Identifying Malignant Tumors

Presented By: Brady Hoskins, Eli Schultz, Don Marek

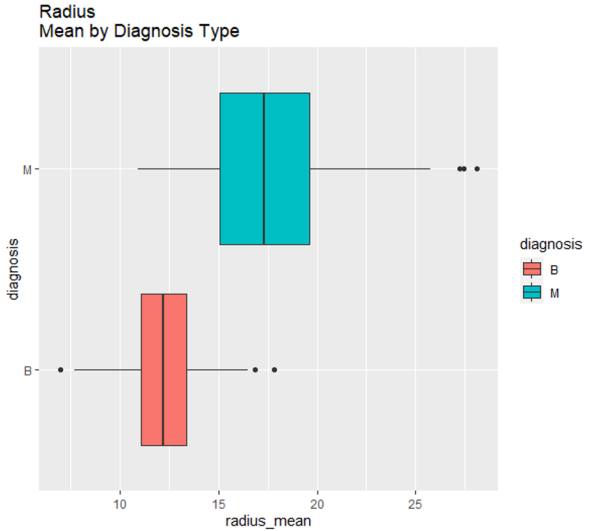
The Problem and Data

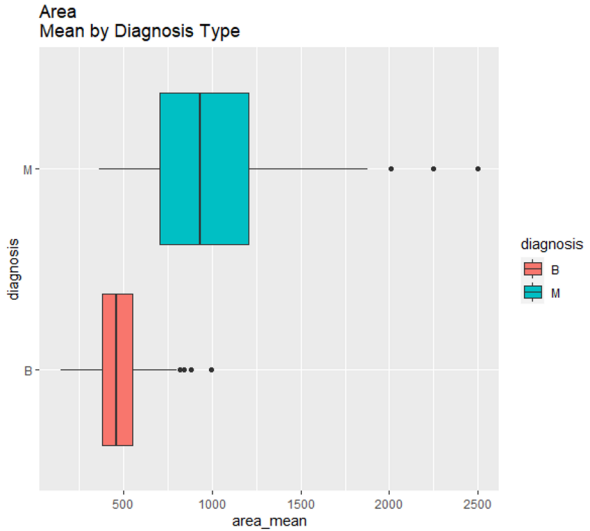
- Can we use machine learning techniques to assist in identifying malignant tumors?
- Sample set of 500 measurements of cell nuclei in breast tissue masses
  - Characteristics for each measurement
    - Mean
    - Standard Error of the mean
    - Maximum
  - Diagnosis
    - 357 Benign
    - 212 Malignant

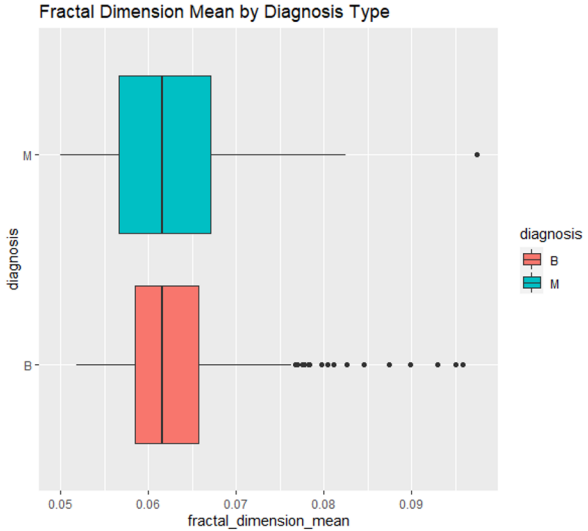
Exploratory Data Analysis

- Look at characteristics in the data
  - Radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimensions
- Plot characteristics vs diagnosis
  - Preliminary look at similarities and differences between each characteristic and the diagnosis
  - Provides an understanding of the data before beginning evaluation of ML models

