HW 3 Report

I'm using Spark with Scala to implement the KNN algirithm

Small Dataset:

- Partition Train dataset to 1
- Partition Test dataset to 1

Medium Dataset:

- Partition Train dataset to 8
- Partition Test dataset to 1

Large Dataset:

- Partition Train dataset to 48
- Partition Test dataset to 1

	Spark	**Threaded Best Time**	**OpenMP Best Time**	**MPI Best Time**	**CUDA Best Time**
small	2,940.717	15.880	6.018	7.469	3.742
medium	6,596.396	152.952	196.287	126.812	6.853
large	69,068.247	924.718	988.412	937.283	19.775

As observed from the runtime, Spark may not be the best option for relatively small datasets. However, its ability to partition datasets and perform distributed computations makes it a better choice for handling large datasets that cannot fit into the memory of a single machine.