Prakriti 2021
Data Analytics

Team Name: Chasing Failure

Raunak Raj
Contact: 9097856670
Institute: IIT Kharagpur

Kirti Vardhan
Contact: 9117691398
Institute: IIT Kharagpur

Chitranshu Ranjan
Contact: 8207279999
Institute: IIT Kharagpur

Table of Content

Brief Workflow

Data Preprocessing

Some values were missing in the Preparation Time, Cooking Time, Flavor, State and Region. For taking care of missing data, mean of Preparation Time and Cooking time were used and mode of Flavor, State and Region were used.

Handling Textual Data

Using nltk library cleaned the text data and then tokenized the data. Ultimately the text data was transformed into a matrix.

Encoding Categorical Data

Numerically encoded dataframes were formed for all the categorical features and later merged with the parent dataframe.

Separating target variables

Different features and target variables were chosen to classify the data according to Food Course Type and Flavor.

Training the model

Data was split into 80-20% ratio for training and testing purpose. Further, three classification models were used namely, Random Forest Classifier, Support Vector Machines Classifier and Gradient Boosting Classifier.

Different score calculations

Produced the confusion matrix, overall accuracy, producer's accuracy, user's accuracy, and kappa coefficient of the classification using all three classification models.

Data Preprocessing

Checking null values

Missing data were identified in the dataset in 5 different features.

Using Simple Imputer

From Scikit Learn, simple imputer was imported for handling the missing values.

