Breast Cancer Type Classification

...

Outline

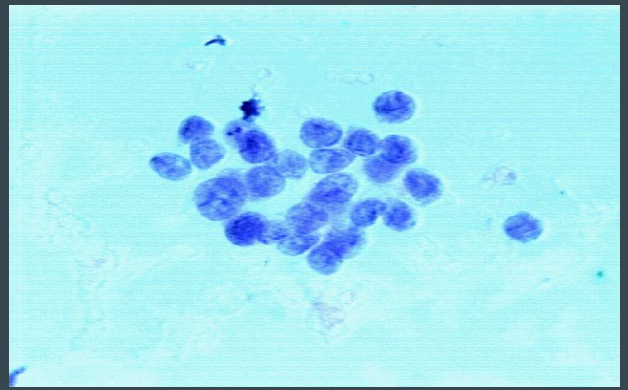
- Motivation and background information
- Data set specifications
- Machine Learning models used
- Evaluation of the model
- Results

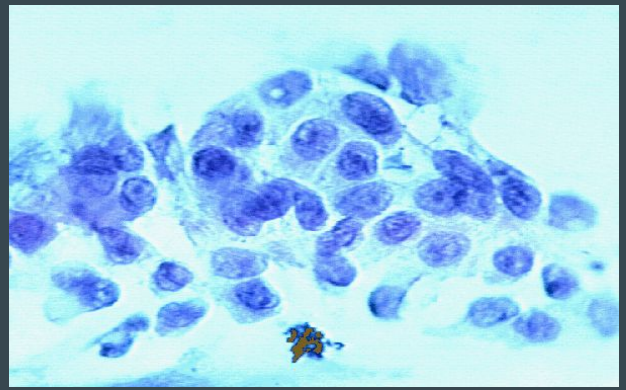
Motivation and Background

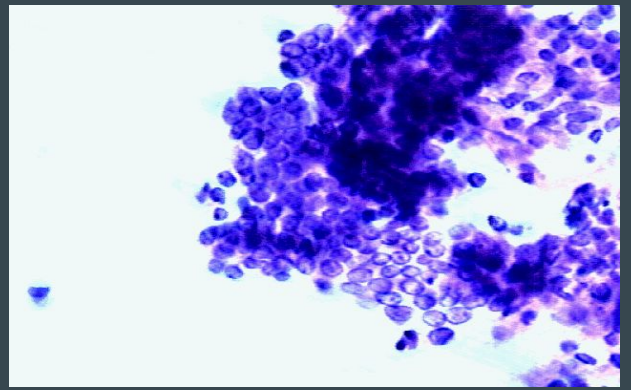
The main goal of the project was to accurately determine whether the breast cancer is malignant or benign using a Fine Needle Aspiration procedure(FNA).

The taken sample is tinted and put under a microscope to take images that are then processed.

