Report Submission

1. Problem Description
   The goal of this project is to predict the number of calories in McDonald's menu items based on their nutritional information and food category. I chose this dataset because I have been eating at McDonald's recently and started to feel a bit heavier—so I got curious about how many calories I might be consuming.

2. Dataset
   The dataset is taken from Kaggle (https://www.kaggle.com/datasets/mcdonalds/nutrition-facts/data), and contains 260 menu items from McDonald's with 24 features, like Total Fat, Protein, Carbohydrates, and food Category, etc.

   Data Cleaning
   I noticed that the "Serving Size" column contained text values such as "113 g" or "12 fl oz". To standardize this, I wrote a regular expression function to extract the gram values. If the value was in ounces or milliliters, I converted it to grams accordingly. I added a new column Serving Size (g) for consistent numerical comparison.
   I also removed any rows with missing values.

   Feature Exploration
   As there are 24 features and to find out which features were most important for predicting Calories, I calculated the correlation between each feature and the target variable. Therefore, I selected the features based on the strong correlation with calories.
   In addition, I one-hot encoded the Category column (e.g., Burgers, Salads, Beverages) to include categorical information about the type of food item.

3. Modeling Approach
   To predict the number of calories in each food item, I used a Linear Regression model, which is a common and interpretable algorithm for numerical prediction tasks. I split the dataset into two parts: 70% for training the model and 30% for testing it.

   Since the features had different units and scales (e.g., grams, mg, percentages), I applied standardization using StandardScaler from sklearn.
   To explore the importance of food types, I tested two versions of the model:
   Model A: Without Category
   I only used numeric nutrition values (e.g., Total Fat, Protein, Sodium, etc.).
   Model B: With Category

I added the Category column using one-hot encoding, which created binary columns for each food type (like Beverages, Breakfast, etc.). This helped the model learn if certain categories tend to have higher or lower calories.
After comparing the results, I found that Model B: With Category feature performed slightly better in terms of R2 score and error metrics.
I also tried polynomial regression (degree = 2), but the performance was slightly not good.

4. Evaluation

To assess the performance of my calorie prediction model, I used these regression evaluation metrics: R2 score, MAE, MSE.

Model	R2 Score	MAE	MSE
Linear Regression (Without Category)	0.9995	4.11	32.63
Linear Regression (With Category)	0.9996	3.81	26.35
Polynomial Regression (degree = 2)	0.9994	4.51	42.24

While the polynomial regression slightly underperformed compared to linear regression in this case, it is still very accurate. However, since linear regression is simpler and performs slightly better, I chose it as the final model.

5. Reflection

One of the main challenges I faced was understanding which features I should choose for the calorie prediction. At first, I thought Serving Size would be an important factor, but during correlation analysis, it showed a very low relationship with calories. So, I learned that the importance of data exploration and not relying on assumptions. I also learned how to: clean and preprocess real-world data, dentify useful features using correlation, compare models using R^2, MAE, and MSE, save and reuse trained models for future predictions. I would also like to explore other models like Random Forest or XGBoost next time.