Capstone Two: Project Proposal
August Perez

- Problem Identification
    - Problem statement formation
        - Produce a model before 27Jun2024 that correctly predicts asthma diagnosis with a sensitivity of at least 90% (reducing false negatives).
    - Context
        - Early diagnosis & treatment of asthma has major quality of life benefits to patients. Symptoms can range from barely noticeable to life-threatening (depending on lifestyle). Using a model to predict a diagnosis would assist doctors in identifying patients in need of treatment before an asthma attack (which can range from manageable to life-threatening).
    - Criteria for success
        - The model achieves at least a 90% sensitivity rate.
    - Scope of solution space
        - Within the jupyter notebook created to build the model.
    - Constraints
        - Some null values within the dataset may have to be imputed or the entire entry removed if multiple nulls are present.
    - Stakeholders
        - Medical personnel (doctors, hospitals, etc.)
    - Data sources
        - The dataset obtained from kaggle
            - https://www.kaggle.com/datasets/rabieelkharoua/asthma-disease-dataset?resource=download

- What is the problem you want to solve?
    - Identification of asthma. Having a model that could assist with early prediction would get patients treated correctly sooner and reduce asthma related suffering.
- Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis?
    - The end client would be medical practitioners (hospitals, doctors). Using the model as a sort of screening they would be more effective in providing treatment where it is needed and would be able to more quickly sort through possible asthma patients.
- What data are you using?
    - An asthma disease dataset that contains 26 possible factors for asthma
- How will you acquire the data?
    - Kaggle dataset (downloaded csv file)

- Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.
    - Data wrangling
    - EDA, check which factors correlate with the diagnosis the most & if any factors co-correlate so that I can reduce total features
    - Pre-processing (deal with null values here)
    - modeling: compare multiple classification models (trees/forests, KNN, logistic regression)
- What are your deliverables? Typically, this includes code, a paper, or a slide deck. For this capstone, the deliverables include:
    - A GitHub repo containing the work you complete for each step of the project, including:
        - A slide deck
        - A project report