Stony Brook University

Breast Cancer Risk Assessment Using Logistic Regression

Orbin Ahmed Acanto

FAR BEYOND

Problem Statement

- Breast cancer is one of the leading causes of cancer-related deaths worldwide.
- Early detection and classification are crucial for improving survival rates.
- There are two major types of breast cancer (e.g., malignant vs. benign).

Goal: Develop a **machine learning model** to classify breast cancer types based on clinical and histological features.

Dataset Overview

- Source: UCI ML Repository (https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic)
- Number of samples: 569
- Target Variable (Diagnosis):
  - "M" → Malignant (cancerous)
  - "B" → Benign (non-cancerous)
- Features: The dataset consists of 30 numerical features derived from breast cancer cell nuclei characteristics, grouped into three categories:
  - Mean Measurements
  - Standard Error Measurements
  - Worst (Max) Value Measurements
- The dataset contains numerical features and a categorical target variable
- The feature values extracted from digital images of fine needle aspirate (FNA) biopsies of breast masses.

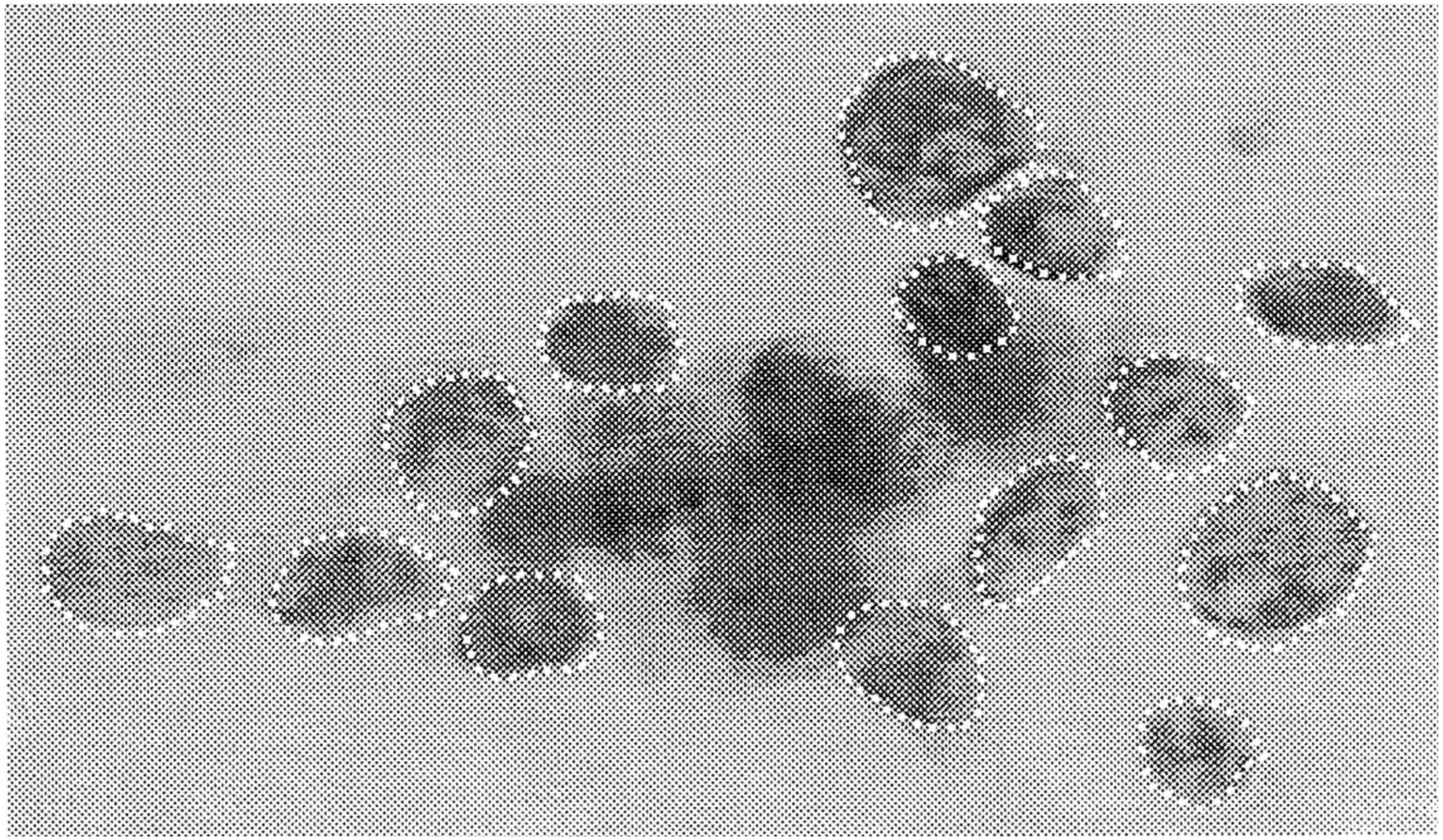


Figure 1: Initial Approximate Boundaries of Cell Nuclei
