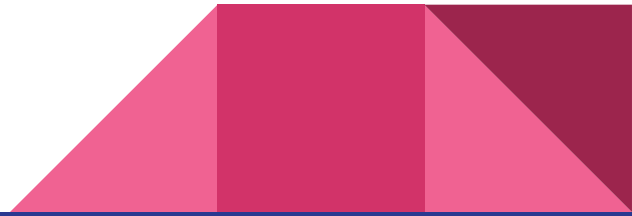
Cuisine Exploration

Erdős Institute, 2022 Bootcamp

Explorer Members: Alejandra Castillo, Anya Michaelsen,
Benjamin Sheller, Karan Srivastava

Overview

- Problem: How can we understand cuisines based on ingredients?
- Stakeholders:
  - Online recipe repositories
  - Apps that implement food / restaurant recommendation systems

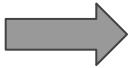


Data

Dataset provided by Yummly for Kaggle competition, “What’s Cooking?”

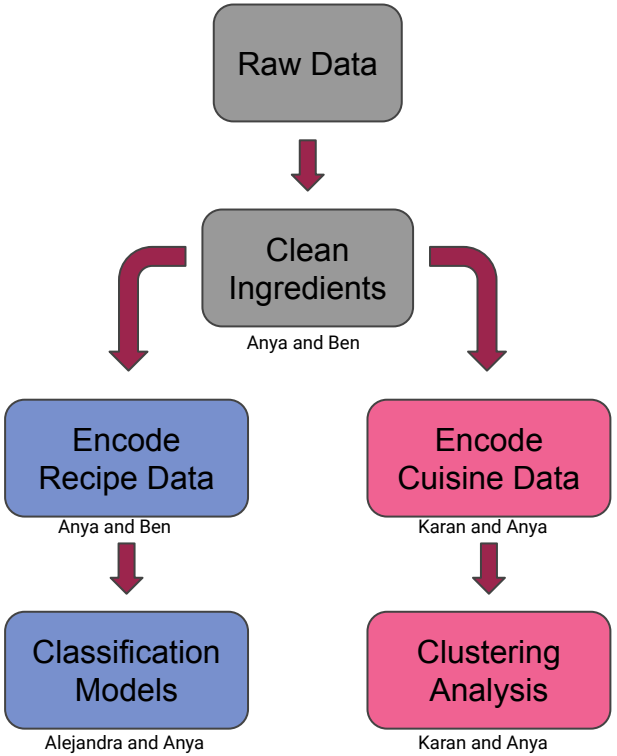
- 20 cuisines, over 6,000 unique ingredients
- Labeled training data, unlabeled testing data

```
{
  "id": 10259,
  "cuisine": "greek",
  "ingredients": [
    "romaine lettuce",
    "black olives",
    "grape tomatoes",
    "garlic",
    "pepper",
    "purple onion",
    "seasoning",
    "garbanzo beans",
    "feta cheese crumbles"
  ]
},
```



id	cuisine	ingredients
10259	greek	romaine lettuce
10259	greek	black olives
10259	greek	grape tomatoes
10259	greek	garlic
10259	greek	pepper
10259	greek	purple onion
10259	greek	seasoning
10259	greek	garbanzo beans
10259	greek	feta cheese crumbles

Workflow



Goals

- **Classify a recipe's cuisine** based on ingredients (supervised)
- **Understand cuisine similarity** through clustering (unsupervised)

Data Preprocessing and Encoding

- Ingredient Cleaning
  - Remove processing instructions e.g. “chopped”, “shredded” or “peeled and diced”
  - Remove modifying adjectives e.g. “small”, “medium”, “fresh”, “frozen”, “low-fat”, “(10 oz.)”
  - Combine variants of common items e.g. “pepper” for “black pepper”, “ground black pepper”
- One-hot encoding
  - Convert each unique ingredient into a binary feature (example below)

id	butter	eggs	feta cheese crumbles	black olives	garlic	purple onion	romaine lettuce	salt	pepper	garbanzo beans
10259	0	0	1	1	1	1	1	0	1	1
20130	1	1	0	0	0	0	0	1	1	0

