**Summer Analytics First Hackathon 2025**

**Overview**

Welcome to First course hackathon of Summer Analytics 2025.  
Hosted by **Consulting & Analytics Club and GeeksforGeeks (GFG)**  
Classify land cover types using NDVI time-series data from satellite imagery and OpenStreetMap (OSM) labels. Your challenge is to build a **Logistic Regression model** that accurately predicts land cover classes despite noisy NDVI signals. Top performers win **GFG Premium memberships**, and all participants get exclusive discounts!

**Description**

Hackathon Problem Statement: NDVI-based Land Cover Classification  
**Key Concepts**

1. NDVI (Normalized Difference Vegetation Index)  
   Measures vegetation health using satellite data:

NDVI=NIR−REDNIR+RED

Where:-

* **NIR** = Near-Infrared reflectance
* **Red** = Red reflectance

**2. Data Challenges**  
**Noise:** The main challenge with the dataset is that both the imagery and the crowdsourced data contain noise (due to cloud cover in the images and inaccurate labeling/digitizing of polygons).

**Missing Data:** Certain NDVI values are missing because of cloud cover obstructing the satellite view.

**Temporal Variations:** NDVI values vary seasonally, requiring careful feature engineering to extract meaningful trends.

**Important Note:**  
The training and public leaderboard test data may contain noisy observations, while the private leaderboard data is clean and free of noise. This design helps evaluate how well your model generalizes beyond noisy training conditions.

**Dataset**  
Each row in the dataset contains:

**class:** Ground truth label of the land cover type — one of {Water, Impervious, Farm, Forest, Grass, Orchard}

**ID:**Unique identifier for the sample

**27 NDVI Time Points:** Columns labeled in the format YYYYMMDD\_N (e.g., 20150720\_N, 20150602\_N) represent NDVI values collected on different dates. These values form a time series representing vegetation dynamics for each location.

**Rules**  
Model: Logistic Regression only (multiclass).

Preprocessing: Denoising, imputation, and feature engineering allowed.

**Leaderboard:**

Public (89% test data): Immediate feedback.

Private (11% test data): Final ranking (avoids overfitting).

**Evaluation**

Submissions will be evaluated on basis of accuracy score of the predicted class.

**Submission format:**  
ID,class  
1,water  
2,water  
3,grass  
4,impervious