The user first draws a rough initial outline of some cell nucleus boundaries. Each outline serves as the initial position for a deformable spline which converges to an accurate boundary of the nucleus.

Street, William Nick et al. “Nuclear feature extraction for breast tumor diagnosis.” Electronic imaging (1993).

Variables

Variable Name	Role	Type	Description	Units	Missing Values
ID	ID	Categorical			no
Diagnosis	Target	Categorical			no
radius1	Feature	Continuous			no
texture1	Feature	Continuous			no
perimeter1	Feature	Continuous			no
area1	Feature	Continuous			no
smoothness1	Feature	Continuous			no
compactness1	Feature	Continuous			no
concavity1	Feature	Continuous			no
concave_points1	Feature	Continuous			no

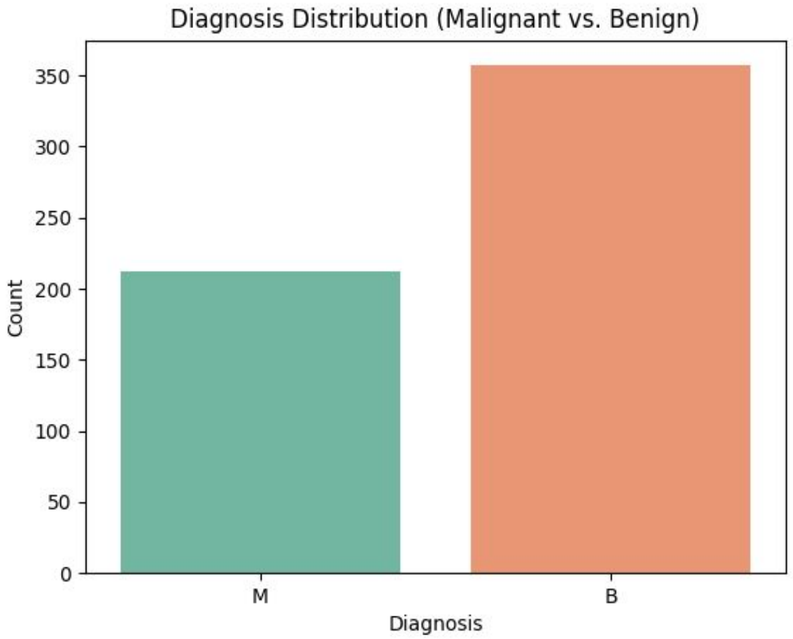
Analytical Techniques & Informatics Methods

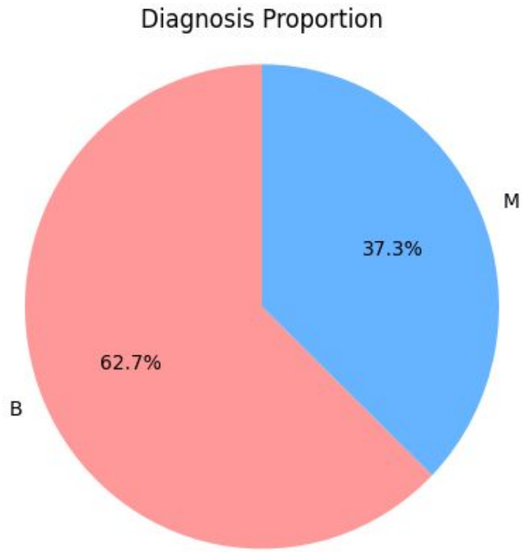
- Machine Learning Model
  - Logistic Regression:
    - Binary classification model (malignant vs. benign).
    - Provides probabilistic cancer risk assessment.
    - Interpretable model useful for medical applications.
- Data Preprocessing Steps
  - Data Normalization: Scaling features (0-1) to ensure model efficiency.
  - Drop some variable which are does not have significant effect over target variables
  - Data set Balancing: SMOTE
  - Splitting Dataset: 70% Training | 15% Validation | 15% Testing for model evaluation.

