HDSC November 2021

Capstone Project Presentation

Plotly Project Group

**Introduction**

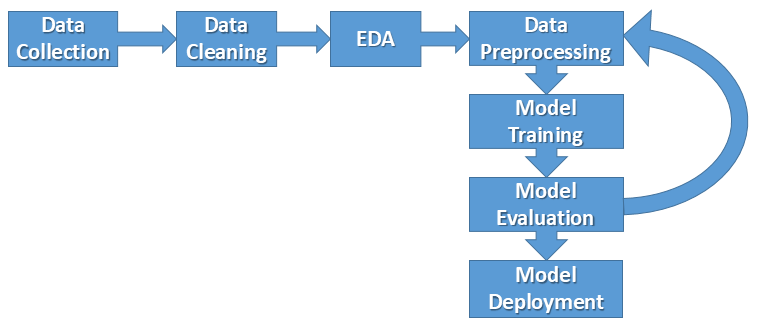
The cultural diversity of culinary practice, as illustrated by the variety of regional cuisines, has evolved over time. We see ingredients mixed to produce dishes that are appealing and edible to humans. These dishes, in most cases, form a cultural identity of people from different countries around the world.

In recent times, various individuals have mislabeled cuisines, due to lack of knowledge or inability to distinguish between dishes as a result of similarity on the blend of ingredients. For instance, while Spain and Mexico might have a common cuisine, other countries might have vastly different cuisines and so one particular dish cannot be associated with them. With this task, a machine learning model would be developed to properly classify these food recipes given the ingredients to their appropriate country or regional cuisine.

**Aims and Objectives**

This project aims at predicting the country of origin of a given cuisine based on a list of ingredients as input variables. The aim of this article is to put you through the project workflow undertook by the plotly team members.

**Project Workflow**



**Data Collection**

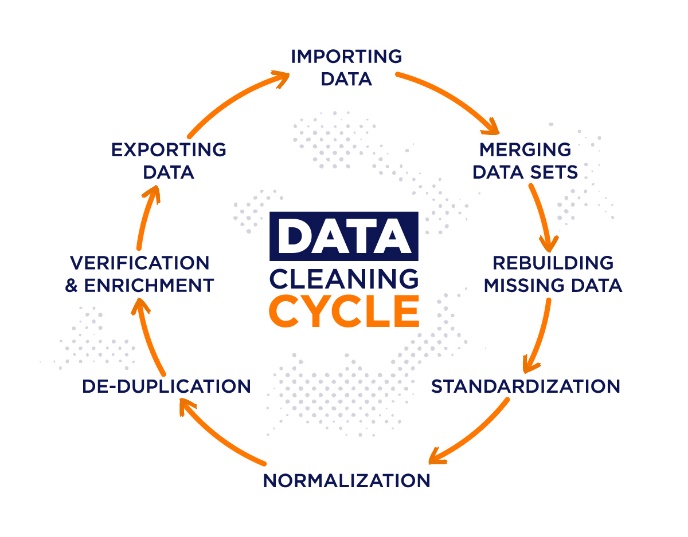
The contents of the dataset for the project was scraped from <https://www.allrecipes.com>, <https://www.epicurious.com>, <https://www.menupan.com>

Link to the dataset: <https://tinyurl.com/2n9aspp8>

The data consists of varieties of recipes for different countries in 57,691 rows and 384 columns. The features are categorical in structure.

**Data Cleaning**

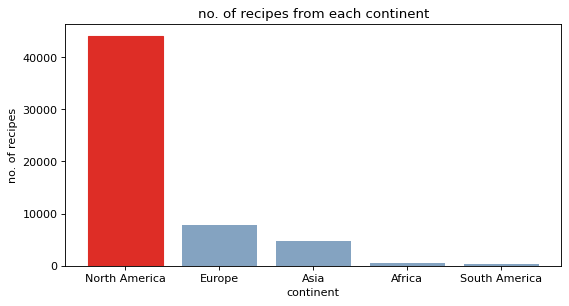
Data cleaning is a vital aspect of a machine learning project. This is because the performance of a trained model depends on the quality of the training data. We spent a major part of our time working on this stage of the project. The concerns we addressed at this stage include null values, outliers, duplicates, redundant columns and wrong data formats. These concerns were fully addressed, and the dataset was ready for exploratory data analysis (EDA). The redundant columns were dropped.



