Wine Origin

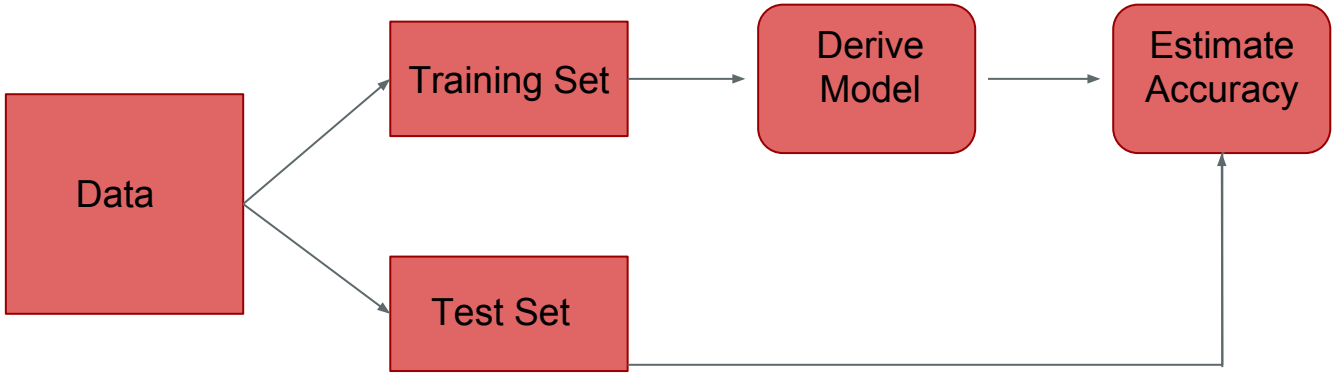
Lizzy Burr, Travis Conte, Spencer Crough,
Rob Gerth, Victor Ramirez, Adam Rosa

Introduction

- Data Set
  - 177 instances (wines)
  - 13 numeric characteristics, 3 possible regional classifications
- Goal
  - Create KNN model that can accurately predict classification of wine using its characteristics

Classification

- Predict origin of wine
- Build model with training set
  - Use remaining data (test data) to test accuracy of model

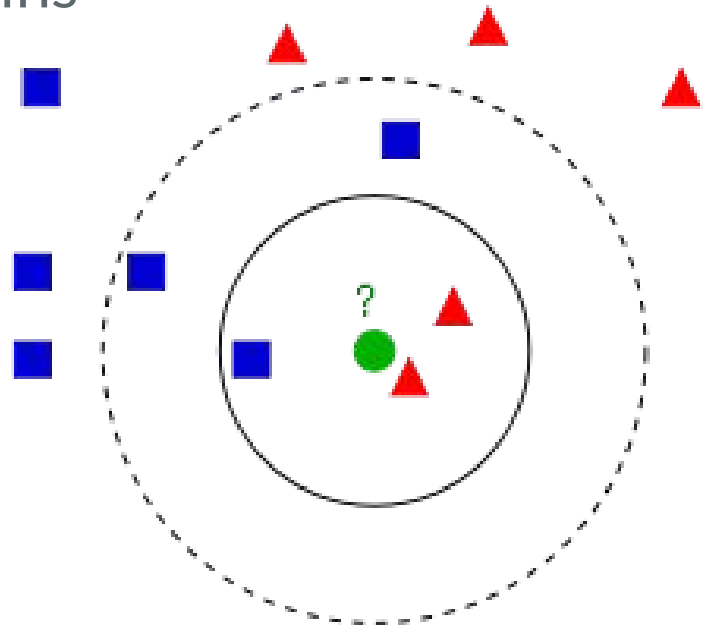


K Nearest Neighbors

- Each characteristic is a dimension on a set of axes
  - Easily visualizable for 2-3 characteristics
- Classifies new points based on coordinates, relative to coordinates of test data

K Nearest Neighbors

- New data points are classified by k nearest points
  - Majority “wins”

