Comparison of three machine learning predictors on diagnostic cancer data

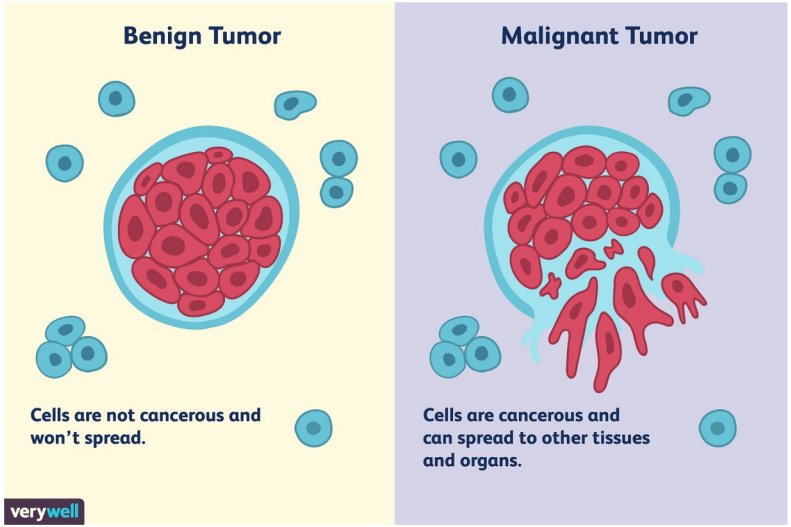
By Thomas Griffin, Mehak Kapoor, and Onyeka Onyenemezu

Talk Overview

- Data Set
- Problem Data
- Three Machine Learning Predictor- hyperparameters, visual graphs, accuracy results.
  - Decision Tree
  - Logistic Regression
  - KNN
- Comparison of Accuracy

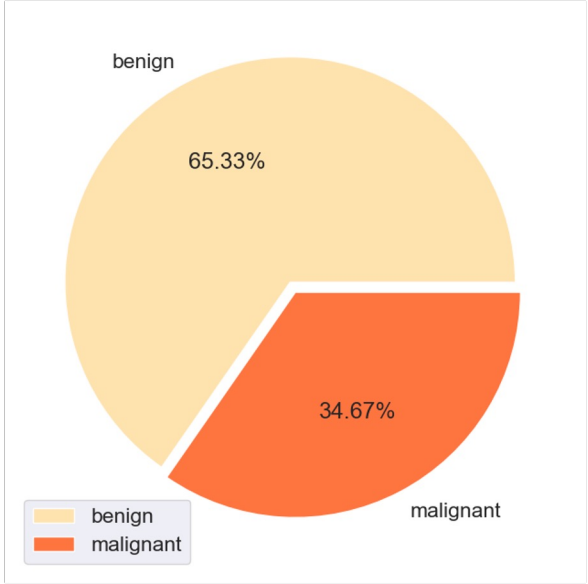
Introduction- Breast cancer

- Most common cause of death in women worldwide.
- Early detection is the most effective way to reduce the breast cancer deaths.
- Benign and Malignant tumors
- Nine real-valued features are computed for each cell nucleus:
  - Clump Thickness
  - Uniformity of Cell Size
  - Uniformity of Cell Shape
  - Marginal Adhesion
  - Single Epithelial
  - Bare Nuclei
  - Bland Chromatin
  - Normal Nucleoli
  - Mitosis

