Predicting Cuisines of Recipes

Francesco Ercole
s2135797
s2135797@ed.ac.uk

Dan Lyth
s2008827
s2008827@ed.ac.uk

Mariel Reyes Salazar
s2123659
s2123659@ed.ac.uk

Haorui Tao
s2039924
s2039924@ed.ac.uk

Abstract

In this report we present the results of the Predicting Cuisines of Recipes project. This project was carried over using a dataset from Bellosi [2] that includes 4236 recipes from 12 different cuisines with 709 ingredients. Due to the high-dimension nature of this dataset, several dimensional techniques were surveyed. A simple neural network model, multiple layer perceptron (MLP) is generated

1 Introduction

Meat patty, chips, cheddar cheese, tomato and a bread bun. What is it? It is most likely a cheeseburger. Where is this recipe from? America, most likely. Humans are able to tell which cuisine a dish may come from using "cues" from the ingredients or by visual inspection.

Can a machine be able to replicate this inference? Cuisine is a strong identity factor that makes people in a certain region connect and share traditions. Each culture has their own style of cooking that is distinctive to others by the usage of different ingredients and cooking techniques.

In this work, we use the dataset from Bellosi [2] to predict the cuisines of recipes. This dataset includes 4236 recipes from 12 cuisines with 709 distinct ingredients. The objectives for this project are:

- Predict the cuisine type when given a list of ingredients in a recipe
- Collaborative filtering to predict ingredients in a partial recipe

This work was done entirely using `Python` in a Jupyter notebook [3]. Several Python libraries were used:

- Pandas [7]
- Sci-kit learn [4]
- seaborn [5]
- Tensorflow[1]

2 Data preparation

Prior to exploring data, it was required to prepare the dataset to ensure it is in the appropriate format to work with it.

1. Rename ingredients columns since many ingredients had whitespaces, e.g. soy sauce. Whitespaces can cause trouble when handling these columns, and therefore all whitespaces in ingredient columns were replaced by an underscore

2. Split data into training set (for exploration and model building), validation set (for adjusting model hyperparameters) and finally a test set for evaluating the generated models using a new dataset. The data splitting was done using Sci-kit learn [4]. The samples for each set are detailed in the below tableL

Table 1: Dataset splitting

Set	Samples
Train	2710
Validation	678
Test	848

3 Exploratory data analysis

The data exploration was done in the training set, and the following discoveries were unveiled:

1. There are several ingredients that appear only in one recipe or in two. These ingredients may add noise to the data, as they do not seem to be impactful in describing a recipe per se.

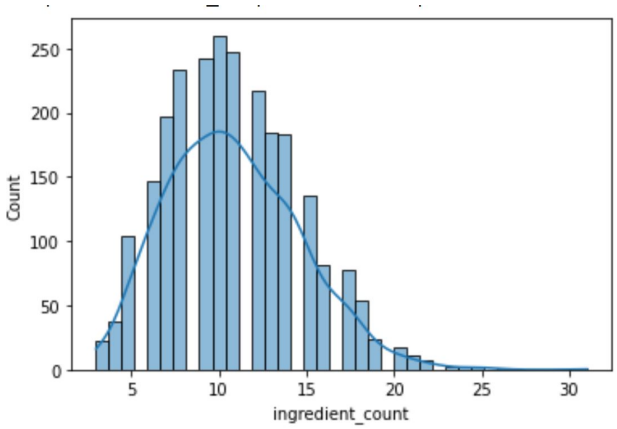

Figure 1: Ingredients per recipe histogram

2. Cuisine types are balanced across the dataset. The median number of ingredients are relatively similar in all cuisine types. However, there are examples of recipes that require a higher number of ingredients. In particular, for Spanish cuisine, there is one recipe that uses 30 ingredients.