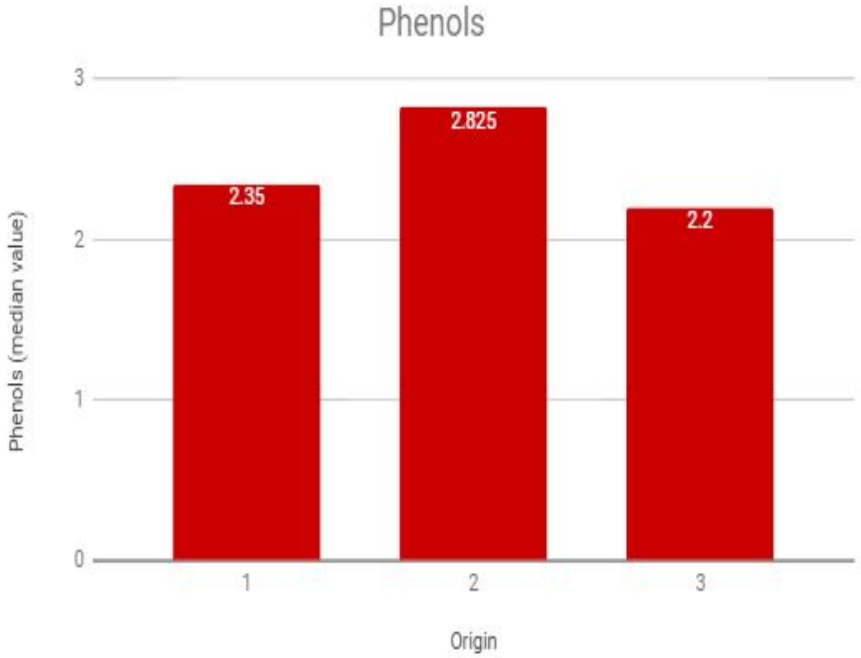


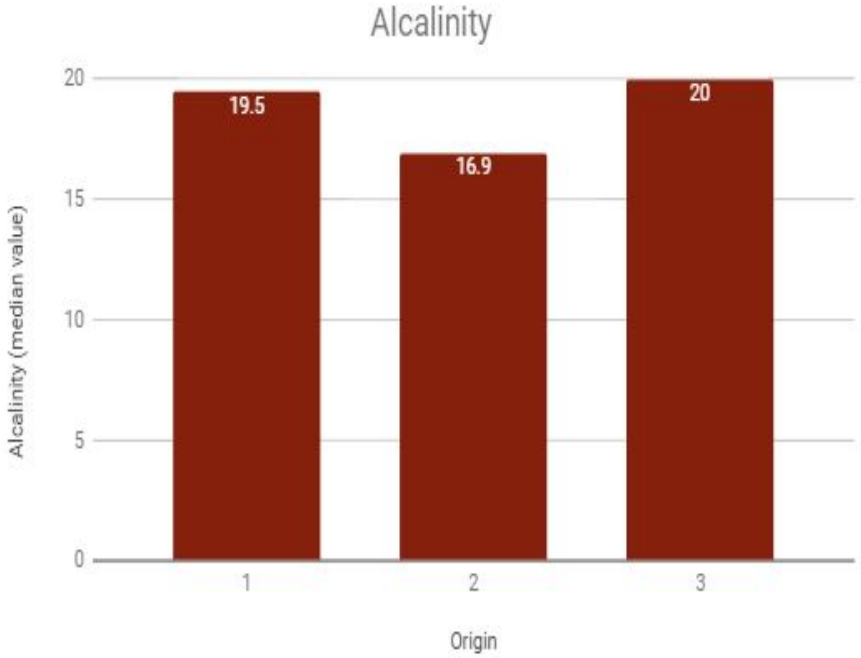
K Nearest Neighbors

How we intend to use it for our data set:

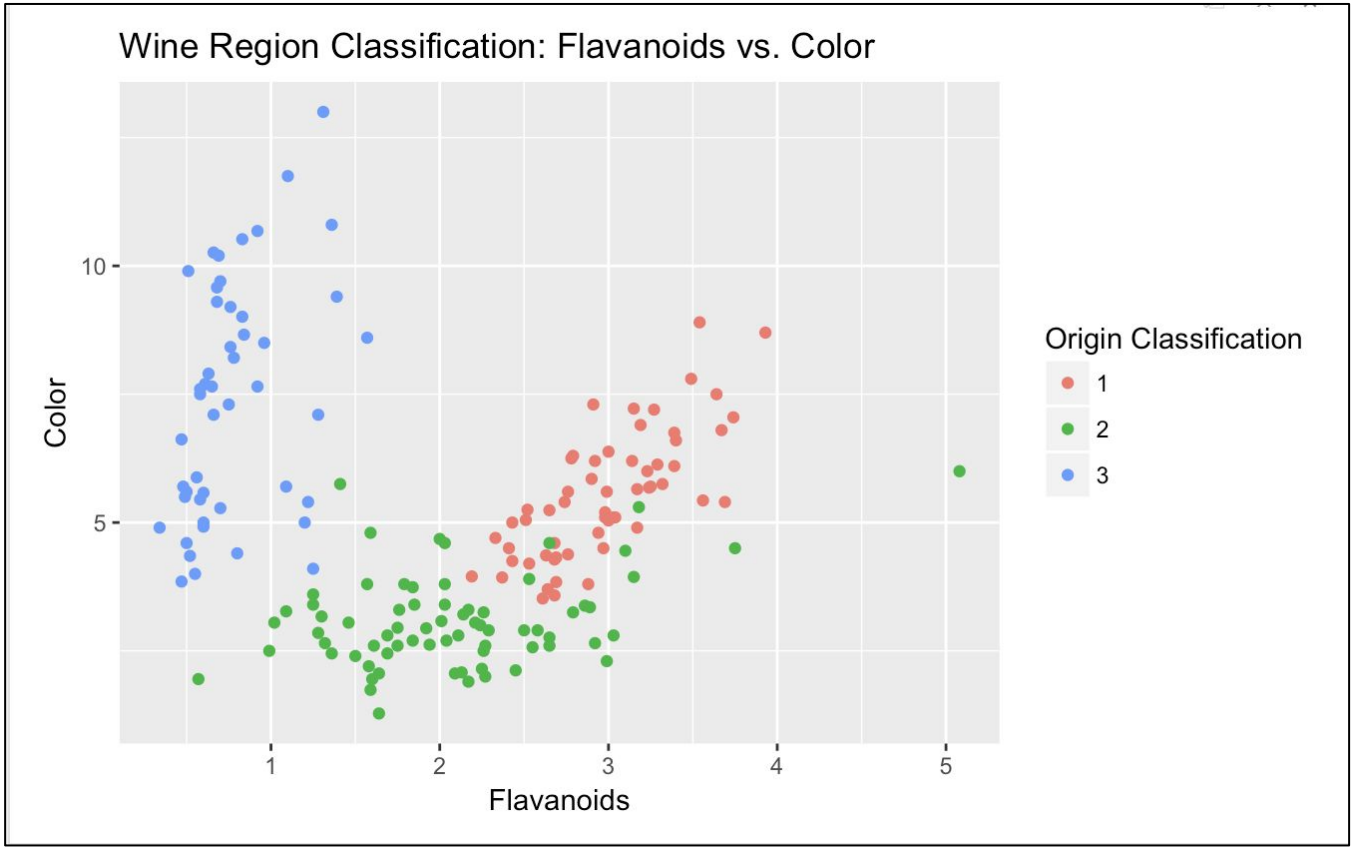
- Split data into two sets
  - Training data (approximately ⅔ of data)
  - Test data (remainder of data)
- Classify test data using models based on training data
- Determine how accurate the models are by comparing assigned classifications of test data to known values provided.