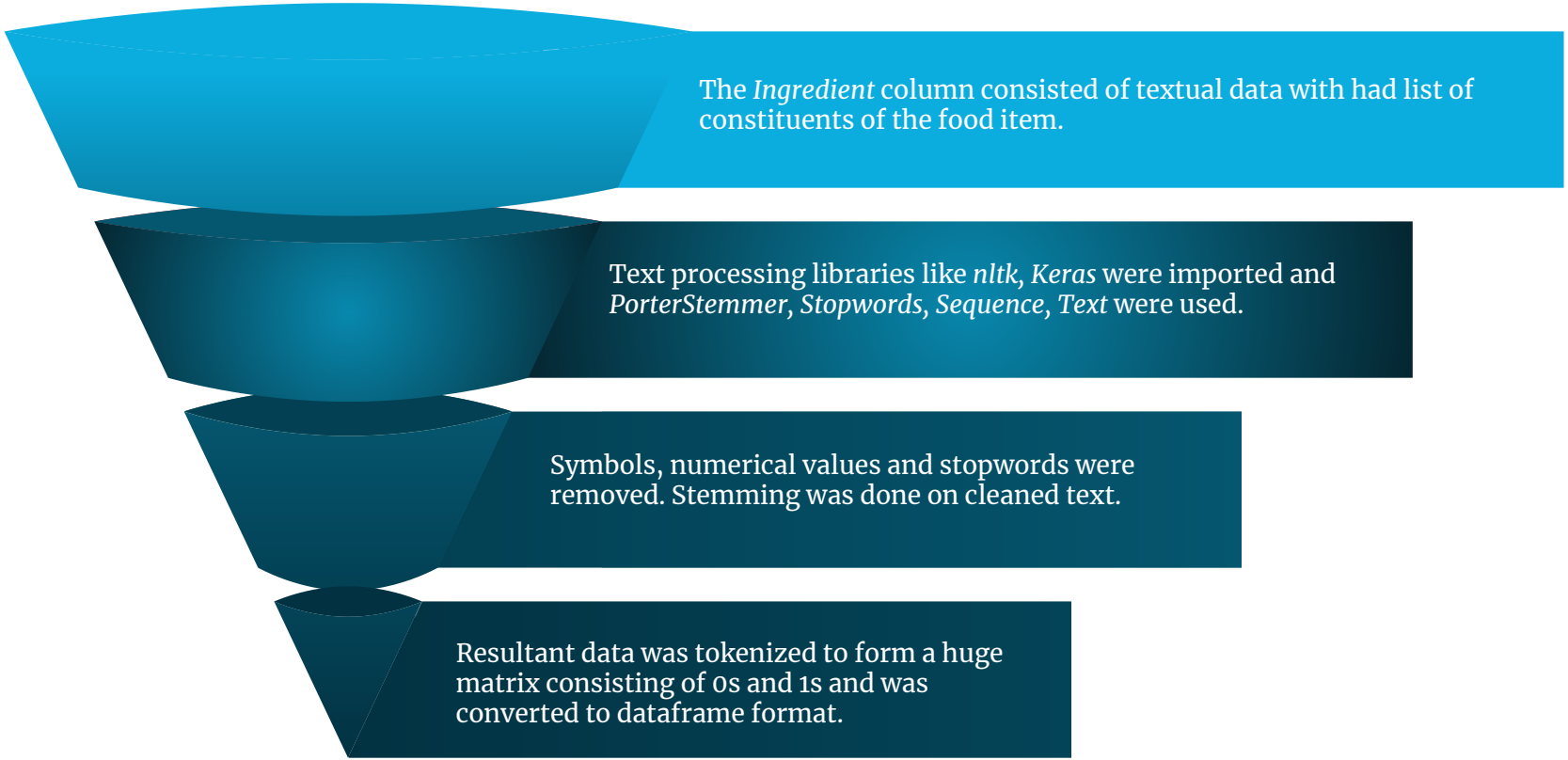
Handling Continuous Data

Mean was used to replace the NaN values in *Preparation Time* and *Cooking Time.*

Handling Categorical Data

Mode was used to replace the NaN values in *Flavor*, *State* and *Region*.

Handling Textual Data



Encoding Categorical Data and Separating Target Variables

Identifying Categorical Data

Diet, Flavor, Course, State and Region

Numerical Encoding of Data

get_dummies function of *Pandas* library was used to form separate encoded dataframes

Preparing two sets of features and target variables

In first set, *Course* was assigned as target variable and rest were used as features whereas in second set *Flavor* was assigned as target variable and rest as features

