COGNITIVE CLASS.ai

(https://cognitiveclass.ai)

From Modeling to Evaluation

Introduction

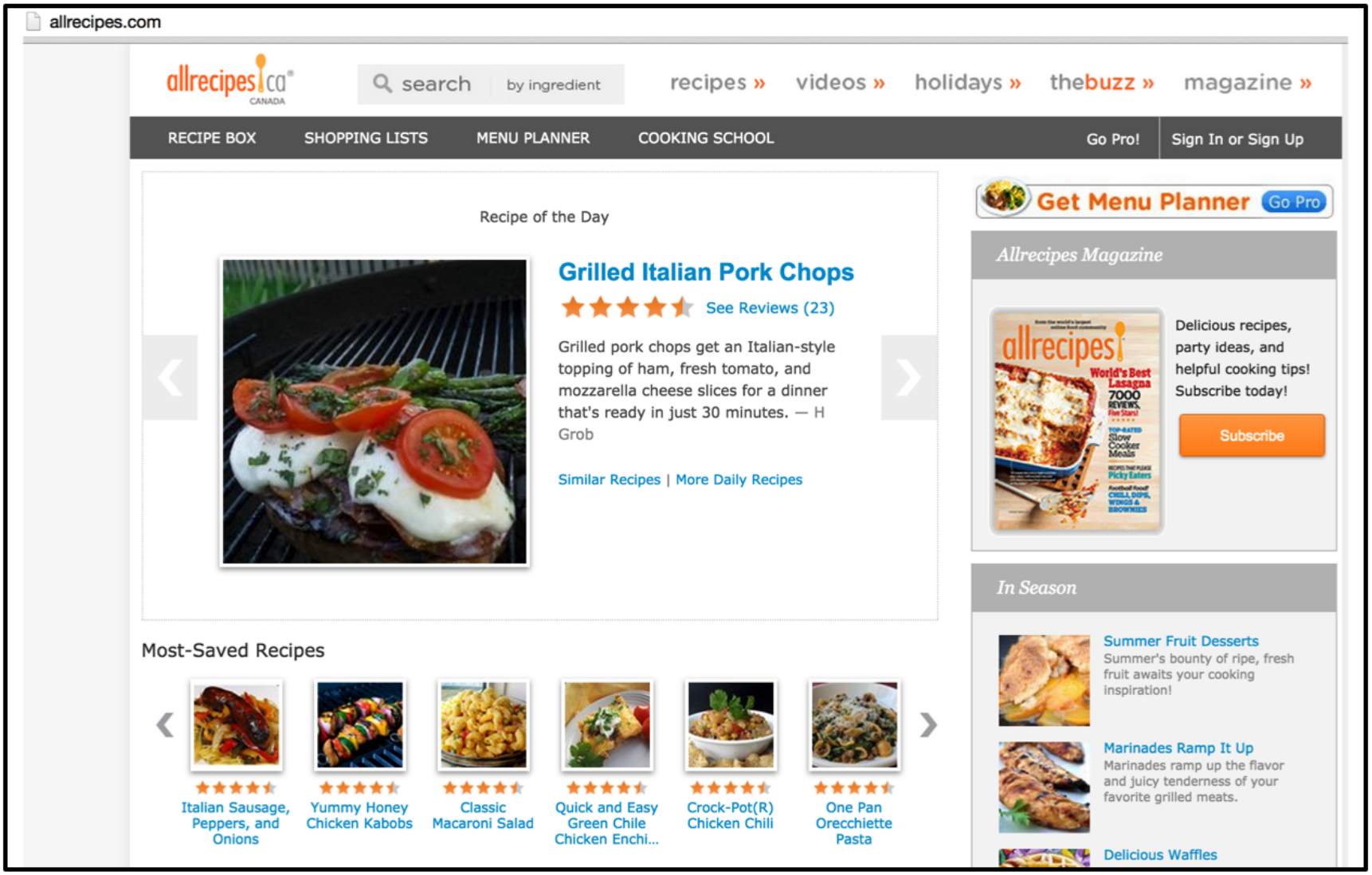
In this lab, we will continue learning about the data science methodology, and focus on the **Modeling** and **Evaluation** stages.