Part 1


Phenols
Phenols (median value)
3
2
1
0
2.35
2.825
2.2
1
2
3
Origin


Alcalinity
Alcalinity (median value)
20
15
10
5
0
19.5
16.9
20
1
2
3
Origin

Part 2


Wine Region Classification: Flavanoids vs. Color
Color
10
5
1
2
3
4
5
Flavanoids
Origin Classification
1
2
3

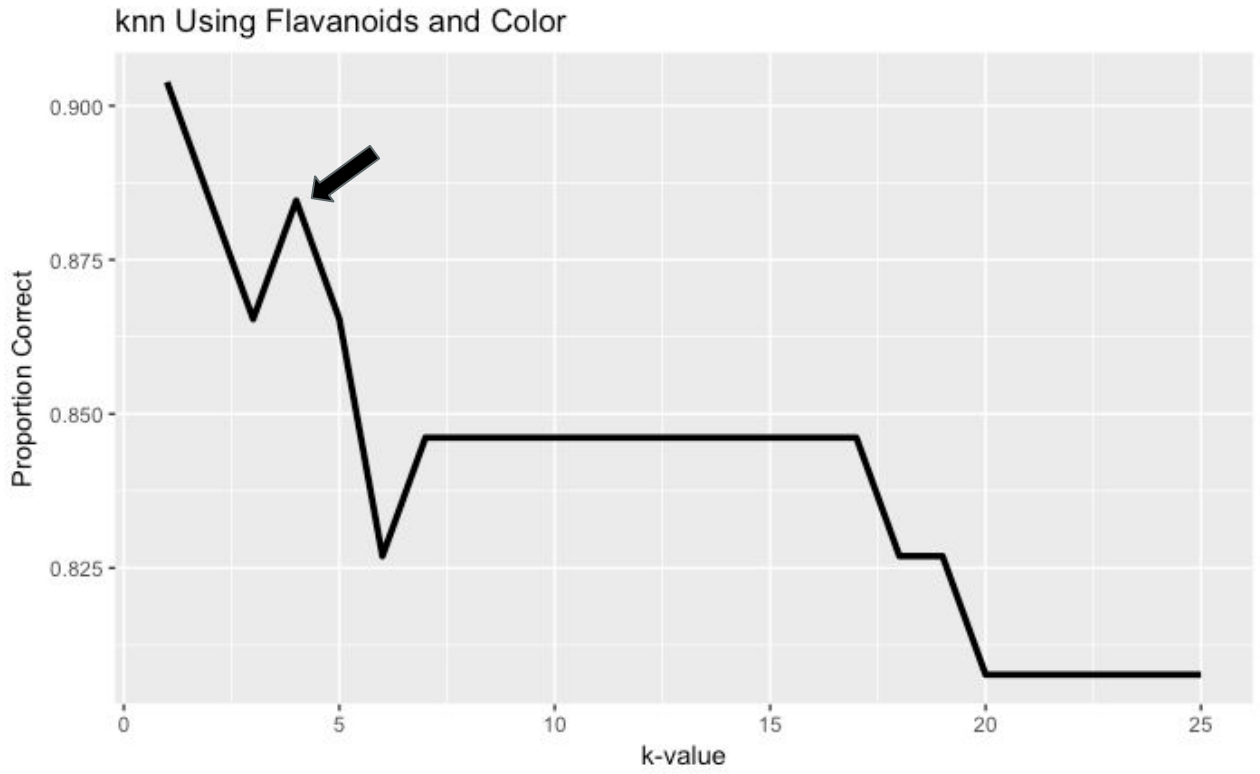
Confusion Matrix

- 2 Characteristics: colors and flavanoids
- k=4

	1	2	3
1	15	4	0
2	0	19	2
3	0	0	11

To find optimal k asses individual confusion matrix for each k value considered.