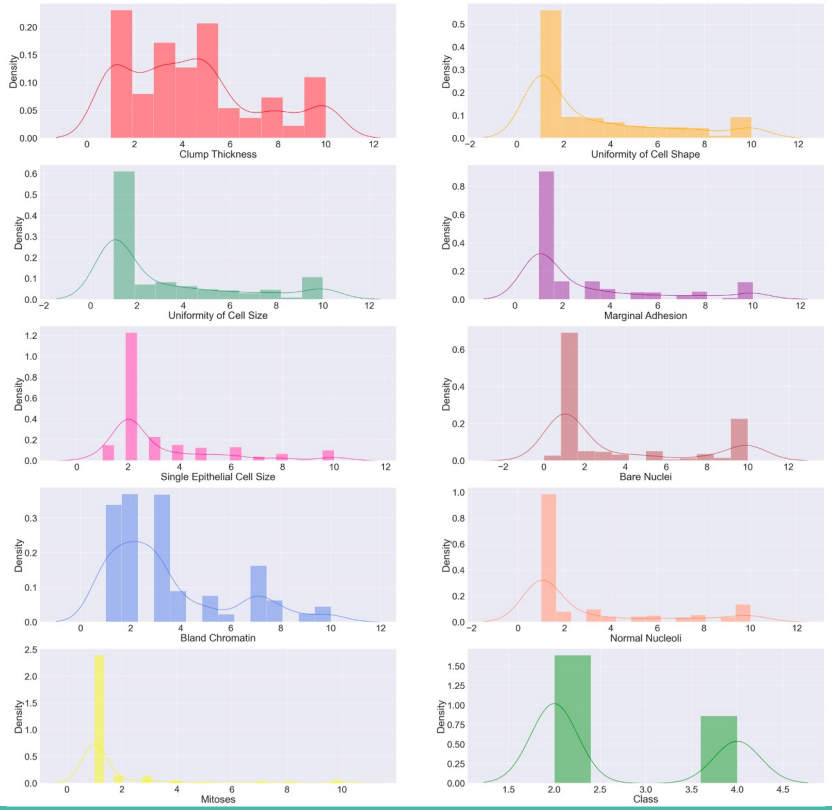


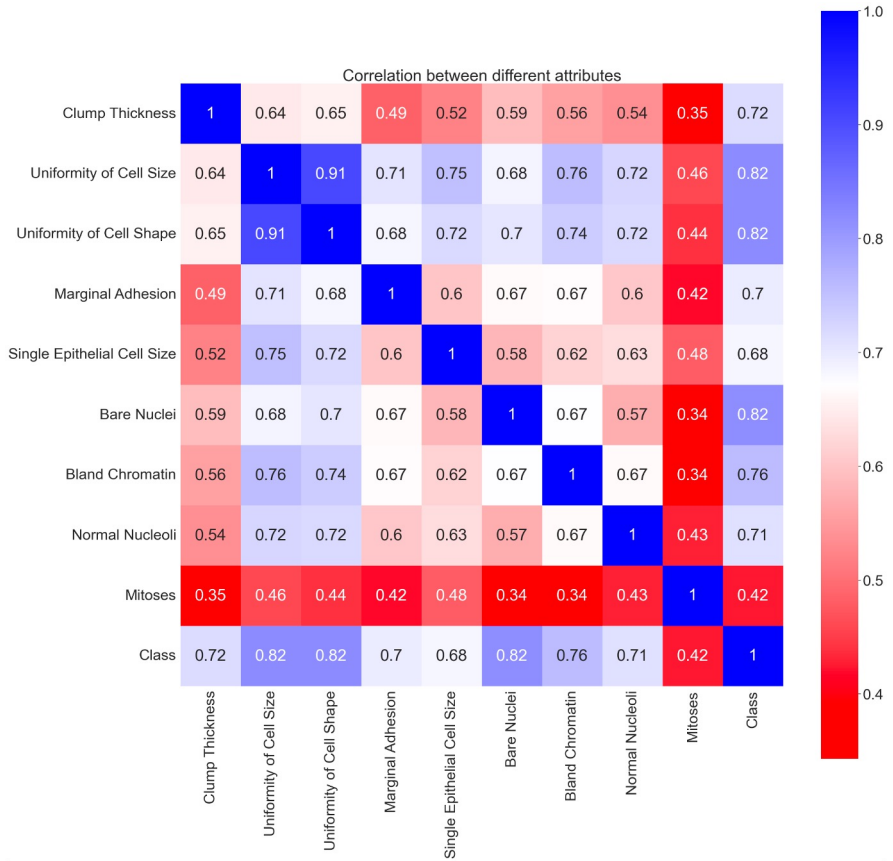
More about the data!

- 699 instances
- 9 features + 1 class
- Missing values - 16 denoted by '?'
- Distribution of class:
  - Benign - 458
  - Malignant - 241



Data Visualization





Handling missing data

- Method of corrections
  - Average dataset : Filled in the missing data points with the average of the column
  - Removed column : All missing data were in the Bare nuclei column so we removed this column
  - Removed Row : All rows with missing data were deleted

Data Split

- Average Dataset
  - Train = 499 samples
  - Test = 100 samples
  - Validation = 100 samples
- Removed Column
  - Train = 499 samples
  - Test = 100 samples
  - Validation = 100 samples
- Removed Row
  - Train = 489 samples
  - Test = 97 samples
  - Validation = 97 samples

For all three predictors we used the same splitting of the data.

Decision Tree - Hyperparameter Tuning

We chose to use depth of 4 as higher depth was overfitting our training data. Dataset was “small” enough to not worry about node sizes.

