EMBRACING HEALTHIER EATING WITH MACHINE LEARNING TECHNOLOGY

Bryce Anthony, Peter Henry, David Mansoir, Dalia Saj, & Amy Williams

IST 736 – Text Mining – Final Project

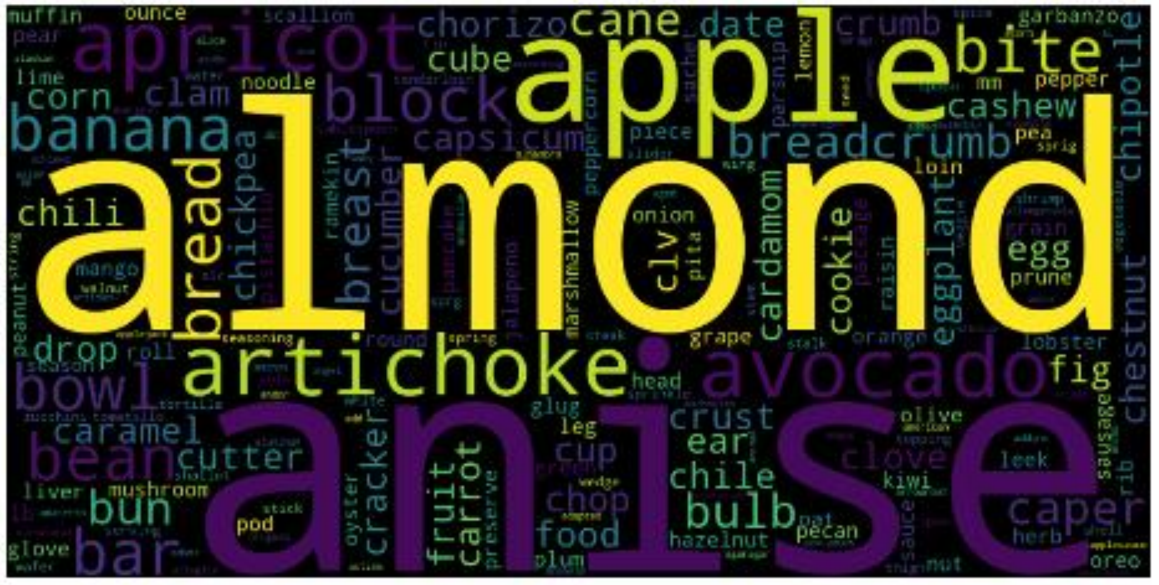
Introduction

- Many people struggle to identify truly healthy recipes
- Importance of healthy eating is ever growing; demand for fresh, whole foods is at an all-time high
- Cuisine varies greatly across cultures, and this diversity presents both opportunities and challenges in making healthy food choices
- Significant opportunity for technology to play a transformative role in our dietary choices
  - Fueled by rising awareness of the links between diet and chronic diseases such as obesity, diabetes and heart disease
  - A healthy diet could prevent 80% of heart disease and 40% of cancer cases globally (WHO)
- Understanding the ingredients essential for maintaining a healthy lifestyle