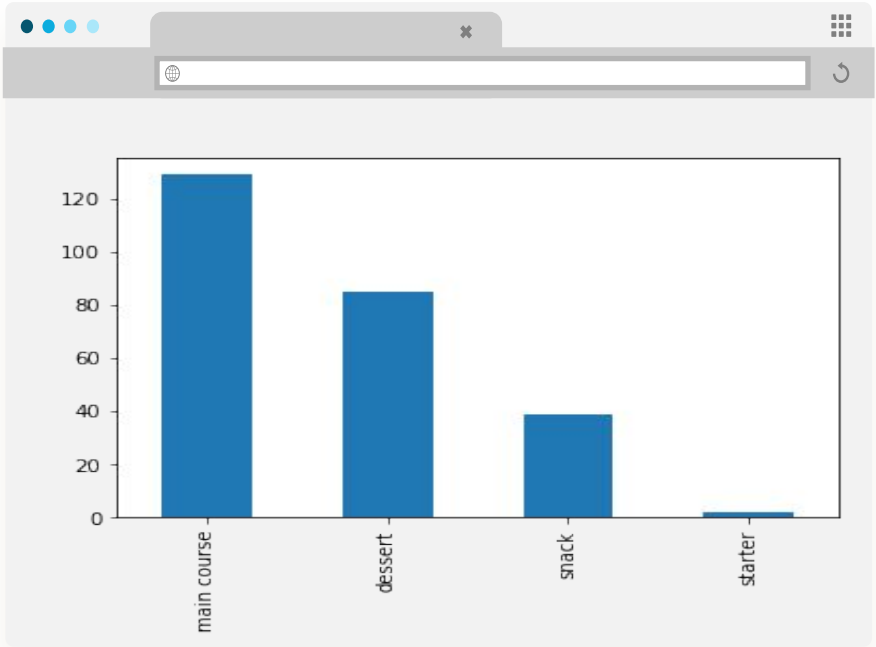
Standardization

StandardScaler was fitted and transformed over the feature Dataframe

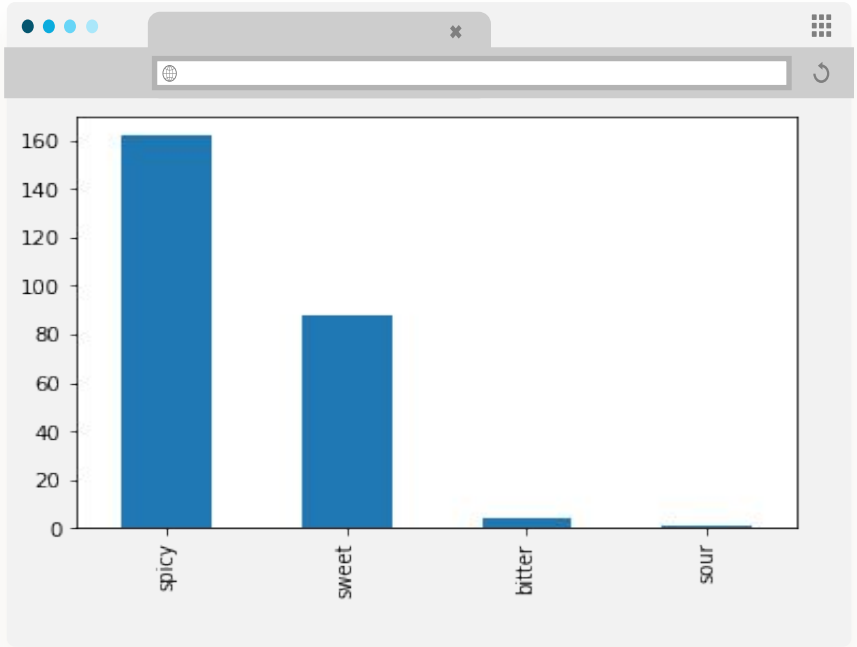
Train-test splitting

X(features) and y(target) was divided into train and test sets with test size as 20%

Value Count Plot for Target Variables



Food Course Type



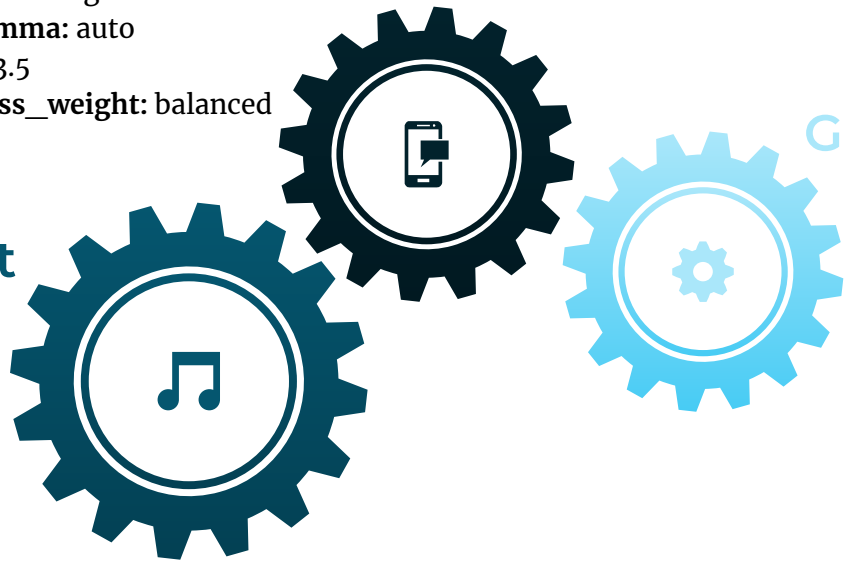
Flavor

Correlation of Numerical Features

Training the model

Support Vector Machine Classifier

- **kernel:** sigmoid
- **gamma:** auto
- **C:** 3.5
- **class_weight:** balanced



Random Forest Classifier

- **n_estimators:** 10
- **random_state:** 42
- **max_depth:** 2
- **criterion:** entropy
- **max_features:** None

Gradient Boosting Classifier

- **n_estimators:** 900
- **learning_rate:** 1.3
- **min_samples_split:** 4
- **validation_fraction:** 0.2

