FACULTAD DE



**CIENCIAS SOCIALES**

**Trabajo 2 – Computación en Paralelo**

Indicaciones generales:

* Este trabajo se realizará en **grupos de** **cuatro estudiantes**. Los grupos serán armados aleatoriamente utilizando las parejas formadas para la Tarea 2.
* Se entregará un solo archivo que contenga los códigos utilizados, resultados y respuestas escritas (cuando sea necesario). Para esto se utilizará un Jupyter Notebook. Estos archivos deben estar ordenados y comentados. Los comentarios deben indicar claramente cada sub-sección de la Tarea y explicar los pasos seguidos.
* Se valorará en la calificación el orden y claridad del código. Este debe ser replicable y llegar a los mismos resultados que se reporten como respuestas a las preguntas.
* Adicionalmente, se entregará el contrato de trabajo grupal que se encuentra al final del documento. Este debe ser firmado por todos los miembros del grupo.
* Para la parte práctica del trabajo deberá utilizarse la librería Dask.
* La fecha de entrega será el **sábado 02/07 hasta las 23:59**. A las entregas que se realicen tarde se les descontará un punto por cada hora transcurrida desde la hora de entrega. Luego de la quinta hora de retraso, ya no se recibirá el trabajo y se calificará con cero.

Indicaciones del Trabajo 2

**Parte I (8 puntos)**

Responda las siguientes preguntas respetando el límite máximo de palabras establecido

1. El equipo del QLAB, para el cual Ud. ahora trabaja, ha decidido comprar 100 tarjetas gráficas (GPUs) para PCs a fin de utilizarlas en computación en paralelo para investigación en Ciencias Sociales. Sabiendo de su experiencia en el tema, el director le ha pedido recomendaciones. Elabore una tabla de comparación de, por lo menos, tres tarjetas gráficas disponibles en Perú donde resuma a detalle las especificaciones de cada una de ellas. Finalmente, diga cuál y por qué le recomendaría al director. Recuerde que el precio es uno de los atributos a considerar. (200 palabras, 3 puntos) / Andrés
2. Revise el artículo de Couper (2013)[[1]](#footnote-1) y la entrevista a Ng (2022)[[2]](#footnote-2). Elabore un comentario informado y crítico con sus apreciaciones sobre “Good Data & Big Data” ¿uno reemplazará al otro y por qué? ¿se pueden complementar y cómo? Dé, por lo menos dos ejemplos/ideas acordes a su argumento (500 palabras, 4 puntos). /Adriana

1. Al trabajar la parte II de esta tarea seguramente encontrará varias dificultades de código o sustanciales (teoría ML). Seleccione la más interesante y que aún no haya podido resolver sin ayuda. Elabore un *post* sobre esta duda en alguno de los portales de preguntas (*StackOverflow, StackExchange, GitHub Issues*), y monitoree las respuestas. Asegúrese de publicar una pregunta no trivial, de leer el manual de etiqueta del portal[[3]](#footnote-3) y mantener el post autocontenido. No se olvide que otros usuarios están invirtiendo su tiempo libre en ayudarlo: interactúe respetuosamente con ellos, repregunte si es necesario y, de ser posible, incluya la solución en su código. Presente como respuesta el *link* al portal de su pregunta. Esta debe tener una antigüedad mayor a 48 horas respecto de la entrega de la tarea. (1 punto)

**Parte II (12 puntos)**

***Atención****: Su aplicación debe tener por lo menos dos variaciones de fondo respecto a los notebooks de ejemplo vistos en clase y del portal oficial de Dask. Podrá utilizar datos de cualquier fuente, pero para algún problema o pregunta ocurrido en el Perú.*

Para esta parte los grupos deberán realizar una aplicación de Machine Learning (ML) utilizando la librería Dask. El tema del ejercicio es libre, pero debe ser propio y de interés del grupo (la justificación del tema será evaluada).

Deberá presentar este ejercicio en un Jupyter o Colab Notebook (.ipynb) donde todas las celdas ya hayan sido ejecutadas; además del mismo documento en formato PDF. Este deberá incluir las siguientes partes:

1. (1 punto) Presentación del problema de predicción que se desea realizar (por ejemplo, predecir la vulnerabilidad de los hogares en el Perú). Defina claramente qué datos utilizara, cuál es su variable *target* y cuáles sus predictores. /R y AP
2. (1 puntos) Describa los pasos a realizar para su aplicación de ML. /R
3. (3 puntos) Describa cómo llevaría a cabo este ejercicio de manera serial y compárelo a su aplicación serial. Como parte de esta descripción, incluir los siguientes aspectos: /R y AP
   1. Explicar qué partes del ejercicio se harán de forma serial y por qué no paralelizo estas tareas.

