**Business context**.

This project focuses on predicting the spread of White Noise Syndrome in bat species across North America, using data found on publicly available datasets with climate features and bat population decline numbers. The goal was to build a machine learning model to classify if a location would be at high risk for WNS spread in the next year or selected year after (2024). A successful model could improve species conservation efforts targeting the species as well as highlight potential features that could correlate to increased or decreased spread. Additionally, this could motivate stakeholders to contribute to the model by offering more data for future model modifications. Protecting bats is important for the agriculture business since they eat many pests, reducing the need for harmful pesticide use. Our model could also be useful for biotech and pharmaceutical companies who may use these findings to develop new ways to treat or prevent White-Nose Syndrome in bats, leading to new opportunities for research and development funding in these fields.

**Data Preparation and Validation**

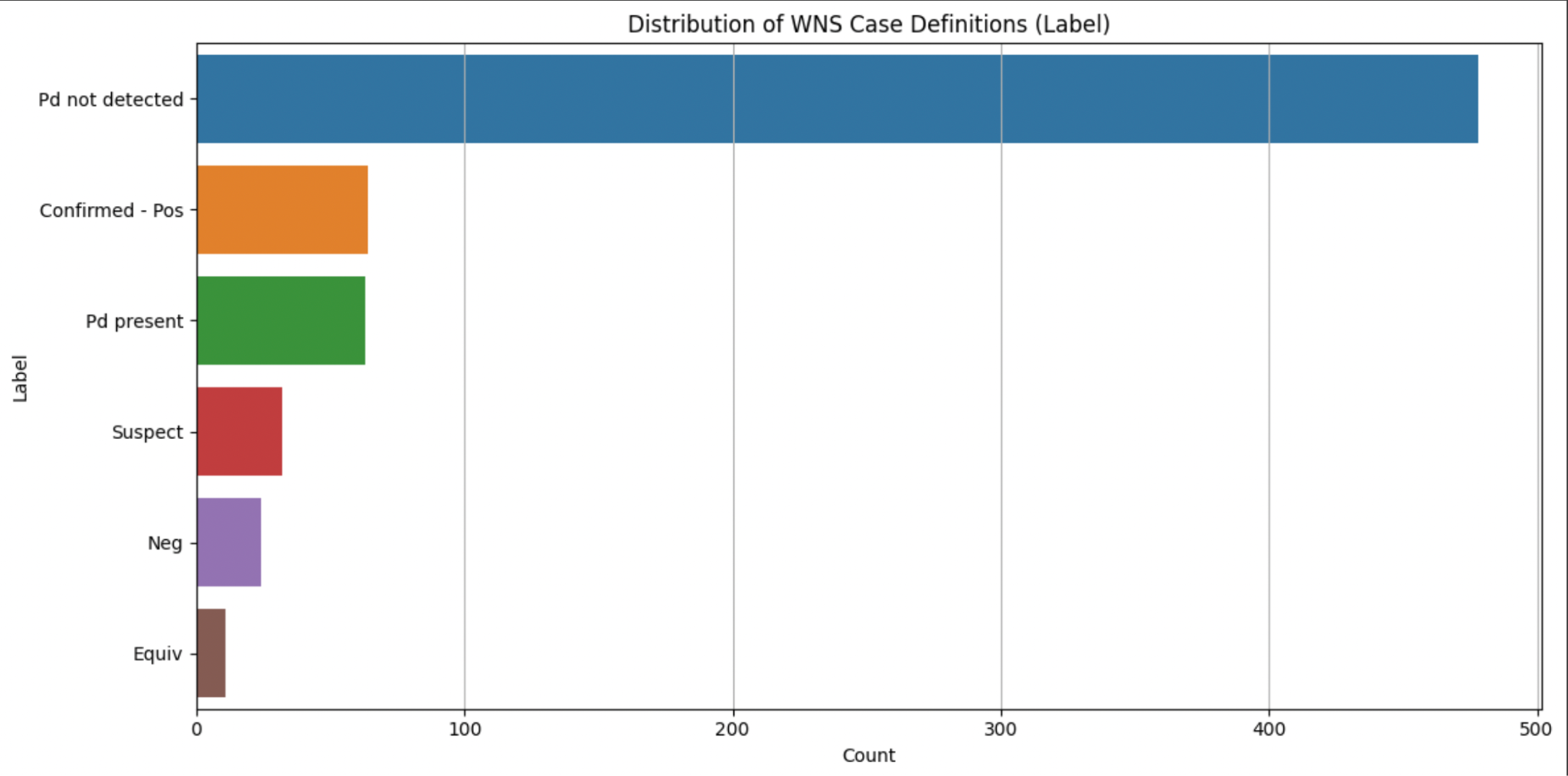
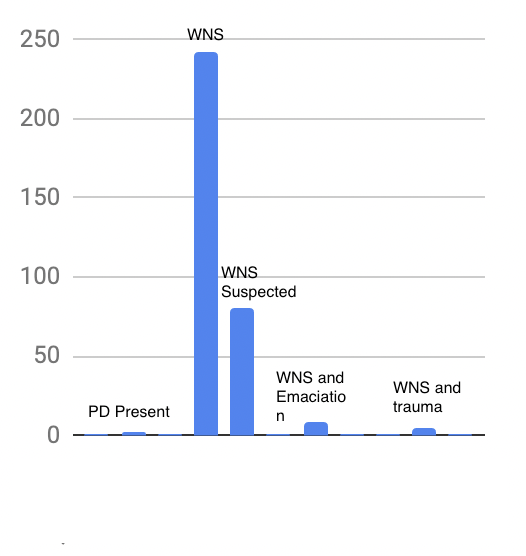
DataSet Description

