STAT 380 PROJECT PROPOSAL
[Gianna Delorenzo & Melissa Kim]

Dataset
[Drug Reviews] https://archive.ics.uci.edu/dataset/461/drug+review+dataset+druglib+com

Goals and Motivations
- To gain insight and understandings on patient reviews and experiences on various drugs
- To apply our understanding of machine learning tasks including regression, classification, and clustering to this dataset.
- To implement model training and evaluation with this dataset.
- **Research Question**: Can machine learning tasks predict the overall satisfaction of the patients with a particular class of drugs for a given medical condition?

About the Data
- **Response Variable**: Rating (Binary, overall patient satisfaction)
    - Negative Reviews: 1-6
    - Positive Reviews: 7-10
- **Snapshot of Explanatory Variables:** urlDrugName, condition, effectiveness

Machine Learning Models
- The type of Machine Learning models that we plan to use include:
    - **Logistic Regression** (Supervised Learning model)
        - Can we predict if a patient's review is positive or negative based on their feedback?
    - **Neural Network** (Supervised Learning model)
        - Can we predict a more nuanced outcome of the review based on patient attributes, dosage, and drug features?
    - **KNN** (Supervised Learning model)
        - Can we classify the sentiment of a new review based on sentiments of similar past reviews?
    - **K-means Clustering** (Unsupervised Learning model)
        - What are the different groups of patient experiences with specific drugs?
        - Can we identify clusters of patients who report similar side effects or benefits from a drug?

Related Work
- Drugs in hospitals
- Mental health effects as a result of drug consumption
- Drug reactions using NLP

Measuring Success
- We will be predicting the **accuracy** of the models to ensure they are working effectively.
- We will also be measuring success (with a focus on Classification) with the usage of:
    - **Precision** (to minimize false positives), **Sensitivity** (to minimize false negatives), **Confusion matrix,** and **AUC.**