Cuisine/Category Prediction using Machine Learning

Group 10

Vaibhav Nauriyal(MT22129)
Kushagra Mittal(MT22108)
Kovida Samir Kshatri(MT22106)

Methodology

Data Processing

- Data Consistency
  - Null Values
  - Un-correlated features
  - Merging the dataset(25% of the dataset used)
- Pivoting the dataset
  - Use only Recipe Id & Ingredient Columns
  - One hot encoding of each ingredient.

Applying TF-IDF

- To give importance or relevance scores to ingredient lists(for each recipe).

Applying Machine Learning Models

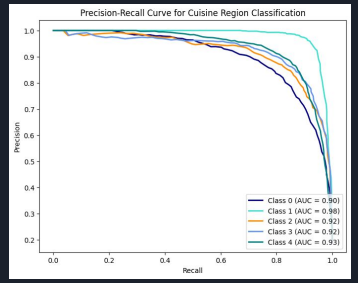
- Random Forest
- Decision Trees
- Gradient Boosting
- Logistic Regression
- Naive Bayes

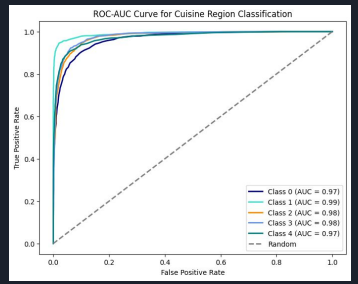
Sampling using SMOTE(for balancing)

Hyperparameter Tuning

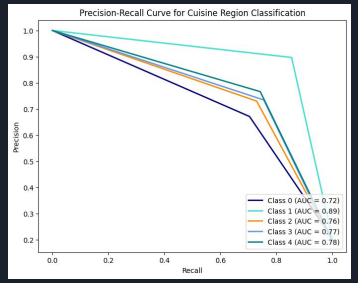
Comparison of ML Models

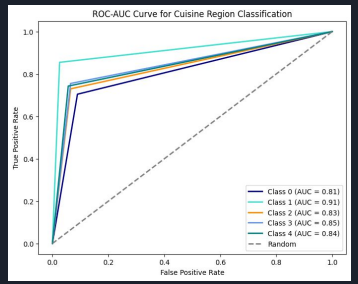
Random Forest



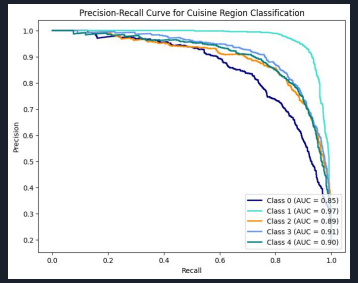


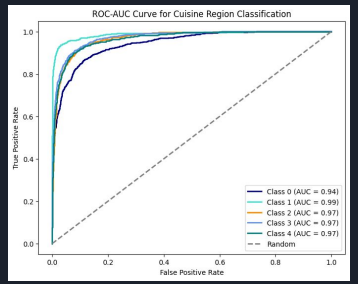
Decision Trees



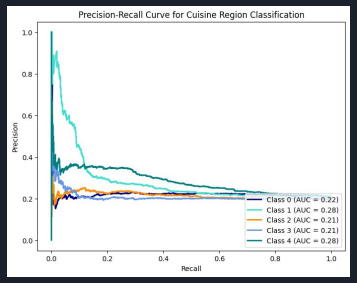


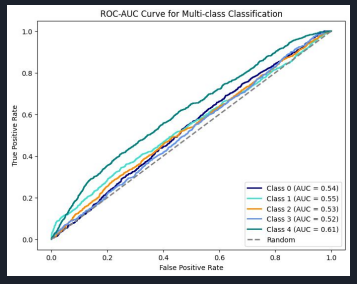
Gradient Boosting



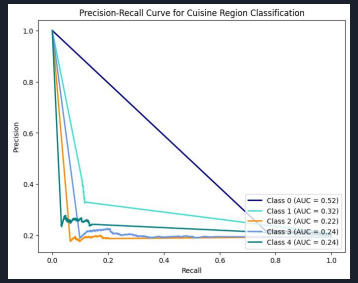


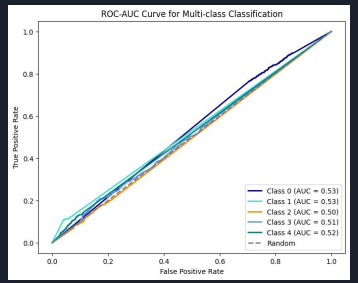
Logistic Regression





Naive Bayes





Results

Models	Random Forest	Decision Trees	Gradient Boosting	Logistic Regression	Naive Bayes
Accuracy	87%	76%	83%	22%	23%
Mean CV Score	85.11%	66.30%	75%	23.02%	21.64%

**Mean Cross Validation is used as a metric.

Conclusion

With hyperparameter tuning of n_estimators of Random Forest, we got a mean cross-validation score of 85.53% which is 0.42% better than base Random Forest model. Ensemble and Boosting algorithm seemed to outperform linear, probabilistic and tree based models. This is again confirmed by Precision and Recall curves.