We were given links to a few publicly available datasets by BioInterphase. These were datasets with locations and populations of bat populations and white nose syndrome detections across North America. We were also encouraged to seek other publicly available datasets, and we utilized regional climate predictions from the NA-CORDEX project.

One of the major challenges of this project was finding relevant data since bat data records tend to be incomprehensive, in a format unfit for analysis, or restricted/classified. The data we did have contained primarily locations, dates, species of bats affected, numbers related to the amount of bats affected, and whether or not there was WNS present at the site.

To pre-process the data, we joined three data sets. We made data types consistent, coding counties by their ID and latitude and longitude,

Exploratory Data Analysis (EDA)



We explored our data set by mapping the number of instances of WNS detection, their location, and time of location, and noting other associated data features such as population count. Because population counts seemed to vary, and an additional data set of WNS detections for bats in our time and space range did not include population counts, we proceeded to drop that feature.

Our sample data was fairly balanced between WNS detections and examples where it wasn’t detected. To fully merge, we had to reorganize our detection label into three categories: WNS, Negative, and Fungus Detected. This was condensed from previous labels of fungus detection with no white nose presence, white-nose syndrome, suspected white-nose syndrome and trauma, and white-nose syndrome and emaciation.

To not bias the samples from reasons for sample collection, we dropped the label of the negative detection, populating another set of negative sample detections from a random selection of points across North America where WNS had not been detected.

Feature Selection

We used climate data from The North American Cordex Program, which utilizes regional climate models to simulate conditions between 2006 and 2100 at daily frequencies with a spatial resolution of 0.22 degrees/25 km. The extracted features were mean/max/min temperature, mean/max/min humidity, mean/max/min precipitation, and surface downwelling shortwave radiation (solar radiation). Feature engineering consisted of scaling the data to enhance compatibility with models employed and introducing a distance parameter to account for the spatial aspect of our data. We also calculated the distance between points and the nearest previous disease detection using the Haversine formula.

**Approach**

Selected models

We selected 3 candidate models to explore guided by interpretability, functionality with low data volume, and low computational demand:

* Decisions Trees - we selected this model for its adjustable complexity, ability to scale well, and interpretability. However, it is prone to overfitting and can get computationally demanding if ensemble models become large. We can reduce overfitting with a random forest.
* Support Vector Machines - We selected this model for its ability to handle small datasets with high feature counts and memory efficiency, and it does not require high data distribution to function effectively. On the con side, it produces probability estimates, which may be inefficient, and the model is not as easily interpretable as decision trees.
* Gaussian Processes - this model was chosen because it allows to calculate confidence intervals and is most suitable for a small sample size and a small to medium number of features that match our data. These models are commonly used when data can be divided into distinct clusters that can be learned by different models, which could reveal differences in climate biomes across the United States.

