Logotipo, nombre de la empresa

Descripción generada automáticamente

Universidad de Montevideo – Introducción a la Ciencia de Datos

Examen 2022

17 de Febrero, 2022

1. Por qué cree que no debe utilizarse el método de validación K-Fold Cross-Validation cuando se tiene una serie de tiempo? Cómo cree que deberían dividirse los conjuntos de entrenamiento y validación?
2. Por qué es necesario el pre-procesamiento o limpieza de datos?
3. Usted esta pre-procesando un dataset de entrenamiento que será utilizado para entrenar un modelo lineal. El dataset contiene cientos de atributos y usted tiene que decidir cuales descartar. Cuales de las siguientes decisiones tomaría usted para cumplir con este objetivo? Seleccione tres de ellas y justifique:
   1. Descartar un atributo si tiene muchos valores faltantes.
   2. Descartar un atributo si tiene pocos valores faltantes.
   3. Descartar un atributo si tiene una alta correlación con el atributo objetivo.
   4. Descartar un atributo si tiene baja varianza.
   5. Descartar un atributo si tiene una baja correlación con el atributo objetivo.
   6. Descartar un atributo si tiene alta varianza.
4. Cómo determinaría el número correcto de clusters en un problema no supervisado?
5. Tf-idf es una técnica estadística que se usa con frecuencia en dominios de aprendizaje automático, como el resumen y la clasificación de texto. Tf-idf mide la relevancia de una palabra en un documento en comparación con todo el corpus de documentos. Dispone de un corpus (D) que contiene los siguientes documentos:
   1. Documento 1 (d1): “A quick brown fox jumps over the lazy dog. What a fox!”
   2. Documento 2 (d2): “A quick brown fox jumps over the lazy fox. What a fox!”

Cual de las siguientes opciones considera correcta? Justifique:

1. Usando tf-idf, la palabra "fox" es igualmente relevante tanto para el documento d1 como para el documento d2.
2. Usando tf-idf, la palabra "fox" es más relevante para el documento d2 que para el documento d1.
3. Usando tf-idf, la palabra "fox" es más relevante para el documento d1 que para el documento d2.
4. Se ha proporcionado información insuficiente para calcular tf-idf.
5. Cuál es la diferencia entre aprendizaje supervisado y no supervisado?
6. Comente una ventaja y una desventaja del modelo de regresión lineal.
7. Qué significa que una serie de tiempo sea estacionaria?
8. Outliers:
   1. ¿Qué formas de detectar outliers conoce?
   2. Una vez detectados, ¿qué tratamiento debe darles?
9. Cuál es la diferencia entre un box plot, un histograma y un violin plot? Explique sus usos.