Background

I am a recent engineering graduate from Ecole Polytechnique de Tunisie, currently working as a quantitative data scientist.

Participation in the Alim' Confiance Challenge

I joined the Alim' Confiance challenge on a friend's recommendation and a keen interest in exploring various challenges across different domains, particularly attracted by this challenge's societal impact.

Building the Winning Model: A Breakdown

Key steps and choices in building the model include:

A. Data Acquisition and Preparation

- **External Data Integration:** The script uses the "Sirene: Fichier Stock-Etablissement du 26 Mars 2024" dataset to enrich the existing data with details about French establishments, adding valuable context.
- **Data Cleaning and Preprocessing:**
  - Dropping columns with a high percentage of missing values or only one unique value.
  - Converting dates to datetime format, extracting day, month, and year.
  - Splitting 'activitePrincipaleEtablissement' into sections, divisions, groups, classes, and subclasses.
  - Imputing missing values with the mean.

B. Feature Engineering

- **Location-Based Features:**
  - Creating a binary feature 'paris' to indicate if an establishment is in Paris.
  - Categorizing address type into 'rue', 'av', 'ecole', or 'autre'.
  - Extracting latitude and longitude from 'geores' to calculate distance to a fixed point in France.
  - Applying KMeans clustering for geographical grouping of establishments.
  - Calculating region center coordinates using mean latitude and longitude.
- **Activity-Based Features:**
  - Standardizing and splitting 'APP_Libelle_activite_etablissement' into individual activities.
  - Creating binary features for various food-related categories.

  - Counting the number of activities per establishment.
- **Other Features:**
  - Splitting 'Numero_inspection' into two numerical features.
  - Using 'Code_postal' for region and sub-region information, and a count feature.
  - Creating a count feature based on 'SIRET'.
  - Extracting month, day, day of the week, and season from 'Date_inspection'.

C. Encoding Categorical Features

- **Label Encoding:** Encoding all categorical features (except the target) using LabelEncoder.
- **Target Encoding:** Encoding the target variable 'Synthese_eval_sanit' with a custom numerical mapping.

D. Model Training and Evaluation

- **XGBoost Classifier:** Selection of XGBoost for its capability with mixed data types and strong performance.
- **Hyperparameter Tuning:** Setting specific hyperparameters including learning rate, max depth, number of estimators, and regularization parameters.
- **Evaluation Metrics:** Using accuracy and confusion matrix for model performance evaluation on the validation set.

E. Prediction and Submission

- Training the model on the entire dataset for predicting the test set 'Synthese_eval_sanit'.
- Converting predictions back to original categories using a reverse encoding dictionary.
- Generating a submission file with predicted categories for each establishment.

Technical and Modeling Choices

- **XGBoost:** Ideal for handling various data types and robust performance.
- **Feature Engineering:** Extensive process critical for extracting meaningful information.
- **External Data:** Enhances the feature space and context, contributing to performance.
- **Hyperparameter Tuning:** Optimizes performance and prevents overfitting.

Potential Improvements

- **Feature Selection:** To identify key features, improve performance and efficiency.
- **Alternative Models:** Experimenting with models like Random Forests or Support Vector Machines for comparative insights.
- **Evaluation Metrics:** Adding precision, recall, and F1-score for a comprehensive performance understanding.

System Specifications Used in the Competition

In The competition moderate hardware was utilized, including an Intel i5-9300HF CPU, 8GB RAM, and an NVIDIA GTX 1050 GPU.

Suggested Improvements for Future Competitions

- **Display Private Scores:** Show private scores of all submissions post-competition.
- **Late Submissions:** Allow for submissions after the official deadline.