Predicting Malignant and Benign Breast Cancer

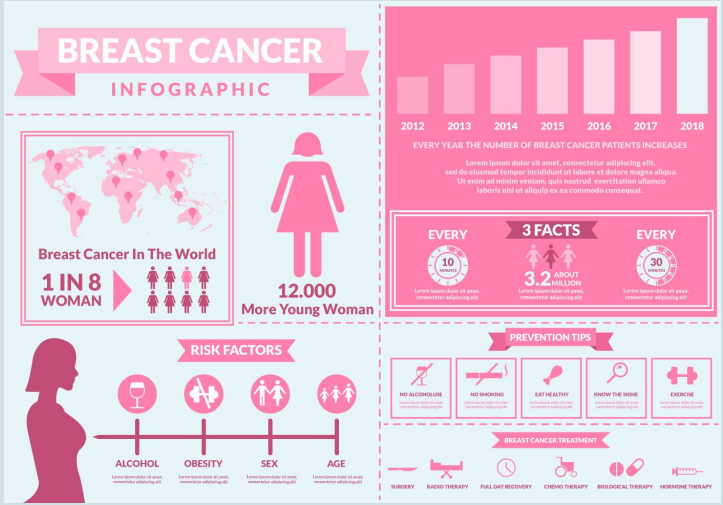
Alexa Kelly, Ryan Folks, and Lyndon Swarey.

About our Dataset

It was created by the **University of Wisconsin** in **1992**, and was collected from *569 breast cancer patients*.

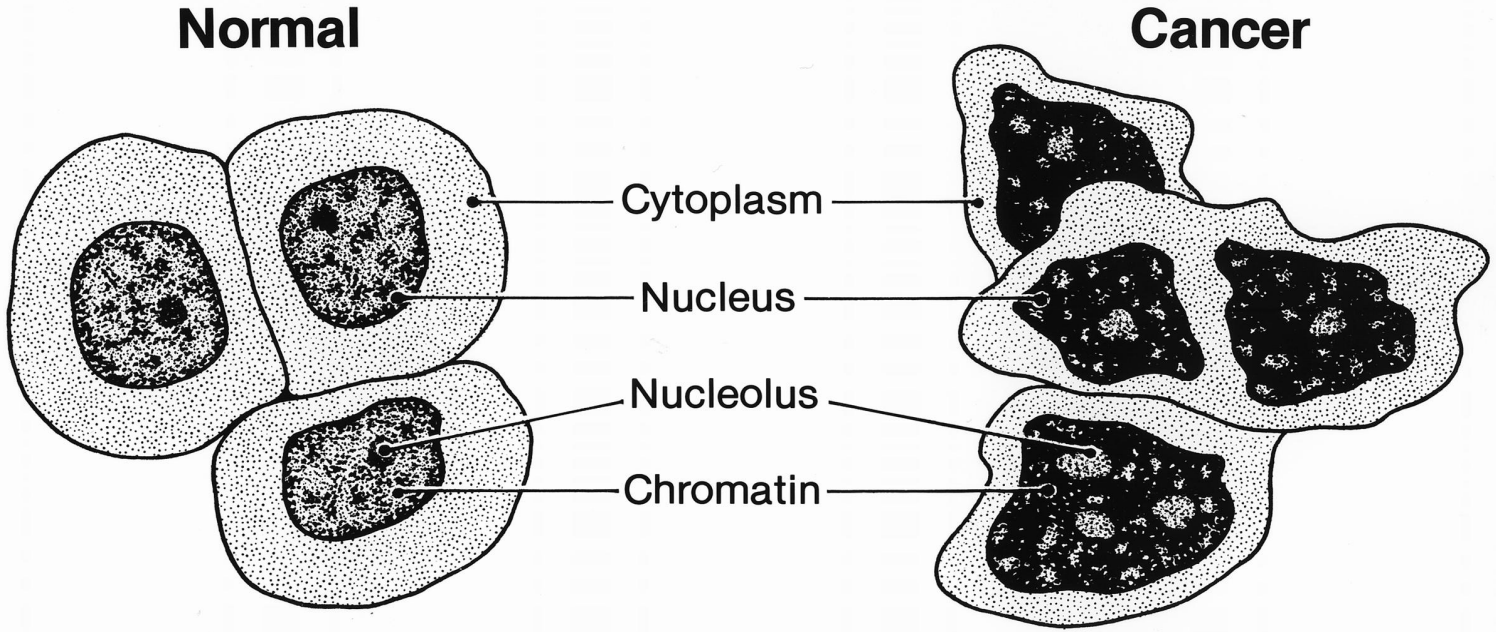
Each datapoint is composed of the attributes of the **nuclei of a cluster of cells** obtained via **fine needle aspiration** and analyzed under a microscope.

