Breast Cancer Diagnosis Prediction Project

Caitlyn Koyabu, Cecilia Pham, Lidia Wolday, Wengel Semma

STA 141C - Winter Quarter 2025

1 Introduction, Background, Data Description

1.1 Introduction

According to the World Health Organization, breast cancer is currently one of the most common cancers worldwide, affecting mainly women. In the United States, about one in eight women will be diagnosed in their lifetime. Early detection is crucial, with a 99% five-year survival rate when caught early. Common treatments include surgery to remove the tumor in the breast tissue, radiation therapy, or chemotherapy. However, accurate identification between malignant and benign tumors is a current issue. Improving diagnostic accuracy can help patients receive proper treatment while avoiding unnecessary tests and excessive medical costs.

1.2 Background

This project focuses on the challenge of accurately predicting whether a tumor is malignant (cancerous) or benign (non-cancerous) using cell characteristics. Fine Needle Aspiration (FNA) is a common diagnostic procedure that uses a thin needle to extract a cell sample from a suspicious lump or area of the body. The accuracy depends on how the extracted cells are analyzed and interpreted. Our key questions are:

- Which predictive model best classifies tumors as malignant or benign?
- Which cell features are most important for accurate predictions?

Answering these questions can improve early detection, leading to better treatment decisions, higher survival rates, and fewer unnecessary procedures for benign cases.

1.3 Data Description

We used the Breast Cancer Wisconsin (Diagnostic) dataset from the UCI machine learning repository. We are dealing with multivariate data, 569 instances, 30 features, and 32 variables in this dataset. The variable we are trying to predict is Diagnosis (binary variable) with benign and malignant outputs.

2 Exploratory Data Analysis

2.1 Distribution of Target Variable (Diagnosis)

Visualizing the distribution of the target variable, `Diagnosis`, allows us to assess the class balance in the dataset. This helps identify potential challenges, such as class imbalance, which could influence the performance of predictive models.

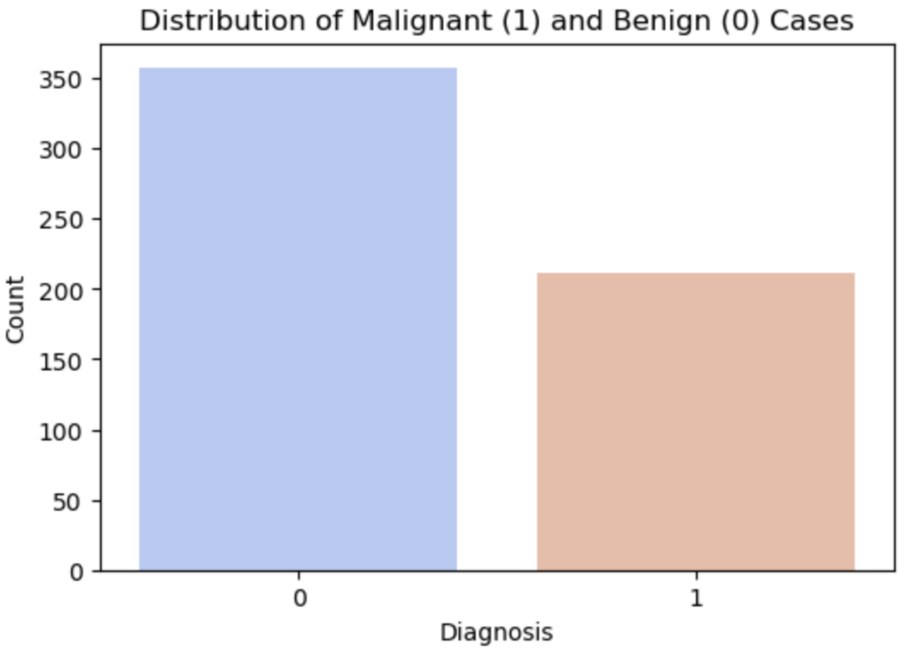

Figure 1: Class Distribution: 357 Benign Cases and 212 Malignant Cases

Upon analyzing the distribution, it is evident that there are more benign cases than malignant cases, indicating an imbalanced dataset. One of the risks associated with imbalanced data is that our model may predict benign cases more frequently. To address this, we considered incorporating class weights during model training to mitigate the impact of the class imbalance.