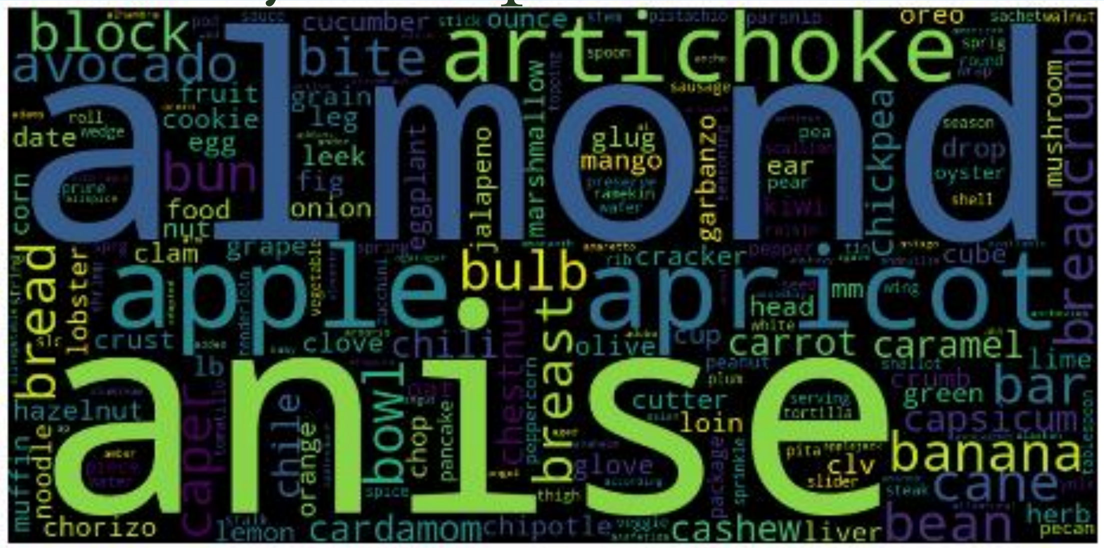
Project Overview

- Prototype for classifying recipes as healthy or unhealthy using machine learning
  - Model considers key nutritional metrics: macronutrients (proteins, fats, carbohydrates), vitamins, and minerals.
- **Application and Benefits**:
  - Valuable for individuals and companies in the health and wellness industry.
  - Potential for integration into products and services, e.g., meal kit delivery services.
  - Helps customers make informed choices and supports health goals.
- **Impact**:
  - Enhances public health by bridging the gap between nutrition science and everyday eating habits.
  - Empowers individuals to make better dietary choices with a user-friendly model.
- **Data Preparation**:
  - Data sourced from two CSV files with healthy and unhealthy recipes; created through scraping the Spoonacular API and website.
  - Includes information on ingredients, calories, fat, protein, and carbs.

