Fadel Alshammasi

Yahan Chang

MATH 495 Final Project Report

Introduction - Motivation

Not many people are familiar with Indian cuisine and even less know how to discern if an Indian dish is a snack or a main course. Indian food is often eaten with hands instead of utensils so it's even harder to recognize the difference. This project helps the Iowa State community become more culturally aware of food from international waters. They will also understand the various dishes that come from all regions of India and be able to master culinary ideas through data analysis and evaluation.

Problem Definition

Our first goal requires us to study the data in order to perform data analysis. Our second objective is to inform people on what kind of course the dish is based on the dish details. In order to complete this objective, we have to study the effect of variables such as diet and flavor profile on the course of meal.

Survey on what has been done and what is different in your project

Currently, there are some machine learning models used on Indian foods. Manwani used machine learning to create an image classifier for North and South Indian food. Nilesh et al. used machine learning to recommend Indian dishes based on ingredient availability. But no projects exist to classify Indian dish courses as a main course or snack. Thus, we will use data from Indian Food 101 to fill this void.

Proposed method

To start off, we will also drop rows that have missing values or values that do not make any sense. This will prevent the model from running into errors and won't affect the overall accuracy of our model. First we will build on what we learned in class about plotting to graph and describe the dataset. We will look at the dataset's correlation heatmap, mean, max, min, compare column values, and various data visuals to understand the data better. This will help visualize the relationships to the audience.

Then, we propose using a logistic regression algorithm in order to classify if a dish is a main course or not. And the same applies to snacks. This infamous algorithm is widely used in machine learning applications today. It is also a relatively easy model to implement and train. These regions will allow novice machine learning students, data scientists, and chefs to be able to use this methodology with ease. Additionally, we will use the one hot encoding technique to encode our categorical data into discrete numerical values. This will allow the computer to utilize qualitative information and thereby increase the accuracy of the prediction because more information is factored into the process. In order to measure our model's performance, we will calculate the accuracy score to see how good the model matches the actual relationship in the dataset. We will also take a look into the confusion matrix so we can see how many false positives and false negatives the model will produce.

Experiments/ Evaluation

1. Analysis Experiment: Note these are just snapshots of what we did here. Please check the code for more figures & analysis.

The analysis shows that Indian foods mostly consist of sweet and savory dishes. Indian foods are not usually made up of bitter or sour dishes as seen in Figure 1. Indians prefer eating spicy food as opposed to other types of flavor that are not that popular in India such as bitter and sour. It looks like sweets are also popular in India. By region, the East Region is the only region that has more dishes that are not spicy as seen in Figure 2.

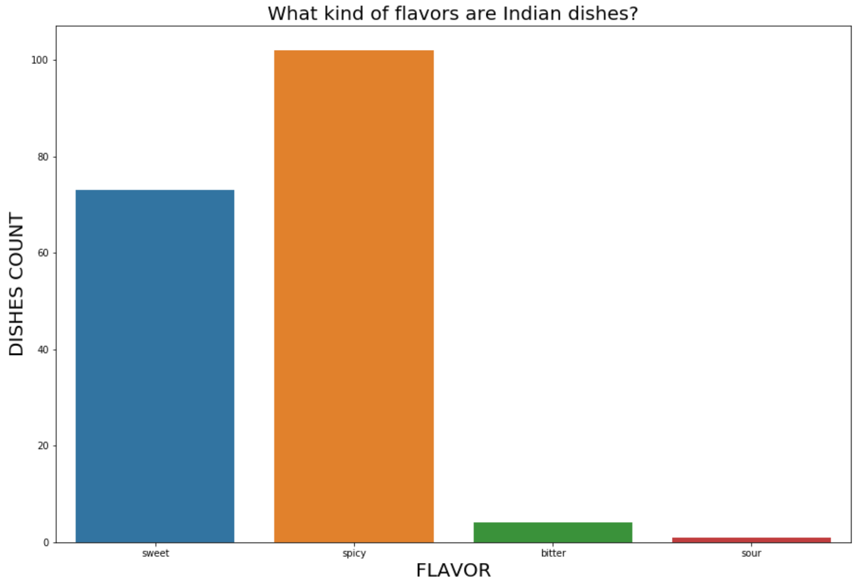

Figure 1.

