Trabajo Sistemas Distribuidos
Prediciendo el Futuro

1) **Descripción de la implementación, discutiendo cómo se ha distribuido y paralelizado el problema: qué tareas se identifican como paralelizadles y/o distribuibles, qué hacen los procesos generados o qué hacen los hilos, entre otras cuestiones.**

A continuación, nos encontramos ante un problema de predicción de series temporales. En este intentaremos predecir los valores que resultarán en un futuro en base a unas series temporales pasadas proporcionadas. Para hacerlo, vamos a usar un algoritmo especial llamado k-NN que busca días similares en el pasado para buscar este futuro predicho.
Las series temporales son una lista de datos que van en orden de tiempo, como las temperaturas cada día o el peso de una persona a lo largo del tiempo.

Vamos a usar un método llamado k-NN, este busca días similares en el pasado y usa lo que pasó después de estos para aplicarlo a el presente actual y predecir lo que va a pasar en el futuro. Por ejemplo, si hoy es un día soleado con temperatura X y queremos saber cómo será mañana, el algoritmo mirará días pasados que también fueron soleados con temperatura X y qué pasó después en esos días.
Para la medición del error existen muchas diferentes métricas; en este caso lo mediremos con:
El Error Porcentual Absoluto Medio (MAPE):

$$MAPE\ (\%) = \frac{100}{h}\sum_{n=1}^{h}\left(\frac{|R_n - P_n|}{|R_n|}\right)$$

Donde h es el horizonte de predicción, Rn el valor real cogido y Pn el valor predicho.

Usaremos diferentes conjuntos de datos, de los cuales desconocemos si son registros meteorológicos o datos de salud. Con estos probaremos nuestro algoritmo y veremos qué tan bien funciona en los diferentes casos. Cada conjunto incluye diferentes cantidades de datos y veremos cuánto tiempo tarda en procesarlos.
Dicho todo esto, una vez comprendido y construido la estructura del código tendremos que paralelizar y distribuir este para ver como se ve afectado por el número de procesadores e hilos que usamos. Para esto haremos uso de las librerías Open-MP y MPI que serán importadas en el código y utilizadas.

Empezamos por inicializar el entorno, recoger argumentos.
Después leer y obtener los datos de los archivos, desarrollar las funciones de cálculo y medición.
Distribución y paralelización:
Hemos utilizado la paralelización para predecir todas las filas para el conjunto de datos seleccionado (archivo de datos).
Hemos distribuido las filas entre el número de procesos seleccionados en el terminal. Cada hilo calcula sus predicciones locales y la medición del error de las mismas.

Finalmente hemos analizado la escalabilidad con gráficos, utilizando los conocimientos de Python de TSI.

2) **Análisis y discusión de tiempos de ejecución. Para ello, se pide estudiar tres dimensiones distintas: tamaño de los datos, número de procesos y número de hilos por proceso.**

   a. **Para el tamaño de los datos, utilice los ficheros de tamaño 1x, 10x y 100x que se encuentran en la plataforma virtual. La evaluación se realizará siempre sobre las últimas 1000 filas de cada fichero, esto es, se realizarán 1000 predicciones correspondientes a las últimas 1000 líneas de los ficheros de entrada.**

   b. **Para cada uno de los tres ficheros, mida los tiempos para 1, 2, 3 y 4 procesos.**

      El programa que hemos desarrollado solo funciona con datos_1X.txt, ya que en los demás archivos da un error de malloc(): corrupted top size, que suponemos que es por exceso de uso de memoria.

   c. **Para cada proceso, utilice y mida los tiempos para 1, 2, 3 y 4 hilos, respectivamente.**

1 PROCESOS:				
	1 HILO	**2 HILOS**	**3 HILOS**	**4 HILOS**
1 X	20,84 s	20,92 s	20,38 s	20,98 s
10 X				
100 X				

2 PROCESOS:				
	1 HILO	**2 HILOS**	**3 HILOS**	**4 HILOS**
1 X	12,44 s	12,45 s	12,38 s	12,42 s
10 X				
100 X				

3 PROCESOS:				
	1 HILO	**2 HILOS**	**3 HILOS**	**4 HILOS**
1 X	2,25 s	2,29 s	2,36 s	2,21 s
10 X				
100 X				

4 PROCESOS:				
	1 HILO	**2 HILOS**	**3 HILOS**	**4 HILOS**
1 X	1,51 s	1,48 s	1,54 s	1,41 s
10 X				
100 X				

3) Descripción hardware de los equipos utilizados para el desarrollo del trabajo.

Hemos utilizado los 4 equipos de cada componente del grupo para desarrollar el código entre todos.

Hemos compilado y hecho las mediciones en el equipo de José Luis Leal ya que era el que más procesadores tenía.

Este equipo tiene las siguientes características:

Nombre del dispositivo: HP 15s-fq1xxx
Procesador: Intel(R) Core (TM) i5-1035G1 CPU @ 1.00GHz
RAM instalada: 8,0 GB (7,4 GB usable)
Tipo de sistema: Sistema operativo de 64 bits, procesador basado en x64