Classifying Breast Cancer Tumors: A Comparative Study of K- Means, Logistic Regression, and SVM

Xienyam Chu, Jack Corley, Beatrice Filart, Jung Huh

Motivation

1 in 8 women in the United States will develop breast cancer in her lifetime.

In 2025, about **316,950 women and 2,800 men** will be diagnosed with invasive breast cancer.

When breast cancer is detected early in the localized stage, the 5-year relative survival rate is **99%**.

Source: National Breast Cancer Foundation, Inc.

About the Data

Title: Breast Cancer Wisconsin (Original)

Source: [UCI Machine Learning Repository]

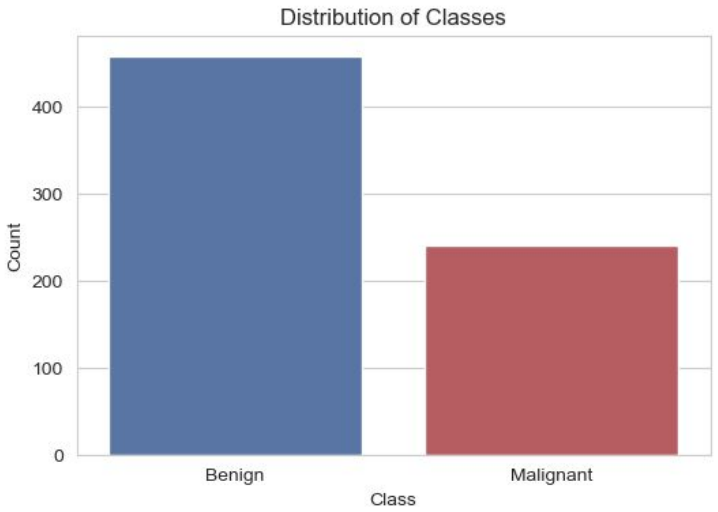
Collected by: Dr. William H. Wolberg, University of Wisconsin Hospitals, Madison