We compared models by analyzing their confusion matrices and balancing precision and recall. We want to predict conservatively, so we aim to reduce false positives (cases where WNS is predicted when there is no presence of WNS).

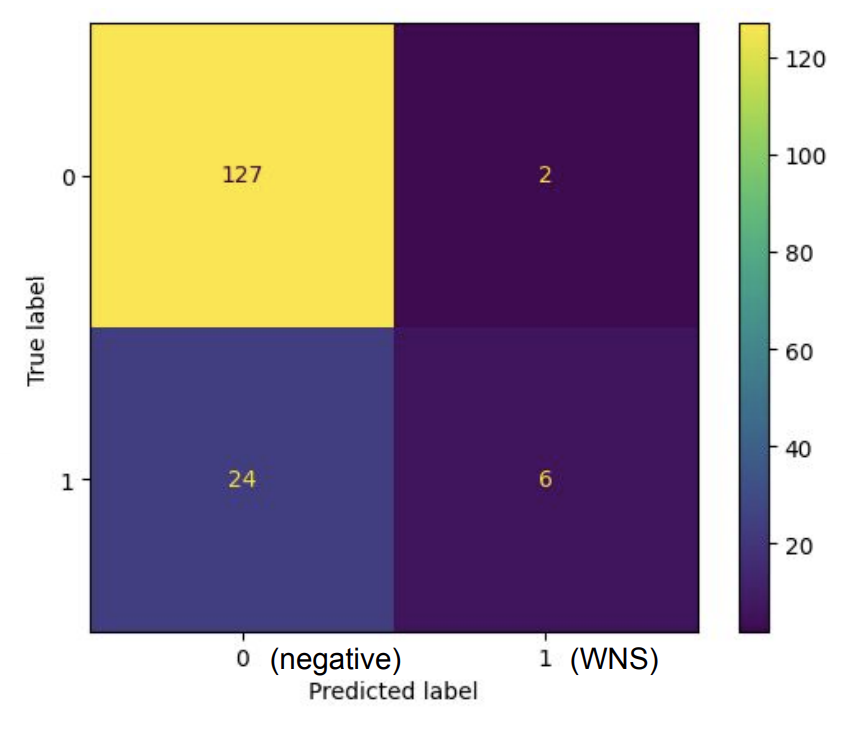
We divided the training data in 2020 into past and future detections to train our models. We aimed to ensure that future detections don't influence past detections.

We first trained and evaluated a decision tree model while observing precision and recall. Then, we fine-tuned hyperparameters using GridSearchCV.

It's worth mentioning that the model with the highest accuracy used a class weight set to 'balanced' to handle the imbalance of positive and negative cases. Our dataset was slightly imbalanced by introducing more negatives through random sampling.