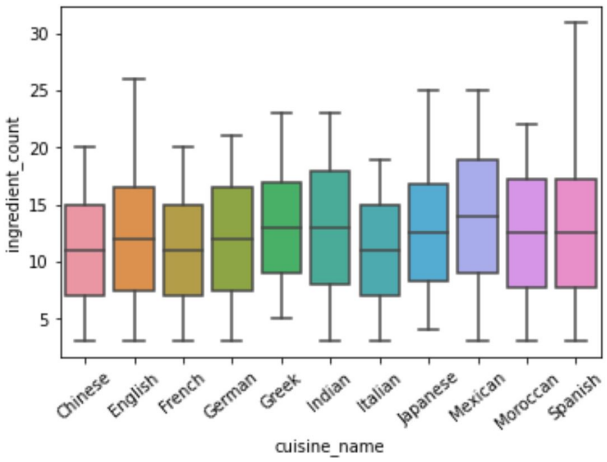

Figure 2: Ingredients per cuisine type boxplot

3. The top 10 most popular ingredients are listed in the following table

Table 2: Fraction of variance explained per principal component

Ingredient	Occurrence in recipes
Garlic	2340
Onion	2157
Olive oil	1390
Salt	1382
Chicken	1324
Pepper	1147
Tomato	953
Water	894
Ginger	855
Butter	78

It can be seen that salt is considered a popular ingredient; however, this ingredient may not be impactful for predicting a recipe, as salt can be added up to taste for most of the recipes, except for sweet dishes or desserts.

4 Dimensionality reduction

Considering this is a high-dimension dataset, it is crucial to reduce the dimensions for a better understanding in 2d or 3d plots, identify patterns or clusters and use this information when building models.

Prior to running any dimensionality reduction technique, the least popular ingredients were removed. The least popularity was defined as an ingredient that appears in only 2 recipes, at most. This reduces the dataset to 538 ingredients.

4.1 PCA

In order to run PCA, the number of components was chosen to be 3. The explained variation per principal component is displayed in the below table:

Table 3: Fraction of variance explained per principal component

Direction	1	2	3
Fraction	0.054	0.043	0.037

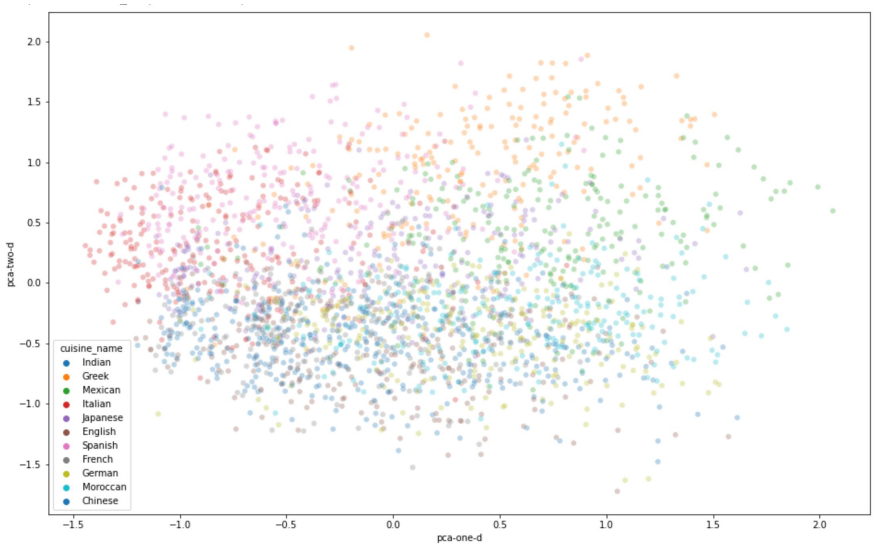

Figure 3: PCA scores visualization

4.2 t-SNE

t-SNE (t-distributed Stochastic Neighbour Embedding) is a cutting edge technique that is explored by experimenting with the perplexity hyperparameter, as per Wattenberg et al. [6]. An insightful representation was achieved with perplexity = 10.