Table of Contents

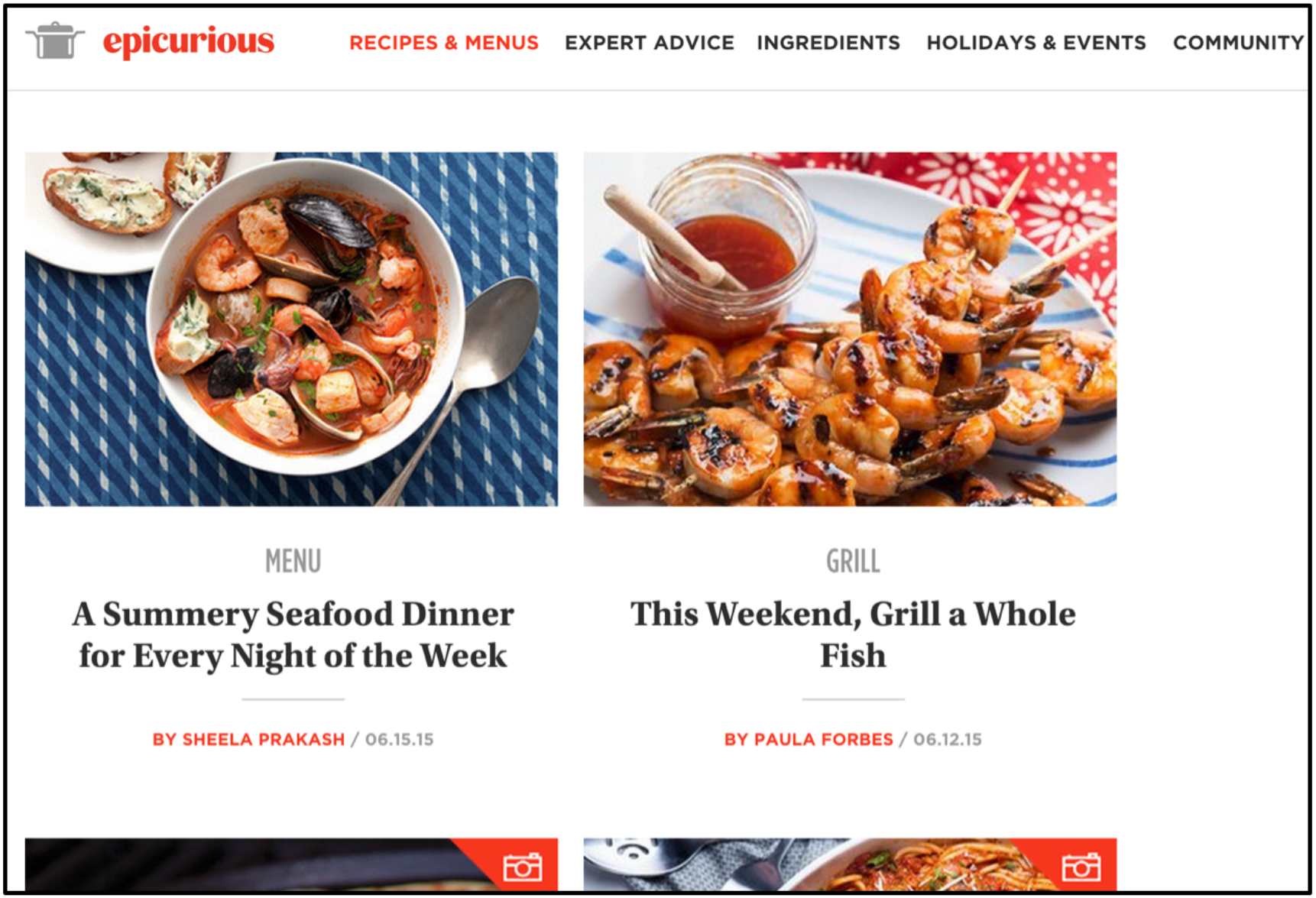
Recap

In Lab **From Understanding to Preparation**, we explored the data and prepared it for modeling.

