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

El rápido crecimiento urbano de Lima Metropolitana se ha regido bajo dos lógicas: la producción de la vivienda informal por parte del Estado y la poca masificación del mercado de la vivienda formal, lo que ha permitido la generación y la habitabilidad de viviendas indignas.

Durante el gobierno de Alberto Fujimori, se realizó una política de titulación, en la que se realizó una masiva entrega de títulos de propiedad, alrededor de 1 millón (Calderón, 2009). Sumado a ello, se consolidaron 400 mil lotes de manera informal (Calderón, 2009). De esta forma, Lima empezó a crecer hacia afuera, formándose así periferias en donde había viviendas precarizadas, un limitado acceso a los servicios básicos y de equipamiento urbano.

Como resultado, Lima Metropolitana, al igual que otras ciudades latinoamericanas, se ha caracterizado por tener una alta segregación urbana. El modelo centro-periferia ha estado marcado por “la producción de territorios diferenciales que consolidan formas de vida antitéticas: por un lado, la segregación auto-inducida de los sectores de más altos ingresos y, por el otro, la segregación estructural (por expulsión) de los pobres urbanos” (Cervio, 2015).

Las políticas sobre el desarrollo urbano y más aún el de las viviendas ha sido todo un reto para los gobiernos de turno y las municipalidades, quienes no han tenido un rol definido sobre dichas políticas. De hecho, con los ajustes políticos y económicos realizados en la década de 1990, el mercado inmobiliario empezó a cobrar mayor protagonismo.

El acceso a la vivienda se dejó en manos del mercado inmobiliario, pues se pensaba que este arreglaría el problema de la vivienda. De esta forma, el mercado se ha constituido como “mecanismo principal de coordinación de la producción de la ciudad, ya sea a través de la privatización de las empresas públicas o por la hegemonía del capital privado en la producción de las materialidades residenciales y comerciales urbanas” (Abramo, 2012, p. 36).

En la actualidad nos encontramos en un mundo más urbanizado y el mercado inmobiliario se ha caracterizado por ser un sector muy dinámico en los últimos años, en el que los precios de los bienes e inmuebles están en constante cambio. De esta manera, este sector constituye un motor de la economía.

En el proceso de toma de decisiones en el sector de vivienda, la valuación de los bienes y la detección de sobreprecios (Zhu, 2014; Mooya, 2016; Fischer, 2017) resulta importante, dado que permite saber la variación de precios de una vivienda en determinado lugar para quienes están interesados en comprar o vender una vivienda.

La vivienda cumple la doble función de ser un bien de inversión, el cual genera rentabilidad a largo plazo, además de ser una fuente de utilidad para quienes lo usan (Grajales, 2019, p.16). Asimismo, es considerado un “bien preferente” (Fischer, 2017), debido a que, las personas buscan acceder a una vivienda con determinadas características que pueda satisfacer sus necesidades. En ese sentido, la valuación de este bien heterogéneo es determinada por sus características observables.

Por un lado, tanto la tendencia de los precios como las características de la vivienda nos brinda información sobre las condiciones de una vivienda en relación a su precio. Por otro lado, tal como lo señala el objetivo de desarrollo sostenible N° 11: ciudades y comunidades sostenibles, podemos acceder y vivir en una vivienda digna, teniendo en cuenta nuestras necesidades, la infraestructura y servicios adecuados.

Si bien la pandemia golpeó a todos los sectores económicos y productivos, el mercado de alquiler y venta ha sido uno de los sectores que ha tenido mayor dinamismo, a pesar de la crisis sanitaria. La recuperación de este sector se empezó a dar a partir del segundo trimestre del 2020 (SEIA, 2020).

Una de las principales características de la oferta inmobiliaria es que el 91% se concentra en departamentos y el 9% en casas (Properati, 2020). Asimismo, el 44% se enfoca en viviendas de 3 habitaciones, mientras que el 32% en 2 (Properati, 2020).

Los distritos que tienen mayor demanda son Santiago de Surco (11%), Miraflores, Los Olivos (9%), Cercado de Lima (7%) y La Molina (7%). Además, las áreas de las viviendas que tienen mayor demanda son entre 70 m² y 90 m² (Properati, 2020).

El precio de alquiler varía según cada distrito. Por ejemplo, en distritos como San Isidro y Barranco, el precio del m² es de U$ 10,50 aproximadamente. En el caso de Jesús María y Lince se registra un valor de alrededor de U$ 8,80 el m² (La Cámara de Comercio, 2020).

En el caso de Lima Centro, el precio de alquiler varía entre U$ 6,50 y U$ 7,60 el m² como en los distritos de La Victoria, Cercado de Lima y Breña. En Lima Este, Ate tiene un valor de U$ 5,36. Por último, Lima Norte tiene el valor más económico. Es decir, distritos como San Martín de Porres y Comas varían entre U$ 3 y U$ 4 (La Cámara de Comercio, 2020).

En ese sentido, cabe analizar cómo varía el precio de las viviendas en base a sus características en las zonas urbanas. La presente propuesta busca responder la siguiente pregunta: ¿Cuál será la tendencia en el precio de venta de la vivienda en Lima Metropolitana? Para ello, se utilizará la base de Properati, la cual contiene información sobre las propiedades que están y estuvieron publicadas en Latinoamérica.

Se hará uso de las herramientas de machine learning o inteligencia artificial. Se propone realizar los modelos de Random Forest, Lasso y Ridge, dado que se busca predecir el precio de venta de la vivienda en Lima Metropolitana. De esta forma, se tiene como primer objetivo: identificar las características de la vivienda que mejor predicen el precio de venta de la vivienda en Lima Metropolitana. En segundo lugar, se pretende encontrar el mejor método que predice el precio de venta de la vivienda en Lima Metropolitana.

1. (1 puntos) Describa los pasos a realizar para su aplicación de ML. /R
2. (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,

- 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)