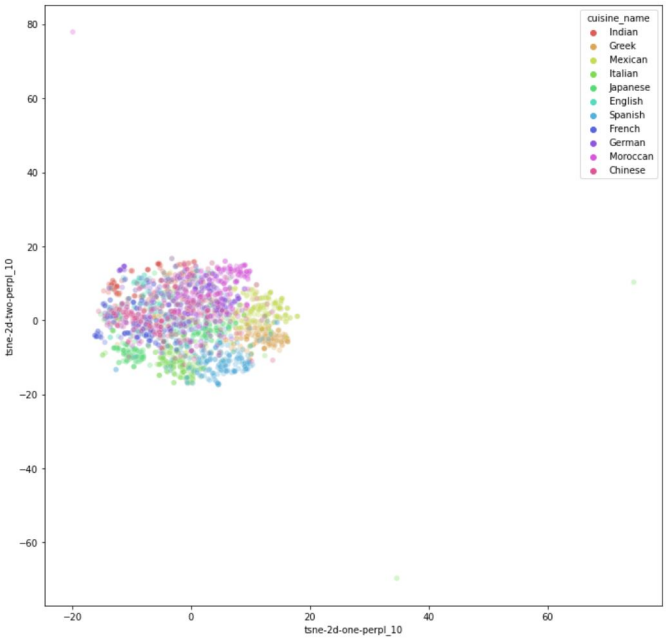

Figure 4: t-SNE using perplexity = 10

4.3 Isomap

Isomap was also surveyed by experimenting with the number of neighbours. The below figure shows the results.

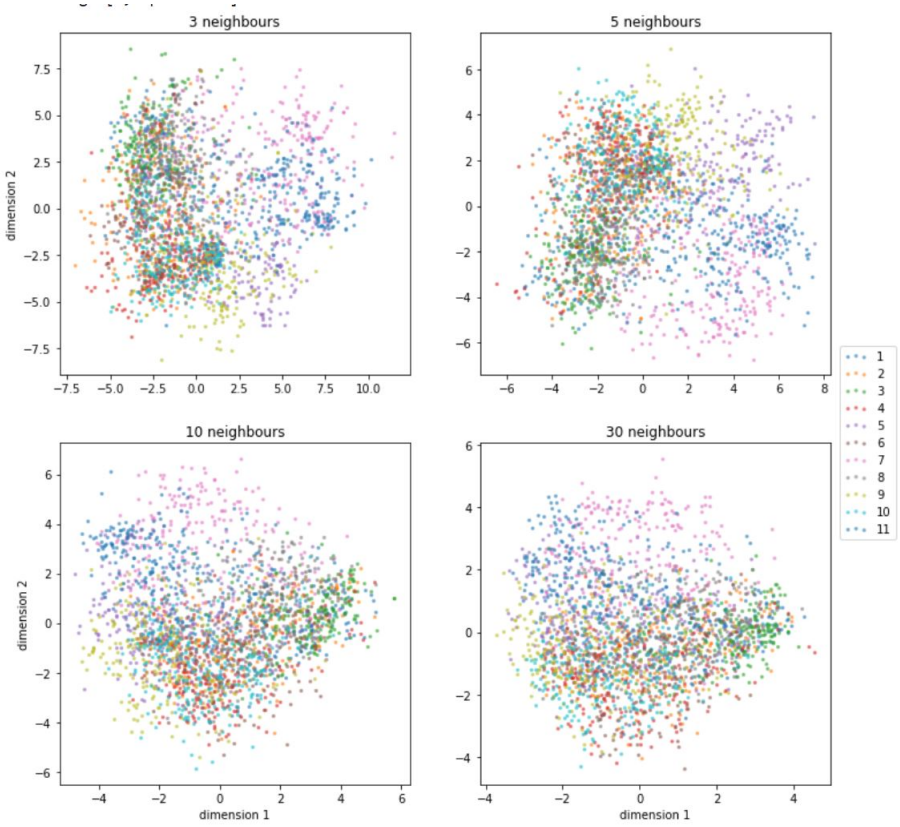

Figure 5: Isomap results

Overall, all techniques utilized are perhaps not as insighful as expected, due to the high-dimension of the dataset. Despite, 171 ingredients were not considered for dimensionality reduction, there are other approaches to be explored, such as removing near-zero variance features. This will be explored in the coming weeks.

