Thyroid Disease

Alex Lai

- **What is the problem you want to solve?**
    - To make a predictive model if the thyroid cancer patient is likely to respond to the treatment or not. The accuracy has to be above 95% to be considered successful. This will be completed within the next 3 months.
    - Stakeholders would be the company making the treatment and doctors.
    - The solution space will most likely be creating the model using Pandas.
    - Constraints would be the limit amount of data to train and the quality of the data.
- **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?**
    - My client would be the company making the treatment, doctors, and hospitals that can locally or remotely access this model to make a quick check up prediction.
        - For the doctors or researchers giving the drug a trial run, they can quickly access the model to roughly see how this patient react to the drug.
- **What data are you using? How will you acquire the data?**
    - I will be using the Thyroid Disease Dataset from Kaggle. (https://www.kaggle.com/datasets/jainaru/thyroid-disease-data),
        - which was provided by the UCI Machine Learning Repository. (https://archive.ics.uci.edu/dataset/915/differentiated+thyroid+cancer+recurrence)
    - This data under the license CC BY 4.0 ATTRIBUTION 4.0 INTERNATIONAL Deed (https://creativecommons.org/licenses/by/4.0/) so it is free and ethical to use.

- **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.**
    - I would first clean this data.
    - explore the data to see what I am working with and if any relationships. And select all or a few of the features which are more relevant. Find one feature that will be the binary marker for having and not having Thyroid Disease, which in this data set will be the 'Thyroid Function' column.
    - Then standardize the data to make it compatible to be inputted into the predictive model.
    - Will split the data set into a training dataset (80% of the total) and a testing dataset (20% of the total). I will design and train the model til I get the desired accuracy.
- **What are your deliverables?**
    - I will deliver a GitHub link to my code, slide deck, and project report.