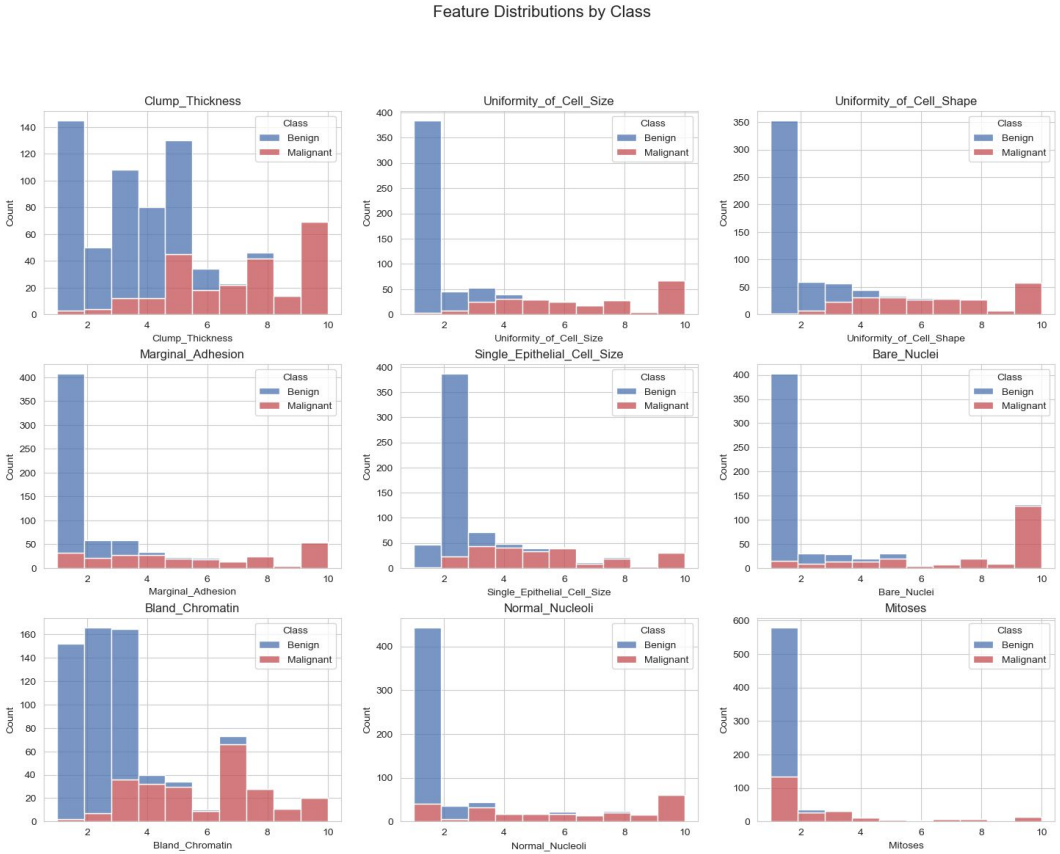
Contains

- 699 patient records
- 10 Numeric Features
- Target Feature: Class
    - Benign vs. Malignant

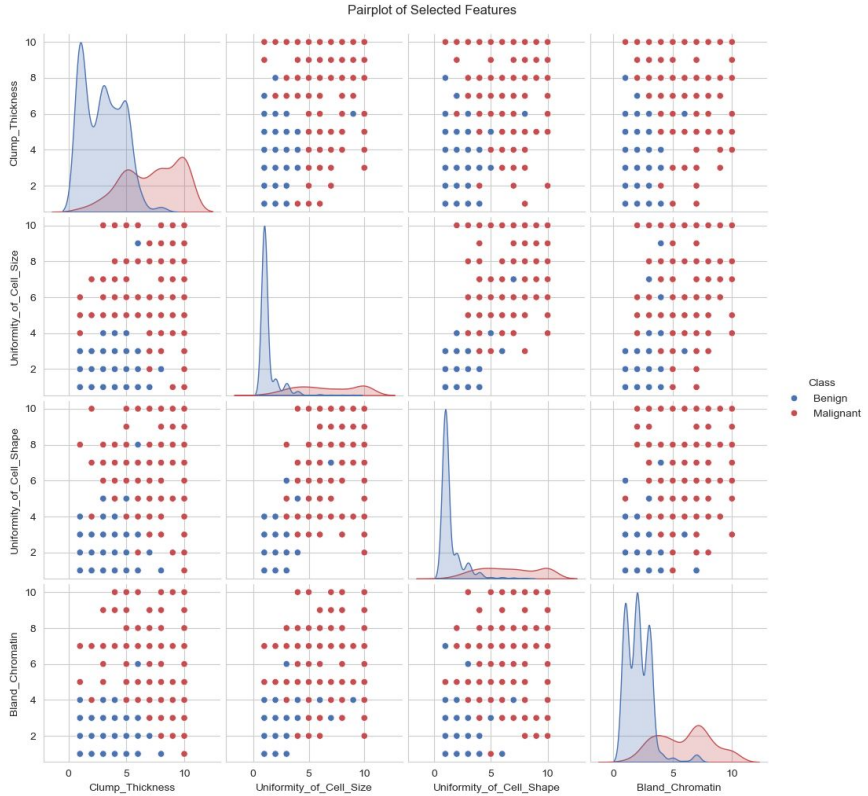
Variable Name	Role	Type
Sample_code_number	ID	Categorical
Clump_thickness	Feature	Integer
Uniformity_of_cell_size	Feature	Integer
Uniformity_of_cell_shape	Feature	Integer
Marginal_adhesion	Feature	Integer
Single_epithelial_cell_size	Feature	Integer
Bare_nuclei	Feature	Integer
Bland_chromatin	Feature	Integer
Normal_nucleoli	Feature	Integer
Mitoses	Feature	Integer