All Recipes



Healthy Recipes

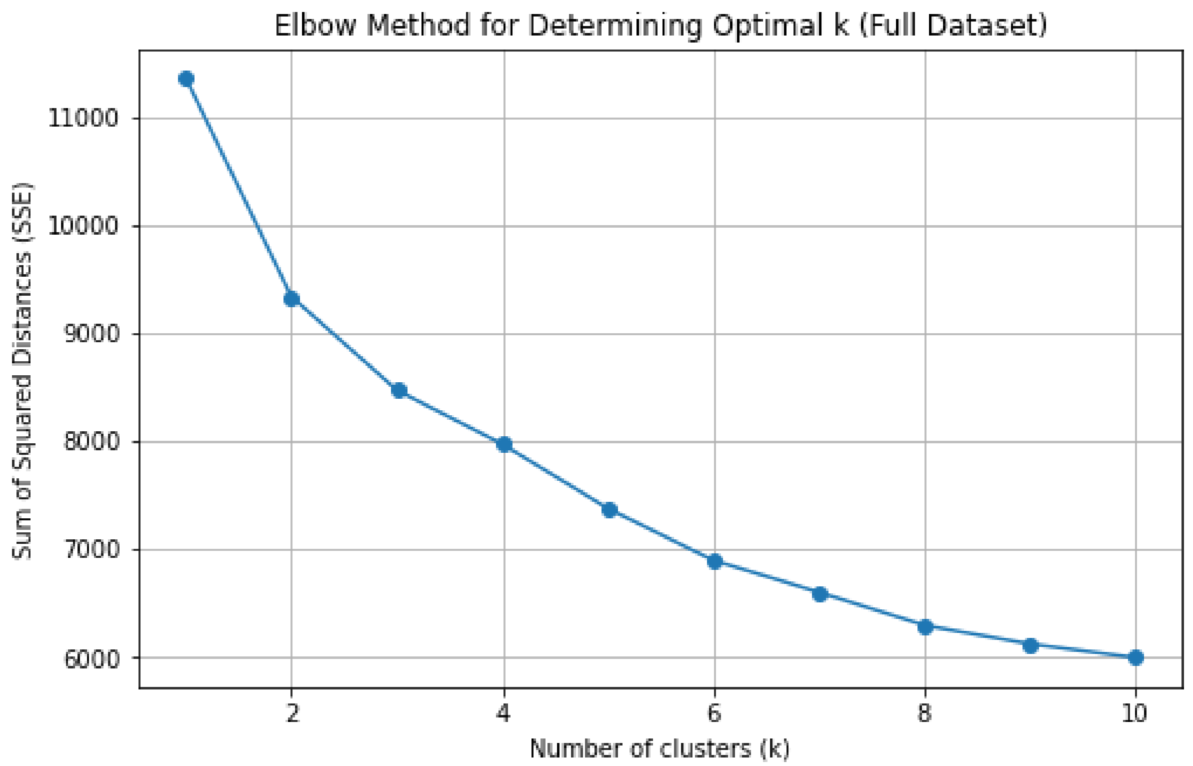


Unhealthy Recipes

- Word clouds appeared almost identical
- Contrasts with the expectation that unhealthy recipes would include more butter, oils, and sugar
    - Hypothesis: It's not the ingredients themselves but their quantities that differentiate healthy from unhealthy recipes
- Suggests that serving size and caloric content may also impact a recipe's healthiness

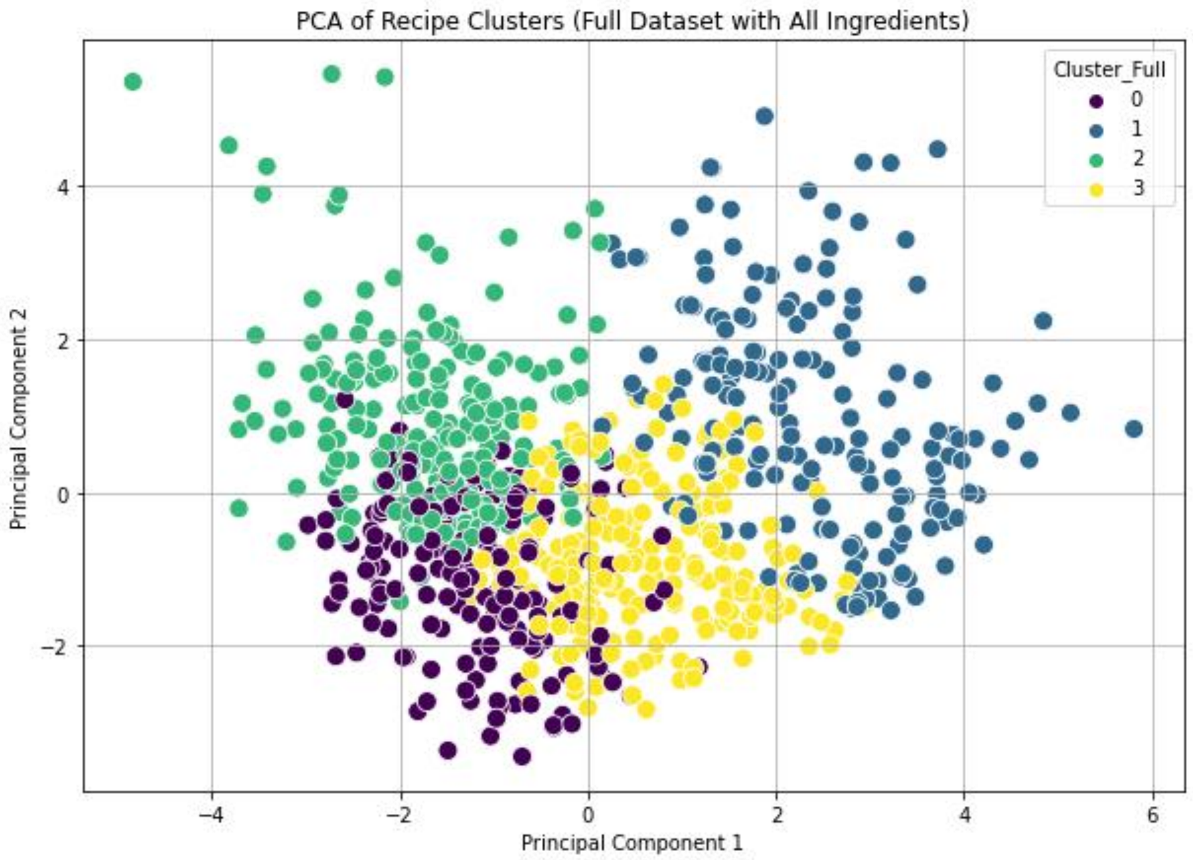
EDA – Elbow Method

k = 2, After this point, the SSE value levels off or decreases more gradually.