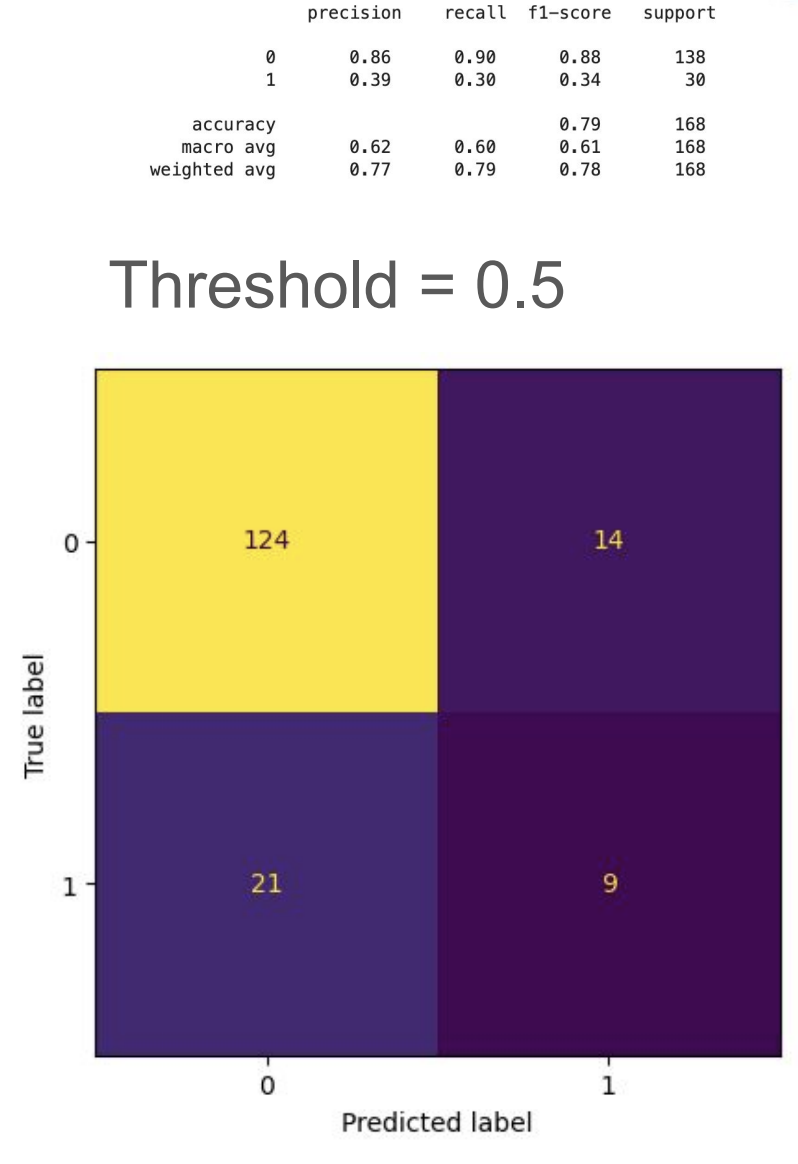
The depth of the tree was set to 'None,' which made us a bit worried about the risk of overfitting, especially considering our small dataset. However, the minimum samples per leaf were set to 1, which suggested a particular tree. This initial tree gave an overall accuracy of 0.72.

We switched to a random forest model to address overfitting issues. As a result, the accuracy improved to 0.84. However, when we examined the classification report, we discovered that the model accurately predicted white-nose syndrome detections, but only 20% were correctly recalled. The final confusion matrix is as follows:



The support vector machine modeling process started with a scaling of features. At first, with no hyper-tuning, the model performs quite poorly at 0.20 accuracy. With a grid search, the accuracy improves to 0.34, but this is just guessing on data categorized into 3 (white-nose detection, negative, fungus detection). By removing that category and simply working on binary white-nose syndrome detections and no detections, the accuracy improves to .45, still randomly guessing.

The Gaussian process start with creating spatial and temporal kernels to capture those aspects of the data. After training, we transformed predictions, which gave us the probability of each label at a certain coordinate to binary classifications with a threshold ○ 0.5 threshold means probabilities above 0.5 are classified as 1 (WNS detection). In contrast, those below are classified as the negative class 0 (No WNS). Scaling this parameter allowed us to play with the precision/recall balance, giving us predictions that were conservative or overpredicting detections.



**Key Findings and Insights**

Key Results

Our final model combined our two highest-scoring models: random forest (0.84) and Gaussian process (0.83). Specifically, we wanted the Gaussian process to inform the labels of our random forest, so we relied on the random forest for future white-nose syndrome predictions. However, the gaussian process model gave us interesting probabilistic results on risk areas.

Our final submission was a future prediction by the random forest model. This also satisfied our need for a model built on interpretability, as we could map each tree and note the key features that led to white-nose syndrome detections or not.

Insights:

1. **Feature Importance (Mean Precipitation Higher Feature Importance)**: We extracted maximum and minimum climate feature sets for features such as humidity and temperature because we were aware that warming temperatures and extreme climate could increase the spread of white-nose syndrome across bats. Even with this initial guess, our highest features still ranked the mean precipitation above mean humidity and solar radiation. It seems that precipitation may correlated with white nose syndrome disease spread. Our feature ranking did confirm that distance to previous detections had the highest importance in predicting WNS.
2. **Federally gathered data:** We also learned that vast resources and initiatives are engaged in data collection! We were so grateful to have found the climate prediction resource NACORDEX. At the same time, we also learned that there are many limitations to doing important conservation research. We faced many blocks in finding specific and enough bat location and disease detection data because its is federally protected to protect bats.
3. **Climate data libraries:** Additionally, we learned to use X-arrays, a popular library for working with climate data. This allows for high-dimensional data with features with coordinates in space and time to be easily manipulated and plotted, which is important for visualizing trends in our data.

