Data Analysis

Nikola Pupovac, nikola.pupovac99@gmail.com

I. Introduction

The purpose of this data analysis is to find correlations between recipes and its origins, so that origin may be predicted as accurately as possible based on recipe ingredients.

II. Database

This database contains a collection of recipes as well as the places where they originated.

There are 10,566 recipes in all, with the majority (22%) coming from South America, followed by Italy (16%) and France (15%). .

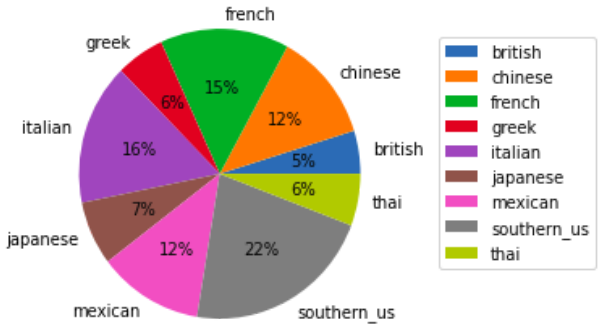


Figure 1 percentage of the data set's recipes corresponding to location of origin.

After removing unnecessary features, the database contains 151 features, 150 of which are binary features (values 0 and 1) that refer to the ingredient and tell us whether the ingredient was used in a specific recipe, and one categorical feature that tells us about the origin of the recipe.

III. Data Analysis

The database has no missing values, so there was no need to fill or drop the data. Database contains the column "Unnamed 0" that represents the unique numeric values irrelevant to the analysis, so the column is dropped.

During the data analysis, it was discovered that there are significant correlations between areas of origin and specific ingredients, as illustrated in **Figure 2**.

Origin to ingredient correlation

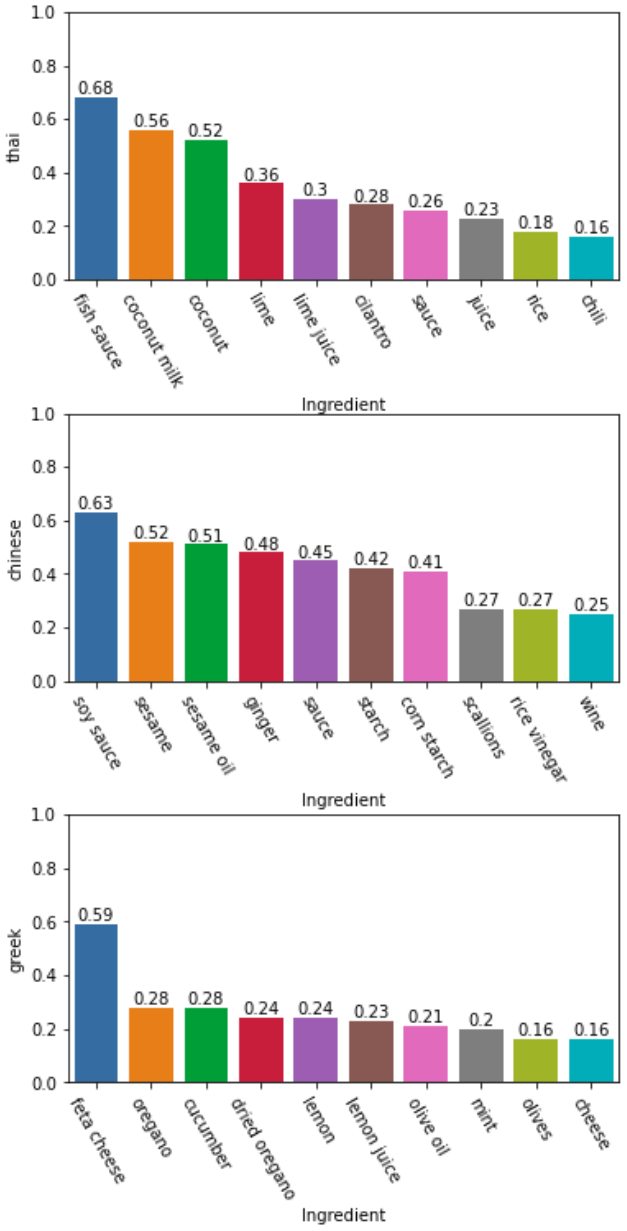


Figure 2. Most significant correlations between ingredients and area of origin

The picture above indicates that if one of the first ingredients shown in the picture above is used for a recipe (for example feta cheese, soy sauce or fish sauce), the recipe is likely to originate from one of the areas listed above.These data can be extremely useful in making predictions.