Software & Tools

- Programming Language
  - Python (preferred for ML & data analysis)
- ML Frameworks & Libraries
  - scikit-learn: Model development (Logistic Regression)
  - pandas & NumPy: Data preprocessing & analysis
  - imbalanced-learn: For class balancing
  - matplotlib & seaborn: Data visualization
- Jupyter Notebook: Experimentation environment
- Outcomes: A highly interpretable model for breast cancer classification.

Diagnosis Distribution: Class Balance Overview

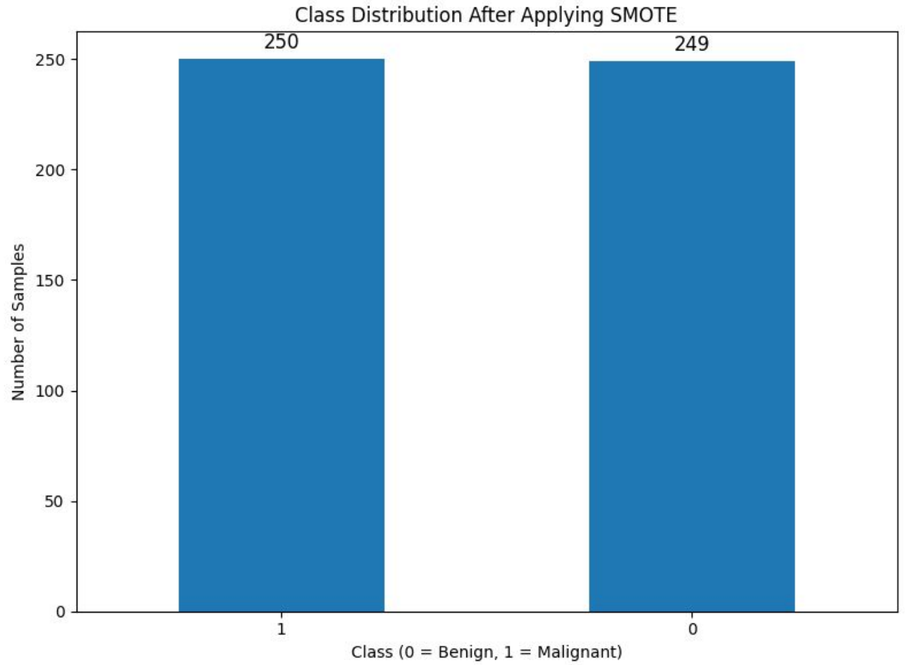




Malignant (M): 212 cases (~37.3%)
Benign (B): 357 cases (~62.7%)

This dataset contains **569 breast cancer cases**, labeled as either **Malignant (M)** or **Benign (B)**.

Diagnosis Distribution: Class Balance Overview (After applying SMOTE)



Train Set

Malignant (1): 250 cases (~50%)
Benign (0): 249 cases (~50%)

