What's in this, cinnamon?

Team Viper

Cynthia Lester, Sanal Shivaprasad, Ignat Soroko

https://github.com/sanalsprasad/ErdosTeamViper

Problem: Given a list of ingredients in a recipe, predict the cuisine of the recipe.

	Input	Output
Pesto Pasta	bow tie pasta, sweet peppers, olive oil, garlic powder, prepared pesto, sliced ripe olives, Parmesan cheese	Italian cuisine
Moroccan Lentils	olive oil, yellow onion, garlic, ground cumin, smoked paprika, ground coriander, cayenne, cinnamon, green lentils, vegetable broth, tomato paste, salt, black pepper, chopped fresh cilantro	Moroccan cuisine

Motivation: Help a recipe aggregator website implement a cuisine tag on recipes.

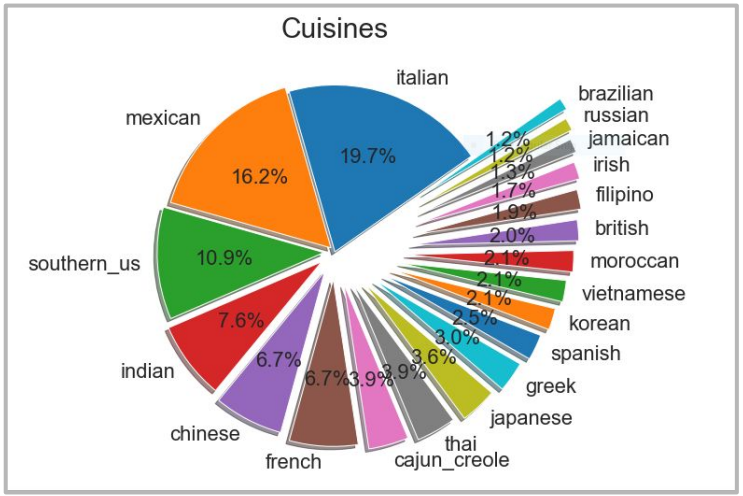
Dataset: From an old Kaggle competition.

Performance Metric: Accuracy

The Data

	id	cuisine	ingredients
0	10259	greek	[romaine lettuce, black olives, grape tomatoes, garlic, pepper, purple onion, seasoning, garbanzo beans, feta cheese crumbles]
1	25693	southern_us	[plain flour, ground pepper, salt, tomatoes, ground black pepper, thyme, eggs, green tomatoes, yellow corn meal, milk, vegetable oil]
2	20130	filipino	[eggs, pepper, salt, mayonaise, cooking oil, green chilies, grilled chicken breasts, garlic powder, yellow onion, soy sauce, butter, chicken livers]
...	...	...	...
39771	2238	irish	[eggs, citrus fruit, raisins, sourdough starter, flour, hot tea, sugar, ground nutmeg, salt, ground cinnamon, milk, butter]
39772	41882	chinese	[boneless chicken skinless thigh, minced garlic, steamed white rice, baking powder, corn starch, dark soy sauce, kosher salt, peanuts, flour, scallions, Chinese rice vinegar, vodka, fresh ginger, egg whites, broccoli, toasted sesame seeds, sugar, store bought low sodium chicken stock, baking soda, Shaoxing wine, oil]
39773	2362	mexican	[green chile, jalapeno chilies, onions, ground black pepper, salt, chopped cilantro fresh, green bell pepper, garlic, white sugar, roma tomatoes, celery, dried oregano]

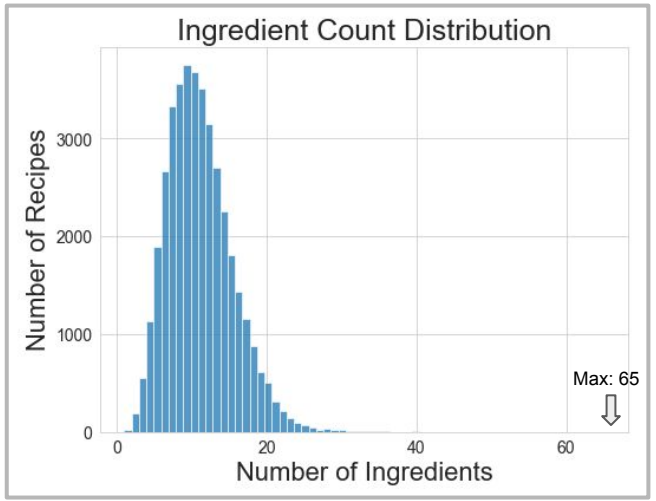
39774 rows × 3 columns



6,714 ingredients

Examples of ingredients:

- Pillsbury™ Crescent Recipe Creations® refrigerated seamless dough sheet
- sweet italian sausag links, cut into
- red food coloring
- food colouring



Cleaning the Data Set

Approach 1: train_trimmed

Removed ingredients that occurred less than 50 times.