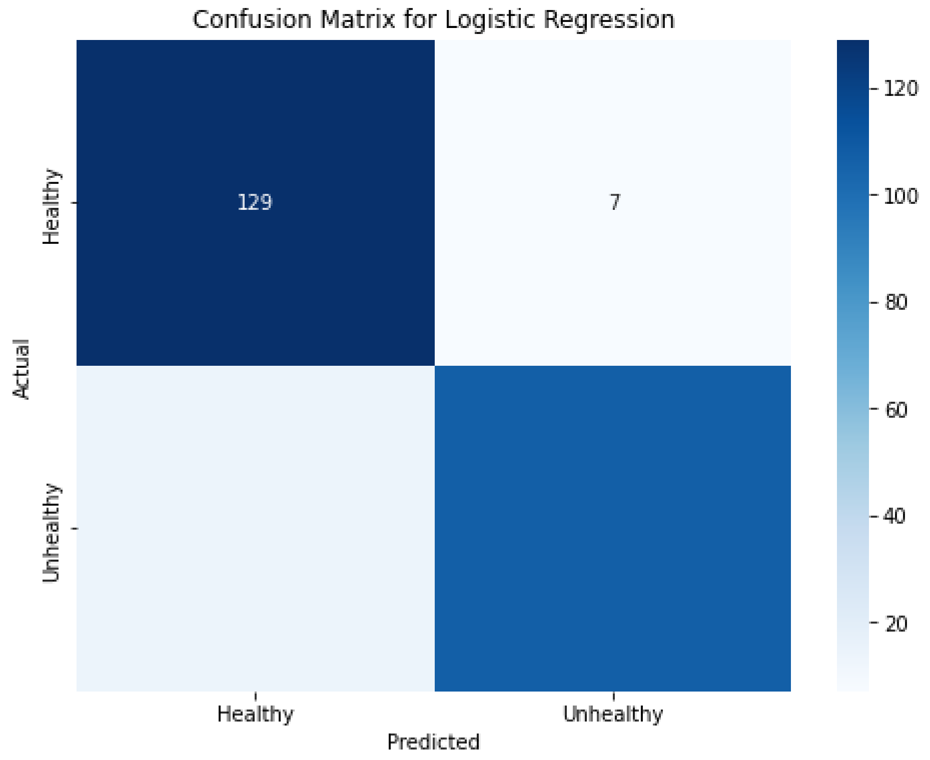
EDA - PCA

- Principal Component 1 is plotted on the x-axis, ranging from -4 to 6
- Principal Component 2 is plotted on the y-axis, ranging from -2 to 4.
- The graph displays four distinct clusters
  - Some overlap between clusters
  - Visible outliers within the cluster



LOGISTIC REGRESSION

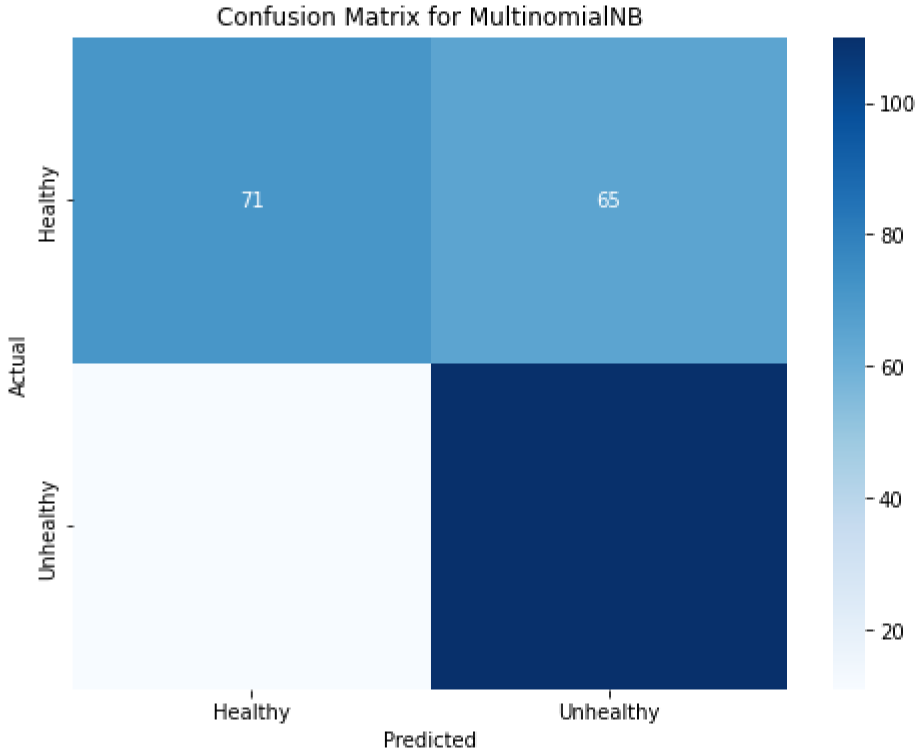
```
Evaluating classifier: Logistic Regression

Cross-validation scores: [0.91747573 0.93170732 0.91219512
0.94634146 0.93170732]
Average cross-validation score: 0.927885389533507
Accuracy on test set: 0.9182879377431906
Classification report:
              precision    recall  f1-score   support

     Healthy       0.90      0.95      0.92       136
   Unhealthy       0.94      0.88      0.91       121

    accuracy                           0.92       257
   macro avg       0.92      0.92      0.92       257
weighted avg       0.92      0.92      0.92       257

--------------------------------------------------
```

