**Week 1:**

**Problem Statement:**

Develop a CNN-based model capable of detecting and classifying plant diseases from images of leaves of various crops such as apples, cherry, grapes, and corn. The model should accurately identify both healthy and diseased leaves while predicting the specific type of disease. This system will aid in precision agriculture by enabling early detection and effective disease management.

**1.Defining the problem**

**What is the problem?**

Plant disease detection is a vital problem to address due to its significant impact on global food security, agricultural economies, and the environment. Plant diseases cause substantial crop losses, threatening the livelihoods of farmers and the stability of food supplies worldwide. Early and accurate detection minimizes these losses, reduces the overuse of harmful chemicals, and promotes sustainable farming practices. With advancements in AI and computer vision, it is now feasible to develop scalable, cost-effective solutions that empower farmers through easy-to-use tools like mobile apps. This interdisciplinary approach bridges agriculture and technology, enabling targeted interventions that benefit both small-scale and large-scale farming operations. Tackling this problem not only improves crop yields but also contributes to economic growth, environmental conservation, and rural development, making it a highly impactful and scalable solution for global challenges.

**How to eradicate the problem?**

Eradicating plant diseases through deep learning algorithms, particularly Convolutional Neural Networks (CNNs), offers a powerful and precise solution. CNNs excel in image recognition, making them ideal for detecting plant diseases from leaf images or other visual crop data. By training CNN models on large datasets of diseased and healthy plant images, the algorithm can accurately identify diseases at early stages. These models can be integrated into mobile applications, enabling farmers to diagnose diseases instantly using smartphone cameras. Coupled with IoT devices and drones for large-scale monitoring, CNN-based systems can provide real-time insights and targeted intervention strategies. The scalability and precision of CNNs, combined with advancements in hardware and data collection, make them a transformative tool for sustainable agriculture, reducing crop losses and promoting effective disease management.

**Objective:**

The objective of this project are:

**Step 1:** Data preparation

* Collect the data which includes the disease plant and also healthy plants in the form of images.
* Preprocess and clean the data
* Label them to their corresponding species and type of the plant.

**Step 2:** Exploratory Data Analysis

Exploratory Data Analysis (EDA) for plant disease detection involves analyzing and visualizing image datasets of healthy and diseased plants to identify patterns, variations, and features crucial for model training and accurate classification.

**Step 3:** Feature Engineering

Feature engineering for plant disease detection involves extracting meaningful features from image data, such as color histograms, texture descriptors (e.g., GLCM), edge detection (using Sobel or Canny filters), and shape attributes, to highlight disease-specific patterns and improve the performance of machine learning models like CNNs.

**Step 4:** Model Development

Model development for plant disease detection using CNN involves designing, training, and optimizing a deep learning model to classify diseases from preprocessed image datasets accurately.

**Step 5:** Model Evaluation

Model evaluation involves assessing the CNN's performance using metrics like accuracy, precision, recall, F1-score, and confusion matrix on validation and test datasets.

**Dataset:**

This dataset consists of about 87K rgb images of healthy and diseased crop leaves which is categorized into 38 different classes. The total dataset is divided into 80/20 ratio of training and validation set preserving the directory structure. A new directory containing 33 test images is created later for prediction purpose.

Data Source:<https://www.kaggle.com/datasets/vipoooool/new-plant-diseasesdataset/data?select=New+Plant+Diseases+Dataset%28Augmented%29>

**Tools we are used:**

1. Google Colab

2. Framework - Streamlit

**2. Data preprocessing**

Data must be preprocessed by using python module called Pandas Dataframe(pd). To preprocess and clean the data first we must check any null values. If there are no null values we must go to further step, else we must remove null values from the dataset. Data preprocessing is a crucial step in any machine learning or data analysis pipeline, and it begins with cleaning the dataset to ensure the quality of the data. To accomplish this, we utilize the Pandas DataFrame module in Python, often imported as pd, which provides powerful and flexible tools for handling structured data. The first step in this process is to check the dataset for any null or missing values, as these can negatively impact the accuracy and reliability of the analysis or model. If no null values are found during this inspection, we can proceed directly to the subsequent steps in the pipeline. However, if null values are detected, it becomes necessary to address them appropriately. Common strategies include removing rows or columns containing null values or imputing the missing data using statistical measures like mean, median, or mode, ensuring the dataset is complete and ready for further analysis.