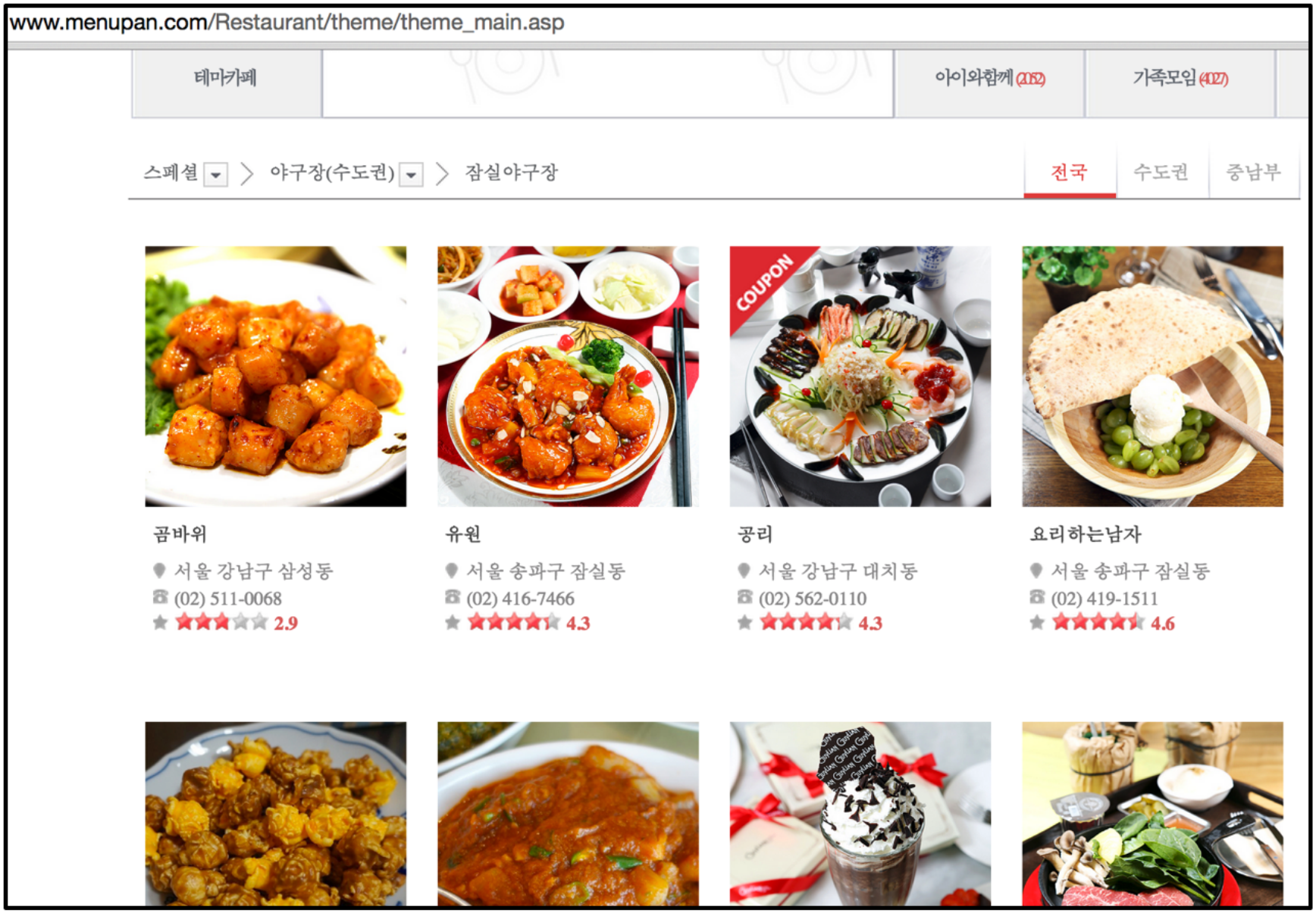
The data was compiled by a researcher named Yong-Yeol Ahn, who scraped tens of thousands of food recipes (cuisines and ingredients) from three different websites, namely:



www.allrecipes.com



www.epicurious.com



www.menupan.com

For more information on Yong-Yeol Ahn and his research, you can read his paper on Flavor Network and the Principles of Food Pairing (http://yongyeol.com/papers/ahn-flavornet-2011.pdf).

Important note: Please note that you are not expected to know how to program in R. This lab is meant to illustrate the stages of modeling and evaluation of the data science methodology, so it is totally fine if you do not understand the individual lines of code. We have a full course on programming in R, R101 (http://cocl.us/RP0101EN_DS0103EN_LAB4_R), so please feel free to complete the course if you are interested in learning how to program in R.

Using this notebook:

To run any of the following cells of code, you can type **Shift + Enter** to excute the code in a cell.

We already placed the data on an IBM server for your convenience, so let's download it from server and read it into a dataframe called **recipes**.

In [1]:

We will repeat the preprocessing steps that we implemented in Lab **From Understanding to Preparation** in order to prepare the data for modeling. For more details on preparing the data, please refer to Lab **From Understanding to Preparation**.

