MAIS 202 Deliverable 1

Jessica Coulson, Brianna Fan, Iris Sun

1. Choice of dataset:
   Our team will be using a combination of datasets found on Kaggle and other websites that provide different recipes based on cuisine preference as well as the ingredients used. We chose these datasets because they are accessible to the public and readily available to use. The following datasets from Kaggle will be used to begin building our dataset: recipe ingredients dataset, food.com recipes and reviews, and food.com recipes with search terms and tags.

2. Methodology:
   a. Data Preprocessing:
      For data preprocessing, we will do the following: handling missing data through techniques such as imputation, deletion and/or prediction models; encoding where we will convert categorical data, such as cuisine type, into numerical values using one-hot encoding or label encoding; and scaling numerical features to have similar ranges using methods like min-max scaling and standardization.

   b. Machine learning model:
      For the model, we will predict and retrieve a recipe for a user based on their preferences and past browsing history. We aim to employ a recommendation system that uses collaborative and content filtering, matrix factorization, and deep learning. We also hope to include support vector regression that analyzes information that provides appropriate details of recipes for users. It will then collect the necessary data from the analysis process and produce an optimal set of data for the recommendation. We might also consider a multilayer perceptron (MLP) model that reduces the error rate in the analysis process that improves the significance and efficiency of the overall recommendation system.

   c. Evaluation Metric:
      <u>Content Based Filtering:</u>
      - Similarity matrix: To evaluate the accuracy of the model in terms of content based filtering, we will consider using cosine similarity. This metric would be beneficial in regards to the accuracy of predictions by comparing data using vectors. A downside to using cosine similarity would be that the frequency of the user's inputs would not be taken into account – only the direction of the vector is considered and not its length. For example, it would be difficult to predict a user's interests depending on how often they search something.
      - Precision and recall with the use of an F1 score could also help train the model to make better predictions by identifying the ratio of what suggestions users liked while also identifying the proportion of liked recipes that were successfully recommended. There can be cases where the precision is high even though the model is not performing well due to negligence of the minority class.

Collaborative Based Filtering

- Jaccard similarity: This metric could be beneficial for improving the model's predictions by comparing how similar two users are using sets of data. This will allow for the model to recommend similar recipes to users who exhibit the same interests. However, it may not work well for very large sets of data.
- Root Mean Squared Error (RMSE): RMSE would help train the model through measuring the difference between predicted ratings and actual ratings. The lower the RMSE value, the better the predictions are. Extreme errors may affect the accuracy of the RMSE which can cause predictions to be not as good.

3. Application:
   We plan to integrate our model into a webapp. The user inputs their cuisine preferences, dietary restrictions, and preferred ingredients on a selection window. The recommendation system will use a hybrid of collaborative and content filtering to display a short personalized list of recipes for the user's meal. After choosing a recipe and cooking, the user will be asked to rate and review the recipe. These ratings will improve future recommendations by better understanding the user's preferences. The recipe recommendations would continue to be tailored to the user inputs provided as well as previous meals the user rated highly. The app will also provide an option for users to keep track of their meals to avoid repeating recipes too frequently. In terms of nutrition, we plan to keep track of the nutrients for each meal that is recommended.