Course: Random Forest Classifier

Accuracy Score

0.86275

Kappa Coefficient

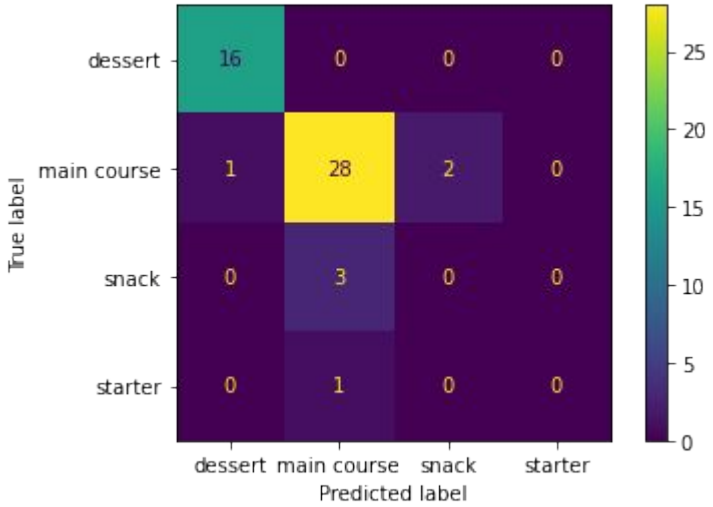
0.73178

Producer's Accuracy

Dessert: 1.0
Main Course: 0.90323
Snack: 0.0
Starter: 0.0

User's Accuracy

Dessert: 0.94118
Main Course: 0.875
Snack: 0.0
Starter: NaN



Course: SVM Classifier

Accuracy Score

0.86275

Kappa Coefficient

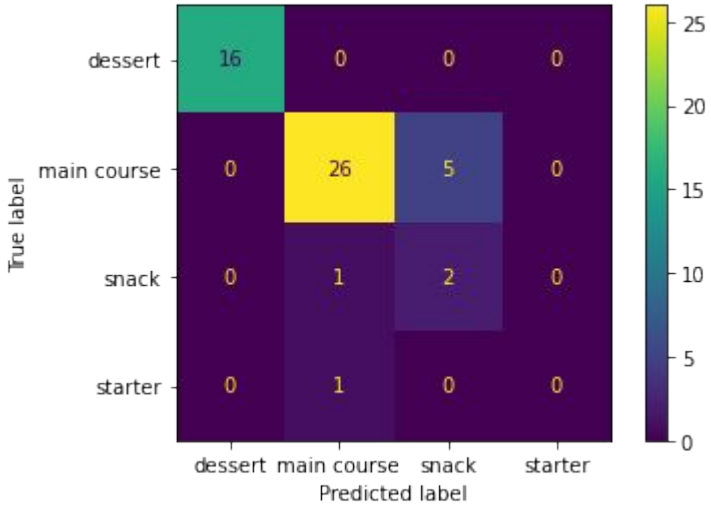
0.75481

Producer's Accuracy

Dessert: 1.0
Main Course: 0.83871
Snack: 0.66667
Starter: 0.0

User's Accuracy

Dessert: 1.0
Main Course: 0.92857
Snack: 0.28571
Starter: NaN



Course: Gradient Boosting Classifier

Accuracy Score

0.82353

Kappa Coefficient

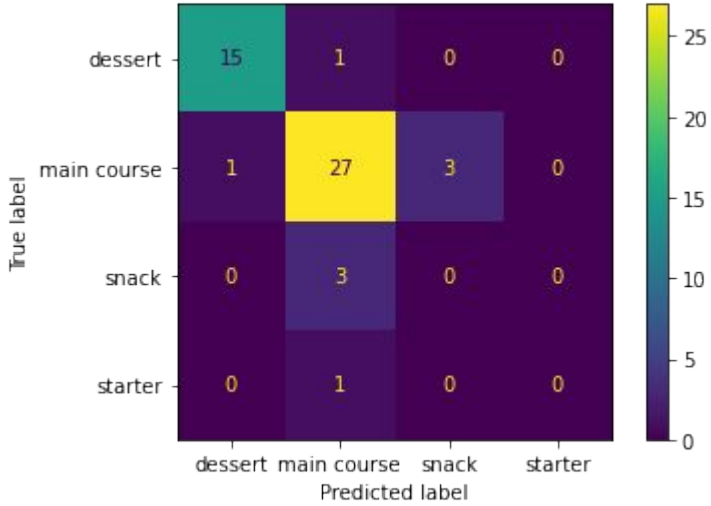