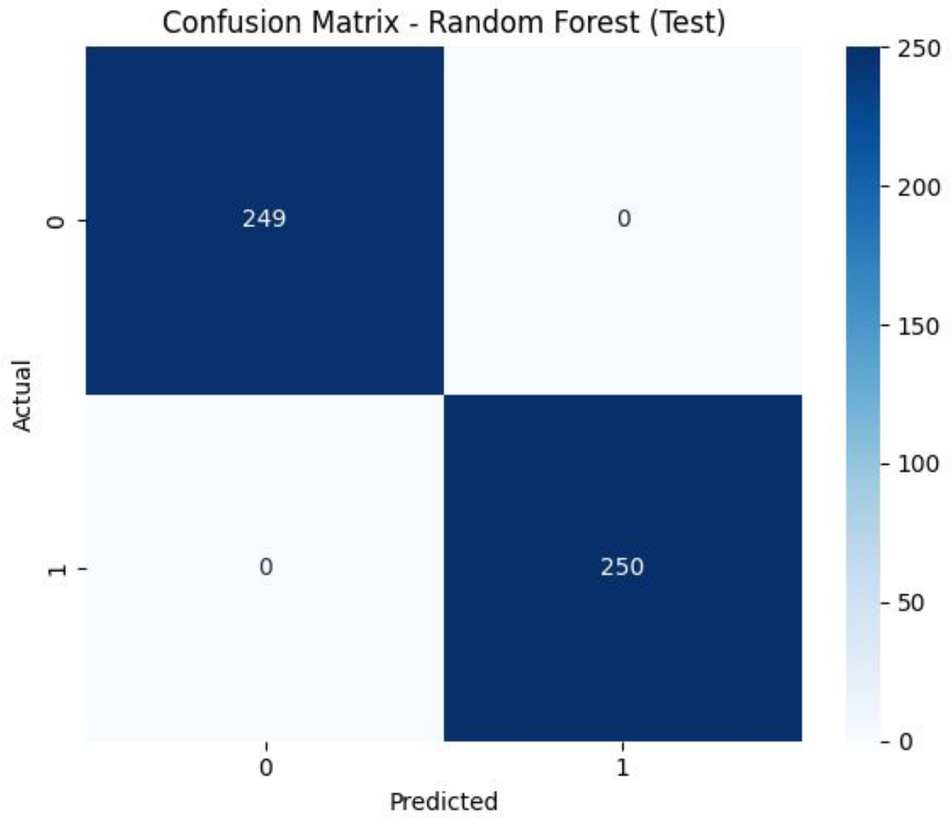
Test Set / Valid Set

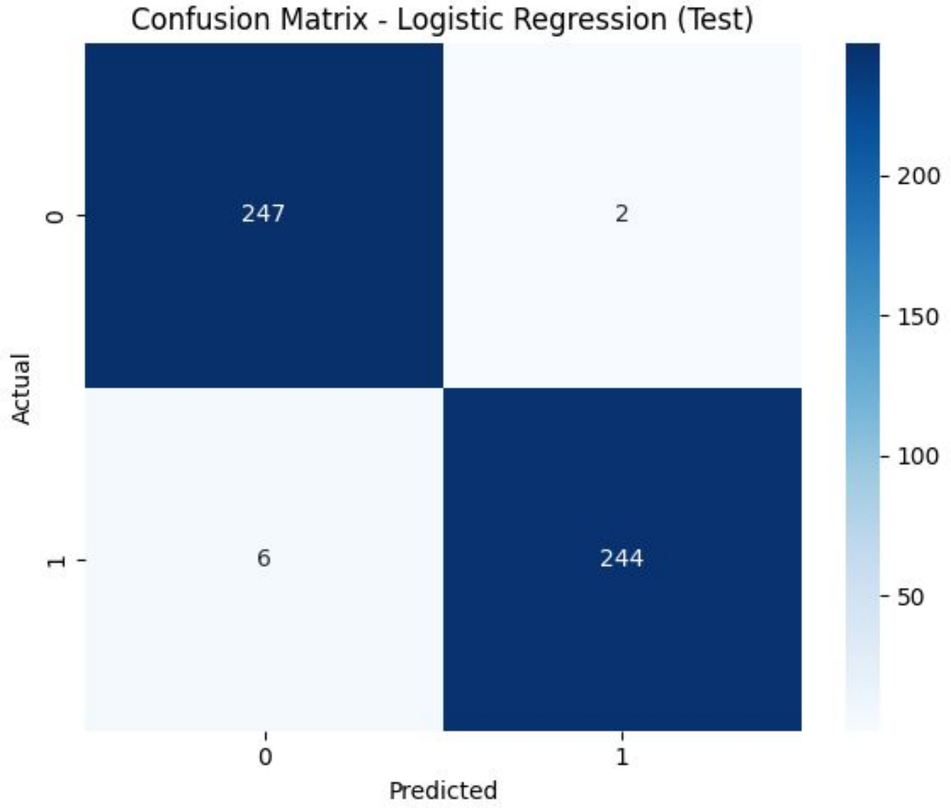
Malignant (1): 54 cases (~50%)
Benign (0): 53 cases (~50%)

New dataset contains **714 breast cancer cases**, labeled as either **Malignant (1)** or **Benign (0)**.