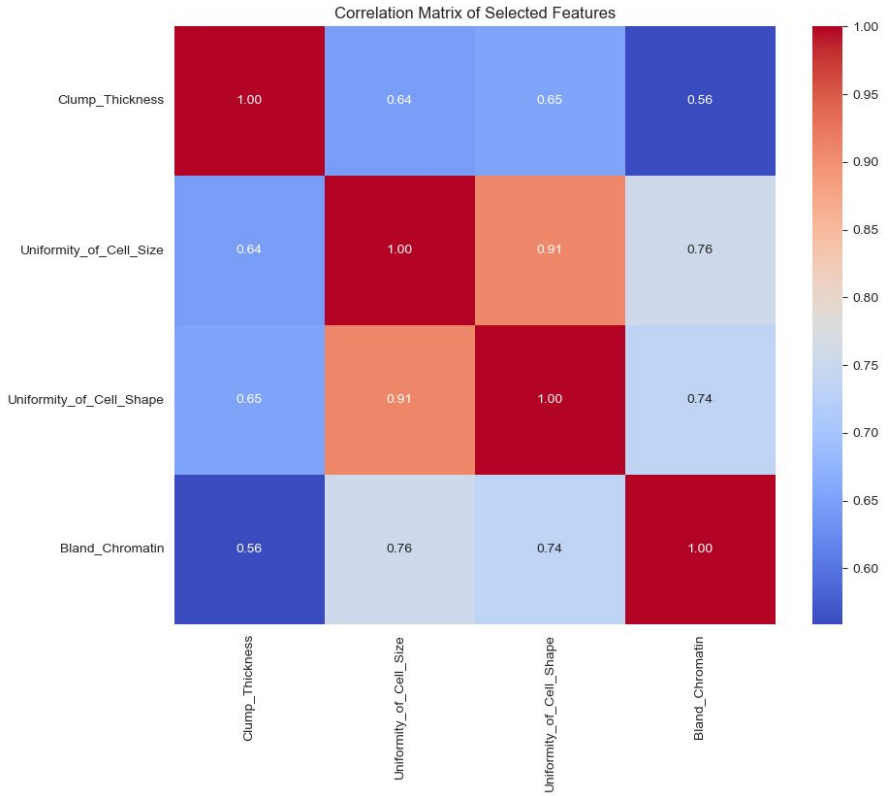
EDA - Class Exploration





EDA - Feature Relationships





Data Preprocessing

- Removed 16 NA values (Bare_Nuclei)
- Split data into train validation test (60-20-20)
- Normalize features
- Converted output variable (Class) to binary (0,1)

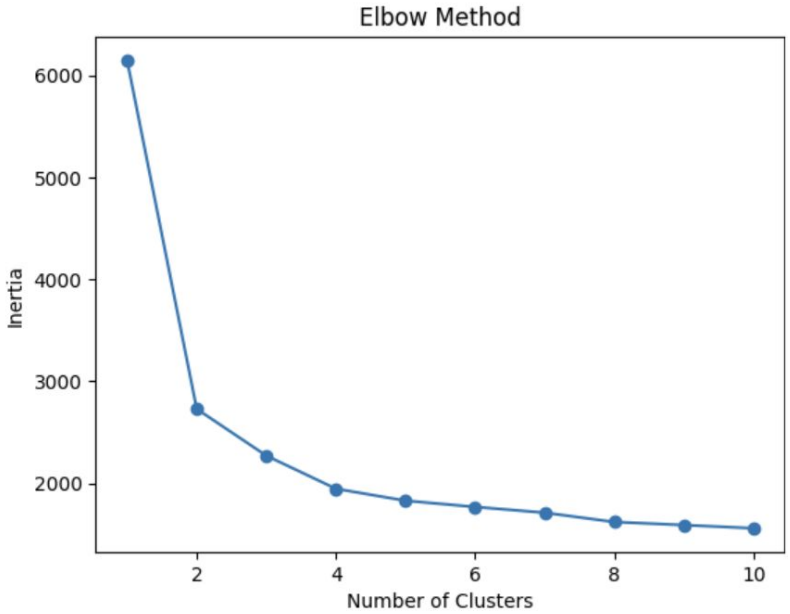
Model 1 - K Means

Key Points to Consider:

Unsupervised Learning

Number of clusters?

- Elbow Curve Method
- Silhouette Analysis
  - Average the S(i), take the K number at maximum score



$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Model 1 - K Means (Cont.)