About ~1000 features.

Ingredient Counts

	ingredient	greek	southern_us	...	moroccan	russian	total
3105	salt	572	2290	...	413	288	18049
2778	onions	185	482	...	280	145	7972
201	olive oil	504	312	...	412	50	7972
1971	water	143	686	...	182	111	7457
2522	garlic	216	259	...	143	20	7380
...	...	...	...	...	...	...	...
77	seedless cucumber	7	1	...	0	2	50
509	frozen spinach	6	0	...	0	0	49
11	plain whole-milk yogurt	5	2	...	1	2	49
251	pitted date	1	1	...	17	1	49
60	salad greens	2	5	...	0	0	49
...	...	...	...	...	...	...	...

Kept

996th Row

Approach 2: key_words

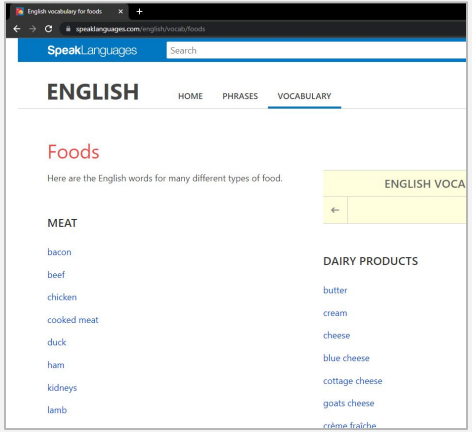
Replaced variations of ingredients with common keywords.

About ~560 features.

Keywords: achiote, adobo, agave, ajwain, ale, alfredo, allspice, almond, amaretto, amchur, …

Obtaining the Keywords

1. Initial list of "keywords": scrape https://www.speaklanguages.com/english/vocab/foods .



2. Find ingredients containing no "keywords":

Ingredient	Count
water	7457
soy sauce	3296
chili powder	2036
scallions	1891
corn starch	1757
...	
whole wheat pita bread rounds	1
clam sauce	1
sparkling sangria tradicional	1
dried hibiscus blossoms	1
lop chong	1

3. Use list to add words to "keywords".

For example, added corn.

Every ingredient that appeared at least 13 times overall or at least 10 times in one cuisine was replaced.

Ingredients containing no keywords were removed.

Multi-class classification models:

k Nearest Neighbors:

- euclidean metric
- jaccard metric (+)

Generative Methods:

- Linear Discriminant Analysis (+)
- Quadratic Discriminant Analysis
- Naïve Bayes