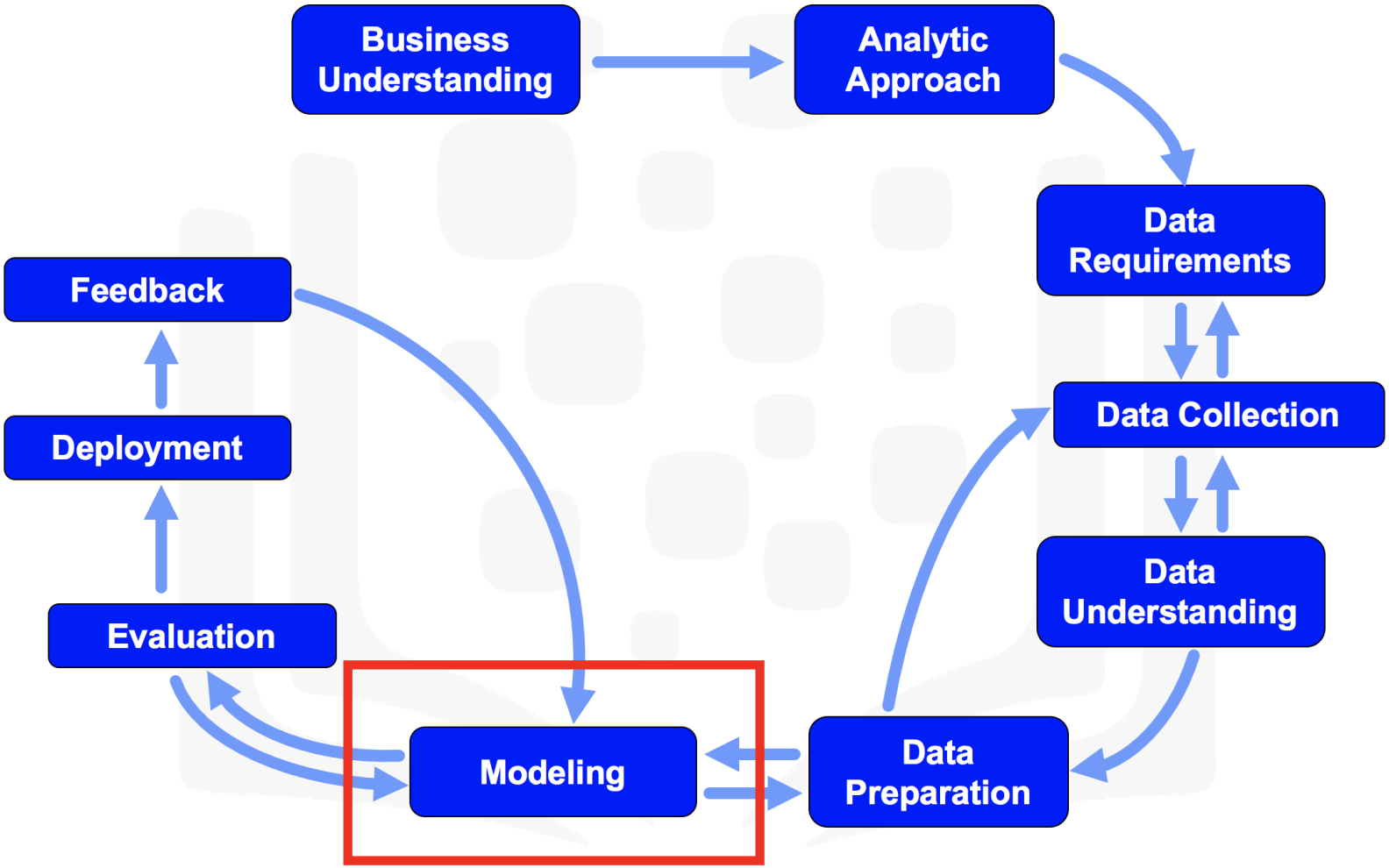
In [2]:

```
                american                    ita
lian                mexican
                 40150
3250                  2390
               french                        a
sian            east_asian
                1264
1193                   951
               korean                    cana
dian                indian
                  799
774                    598
             western                     chi
nese     spanish_portuguese
                450
442                    416
         uk-and-irish       southern_soul
food                jewish
                368
346                    329
              japanese                     ge
```

```
rman            mediterranean
                          320
289                       289
                     thai            scandina
vian            middleeastern
                          289
250                       248
  central_southamerican          eastern-eu
rope                    greek
                          241
235                       225
       english_scottish                carib
bean             cajun_creole
                          204
183                       146
easterneuropean_russian                 moro
ccan                   african
                          146
137                       115
            southwestern           south-ame
rica                vietnamese
```

```
                    108
103                      95
         north-african
                    60
```

Data Modeling


Business Understanding
Analytic Approach
Data Requirements
Data Collection
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Feedback

Let's start by importing and installing the R libraries relevant to decision trees

In [3]:

```
[1] "Libraries loaded!"
```

Check the data again!