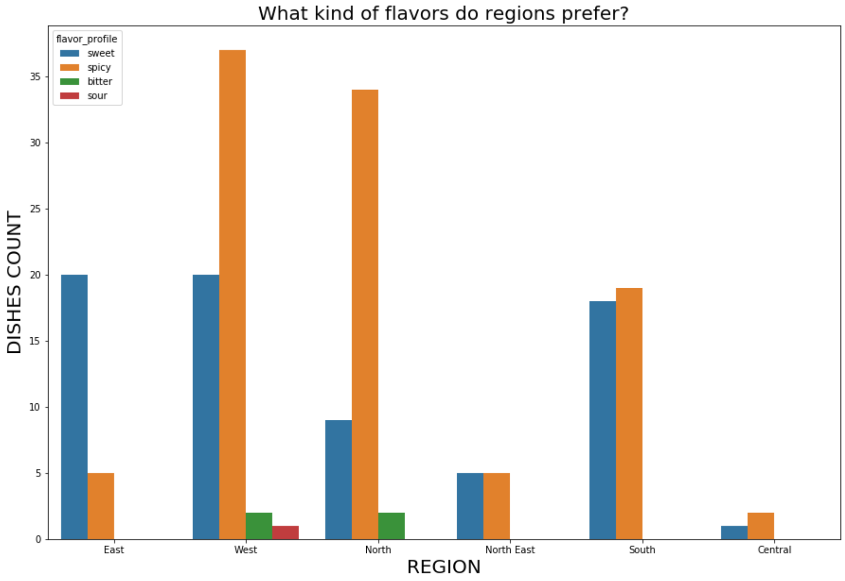

Figure 2.

From Figure 3, the states with the most dishes were Punjab, Gujarat, and Maharashtra. These states are all states in the West of India. Panjab is the home of most Indian dishes with 30 dishes (beating Gujarat by only 1)! There are an overwhelming number of Indian dishes in the West, North, and South regions of India compared to Central or East India. The West region is the region where most of Indian dishes come from.

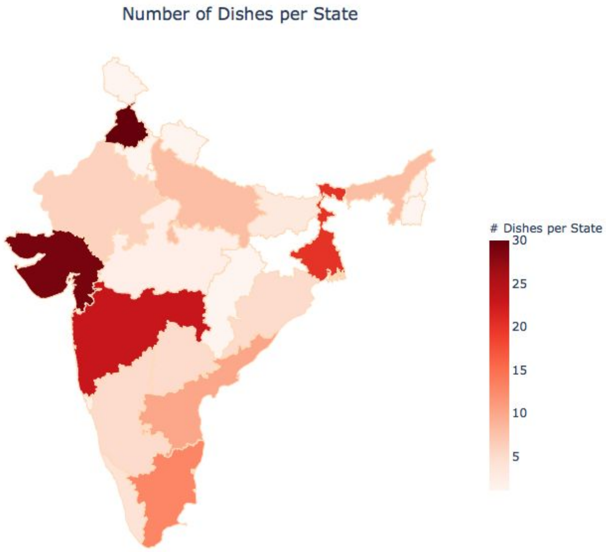

Figure 3.

There are 326 unique ingredients. This means that Indian dishes are made in a diverse amount of ingredients which isn't surprising since Indian spices were sought out by many in the 15th and 16th century such as Christopher Columbus. The most popular ingredients are sugar, garam masala(a spice blend), ghee, and ginger.

Figure 4. shows the total cooking time and ingredients count plotted in a scatter plot. There's a weak negative correlation between the number of ingredients and the total time! This is quite shocking because we expected to see a strong positive correlation.

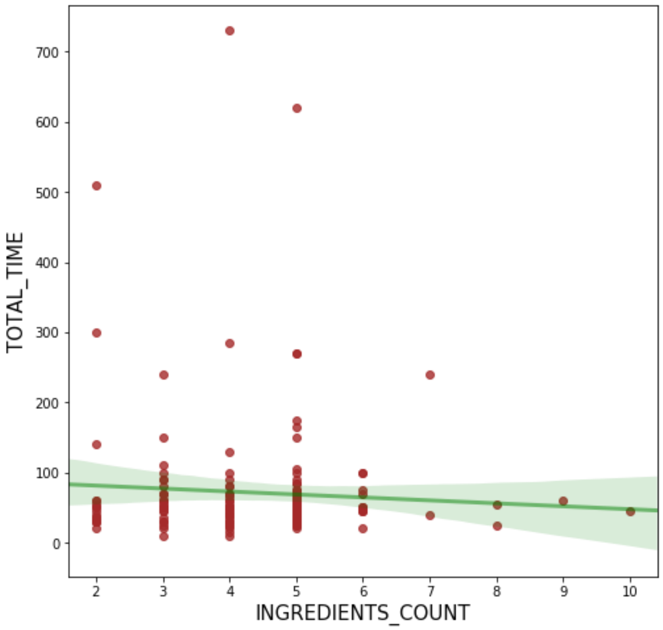

Figure 4.

