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 domingo 17/09 hasta las 11:59am**. 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. Explique, de manera sencilla, cómo el algoritmo de Dask se relaciona con el método de Foster. Agregue ejemplos. (150 palabras, 1 puntos)
2. No todo se debe paralelizar. ¿Bajo qué circuntancias la paralelización podría ser ineficiente? Provea al menos dos ejemplos. (150 palabras, 1 puntos)
3. Revise el artículo de Desouza y Smith (2014) en Stanford Social Innovation Review[[1]](#footnote-1) y la presentación de Ng en el Data+AI Summit 2022[[2]](#footnote-2). Elabore un comentario informado y crítico sobre “Good Data & Big Data”. ¿uno reemplazará al otro y por qué? ¿se pueden complementar y cómo? Agregue al menos dos ejemplos/ideas acordes a su argumento relacionados al caso peruano (400 palabras, 4 puntos).
4. Imagina que estás trabajando en un proyecto de análisis de datos para una empresa que maneja una gran cantidad de registros de ventas en línea. Explica cómo Dask podría aplicarse de manera específica para acelerar el procesamiento y análisis de estos registros. Proporciona un ejemplo detallado de cómo Dask podría utilizarse para realizar una tarea específica en este contexto y menciona las ventajas que ofrece Dask en términos de paralelización y escalabilidad en comparación con enfoques tradicionales. (300 palabras, 2 puntos)

**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 y relevancia 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.
2. (1 puntos) Describa los pasos a realizar para su aplicación de ML.
3. (3 puntos) Describa cómo llevaría a cabo este ejercicio de manera paralelizada y compárelo a su aplicación serial. Como parte de esta descripción, incluir los siguientes aspectos:
   1. Explicar qué partes del ejercicio se harán de forma serial y por qué no paralelizo estas tareas.
   2. 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.
   3. Discuta qué tipos de procesadores podría utilizar para cada parte. (No es necesario que utilice los GPU pese a que señale su mejor desempeño)
   4. Identificar los cuellos de botella del ejercicio y comente hasta qué punto la paralelización puede ayudar a resolverlos.
4. (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:

- Creación de por lo menos dos 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

|  |
| --- |
| **Recomendaciones**   * No se evaluará la complejidad de modelo de ML, al no ser la intención del curso. En realidad, lo importante es demostrar el dominio del uso de Dask para seguir todos los pasos de un *pipeline* de ML con paralelización. * 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. Rodrigo Soto
2. Gabriela Calvo
3. Cesar Nuñez
4. Angel Ibañez

**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 1 – pregunta 1 al 3 | Rodrigo Soto | 17-09-2023 | Sí |
| 2. Parte 1 – pregunta 4 / parte 2 pregunta 1 a 3 | Gabriela Calvo | 17-09-2023 | Sí |
| 3. Parte 2 pregunta 4a | Cesar Nuñez | 17-09-2023 | Sí |
| 4. Parte 2 pregunta 4b | Angel Ibañez | 17-09-2023 | Sí |
| Añadir según sea necesario |  |  |  |

**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.

|  |
| --- |
| comunicar |

Toma de decisiones: Tomaremos decisiones de la siguiente manera:

|  |
| --- |
| Coordinación whatsapp |

**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.

|  |
| --- |
| Por whatsapp |

**Resolución de conflictos**

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

|  |
| --- |
| Por whatsapp |

**Firmas**

1. Rodrigo Soto
2. Gabriela Calvo
3. Cesar Nuñez
4. Angel Ibañez

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

1. https://ssir.org/articles/entry/big\_data\_for\_social\_innovation [↑](#footnote-ref-1)
2. https://www.youtube.com/watch?v=avoijDORAlc [↑](#footnote-ref-2)