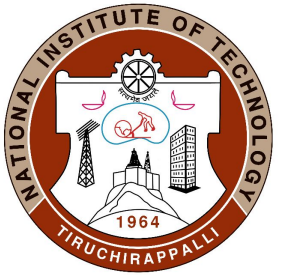
Using Machine Learning to predict Economic Threshold Level for crops from pest damages

Shubham
M.Tech Data Analytics
Department of Computer Analytics
National Institute of Technology,
Tiruchirappalli



Introduction
Insect, abundance and diversity

How many species of Insects exist? Compared to other life forms.

Insects	10,00,000	Fishes	19,000
		Amphibians	4,200
		Birds	9,000
		Flowering plants	2,90,000

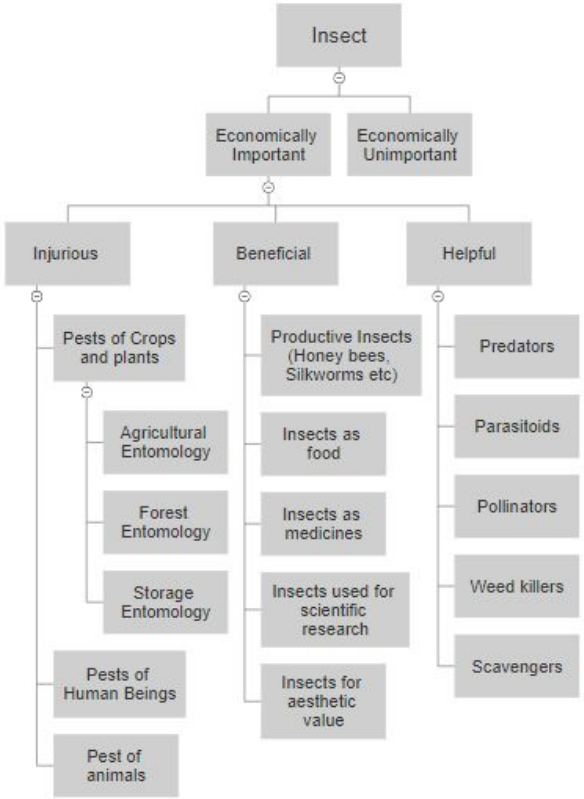
Insects constitute of more than 60% of life form present on planet earth.

Why Insects have been proven successful as species?

Darwin theory of ‘Evolution by natural selection’ has five basic premises which very well fits to insects.

- Many more individuals are born in each generation than will survive and reproduce.(natality)
- There is variation among individuals; they are not identical in all their characteristics.(variability)
- Individuals with certain characteristics have a better chance of surviving and reproducing than individuals with other characteristics.(survivorship)
- At Least some of the characteristics resulting in differential reproduction are caused by having different genes.(heritability)
- Enormous spans of time available for slow gradual change.(Time)

Insect classification (based on economic importance)



What is a Pest?

“Pest is an any organism whose population increases to such an extent as to cause economic loss to crops or a nuisance and health hazards to man and his livestock.”

The pest status of an insect species may be determined by number of ways such as:

- Increase in the number of insects species.
- Change in the type of damage inflicted on the crop.
- Change in the method of cultivation or harvesting.
- Fluctuation in the market value of the crops etc.

Pests are organisms which impose burdens on human population by causing:

- Injury to crop plants, forests and ornamentals.
- Annoyance, injury and death to humans and domesticated animals.
- Destruction or value depreciation of stored products.

Pests include insects, nematodes, mites, snails, slugs, etc and vertebrates like rats, birds etc.

- Average 18% of the crop yield is lost due to pests.

- Annual monetary loss in India is RS 60,000 crores.

How Insects become Pests? (Reason for outbreak of pest)

A. Destruction of forest or bringing forest area under cultivation.
B. Destruction of natural enemies.
C. Intensive and extensive cultivation of crops. E.g stem borers in rice and sugarcane.
D. Introduction of new crops and improved variations. Introduction of new crop may seem as new host for the pest. E.g sunflower and head borer.
E. Improved agronomic practices.(use of nitrogeneous fertilizers)
F. Introduction of new pest in new areas.
G. Accidental Introduction of foreign pests
H. Resurgence of sucking pests.
I. Large scale storage of food grains.