An observation consists of an ID number, a diagnosis (M or B), mean, standard error, and “worst” (average of the three largest values) for a 10 factors.



Normal and Cancer Cells

Structure



— Large cytoplasm
— Single nucleus
— Single nucleolus
— Fine chromatin

— Small cytoplasm
— Multiple nuclei
— Multiple and large nucleoli
— Coarse chromatin

Objective

- determine the type of tumor a breast cancer patient has, **malignant (M) or benign (B),** by using data procured from biopsy samples of cancer patients.

- compare the results from **K-NN, Linear Discriminant Analysis** and **Quadratic Discriminant Analysis** to determine which ML algorithm is best for predicting the type of cancer in a small dataset, using *all the variables* and *two subsets of the variables*
    - The subsets of variables are selected using the **Chi-square test and Classification Tree**.

- Two biggest concerns were **accuracy** and **recall**.

Approach overview

Our algorithm starts by **K-means clustering** to bin the variables .

The **Elbow method** was used to find the optimal k

Then, we then ran a **Chi-square test** and a **Classification Tree** on the newly created binned variables.

A **Box Cox transformation** was used to make these variables normally distributed.

Performed **Linear Discriminant Analysis** and **Quadratic Discriminant Analysis.**

K-Means Clustering Results

To select k, k-means was run using different k's and the within-cluster sum of squares sum (WSS) was computed for each k.

The WSS for each k was plot and the elbow method was used to select an optimal k. This optimal k determined the number of bins for that feature.

The optimal k was 3 or 4 depending on the feature.

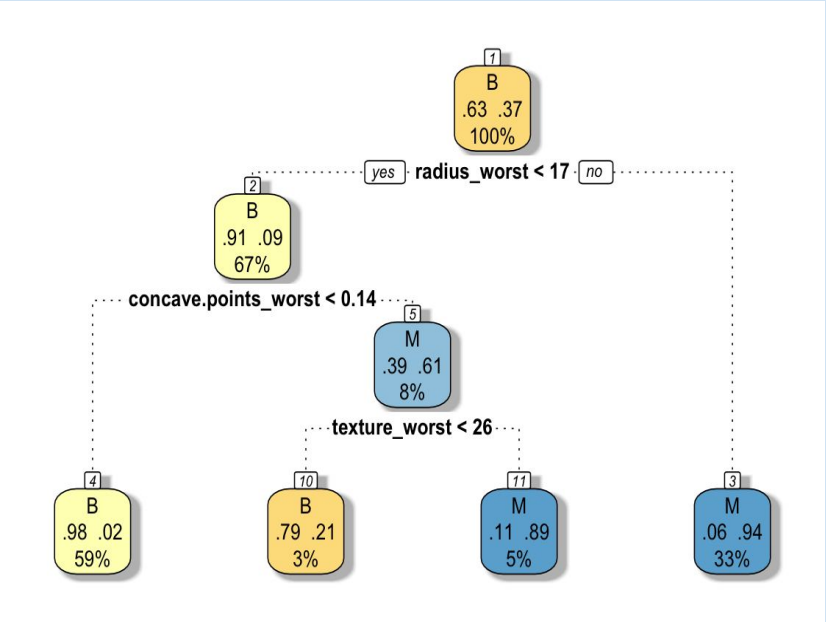
```
# Kmeans wss method
k_wss <- kmeans(data_3,
centers = 3, nstart = 10)
```

Variable selection

A **Chi-Squared test** and **Classification Tree** was used for feature selection, creating three different ways we looked at our dataset.

The objective of a classification tree is to minimize the cross-entropy or the Gini index of the target variable. The cross-entropy and Gini index are measurements of purity within a node.

Recall:
M stands for Malignant, and
B stands for Benign!