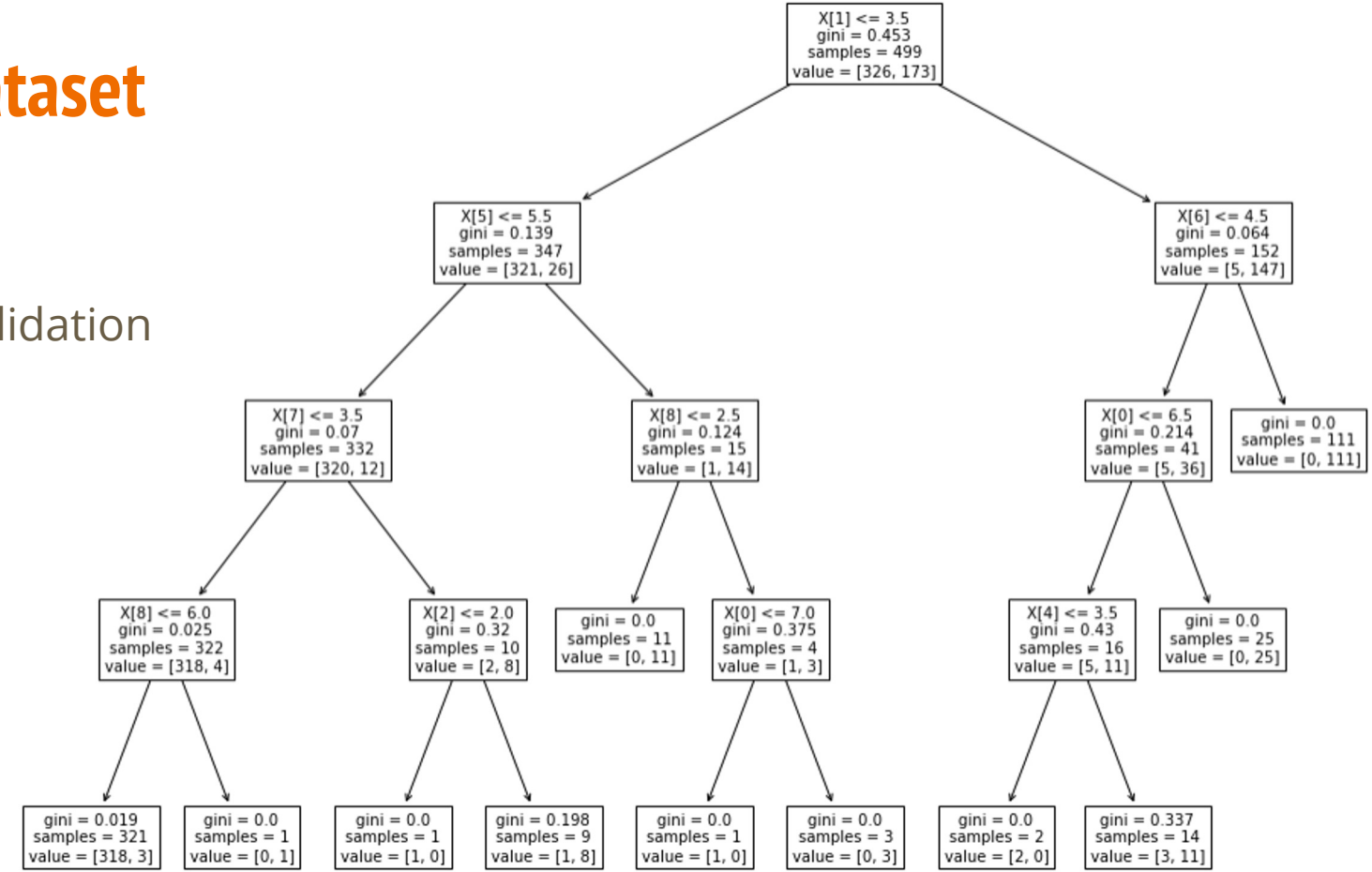
Depth of Tree	Accuracy of Test Data (100 points)
1	.87
2	.89
3	.90
4	.91
5	.9
6	.89
7	.89