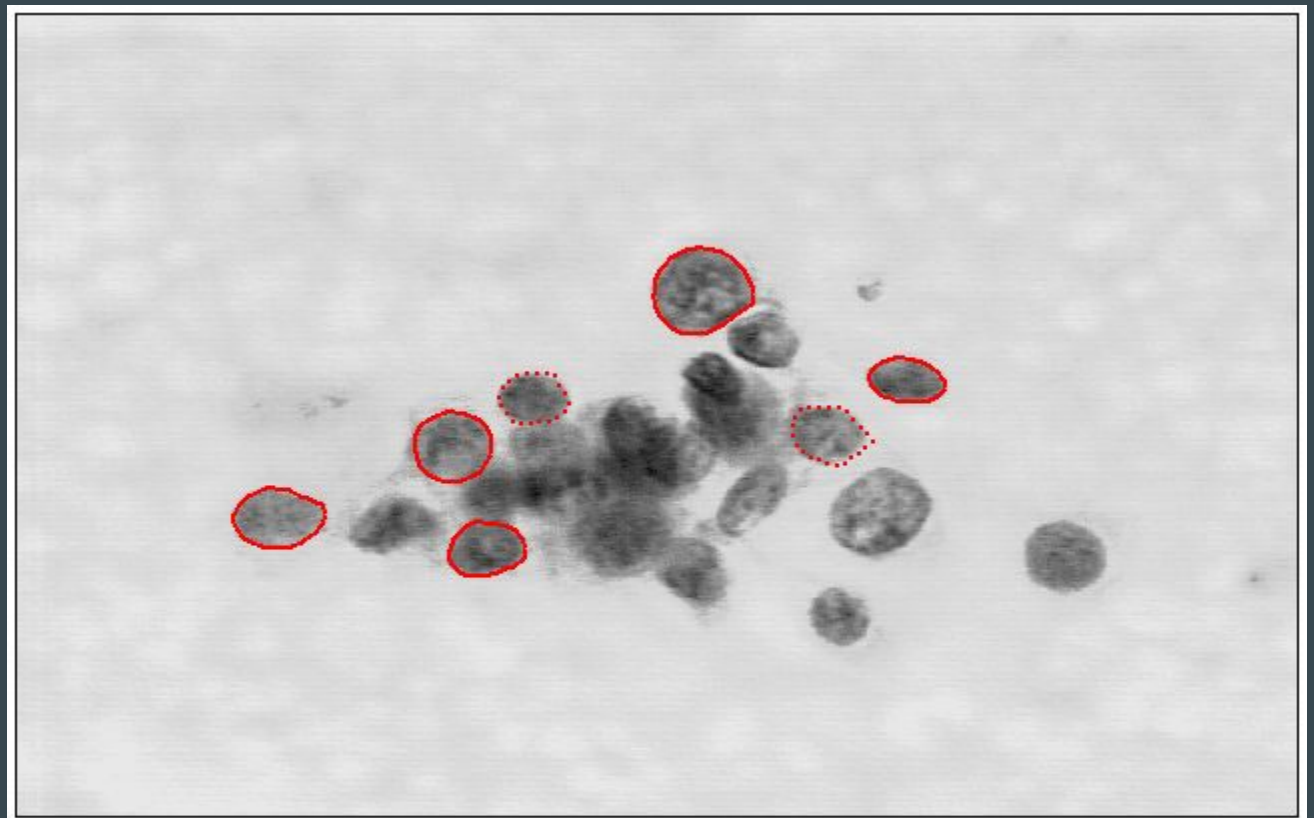
Characteristics such as size, shape and texture are measured from the images.