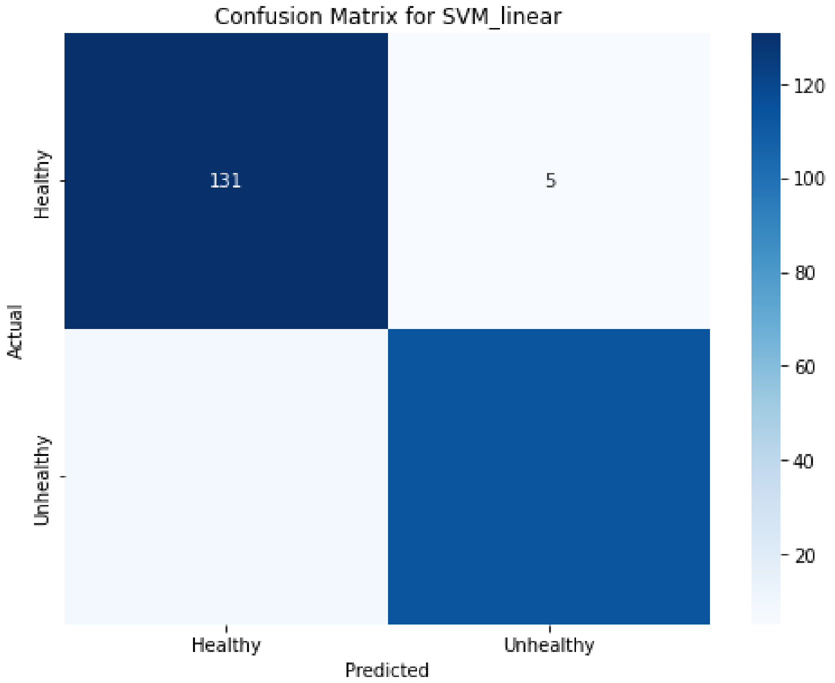


NAÏVE BAYES

Multinomial and Gaussian NB
Best performer: Multinomial

```
Evaluating classifier: MultinomialNB

Cross-validation scores: [0.65048544 0.69268293 0.62439024
0.65365854 0.71219512]
Average cross-validation score: 0.6666824532322992
Accuracy on test set: 0.7042801556420234
Classification report:
              precision    recall  f1-score   support

     Healthy       0.87      0.52      0.65       136
   Unhealthy       0.63      0.91      0.74       121

    accuracy                           0.70       257
   macro avg       0.75      0.72      0.70       257
weighted avg       0.75      0.70      0.69       257

--------------------------------------------------
```






```
Evaluating classifier: SVM_linear

Cross-validation scores: [0.9368932  0.93658537 0.95121951
0.96585366 0.94634146]
Average cross-validation score: 0.9473786407766991
Accuracy on test set: 0.9494163424124513
Classification report:
              precision    recall  f1-score   support

     Healthy       0.94      0.96      0.95       136
   Unhealthy       0.96      0.93      0.95       121

    accuracy                           0.95       257
   macro avg       0.95      0.95      0.95       257
weighted avg       0.95      0.95      0.95       257

--------------------------------------------------
```