Classification Models

Modeling

Classification Models

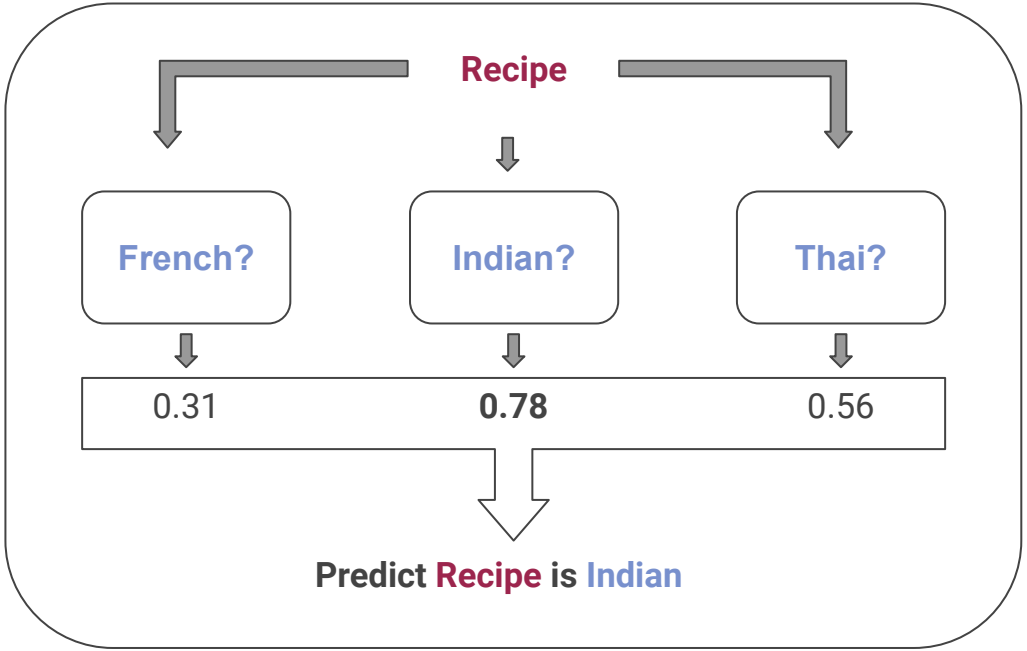
Classification models take sample points (recipes) and predicts classes (cuisine)

Model Type	Accuracy*
Logistic Regression	**77.9%**
LinearSVC	**76.8%**
Linear Discriminant Analysis	73.8%
k-Nearest Neighbors	66.5%
Random Forest	68.6%

*Accuracy based on Kaggle test dataset

Logistic Regression

Multi-Class Logistic Regression



77.9%

accuracy

Model Interpretability

- The model learns a linear combination of the features for each cuisine
- Large, positive coefficients showing strong associations with a cuisine class

Top 10 Ingredients for Thai

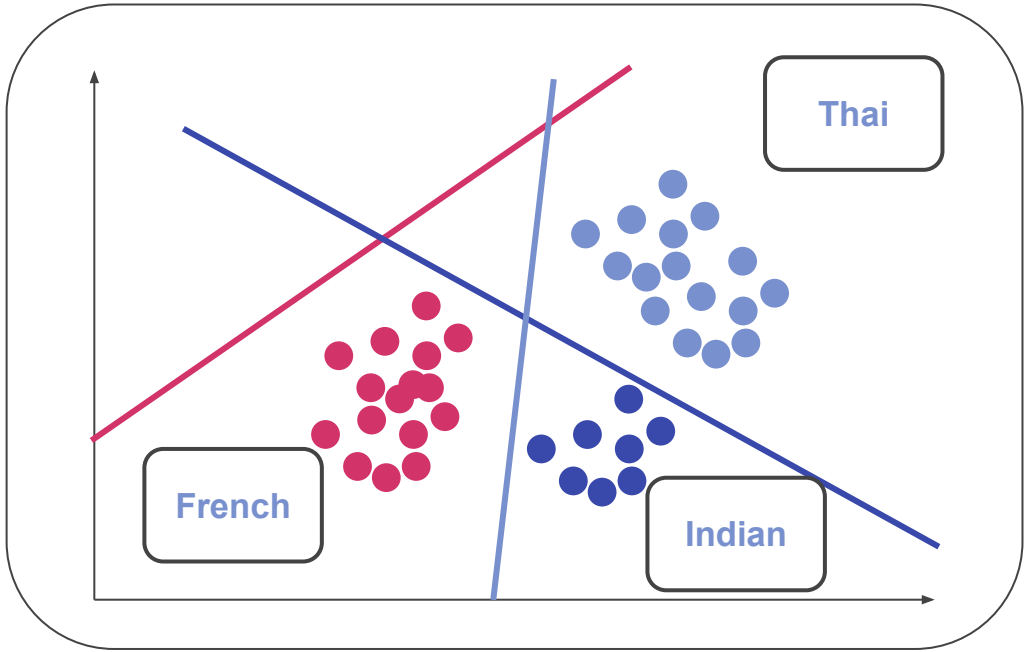
- Thai red curry paste, chunky peanut butter, unsweetened coconut milk, sweet chili sauce, fish sauce, red curry paste, thai basil, coconut milk, palm sugar, Thai fish sauce