Average Dataset

Depth=4

Accuracy on Validation

Data set = 92%

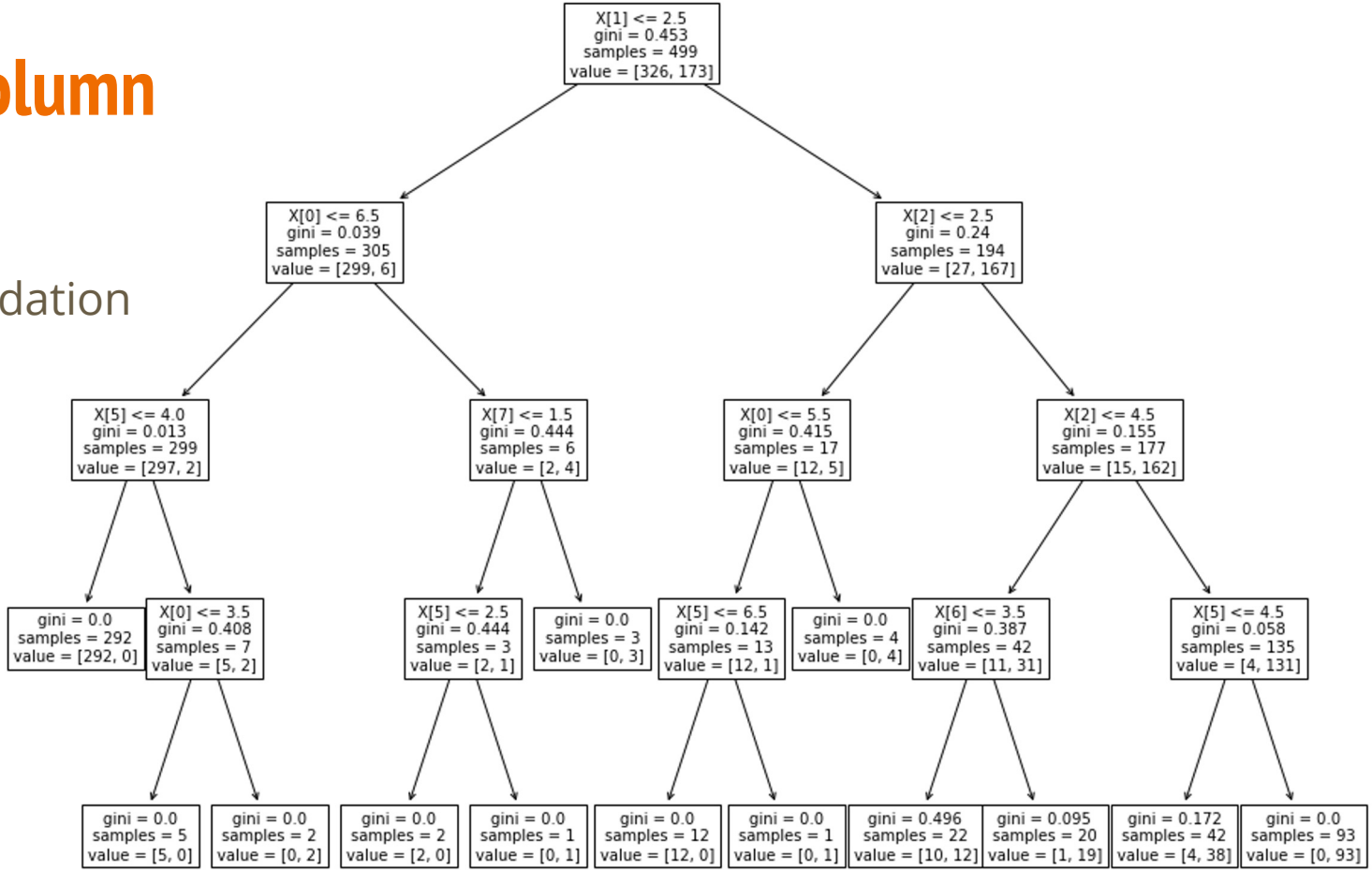


Removed Column

Depth=4

Accuracy on Validation

Data set = 94%



