Project Proposal

1. The problem I would like to solve is to predict whether a restaurant is fine dining or not using a combination of factors: customer reviews, business attributes, and images of a restaurant's food/beverages, interior/exterior, & menus.

   I believe this is an interesting problem as the project will integrate multiple methods (image, text, and structured data) to influence binary predictions. Both traditional machine learning and deep learning approaches will be utilized to predict whether a restaurant is classified as fine dining.

   My proposed ensemble model could be used by potential restaurant owners to better understand how their restaurant is perceived by their photos and customer reviews as well as assist owners in positioning themselves in their desired market segment. This model could also be used by consumers seeking specific dining experiences and to aid in the discovery of new upscale dining options.

2. For this project I will be using the Yelp Open Dataset from https://www.yelp.com/dataset which includes:
   a. Business Data - tabular data on 150K businesses covering 11 metropolitan areas.
   b. Review Data - 6.9 million reviews by Yelp users with star ratings and text-based reviews.
   c. Image Data - 200K photos of food, interior, exterior, drinks, and menus.

   Since the Yelp Data includes information on multiple businesses such as nail salons, grocery stores, hotels, and hair salons, the business data is going to be filtered to restaurants only for this project. In addition the metropolis area will be filtered specifically to the Florida Tampa Bay area to reduce the size of the dataset.

   I will not be using the user data, tips data, or check-in datasets for this project as these would be more appropriate for a recommendation system and/or time series analysis problem.

3. Approach to solving the problem:
   a. This is a supervised learning problem as the available tabular data is labeled and the target variable is known (i.e., fine dining or not). The ensemble model will be trained to predict this label.
   b. This is a classification problem since the goal is to classify a restaurant as fine dining (1) or not fine dining (0). However one of the models in the ensemble will be using linear regression to predict the star rating of the restaurant based on its attributes.
   c. I am trying to predict whether a restaurant is fine dining or not based on images, reviews, and business data using a weighted average ensemble.

   d. Predictors:
      i. Images - using CNN models (OpenAI CLIP) for image classification purposes. Specifically classifying the image as fine dining or not.
      ii. Text Reviews - using a pre-trained BERT model for NLP purposes, with the goal of extracting sentiment analysis from the Yelp user reviews. Will focus on 1 & 5-star reviews to determine if the text is positive or negative.
      iii. Business Attributes - will use the XGBoost model to handle the business tabular data and ultimately predict overall restaurant star ratings. This predictor in contrast to the two above will be using linear regression.

   e. I will be using a hybrid approach which will use Deep Learning for the image classification (CNN) and review sentiment analysis (NLP) as well as Traditional Machine Learning (XGBoost for regression) for the business attribute analysis. Will also be using ensemble methods to combine the predictions from the different models.

4. The final deliverable will be an application deployed as a web service with an API. I plan to train and deploy my model on SageMaker and deploy it as an endpoint. This will be my backend model inference and the web framework will be designed using FastAPI.

   The user will be able to use the basic front-end user interface to upload images, submit reviews, and import restaurant tabular data (i.e., a flat file/csv) which FastAPI will process and then send to the SageMaker endpoint for predictions (fine dining or not).

5. Computational Resources
   a. Processing Power (CPU) - moderate CPU resources for handling tabular and text data. Will use Google Colab or SageMaker for training and evaluating my datasets. Need to determine what kind of charges may accrue using SageMaker and S3.
   b. Memory - minimum of 16 GB of RAM which is available on my personal laptop to handle the large datasets.
   c. Specialized Hardware such as GPUs - these will be necessary for training my deep learning models (e.g., image classification (CNN) & text sentiment analysis (BERT)). SageMaker will come into play again for training the deep learning models as the availability of GPUs in Google Colab is not consistent.