## Model Card for Ryan Seaman Random Forest Model

## Introduction

This model card template is taken directly from Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* (pp. 220-229).

## Model Details

* **Person or organization developing model**:
  + Ryan Seaman, Colby College, CS498: Seminar in Artificial Intelligence
  + Brown Tail Moth Research Project
* **Model date**:
  + December 7, 2022
* **Model version**:
  + 1.1
* **Model type**:
  + Random Forest
* **Paper or other resource for more information**:
  + Details on the model:
  + <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
* **Citation details**:
  + No citations for the model
* **License**:
  + na
* **Feedback on the model**:
  + The model was efficient and effective at determining BTM with provided data.

## Intended Use

* **Primary intended uses:** 
  + Using BTM infestation data and metadata on specific trees to classify infestation and ultimately treatment options.
* **Primary intended users:** 
  + Researchers studying BTM infestation and government officials seeking to address the BTM infestation, specifically within Maine.
* **Out-of-scope uses:** 
  + People learning about BTM infestation and people learning about the potential applications for a Random Forest.

## Factors

* **Relevant factors:** 
  + na
* **Evaluation factors:** 
  + na

## Metrics

* **Model performance measures:**
  + Confusion matrix accuracy percentages.
* **Decision thresholds:**
  + na
* **Approaches to uncertainty and variability:**
  + na

## Evaluation Data

* **Datasets:** 
  + The BTM dataset that was provided for the class: BTM.xlsx
  + 90 to 10, training to test split.
* **Motivation:**
  + Measure the effectiveness of the Random Forest Model to predict BTM infestation.
* **Preprocessing:** 
  + Normalization and transforming all data to floats
  + Filling in all NA values with median.
  + Removing fields that were directly indicative of BTM infestation such as ‘Proposed Treatment Type.’
  + Manually switching some of the cells within excel from Yes and No to Y and N.

## Training Data

* **Datasets:** 
  + The BTM dataset that was provided for the class: BTM.xlsx
  + 90 to 10, training to test split.
* **Motivation:**
  + Measure the effectiveness of the Random Forest Model to predict BTM infestation.
* **Preprocessing:** 
  + Normalization and transforming all data to floats
  + Filling in all NA values with median.
  + Removing fields that were directly indicative of BTM infestation such as ‘Proposed Treatment Type.’
  + Manually switching some of the cells within excel from Yes and No to Y and N.

## Quantitative Analyses

* **Unitary results:** 
  + With the best chosen hyperparameters (max features = sqrt or log2, max depth = 5, n estimators = 100), the model was shown to have an accuracy of 0.989.
* **Intersectional results:** 
  + na

## Ethical Considerations

There are no ethical considerations needing to be mentioned.

## Caveats and Recommendations

Recommendations:

Collecting more data with more data fields (columns) and working to reduce the number of NA inputs. This will allow the model to have more fields with useful information to draw from.

Credits:

The data was prepared referencing Professor Amanda Stent’s sample code for data preparation.

The decision trees and random forests were made referencing Professor Amanda Stent’s sample code (on the iris dataset), along with the YouTube video made by Data 360 YP walking through a random forest workflow, and a stack overflow tread on visualizing decision trees.