Data Analysis Plots







Noticeable Difference Between Benign & Malignant

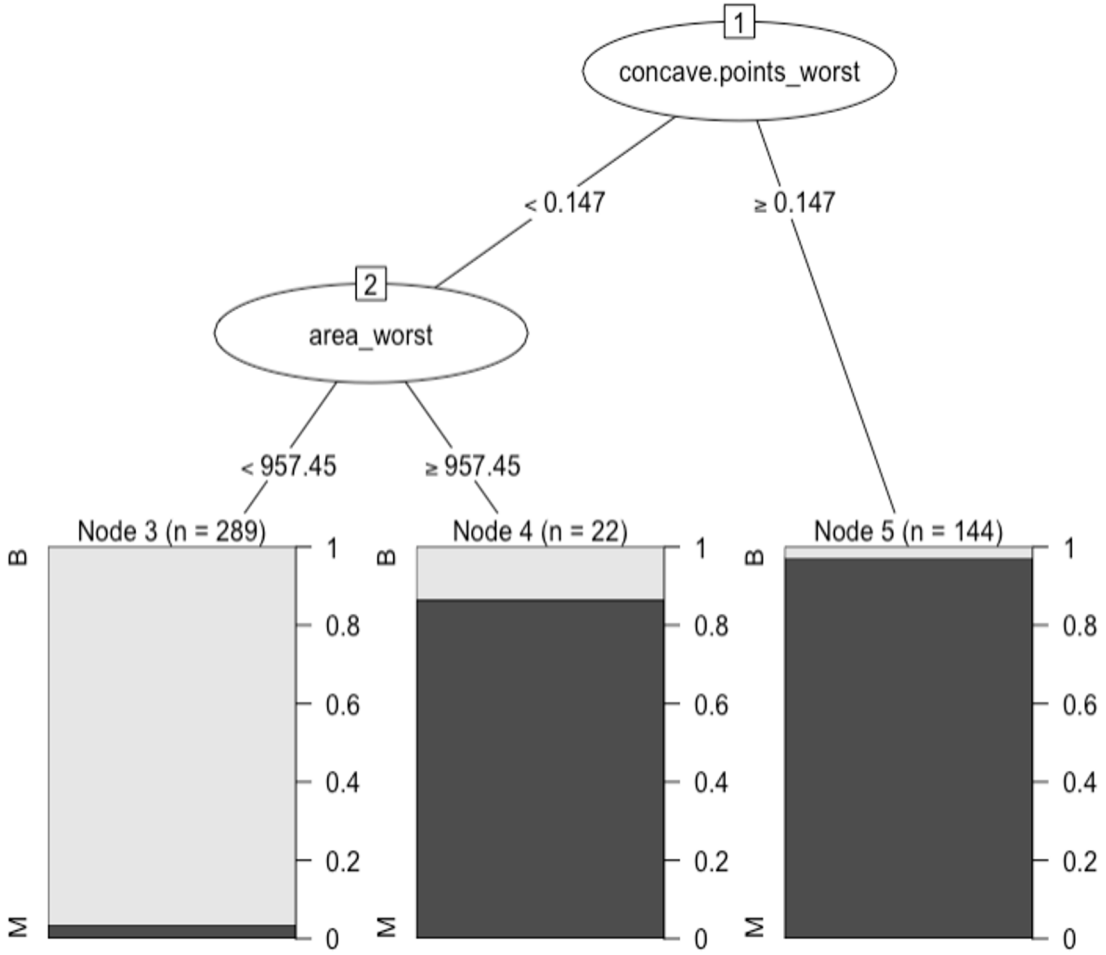
No Noticable Difference Between Diagnoses

Evaluation Criteria

- Confusion Matrices
  - Actual vs Predicted
- Misclassification Rate
  - % Incorrectly classified
- Sensitivity
  - % of Actual Malignant Correctly Classified

Decision Tree Model

- Classification based on logical splits
  - concave.points_worst < 0.147 and area_worst < 957.45 = Benign, else Malignant



