Sprint 0: Crystal Gil Herrera

Requirements

Choose a subject area that interests you, identify potential problems within it you would like to address, and seek out potential datasets you could build solutions on. Address the following writing prompts with a paragraph each.

The Problem area: What is your area of interest? Within it, what challenges or opportunities could your project address?

My main project idea is to build a recipe recommendation tool that provides recipe options based on calories needed and by ingredients, so that it can help the user on their weight goals. An opportunity that I see is that the system could provide a breakdown of the nutritional values and calories so that these can be more easily inputted and tracked on an app like MyFitnessPal. This would help the user track macronutrients and calories based on their goals. I think another opportunity would be to provide a link to the recipe instructions so that the user can follow a guide if needed, which opens up the option to try new recipes. I think one of the challenges at this time is that I would only be able to account for an estimated amount of calories and macronutrients, rather than an exact measure as I would then need to account for the amount of grams per ingredient in the recipe.

The User: Who experiences these problems? How would they benefit from the outcomes of your project?

These problems would be experienced by people that are trying to lose, gain or maintain their weight, as they would have to track their calories and macronutrients for every meal everyday. They would benefit from the outcome by saving time on finding recipes that fit their calorie and nutritional needs. This would also benefit the user by providing more recipe ideas, as sometimes meals can get repetitive which can discourage the user from sticking to diets.

The Big Idea: How can machine learning bring solutions to these areas? Research how other people have approached the problem previously. Refer to the "Intro to Capstone" slides on synapse for an overview of different machine learning approaches.

Some of the research I found showed that machine learning has been used to create personalized diet recommendation systems based on certain input parameters, such as age, gender, daily meals, exercise intensity and weight goals, to then generate a diet plan for the day which includes breakfast, lunch and dinner. These plans included nutritional information for each meal as well as preparation instructions. Another approach that was also used was to recommend recipes based on preferences of the user given in the form of ratings and then comparing them.

The Impact: What societal or business value do you anticipate your project to add? If possible, try to quantify the scale of the problem (in dollars, in CO2, in time spent, ...)

The value of this project is that it would help users save time when planning their meals around their diet, which can be a time consuming task. Also, it could encourage home meal preparation which can help improve the diet quality of users.

The Data: Identify several possible datasets in this subject area and describe them at a high level. Include references. If you struggle to find more than one or two datasets, this might mean a Data Science approach to the problem will be challenging. Check in with your Educator.

Recipe datasets.
https://www.kaggle.com/datasets/paultimothymooney/recipenlg
Dataset of cooking recipes in Kaggle which includes about 2.2M recipes. This dataset includes some variables of interest such as the name of the recipe, ingredients, directions, recipe links, as well as the named entity recognition (NER) which can help match the name of the ingredients to their calorie and nutritional values in another dataset.

https://apify.com/dtrungtin/allrecipes-scraper/api/client/nodejs
An additional dataset would be from the website allrecipes.com which contains 113K original recipes and could be accessed through API.

Nutritional value datasets.
https://www.kaggle.com/datasets/georgiyfedorov/nutricients?select=CSV_VERSION.csv
https://fdc.nal.usda.gov/api-guide.html

The database from the USDA contains data for nutrient composition of the ingredients. I was able to find a dataset in Kaggle but this can also be accessed through the USDA Database API. This contains several features of interest such as calories, proteins, carbohydrates, lipids, fiber, and sugar values for approximately 8.7K different ingredients.

The Alternative: In a few sentences, summarize a problem in an alternative subject area that also interests you.

An alternative idea would be to build a recommendation tool for music to enhance concentration/increase productivity based on input and past history/preference. This would involve understanding the user preferences and recommending suitable music tracks or playlists that enhance concentration and productivity.