Result

Model	Dataset	Accuracy (%)	Precision (Benign)	Recall (Benign)	Precision (Malignant)	Recall (Malignant)
Logistic Regression	Validation	96.26	0.95	0.98	0.98	0.94
Random Forest	Validation	98.13	0.96	1.00	1.00	0.96
Logistic Regression	Test	98.39	0.98	0.99	0.99	0.98
Random Forest	Test	100.00	1.00	1.00	1.00	1.00


Confusion Matrix - Random Forest (Test)
249
0
0
250
Actual
Predicted
0
1
250
200
150
100
50
0

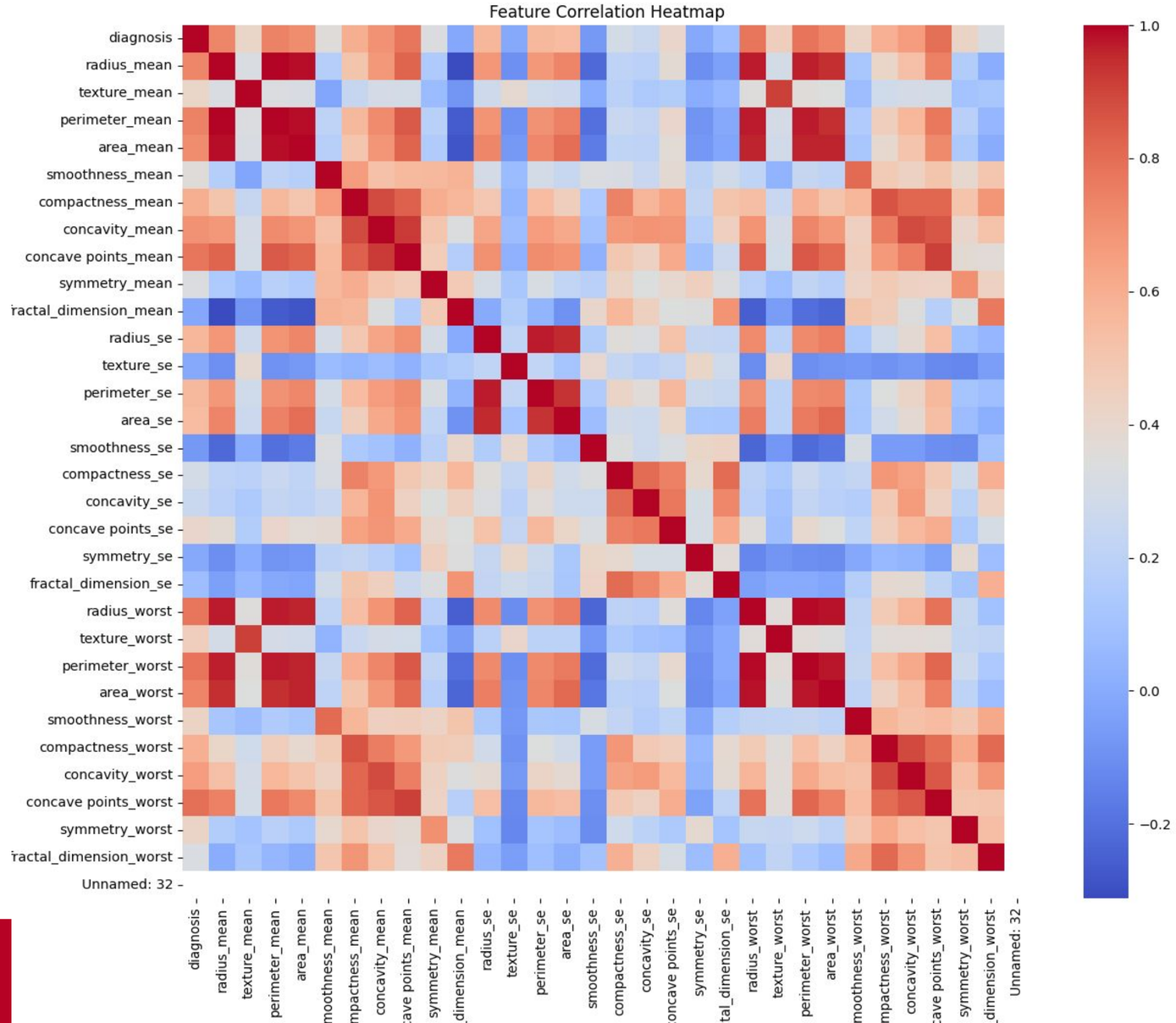

Confusion Matrix - Logistic Regression (Test)
247
2
6
244
Actual
Predicted
0
1
200
150
100
50

