ML Project – Restaurants in France

# Overview

The objective of this project is to predict whether a restaurant will succeed based on a dataset obtained from TripAdvisor.

For the purpose of this analysis, a restaurant is considered successful if it receives more than 100 votes with an average score higher than 4.

This prediction can assist in understanding the factors that contribute to the success of restaurants and can be leveraged by restaurant owners or stakeholders to improve their services or strategies.

A restaurant with tables and chairs

Description automatically generated

# Data Preparation

The dataset, originally encompassing restaurants across Europe, was filtered to focus specifically on France, resulting in a dataset of **6,522,096** restaurants.

**Key Data Preparation Steps:**

1. **Data Filtering and Focus**:
   * The dataset was narrowed down to include only restaurants located in France.
2. **Reduction of Large Categories**:
   * Simplified the dataset by removing redundant columns like Country, Popularity\_Detailed, and address-related columns, focusing on city-level information instead.
3. **Categorization and Transformation**:
   * **price**\_**level** - converted symbols (e.g., €, €€-€€€, €€€€) to integers and categorized them.
   * **meals** – this column listed different types of meals offered by the restaurants (e.g., Brunch, Lunch, Dinner). Since a restaurant could offer more than one type of meal, I created dummy columns for each meal type.
   * **Boolean** columns (gluten\_free, vegetarian\_friendly, vegan\_options, and claimed) transform into binary format and marked them as categorical.
   * **awards** - Processed by standardizing and creating dummy columns for different award types.
   * **cuisines** - Simplified the cuisines column by retaining the 15 most common cuisines, grouping others under "Other," and creating dummy columns to reflect multiple cuisine types offered by each restaurant.
   * **openning\_hours** - Extracted key features from complex opening hours data: opens\_days\_count, open\_on\_weekend, and total\_open\_hours.
4. **Handling Data Duplication**:
   * Identified and retained brands restaurants appearing more than 10 times.
   * Removed duplicates for non- brands restaurants within a 100-meter radius. For this, I use the latitude, longtitude data I have.
5. **Creating the Prediction Column**:
   * Defined the target variable high\_rated\_popular based on whether a restaurant had more than 100 reviews and an average rating of 4 or higher.

# Exploratory Data Analysis (EDA)