Deliverable 1: Model for disease risk evaluation of crops

Datasets

We will use historical weather data (temperature, humidity, rainfall, wind, etc.) combined with crop outcome data such as disease occurrence or levels. The dataset must include dates and location information so the model can learn patterns over time and predict future agricultural disease risk. If available, soil, and irrigation data will be included to improve prediction accuracy and provide more precise recommendations.

Methodology

a) The dataset is feasible because it contains both environmental variables and disease labels, which are necessary to train a predictive model. The most useful features include temperature, relative humidity, rainfall, and past disease occurrences. To transform the dataset into a forecasting problem, we will create rolling averages (such as three-day and seven-day averages of humidity), cumulative rainfall measures, and lag features that capture recent environmental trends. We will shift the disease label forward in time so the model predicts outbreaks in the following week rather than on the same day. During preprocessing, we will handle missing values, scale numerical features if needed, and encode categorical variables such as crop type or region.

b) From these datasets, the primary goal is to predict crop health and disease presence. The ideal choice would be to use a Convolutional Neural Network (CNN), which is a great tool for handling and processing leaf images. (or Regression with a probability score/classification technique)

c) Regarding evaluation metrics, we will be able to use a confusion matrix to compute the Accuracy/F1 score. In the case of a non classification ML technique, we will be able to convert the probability of disease score / estimated number of days before disease (/ estimated yield) to a class before estimating the accuracy of the model. It is coherent as we only have a finite set of values in labels. We will try to optimize the model by minimizing economic losses (hence minimize False Negatives at first, as losing the crops is more costly than just spreading the disease reduction method)

Application

Our model will first be integrated into a web application with user login. The user will be able to register their fields and farming habits, allowing for long-term monitoring rather than one-time predictions. The application can also be connected with soil sensors to extract further localised features such as soil moisture/temperature, electrical conductivity, and even soil nutrient levels. The user will input their fields with their associated:
- Location/size
- Crop type and variety

- Planting date/stage
- Soil sensor (optional but really useful)
- Irrigation type (optional)

Over time, the user can also provide more information for more accurate predictions. Information such as:
- Recent observations (pictures, ex leaf spotting observed in field C)
- Treatment history

As output, the user will have access to:
- Disease risk per field (as a probability/risk-tier) with forecast/trend visualization
- The expected severity of the disease
- The contributing factors of the predictions
- The potential type of predicted disease and their respective treatments
- Preventative measures adapted to the current situation/risk -> action recommendation
- Tracking of disease and risk history

The output will be displayed through dashboards for each field, with their associated:
- Graphs showing risk trends over time
- History log of predictions, treatments, imputed observations…

Ease of visualization will be a priority, allowing for an effective overview of trends in disease risks in each user's fields.