In [4]:

A data.frame: 6 × 384

	cuisine	almond	angelica	anise	anise_seed	a
	<fct>	<fct>	<fct>	<fct>	<fct>	<
1	vietnamese	No	No	No	No	
2	vietnamese	No	No	No	No	
3	vietnamese	No	No	No	No	
4	vietnamese	No	No	No	No	
5	vietnamese	No	No	No	No	
6	vietnamese	No	No	No	No	

[bamboo_tree] Only Asian and Indian Cuisines

Here, we are creating a decision tree for the recipes for just some of the Asian (Korean, Japanese, Chinese, Thai) and Indian cuisines. The reason for this is because the decision tree does not run well when the data is biased towards one cuisine, in this case American cuisines. One option is to exclude the American cuisines from our analysis or just build decision trees for different subsets of the data. Let's go with the latter solution.

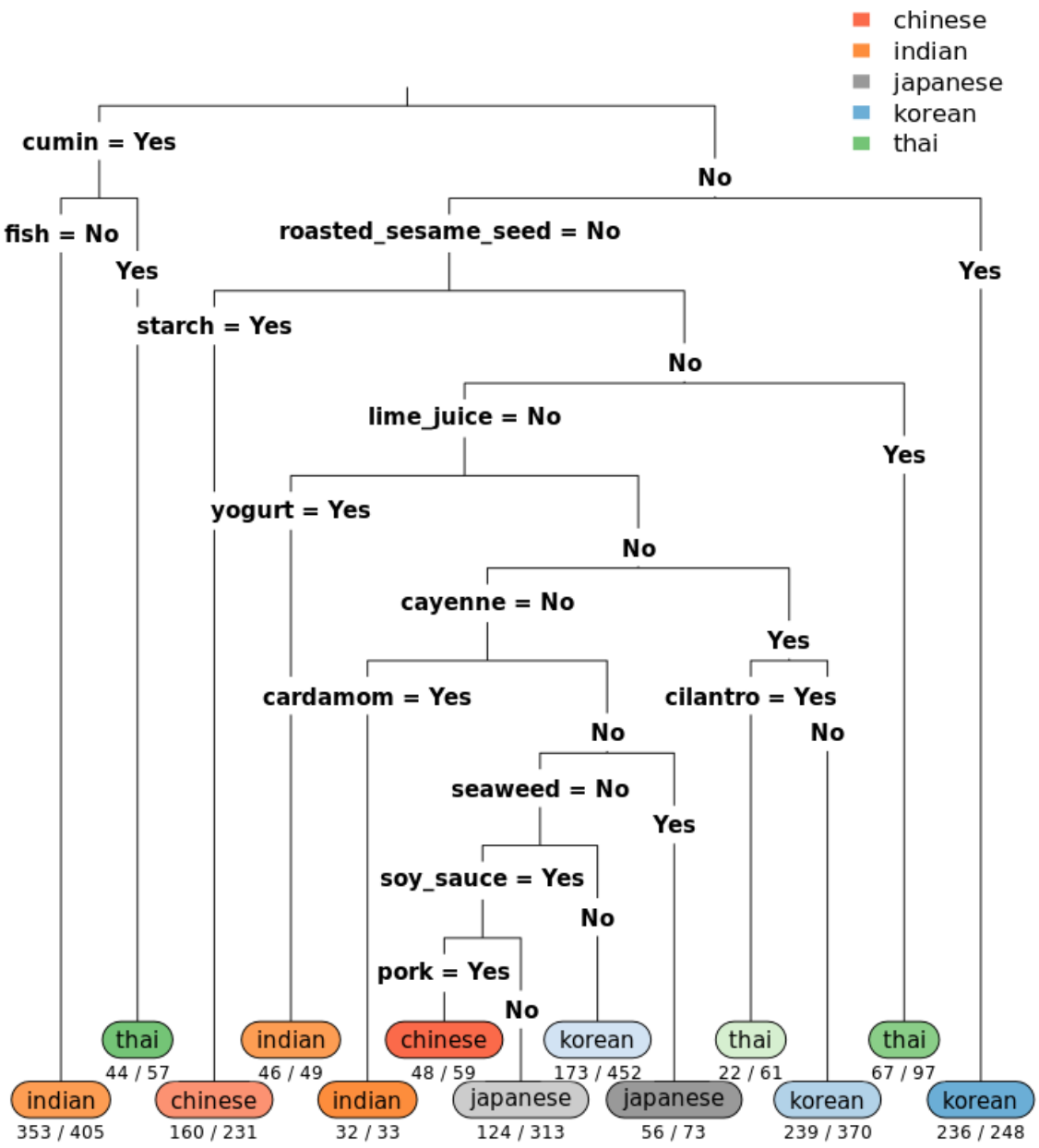
Let's build our decision tree using the data pertaining to the Asian and Indian cuisines and name our decision tree *bamboo_tree*.

