Breast Cancer Classification

(Determining the presence of Breast Cancer using a digitized image dataset)

Questions to Explore: What accuracy can these models achieve? Specificity? Sensitivity? What performance metrics should be of focus? What models make the most sense in classifying this dataset

Methods:
Data exploration: Correlation heat maps
Model Selection: Logistic Regression (simple and interpretable), Random Forest (Highly accurate and moderately interpretable), and Feedforward Neural Network (complex, no interpretability).

Results:

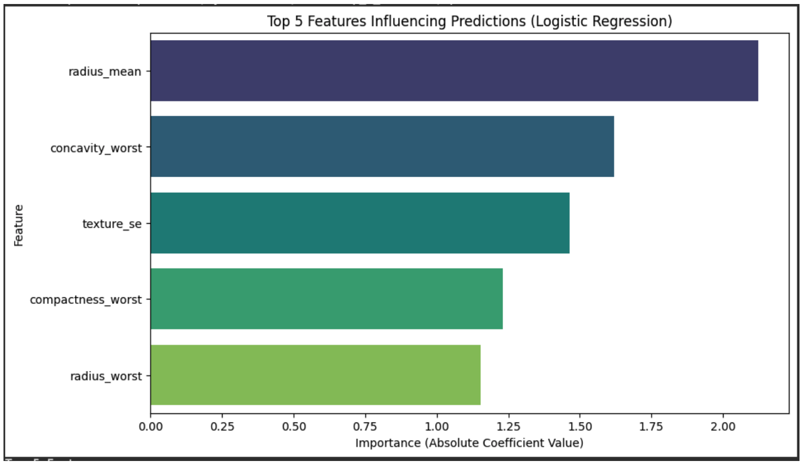
Model 1:

Logistic Regression ~

Accuracy: 96.5%, Sensitivity: 93.02%, Specificity: 96.8%

While logistic regression performed relatively well, it's important to consider the dataset we are training on. In the context of breast cancer detection, achieving high accuracy is crucial, and logistic regression alone may not be sufficient for reliable diagnosis.

We did, however, find these interesting results:



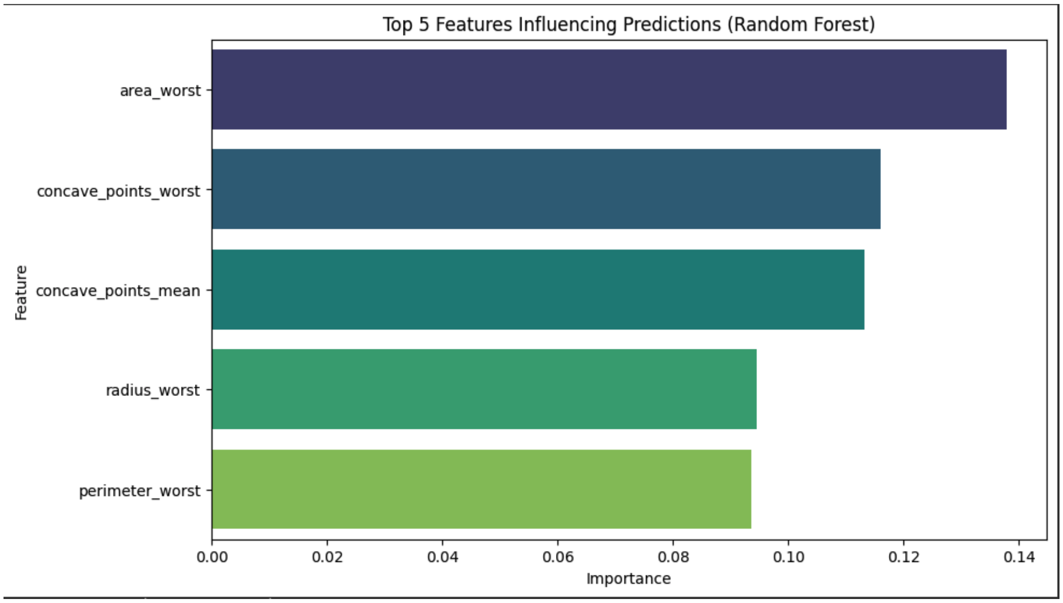
Model 2:

Random Forest ~

Accuracy: 98.2%, Sensitivity: 97.6%, Specificity: 98.7%

Random Forest showed promising results, missing only two cases. However, further testing on a much larger dataset is needed to ensure consistency and reliability.

Random Forest determined these variables had the most influence:



Model 3:

Feed Forward Neural Network ~

Accuracy: 90.4%, Sensitivity: 1.00%, Specificity: 84.5%

The FNN continuously overfit even after many tweaks. An FNN model may just not be necessary for this type of problem and we can likely stick with simple, more interpretable models.

1. Is this feasible?

Yes, I believe using Machine learning to detect breast cancer is extremely useful. It is already being used to help doctores to catch signs earlier and provide better diagnosis so patients can get better treatment. So, this is more than feasible, and will be a vital intersection in A.I in the future.

2. What sorts of other cases do you think you could detect, given these results and why?

As I mentioned above, image classification will prove its usefulness across the medical field. It will help doctors tremendously and will allow them to do their jobs even better.

3. What is the best model you were able to construct? Was it a neural network? How do you define "best"?

In this case, the random forest would be considered the best. It not only performed the best, but was very balanced, minimizing errors in each class.

4. (Bonus) How much compute power is necessary to make predictions with your model (e.g. CPU and RAM)?

2 gigs of Ram