STA 141C- Winter Quarter

Breast Cancer
Diagnosis Prediction Project

Group Members: Caitlyn Koyabu
Cecilia Pham
Lidia Wolday
Wengel Semma

Introduction 1

We want to accurately predict whether a tumor is malignant or benign based on cell characteristics using data from the Fine Needle Aspiration (FNA) procedure.

Research Question

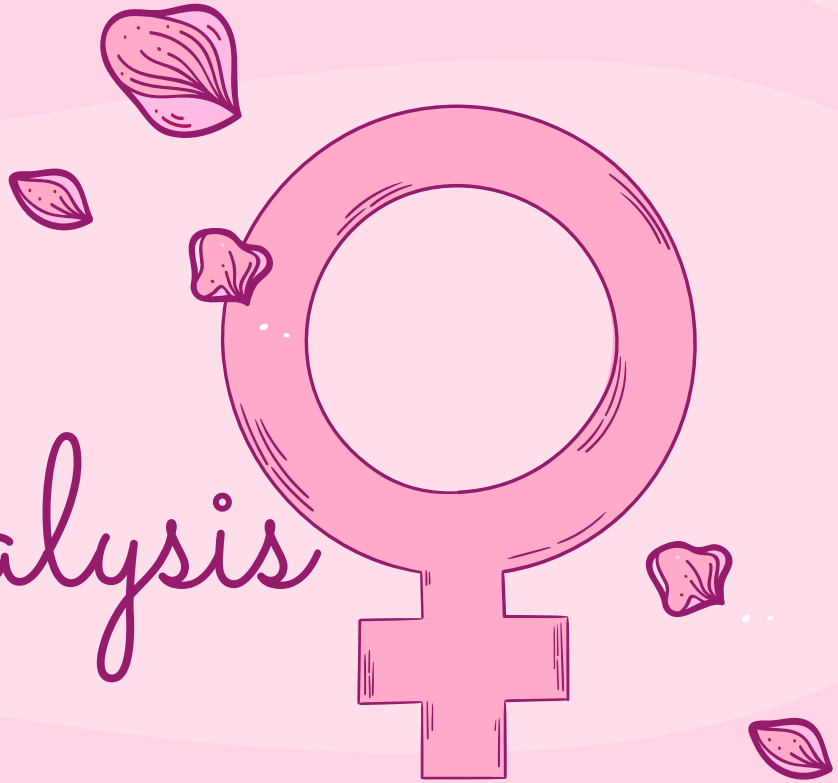
- Which predictive model best classifies tumors as malignant or benign?
- Which features are most important for accurate predictions?

Objectives

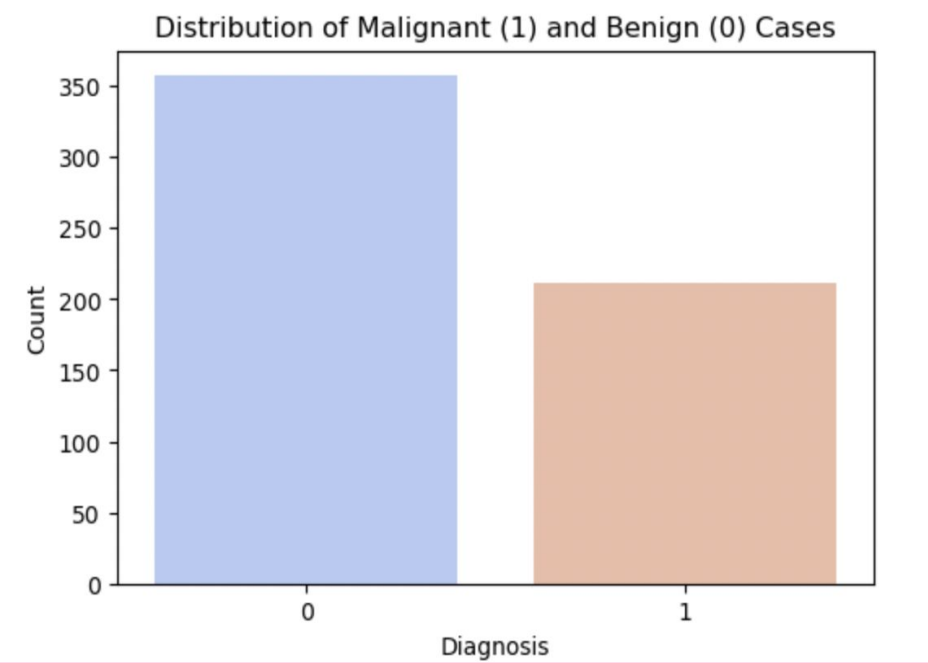
- Improve early detection, survival rates, and treatment decisions.
- Identify malignant and benign tumor for new data set

