# Weekly Report 3 - XGBoost

Ganji Varshitha
AI20BTECH11009

---

## Introduction

XGBoost stands for Extreme Gradient Boosting algorithm. It is an enhanced GBM model. It improves the performance and speed of model. It provides parallel tree boosting and works very fast than GBMs.
It is a supervised learning ensemble model.

## Algorithm

The GBM algorithm is modified with few regularization factors in XGBoost algorithm.

**Algorithm 1:** XGboost algorithm

**Data:** Dataset and hyperparameters
Initialize $f_0(x)$;
**for** $k = 1, 2, \ldots, M$ **do**
  Calculate $g_k = \frac{\partial L(y,f)}{\partial f}$;
  Calculate $h_k = \frac{\partial^2 L(y,f)}{\partial f^2}$;
  Determine the structure by choosing splits with maximized gain
  $\mathbf{A} = \frac{1}{2}\left[\frac{G_L^2}{H_L} + \frac{G_R^2}{H_R} - \frac{G^2}{H}\right]$;
  Determine the leaf weights $w^* = -\frac{G}{H}$;
  Determine the base learner $\hat{b}(x) = \sum_{j=1}^{T} wI$;
  Add trees $f_k(x) = f_{k-1}(x) + \hat{b}(x)$;
**end**
**Result:** $f(x) = \sum_{k=0}^{M} f_k(x)$

## Optimizations

### System optimizations

- Parallelization: XGboost can take use of multiple cores of CPU. Data stores in memory units called blocks which can be reused in the algorithm.
- Distributed Computing for training very large models using a cluster of machines.
- Out-of-Core Computing for very large datasets that don't fit into memory.
- Hardware Optimisation: Cache Optimization of data structures and algorithm to make best use of hardware.

### Algorithm features

- Regularization: It uses LASSO(L1) and ridge(L2) regularization. This prevents overfitting.
- Cross validation: The algorithm has inbuilt cross validation at every iteration of training.
- Tree pruning: It follows depth first search approach. The algorithm prunes the tree in backward direction.

## Questions