Logotipo, nombre de la empresa

Descripción generada automáticamente

Universidad de Montevideo – Introducción a la Ciencia de Datos

Examen 2021

23 de Noviembre, 2021

1. Suponga que usted tiene dos tablas A y B. Para cada fila de A usted quiere traer los datos asociados al registro en la tabla B si es que existe, pues no quiere perder registros de la tabla A tras hacer la intersección. ¿Qué tipo de operación debe hacer y como la haría?
2. ¿Cuál es la diferencia entre escalar y estandarizar una columna? ¿Cuándo se aplica cada método?
3. Outliers:
   1. ¿Qué formas de detectar outliers conoce?
   2. Una vez detectados, ¿qué tratamiento debe darles?
4. Usted quiere segmentar sus datos haciendo clustering pero sabe que sus datos tienen outliers. ¿Es K-Means un algoritmo apropiado? Si o No. Justifique su respuesta.
5. Cross Validation:
   1. ¿Cuál es la diferencia entre el Validation Set y el Test Set?
   2. ¿Cuándo se debería hacer el pre-procesamiento de los datos, antes o después del split? Explique su respuesta.
6. Comente que entiende por efecto neto y bruto en una regresión lineal múltiple. Incluya un ejemplo diferente al visto en clase.
7. ¿Cómo se puede calcular el *accuracy* utilizando una matriz de confusión?
8. Comente una ventaja y una desventaja del modelo de regresión lineal.
9. En un problema de clasificación binaria fuertemente desbalanceado si la clase minoritaria es la clase positiva:
   1. ¿Cómo espera que sean las métricas Sensitividad y Especificidad?
   2. ¿Cómo solucionaría el problema del desbalance de clases?
10. ¿Cuáles son las diferencias entre el aprendizaje supervisado y no supervisado?