Kai Tiede, Irfan Radarma, Jack Jacobs, Matt Lampl, Sara Maillacheruvu

Project Title: To Bee or not to Bee: Colony Collapse in the U.S.

Team Name: Honey Bears aka Busy Bees aka The Swarm™ aka The Bee's Knees

Data and Model Plan

Data Summary

To complete both the Machine Learning and Simulation tasks, we will need a wealth of data describing honey bees and honey production. We will be using the following datasets to support these tasks:

- **Honey Neonic Dataset:** Sourced from Kaggle, this dataset includes our target variables (number of honeybee colonies and honey production levels) as well as feature variables, such as levels of exposure to various toxins that harm honeybee populations.
- **Temperature dataset:** Sourced from the National Oceanic and Atmospheric Administration, this dataset contains historical temperature at the state level
- **APHIS Dataset:** Sourced from the National Honey Bee Survey, this dataset contains feature data on factors that can impact honeybee populations, including levels of varroa mites, Nosema spores, and more.
- **Urbanicity Dataset:** Sourced from the US Census Bureau, this dataset contains information on the level of urbanicity per state across decennial census years to enable us to see whether states with higher levels of urbanization have impaired honey production levels.
- **Air Quality Dataset:** Sourced from the Environmental Protection Agency, this series of datasets contains county-level information on air quality, such as the number of days in a year with good air quality, hazardous air quality, and levels of PM 2.5 and PM10.
- **Pesticide Dataset:** Sourced from the US Department of Agriculture, this dataset contains data from 1994 to 2020 on levels of various pesticides, such as Metribuzin and Trifloxystrobin.

The main resource for the Machine Learning task will be the Honey Neonic dataset. This dataset includes information about the levels of pesticides used, as well as the level of honey production (the target variable). The Simulation task will take a slightly different approach. Using

weather data, as well as known distributions from prior studies about pesticides and honeybees, we will be leveraging data to build simulation models.

Model Plan

Machine Learning Modeling Plan

The model is predicting the *number of honey producing colonies* using 11 features. Because the features were from different data sources with different format and granularity, we opted to normalize the data on an annual (except the urbanicity data) and state-by-state basis.

Code	Description	Data Source
X_1	The amount in kg of clothianidin applied	USGS's Pesticide National synthesis Project
X_2	The amount in kg of imidacloprid applied	
X_3	The amount in kg of thiamethoxam applied	
X_4	The amount in kg of acetamiprid applied	
X_5	The amount in kg of thiacloprid applied	
X_6	Annual Average Temperature	National Centers for Environmental Information
X_7	Annual Average Temperature Anomaly	
X_8	Decade Average Urbanization	U.S. Census Bureau
X_9	Annual Median AQI	U.S. Environmental Protection Agency

X_{10}	Number of varroa mites per 100 bees	APHIS National Honey Bee Survey
X_{11}	Millions of Nosema spores per bee	

Because the label to predict is continuous data, we will use regression models. Two regression models we employ in this project are:

1. Linear Regression:

$$\hat{y} = \beta_0 + \sum_{i=1}^{11} \beta_i x_i$$

Note: regularization may be implemented, depending on the training process

2. Regression Trees with ID3 algorithm.

- Entropy:

$$H(S) = \sum_{x \epsilon X} -p(x) \log_2 p(x)$$

S : The dataset for which entropy is being calculated

X : The set of classes in S

p(x) : The proportion of the number of elements in class x to the number of elements in set S

- Information gain:

$$IG(S, A) = H(S) - \sum_{t \epsilon T} p(t) H(t)$$

H(S) : Entropy of set S

T : The subsets created from splitting set S by attribute A

p(t) : The proportion of the number of elements in *t* to the number of elements in set S

H(t) : Entropy of subset *t*

Simulation Modeling Plan (& Decision Process)

Honey production by agricultural bee colonies does not proceed deterministically from the levels of the features detailed above. Even if it did, though, predicting either agricultural yield or colony collapse in years that are outside our training data would be complicated by the stochastic nature of the contributors and impediments to colony success. Prior work by scientists at the EPA provides us with a framework to simulate distributions for our predictions using Monte Carlo methods for our features (and their interactions). Specific distributional assumptions for relevant features are listed below (all sampled by state), though we may add to these if we find in our modeling process that interactions between features produce unexpected distributions when explored, or if disaggregated data suggests a data-generating process with a different distribution from below.

Feature category	**Distribution sampled**
Pesticide application	Normal
Annual average temperature	Normal
Annual average temperature anomaly	Poisson
Decade average urbanization	**Not modeled stochastically**
Annual median AQI	log-Normal
Pests per bee	Poisson

Monte Carlo simulation will be one of two primary methods (the other being regularization) by which we will try to achieve external validity from our predictions. Also, due to our state-by-state

sampling process, our model should produce relatively larger confidence intervals for state-level predictions and smaller confidence intervals for nation-level predictions.