2 Data Analysis

Exploratory Data Analysis (EDA) & Model Fitting



Class Distribution & Imbalance Issue



EDA shows that

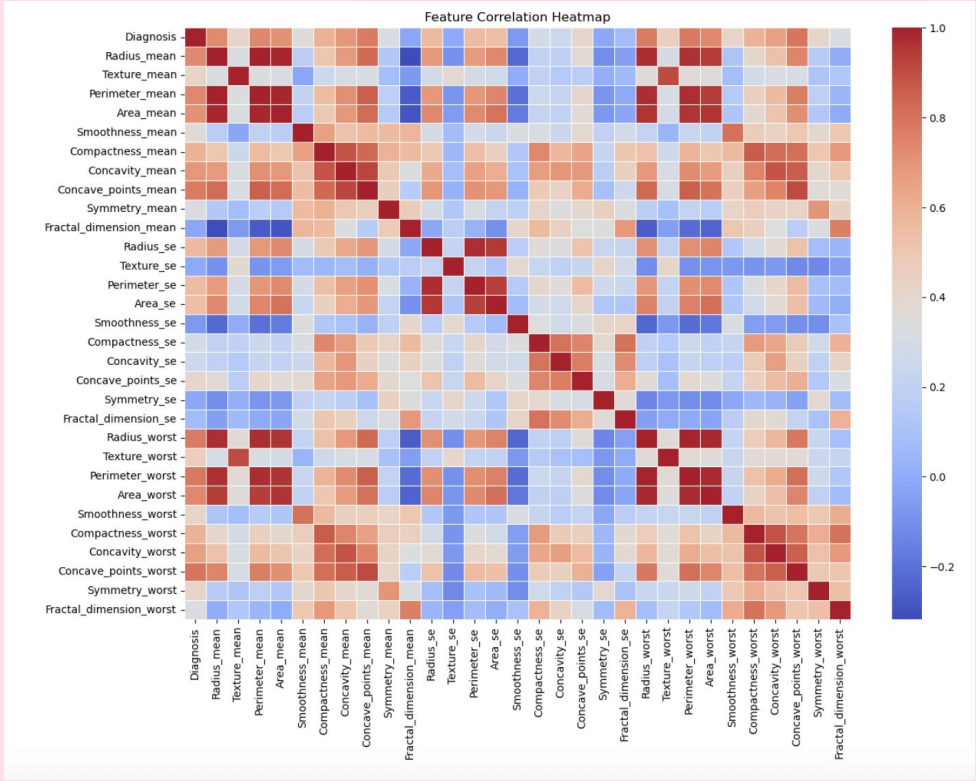
- More benign cases than malignant → **Imbalanced dataset**
- Risk: Model may predict benign cases more often

Benign (357 cases)

Malignant (212 cases)

Feature Correlation Matrix: Key Relationships

Heat Map



- Highly correlated features (strong positive correlation)
- Moderately correlated features.
- Weakly correlated but still relevant features.
- Negatively correlated features (inverse relationship)
- Strong negative correlation.

Model Fitting

Logistic Regression:

- Simple, interpretable
- Correlation: L1, L2 regularization
- Imbalance: class-weight adjustment

Forward Selection:

- Starts with no features, adds important ones step by step
- Helps pick the best features

Ridge vs. Lasso

Lasso

- Applies L1 penalty (shrinks some coefficients to zero for feature selection).
- Helps with automatic feature selection.
- Can be unstable if features are highly correlated.