5 Learning methods

5.1 Multi-layered perceptron (MLP) neural network

A multi-layered perceptron neural network model was created as a first attempt model, using Tensorflow. We evaluated the model performance by using the validation set and obtaining the confusion matrix

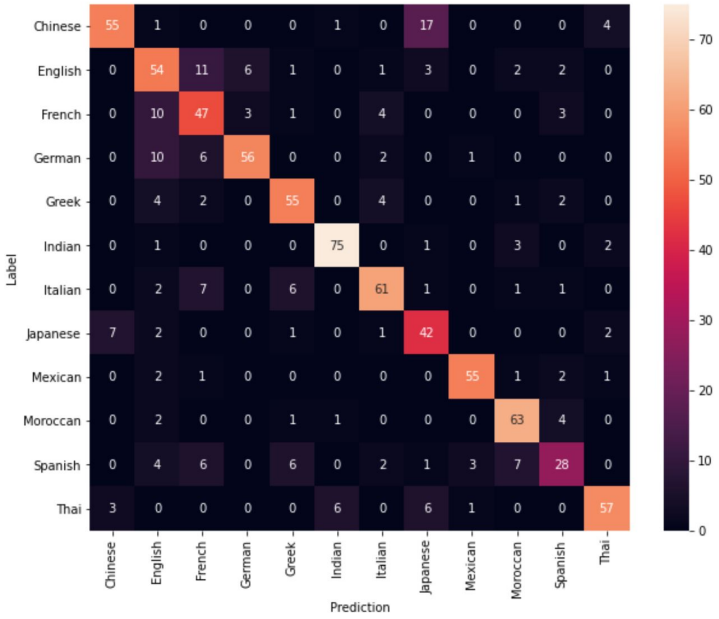

Figure 6: MLP model confusion matrix

This model appears to perform well in predicting the cuisine type; however, it is important to compare with other modelling techniques that may perform similar or better, while being simpler that may be less prone to over-fitting and hence, more robust for generalization.

6 Next steps

There are still items to cover, such as:

- Remove near-zero variance features
- Explore other models such as logistic regression, support vector machines and random forests.
- Collaborative filtering for completing recipes

References

[1] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. and Zheng, X. [2015], 'TensorFlow: Large-scale machine learning on heterogeneous systems'. Software available from tensorflow.org.
URL: *https://www.tensorflow.org/*

[2] Bellosi, F. [2012], Machine learning for cuisine discovery, Master's thesis, University of Edinburgh.

[3] Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S. and Willing, C. [2016], Jupyter notebooks – a publishing format for reproducible computational workflows, *in* F. Loizides and B. Schmidt, eds, 'Positioning and Power in Academic Publishing: Players, Agents and Agendas', IOS Press, pp. 87 – 90.

[4] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. [2011], 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

[5] Waskom, M. and the seaborn development team [2020], 'mwaskom/seaborn'.
URL: *https://doi.org/10.5281/zenodo.592845*

[6] Wattenberg, M., Viégas, F. and Johnson, I. [2016], 'How to use t-sne effectively', *Distill* . (Accessed: 13 March 2021).
URL: *http://distill.pub/2016/misread-tsne*

[7] Wes McKinney [2010], Data Structures for Statistical Computing in Python, *in* Stéfan van der Walt and Jarrod Millman, eds, 'Proceedings of the 9th Python in Science Conference', pp. 56 – 61.