What's Cooking

A Data Science Project

Daniel Felipe Barrera Suarez, Andres Julian Vargas Medina, Santiago Reyes Gomez

Universidad Distrital Francisco Jose De Caldas Bogota D.C.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Abstract

This project aims to predict the cuisine type of a dish based on its ingredients. Python was selected for its accessibility and robust libraries such as pandas and sklearn. Through meticulous data collection and prepossessing, the investigation reveals valuable insights into ingredient usage, cuisine calorie profiles, and the predominant culinary styles based on ingredient composition.

Background

With globalization, the proliferation of dishes and culinary traditions has become nothing short of impressive. Many of these dishes have seamlessly integrated into daily diets worldwide, prompting questions among chefs about the origins of certain dishes and potential alternative uses for familiar ingredients.

This data science project focuses on collecting a dataset of ingredients sourced from Yummli, a company dedicated to curating culinary data for home cooking. Once acquired, the choice of programming language was determined.

Python was selected for its straightforward variable handling and extensive libraries, which facilitate data manipulation and analysis. The project utilizes graphical and tabular representations to effectively manage and interpret data and outcomes.

Main Objectives

1. Classification of kitchen ingredients.
2. Provide accurate dish predictions based on ingredients.
3. Offer a comprehensive view of the usage of each ingredient across dishes.
4. Provide nutritional information for each ingredient listed.
5. Utilize classroom knowledge such as data handling and web scraping.

Materials

In this project, several Python libraries were employed for effective data handling, cleaning, visualization, and machine learning tasks:

General

- **Pandas** (`import pandas as pd`)
- **NumPy** (`import numpy as np`)
- **Regular Expressions** (`import re`)

Web Scraping

- **Requests** (`import requests`)
- **BeautifulSoup** (`from bs4 import BeautifulSoup`)

Charts

- **Matplotlib** (`import matplotlib.pyplot as plt`)
- **Seaborn** (`import seaborn as sns`)

Cleaning

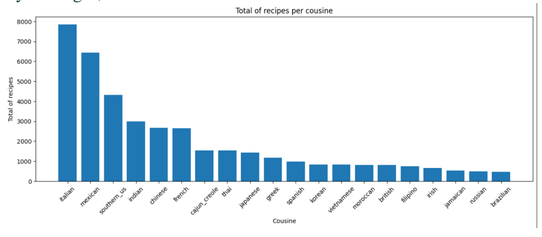
- **NLTK** (`import nltk`)
- **Unidecode** (`import unidecode`)

Machine Learning

for Machine Learning we will use Scikit-learn, this is a library used extensively for machine learning tasks. It offers a comprehensive suite of tools and algorithms for building and deploying machine learning models.

Metods

Thanks to the data collection from both Kaggle and web scraping, we have gathered sufficient data to leverage the capabilities of each library, facilitating the development of all project activities. In this case, the primary focus has been on the number of ingredients in each dish. Therefore, it is essential to clean up redundant data found in each list. By printing tables with ingredient information, we can determine how many dishes use each ingredient by dish type and country of origin, as well as their caloric content and other characteristics.



The use of this data within the project leads to the following: the data is already organized along with supplementary information from the train.json file, which provides a wide range of dishes categorized by region. This means that calling a specific dish code will provide us with its ingredients along with detailed information on caloric content.

	id	cuisine	ingredients	ingredients quantity	estimated calories
0	10259	greek	romaine lettuce black olive grape tomato garli...	9	44
1	25693	southern_us	plain flour ground pepper salt tomato ground b...	11	125
2	20130	filipino	egg pepper salt mayonaise cooking oil green ch...	12	281
3	22213	indian	water vegetable oil wheat salt	4	0
4	13162	indian	black pepper shallot cornflour cayenne pepper ...	20	601

The data resulting from this feature engineering process goes into the training phase, This process illustrates how to prepare, train, and evaluate a machine learning model using the specific libraries mentioned to process tabular data and text, and apply a classification algorithm to predict the type of cuisine based on ingredients and other dish characteristics.

Results

A key factor in the success of this project is the type of business questions it can primarily address. 1. What are the top 10 most common ingredients in all recipes?

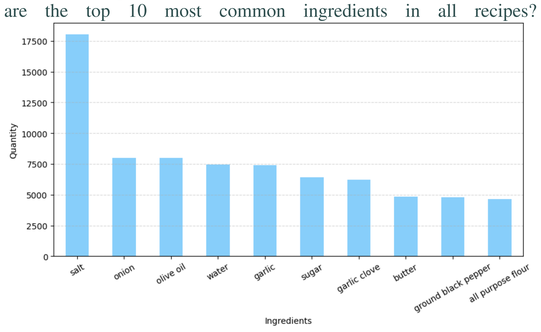


Figure 1: The top 10 most common ingredients

2. What are the top 5 cuisines with the highest mean estimated calories?

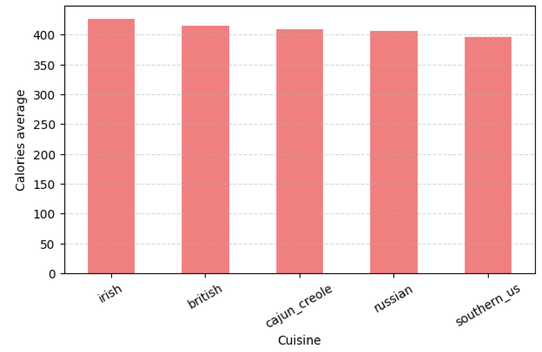


Figure 2: Top 5 cuisines with the highest mean calories

3. which ingredients are most common in each cuisine

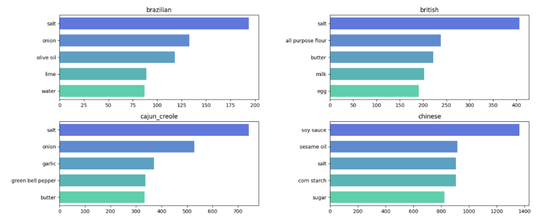


Figure 3: Top 5 ingredients most common per cuisine

Conclusions

- The data was successfully classified along with the ingredients and dishes.
- Salt is the most used ingredient in the vast majority of dishes worldwide.
- There is a higher probability that a good portion of the dishes to be prepared will be Italian.

References

[1] Wendy Kan. What's cooking?, 2015.