Stony Brook Medicine

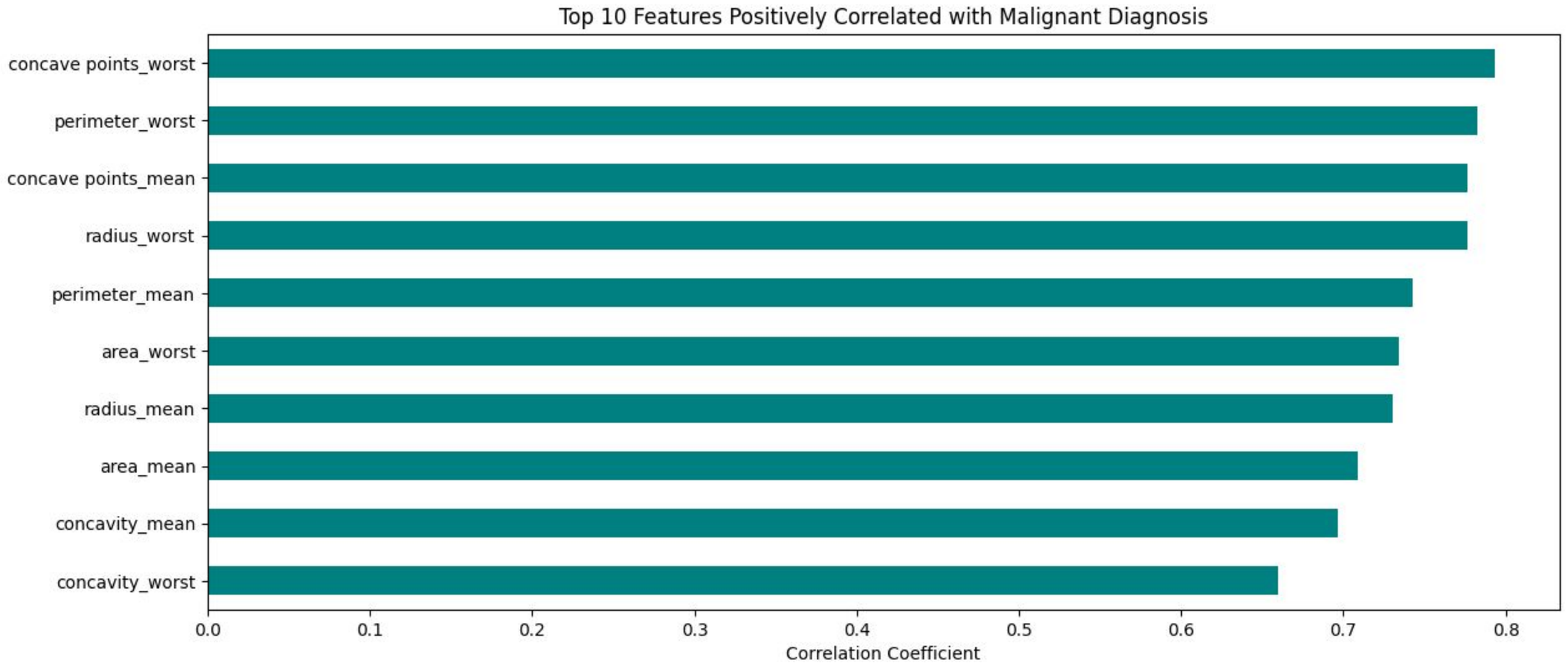
SUNY

Feature Correlation Heatmap

- A heatmap was generated to understand how features relate to each other and the target (diagnosis).
- Red indicates strong positive correlation, blue indicates negative correlation, and lighter shades indicate weaker correlations.
- Some features are highly correlated with the target (like radius, area, concavity).
- This insight will help in feature selection, reducing noise and improving model performance.



Top Predictive Features



Model Evaluation

Train-Test Strategy:

- Dataset split: 70% training, 15% testing, 15% validation
- **Stratified split** to preserve class distribution

Evaluation Metrics:

- Accuracy: Overall model correctness
- Precision: Correctly identified malignant cases among predicted malignant
- Recall (Sensitivity): Malignant cases correctly detected
- F1 Score: Balance of precision and recall

Result (Top 10 Features)

Experiment	Features	Logistic Regression (Test Acc)	Random Forest (Test Acc)
All Features (30)	30	98.39%	100%
Top 10 Correlated Features Only	10	94.79%	100%

- **Random Forest** is extremely robust, it performed perfectly with both 30 and 10 features.
- **Logistic Regression** benefits from **more features** to maintain very high accuracy.
- Feature selection **simplifies models** but sometimes **costs small accuracy loss** for simpler models like Logistic Regression.
- **Training time** becomes **faster** with fewer features.

Thank You.