Hyperparameters:

- Initialization Method: K-means++ (default by Scikit)
- Clustering Method: Elbow Curve Method (ECM)
- Number of initialization: 10 (default by Scikit)
- Max iteration: 300
- Appropriate K number: 2 given by ECM

Silhouette Score: 0.5732450609290859

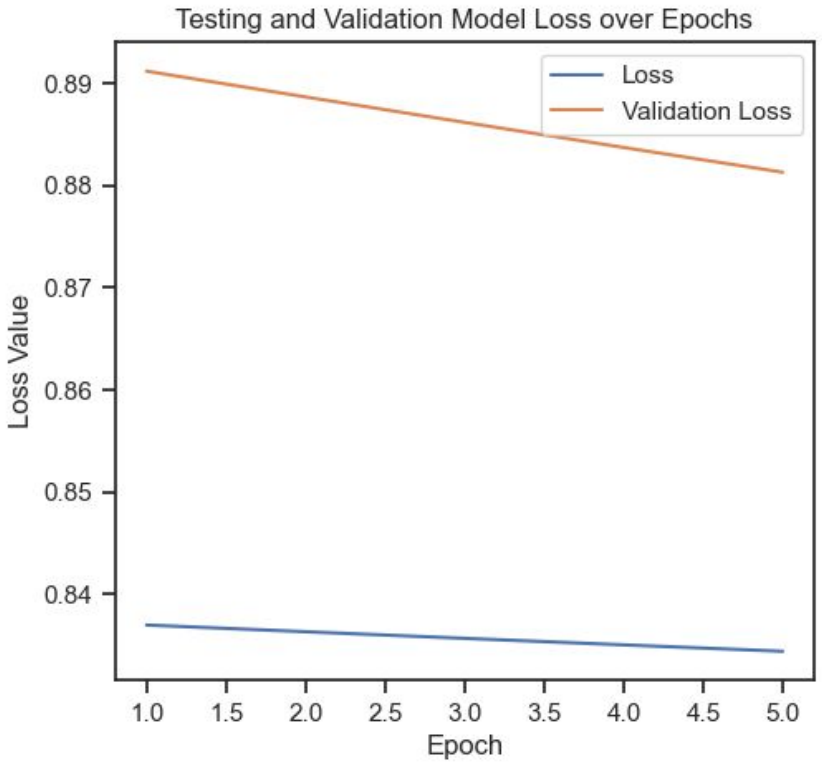
Model 2 - Logistic Regression

Model reasoning:

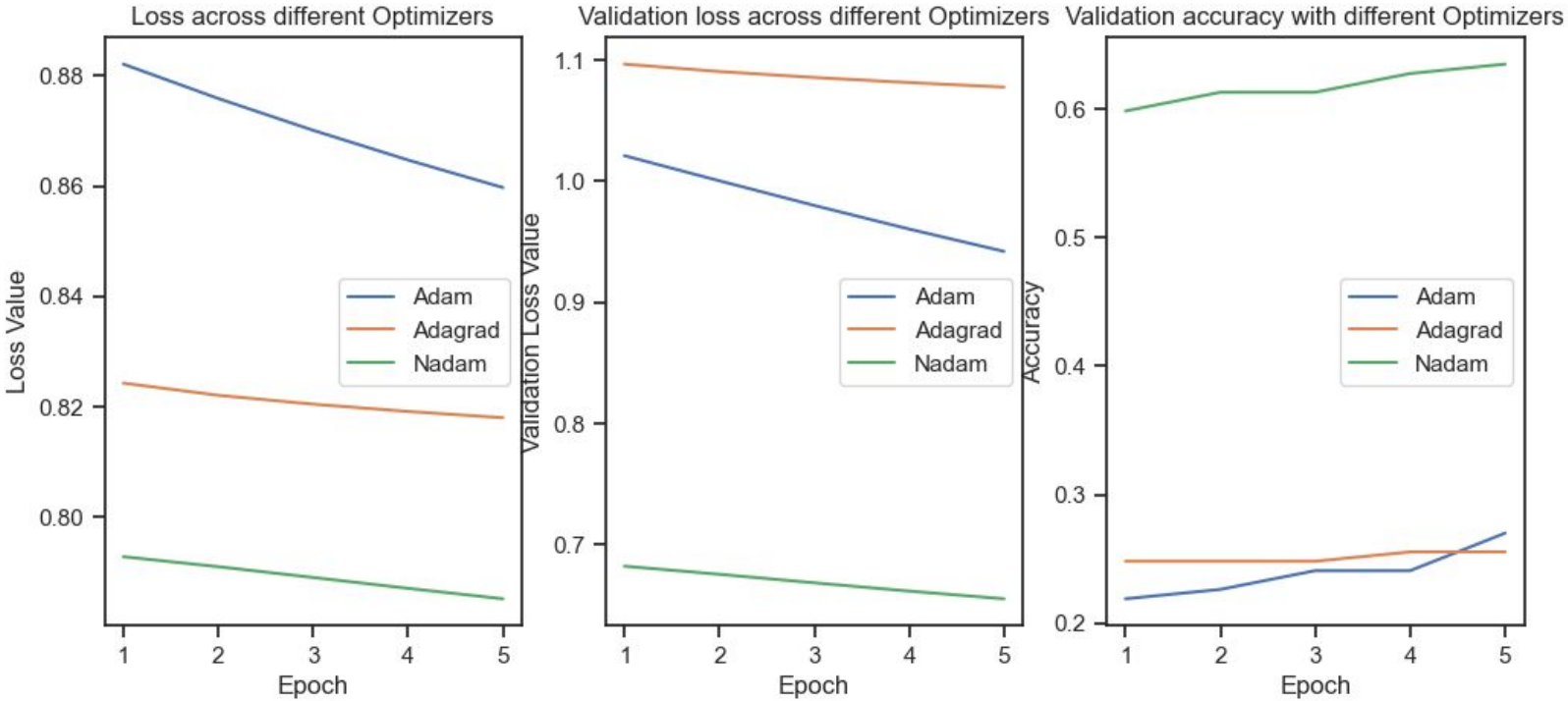
- User understandable

Key considerations:

- Balanced groups
- Supervised learning
  - Risk overfitting with small dataset

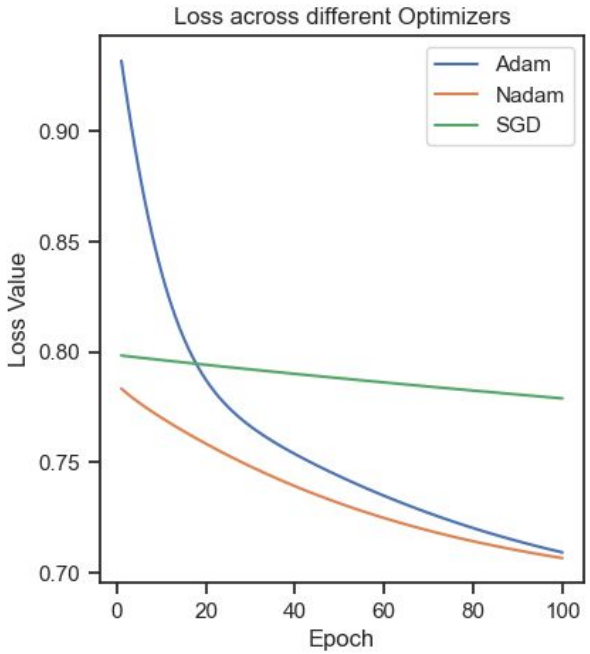


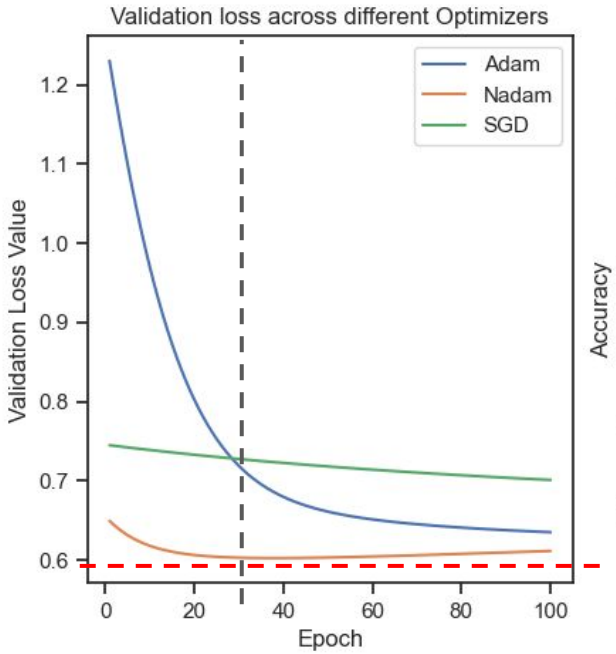
Model 2 - Logistic Regression Experiments