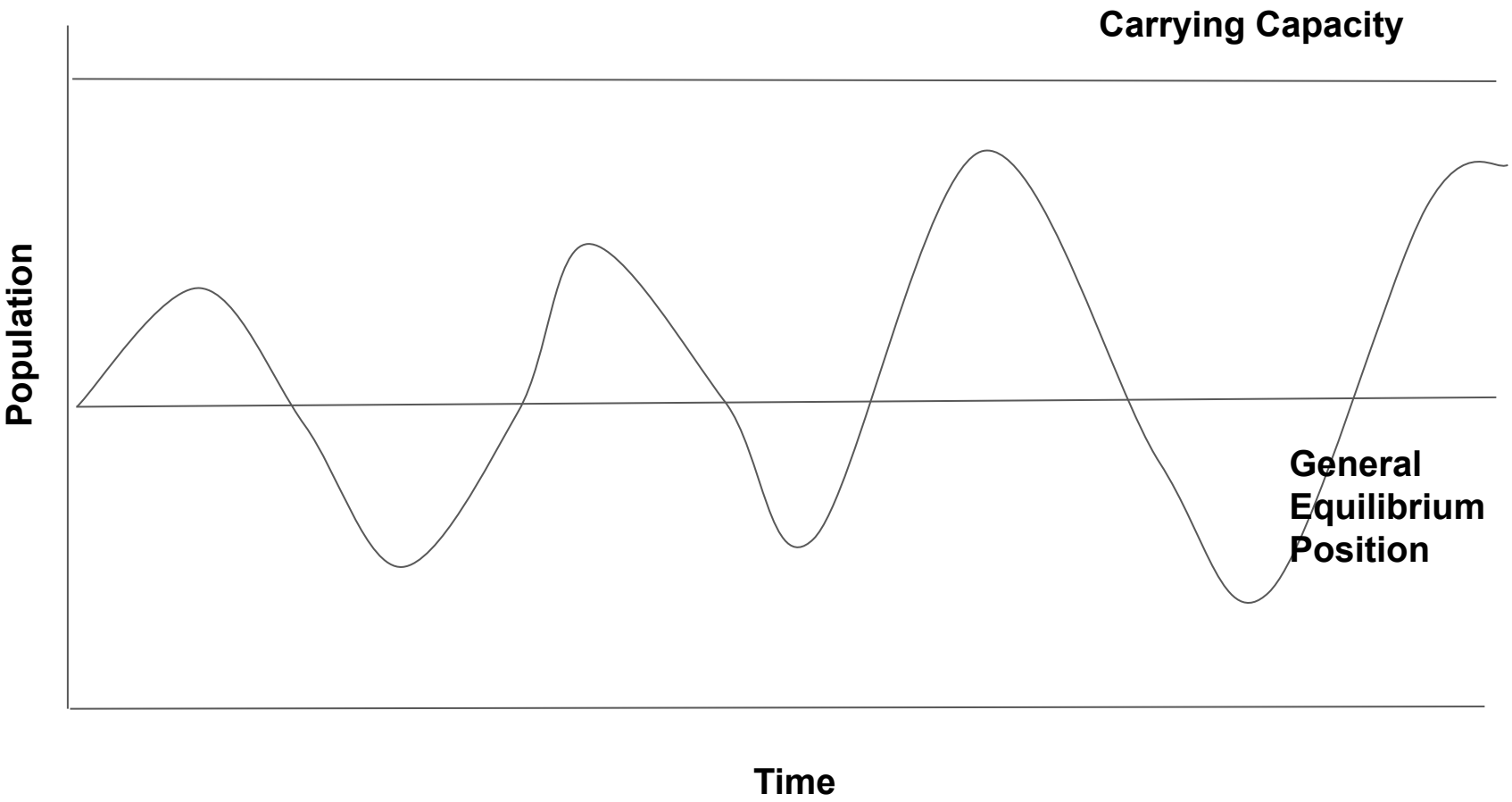
When an Insect is called Pest?