In [5]:

```
[1] "Decision tree model saved to bamboo_tr
ee!"
```

Let's plot the decision tree and examine how it looks like.

In [6]:


chinese
indian
japanese
korean
thai
cumin = Yes
No
fish = No
Yes
roasted_sesame_seed = No
Yes
starch = Yes
No
lime_juice = No
Yes
yogurt = Yes
No
cayenne = No
Yes
cardamom = Yes
No
cilantro = Yes
No
seaweed = No
Yes
soy_sauce = Yes
No
pork = Yes
No
thai
44 / 57
indian
46 / 49
chinese
48 / 59
korean
173 / 452
thai
22 / 61
thai
67 / 97
indian
353 / 405
chinese
160 / 231
indian
32 / 33
japanese
124 / 313
japanese
56 / 73
korean
239 / 370
korean
236 / 248

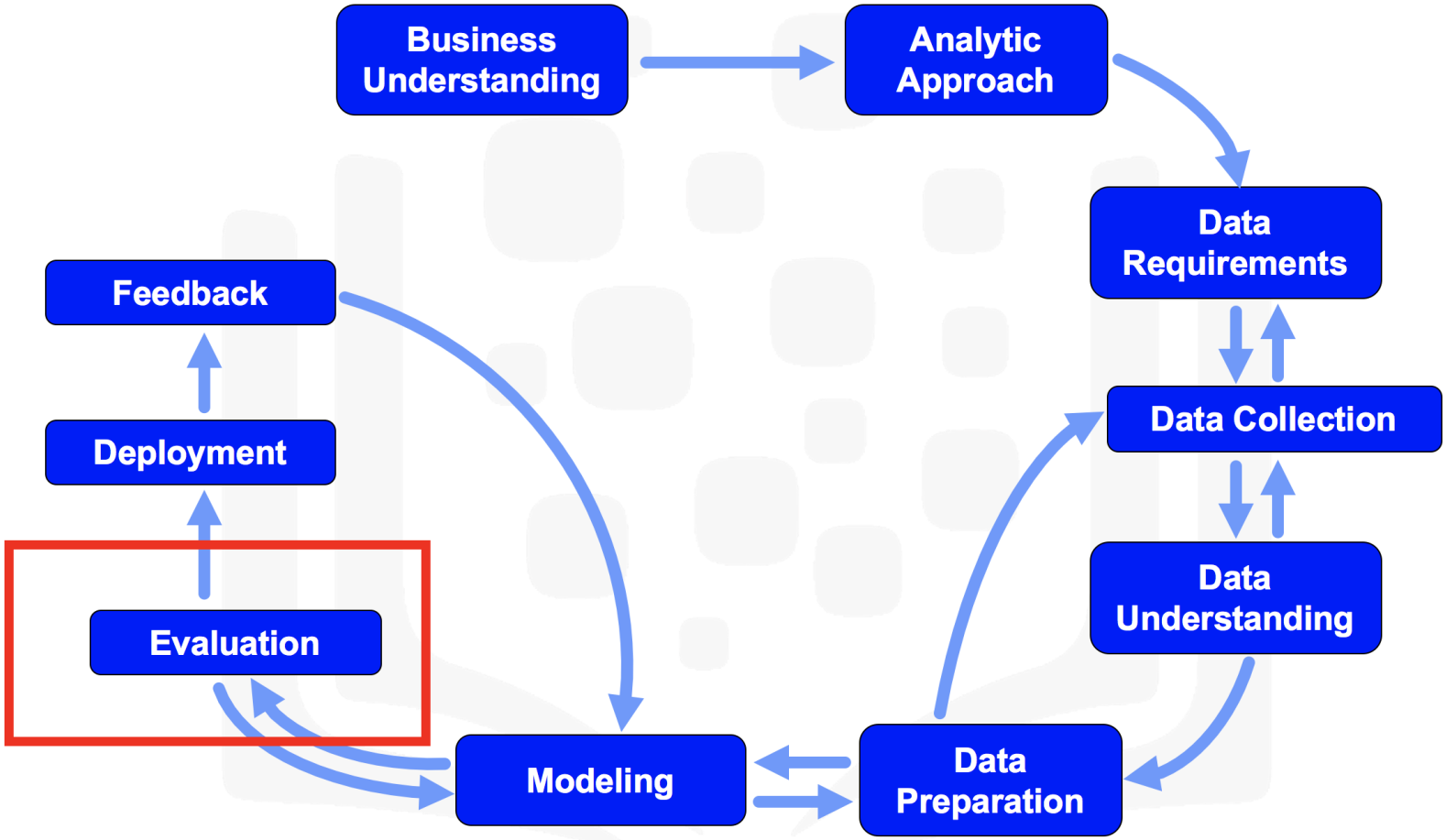
According to the above decision tree:

- If a recipe contains *cumin* and *fish*, then it is most likely a **Thai** recipe.
- If a recipe contains *cumin* but no *fish*, then it is most likely an **Indian** recipe.
- If a recipe does not contain *cumin* and contains *roasted_sesame_seed*, then it is most likely a **Korean** recipe.

