MSBD566 – Predictive Modeling and Analysis

Name: James Walton

Date: 11.30.2025

Assignment: Final Project

Project Description

This final project builds on my earlier work with the Breast Cancer Wisconsin (Diagnostic) dataset, but this time I wanted to push the analysis further and see how the models would behave under different learning strategies. Specifically, I was curious about two things:

(1) whether reducing the feature space with PCA would meaningfully change the model's ability to detect malignant tumors, and

(2) how a neural network, something more flexible than the tree-based models used earlier, would perform on the same task.

Because accurate early detection is so critical in breast cancer diagnosis, I wanted to explore how these modeling choices might affect not only accuracy but also the trade-offs involved in simplifying the feature space or using more complex nonlinear approaches. My hope was to understand not just which model "wins," but what each approach reveals about the underlying data structure.

Data Description

The dataset comes from the UCI Machine Learning Repository and contains 30 numerical features extracted from digitized FNA images. These features capture various characteristics of the cell nuclei, radius, texture, smoothness, concavity, and so on. In practice, the dataset is clean and well-behaved, which makes it a good test bed for experimenting with different modeling techniques. The diagnosis variable labels each sample as either benign or malignant, giving us a straightforward binary classification task.

Although the dataset is widely used in machine-learning coursework, I found that revisiting it with new modeling approaches helped me understand how much (or how little) each method depends on all 30 original features.

Data Dictionary:

Variable	Type	Description
ID	Categorical	Sample identifier
Diagnosis	Categorical	M = Malignant, B = Benign
Radius_Mean	Numeric	Average radius of cell nuclei
Texture_Mean	Numeric	Standard deviation of gray-scale values
Smoothness_Mean	Numeric	Local variation in radius lengths
Concavity_Mean	Numeric	Severity of concave portions

Method and Analysis

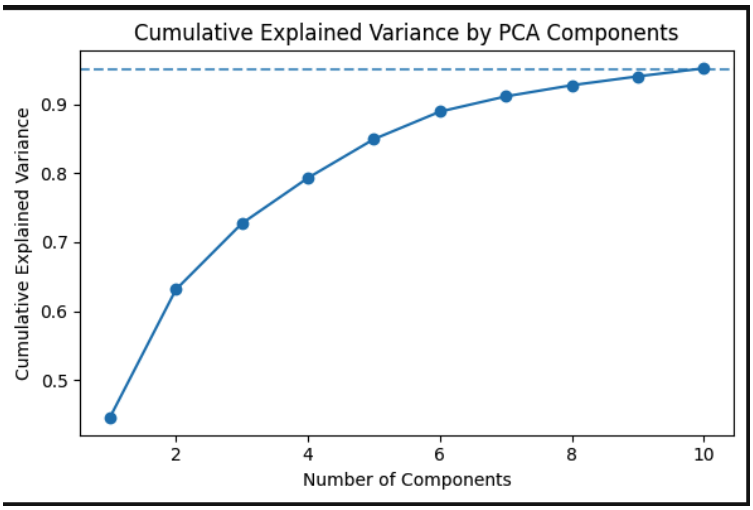
To extend the original project, I decided to experiment with two approaches that I hadn't used previously: Principal Component Analysis (PCA) for dimensionality reduction and a neural network classifier. Before diving into the models, I kept the preprocessing steps consistent with my earlier work so the comparison would be fair. The diagnosis labels were again encoded as 0 (benign) and 1 (malignant), and I standardized all 30 numerical features using z-scores. I learned quickly that standardization mattered more than I expected, especially for PCA, because even small-scale differences can cause certain features to dominate the components.