General Equilibrium positions(GEP): It is the average of a population over a long period of time, around which the insect population tends to oscillate due to biotic and abiotic factors.

-> Biotic factor : Fecundity, Food availability, natural enemies etc.

-> Abiotic factor : Temperature, Rainfall, humidity, wind etc.

- Whenever, the population crosses the carrying capacity certain biotic and abiotic factors come in play to bring down the population.
- Whether this GEP of a pest will be a matter of economic concern for crops? If yes, bring down the population artificially.


Carrying Capacity
Population
General
Equilibrium
Position
Time

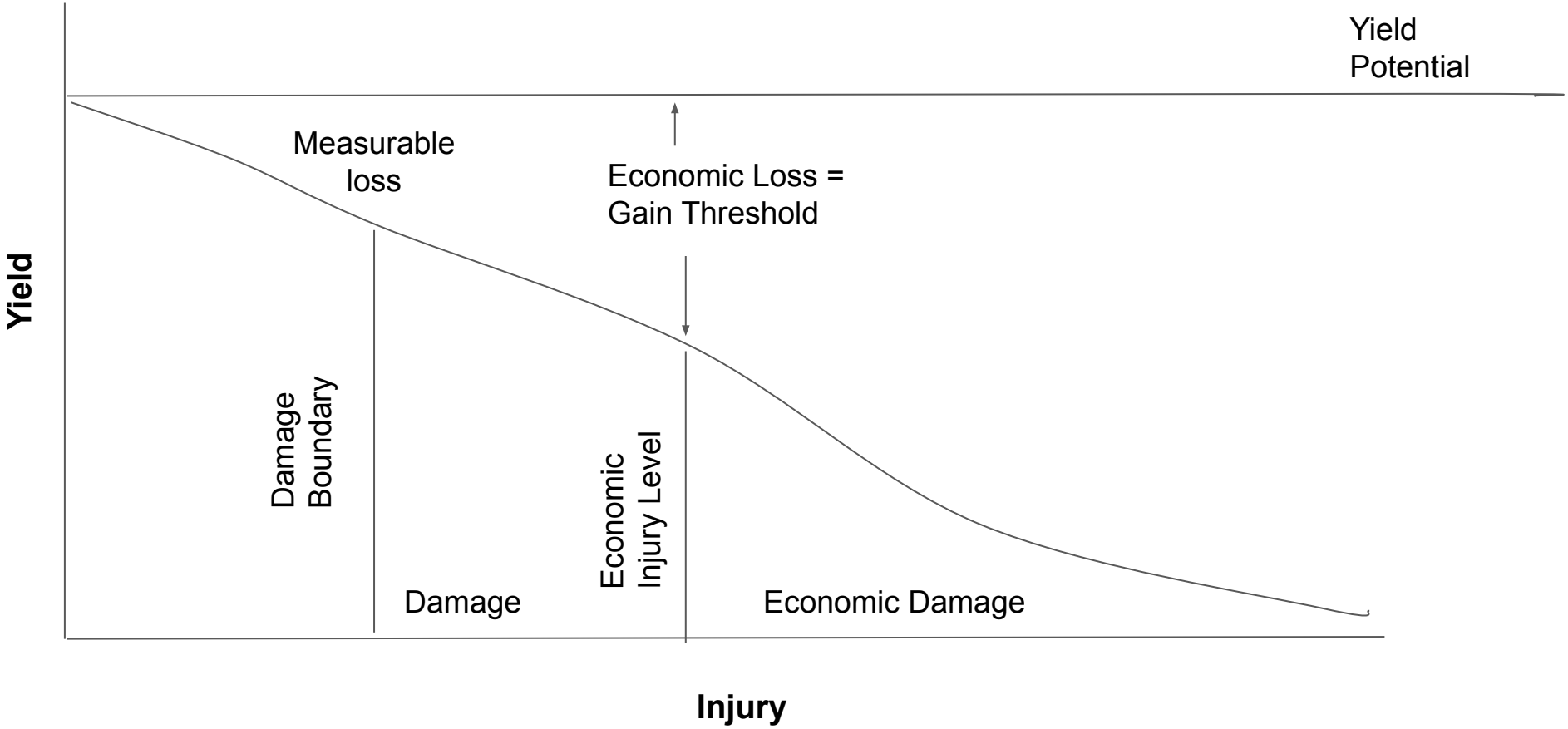
Important Terms

-> Damage boundary (DB) : It is the lowest level of damage which can be measured .

-> Economic Injury Level (EIL) : It is defined as the lowest population of insect that will cause economic damage.

Or,

The critical density of insect population where the loss caused by the pest equals the loss of control measures.


Yield
Potential
Measurable
loss
Economic Loss =
Gain Threshold
Yield
Damage
Boundary
Economic
Injury Level
Damage
Economic Damage
Injury

How to calculate economic injury level?

EIL can be calculated by using the following formula.

EIL = P’ = C/V*I*D*K

- P’ = Economic injury level in insects/ production or insects/ha
- C = cost of management activity per unit production(RS/ha)
- V = market value per unit yield (RS/tonne)
- I = crop injury per insect(percent defoliation or percent fruit damage)
- D = Damage or yield loss per unit injury (Tonne loss or % defoliation)
- K = proportional reduction in injury from pesticide use

An Example

Problem: Calculate EIL in terms of pest population/ha with the following figures.

C = management cost per unit = RS 3000/ha

V = market value (in RS / Unit product) = RS 1000/tonne

I = crop Injur / Pest density = 1% defoliation / 100 insect

D = Loss caused by unit injury = 0.05 tonne loss / 1% defoliation

K = Proportionate reduction in injury by pesticide application = 0.8(80% control)

