Problem or Experiment Description:

I want to develop a machine learning classification model to predict species extinction risk based on biological and ecological factors. By analyzing species traits, I will classify species into different risk categories (e.g., Least Concern, Vulnerable, Endangered, or Critically Endangered) as defined by the IUCN Red List. While many at-risk species are already identified, this experiment will show if there are biological similarities between those species, and will hopefully create a model that will predict whether additional species need conservation.

Dataset:

I have identified two main datasets to use. The first is data from the IUCN Red List. This data shows classification labels for species based on how at-risk they are for extinction. It is publicly available online, following a request to use it. I currently have a request pending. The second dataset I want to use is from the PanTHERIA Database. This is also publicly available, and includes species and their biological and ecological traits. I will combine these two datasets to make one dataset that has each species' name, traits, and extinction risk.

As I'm still waiting to see the data and variables available from the IUCN Red List, I may need to find additional datasets with more species or environmental data. Once my data request is approved, I will have a more clear idea of what I have and if/what I need to fill in any gaps.

Algorithm or Methodology:

For this project, I will build a few different classification models and evaluate them to choose the best one. I would like to create a support vector machine, and perhaps a decision tree and/or random forest model. We haven't learned them yet, but I think they would be relevant. Once I have these models created, I will review summary statistics and report out results, including how the models compare and which model was the most effective.

Expected Results:

At the end of this project, I will have at least two trained classification models that predict the extinction risk categories of different species based on biological and ecological data. I will also have a recommendation for which model should be used. I will use matplotlib and seaborn to create visualizations of the models and data. If my model performs well, I will hopefully also have insight into the most influential predictors of species vulnerability. This data could then be used to make recommendations on which species are in need of future conservation efforts or research.