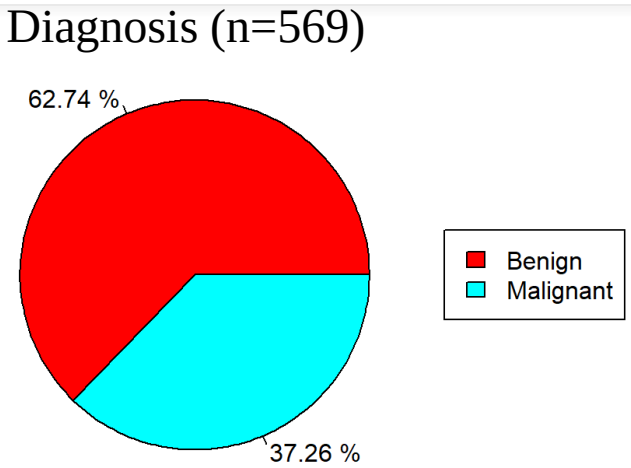
Comparison of Machine Learning Models for Predicting Breast Cancer Malignancy

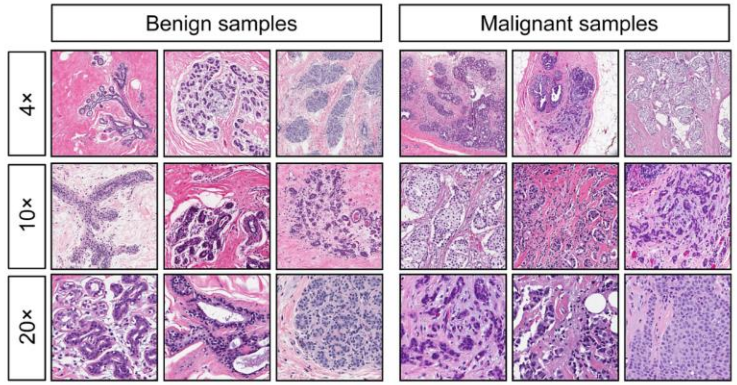
Breast Cancer Wisconsin (Diagnostic)

Andrew Taranto, May 12, 2023

Data: Breast Cancer Wisconsin (Diagnostic), UC Irvine

- 569 observations split roughly between 2/3^{rds} benign and 1/3rd malignant tumors, our target variable
- 30 continuous quantitative metrics of size and quality of cancerous tissue
- "Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image."





Sources: Data set description https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic), image https://medium.com/analytics-vidhya/breast-cancer-diagnostic-dataset-eda-fa0de80f15bd

Method

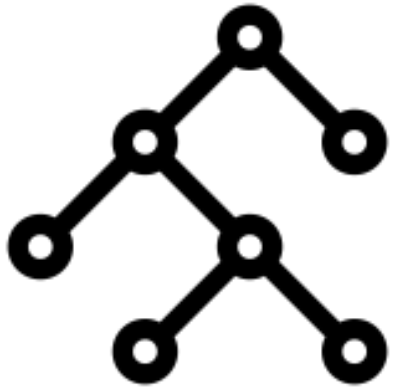
- Models were “trained” on a randomly selected subset of 80% of 569 observations (n=455)
- Trained models would predict the diagnosis, benign or malignant, of the remaining 20% of the observations (n=114)
- Models were iterated on, and final candidates chosen based on accuracy scores (total correct predictions / total predictions)
- Leading models of each tested algorithm compared using several standard performance metrics, including accuracy

Models

Data tested with three popular machine learning algorithms:

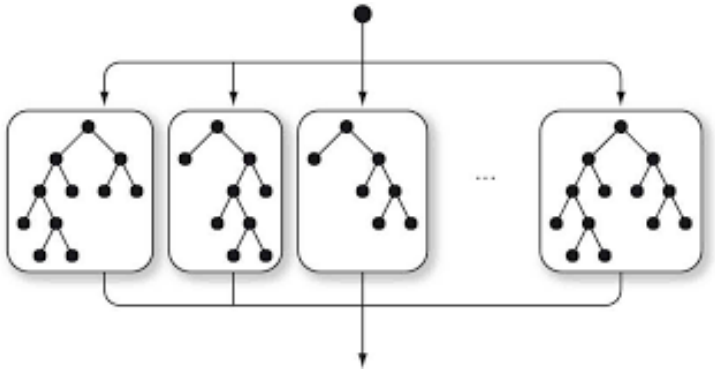
Decision tree

Hierarchical branching decision-making structure of one or more independent variable nodes that terminate in predictive "leaves"



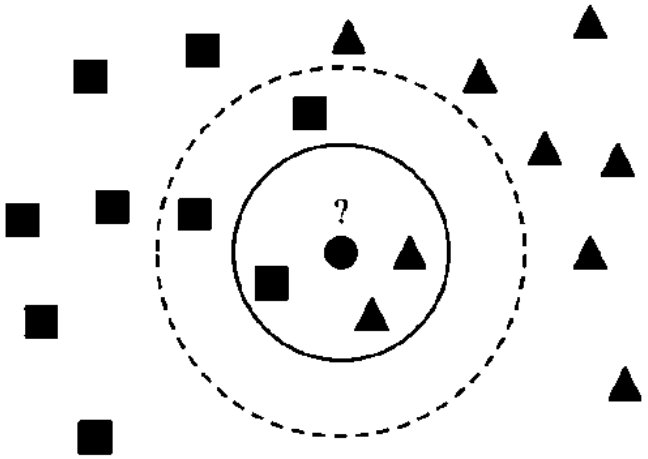
Random forest

"Wisdom of crowds" model that averages the individual predictions of an ensemble of decision trees



K-nearest neighbors

Grouping of data into clusters based on geometrical closeness, used to predict in which clusters new observations belong



Results

Models had comparable accuracy scores, though notable difference in avoiding false positives/negatives

Decision Tree Accuracy = 94.7% (1 variable, "radius_worst")		
	Predicted negative	Predicted positive
Actual negative	71	0
Actual positive	6	37

Random Forest Accuracy = 93.9% (8 variables, 200 trees)		
	Predicted negative	Predicted positive
Actual negative	70	1
Actual positive	6	37

K-Nearest Neighbors Accuracy = 93.9% (8 variables, k=16)		
	Predicted negative	Predicted positive
Actual negative	69	5
Actual positive	2	38

Metrics

Decision tree and random forest out-perform K-nearest neighbors on precision and specificity, under-perform on recall

Metric	Decision Tree	Random Forest	K-Nearest Neighbors
Accuracy (TP+TN/TP+TN+FP+FN)	0.9473684	0.9462863	0.9385965
Precision (TP/TP+FP)	1.0000000	0.9743590	0.8837209
Recall (TP/FN+TP)	0.8604651	0.8813376	**0.9500000**
Specificity (TN/FP+TN)	1.0000000	0.9857200	0.9324324
F1 Score (2 * (precision * recall)/ (precision + recall))	0.9250000	0.9255169	0.9156627

Conclusion

- While all three models had comparable overall predictive success based on accuracy and F1 scores, the K-nearest neighbors appeared better at avoiding false negatives (i.e., notably better recall score).
- K-nearest neighbors is recommended, on grounds that a higher false positive rate is preferable to a higher false negative: a patient is more likely to seek a second opinion for a false positive, whereas minimizing false negatives helps ensure patients with actually malignant tumors seek treatment as early as possible

Next steps

- Iterate on models with different hyperparameters, variable sets, train/test subsets
- Verify model performance, explore potential to tune for false negative avoidance,
- Test against additional data, as available