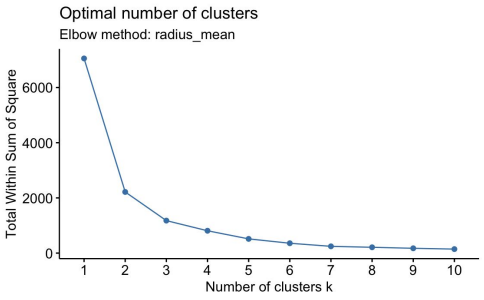


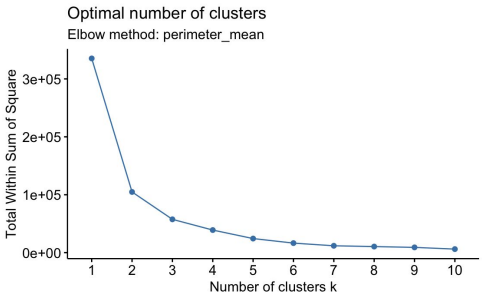
Elbow Method  vs. Silhouette method 

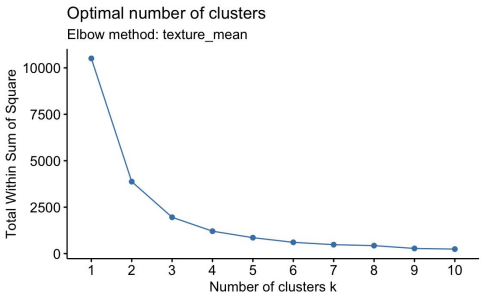
We considered both the elbow method and silhouette method in selecting the optimal k. The elbow and silhouette methods are really just two ways of achieving the same goal.

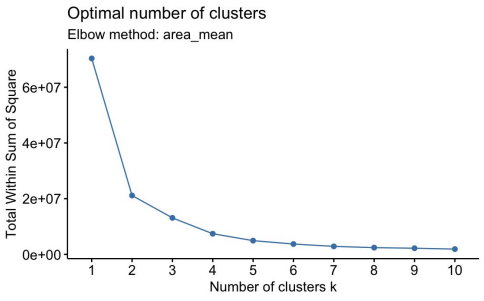
We decided on pursuing the elbow method, because we saw that it created varying bins for the continuous variables.

Elbow Method









Plots the sum of squares vs a growing number of k clusters.

The optimal number will end up being in the “elbow” of this graph.

```
# Graph for wws method
fviz_nbclust(data_3, kmeans,
method = "wss") +
  labs(subtitle = "Elbow method:
radius_mean")
```

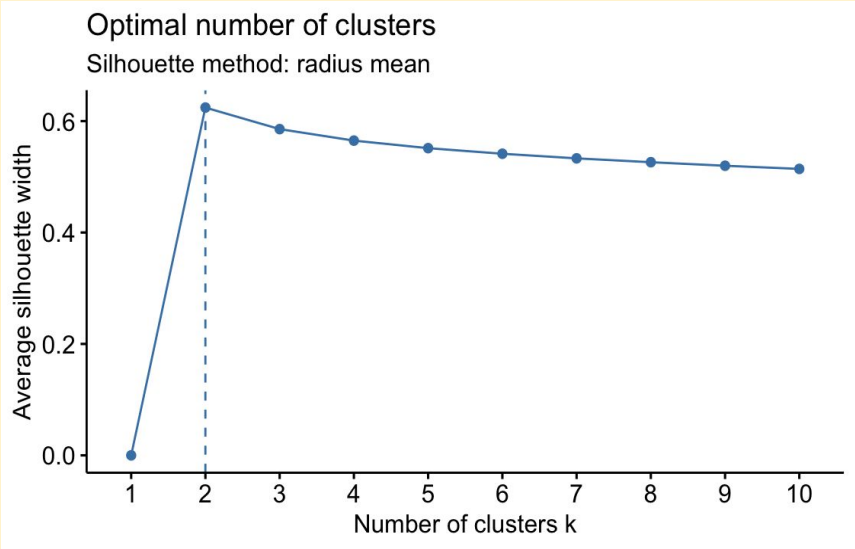
Silhouette method

(abandoned)

The silhouette method is calculated using the mean intracluster distance and the mean nearest-cluster distance for each sample.