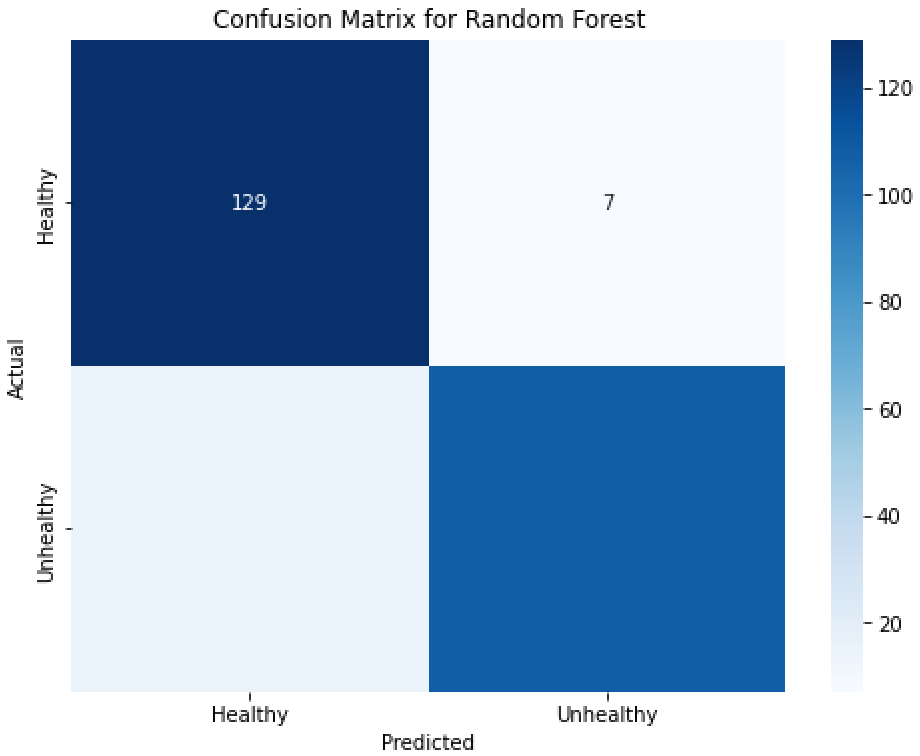
SUPPORT VECTOR MACHINES

Kernels ran: Linear, Polynomial, Radial Basis Function

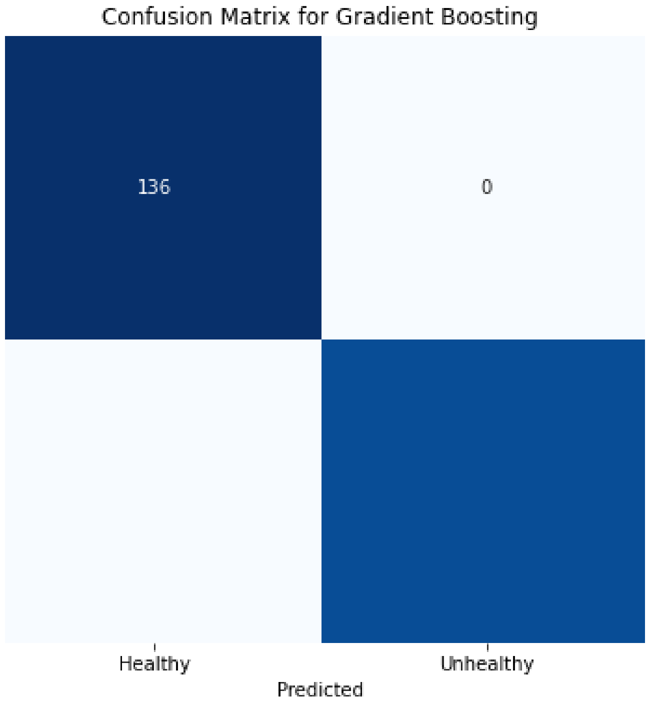
Best performer: Linear Kernel

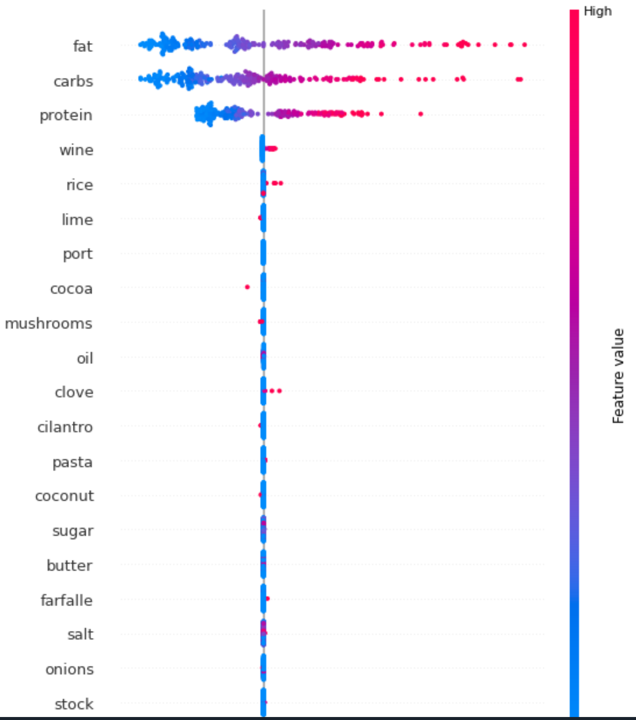
RANDOM FOREST

```
Evaluating classifier: Random Forest

Cross-validation scores: [0.94660194 0.92195122 0.89268293
0.92195122 0.93170732]
Average cross-validation score: 0.9229789249348805
Accuracy on test set: 0.9182879377431906
Classification report:
              precision    recall  f1-score   support

     Healthy       0.90      0.95      0.92       136
   Unhealthy       0.94      0.88      0.91       121

    accuracy                           0.92       257
   macro avg       0.92      0.92      0.92       257