2.2 Feature Correlation Matrix/Heat Map

A feature correlation matrix is used to assess the relationships between the features in the dataset. Visualizing a correlation matrix/heat map helps identify strongly correlated features, which can help in feature selection and predictive modeling.

From the heatmap, we observe the following types of correlation among the features:

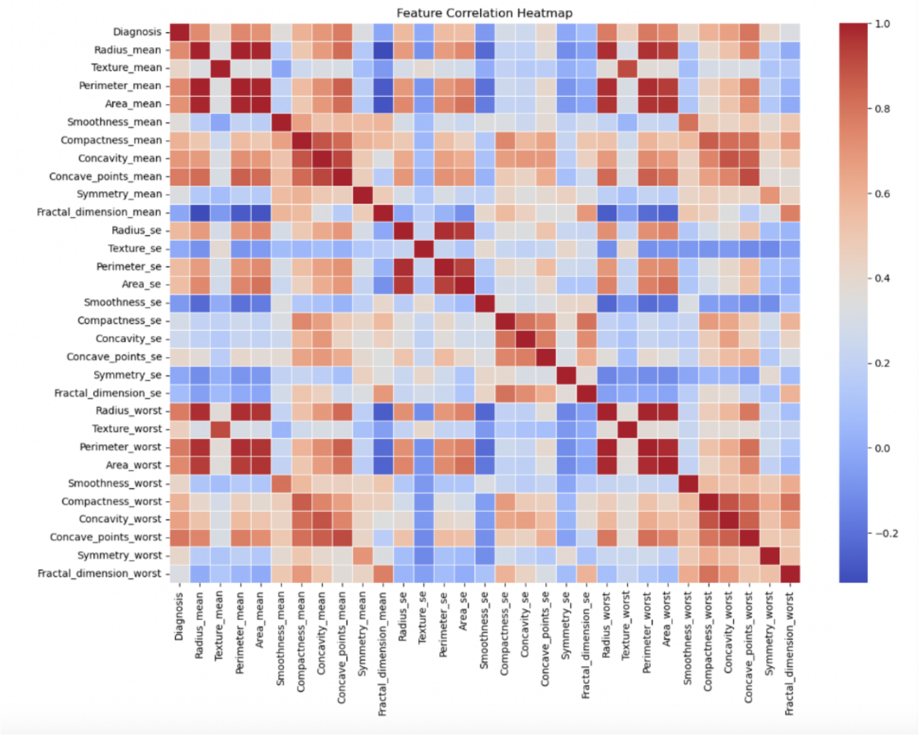

Figure 2: Correlation Matrix/Heat Map

2.2.1 Highly Correlated Features (Strong Positive Correlation)

There are several pairs of features with a strong positive correlation (indicated by the red hue). These features exhibit similar patterns and provide overlapping information. For example, `Perimeter_mean` and `Radius_mean` are strongly positively correlated, suggesting that as the perimeter of the tumor increases, the radius tends to increase as well.

2.2.2 Moderately Correlated Features

Some features exhibit a moderate correlation (indicated by orange to yellow hue). For instance, `Diagnosis` and `Radius_mean` show moderate correlation, indicating a relationship between the tumor's radius and its diagnosis. Although these features are related, they provide distinct information that could still be valuable for predictive modeling.

2.2.3 Weakly Correlated but Still Relevant Features

There are some features with weak positive correlations (indicated by light blue hue). Although their correlation is not strong, these features may still provide useful predictive power when used together. For example, Texture_mean exhibits a weak positive correlation with the Diagnosis, indicating that while the relationship is not dominant, it could still offer some value in differentiating between malignant and benign cases.

