By: Marco Palermo, Liliya Panfilova ,Krupal Patel, Aarthie Senathirajah

WHAT’s Cooking?



3251 Term Project

What’s Cooking?

By Group 5: [Marco Palerm](https://q.utoronto.ca/groups/115137/users/368077)o, [Liliya Panfilova](https://q.utoronto.ca/groups/115137/users/364109) ,[Krupal Patel](https://q.utoronto.ca/groups/115137/users/365678), [Aarthie Senathirajah](https://q.utoronto.ca/groups/115137/users/364112)

Introduction

Have you ever found yourself recalling a delicious dish you consumed and trying to find it again but have no clue what it’s called? The best way to start would be to find the type of cuisine, you might ask yourself “If I don’t know what it is, how can I figure out where the dish is from?” Not to fear! Based on the ingredients it may be possible to identify the exact type of cuisine.

The objective of this study was to predict the type of cuisine based on the ingredients in the recipe. Using logistic regression best describes the relationship between a set of independent variables. and a categorical dependent variable. In the case of this study the independent variables being the ingredients and the type of cuisine as the dependent variable.

Exploring the Data

The data set is a json file from Kaggle, containing the recipe id, type of cuisine and the list of ingredients. The data was separated into a training data Json file and test data Json file. The training file contained 39773 recipes and the testing file contained 9943.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **id** | **cuisine** | **ingredients** |
| 39769 | 29109 | irish | [light brown sugar, granulated sugar, butter, ... |
| 39770 | 11462 | italian | [KRAFT Zesty Italian Dressing, purple onion, b... |
| 39771 | 2238 | irish | [eggs, citrus fruit, raisins, sourdough starte... |
| 39772 | 41882 | chinese | [boneless chicken skinless thigh, minced garli... |
| 39773 | 2362 | mexican | [green chile, jalapeno chilies, onions, ground... |

Fig 1, Example of data structure

The count for each type of cuisine was then produced from the training set and a bar plot was created to cleanly observe the distribution of the recipes from the training file. There was a total of 20 different types of cuisine, with Italian recipes being the most popular with 7838 recipes and Brazilian being the least popular with 467.

|  |  |  |  |
| --- | --- | --- | --- |
| italian 7838 | french 2646 | spanish 989 | filipino 755 |
| mexican 6438 | cajun\_creole 1546 | korean 830 | irish 667 |
| southern\_us 4320 | thai 1539 | vietnamese 825 | jamaican 526 |
| indian 3003 | japanese 1423 | moroccan 821 | russian 489 |
| chinese 2673 | greek 1175 | british 804 | brazilian 467 |

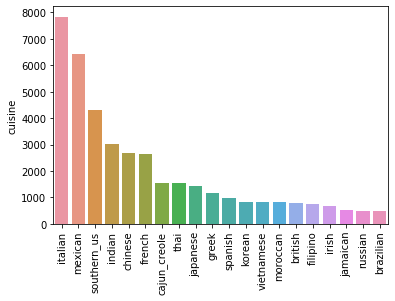
Fig 2. Number of recipes per type of cuisine

Fig 3 Barplot representing frequencies of recipes

These were appropriate sample sizes for each type of cuisine to perform an analysis. Using the training data, a list of unique ingredients and a list of complete ingredients were created. There were 6714 unique ingredients found across all recipes, and a total of 428275 ingredients were used across all the recipes in the training data. Additional analysis showed what the top ingredient was for each cuisine. Interestingly salt was predominantly the most commonly used ingredient across the various cuisine types as demonstrated in Fig (4)

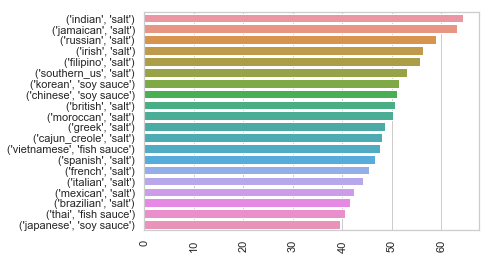


Fig 4- Top ingredient per cuisine type (percentage used per total cuisine recipes)

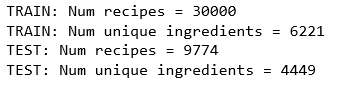
In Indian cuisine alone, salt is used in 64.4% of Indian recipes. Not surprising using the count function, salt was the most popular ingredient. The top 5 ingredients are displayed in Fig (5).

|  |  |
| --- | --- |
| **Ingredient** | **Count** |
| **salt** | **18049** |
| **onions** | **7972** |
| **olive oil** | **7972** |
| **water** | **7457** |
| **garlic** | **7380** |

Fig 5: Top 5 ingredients

Training and validation data

Because the test data was not labeled, the train dataset was split into train and validation datasets. This approach allowed us to calculate accuracy score without submitting results to Kaggle. Out of 39774 recipes, 30000 are used for training the model and 9774 for testing the prediction accuracy. The original dataset is being shuffled each time before split.