0.65848

Producer's Accuracy

Dessert: 0.9375
Main Course: 0.87097
Snack: 0.0
Starter: 0.0

User's Accuracy

Dessert: 0.9375
Main Course: 0.84375
Snack: 0.0
Starter: NaN



Flavor: Random Forest Classifier

Accuracy Score

0.94118

Kappa Coefficient

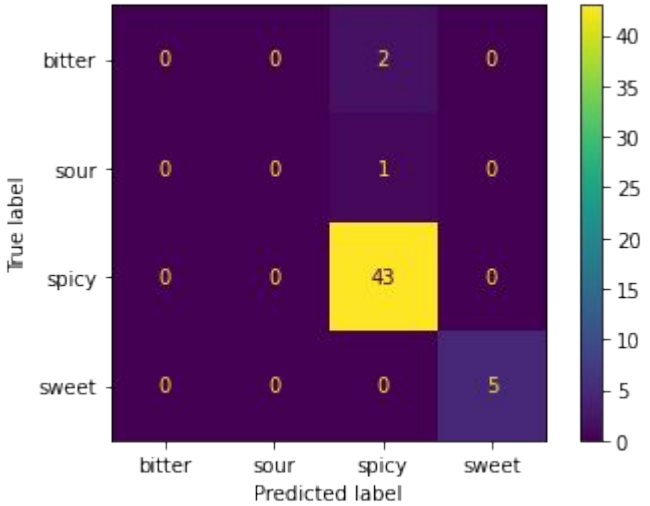
0.74415

Producer's Accuracy

Bitter: 0.0
Sour: 0.0
Spicy: 1.0
Sweet: 1.0

User's Accuracy

Bitter: NaN
Sour: NaN
Spicy: 0.93478
Sweet: 1.0



Flavor: SVM Classifier

Accuracy Score

0.92157

Kappa Coefficient

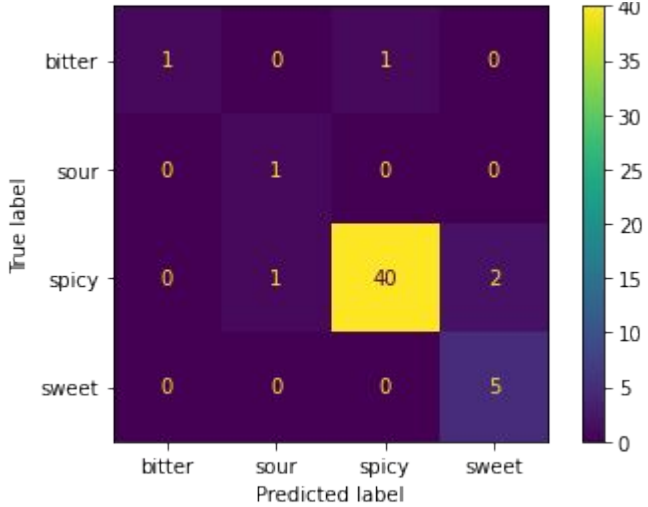
0.74468

Producer's Accuracy

Bitter: 0.5
Sour: 1.0
Spicy: 0.93023
Sweet: 1.0

User's Accuracy

Bitter: 0.9375
Sour: 0.84375
Spicy: 0.0