Decision Trees

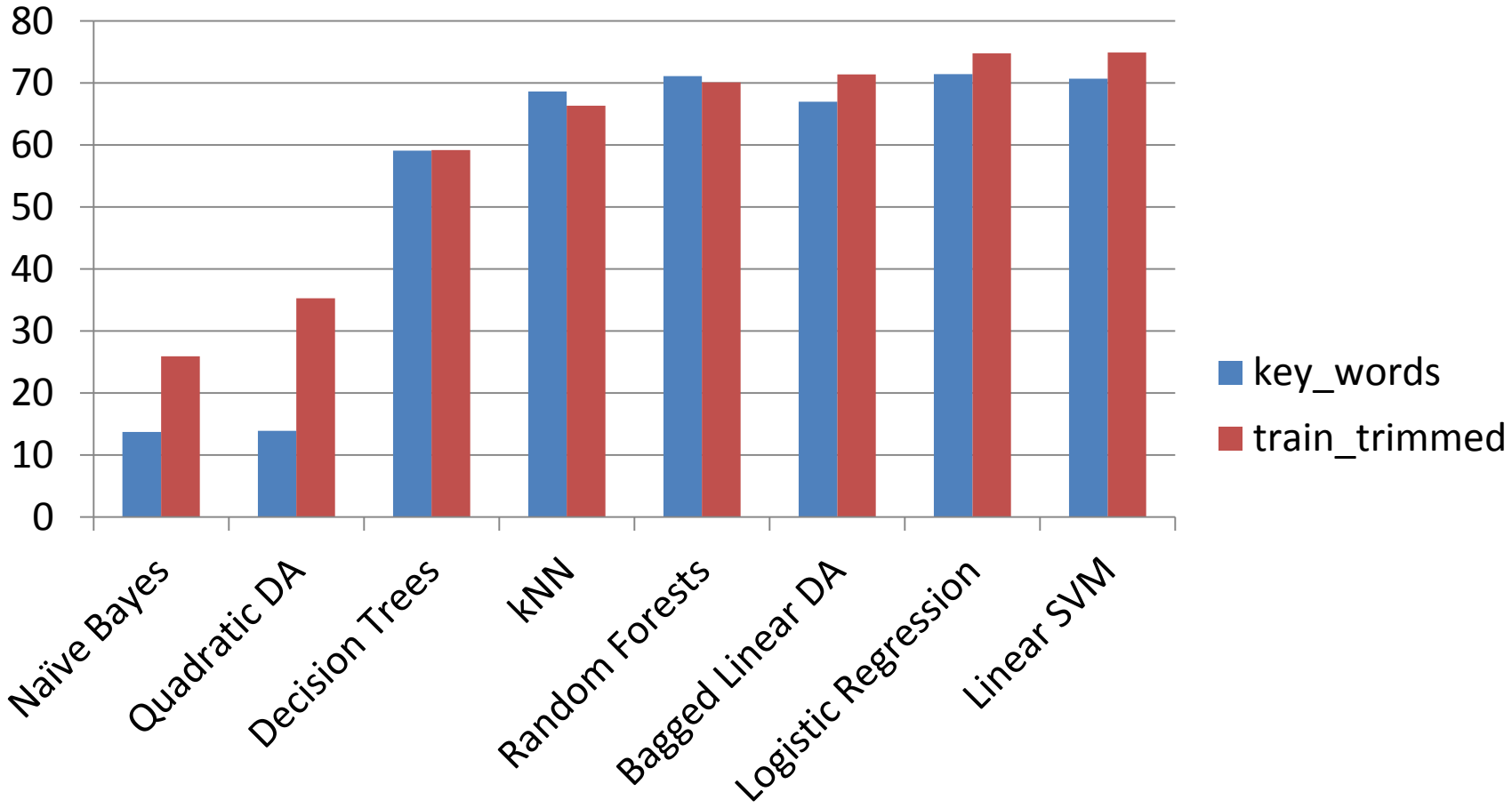
Logistic Regression

Random Forests

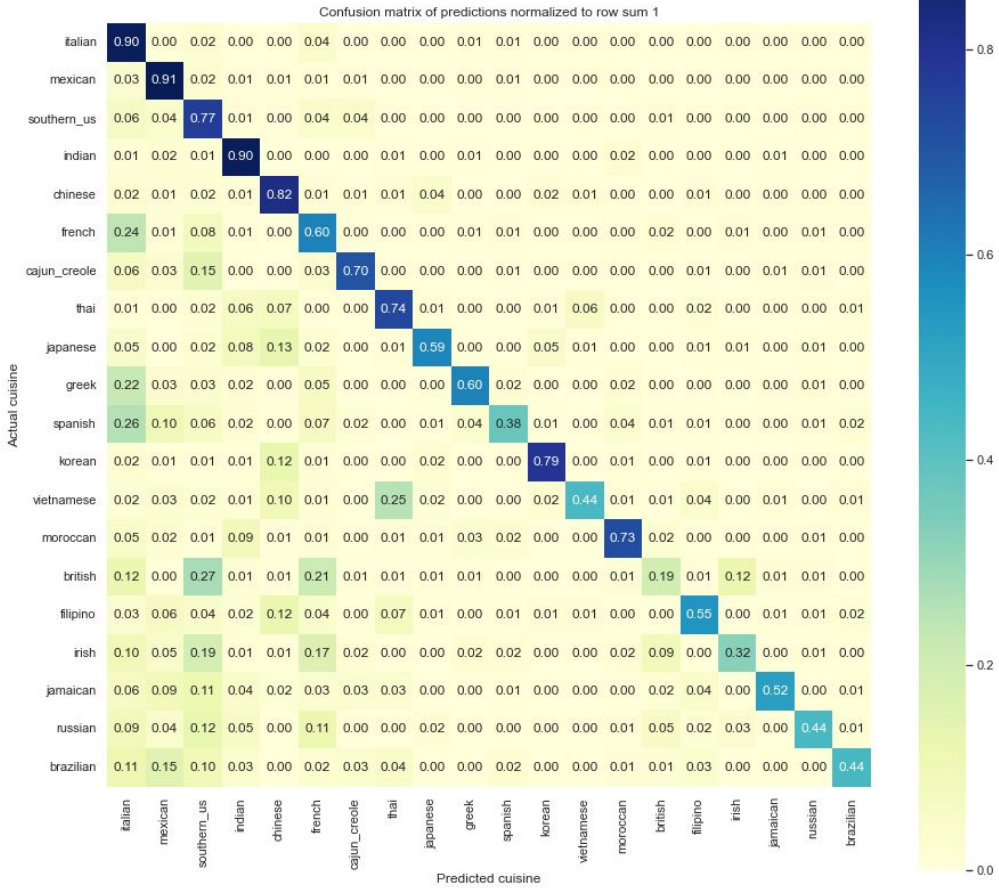
Linear Support Vector Machines

using **K-fold cross-validation** with 5 splits to fine tune hyperparameters and **Bagging** for Linear Discriminant Analysis and Linear SVM

Performance Analysis: Accuracy



Findings:



Future work: Implement a better data cleaning technique.

Thank You!

https://github.com/sanalsprasad/ErdosTeamViper