Preprocessing recipes

Original train dataset showed high variability of the ingredient records. Same or similar ingredients could have been written in multiple ways. For example, let’s observe the most popular ingredient - salt. There are 87 unique ingredients containing the word “salt”. Some examples are cooking salt, fine salt, iodized salt, low sodium salt, table salt. Also, there are multiple other ingredients containing “no salt added”, “salted”, “unsalted”, “salt free”. Moreover, some ingredients were written using capital letters, numbers and special characters. For example: “1% low-fat chocolate milk” or “33% less sodium cooked ham”.

To reduce the dimensionality of the dataset and improve the modelling the following data preprocessing steps were taken: removal of special characters, numbers and capital letters, removal of stop words, and stemming.

1. Special characters, numbers, capital letters

In the original dataset 31 unique special characters and numbers were found. Out of 39774 recipes 11440 contained special characters, 10080 contained numbers and 10277 contained capital letters. All ingredients were converted to the lower case and all the following characters and numbers were removed from the data:

['!', '%', '&', "'", '(', ')', ',', '-', '.', '/', '0', '1', '2', '3', '4', '5', '6', '7', '8', '9', '®', 'â', 'ç', 'è', 'é', 'í', 'î', 'ú', '’', '€', '™'] .

1. Stop words

In addition, it was discovered that certain recipes are written in a very detailed form, containing multiple descriptive words. For example: “condensed reduced fat reduced sodium cream of chicken soup”, where the useful words are “cream chicken soup” or “hellmann' or best food real mayonnaise”, where the only useful word is “mayonnaise”. Therefore, during data exploration the list of stop words was created. This list is used to remove certain words which do not contribute to the defining of a cuisine. This list includes measurements, such as “oz” or “g’, size, different ways of cutting, words like: “fine”, “good”, “ready”, “fresh”, etc.

1. Stemming

Finally, the stemming technique was applied to every word within the ingredients. All ingredients were written in multiple different forms like singular and plural, present and past tense, as an adjective or a verb. Moreover, some words were missing the endings and were written with spelling errors. The stemming technique allowed us to bring some words to the same form. It is reducing the word to the root form “stemma”, excluding the ending.

After applying all the above techniques of data preprocessing to training and test dataset, the number of unique ingredients was significantly reduced.