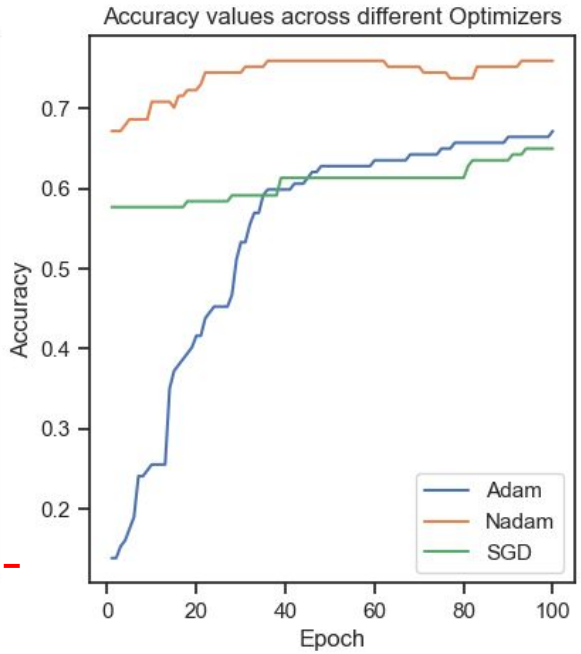


Optimizer parameter tuning

Model 2 - Logistic Regression Experiments (Cont.)



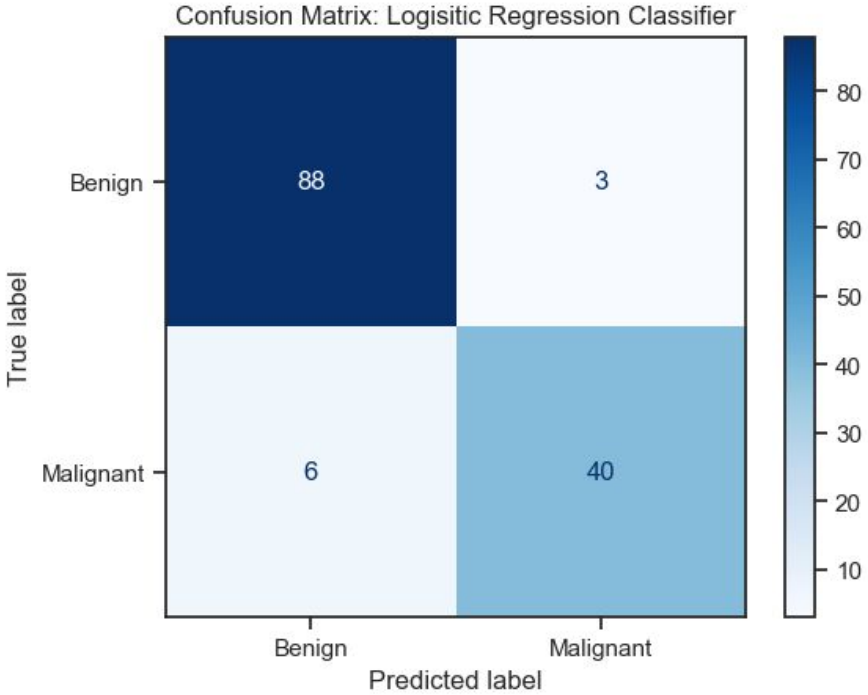




Epoch tuning parameter tuning

Model 2 - Logistic Regression Conclusion

Conclusion: Nadam optimizer with 25 epochs and 0.077 learning rate optimal




```
Accuracy: 0.9343
precision: 0.93
recall:  0.87
```

Model 3 - Support Vector Machines

Data Preprocessing Steps

- Binary Labels Encoded
- Split into train/val/test
- Stratify to maintain class balance
- Scaled numerical measurements post split