Table 1: Ingredient correlation with Thailand

Ingredient	Correlation
fish sauce	0.68
coconut milk	0.56
coconut	0.52
lime	0.36
lime juice	0.30
cilantro	0.28
sauce	0.26
juice	0.23
rice	0.18
chili	0.16

IV. 1ST MODEL SELECTION

As requested in the task, the first selected model is KNN - k nearest neighbors.

This algorithm belongs to the category of supervised learning algorithms, which means that it uses already known output values.

The sample is classified using KNN based on the "k" nearest samples.

"K" denotes the sample's number of closest neighbors, which are also defined by predefined distance metrics.

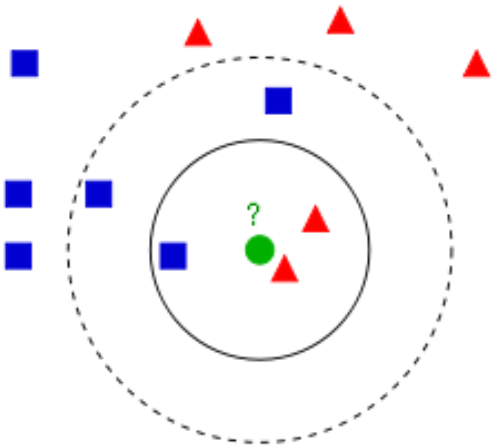


Figure 4. KNN classification

The algorithm was evaluated with a variety of "k" values, "jaccard," and "dice" metrics.

The database is split into two sets: one for training, which contains 90% of the data, and another for testing, which covers 10% of the data. Sets were sampled randomly and evenly.

The training set was divided using the "KFold" method, in which the data were mixed and the database was partitioned into 6 parts, each having an equal percentage of samples from each place of origin of the recipe.

One part is utilized as a validation set, while the rest is used to train the mode.This would repeat another 5 times with each part being a cross validation set once.

The highest accuracy on the test set was shown using the parameter k = 25 as well as the jaccard distance to be approximately 0.69. It can also be observed that all micro accuracy measures indicate approximately the same accuracy.

Table 2: model evaluation

Mesure	Score
Accuracy	0.69
Micro precission	0.69
Micro recall	0.69
Micro f1 score	0.69
Macro precission	0.73
Macro recall	0.63
Macro f1 score	0.73

With the confusion matrix (Figure 5), we can observe that the accuracy in the case of Britain is quite low, while the highest accuracy is in the case of South America, indicating that there is a more obvious pattern in South American recipes.

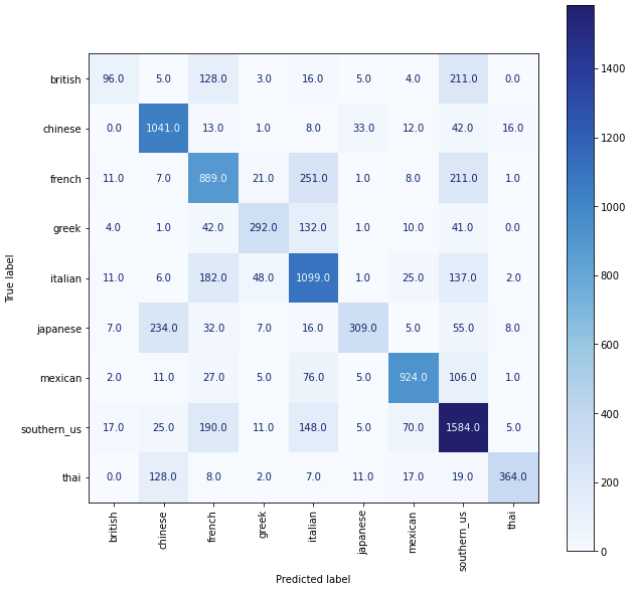
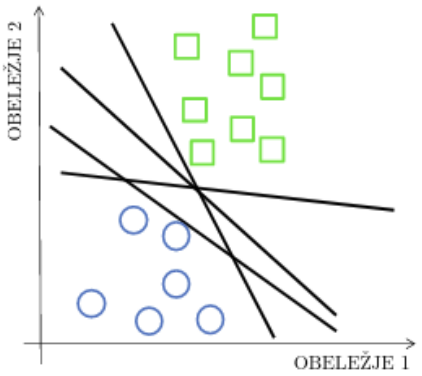