Sample Analysis



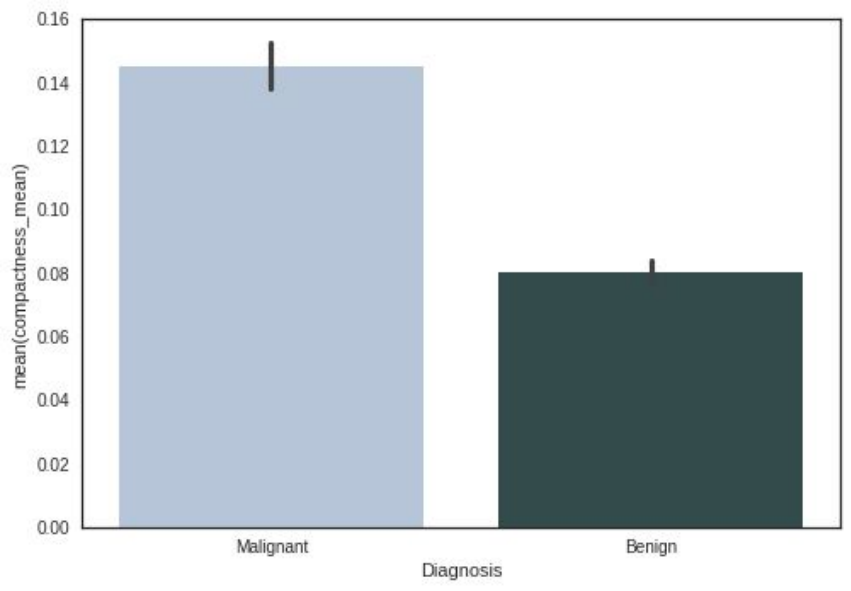
Data Set

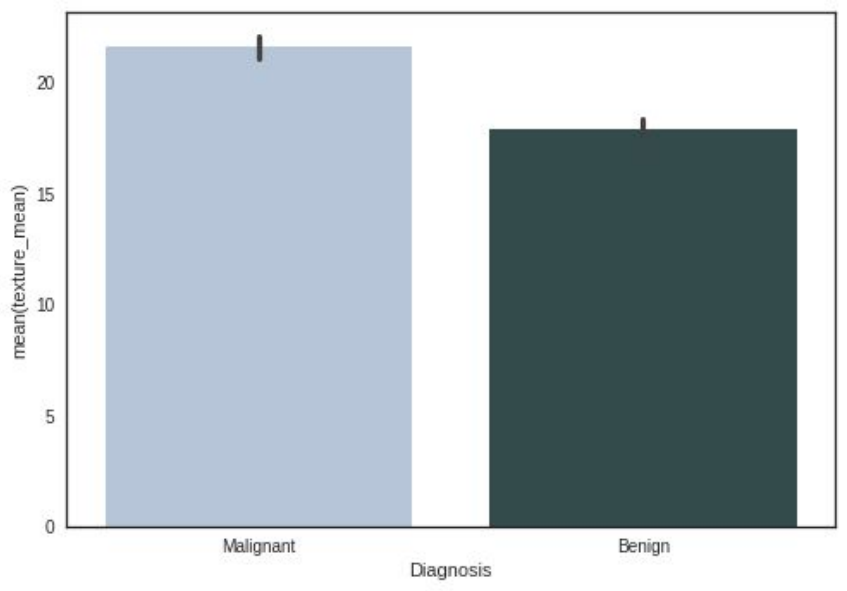
- 569 data samples
  - 357 Benign
  - 212 Malignant
- 30 features and target
- Splitting the data set:
  - 70 % Training
  - 15 % Validation
  - 15 % Test

Sample of the Data Set

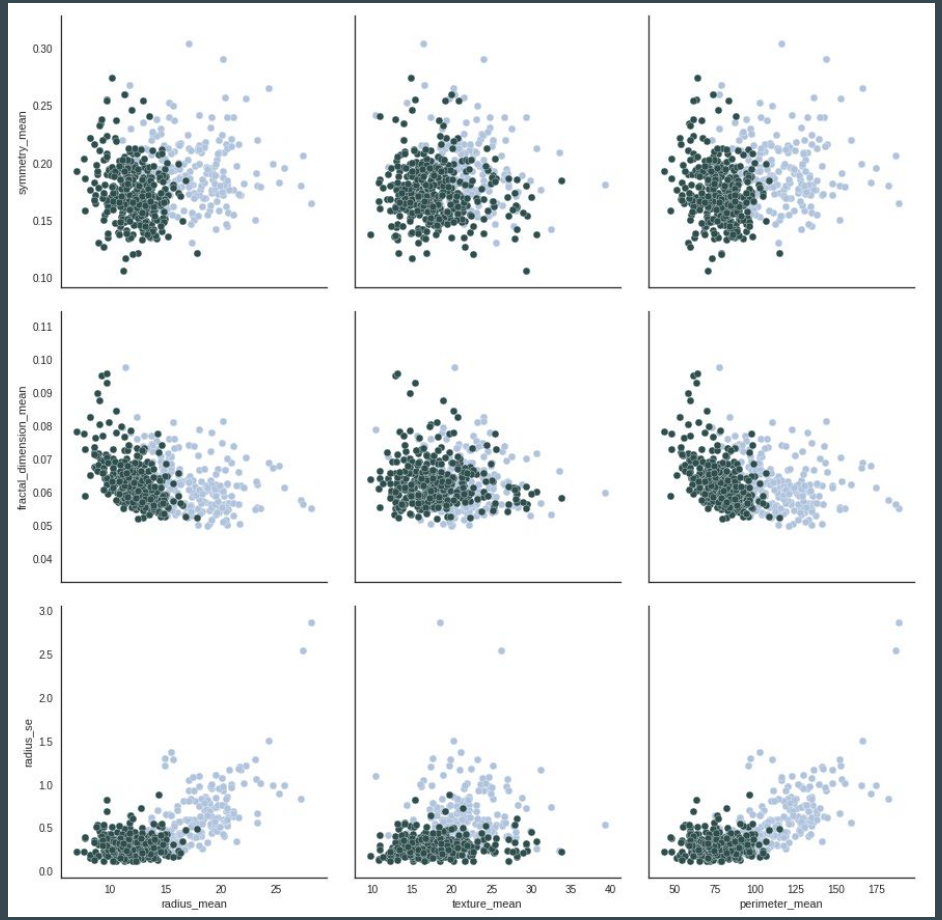
diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean
M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419
M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812
M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069