En el caso del procesamiento de operaciones secuenciales estas deben realizarse de manera serial. Dentro del ejercicio planteado se propone identificar las variables que mejor predicen el precio de alquiler de la vivienda en Lima Metropolitana, para ello se utilizará la ENAHO 2020. Para tomar la muestra de la base de datos a utilizar se realizará una partición de la base, tomando únicamente la parte que corresponde a Lima Metropolitana.

* 1. Para las tareas en paralelo, explique usando el método de Foster como se dan las etapas de partición, comunicación, aglomeración y mapeo (PCAM) para su aplicación.

El método Foster es un enfoque exploratorio de diseño en el que se busca que distintas partes puedan ejecutarse de forma paralela sin seguir un orden y sin afectar el resultado final. En el caso de la aplicación de este al ejercicio.

* 1. Discuta qué tipos de procesadores podría utilizar para cada parte. (No es necesario que utilice los GPU pese a que señale su mejor desempeño)
  2. Identificar los cuellos de botella del ejercicio y comente hasta qué punto la paralelización puede ayudar a resolverlos.

1. (7 puntos) Uso de Dask.
   1. Ejercicio de ETL: Usar Dask Dataframes para cargar la(s) base(s) de datos que se utilizará(n) y presentar lo siguiente:

- Dataframe con DASK (limpiar NAs y sacar lo que no es ventas)

- Creación de por lo menos dos variables

Variable dependiente: precio de la vivienda (venta) -> la mayoría en Lima es propietario de su vivienda.

Independientes: características de la vivienda (baños, dormitorios, acceso a servicios básicos), calidad de materiales,

La variable dependiente a utilizar será el precio de alquiler. En la ENAHO esta variable (GA03HD - Gasto por alquiler) es recogida a partir de la declaración del encuestado de su gasto mensual por el alquiler de su vivienda.

Variables predictoras:

1. UBIGEO - Distrito
2. P22 - Además de esta vivienda, ¿Existe otra vivienda en la que usted o algún miembro de su hogar viven regularmente?
3. P101 - Tipo de vivienda
4. P24A - Las paredes exteriores de la vivienda tienen fachada: Con tarrajeo
5. P24B - Las paredes exteriores de la vivienda tienen fachada: Pintada
6. P25\_1 - La calle, av., jr., carretera, donde se ubica la vivienda tiene: Pista asfaltada
7. P25\_2 - La calle, av., jr., carretera, donde se ubica la vivienda tiene: Pista afirmada/tierra
8. P25\_3 - La calle, av., jr., carretera, donde se ubica la vivienda tiene: Veredas
9. P25\_4 - La calle, av., jr., carretera, donde se ubica la vivienda tiene: Poste alumbrado público
10. P25\_5 - La calle, av., jr., carretera, donde se ubica la vivienda tiene: Ninguno
11. P102A - ¿El material predominante en las paredes exteriores es?
12. P103 - ¿El material predominante en los pisos es?
13. P103A - ¿El material predominante en los techos es?
14. P104 - ¿Cuántas habitaciones en total tiene la vivienda?
15. P104A - ¿Cuántas habitaciones se usan exclusivamente para dormir?
16. P104B1 - ¿La vivienda cuenta con licencia de construcción
17. P104B2 - Fue construída con asistencia técnica de un ing. Civil o arquitecto?
18. P105A - ¿La vivienda que ocupa su hogar es? - Alquilada
19. P105B - Monto mensual por alquiler o compra de la vivienda (en S/.)
20. P106A - ¿Esta vivienda tiene título de propiedad?
21. P106B - ¿El título de la vivienda está Registrado en la SUNARP?
22. P110 - El abastecimiento de agua en su hogar procede de?
23. P110A1 - ¿El agua es potable?
24. P110A - Calidad bacteriológica del agua
25. P110A\_MODIFICADA - Nivel de Cloro Residual del agua (Transcripción de valores de kit comparador)
26. P110C - ¿El hogar tiene acceso al servicio de agua todos los días de la semana?
27. P110C1 - ¿Cuántas horas al día?
28. P110C2 - ¿Cuántos días a la semana tiene este servicio?
29. P110F - ¿Pagan por el servicio de agua?
30. P110G - ¿A qué empresa o entidad se paga por el servicio de agua?
31. P111A - El baño o servicio higiénico que tiene su hogar esta conectado a red pública/letrina, pozo, río, etc.
32. P1121 - ¿Cuál es el tipo de alumbrado que tiene su hogar? - Electricidad
33. P1123 - ¿Cuál es el tipo de alumbrado que tiene su hogar? - Petróleo/gas (lámpara)
34. P1124 - ¿Cuál es el tipo de alumbrado que tiene su hogar? - Vela
35. P1125 - ¿Cuál es el tipo de alumbrado que tiene su hogar? - Generador
36. P1126 - ¿Cuál es el tipo de alumbrado que tiene su hogar? - Otro
37. P1127 - ¿Cuál es el tipo de alumbrado que tiene su hogar? - No utiliza
38. P112A - ¿El servicio eléctrico de su hogar es: medidor individual/colectivo?
39. P113A - Combustible que usan en el hogar para cocinar: Mayor frecuencia
40. P1141 - Su hogar tiene: ¿Teléfono (fijo)?
41. P1142 - Su hogar tiene: ¿Teléfono Celular?
42. P1143 - Su hogar tiene: Conexión a TV cable o satelital?
43. P1144 - Su hogar tiene: ¿Conexión a Internet?
44. P1145 - Su hogar tiene: NO TIENE
45. P1173$01-P1173$16 - Último gasto mensual (16 variables)
46. P1175$01 - El último gasto mensual por consumo de Agua fue incluido en el alquiler/no gastó
47. P1175$02 -  El último gasto mensual por consumo de Electricidad fue incluido en el alquiler/no gastó
48. P1175$03-P1175$16 - Último gasto mensual (14 variables)