Top 10 Ingredients for Italian

- polenta, arborio rice, italian sausage, marsala wine, mascarpone, spaghetti, gnocchi, pesto, marinara sauce, fettucine

Linear Support Vector Classifier

Multi-Class Supervised Classifier



76.8%

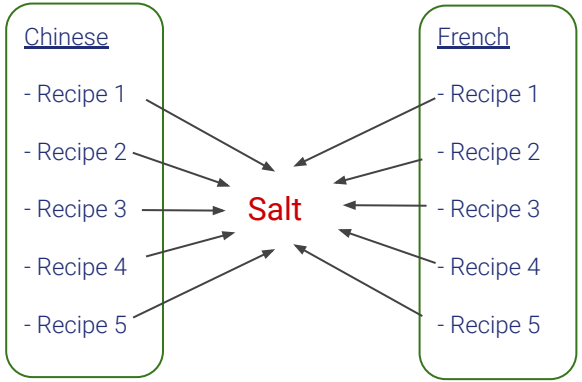
accuracy

Cuisine Clustering

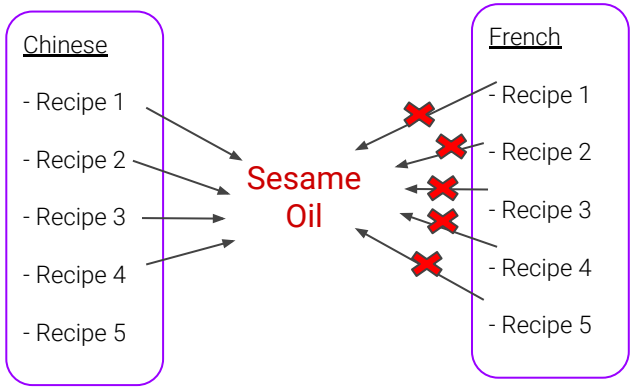
Cuisine Data Encoding

- Used Term Frequency - Inverse Document Frequency (TF-IDF) to encode data

Term Frequency



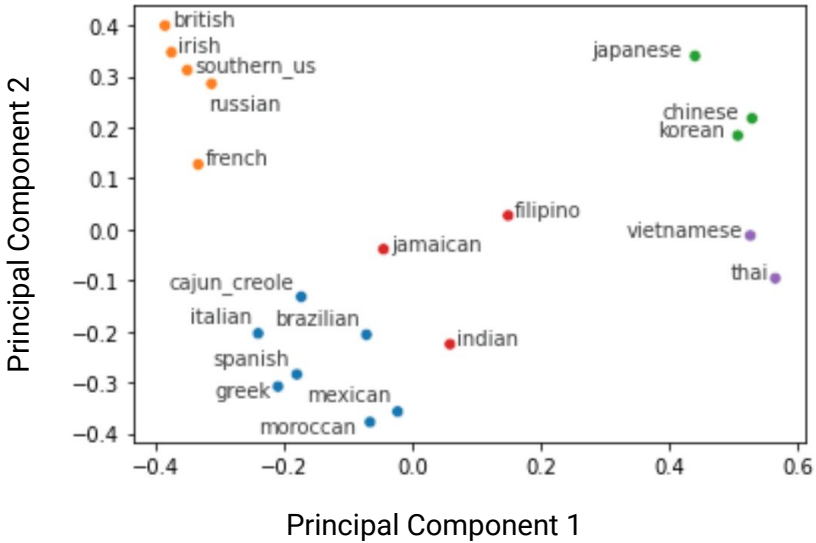
Inverse Document Frequency



- Ingredient x $\xrightarrow{\text{Encode}}$ T.f (x) · I.d.f (x)

Clustering and PCA

- Clustered the data using K-Means clustering
- Reduced the dimensionality using PCA to visualize the clusters