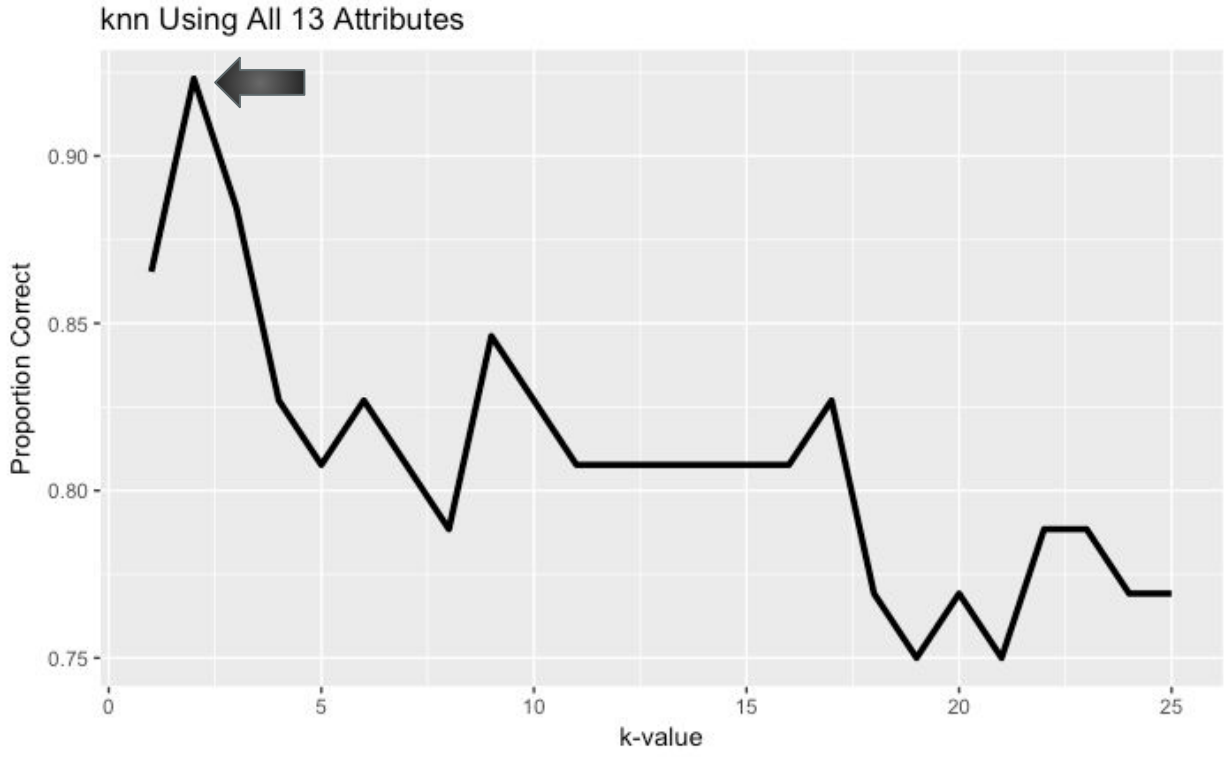
Best k for Flavanoids & Color



Optimal k: 3

Accuracy:89%

Best k for All Characteristics



Optimal k: 4

Accuracy: 95%

Results from repeated application

Color and Flavanoid accuracy with optimal k:

89%

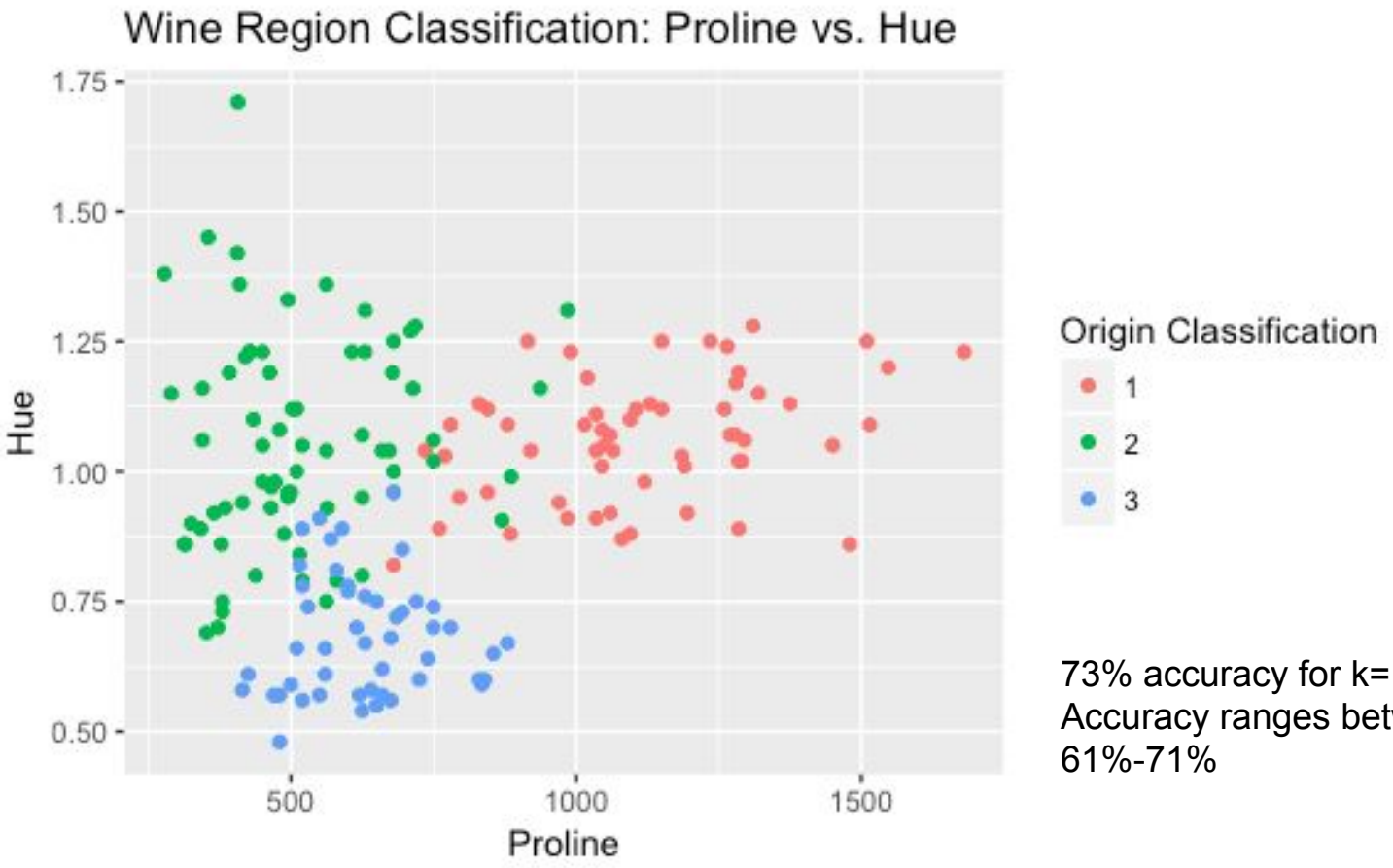
All characteristics accuracy with optimal k:

95%

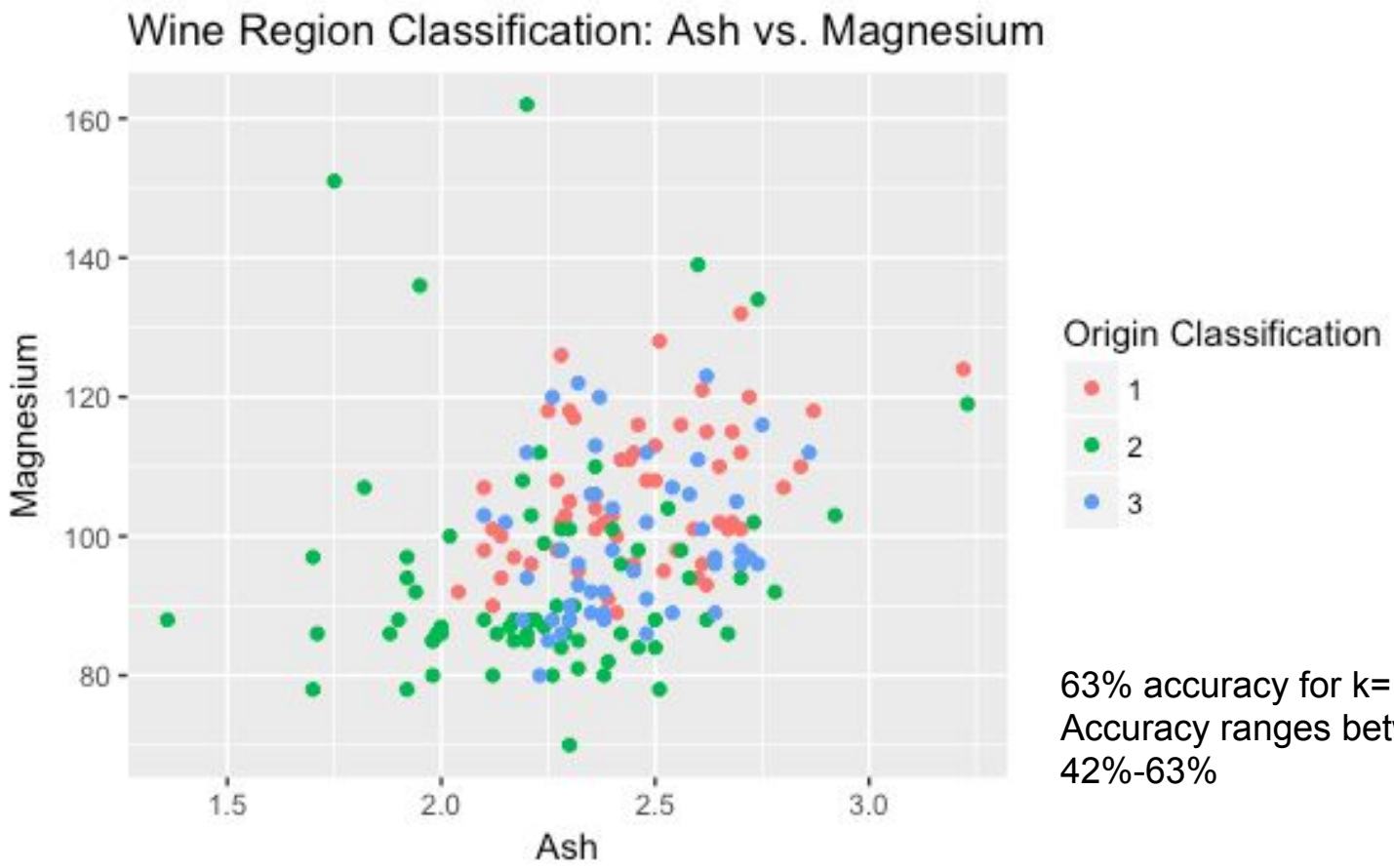
Tradeoff between accuracy and complexity

Further Experimentation

- Different two variable combinations for the simpler model
- Variable amounts between 2 and 13
- Change test and training set proportions
    - Smaller data set limits viable proportions to choose from



73% accuracy for k=12
Accuracy ranges between 61%-71%



63% accuracy for k=17 and 18
Accuracy ranges between 42%-63%

References

https://upload.wikimedia.org/wikipedia/commons/thumb/e/e7/KnnClassification.svg/220px-KnnClassification.svg.png