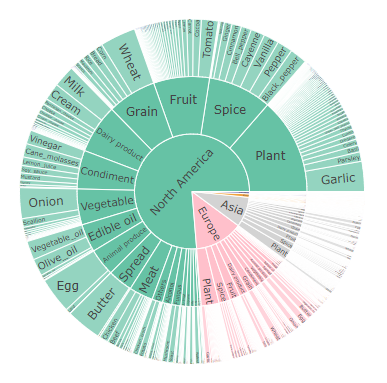
**Exploratory Data Analysis (EDA)**

Exploratory data analysis is performed on a given dataset to gain more insights about the dataset, in terms of its summary statistics and visualizations of relationships that exist among the various variables in the dataset. During this stage of the project, we worked closely with matplotlib and plotly express for data visualization.

As seen below, majority of our data points are North American cuisines.

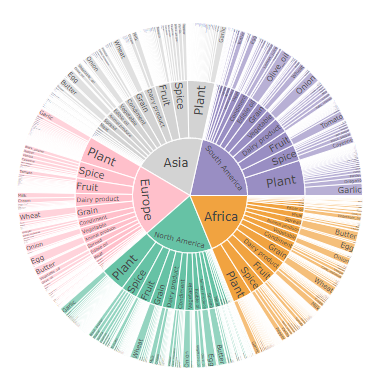


So when checking ingredient counts per continent, North America dominates because it has way more recipes hence higher chances of using all the ingredients multiple times. This is clear in the chart below.

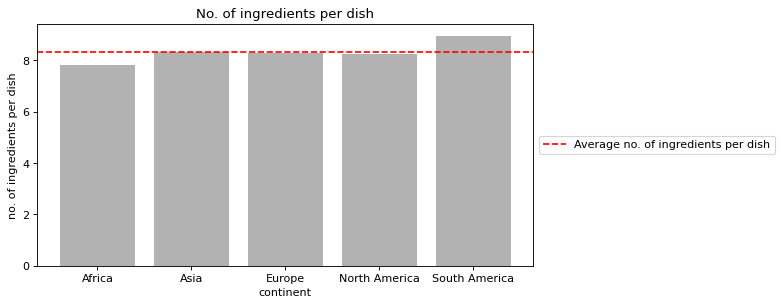


A better measure is to calculate each ingredient’s count per dish for each continent. So we get the prevalence score for every ingredient in each continent.

In the previous chart, garlic is in 13,411 North American dishes and 651 in South American dishes. But in the chart below, garlic is in 0.3 of every North American dish and 0.48 of every South American dish.



Noticeably, there are an average of eight ingredients in every recipe.



**Data Preprocessing**

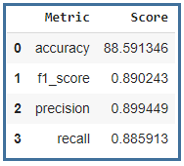
To get the data ready for model training, some techniques were used to perform final preprocessing on the cleaned dataset. During this stage of the project we worked closely with the preprocessing package of python’s sklearn library. We balanced the dataset using RandomOverSampler and also encoded the categorical labels in the dataset. The above was done in order for our model not to be biased.

**Model Training and Evaluation**

The cleaned and processed dataset was first separated into a target variable set and features variable set. The approach we employed for the project was classification machine learning, which is a supervised machine learning method. Which chose this approach because we are working with categorically labeled dataset, and also considering the predefined problem the model is to solve.

To train and test the model, we split the dataset into train set (80%) and test set (20%). Several classification models were trained with the train set and the best performer was selected. The model we selected was ExtraTreesClassifier with accuracy scores of 92.92% and 88.56% on the training and test sets.

The trained model was used to make predictions with the test feature values, and the predictions were evaluated against the test target variables. The scores of the evaluation metrics are given in the figure below.



**Conclusion**

The project’s predefined objectives were met with high levels of accuracy. With the web app associated with this model, you can easily predict the country of origin of a combination of the ingredients in the recipes you wish to try out. Variety is the spice of life, but knowing the origin of what you feed on can make a lot of difference.

**Plotly Team Contributing Members**:

1. John Kennedy Kalu
2. Lois Wakili
3. Chukwuka Kennedy Uche
4. Adedamola Adesoye
5. Rakiya Emmanuel
6. Taiwo Kuponu
7. Humphery Otuniyo
8. Ikechukwu Okoye
9. Hari Prasad

Thanks for reading…