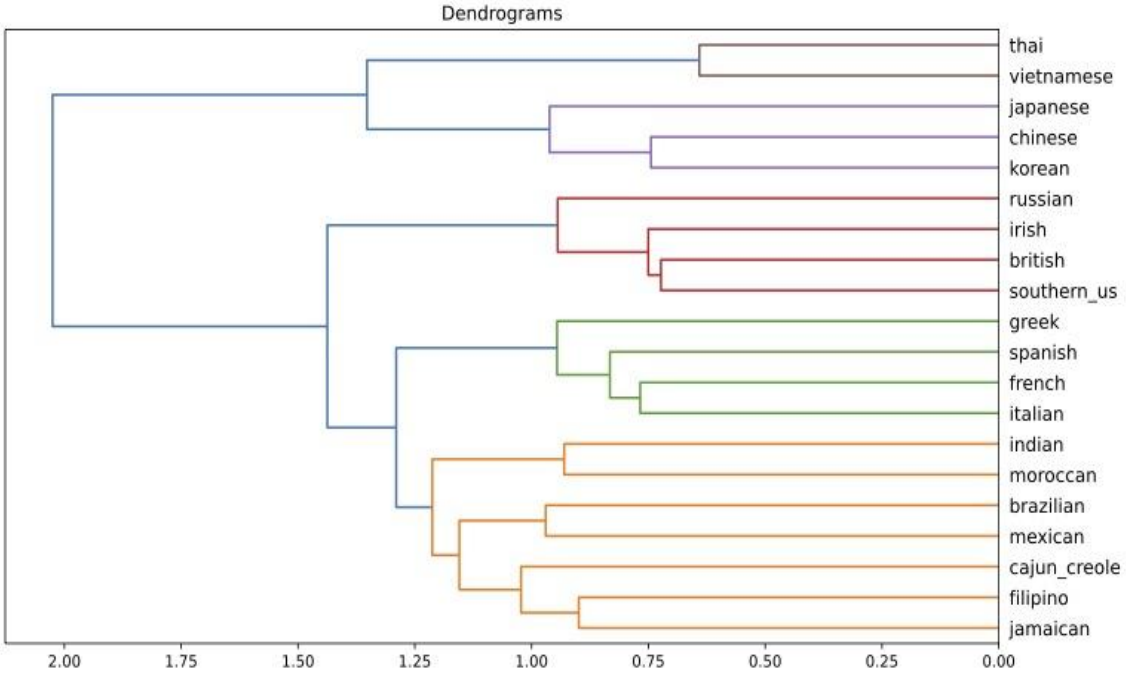


Principal Component Analysis

Higher Dimensional Data
(Hard to Visualize)

Lower Dimensional Data
(Easy to visualize)

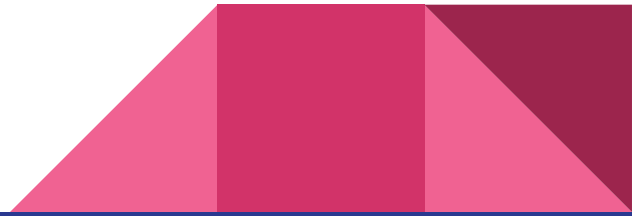
Clustering



- Used hierarchical clustering to find out how closely related any two cuisines are.

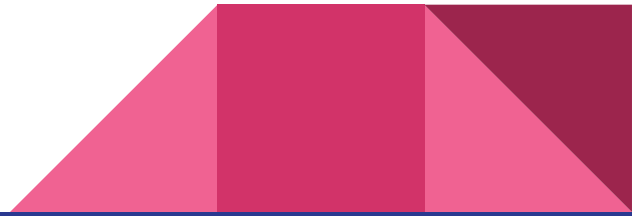
Deliverables

- Classification Model (≈78% accuracy)
  - Automatically classify recipes into cuisines based on ingredients
  - Use model coefficients to generate lists of typical ingredients for each cuisine
  - Suggest a cuisine type based on ingredients available, and potentially other ingredients to use
- Clustering
  - Recommend new cuisines similar to the users preferences



Future Work

- Further data cleaning
- Additional data, particularly from under-represented cuisines
- Inclusion of ingredient amounts and preparation methods



LinkedIn Profiles

http://www.linkedin.com/in/alejandra-m-castillo

https://www.linkedin.com/in/anya-michaelsen/

https://www.linkedin.com/in/benjamin-sheller-6b49251b9/

https://www.linkedin.com/in/karan-srivastava-a01ab8240/