weighted avg       0.92      0.92      0.92       257

--------------------------------------------------
```



GRADIENT BOOSTING





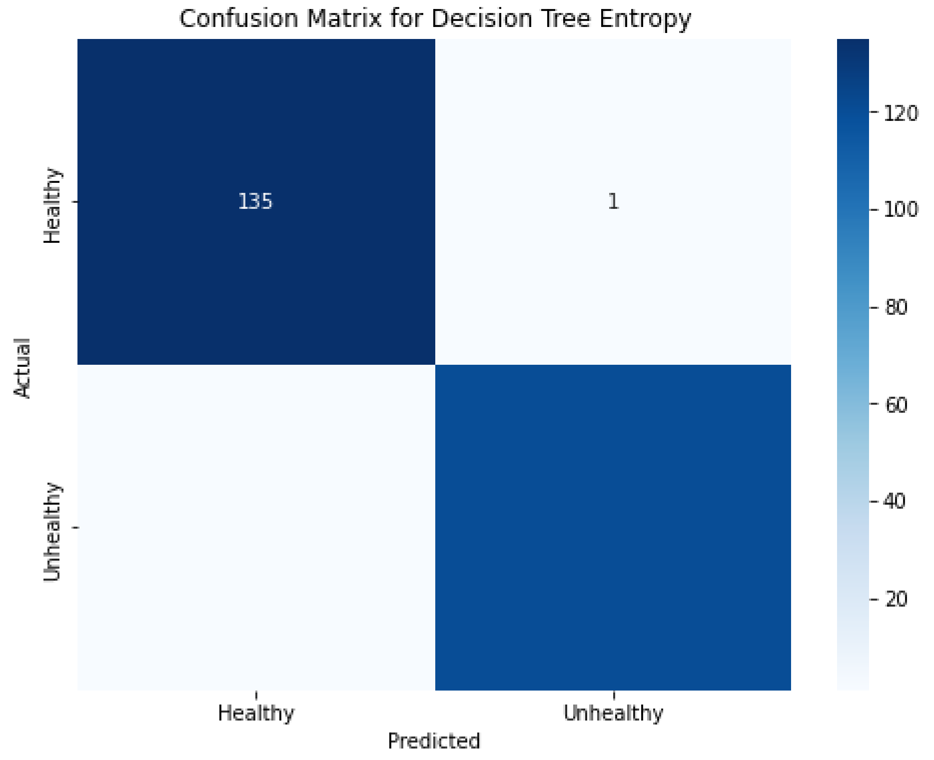

```
Evaluating classifier: Gradient Boosting

Cross-validation scores: [0.99514563 0.9804878  0.98536585
0.99512195 0.9902439 ]
Average cross-validation score: 0.9892730286526167
Accuracy on test set: 1.0
Classification report:
              precision    recall  f1-score   support

     Healthy       1.00      1.00      1.00       136
   Unhealthy       1.00      1.00      1.00       121

    accuracy                           1.00       257
   macro avg       1.00      1.00      1.00       257
weighted avg       1.00      1.00      1.00       257

-----------------------------------------------------
```