- Por lo menos dos estadísticos descriptivos que vayan en línea con el tema y argumento. Explíquelos

- Por lo menos dos gráficos que vayan en línea con el tema y argumento.

Tanto los gráficos como los descriptivos deben estar en calidad para ser incluidos en un reporte final. Se descontarán puntos por presentación descuidada.

* 1. Implementación de Machine Learning: Utilice Dask para entrenar por lo menos un modelo de Machine Learning supervisado. Este acápite debe contener, por lo menos, los siguientes elementos:
     + Definición de predictores (X) y vector de *target* (y)
     + *Train-test splitting*
     + *Cross-Validation*
     + *GridSearch*
     + *Model fit* y selección del modelo óptimo
     + Computo de dos indicadores de la calidad de ajuste en muestra entrenamiento
     + Computo de dos indicadores de la calidad de ajuste fuera de la muestra (*test*)
     + Limitaciones y posibles extensiones

Recomendación: no se evaluará la complejidad de modelo de ML al no ser la intención del curso. Lo crucial es demostrar el dominio del uso de Dask para seguir todos los pasos de un *pipeline* de ML. Se recomienda el uso de *dask\_ml.* Si deciden utilizar datos provenientes de encuestas complejas, no es necesario que considere el diseño ni los pesos muestras en ninguno de los pasos a fin de no complejizar el ejercicio.

**Contrato de grupo de trabajo**

**Nombres completos de cada miembro:**

1. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
4. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Objetivos y fechas**

Rellenar con las tareas que consideren necesarias para completar el trabajo grupal. Asignar un encargado/a, fecha de entrega y si se cumplió con la tarea asignada.

|  |  |  |  |
| --- | --- | --- | --- |
| **Tareas** | **Encargado/a** | **Fecha de entrega** | **¿Se cumplió?** |
| 1. Parte I, Pregunta 1 | Andrés | 29/06/22 |  |
| 2. Parte I, Pregunta 2 | Adriana | 29/06/22 |  |
| 3. Parte II, Pregunta 1 | Raysa | 29/06/22 |  |
| 4. Parte II, Pregunta 3 | Ana Paula | 29/06/22 |  |
| 5. Parte II, Pregunta 2 | Andrés | 29/06/22 |  |
| 6. |  |  |  |

**Sobre la relación entre miembros del grupo**

Normas grupales: Consideramos importantes las siguientes actitudes y comportamientos y nos esforzaremos en mantenerlas a lo largo del trabajo grupal.

|  |
| --- |
|  |

Toma de decisiones: Tomaremos decisiones de la siguiente manera:

|  |
| --- |
|  |

**Reglas de comunicación**

Los siguientes medios serán utilizados para coordinar los aspectos relacionados al trabajo grupal. Si el grupo tendrá reuniones de grupo (presenciales o virtuales), especificar las fechas de reunión.

|  |
| --- |
|  |

**Resolución de conflictos**

Especificar cómo el grupo lidiará con los conflictos que pueda haber o situaciones en los que no se cumpla este contrato.

|  |
| --- |
|  |

**Firmas**

1. Ana Paula Carrillo Miranda
2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
4. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Al escribir nuestros nombres completos, aceptamos la responsabilidad de completar las tareas del trabajo grupal.

1. Couper, M. P. (2013, December). Is the sky falling? New technology, changing media, and the future of surveys. In Survey Research Methods (Vol. 7, No. 3, pp. 145-156). [↑](#footnote-ref-1)
2. https://spectrum.ieee.org/andrew-ng-data-centric-ai [↑](#footnote-ref-2)
3. https://stackoverflow.com/help/how-to-ask [↑](#footnote-ref-3)