EIL = C/V*I*D*K = 3000 / 1000 * 0.01 * 0.05 *0.8 = 7500 insects/ha

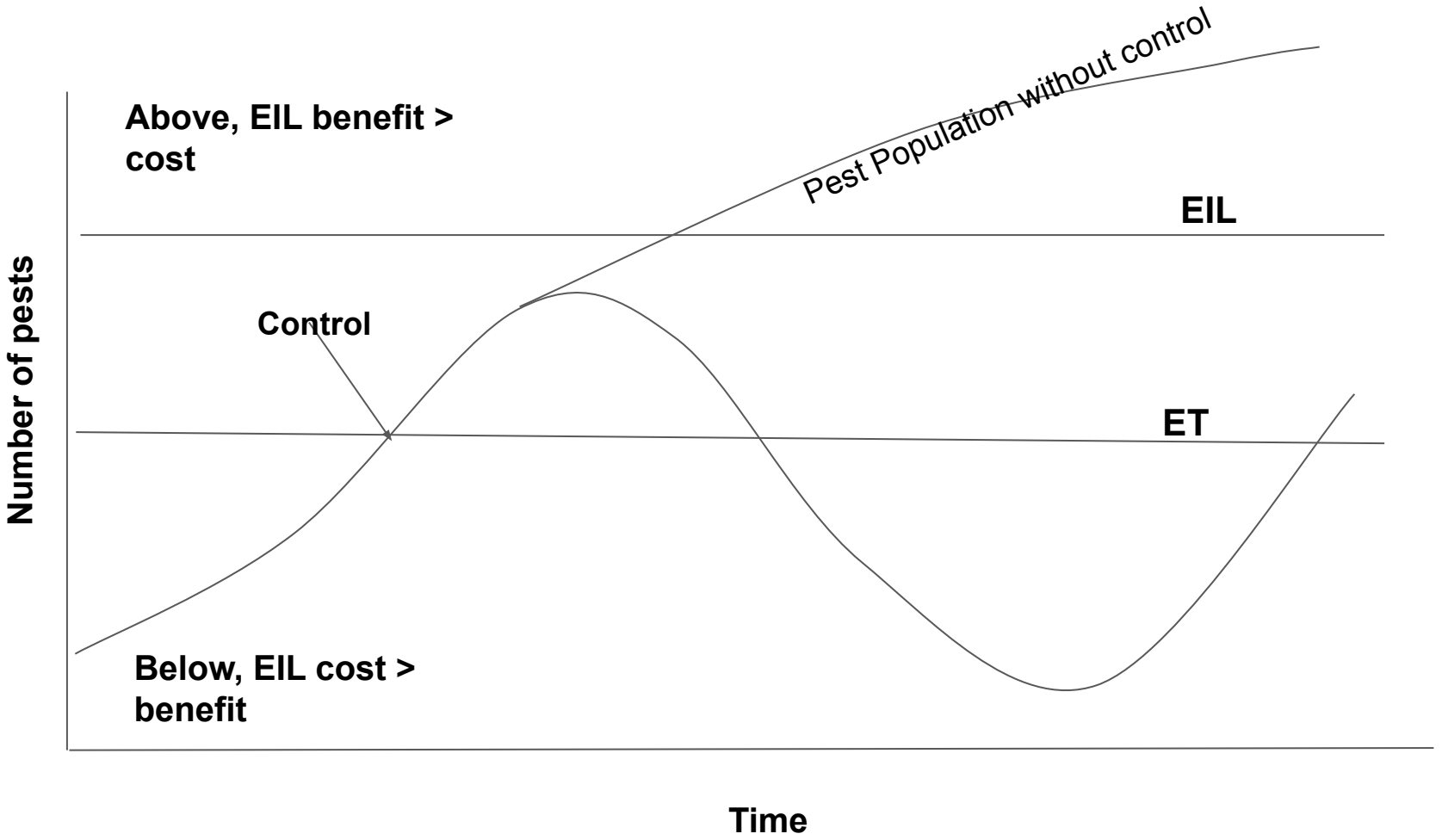
Is EIL of a pest constant at all time? NO

EIL is influenced by:

1. Market value of the crop(Primary factor): When crop value increases, EIL decreases and vice versa.
2. Management os injury by insect(Primary factor): when management cost increases, EIL also increases.
3. Degree of injury per insect(secondary factor): Insect damaging leaves or reproductive parts will have different EILs. For instance, defoliators will have higher EIL.
4. Crop Susceptibility to injury(secondary factor): If the crop can tolerate the injury and gives good yield, EIL can be fixed at higher volume. When the crop is older it can withstand high population - EIL can be high

Should we start Pest Management at EIL? NO

- Pest Management can’t be started at EIL because by the time the measures turn out to be effective crop would have crossed EIL.
- Hence, Pest management should be started at Economic Threshold Level(ETL) or Action Threshold.
- Economic Threshold Level(ETL) or Action Threshold: It is defined as the pest density at which control measures should be applied to prevent an increasing pest population from reaching EIL.


Above, EIL benefit > cost
Pest Population without control
EIL
Control
ET
Below, EIL cost > benefit
Number of pests
Time

How Machine learning can help?

- As, it can be seen finding Economic threshold is an optimization problem which can be solved using Machine learning.
- To start pest management farmers have to make a decision i.e a point in time at which they should start the use of insecticides/pesticides and/or other methods for pest management.
- By analyzing data records and having a prenotion of economic injury level we can give a confidence level for Economic threshold.
- This prediction can thus give us better cost benefit.

Plan of Attack

- Scraping the data of internet for the prices of Cotton
- Preprocessing the data available for incidence of pest on cotton plant which includes encoding for categorical data and scaling of different features using normal distribution.
- Applying different regression techniques to model the data.
- Applying hyperparameter tuning such using grid search
- Applying K-fold-cross validation
- Using some ensemble methods to increase model performance like XGBoost and Random Forest.
- Calculating ET of crop using Live market price.
- Predicting Pest occurrence using best model performance, and checking whether it has crossed ET or not.

