# Lab Experiment: 6
# Birch Clustering Algorithm

Course: Machine Learning Lab

September 10, 2025

## Aim

Implement Birch from scratch with your own function and compare the result with SKlearn Birch clustering

## Theory: BIRCH Clustering Algorithm

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is a hierarchical clustering method designed for large-scale datasets. Instead of processing all data points directly, BIRCH incrementally builds a compact tree structure called the **Clustering Feature Tree (CF Tree)** that summarizes the dataset efficiently.

### Clustering Feature (CF)

Each cluster in BIRCH is represented by a **Clustering Feature (CF)**, defined as:

$$CF = (N, \mathbf{LS}, \mathbf{SS})$$

where:

- $N$ = number of data points in the cluster
- $\mathbf{LS} = \sum_{i=1}^{N} \mathbf{x}_i$ = linear sum of data points
- $\mathbf{SS} = \sum_{i=1}^{N} \mathbf{x}_i^2$ = squared sum of data points

From these values, cluster statistics can be computed:

$$\mu = \frac{\mathbf{LS}}{N} \quad \text{(Centroid)},$$

$$R = \sqrt{\frac{\mathbf{SS}}{N} - \mu^2} \quad \text{(Radius)}.$$

### CF Tree Structure

A CF Tree is a height-balanced tree that stores hierarchical clustering information:

- **Branching factor** ($B$): maximum number of entries per node.
- **Threshold** ($T$): maximum cluster radius allowed for each entry.

Internal nodes store summaries of their children, while leaf nodes contain CF entries representing clusters.

### Insertion Process

When inserting a new point:

1. Descend the tree by choosing the closest child (based on centroid distance).
2. At the leaf, attempt to merge the point into the nearest CF entry.
   - If the updated radius $\leq T$, merge into the CF.
   - Otherwise, create a new CF entry.
3. If the leaf exceeds $B$ entries, split the node using farthest-pair seeds and redistribute entries.
4. If the root splits, the tree height increases.

Thus, the CF Tree provides a compact, incremental representation of clusters and allows efficient clustering on large datasets.

## Dataset

Use the **titanic** dataset `https://shorturl.at/Be7w7`

## Tasks

1. Import required libraries.
2. Load the Titanic dataset
3. Select features ['pclass', 'sex', 'fare', 'embarked'] and the target survived.
4. Fill missing values in 'embarked' and 'fare'.
5. Encode categorical features 'sex' and 'embarked' into numerical values.
6. Scale the features using StandardScaler
7. show the Dendrogram
8. Perform BIRCH clustering from scratch on the scaled data.
9. Add new columns for true survival and predicted cluster labels.
10. Map predicted clusters to true labels for comparison
11. Show and calculate the confusion matrix, accuracy score and classification report
12. Generate scatter plots to visualize predictions against true survival labels.
13. Print a textual or visual representation of the BIRCH tree structure.
14. Compare your result with sklearn birch clustering

## Results (to be filled by students)

- Custom Birch clustering accuracy: _____
- Sklearn Birch clustering accuracy: _____

## Submission

Submit: A short report (2-3 pages), which contains an introduction about Birch clustering techniques, an algorithm, a code link, a result screenshot, and a conclusion about the result and model.

Upload link: `https://forms.gle/sukSvLUud4GC1U789`