Hyperparameter Experimentation

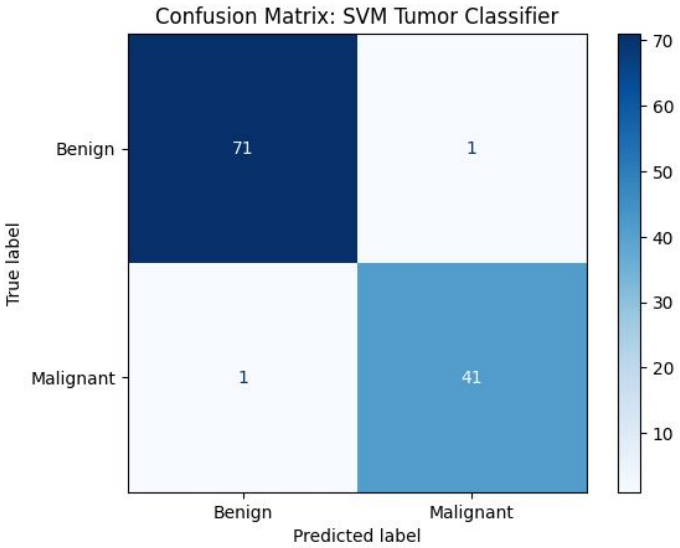
- Kernel Function
- ~~Kernel Coefficient, Degree~~
- ~~Regularization Parameter~~

Post-Tuning

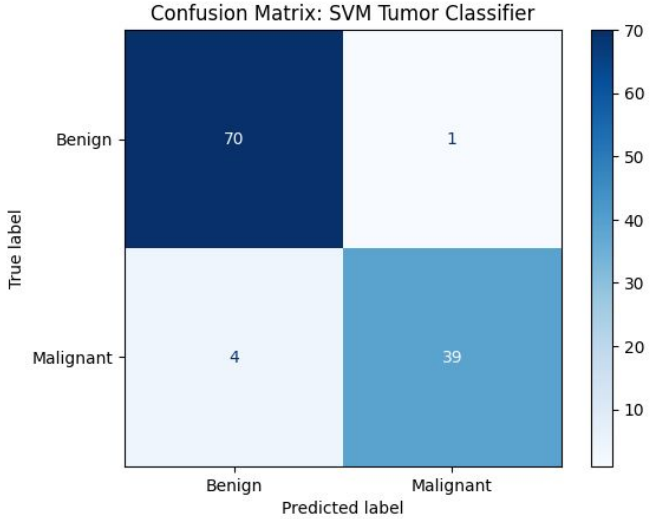
- Trained final model on 80/20 train/test data due to limited data points