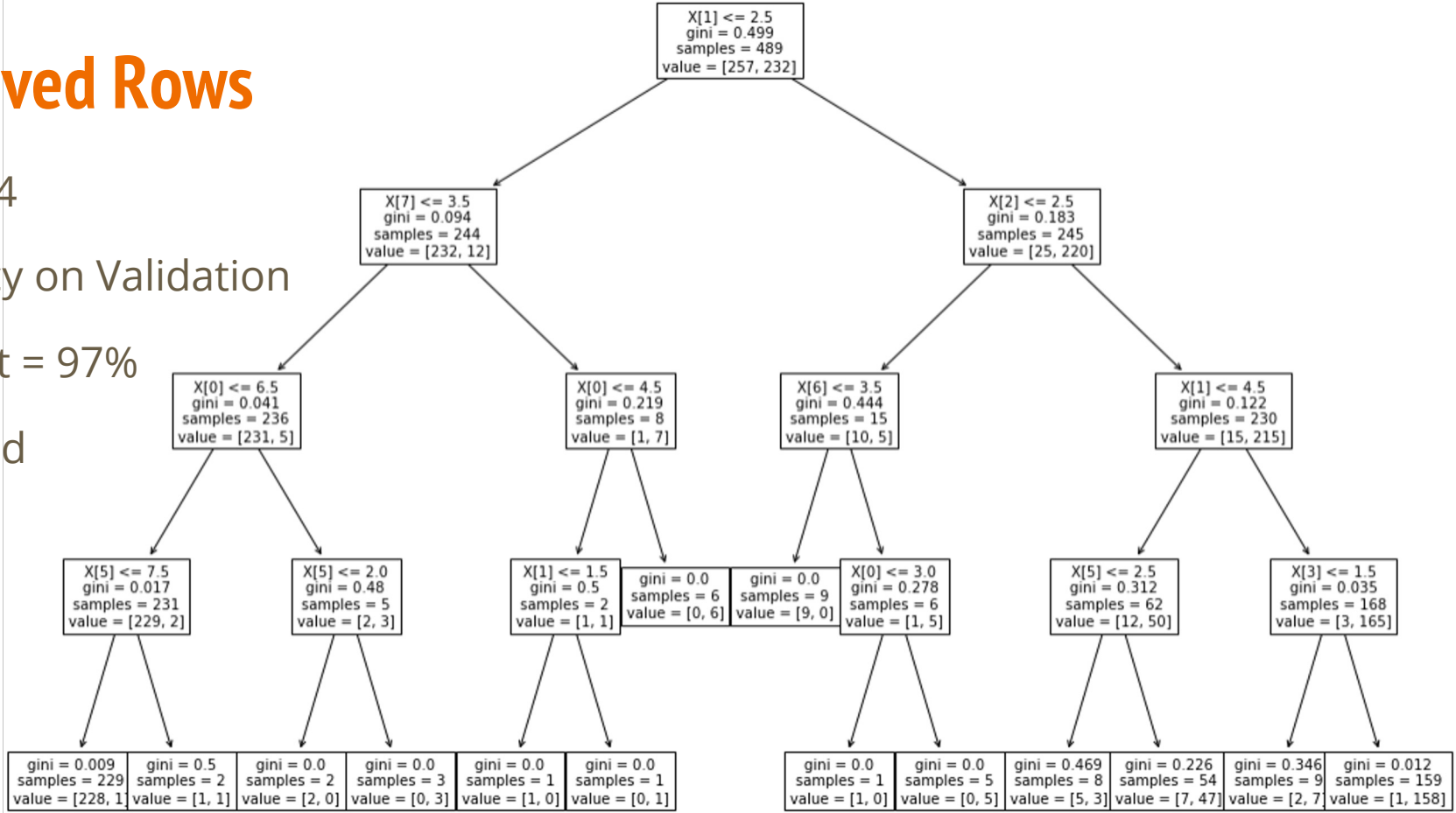
Removed Rows

Depth=4

Accuracy on Validation

Data set = 97%

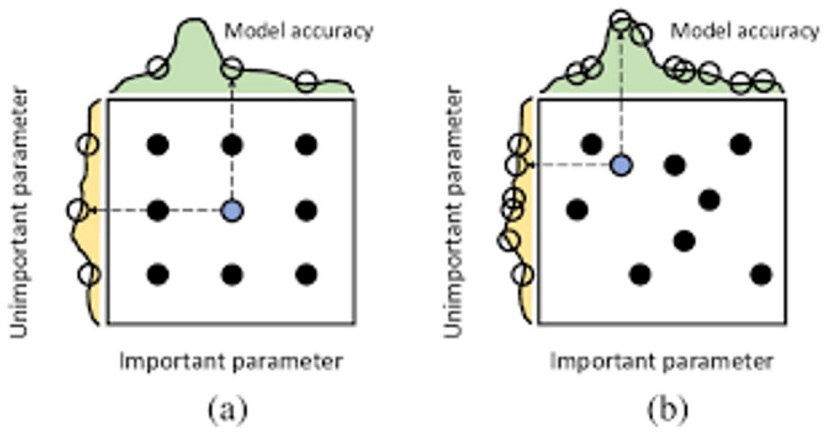
Rounded



Logistic Regression - Hyperparameter Tuning

