Deliverable 1
Team Member: Shiqi Tan, Theresa Quan, Angela Hu

1. Choice of dataset

Our primary dataset is NELL: NELL is a dataset which provides access to information sources such as 2, 78, 388 clinical events, 17,898 symptoms and 230 medicines. We choose this dataset mainly because it is large enough and contains the parameter symptoms and medicines that we want. We are currently contacting the director of project NELL to get access of the data.
see website: https://www.nursing.emory.edu/pages/project-nell

supplementary datasets:

1. UCI ML Drug Review dataset(contains parameters: drugName, condition, review)https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon-winter-2018
2. 250k Medicines Usage, Side Effects and Substitute(contains parameters:Drug name, Adverse reactions and side effects, Drug interactions,Substitute drugs):https://www.kaggle.com/datasets/shudhanshusingh/250k-medicines-usage-side-effects-and-substitutes
3. symptom-disease related:
   https://www.kaggle.com/datasets/mirunaandreeagheata/medicinet-diseases-and-symptoms
   https://www.kaggle.com/datasets/uom190346a/disease-symptoms-and-patient-profile-dataset
   https://www.kaggle.com/datasets/niyarrbarman/symptom2disease
   https://www.kaggle.com/code/plarmuseau/symptom-disease-recommender
4. side effect data:
   a. https://www.rxlist.com/
   b. https://www.webmd.com/drugs/2/index
5. National Library of Medicine: https://pubmed.ncbi.nlm.nih.gov/
6. clinicalTrials.gov: https://clinicaltrials.gov/
7. Canada Institutes of Health Research: https://cihr-irsc.gc.ca/e/49941.html

2. Methodology
   a. **Data Preprocessing**
      i. **Inputs:** The column containing the respective drugs will serve as our Y_label, while the remaining columns containing symptom and patient information will be considered as X_features.
      ii. **Drug column:** The drugs are transformed into standardized codes using either the ATC Third Level or AHFS

(Pharmacologic-Therapeutic Classification System) coding systems.

1. This standardization ensures uniformity and consistency in drug representation.

iii. **Symptom column:**

1. Feature Representation: For symptom representation, we consider multiple techniques such as one-hot encoding, Bag-of-Words (BOW), or Term Frequency-Inverse Document Frequency (TF-IDF) based on the specific requirements of the task.
2. Feature Selection: To manage computational complexity and focus on the most relevant information, we narrow down our symptom features to the top 500 most popular symptoms. This selection ensures a balance between informativeness and computational efficiency.

iv. **Outputs:**

v. Step 1: Drug-Probability Vector Generation: The processed inputs are utilized to generate a probability vector for all drugs from the given drug set, considering the provided symptom set.

1. This vector signifies the likelihood of each drug being a suitable recommendation based on the input symptoms.

vi. Step 2: Scaling Using Sigmoid Function: To ensure the output probabilities fall within a consistent and interpretable range, a sigmoid function is applied. This function scales the output to a probability measure ranging between 0 and 1.

1. This transformation enhances the interpretability of the model's predictions regarding the confidence level associated with each drug recommendation.

b. Machine Learning Model

i. We will assess a total of three models and select the most suitable one from the following options(If time permits, we will also consider the alternative options):

ii. 1. Multi-class Logistic Regression:

1. Pros: works well when the relationship between the symptoms and drug recommendations is approximately linear.
2. Cons: Assumes a linear relationship, which might not capture complex patterns in the data. (It might not perform well if the data is highly non-linear.)

iii. 2. Decision Tree:

1. Pros: Can capture complex non-linear relationships in the data.
2. Cons: Prone to overfitting, especially if the tree is deep.

iv. 3. Random Forest:

1. Pros: Combines multiple decision trees to improve accuracy and reduce overfitting. Handles non-linearity well. Provides feature importance scores.
2. Cons: Can be computationally intensive and complex.

v. **Considering Alternatives:**

vi. Support Vector Machines (SVM):
1. Pros: Effective in high-dimensional spaces, versatile as different kernel functions can be specified.
2. Cons: Can be sensitive to noise in the data

vii. Gradient Boosting Machines (GBM):
1. Pros: Builds trees sequentially, focusing on the errors of the previous trees, leading to high accuracy.
2. Cons: Prone to overfitting

viii. bagging and boosting
1. xgboost model, expected to achieve better accuracy than decision trees

c. Evaluation Metric

Our project is a classification problem, where based on certain symptoms, our model will classify or recommend a set of drugs. Therefore, the primary evaluation metric that would be relevant here is the confusion matrix, accuracy, precision-recall, and logistic loss for the classification problem.

- Confusion Matrix and Related Metrics:
  - The confusion matrix will give a clear picture of the True Positives, True Negatives, False Positives, and False Negatives.
  - Accuracy = (TP+TN) / (TP+TN+FP+FN)
  - Precision = TP / (TP + FP)
  - Recall = TP / (TP + FN)
  - F1-Score = 2 * (Precision * Recall) / (Precision + Recall)

- Logistic Loss (Log Loss): Since our model outputs probabilities, log loss will give an idea about the confidence of the predictions. It penalizes wrong predictions that are made with high confidence.

- Jaccard Coefficient and F1 score: these are set-based metrics which measure the similarity between the predicted set of drugs and the true set of drugs.

Average Baseline:

We should have a simple model or heuristic-based system as a baseline. For example, a model that recommends the most popular

drugs for a given symptom can be a baseline. We then aim to beat this baseline with our sophisticated model

3. Application

   User Input:

   - The user inputs their symptoms, age and gender.
   - Input Method: Through a user-friendly interface on a landing page web app. They can choose symptoms from a dropdown menu or select from a list. Age and gender can be input through text fields and dropdowns respectively.

   User Output:

   - The user receives a list of recommended drugs based on their input symptoms.
   - Each drug in the list is accompanied by a score of 5, representing its effectiveness based on past reviews. The list is sorted based on these reviews.
   - Display Method: A list format where each drug name is clickable (maybe redirect to a page with more information about the drug). Beside each drug name, the score out of 5 is displayed prominently. There can also be an option for users to read the reviews or more details about potential side effects and drug-drug interactions.