Model 3 - SVM Results

Train: 80%
Test: 20%



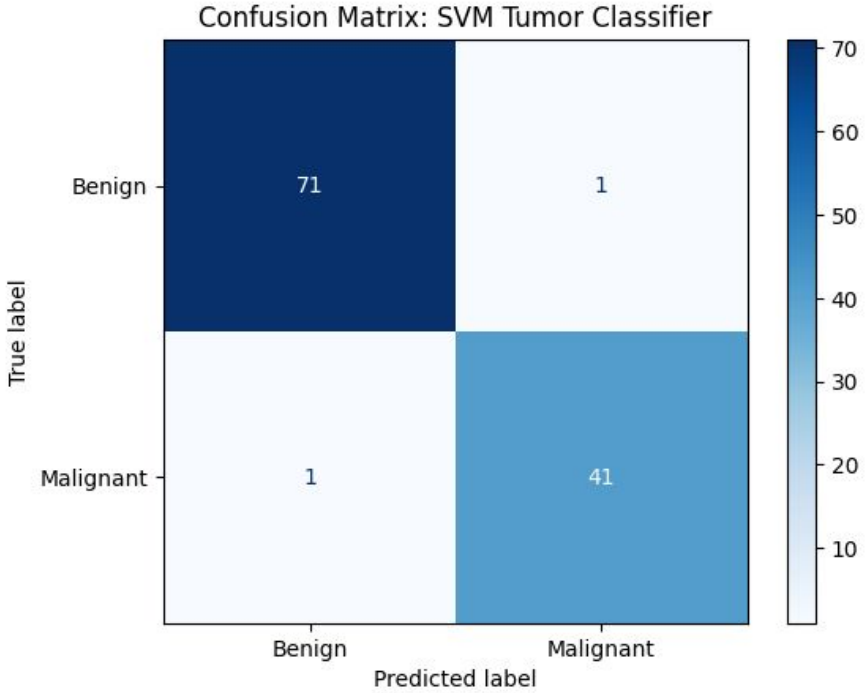
Train 60%
Validation: 20%
Test: 20%



Model 3 - SVM Conclusion

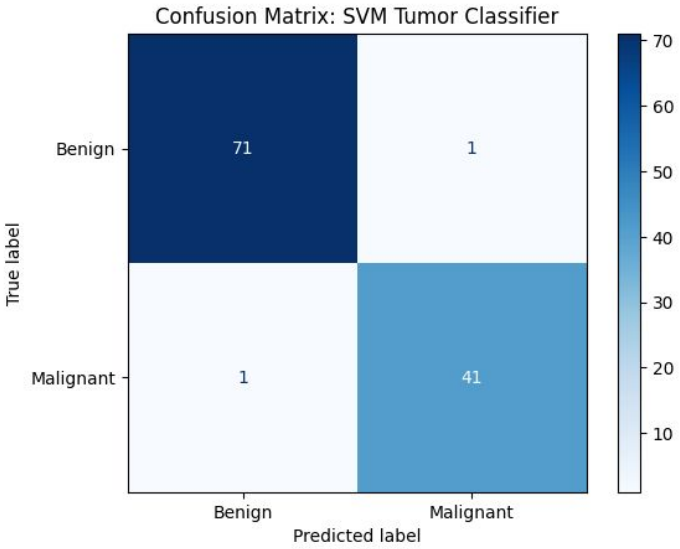
Conclusion: SVM with a Radial Bias Function(RBF) is optimal

```
Accuracy: 0.9824561403508771
precision: 0.98
recall:  0.98
F1-score: 0.98
```



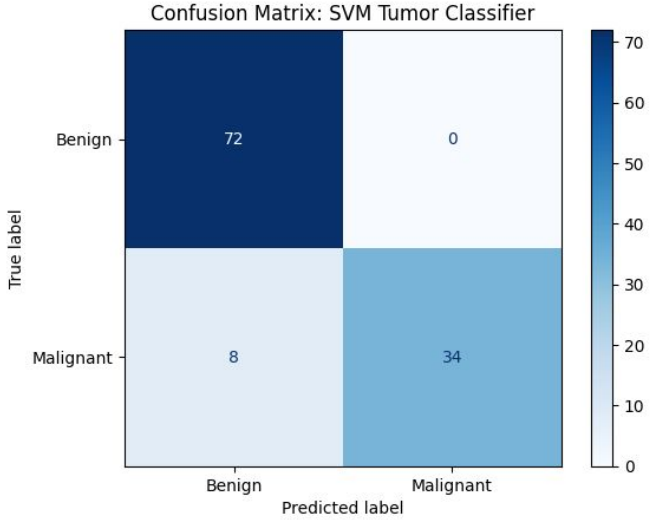
Model 3 - SVM Results cont.

Radial Bias Function
Recall: 97.6



$$K(x, x') = \exp\left(-\gamma \|x - x'\|^2\right)$$

Polynomial
Recall: 1.00



$$K(x, x') = (\gamma x^T x' + r)^d$$

Final Comparison, Model Choice

Model	**Accuracy**	**Disqualifier**
K-means	S.S: 0.573	Not suited for binary labeled data
Logistic Regression	93.43%	Accuracy too low, inherently linear.
SVM w/ Polynomial Activation	92.98%	Accuracy too low, but excels in recall
SVM w/ RBF Activation	98.24%	

Lessons Learned

- Human/Expert annotations are very important contributors to highly accurate models
- Importance of aligning your data characteristics with the assumptions of your algorithm
- It can be very useful to re-split data after evaluating effectiveness of hyperparameters when building the final model.

Data Source

Wolberg, W. (1990). Breast Cancer Wisconsin (Original) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5HP4Z.

Acknowledgments & References

O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.

O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.

K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).

Analytics Vidhya. (2021, May 17). K-Means: Getting the optimal number of clusters. https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/

Contributions

Xienyam Chiu

- Logistic Regression Model

Jack Corley

- Support Vector Machines Model

Beatrice Filart

- Motivation, Current Implementations, EDA & Visualizations

Jung Huh

- K-Means Model