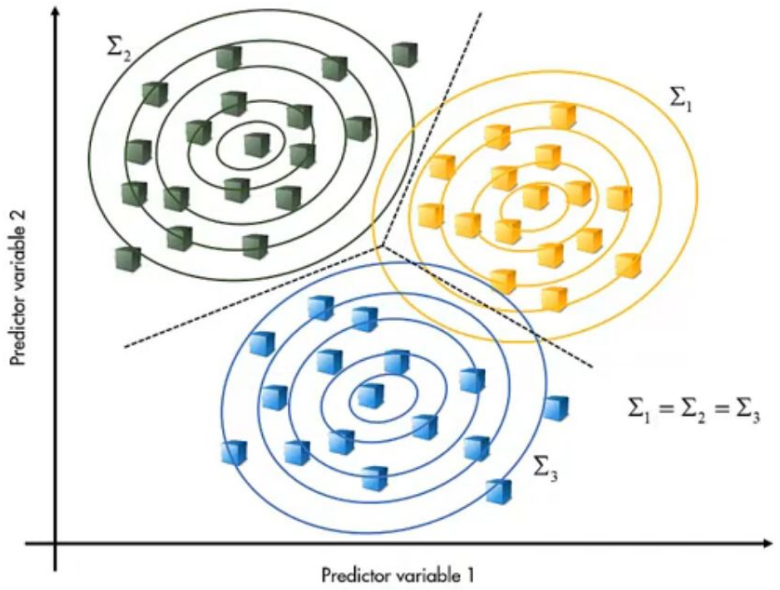
The optimal number of clusters k is the one that maximizes the average silhouette over a range of possible values for k.

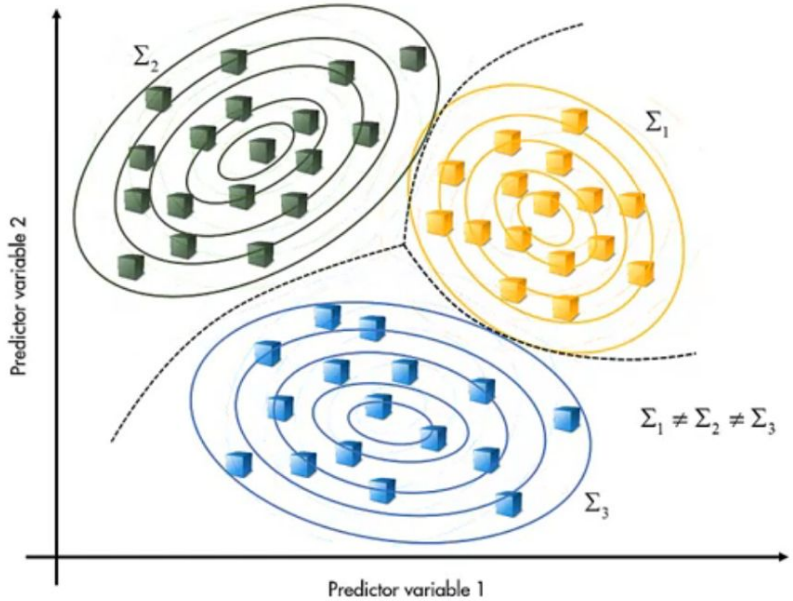
The average silhouette approach measures the quality of a clustering. A high value is desirable and indicates that the point is placed in the correct cluster.



LDA and QDA

- Attempt to find the optimal linear/quadratic combination of the predictors that gives the best separation of the predictor class.
- Attempt to maximize the distance between the mean of the response variable's categories, while minimizing the variance (or scatter)
- LDA and QDA made a few initial assumptions (normality, equal variance (lda), non-collinear)
- Boundary is where the probability of two multidimensional normal distributions is equal.


Σ_2
Σ_1
$\Sigma_1 = \Sigma_2 = \Sigma_3$
Σ_3
Predictor variable 2
Predictor variable 1


Σ_2
Σ_1
$\Sigma_1 \neq \Sigma_2 \neq \Sigma_3$
Σ_3
Predictor variable 2
Predictor variable 1

Results from LDA and QDA

The results of the three best models are shown.

