K Nearest Neighbour implementation on GPU and CPU.

Done by :- Mr Saurav Chakraborty.
Roll No :- CS19S015.

K-NN Implementation.

K nearest neighbour model is used in Classification and regression problem .In this project I am using KNN as a classifier.Where all the feature vector belongs to certain distinct classes .So based on a query vector it predicts in which class it may belong.

K NN predicts a suitable class for a query points by calculating its k nearest neighbour in the training set and then among the K nearest neighbour in which class they belong most.

K NN is a lazy learner for a given point we have to iterate through the whole dataset.

There is no pre processing in KNN.

There are some heuristics which partition the data set so that we have to search on less no of points

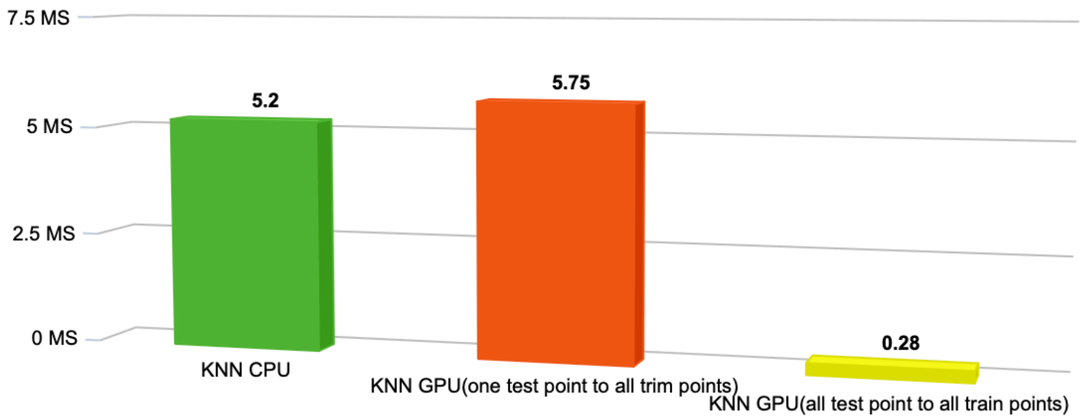
Like KD tree , Ball tree , LSH .

In this project I implemented the basic KNN on both GPU and CPU.

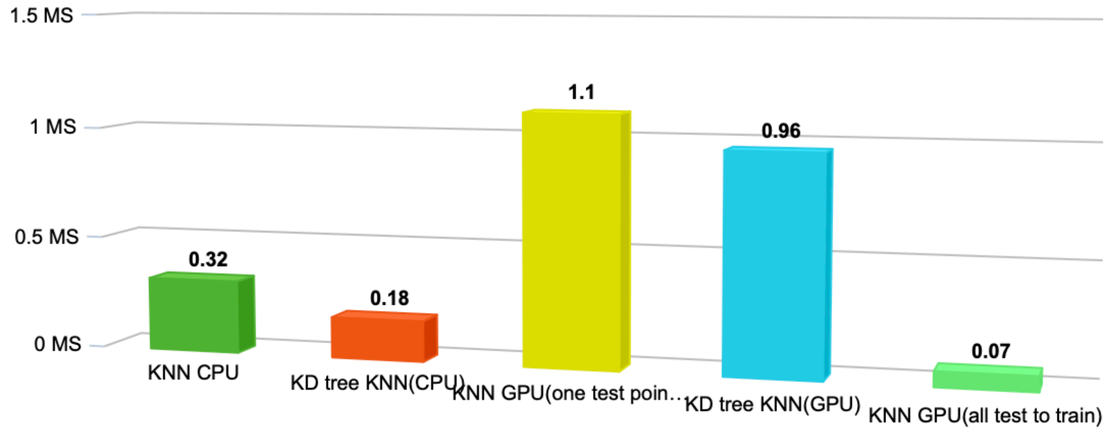
I also implemented the the KD tree approximation on both GPU and CPU.

- KNN Model is Coded on two data sets taken from UCI Website.
- Data sets :- 1. Iris Data Set :- https://archive.ics.uci.edu/ml/datasets/iris
- 2. Breast Cancer Data set :-https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

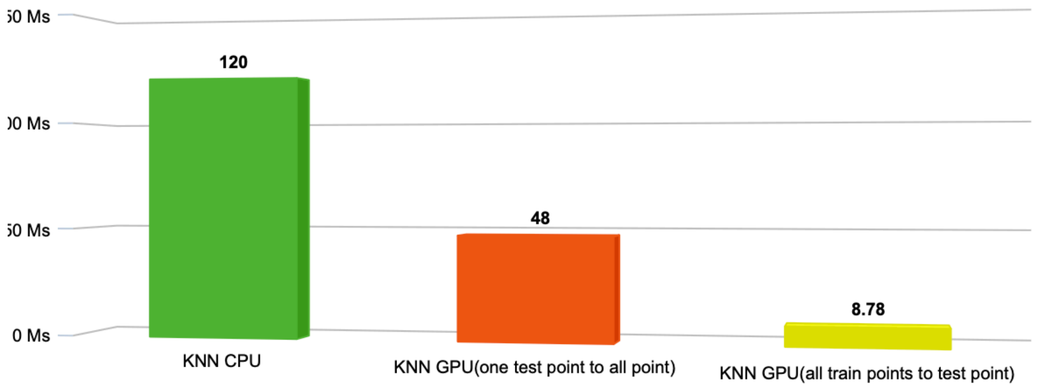
Results.



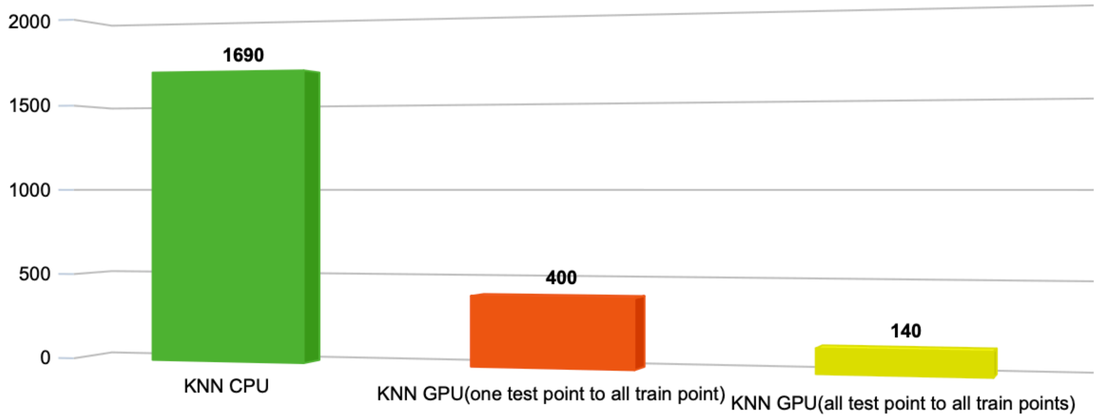
Breast cancer Data set



IRIS DATA SET



Randomly generated dataset(10000 train points)



Randomly generated dataset(160000 train points)

Challenges .

- KD tree for KNN I implemented it in a Hash-map way For implementing that It took some time.
- Used thrust library function to sort the data in some implantation of KNN which I saw have used their own sort function. It makes the job done with just one line of code.
- Spend time in searching data set suitable for my implementation. Most of the data set do not contain explanation about the data they provide and most of them are unsupervised.
- I need a supervised set so I used a IRIS data set ,breast cancer data set.
- I have also test the performance on randomly generated data having simillar attribute as breast cancer dataset having 10000 rows.