DECISION TREES

Entropy and Gini Impurity Criteria

Best performer: Entropy

```
Evaluating classifier: Decision Tree Entropy

Cross-validation scores: [0.98543689 0.96585366 0.98536585 1.
0.98536585]
Average cross-validation score: 0.9844044518115084
Accuracy on test set: 0.9922178988326849
Classification report:
              precision    recall  f1-score   support

     Healthy       0.99      0.99      0.99       136
   Unhealthy       0.99      0.99      0.99       121

    accuracy                           0.99       257
   macro avg       0.99      0.99      0.99       257
weighted avg       0.99      0.99      0.99       257

--------------------------------------------------
```



LATENT DIRICHLET ALLOCATION

```
Evaluating classifier: LDA

Accuracy on test set: 0.5652173913043478
Classification report:
              precision    recall  f1-score   support

           0       0.82      0.34      0.48        41
           1       0.48      0.89      0.62        28

    accuracy                           0.57        69
   macro avg       0.65      0.62      0.55        69
weighted avg       0.68      0.57      0.54        69

--------------------------------------------------
```

```
words representative of their class as identified by the lda
Topic #0:
sugar butter flour egg bake powder salt vanilla brown chocol
Topic #1:
pepper oil salt garlic onion chicken oliv sauc tomato chees
['latentdirichletallocation0' 'latentdirichletallocation1']
```

Methods and Models – Overall Results

Models and Accuracies:

- Logistic Regression (0.92)
- Multinomial Naive Bayes (0.70)
- Gaussian Naive Bayes (0.64)
- SVM (Linear Kernel) (0.95)
- SVM (Polynomial Kernel) (0.70)
- SVM (Radial Basis Function Kernel) (0.86)
- Random Forest (0.92)
- Gradient Boosting (1.0)
- Decision Trees (Entropy: 0.99, Gini: 0.98)
- Latent Dirichlet Allocation (0.57)

Top Three Models:

- **Gradient Boosting:**
  - Perfect accuracy: 1.0
  - Excellent precision, recall, and F1-score
  - Potential for overfitting, needs further validation
- **Decision Tree (Entropy):**
  - High accuracy: 0.99
  - Balanced performance
  - May need validation to ensure generalizability
- **Support Vector Machine (Linear Kernel):**
  - High accuracy: 0.95
  - Strong performance across classes
  - Recommended for further tuning

Results – Continued

Ingredient Patterns:

- Similar word clouds for healthy and unhealthy recipes
- Ingredient lists may not be the primary differentiator
- Other factors like ingredient quantities or serving sizes may be more significant

Quantitative Factors:

- Hypothesis: Quantity of ingredients (e.g., fats, sugars) differentiates healthy from unhealthy recipes
- Healthiness may be related to the amount consumed rather than ingredient presence

Caloric Content and Serving Size:

- Importance of serving size and caloric content in determining healthiness
- Recipes with similar ingredients may have different health implications based on portion sizes and calorie counts

Further Analysis Needed:

- Focus on ingredient quantities, calorie content, and serving sizes
- Refine model tuning and validate with new data to improve accuracy and robustness

Conclusion & Future Directions

- Complexity of modern eating habits and the influence of cultural traditions underscore the importance of effective tools and resources
- Recipes can guide individuals toward healthier choices, enhance cooking skills, and support a balanced lifestyle
- Foster positive lifestyle changes and improve overall quality of life
- Individuals can navigate dietary options more effectively and make meaningful strides toward achieving a healthier and more fulfilling life
- Can be useful for people with various health concerns (diabetes, hyperglycemia, etc.)
- Diet planning for kitchens at senior living facilities, schools, individual consumers
- Potential Use: app can be developed to use this data to help users track calories, fat, protein, etc.

*Note about healthy/unhealthy recipes