We will see some of the labels we assigned to the dataset through the preprocessing:

{'Apple\_\_\_Apple\_scab': 0,

'Apple\_\_\_Black\_rot': 1,

'Apple\_\_\_Cedar\_apple\_rust': 2,

'Apple\_\_\_healthy': 3,

'Blueberry\_\_\_healthy': 4,

'Cherry\_(including\_sour)\_\_\_Powdery\_mildew': 5,

'Cherry\_(including\_sour)\_\_\_healthy': 6,

'Corn\_(maize)\_\_\_Cercospora\_leaf\_spot Gray\_leaf\_spot': 7,

'Corn\_(maize)\_\_\_Common\_rust\_': 8,

'Corn\_(maize)\_\_\_Northern\_Leaf\_Blight': 9,

'Corn\_(maize)\_\_\_healthy': 10,

'Grape\_\_\_Black\_rot': 11,

'Grape\_\_\_Esca\_(Black\_Measles)': 12,

'Grape\_\_\_Leaf\_blight\_(Isariopsis\_Leaf\_Spot)': 13,

'Grape\_\_\_healthy': 14,

'Orange\_\_\_Haunglongbing\_(Citrus\_greening)': 15,

'Peach\_\_\_Bacterial\_spot': 16,

'Peach\_\_\_healthy': 17,

'Pepper,\_bell\_\_\_Bacterial\_spot': 18,

'Pepper,\_bell\_\_\_healthy': 19,

'Potato\_\_\_Early\_blight': 20,

'Potato\_\_\_Late\_blight': 21,

'Potato\_\_\_healthy': 22,

'Raspberry\_\_\_healthy': 23,

'Soybean\_\_\_healthy': 24,

'Squash\_\_\_Powdery\_mildew': 25,

'Strawberry\_\_\_Leaf\_scorch': 26,

'Strawberry\_\_\_healthy': 27,

'Tomato\_\_\_Bacterial\_spot': 28,

'Tomato\_\_\_Early\_blight': 29,

'Tomato\_\_\_Late\_blight': 30,

'Tomato\_\_\_Leaf\_Mold': 31,

'Tomato\_\_\_Septoria\_leaf\_spot': 32,

'Tomato\_\_\_Spider\_mites Two-spotted\_spider\_mite': 33,

'Tomato\_\_\_Target\_Spot': 34,

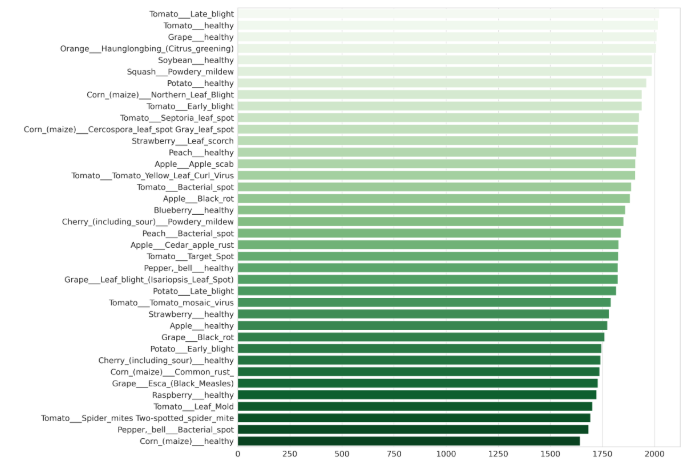
'Tomato\_\_\_Tomato\_Yellow\_Leaf\_Curl\_Virus': 35,

'Tomato\_\_\_Tomato\_mosaic\_virus': 36,

'Tomato\_\_\_healthy': 37}

**3. Exploratory Data Analysis(EDA)**

The bar chart illustrates the class distribution in a plant disease detection dataset, highlighting the frequency of various plant diseases and healthy conditions. This visualization aids in identifying potential class imbalances, which is crucial for ensuring unbiased model training and evaluation.

**Fig:** Performed EDA on train set of new plant disease detection highlighting frequency of various disease

**Step 3:** Splitting the dataset into Train, Valid and Test sets

**Training set:** The training set consists of labeled images representing various plant conditions, including healthy plants and plants affected by specific diseases. Each label corresponds to a unique class, such as a disease type (e.g., late blight, powdery mildew) or a healthy state. The dataset is structured to provide sufficient examples for each class, enabling the model to learn patterns effectively. However, as shown in the distribution chart, certain classes may dominate in frequency, indicating a potential class imbalance that needs to be addressed through techniques like data augmentation, oversampling, or weighted loss functions during training.

**Ex:** Found 63282 images belonging to 38 classes

2.Validation Set: The validation set comprises a smaller subset of the dataset, distinct from the training data, and is used to evaluate the model's performance during training. It includes images of plants representing the same classes (healthy and diseased conditions) as in the training set. The validation set helps monitor the model's generalization ability, tuning hyperparameters, and detecting issues like overfitting. Ideally, the validation set maintains a similar class distribution as the training set to ensure reliable performance assessment.

**Ex:** Found 1742 images belonging to 38 classes

3.Test Set: The test set is an independent subset of the dataset used to evaluate the final performance of the trained model. It contains unseen images of plants representing various classes, including healthy conditions and specific diseases. The test set serves as a benchmark to assess the model's generalization capability on new, real-world data. Unlike the training and validation sets, the test set is not involved in the model's training process, ensuring an unbiased evaluation of its accuracy, robustness, and reliability.

**Ex:** Found 17572 images belonging to 38 classes.