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.

Competencias

En esta PEC se desarrollan las siguientes competencias del Máster de Data Science:

- 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 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.

Descripción de la PEC a realizar

Ejercicio 1 [20%]

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. ¿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 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.

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]

   - Muestra aleatoria simple sin sustitución: de todas las muestras o registros que forman el conjunto, se escoge un subconjunto donde la probabilidad de cada registro de ser seleccionado es la misma. En el ejemplo anterior,

se seleccionaría un subconjunto de pacientes, donde la probabilidad de cada paciente de ser seleccionado sería la misma.

- Muestra aleatoria simple con sustitución: en este caso, cada vez que se selecciona un registro, este se vuelve a tener en cuenta en la siguiente selección, por lo que cada registro puede escogerse varias veces. En el ejemplo anterior, se seleccionaría un subconjunto de pacientes, donde un mismo paciente podría seleccionarse varias veces.

- Muestra de clústeres: la selección se realiza por clústeres, es decir, si los registros en el conjunto se agrupan en M clústeres diferentes, se selecciona un subconjunto de s clústeres de forma aleatoria, donde s < M. En el ejemplo anterior, se dividiría el conjunto de pacientes en clústeres (por ejemplo, según la enfermedad diagnosticada), y se seleccionaría solo un subconjunto de clústeres (es decir, solo algunas de las enfermedades diagnosticadas).

- Muestra estratificada: si el conjunto se divide en partes perfectamente disociadas llamadas estratos, este método selecciona una muestra aleatoria para cada estrato del conjunto. En el ejemplo anterior, se seleccionaría un subconjunto de pacientes, donde las proporciones de cada enfermedad presente en los datos se mantendrían; es decir, si en el conjunto inicial hay un 20% de la enfermedad A, 30% de la enfermedad B y 50% de la enfermedad C, el subconjunto de enfermedades seleccionado mantendría esas mismas proporciones.

Ejercicio 2 [30%]

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]

Los valores extremos (*extreme scores* o *outliers*) son aquellos datos que se encuentran muy alejados de la distribución normal de una variable o población. Son observaciones que se separan notablemente del resto de valores de la muestra, que pueden levantar sospechas sobre si fueron generadas mediante el mismo mecanismo.

Los *outliers* pueden afectar de forma negativa los resultados de los análisis posteriores, puesto que incrementan el error en la varianza de los datos y sesgan significativamente los cálculos y estimaciones.

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 práctico de cada una. [Máximo 200 palabras]

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

1. ***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).

2. ***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.

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].

1. **Reemplazar por una misma constante o etiqueta.** Se utiliza, generalmente, cuando los datos perdidos tienen un significado común por lo que la etiqueta o constante tiene un significado y facilita una interpretación. Por ejemplo, en un formulario clínico si se pregunta si hay riesgo de embarazo, hay hombres que dejarán vacio (dato perdido) la casilla en vez de colocar “No”. Por lo tanto, estos datos pueden ser remplazados por una etiqueta unica de “No procede” (o cualquier otra que se desee).

2. **Reemplazar por una misma medida de tendencia central.** Como su nombre lo indica, se remplazan los valores perdidos por una única medida de tendencia central sobre un cierto conjunto de datos. Dicha medida se puede calcular tanto para toda la muestra como para cada una de las clases

o categorías que la describan. Sin embargo, su utilización depende de la distribución de los datos. Por ejemplo, si se desea buscar los niveles de colesterol en un grupo de pacientes, los valores perdidos pueden ser remplazados por la media de un subgrupo similar al del paciente en cuestión. Es decir, se buscaría la media de los niveles de colesterol en pacientes en un rango de edad similar al paciente con el dato perdido para que este sea remplazado.

3. **Implementación de métodos probabilistas para predecir los valores perdidos.** Son técnicas basadas en métodos o modelos probabilistas que permiten predecir los valores perdidos, a partir del resto de atributos o de muestras anteriores. Entre los más utilizados tenemos las regresiones, los modelos bayesianos y los árboles de decisión. Se utilizan cuando se requiera una aproximación más precisa del dato perdido. Por ejemplo, al analizar una señal de ritmo cardiaco, podemos imputar (predecir) los datos perdidos basándonos en las medidas previas a través de la utilización de algoritmos como los mencionados previamente.

Ejercicio 3 [10%]

Después de leer el capítulo 2.2 del recurso “Introducción a la limpieza y análisis de los datos”, contesta 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 empleadas 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]

   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.

Ejercicio 4 [40%]

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]

   Los **modelos supervisados** estiman una función o modelo a partir de una serie de datos de entrenamiento, con el objetivo de predecir posteriormente el resultado de nuevos datos desconocidos. Los conjuntos de datos de entrenamiento están formados por pares de objetos que representan los datos de entrada y los resultados deseados. Estos resultados pueden ser un valor numérico o una etiqueta de clase. El conjunto de datos se dividirá por tanto en los subconjuntos de entrenamiento (*training*) y de prueba o test (*testing*). Gracias a los primeros, se entrenará un modelo que aprenderá a predecir el resultado adecuado. Ejemplos de algoritmos: *Support Vector Machine* (SVM), random forest y *k-nearest neighbors* (k-NN).

   El **modelo no supervisado** tiene lugar cuando no se dispone de datos "etiquetados" para el entrenamiento. Sólo conocemos los datos de entrada, pero no existen datos de salida que correspondan a un determinado *input*. Por lo tanto, consiste en adaptar un modelo a las observaciones dadas. Ejemplos de algoritmos: *k-means*, *Principal Component Analysis* (PCA) y *Singular Value Decomposition* (SVD).

2. A lo 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. Mencione y explique 3 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 de 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.

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: 20%
- Ejercicio 2: 30%
- Ejercicio 3: 10%
- Ejercicio 4: 40%

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 Word, Open Office o **PDF** (preferiblemente este último) con las respuestas a las diferentes preguntas.

Este documento debe entregarse en el espacio de Entrega y Registro de AC del aula antes de las **23:59** del día **3 de mayo**. No se aceptarán entregas fuera de plazo. Todas las actividades son obligatorias.