# InchTech's Intern
# Data Science Project

**Project Description**
Smallholder farmers are crucial contributors to global food production, and in India often suffer most from poverty and malnutrition. These farmers face challenges such as limited access to modern agriculture, unpredictable weather, and resource constraints. To tackle this issue, Digital Green collected data via surveys, offering insights into farming practices, environmental conditions, and crop yields.

The objective of this project is to create a machine learning solution to predict the crop yield per acre of rice or wheat crops in India. Our goal is to empower these farmers and break the cycle of poverty and malnutrition.

A crop yield model could revolutionise Indian agriculture, and serve as a global model for smallholder farmers. Accurate yield predictions empower smallholder farmers to make informed planting and resource allocation decisions, reducing poverty and malnutrition and improving food security. As climate change intensifies, adaptive farming practices become crucial, making precise yield predictions even more valuable. Solutions developed here can drive sustainable agriculture and ensure a stable food supply for the world's growing population. This challenge offers data scientists and machine learning enthusiasts a unique chance to make a real difference in vulnerable populations' lives while advancing global food security in a concise, impactful way.

**About Digital Green (digitalgreen.org)**
Digital Green is a global development organization that empowers smallholder farmers to lift themselves out of poverty by harnessing the collective power of technology and grassroots-level partnerships.

**Project Evaluation**
The evaluation metric for this competition is **Root Mean Squared Error**.
For every row in the dataset, submission files should contain 2 columns: ID and Target.
Your submission file should look like this (numbers to show format only):

| ID | Yield |
|---|---|
| ID_F9XXEXN2ADR2 | 437.9 |
| ID_SO3VW2X4QO93 | 100.56 |

**About the dataset**
The data was collected through a survey conducted across multiple districts in India. It consists of a variety of factors that could potentially impact the yield of rice crops. These factors include things like the type and amount of fertilizers used, the quantity of seedlings planted, methods of preparing the land, different irrigation techniques employed, among other features. The dataset comprises more than 5000 data points, each having more than 40 features.

For evaluating the performance of predictive models, the dataset has been split into a training set and a test set. About 25% of the data is reserved for testing. Within this test set, there is a further division into a public subset and a private subset. This split follows a 25:75 ratio, allowing for both a preliminary assessment of model performance and a later, more comprehensive evaluation.