Ridge

- Applies L2 penalty (shrinks coefficients but keeps all features).
- Better for handling multicollinearity and stabilizing models.
- Retains all features, making it less interpretable.

Accuracy for All Models

Model	Accuracy
Logistic Regression (Forward Selection)	0.99
Lasso	0.98
Ridge	0.98

Since the accuracy for lasso and ridge are the same, we decided to use ridge selected features because it retains more important features for our models.

Model Diagnostics

Ridge Regression:

- Handles multicollinearity well by applying an L2 penalty.
- Performs well for linear relationships between features.
- Less flexible in capturing complex, nonlinear patterns.
- More stable and interpretable compared to GAM.

Generative Additive Model (GAM):

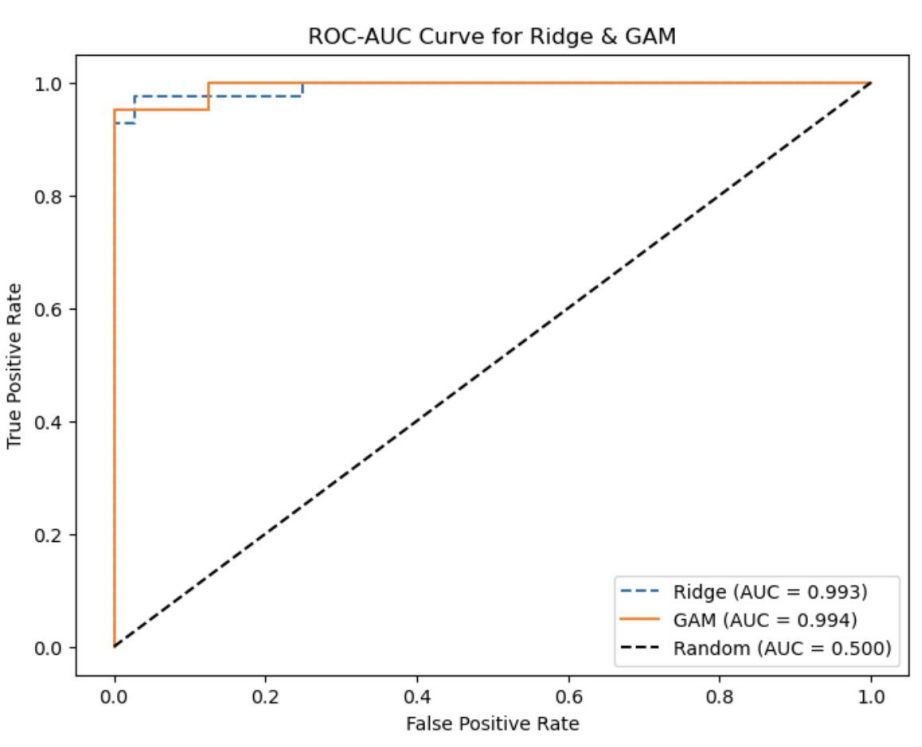
- Flexible for nonlinear relationships by using smooth functions.
- Can capture interactions between features better than Ridge.
- More complex and computationally expensive compared to Ridge.
- Requires tuning of smoothness parameters for best performance.

Model Performance

Model	Accuracy	Precision	Recall
Ridge Regression	97%	96% (B) / 100% (M)	100% (B) / 93% (M)
GAM	97%	97% (B) / 98% (M)	99% (B) / 95% (M)

- Both models achieved 97% accuracy.
- GAM had higher recall for malignant cases (0.95) compared to Ridge (0.93), finds more cancer cases, which is very important.

ROC-AUC Curve: Ridge vs. GAM



- Both models work very well, with high AUC scores (Ridge: 0.993, GAM: 0.994).
- GAM is the better choice to avoid missing cancer cases.

Conclusion

- We tested feature selection methods like Lasso and Ridge Regression and found Ridge Regression performed better. It penalizes coefficients without removing features and effectively handles multicollinearity.
- We compared ridge regression and GAM for classifying benign and malignant cases and found GAM performed better.
- If recall is the priority, GAM is the best model for identifying malignant tumors.