Model Comparison: Decision Tree

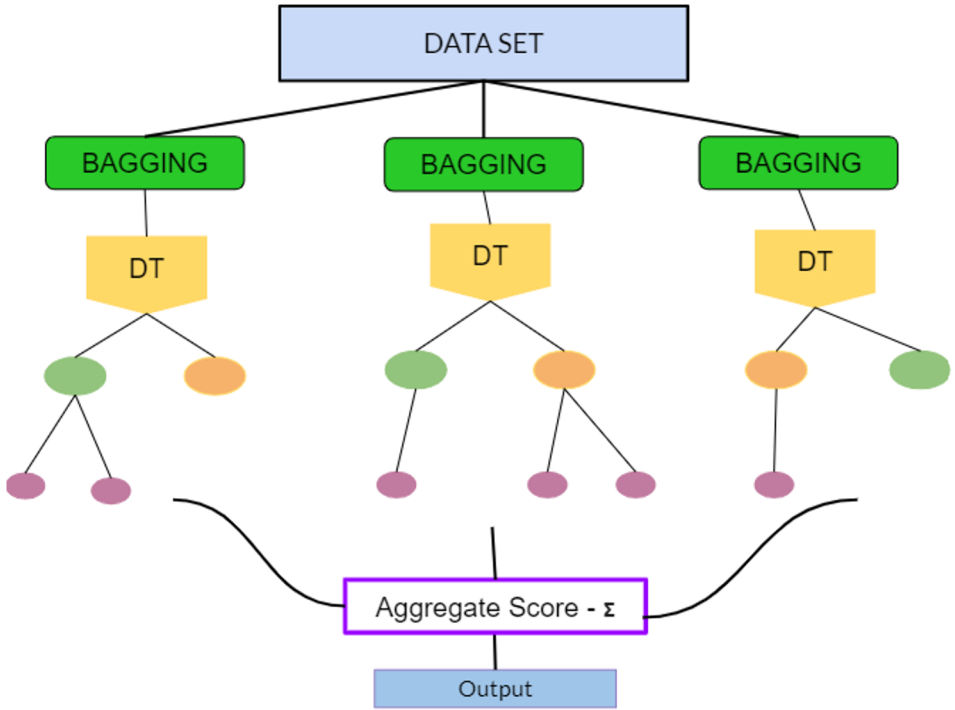
Predicted

Actual		Benign	Malignant
	Benign	67	6
	Malignant	4	37

Misclassification Rate: 8.77%
Sensitivity: 0.8605

Bagged Trees Model

- Combines 400 different decision trees
  - Each tree uses a random subset of data
  - Uses all features
  - More robust than single tree
- Important Features
  - Concave Points (Max, Mean)
  - Perimeter (Max)
  - Area (Max)



https://medium.com/analytics-vidhya/ensemble-models-baggings-vs-boosting-8affa6d18098

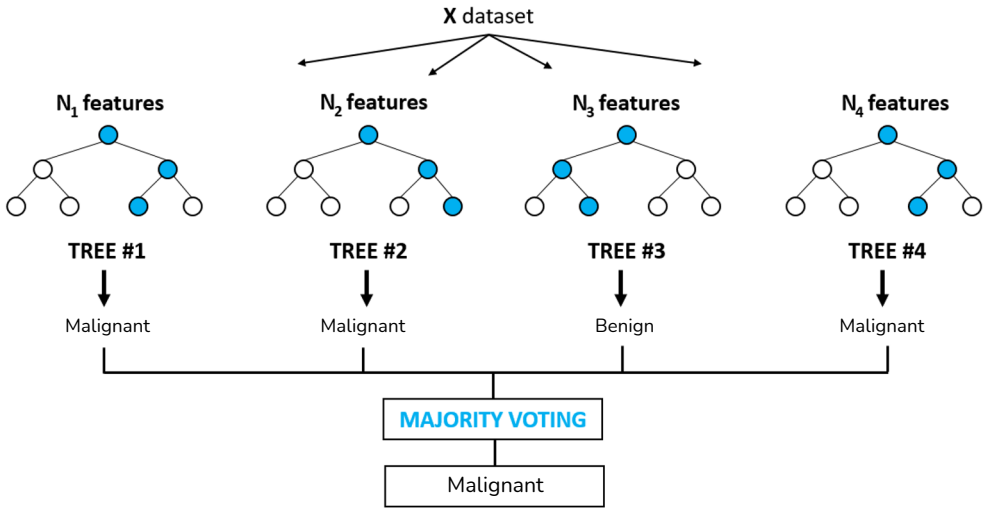
Model Comparison: Bagged Trees

Predicted

Actual		Benign	Malignant
	Benign	67	5
	Malignant	4	38

Misclassification Rate: 7.89%
Sensitivity: 0.8837

Random Forest Model



https://medium.com/@ar.ingenious/applying-random-forest-classification-machine-learning-algorithm-from-scratch-with-real-24ff198a1c57

- Combines 400 different decision trees
  - Similar to Bagged Trees
  - Optimize # of features per tree
  - 16 features per tree
- Important Features
  - Concave Points (Max, Mean)
  - Perimeter (Max)
  - Radius (Max)
  - Area (Max)