Data Analysis

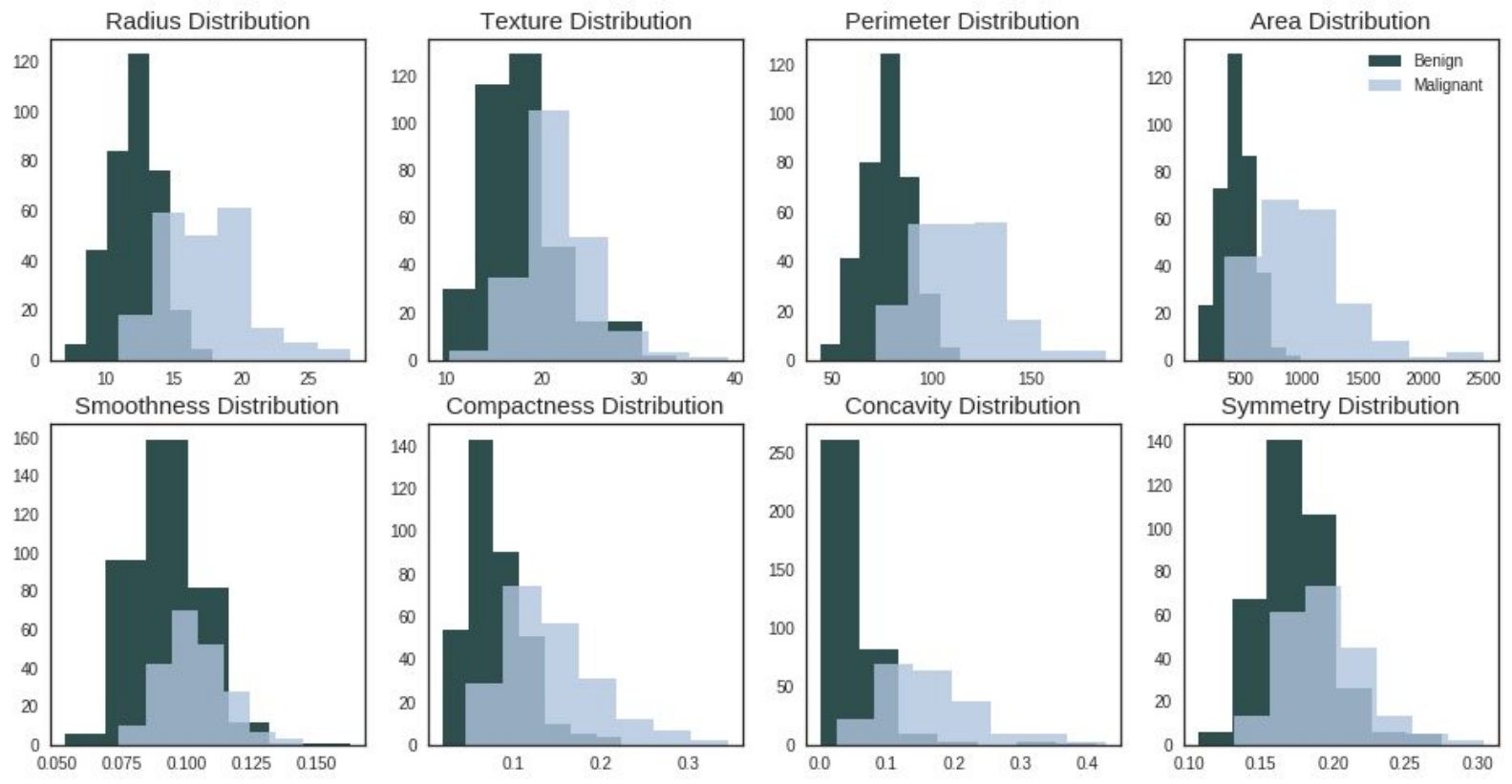




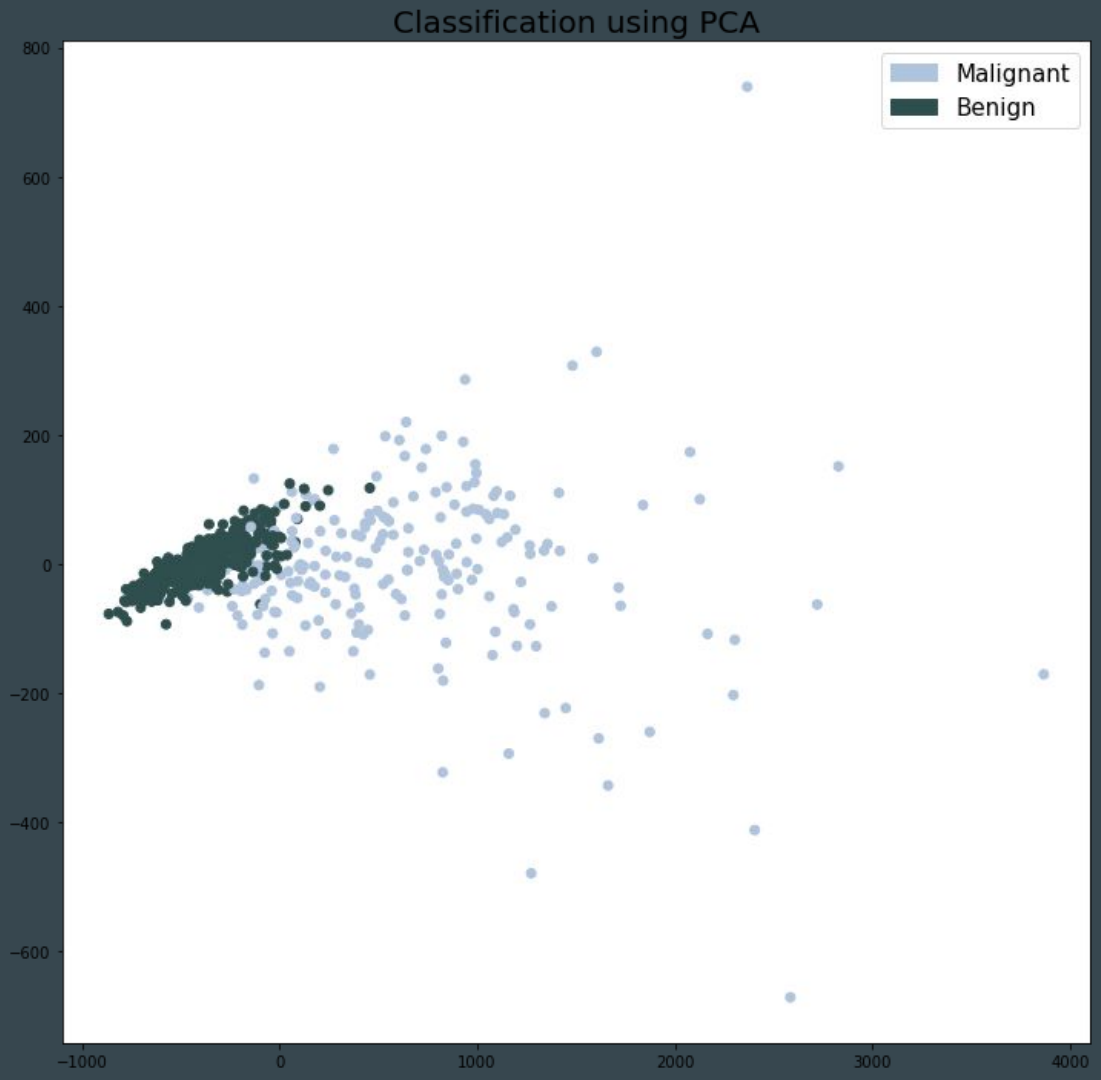
Correlation Pair Plots



Distribution plots



Clustering using PCA



Machine Learning: Classification

Algorithm	Validation f-1