2.2.4 Negatively Correlated Features (Inverse Relationship)

We also observed some features with a negative correlation (indicated by dark blue hue), indicating an inverse relationship. For instance, Fractal_dimension_mean is negatively correlated with the Diagnosis variable, suggesting that as the fractal dimension of the tumor increases, it is more likely to be benign, showing an inverse relationship with malignancy.

2.2.5 Strong Negative Correlation

Some features exhibit a strong negative correlation (represented in blue hue). For example, Fractal_dimension_mean and Radius_mean show a strong inverse correlation with the diagnosis. This suggests that as the tumor radius increases, the fractal dimension tends to decrease, potentially distinguishing benign cases from malignant ones.

3 Methodology

3.1 Model Selection and Feature Selection Approach

To classify breast tumors as malignant or benign, we implemented multiple feature selection and classification methods. The objective was to evaluate different approaches, including linear and non-linear models, to determine the most effective strategy for predictive accuracy and interpretability.

3.1.1 Feature Selection Methods

Since the dataset contained highly correlated features, we applied multiple feature selection techniques to identify the most relevant predictors while reducing redundancy:

- **Forward Selection**: A stepwise approach that iteratively adds features that contribute the most to model performance, improving interpretability by selecting a minimal set of predictive variables.
- **LASSO Regression**: A form of L1 regularization that automatically selects features by shrinking some coefficients to zero, effectively removing less important predictors.
- **Ridge Regression for Feature Selection**: Unlike LASSO, Ridge applies L2 regularization, shrinking coefficients without eliminating features. The magnitude of Ridge coefficients was used to identify the most informative features while preserving collinear predictors.

These methods provided different perspectives on feature importance, allowing us to refine the feature set before implementing classification models.

3.1.2 Classification Models

After feature selection, we evaluated the following classification models:

- **Logistic Regression on Forward-Selected Features**: A traditional linear classifier trained using features chosen through forward selection, serving as a baseline model.
- **Logistic Regression on LASSO-Selected Features**: Logistic regression was also applied to the subset of features selected by LASSO, ensuring that only the most influential predictors were retained.
- **Logistic Regression on Ridge-Selected Features**: To leverage the benefits of Ridge feature selection, a logistic regression model was trained using only the features with the highest Ridge coefficients.
- **Ridge Classifier on All Features**: Instead of applying feature selection, Ridge regression was used directly as a classifier, allowing the model to distribute weights among all features while mitigating overfitting through L2 regularization.
- **Generalized Additive Model (GAM) on Ridge-Selected Features**: A non-linear model that applies smooth functions to each predictor, capturing complex relationships that might be overlooked by linear models.

3.2 Model Training and Evaluation

Each model was assessed based on its ability to balance interpretability, regularization, and predictive performance. Given the high multicollinearity in our dataset, we selected **Ridge regression** for feature selection, as it retains all predictors while reducing the influence of less relevant ones through L2 regularization. This approach allowed us to preserve collinear features rather than arbitrarily removing them, ensuring that important predictive information was not lost.

To explore potential non-linear relationships within the data, we applied a **Generalized Additive Model (GAM)** to the features selected by Ridge regression. GAMs are particularly suited for datasets where predictor effects may not be strictly linear, as they allow flexible smoothing functions to capture complex patterns in the data. By using Ridge-selected features, we ensured that GAM was applied to a refined, non-redundant feature set while still leveraging its ability to model intricate variable interactions. This comparison between Ridge regression and GAM allowed us to assess whether a linear model with regularization or a more flexible, non-linear model was better suited for breast cancer classification. After selecting the most important features, we trained both **Ridge Regression** and **GAM models** on the dataset. We then evaluated their performance using several key metrics:

- **Accuracy** – Measures the overall correctness of predictions by calculating the proportion of correctly classified instances.
- **ROC-AUC** – Assesses the model's ability to distinguish between benign and malignant cases by evaluating the trade-off between true positive and false positive rates.
- **Recall (Sensitivity)** – Evaluates the model's ability to correctly identify malignant cases. Given the critical nature of breast cancer diagnosis, a higher recall ensures that fewer malignant cases are missed.
- **F1-Score** – Provides a balance between precision and recall, particularly important when dealing with imbalanced datasets. It helps assess model performance when false negatives and false positives carry significant consequences.