Model Comparison: Random Forest

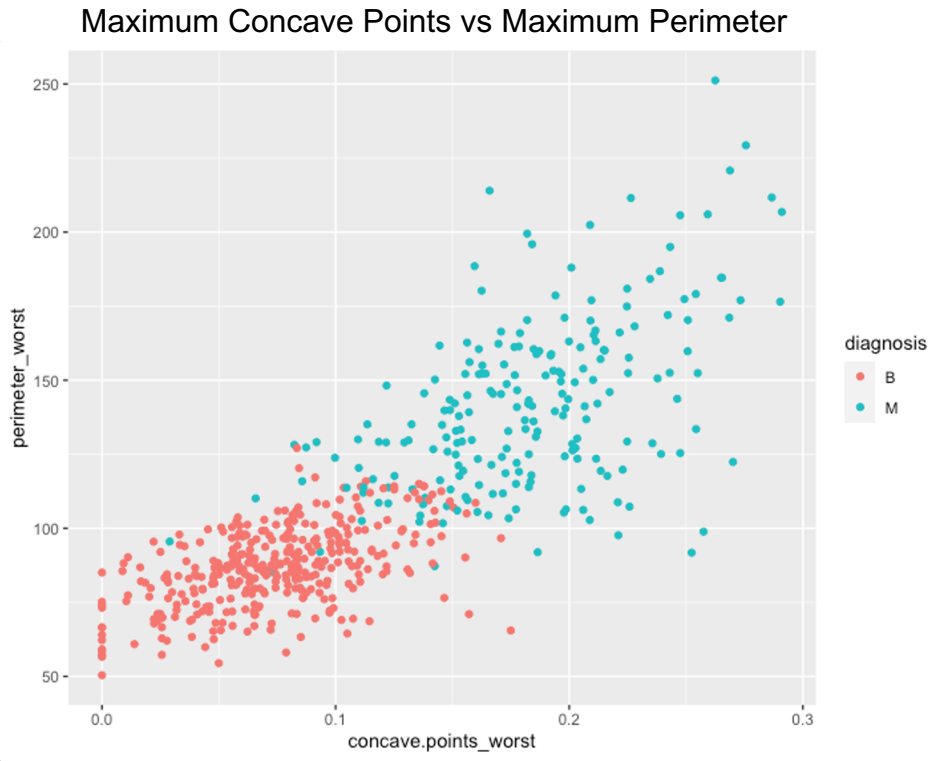
Predicted

Actual		Benign	Malignant
	Benign	68	5
	Malignant	3	38

Misclassification Rate: 7.02%
Sensitivity: 0.8837

K-Nearest Neighbors Model

- Classification based on proximity of patient's data to K number of closest patients
  - Experiment to find optimal K value
  - K = 5
  - Majority vote of 5 nearest patients



Model Comparison: K-Nearest Neighbors

Predicted

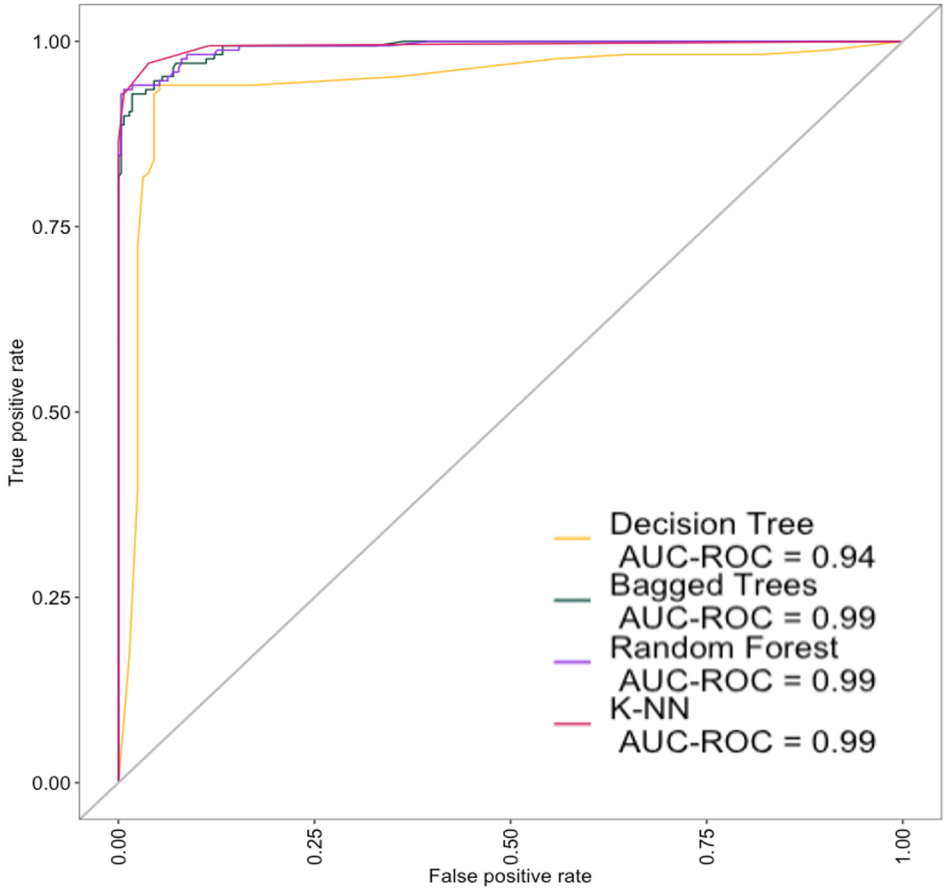
Actual		Benign	Malignant
	Benign	70	5
	Malignant	1	38

Misclassification Rate: 5.26%* (Best Performance)
Sensitivity: 0.8837

Model Comparison: ROC Curves

True positive rate vs false positive rate

- Find cutoff for optimal sensitivity
- AUC = Area Under the Curve
  - # to quantify curve



Recommendation:

- KNN Model
  - Lowest misclassification rate
  - Highest sensitivity