Examen I.III - Análisis de Datos II

Universidad de Costa Rica
Curso CA0305 I-2024
Parte III: Clases - Funciones
Encargado: Potoy Juárez Luis Alberto
Correo: luis.juarez@ucr.ac.cr

Considere a las siguientes indicaciones para el desarrollo del examen:

Indicaciones

1. El examen debe llevar el nombre: *Examen_I_III - Carnet - NombreApellidos*
2. Puede usar todo material visto en clases, tareas y material de apoyo virtual
3. Valor de la prueba es de 30%
4. El examen es de forma individual. En caso de utilizar alguna fuente o referencia documente. Todo acto de plagio será motivo invalidación de la prueba.
5. La entrega de la tarea es por medio de mediación virtual y con fecha límite del día 5 mayo 2024 3:00 pm. No se aceptaran entregas después de la hora establecida sin justificación.
6. El desarrollo de la prueba debe realizarlo en Jupyter Notebook y en Spyder. Los módulos deben ser programados en Spyder y las pruebas - resultados de testeo en Jupyter Notebook.
7. Al no cumplir algunas de las normas estándares de programación vista en el curso y items anteriores será motivo de reducción de 0 pts a 5 pts, por ejercicio.

1. (25 pts) Cree una función que reciba una lista l con dimension n e imprima un diccionario con todas las combinaciones, de los valores de la lista, que cumple con ser una terna pitagórica, es decir datos a, b, c números reales verifique:

$$a^2 + b^2 = c^2$$

2. (50 pts) Tratamiento de datos nulos. Una de las tareas más importantes en el análisis de datos es el tratamiento que se le dé a los datos nulos (datos vacíos) por lo tanto es indispensable estudiar en el curso metodologías que permitan bajo un criterio experto tomar la mejor decisión con respecto a estos datos.

- Programe un módulo en python que permita al usuario realizar y decidir cuál de las siguientes metodologías usar para el tratamiento de datos faltantes:
  - Eliminar las observaciones de las columnas que contengan datos nulos. Las columnas pueden ser seleccionadas por el usuario.
  - Imputación de valores nulos con información general: este proceso sustituye el valor faltante (dada una columna) con algún dato estadístico (de la misma columna). Esta sustitución puede ser con: algún valor aleatorio (dentro del rango de las observaciones), valor promedio, mediana, valor máximo, mínimo, valor ingresado por el usuario (comúnmente es criterio experto).
  - Imputación por medio de agrupación: el usuario por medio una agrupación (una o unas variables categóricas) pueda imputar los valores nulos. Sugerencia: es lo mismo que el item

anterior solo que la información para la estimación del promedio, mediana, máximo o mínimo depende de la agrupación y de la columna dada.

- Imputación por promedio móvil: dada una columna de la tabla de datos se sustituye los valores faltantes por medio del promedio móvil. Ejemplo: $[1, 2, 3, 4, None, 5, 6, None, 7]$ se desea imputar el dato por medio de promedio móvil de $banda = 2$, entonces el promedio de los 4 valores (2 a la derecha y dos a la izquierda) de el valor nulo es $3 + 4 + 5 + 6 = 4.5$. Así, se tiene que $[1, 2, 3, 4, 4.5, 5, 6, None]$. En caso de salirse de los márgenes tome solo los valores que tiene, esto es si $banda = 3$ se realiza el promedio de $2, 3, 4, 5, 6, 7$ y en el caso del último valor nulo se tomaría el promedio $4, 5, 6, 7$ Este ejercicio debe programarse desde cero
- (Opcional 5 pts). Investigue qué librerías de python realizan lo anterior y si hay más métodos. Sea concreto.

3. (25 pts) Pruebe los métodos anteriores con los datos Base_salarios.csv, debe imputar los datos nulos que se presentan en la columna *Salario base*. Utilice material gráfico y tablas estadísticas para demostrar los cambios que se tiene a nivel de datos el utilizar un criterio de imputación de los datos.
    - ¿Cómo se determina una metodología razonable? ¿Cuál usaría usted y por qué?

Información adicional

- No es necesario la documentación de los métodos constructor, get y set.