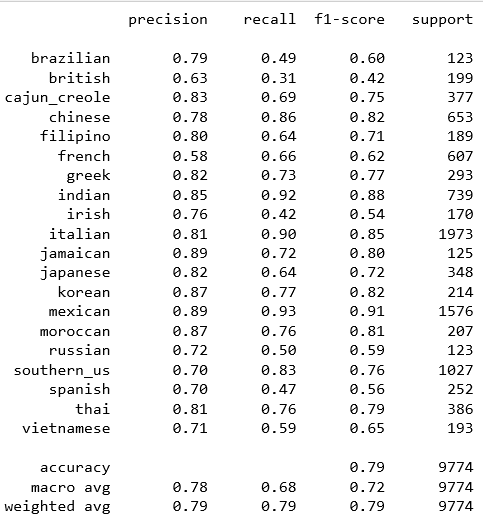
Encoding cuisines and recipes

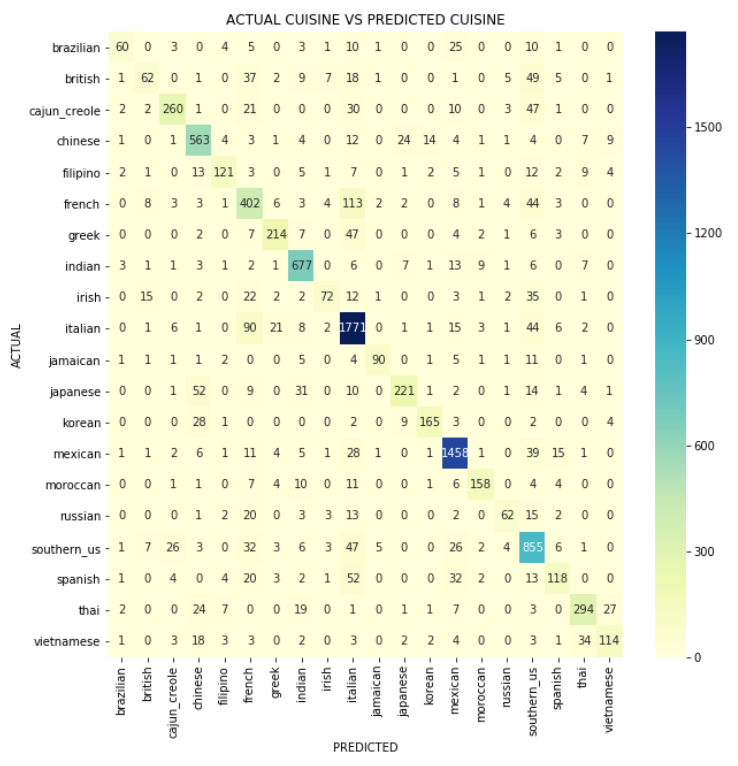
Cuisines and recipes had to be encoded because all the algorithms which we can potentially leverage can only consume numerical data. Cuisines were encoded in the simplest way, i.e. as integers. For this, we used LabelEncoder from sklearn.preprocessing package. Recipes were encoded using one-hot encoding where each recipe is represented by a vector of zeros and ones. Each zero or one stands for whether an ingredient was or was not used in the recipe. For this encoding, we used MultiLabelBinarizer from sklearn.preprocessing Python package.

We only encoded ingredients which were in the training set. Therefore, ingredients which were in the test set but not in the training set were not encoded and were ignored during model evaluation on the test set.

Logistic regression

We have a multiclass classification problem of predicting one of 20 cuisines by the recipe. The logistic regression analysis is used for classification problems with binary dependent variables. Therefore, we converted the original problem of predicting cuisine for a recipe into a problem of predicting likelihood of the recipe belonging to each of the 20 cuisines in the dataset with the highest likelihood cuisine being the final answer. To predict the likelihood of the recipe belonging to a specific cuisine, we created 20 logistic regression models, one for each cuisine. All these 20 models were created using the Logistic Regression object from sclera package. The below figure represents the performance of our model on the test dataset:



This classification report shows the precision (percentage of true positives within all predicted positives), recall (percentage of correctly predicted true positives) and f1-score (a weighted average of precision and recall). The overall accuracy of our model is 79%, which appears to be a good result given that we have 20 cuisines. A random guessing would only have 5% accuracy. We can observe the highest precision and recall for Indian, Italian and Mexican cuisines, probably because of the large amount of training data and the ratio of unique ingredients defining it. Indian, Italian and Mexican cuisines had 9.2%, 15.9% and 16.7% of unique for this cuisine ingredients respectively. British, French and Vietnamese cuisines had the worst f1-score due to the small data samples.

The confusion matrix represents how cuisines were predicted by our model. On the vertical axis there are actual cuisines (labels), while on a horizontal axis there are predicted cuisines. So, for example for Brazilian cuisine, 60 recipes were predicted correctly, 25 were falsely predicted as Mexican, 10 as Southern US, and 10 as Italian. We can see in this matrix that some cuisines probably have similarities, as higher values are observed on the both sides of the diagonal. For example, French cuisine was falsely predicted as Italian 113 times while Italian cuisine was falsely predicted as French 90 times.

Conclusion

In conclusion we were successfully able to create a model with approximately 79% accuracy of determining a type of cuisine based on ingredients. There are several ingredients such as Kimchi, turmeric and snails that are unique to a specific cuisine, which aids the model. Based on the values of precision and accuracy that the model appears to be enough.

However, as identified previously there are many cuisines with similar ingredients in result there was significant false prediction. Therefore, with more data it is possible to reduce error and create clusters of a combination of ingredients and quantities, specific to types of cuisines. For example, French and Italian cuisine both countries are regionally close therefore have access to similar ingredients, however the food tends to be prepared differently. This information could increase precision and accuracy of the model.

This model is a suitable steppingstone in predicting types of cuisines of a recipe using ingredients, with enough accuracy. With additional recipes the model will only improve and yield the intended results. Never forget a dish again!