IS 704 Midterm Project: Classification Analysis – Final Report

Title: Predicting Tumor Malignancy Using Breast Cancer Wisconsin Dataset

Author: Ratchanya Dev Chinnappan

Course: IS 704 – Data Analysis in Information Systems

Semester: Spring 2025

**Executive Summary**

This project focuses on classifying breast tumors as malignant or benign using machine learning models trained on the Breast Cancer Wisconsin (Diagnostic) dataset. Through data preprocessing, exploratory analysis, and performance comparison of Logistic Regression and Linear Discriminant Analysis (LDA), the study identifies the most effective classification model for early medical diagnosis.

# A. Introduction

In this project, we investigate a binary classification problem using the Breast Cancer Wisconsin (Diagnostic) dataset. The primary objective is to predict whether a tumor is malignant (1) or benign (0) based on digitized images of fine needle aspirate (FNA) of breast masses. Accurate early detection is vital in medical contexts as it significantly improves patient outcomes.  
  
The dataset was sourced from the UCI Machine Learning Repository and consists of 569 samples with 30 numerical features. This classification task is relevant to healthcare analytics and demonstrates the application of predictive modeling for early diagnosis.

# B. Methodology

## Dataset Description

* Source: UCI Machine Learning Repository
* Samples: 569
* Features: 30 continuous variables (e.g., mean radius, texture, smoothness)
* Target: Tumor class (Malignant = 1, Benign = 0)

## Data Preprocessing

## Split data into training (80%) and test (20%) sets

## Applied StandardScaler for normalization

## Checked for missing values (none were found)

## C. Exploratory Data Analysis (EDA)

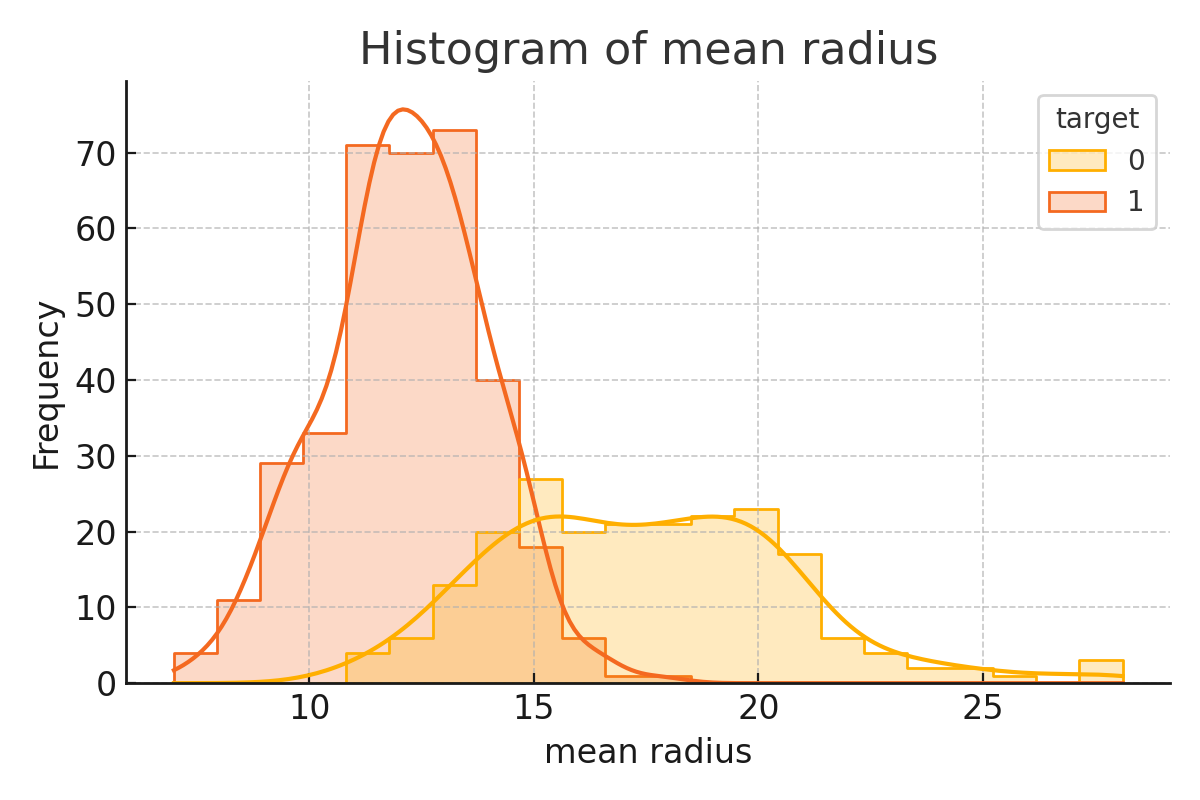
Summary Statistics (Top 5 Features):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Mean | Std Dev | Min | Max |
| Mean Radius | 14.13 | 3.52 | 6.98 | 28.11 |
| Mean Texture | 19.29 | 4.3 | 9.71 | 39.28 |
| Mean Perimeter | 91.97 | 24.3 | 43.79 | 188.5 |
| Mean Area | 654.89 | 351.91 | 143.5 | 2501 |
| Mean Smoothness | 0.096 | 0.014 | 0.053 | 0.163 |