Logistic Regression	**97%**
Support Vector Machine	94%
Naive Bayes	93%
K-Nearest Neighbors	95%
Decision Tree	90%

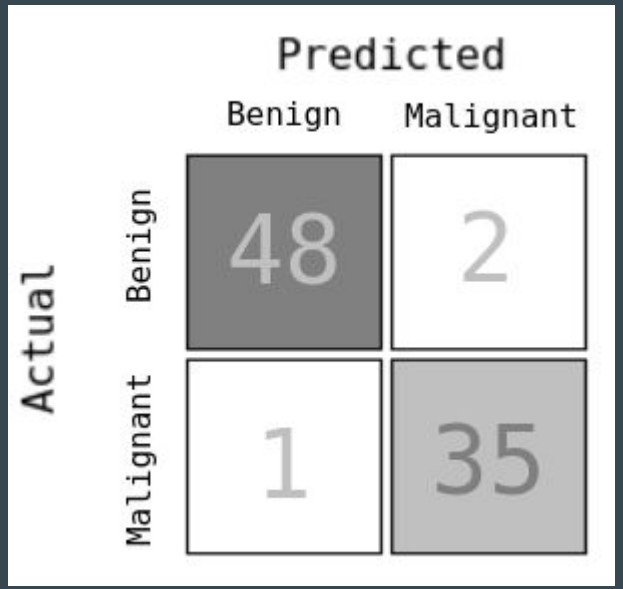
Model selection was done over the validation set using confusion matrix and precision recall.