Figure 5. Confusion matrix of KNN algorithm

V. 2ND MODEL SELECTION

As the second model, SVM - Support Vector Machine was chosen.

This algorithm also belongs to the category of supervised learning algorithms. It seeks optimal margins and thus achieves the desired accuracy.



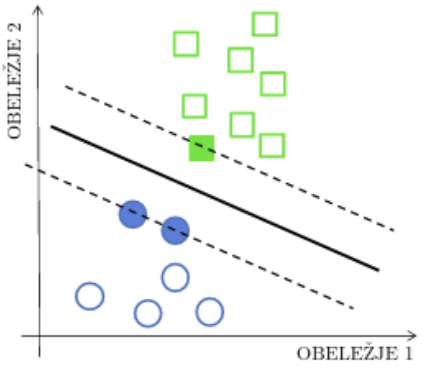


Figure 6. SVM

The algorithm was evaluated with a variety of parameters, with the following being the most effective:

- Regularization parameter: 1
- Kernel: “rbf”
- Decision function: “one vs one”

Table 3: model evaluation

Mesure	Score
Accuracy	0.73
Micro precission	0.73
Micro recall	0.73
Micro f1 score	0.73
Macro precission	0.77
Macro recall	0.68
Macro f1 score	0.70

The algorithm has a higher accuracy than the KNN algorithm, which is 0.73, where all micro accuracy measures also show a value of 0.73.

With the confusion matrix, we can observe high accuracy in South America and China, considering the number of samples that contain them as a place of origin.

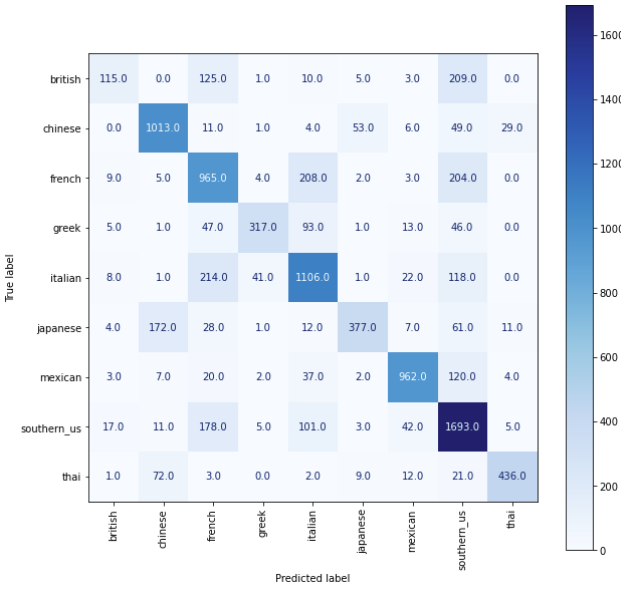


Figure 7. Confusion matrix of SVM algorithm

VI. Model comparison

KNN is a better approach when the number of samples is much greater than the number of features (m >> n), whereas SVM works much better when the number of features is large but the number of samples is small, as seen in this case.

When looking at the accuracy, it can be noticed that the SVM algorithm has a much greater accuracy.

Because the database does not have an unequal data distribution, accuracy may be used to measure the algorithm's performance.