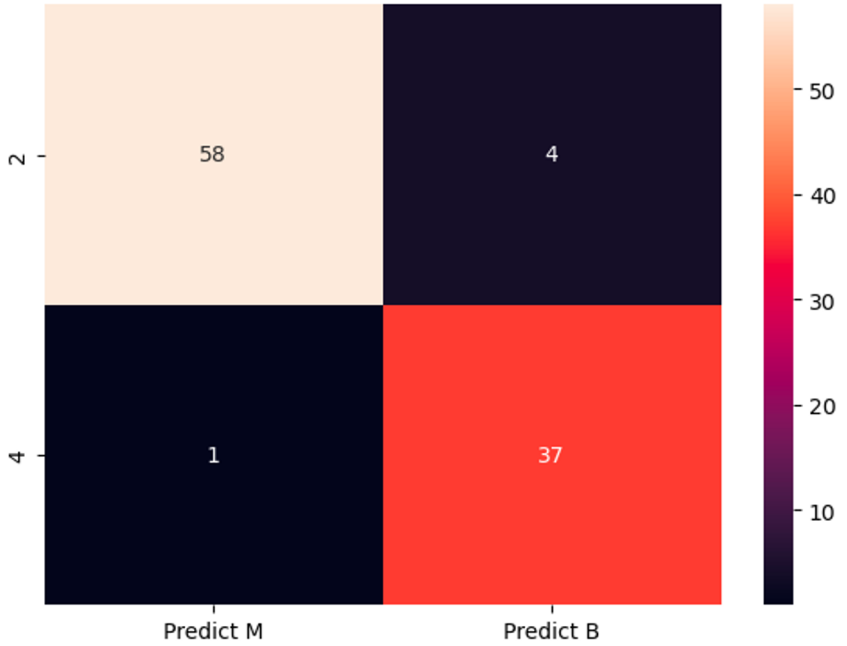
- Penalty
  - L2
  - L1
- C : regularization strength
  - 0.000001
  - 0.001
  - 0.01
  - 0.1
  - 1
  - 10
  - 100
  - 1000

