When a Plant Disease Epidemiologist works with data science

Sith Jaisong
Plant Disease Management Group, CESD, IRRI
Los Baños, Philippines
s.jaisong@irri.org

Data generally are raw and need to be cleaned in order to extract meaning from them or solve the problems. There are many methods in data processing that botanic epidemiology can borrow from data science. Epidemiologists often deal with more than one factor possibly causing or contributing to plant disease. Temperature, moisture, soil pH, soil type, plant varieties, crop density are some examples of variables considering the causes of plant disease development. When this number of variables are added to a study, the dataset becomes massive and requires powerful tools to manage it. Data science is emerging to meet the challenge of processing very large data. It applies techniques from statistics and computer science, extracting the meaning from the data and produce graphs, models, and other representations of the data, etc. in order to answer questions. The process starts with collecting the data that are able to answer the questions or hypothesis. Joining, combining, or wrangling the data are the processes after collection. In reality data may need to be cleaned (outlier and missing value removal or imputation). Once we have a clean dataset, exploratory data analysis may be applied. Next, designing models by using an algorithm such as k-nearest neighbor (k-NN), linear regression etc. may be used. The model selected depends upon the type of questions being asked. Then, we can interpret, visualize, report or communicate the result. The process of data science is in common with plant disease epidemiology. Epidemiologists often deal with more than one variable possibly causing the plant disease. Plant diseases develop from interaction between plant and its pathogen but also environment (weather) and human activities referring to the agricultural practices. The models are effective features capturing these relationships. They are developed for many uses or objectives. The most common ones are description, understating, prediction, comparison, and communication.The ideas and the process of tackling the problem are in the same concept. Therefore, data science process is the promising framework that epidemiologist should have in mind.