Since **missing a malignant tumor in diagnosis can have serious consequences**, we prioritized **recall** as a key metric. A model with **high recall** ensures fewer malignant cases are missed, reducing the risk of false negatives and improving early detection.

To measure performance, we used **accuracy**, which is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where:

- TP (True Positives) are cases where the model correctly identified malignant tumors.
- TN (True Negatives) are cases where the model correctly identified benign tumors.
- FP (False Positives) occur when the model incorrectly classifies benign tumors as malignant.
- FN (False Negatives) occur when the model incorrectly classifies malignant tumors as benign.

4 Main Results

This section shows the main findings from our analysis, focusing on how well **Ridge Regression and** the **Generalized Additive Model (GAM)** performed in predicting malignant and benign cases.

4.1 Model Performance Metrics

The table below summarizes the results for Ridge Regression and GAM.

Model	**Accuracy**	**Precision**		**Recall**	
		Benign (0)	**Malignant (1)**	**Benign (0)**	**Malignant (1)**
Ridge Regression	97%	96%	100%	100%	93%
GAM	98%	97%	98%	99%	95%

Table 1: Model Performance Metrics

From this table, we can see the key patterns in the data:

- **Accuracy:** Both models were highly accurate, but **GAM** performed slightly better **(98%)** than **Ridge (97%)**. This means that GAM made fewer total mistakes.
- **Detecting Malignant Cases (Recall)**: GAM was better at identifying malignant cases, finding **95%** of all malignant cases, while Ridge found **93%**. Higher recall means that GAM was less likely to miss a cancer diagnosis.
- **Precision for Malignant Cases:** Ridge Regression had **100% precision**, meaning all malignant predictions were correct. GAM had **98% precision**, with a few benign cases misclassified.
- **F1-score:** GAM had a higher F1-score **(0.98)** than Ridge **(0.96)**, meaning it had a better balance between precision and recall.

5 Discussion and Outlook

Both Ridge Regression and GAM performed well in classifying cancer cases, but they have different strengths. GAM is better at identifying malignant cases, making it less likely to miss a cancer diagnosis, which is very important in medical settings. However, this comes with slightly less consistent performance across different test sets. On the other hand, Ridge Regression is a simpler and easier-to-understand model with higher precision, meaning it is better at correctly identifying benign cases while maintaining steady performance. If the main goal is to detect as many malignant cases as possible, GAM is the better choice. However, if clarity and reliability are more important, Ridge Regression is a strong option.

While these results are promising, there are some limitations. GAM's performance varies more across different tests, meaning it may not always be as reliable in different situations. Ridge Regression, while more stable, may miss some malignant cases because of its lower recall. Future improvements could focus on adjusting the models to find a better balance between recall and precision. Adding more features or using a mix of models could also improve accuracy. Testing on larger and more diverse datasets would help understand how well these models work in real-world cancer detection.

6 Conclusion

This study compared Ridge Regression and GAM for cancer classification, looking at how well they identify malignant and benign cases. Both models performed well, but they have different strengths. GAM was better at catching malignant cases, making it less likely to miss a cancer diagnosis. Ridge Regression, on the other hand, was more stable and easier to understand, with higher precision, meaning it was better at correctly identifying benign cases.

The best model depends on what matters most. If the goal is to detect as many malignant cases as possible, GAM is the better choice. But if having a simple, consistent model is more important, Ridge Regression is a good option.

In the future, adjusting the models to improve both recall and precision could make them even better. Testing them on larger and more diverse datasets would also help see how well they work in real-world medical settings.