GridsearchCV



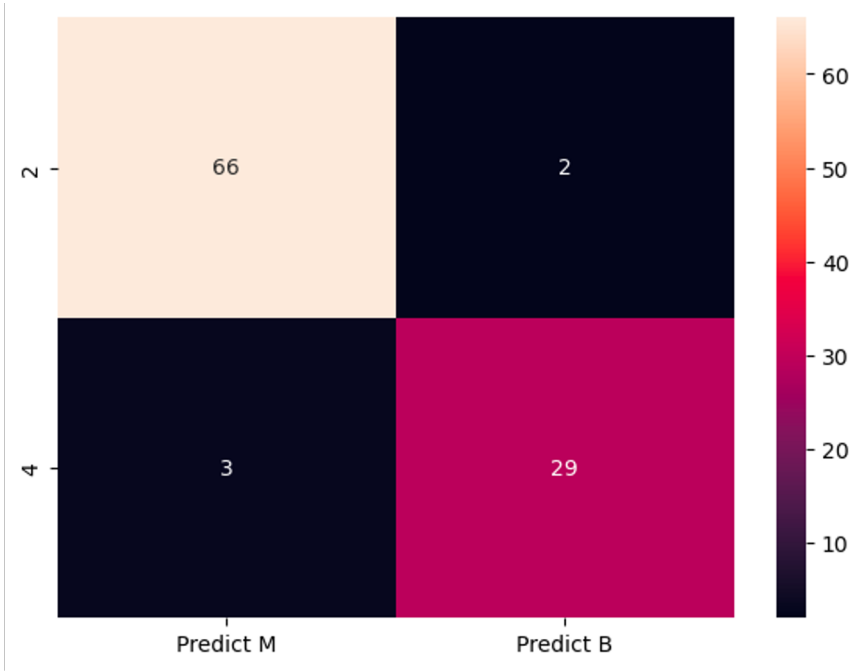
Average Dataset

- Accuracy : 95%
- Best parameters
  - C= 10
  - l2 penalty



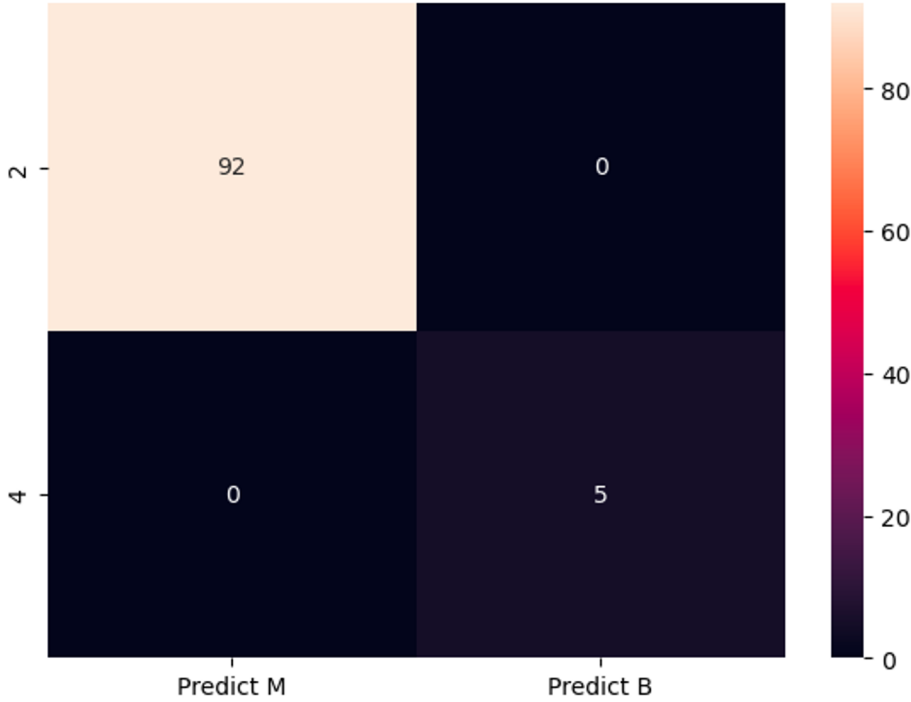
Removed Column

- Accuracy : 95%
- Best parameters:
  - C= 0.1
  - l2 penalty



Removed Rows

- Accuracy : 100%
- Best parameters
  - C= 0.01
  - l2 penalty



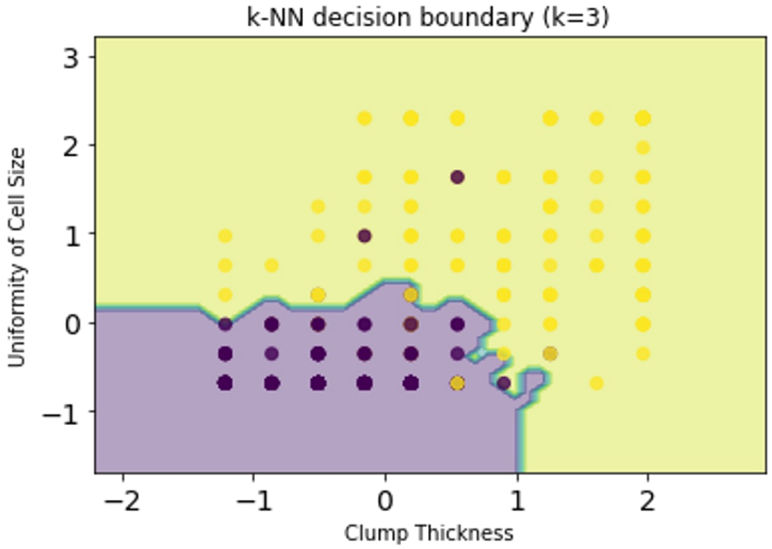
Hyperparameter tuning - KNN

Using GridsearchCV found the best combination of parameters for :

- n_neighbors
  - Range 1 to 21
- weights
  - Uniform
  - Distance
- metric
  - Euclidean
  - Manhattan
  - Minkowski