You can analyze the remaining branches of the tree to come up with similar rules for determining the cuisine of different recipes.

Feel free to select another subset of cuisines and build a decision tree of their recipes. You can select some European cuisines and build a decision tree to explore the ingredients that differentiate them.

Model Evaluation


Business Understanding
Analytic Approach
Data Requirements
Data Collection
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Feedback

To evaluate our model of Asian and Indian cuisines, we will split our dataset into a training set and a test set. We will build the decision tree using the training set. Then, we will test the model on the test set and compare the cuisines that the model predicts to the actual cuisines.

Let's first create a new dataframe using only the data pertaining to the Asian and Indian cuisines, and let's call the new dataframe **bamboo**.

In [7]:

Let's see how many recipes exist for each cuisine.

In [8]:

```
 chinese   indian japanese   korean     tha
i
     442      598      320      799       28
9
```

Let's remove 30 recipes from each cuisine to use as the test set, and let's name this set **bamboo_test**.

In [9]:

Create a dataframe containing 30 recipes from each cuisine, selected randomly.

In [10]:

Check that there are 30 recipes for each cuisine.

In [11]:

```
 chinese   indian japanese   korean     tha
i
      30       30       30       30       3
0
```

Next, let's create the training set by removing the test set from the **bamboo** dataset, and let's call this set **bamboo_train**.

In [12]:

Check that there are 30 *fewer* recipes now for each cuisine.

In [13]:

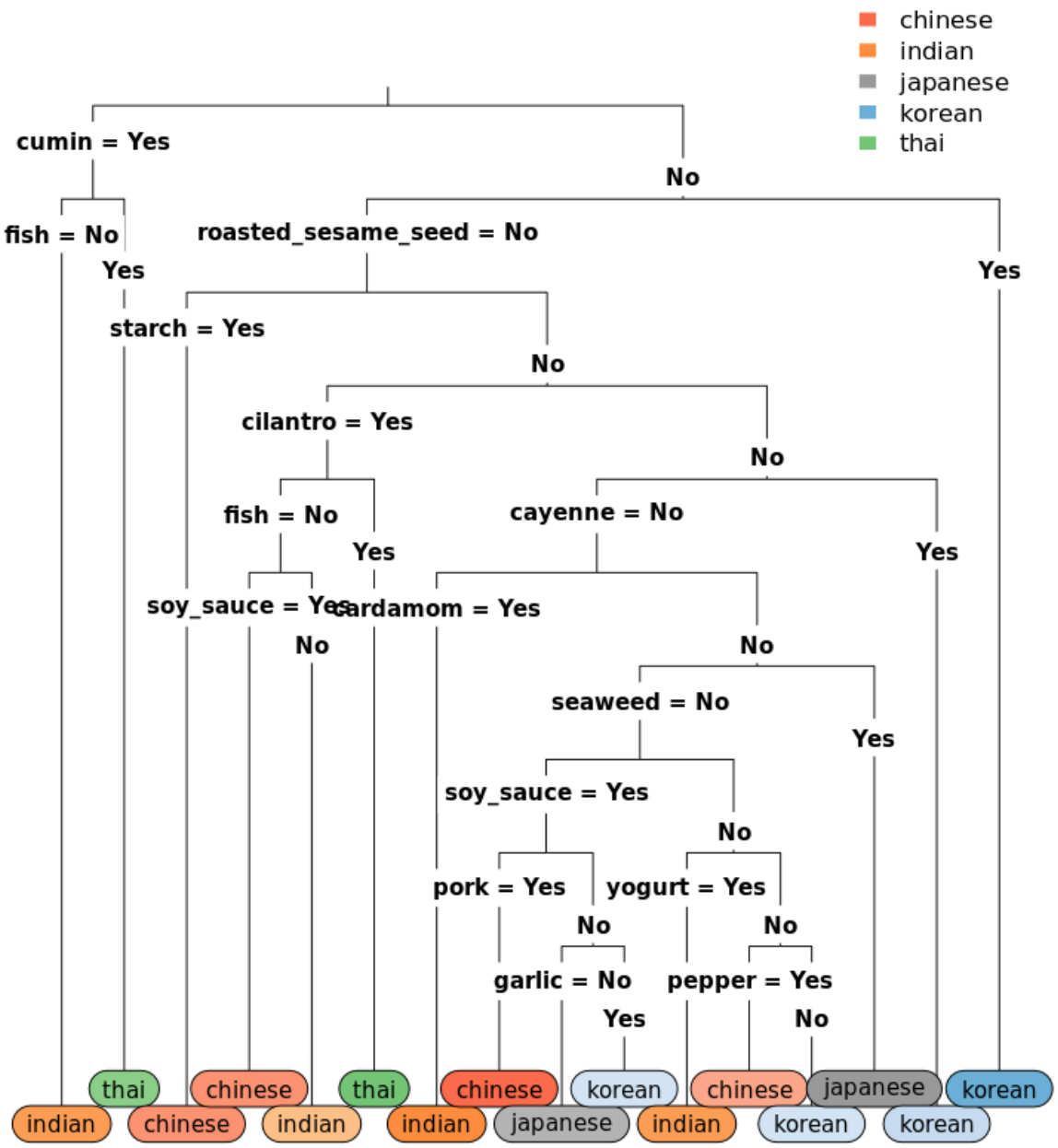
```
 chinese   indian japanese   korean     tha
i
     412      568      290      769       25
9
```