Sweet: NaN



Flavor: Gradient Boosting Classifier

Accuracy Score

0.92157

Kappa Coefficient

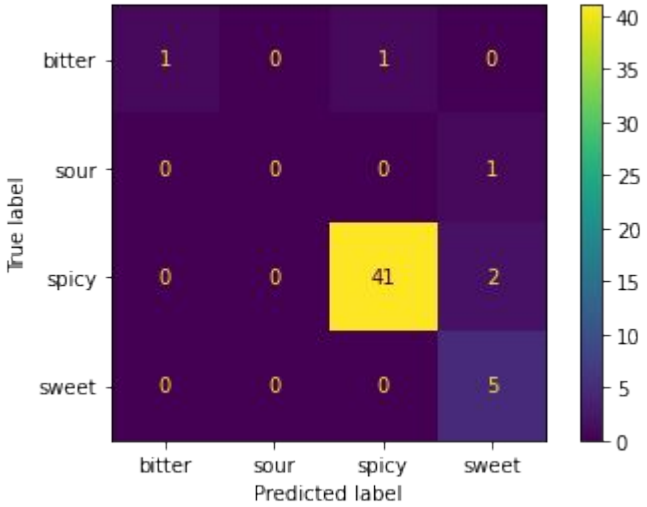
0.72908

Producer's Accuracy

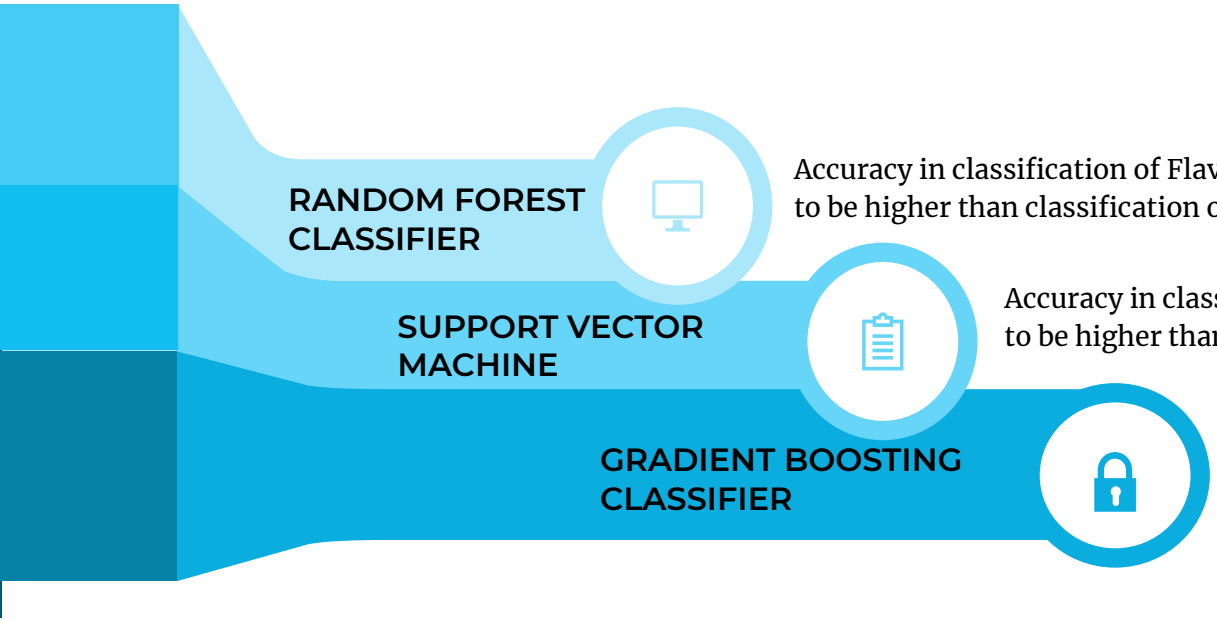
Bitter: 0.5
Sour: 0.0
Spicy: 0.95349
Sweet: 1.0

User's Accuracy

Bitter: 1.0
Sour: NaN
Spicy: 0.97619
Sweet: 0.625



Comparison between two target variables



Accuracy in classification of Flavor(0.94) was found to be higher than classification of Course(0.86)

Accuracy in classification of Flavor(0.94) was found to be higher than classification of Course(0.86)

Accuracy in classification of Flavor(0.92) was found to be higher than classification of Course(0.82)

Thank You