generando así un *outlier*.

3. *Describe tres técnicas utilizadas para el tratamiento de los datos perdidos y pon ejemplos donde aplicarías cada una de estas técnicas. (máximo 400 palabras)*

   Los datos perdidos o *missing data* es aquella información que se ha perdido por diferentes motivos, y para solventar dicho problema podemos hacer uso de diferentes técnicas [1].

   La primera de ellas es hacer uso de medidas de tendencia central para imputar el valor, es decir, podríamos por ejemplo calcular la media de una variable y todas aquellas observaciones que no tenga valor imputamos la media. Esto mismo lo podríamos realizar para la mediana, y calcular tanto la media como la mediana para toda la muestra o por categorías. Aplicaría esta técnica si la variable fuese numérica, en caso contrario, si es categórica podríamos imputar el valor más frecuente de la muestra o por categorías.

   La segunda técnica sería categorizar los valores perdidos bajo una misma etiqueta, por ejemplo, si alguna nacionalidad de los jugadores de baloncesto no tiene un valor, podríamos categorizar dicho valor perdido por un "Desconocido" o valor "NA"… Esto es realmente útil porque si tenemos diferentes tipos de valores perdidos dentro de un mismo atributo, al categorizarlos todos como lo mismo establecemos un significado común para todos ellos. En nuestro caso, como bien comentábamos la nacionalidad puede venir como una cadena de caracteres vacía, como un valor "NA", por lo que crear una etiqueta, como por ejemplo "Desconocido" para agrupar todos los valores perdidos dentro de las misma categoría, nos facilitaría luego el manejar esos datos.

   La tercera técnica es hacer uso de algún método probabilístico para imputar dichos valores perdidos, es decir, si tenemos jugadores de baloncesto con una altura de 0 (lo cual no tiene sentido), podríamos hacer uso de un modelo de regresión lineal, en el cual, entrenamos al modelo con los valores de altura que conocemos, y una vez entrenado predecimos para todos aquellos registros que no tengan valores qué valor tendrían, de esta forma dejamos de tener valores perdidos y tendríamos registros con más información aunque no sea la verdadera.

2.3. Ejercicio 3

Después de leer el capítulo 2.2 del recurso "Introducción a la limpieza y análisis de los datos", contesta a la siguiente pregunta con tus propias palabras:

1. *¿En qué se diferencian los modelos de regresión lineal y regresión logística, suponiendo que las variables independientes empleados fueran las mismas? ¿Cuáles son las métricas que nos permiten evaluar la calidad de estos modelos y de qué forma lo indican? (máximo 150 palabras)*

   La regresión lineal tiene como objetivo predecir un valor (variable objetivo o dependiente) a partir de las variables independientes [3], para ello se calcula una recta, en nuestro caso, podríamos predecir el valor de la altura de los jugadores de baloncesto.

   La métrica usada para evaluar la calidad de los modelos de regresión lineal es R^2, es decir, cómo de cerca están los datos de la línea de regresión que se ha calculado, este valor va entre 0 y 1, a mayor R^2 mejor es el modelo [1].

   La regresión logística por el contrario también busca predecir, pero en este caso el valor solo puede ser o 0 o 1, es decir, nos ayuda a resolver problemas de clasificación [3]. Siguiendo con el ejemplo, ahora el problema sería predecir si un jugador es alto o no.

   La métrica usada para la regresión logística es AIC, se calcula a partir de la bondad de ajuste y la complejidad del modelo, a menor AIC mejor es el modelo.

2.4. Ejercicio 4

Después de leer los capítulos 2.4 del recurso "Introducción a la limpieza y análisis de los datos", y en el recurso complementario "Data mining: concepts and techniques", contesta las siguientes preguntas con tus propias palabras:

1. *Explica las diferencias entre modelos supervisados y no supervisados. Para cada tipo de modelo, da tres ejemplos de algoritmos. (máximo 200 palabras)*

   La diferencia entre un modelo supervisado y no supervisado es básicamente que, en el primero tenemos un conocimiento del dominio del problema mientras que en el segundo no.

En los modelos supervisados, tenemos pleno conocimiento de todos los datos, es decir, en todo momento conocemos el valor de la variable objetivo o dependiente. Es por ello que el *dataset* original se divide en dos conjuntos, uno para entrenar el modelo, es decir, le pasamos tanto las variables independientes como la variable dependiente para que el modelo pueda calcular el error que genera y así mejorarlo, y una vez entrenado el modelo predecimos, a partir de un conjunto de *testing* podemos validar el modelo y saber cómo de bueno es.

Tres ejemplos de algoritmos de aprendizaje supervisado son: la regresión logística, clasificadores bayesianos, árboles de decisión…

Respecto a los modelos no supervisados, en estos no tenemos el conocimiento sobre la variable objetivo, por lo que estos modelos se ajustan solo a partir de los datos de entrada, es decir, no tenemos observaciones etiquetadas/clasificadas.

Tres ejemplos de algoritmos de aprendizaje no supervisado son: *clustering*, *clustering* jerárquico, mapas auto-organizados…

2. *A la hora de 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. Menciones y explique 3 de ellas. (máximo 300 palabras)*

   Las particiones de datos más usadas en clasificación son [1]:

   - El método de exclusión: es la forma más simple de dividir un conjunto de datos en dos subconjuntos, un para el *training* y otro para el *testing*. Para ello, dividimos de forma aleatoria las observaciones de dos conjuntos, por norma general, el conjunto de *training* suele se doble que el de *testing*. El principal problema de este método es que el modelo se entrena a partir del conjunto de entrenamiento, por lo que dependiendo de cómo de bueno sea el conjunto de entrenamiento va a ser mejor el modelo o no, por lo que es método como primera aproximación está bien.

   - El método de submuestreo aleatorio: este método es muy parecido al anterior, pero como bien hemos mencionado el método de exclusión depende de cómo de bueno sea la partición que se realiza, ya que solo se realiza una vez. Es por ello que este método es en ese caso igual que el anterior, con la peculiaridad de que se va a ejecutar

el modelo x veces y se va a hacer un promedio del resultado, para ver qué partición es mejor.

- El método de validación cruzada: el principal problema que tienen los dos anteriores métodos es que no se utiliza todo el conjunto de datos para entrenar el modelo, es aquí donde entre en juego la validación cruzada. Tomando como ejemplo la validación cruzada de tipo *k-fold*, se divide el *dataset* original en k subconjuntos con datos exclusivos y de tamaño similares, entonces el *training* se realiza tantas veces como combinaciones posibles haya a partir de los subconjuntos k – 1, y ese otro conjunto se deja para el *testing*. De tal forma que se calcula este proceso k veces y se promedia el resultado.

3. Bibliografía

[1] «PID_00265704.pdf». Accedido: abr. 28, 2021. [En línea]. Disponible en: https://materials.campus.uoc.edu/daisy/Materials/PID_00265704/pdf/PID_00265704.pdf.

[2] «Outlier - Definición, qué es y concepto», *Economipedia*. https://economipedia.com/definiciones/outlier.html (accedido abr. 28, 2021).

[3] «Diferencia entre regresión lineal y logística: una pregunta común en una entrevista para científicos de datos», *ICHI.PRO*. https://ichi.pro/es/diferencia-entre-regresion-lineal-y-logistica-una-pregunta-comun-en-una-entrevista-para-cientificos-de-datos-66213244436213 (accedido abr. 28, 2021).