Logistic Regression

- The dependant variable needs to be binary (Benign or Malignant)
- Estimates the probability of of a binary response based on one or more independent variables
- Measures each independent variable's partial contribution to variations in the dependent variable
- The goal is to correctly predict the category of the outcome
- Logistic regression is like linear regression, but it's predicting probabilities between 0 and 1 instead of numbers.

Model Evaluation and Results

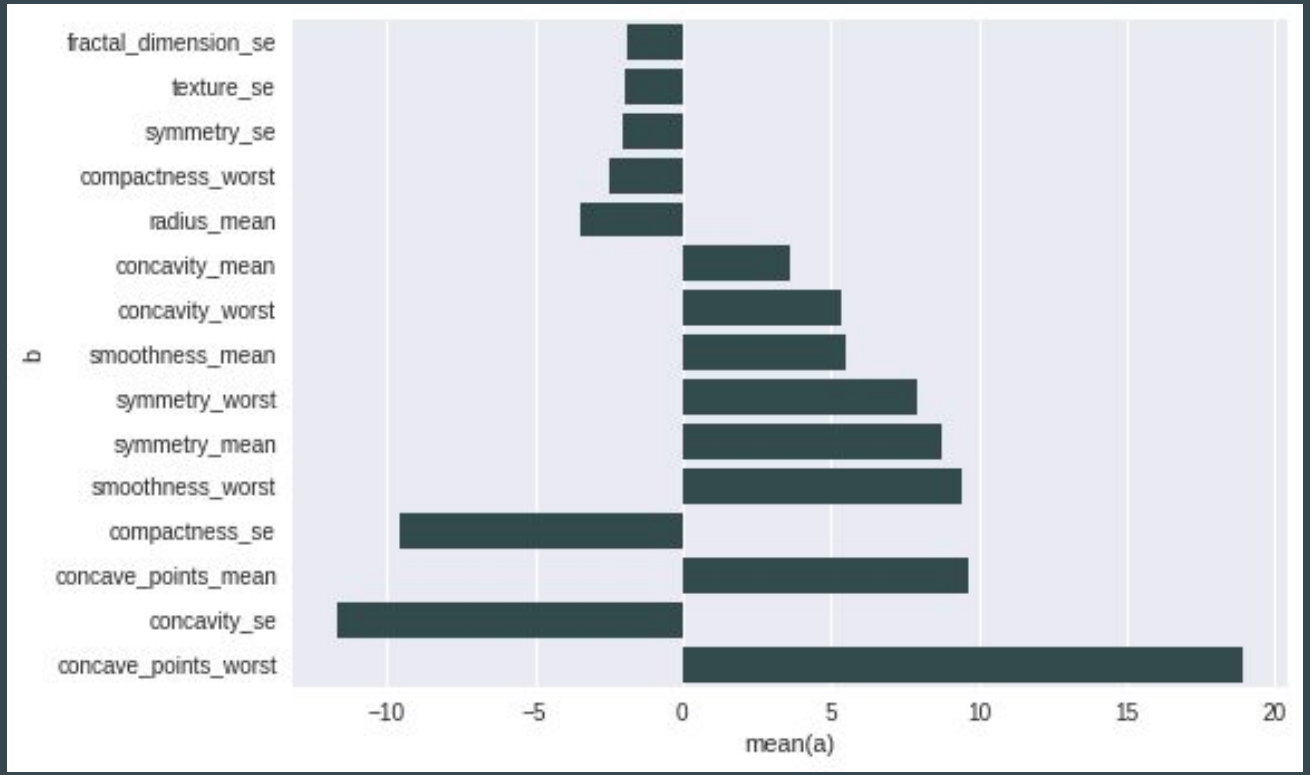
Confusion Matrix



Classification Report

	precision	recall	f1-score	support
B	0.98	0.96	0.97	50
M	0.95	0.97	0.96	36
avg / total	0.97	0.97	0.97	86

Model Evaluation and Results



Conclusion

- ML models can be used to classify inconclusive FNA test results with high precision
- 5 Classification ML models were applied
- The best results: Logistic Regression and K - Nearest Neighbors
- All of the models above 90% accurate
- Computer are better at classifying cancer type where humans are unable to distinguish between malignant or benign.