What’s Cooking? – Predictive Modelling and Analysis

Executive Summary

The project is based on the "What's Cooking?" challenge, in which one has to predict the cuisine of a recipe based on its list of ingredients. Just as we recognize a dish's origin by its unique spices like cumin in Indian food or basil in Italian dishes. We trained a model to detect these patterns and associate them with specific cuisines. First, we loaded and preprocessed the zip file to have consistent dataframe. We then used Explanatory Data Analysis to find the top cuisines, top ingredients, and other trends in the data. Data was transformed into numerical features using TF-IDF vectorization and Countvectorizer to feed the models. We tried three models Logistic Regression, Random forest Classifier and Support Vector Classifier, and among them, Support Vector Classifier after hyperparameter tuning gave the best results with a validation accuracy of **81%** and a score of **0.80410** (80.41%) on the Kaggle Leaderboard.

Data Pre processing

The first step in this project was to validate some key data files that were vital for our analysis. Once the datasets were secured, we plunged right into the exploration of data: distribution of cuisines and ingredients. In doing so, we built bar charts for visualization to help uncover major trends and outliers in the data. Further analysis, by looking at the frequency, underlined the most characteristic ingredients for each cuisine. This analysis not only gave insight into typical flavor profiles but also helped to guide feature selection for our predictive models. Finally, we standardized the ingredient data by cleaning and simplifying the text, which included removing special characters and consolidating ingredients into a unified format. This was a very important step because now our model could process and learn from the data with consistency and accuracy.

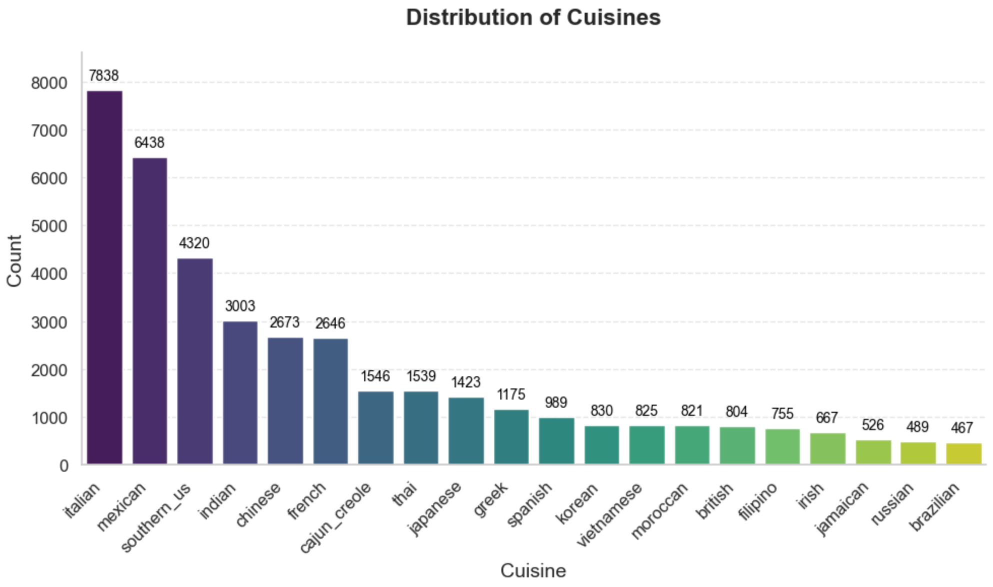

Fig1: Top 20 Cuisines in the dataset

Modelling

We train three models; Logistic Regression, Random Forest Classifier, and Support Vector Classifier. After a careful evaluation based on accuracy and consistency in prediction, we ultimately selected the Support Vector Classifier as our preferred model. This decision was backed by its impressive performance, achieving a validation accuracy of **81%** and a score of **0.7964** (79.64%) on the Kaggle Leaderboard, making it the standout choice among the models we considered. For further refining our model's performance, we conducted hyperparameter tuning using a GridSearch technique. This approach resulted in a validation accuracy of 81%, which was higher than the performance achieved with the initial setup of the Support Vector Classifier.

By tuning our initial model setup, we ensured that our newly tuned predictive model is not only robust but also finely tuned to the specific characteristics of our dataset. This strategic approach was crucial in optimizing the model’s performance, allowing us to accurately predict cuisines with a high degree of reliability. Our systematic yet flexible methodology confirmed that we had selected a model that was exceptionally well-adapted to meet the demands of our project, proving essential in our success in the "What's Cooking?" challenge.

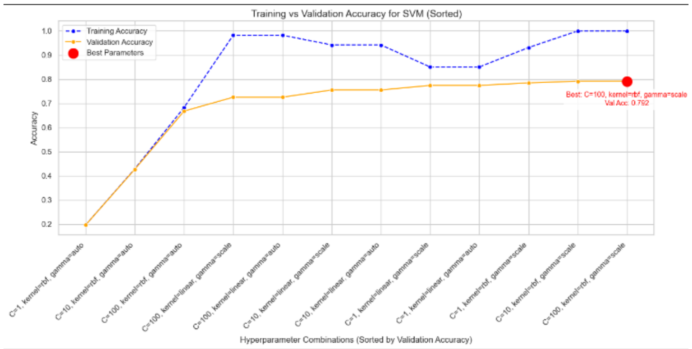

Fig 2: SVM accuracy comparison across hyperparameters.

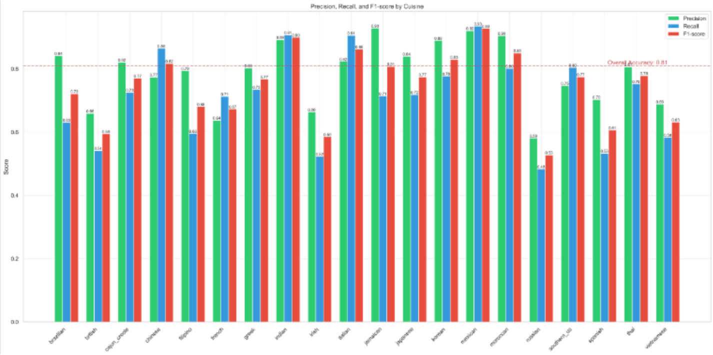

Fig 3: Precision, Recall, and F1-score analysis.

Conclusion

This project aimed to predict a recipe’s cuisine based on its ingredients, inspired by the "What's Cooking?" challenge. We prepared and explored data to find trends in both cuisine distribution and ingredient usages, then compared various models from simple classifiers to advanced ensemble methods. The model which performed best was the tuned Support Vector Classifier which had a validation accuracy of 81% and a score of 0.8041 on Kaggle Leaderboard. It learned to identify patterns in ingredient lists indicative of different cuisines. In general, this project showcases the power of data analysis and model comparison in solving practical problems. By leveraging ingredient data, we developed a robust tool for classifying cuisines, fusing culinary insight with predictive techniques.