Let's build the decision tree using the training set, **bamboo_train**, and name the generated tree **bamboo_tree_pred** for prediction.

In [14]:

Let's plot the decision tree and explore it.

In [15]:


chinese
indian
japanese
korean
thai
cumin = Yes
No
fish = No
roasted_sesame_seed = No
Yes
Yes
starch = Yes
No
cilantro = Yes
No
fish = No
cayenne = No
Yes
Yes
soy_sauce = Yes
cardamom = Yes
No
No
seaweed = No
Yes
soy_sauce = Yes
No
pork = Yes
yogurt = Yes
No
No
garlic = No
pepper = Yes
Yes
No
thai
chinese
thai
chinese
korean
chinese
japanese
korean
indian
chinese
indian
indian
japanese
indian
korean
korean

It is obvious how removing 30 recipes from each cuisine resulted in more decisions in the tree.

Now let's test our model on the test data.

In [16]:

To quantify how well the decision tree is able to determine the cuisine of each recipe correctly, we will create a confusion matrix which presents a nice summary on how many recipes from each cuisine are correctly classified. It also sheds some light on what cuisines are being confused with what other cuisines.

So let's go ahead and create the confusion matrix for how well the decision tree is able to correctly classify the recipes in **bamboo_test**.

In [17]:

```
                chinese_pred indian_pred ja
panese_pred korean_pred thai_pred
  chinese_true          60.0         0.0
3.3         36.7        0.0
  indian_true            0.0        90.0
0.0         10.0        0.0
  japanese_true         20.0         3.3
33.3         40.0        3.3
  korean_true            6.7         0.0
16.7         76.7        0.0
  thai_true              3.3        20.0
0.0         33.3       43.3
```

The rows represent the actual cuisines from the dataset and the columns represent the predicted ones. Each row should sum to 100%. We make the following observations:

- Using the first row in the confusion matrix, 60% of the **Chinese** recipes in **bamboo_test** were correctly classified by our decision tree whereas 36.7% of the **Chinese** recipes were misclassified as **Korean** and 3.3% were misclassified as **Japenese**.
- Using the Indian row, 90% of the **Indian** recipes in **bamboo_test** were correctly classified by our decision tree and 10% of the **Indian** recipes were misclassified as **Korean**.

How many **Japanese** recipes were correctly classified by our decision tree?

Your Answer: 33.3%

Double-click **here** for the solution.

How many **Korean** recipes were misclassified as **Japanese**?

Your Answer: 16.7%

Double-click **here** for the solution.

What cuisine has the least number of recipes correctly classified by the decision tree?

Your Answer: Japanese 33.3%

Double-click **here** for the solution.

Thank you for completing this lab!

This notebook was created by Polong Lin (https://ca.linkedin.com/in/polonglin) and Alex Aklson (https://www.linkedin.com/in/aklson/). We hope you found this lab session interesting. Feel free to contact us if you have any questions!

This notebook is part of the free course on **Cognitive Class** called *Data Science Methodology*. If you accessed this notebook outside the course, you can take this free self-paced course, online by clicking here (http://cocl.us/DS0103EN_LAB4_R).