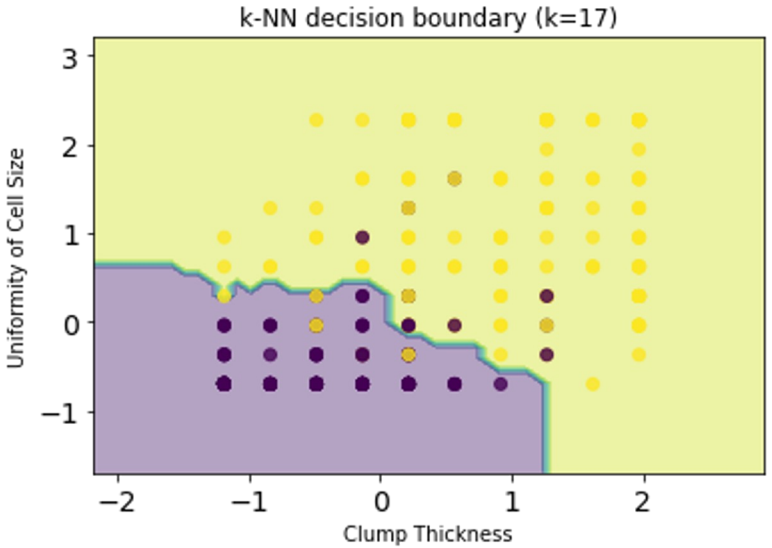
Average Dataset

Accuracy = 96%



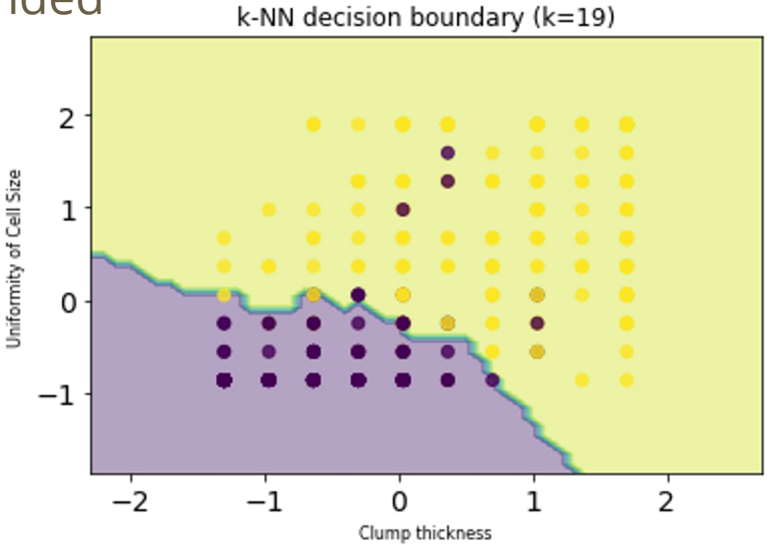
Removed Column

Accuracy = 95%



Removed Rows

Accuracy : 99% Rounded



Conclusion : Accuracy table/matrix

Comparing the results we see a slight increase in accuracy for Logistic Regression and KNN over the Decision Tree models.

	Average	Removed Column	Removed Row
Logistic Regression	95%	95%	100%
Decision Tree	92%	94%	97%
KNN	96%	95%	99%

Possible Future Work

- Testing out different ways to handle the missing values.
  - Median
  - Mode
- Un-Scale the data
  - Currently the categories are scaled 1-10 on different metrics, we could instead use the raw unsorted data.
- Other Predictors
  - Neural Networks
  - Random Forests
  - Gradient Boost

Dataset

Creators:

1. Dr. William H. Wolberg, General Surgery Dept.
University of Wisconsin, Clinical Sciences Center
Madison, WI 53792
wolberg '@' eagle.surgery.wisc.edu

2. W. Nick Street, Computer Sciences Dept.
University of Wisconsin, 1210 West Dayton St., Madison, WI 53706
street '@' cs.wisc.edu 608-262-6619

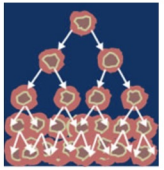
3. Olvi L. Mangasarian, Computer Sciences Dept.
University of Wisconsin, 1210 West Dayton St., Madison, WI 53706
olvi '@' cs.wisc.edu

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Check out the beta version of the new UCI Machine Learning Repository we are currently testing! Contact us if you have any

Breast Cancer Wisconsin (Diagnostic) Data Set

Download: Data Folder, Data Set Description



Abstract: Diagnostic Wisconsin Breast Cancer Database

Data Set Characteristics:	Multivariate	**Number of Instances:**	569	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	32	**Date Donated**	1995-11-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	1976500

Source:

Creators:

1. Dr. William H. Wolberg, General Surgery Dept.
University of Wisconsin, Clinical Sciences Center
Madison, WI 53792
wolberg '@' eagle.surgery.wisc.edu

2. W. Nick Street, Computer Sciences Dept.
University of Wisconsin, 1210 West Dayton St., Madison, WI 53706
street '@' cs.wisc.edu 608-262-6619

3. Olvi L. Mangasarian, Computer Sciences Dept.
University of Wisconsin, 1210 West Dayton St., Madison, WI 53706
olvi '@' cs.wisc.edu


THANKS