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].

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]

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]

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]

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].

Ejercicio 3 [20%]

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]

Ejercicio 4 [30%]

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]

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]

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: 20%
- Ejercicio 4: 30%

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.