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

# Domain Knowledge

Understanding the restaurant industry is crucial for predicting restaurant success. Several key factors were considered:

* **Customer Reviews and Ratings**: The project considers success as a combination of two factors: high ratings and high number of voters
* **Location**: Geographical data was carefully analyzed, and duplicate entries within close proximity were removed.
* **Cuisine Offerings**: The dataset was refined to focus on the 15 most common cuisines, with others grouped as "Other" to accurately reflect their impact on success.
* **Operational Hours**: Availability during peak times, like weekends and dinner hours, is critical. Features such as opens\_days\_count, open\_on\_weekend, and total\_open\_hours were derived to capture these aspects.
* **Brand Presence**: Chain restaurants often have more resources and recognition, contributing to their success. Restaurants appearing more than 10 times were flagged as brands and treated accordingly.

Incorporating these industry insights into the data preparation process allowed for more relevant feature selection, ensuring that the predictive model reflects real-world dynamics in the restaurant industry.

# 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)

For the Exploratory Data Analysis (EDA), an in-depth report has been generated using various reporting libraries. You can access the detailed EDA report [here](https://drive.google.com/file/d/1l5nf-hdVOHQcVimm6-r8jKonBj_pltOH/view?usp=drive_link).

# Data Cleansing

## Filling Missing Values

I have filled the following values using KNN model:

* + **province** 18172
  + **claimed** 171
  + price\_**level** 29682
  + working\_shifts\_per\_week 69611

# Map

For better visualization I created a map.

The code to the map is [here](MAP.ipynb).

A map of france with orange circles

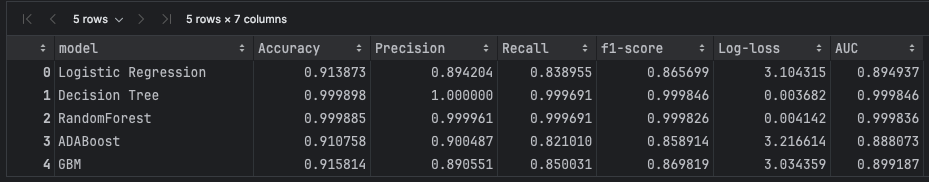
Description automatically generated

# Model Selection

The following machine learning models were evaluated for predicting restaurant success:

* **Decision Tree**
* **RandomForest**
* **XGBoost (XGB)**
* **Gradient Boosting Machine (GBM)**
* **Logistic Regression**
* **AdaBoost**

## Model Performance Metrics



## Analysis

* **Decision Tree** and **RandomForest** models achieved the highest accuracy and F1-score. They exhibit near-perfect accuracy, which raises a red flag for potential **overfitting**.
* **XGBoost (XGB)**, **Gradient Boosting Machine (GBM)**, and **Logistic Regression** showed slightly lower but still strong performance.
* **AdaBoost** performed well, though it had the lowest recall among the models, which could suggest it misses some positive cases.

Based on these results **XGBoost** or **Logistic Regression** would be strong candidates for final model selection due to their high accuracy and overall balanced performance across metrics.