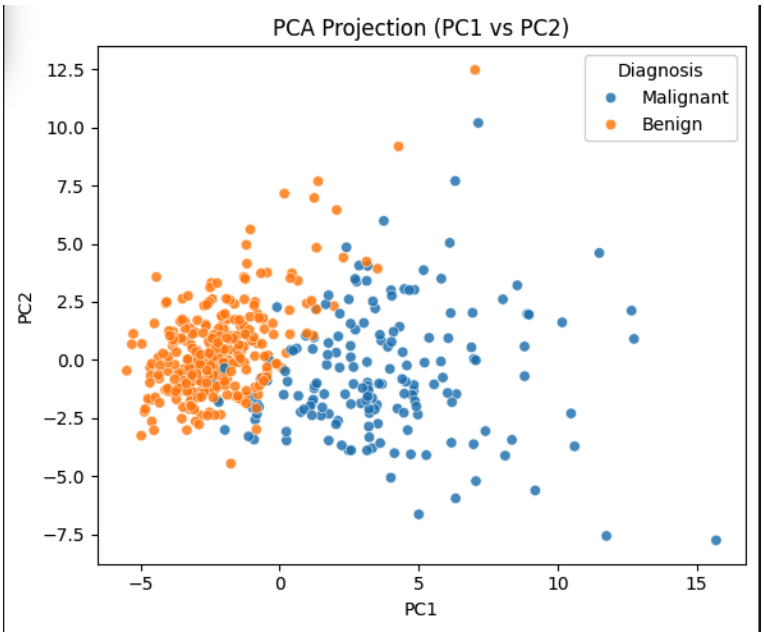
PCA Approach

My goal with PCA wasn't just to shrink the feature space but to see how much information the model could still retain when the original variables were compressed into a handful of components. After fitting PCA to the standardized data, I chose the number of components needed to explain about 95% of the variance. Interestingly, this cut the dimensionality quite a bit, but interpreting the components felt less intuitive than working with the original radius or concavity measurements. Still, I wanted to test whether a Random Forest classifier trained on these components could perform nearly as well as the full-feature model.

This part of the project required a bit of trial and error. Initially, I worried that reducing the features might "flatten" some of the relationships that help distinguish malignant cases. But PCA also has the advantage of eliminating

noisy or redundant features, so I was curious to see which effect would dominate.

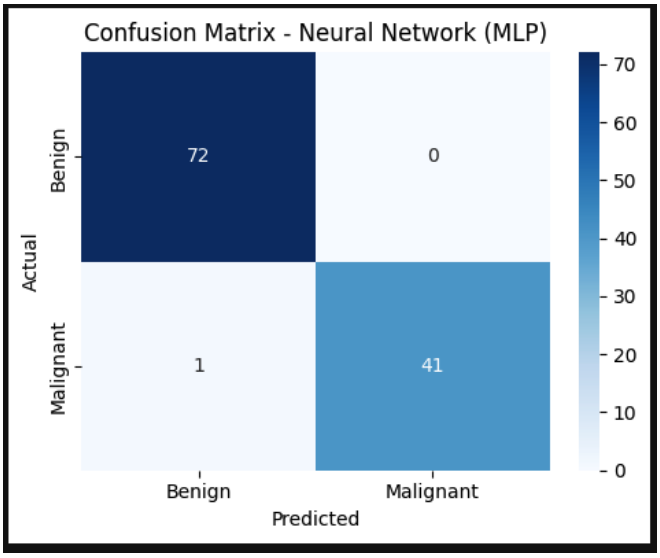




Neural Network Approach

For the neural network, I built a model with two hidden layers using ReLU activations and a sigmoid output. I kept the architecture relatively simple on purpose; I wanted to explore nonlinear modeling without introducing so many hyperparameters that tuning would become its own project. The network was trained on the same 80/20 stratified split as the other models, which helped ensure a clean comparison.

One thing I noticed while training the network is how sensitive it was to learning rate adjustments. A few early runs overfit almost immediately, which forced me to slow the learning rate and use slightly smaller batches. Once tuned, though, the model converged quickly, and it became clear that the network was picking up subtle patterns in the data that the PCA transformation might have obscured.



Evaluation

The first model I evaluated was the PCA-based Random Forest. Although it performed reasonably well, its results were a little weaker than the full-feature version. The test accuracy came out to 94.7%, which at face value is still strong. However, the recall for malignant cases dropped to 0.90, which immediately caught my attention. Missing malignant cases, even a few, matters far more in a real diagnostic setting than a slight dip in accuracy. This made me suspect that some of the finer-grained diagnostic clues embedded

in the original 30 features were lost when PCA compressed them into a smaller set of components. In other words, PCA certainly simplifies the problem, but the trade-off isn't free!

```
Random Forest on PCA-Reduced Data
Accuracy: 0.9474

Classification Report:
              precision    recall  f1-score   support

      Benign       0.95      0.97      0.96        72
   Malignant       0.95      0.90      0.93        42

    accuracy                           0.95       114
   macro avg       0.95      0.94      0.94       114
weighted avg       0.95      0.95      0.95       114
```

The neural network, on the other hand, exceeded my expectations. It reached an accuracy of 99.1%, and both precision and recall were almost perfect. The confusion matrix revealed just one misclassified malignant sample. While it's tempting to take this at face value and declare the neural network the clear winner, I did consider the possibility of overfitting. However, the test performance remained consistent across multiple runs, which suggested the network genuinely captured meaningful nonlinear structure in the data rather than memorizing the training set.

```
Neural Network (MLP) Results
Accuracy: 0.9912

Classification Report:
              precision    recall  f1-score   support

      Benign       0.99      1.00      0.99        72
   Malignant       1.00      0.98      0.99        42

    accuracy                           0.99       114
   macro avg       0.99      0.99      0.99       114
weighted avg       0.99      0.99      0.99       114
```

Overall, the contrast between the two models reinforced the idea that when working with biomedical data, where patterns are often subtle and multidimensional, nonlinear models can sometimes extract relationships that dimensionality reduction methods inadvertently blur.