The heat map of the column values show a strong correlation for prep time, cook time, and total time which is to be expected. At this stage we encoded the state, region, course, and other categorical variables in order to complete our second objective of prediction.

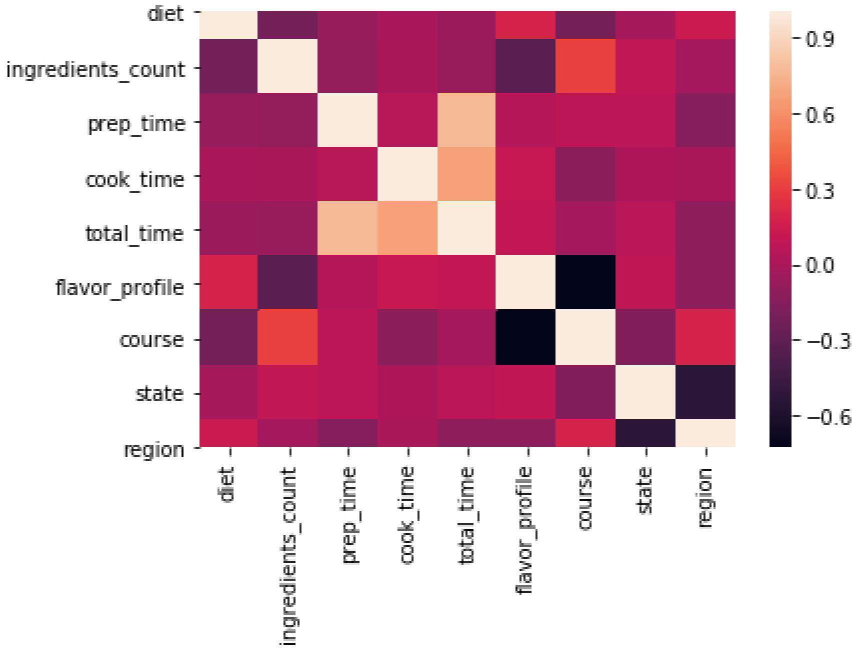

Figure 5.

2. Classification Experiment: two classification models:
    a. Whether or not a dish is a main course:
        - Algorithm: logistic regression
        - Independent variables: prep_time, cook_time,ingredients_count, diet, the encoding of sweet, and the encoding of spicy.
        - Dependent variable: main_course (encoded using one hot)
        - Accuracy score: ≈ 80.55%

```
Accuracy: 0.8055555555555556
```

    b. Whether or a dish is a snack:
        - Algorithm: logistic regression
        - Independent variables: prep_time, cook_time,ingredients_count
        - Dependent variable: snack (encoded using one hot)
        - Accuracy score: ≈ 86.11%

```
Accuracy: 0.8611111111111112
```

Conclusion

In conclusion, we have conducted 2 experiments when it comes to the Indian food dataset. We first analyzed the data carefully and made a lot of observations about the data. Some results were expected such as the distribution of vegetarian vs non-vegetarian dishes. Some other results were surprising such as the negative correlation between total time and the number of ingredients. We learned what state(s) most of Indian dishes come from. Additionally, we utilized machine learning classification models to classify the type of dish (main course & snack).We used one-hot encoding to convert some variables to numerical. We obtained a high accuracy score for both.

As a reflection, working with this data was not easy. We would like to say that we have found that it's easier to work with a long and in depth dataset. The Indian Food 101 dataset was a little sparse in rows and columns. When it came time to clean the dataset, more information was lost. This stresses the importance of finding a dataset that has little null values, is easy to work with, and has a lot of information. Before starting any machine learning problem, be careful with what dataset you're planning to model with.

Distribution of team member effort:

Both worked on the code, report, and presentation.