Problem Statement Description

1. Problem Statement - Apply ML techniques to study the pest population dynamics in agriculture.

2. Problem description - To develop Machine Learning models to study the pest population dynamics. Pests are harmful animals or plants which pose a threat to human concerns. In our problem statement we are studying pests that pose a threat to agricultural crops. We need to have a strict check on pest population in order to maximize the crop yields. Crops that constitute food, fibre etc are essential for human existence and thus their destruction poses a threat to human kind. In this project we will be discussing how several independent parameters lead to pest population growth and using those parameters we will predict pest population for the next crop cycle and so that their incidence can be reduced.

3. Dataset Description - Data used for this project is collected by
ICAR - National Bureau of Agricultural Insect Resources (NBAIR)

   Independent Variables in the dataset:
   - Season - season of the year.
   - Crop stage - stage of the group.
   - MaxT - Maximum temperature during that season.
   - MinT - Minimum temperature during that season.
   - RF - Rainfall (in mm)
   - RFD - Rainfall days i.e no of days rain fell in a week.
   - RH1 - Relative humidity maximum.
   - RH2 - Relative humidity minimum.
   - NE1 - natural enemy of pests, 1 i.e spider
   - NE2 - natural enemy of pests, 2 i.e Coccinellids

   Dependent Variables in the dataset:
   - PI - Pest incidence

Preprocessing

1. Data Cleaning:
    - Outliers would be removed from data.
    - Data would be verified for any missing values
    - If found missing values are found, proper imputations would be done using various statistical methods.
    - Missing values would be replaced with mean, median or mode of that particular column.
    - Missing value can also be replaced with global constant or with most probable value derived from bayesian inference.
    - However, the dataset acquired for this project from **ICAR - National Bureau of Agricultural Insect Resources (NBAIR)** doesn't have any missing values.
    - In Spite of no missing values in the dataset we will check for outliers using clustering and hypothesis testing.

2. Encoding of categorical variables:
    - Two independent variables in our dataset are categorical.
        i. Season - season of the year.
        ii. Crop stage - stage of the group.
    - Categorical variables can only take one of a limited, and usually fixed, number of possible values. Thus a higher categorical value can distort our result. Thus we need to encode these variables to neutralise this effect.
    - We will perform Integer encoding to convert these categorical variables into dummy integer variables.

Contd...

3. Feature scaling:
    - Since the range of values of independent variables like temperature and relative humidity will vary widely we will scale their values.
        i. Here feature scaling will be done using standardisation.
        Where, Xstand = (x - Xmean) / Standard Deviation(X)

4. Data Discretization:
    - To model our prediction using decision trees and random forest algorithms we would need discretized values for variables which are continuous and cover a large range of values. For our data set temperature and humidity can be discretized.
    - Data could we discretized by using various methods we would be using following three approaches:
        - Equal width binning- In equal-width binning, the algorithm partitions the data in ‘k’ intervals, and each interval has the same size.
        - Equal frequency binning- Here we will be dividing data into ‘k’ groups where each group will have approximately the same number of data points.
        - Max diff method- Here the column which needs to be discretized will be sorted in ascending order and bins boundaries will be defined at those points where there is maximum difference between the adjoining values.

Model Selection

1. Data set is divided into training set and test set using train test split method of sckit learn module, random state is set to 42 to increase the spatial variance.
2. Various Regression models are selected to train the data. Starting from polynomial regression, SVM, Decision tree etc.
3. Trained model are fed encoded data which are already scaled using standard scaler.
4. Predicted results are thus compared with test set.
5. We calculate r2 score of the test data to measure model performance.
6. Further Ensemble based techniques are used to increase model performance. Algorithms used for our dataset are Decision tree and XGBoost.
7. K-Fold-validation is applied to check if the performance on different subsets of data.

Predicting whether crop has crossed ET or not

1. We fetch live cotton prices from internet by scraping websites.
2. Using the price we calculate ET.
3. We put sample data in one of our best performing models to predict pest incidence level. If that pest incidence level has crossed ET of that crop we alert the user stating the found fact.

Thank you