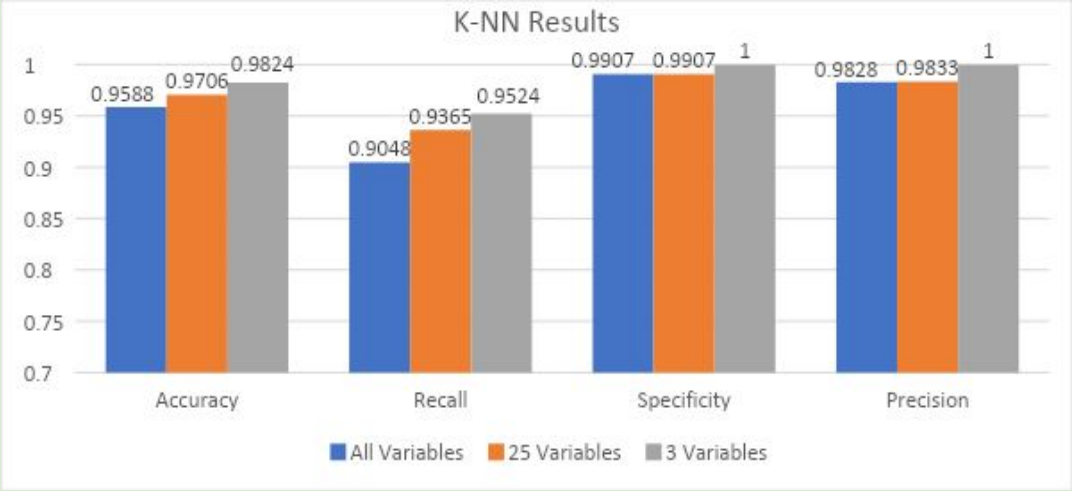
Confusion Matrix: All Variables (LDA)		
Predicted	Reference	
	Benign	Malignant
Benign	107	5
Malignant	0	58

Confusion Matrix: 3 Variables (LDA)		
Predicted	Reference	
	Benign	Malignant
Benign	104	5
Malignant	3	58

Confusion Matrix: 3 Variables (QDA)		
Predicted	Reference	
	Benign	Malignant
Benign	104	5
Malignant	3	59

KNN

- The results indicate that K-NN does the best at classifying cancer as **malignant** with *3 variables*.
    - Has the highest accuracy of **0.9824** with also the highest recall of **0.9524.**



Confusion Matrix

Confusion Matrix: All Variables		
Predicted	Reference	
	Benign	Malignant
Benign	106	6
Malignant	1	57

Confusion Matrix: 3 Variables		
Predicted	Reference	
	Benign	Malignant
Benign	107	3
Malignant	0	60

Confusion Matrix: 25 Variables		
Predicted	Reference	
	Benign	Malignant
Benign	106	4
Malignant	1	59

Summary of Results

All methods had a predictive accuracy of >94%. (High? Perhaps not high enough.)

In many cases, having all variables present in the models improved accuracy slightly.

> *Predicting benign tumors correctly, while important, is secondary to detecting malignant tumors.*

Most methods had around a 92% predictive accuracy in the malignant class.

This dataset turned out to yield highly accurate models (intrinsic to the kind of data, feature engineering)

Tried or Thought About

Naive Bayes

- model is easy to build and particularly useful for very large data sets. There are two parts to this algorithm: Naive and Bayes!
- Although it is a decent classifier, it is known to be a *bad estimator*

and

Logistic Regression

- measures the relationship between the dependent variable and the one or more independent variables (our features), by estimating probabilities using it's underlying logistic function.
- Yields a discrete binary outcome between 0 and 1.
- It is an algorithm that is known for its *vulnerability to overfitting*.

Further Work and Considerations

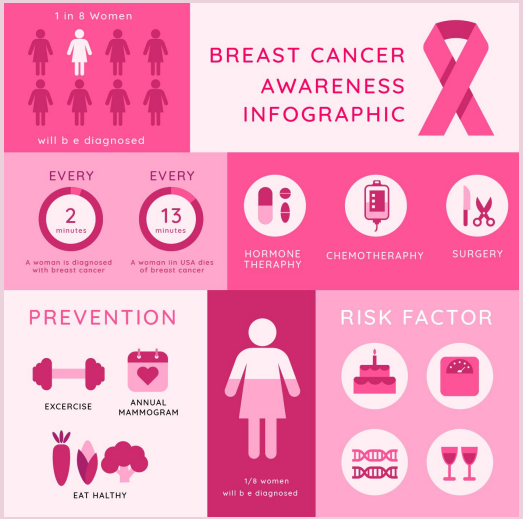
- The model did not have perfect accuracy, so of course there is room for improvement.
  - It does predict all the benign cancer correctly and the model does not give any false positives.
  - Based on the confusion matrix, the model with 3 variables has a false negative rate of .0272. Which means we had recall of 99.9728%
- *So, did we obtain accuracy and recall?*
  - 'High' accuracy models ✓
  - 'High' recall ✓

Of course, even though most of our models had 92% predictive accuracy on malignant tumors, it still may not be high enough for hospitals and healthcare professionals.

Citation

Credit to Dua, D. and Gra, C. (2019). UCI Machine Learning Repository, Irvine, CA: University of California (UCI), School of Information and Computer Science .Predicting Malignant and Benign Breast Cancer| ML Algorithm Report7

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29.



Questions?