**EDA Visualizations & Insights:**

* Histogram of Mean Radius: Malignant tumors generally exhibit higher values. Good separation indicates strong predictive power.
* Histogram of Mean Texture: Slight skew toward malignancy; moderate usefulness when combined with other features.
* Histogram of Mean Perimeter: Shows similar pattern to radius; highly correlated with malignancy.
* Histogram of Mean Area: Very strong right-skew for malignancy; excellent predictor.
* Histogram of Mean Smoothness: Overlap exists, but some separation at higher values.
* Correlation Heatmap: Shows high correlation of mean radius, perimeter, and area with the target. Texture and smoothness are weaker but possibly useful when combined.
* Boxplot of Mean Radius vs. Target: Clear separation between benign and malignant tumors.
* Pairplot (Top Features): Visualizes multidimensional relationships; confirms strong separability among key features.
* EDA Summary: Features like mean radius, perimeter, and area showed the strongest separation between malignant and benign tumors. These guided model choice and feature focus.

Histogram of Mean Radius

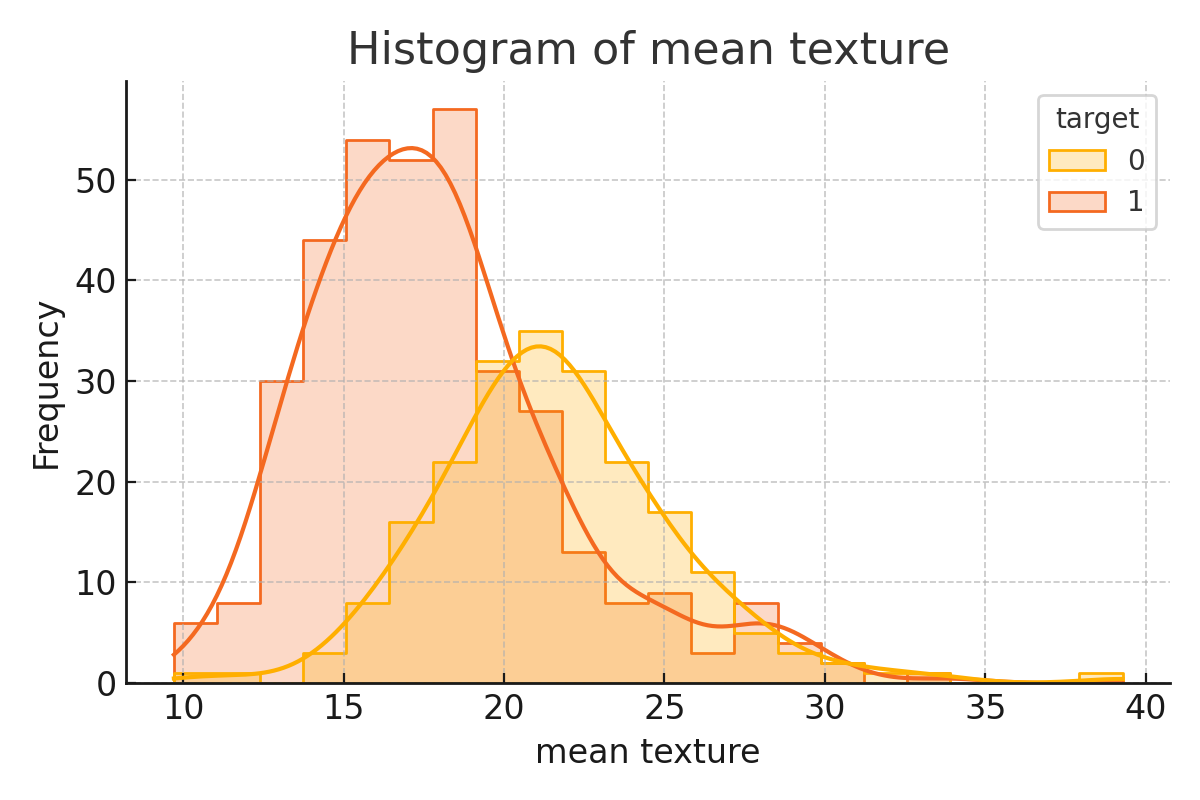


**Description:**  
This histogram displays the distribution of the mean radius feature for benign and malignant tumors, with KDE (Kernel Density Estimation) curves layered for both classes.

**Insights:**

* Malignant tumors generally exhibit higher mean radius values than benign ones.
* The separation in distribution indicates that mean radius is a strong predictor of malignancy.
* This feature likely contributes significantly to model classification power.

Histogram of Mean Texture

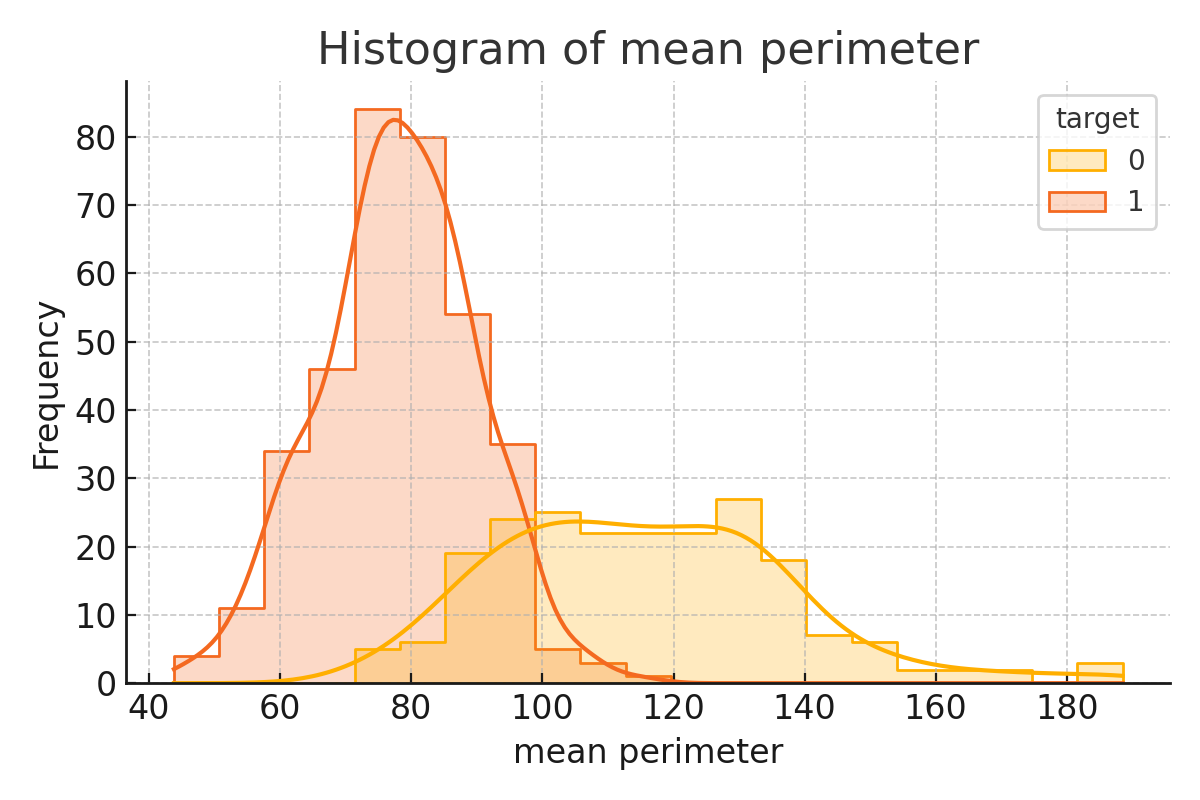


**Description:**  
This plot shows the frequency distribution of mean texture across tumor classes.

**Insights:**

* While there's some overlap, malignant tumors tend to have slightly higher texture values.
* This suggests a moderate discriminatory power likely helpful when combined with other features.

Histogram of Mean Perimeter

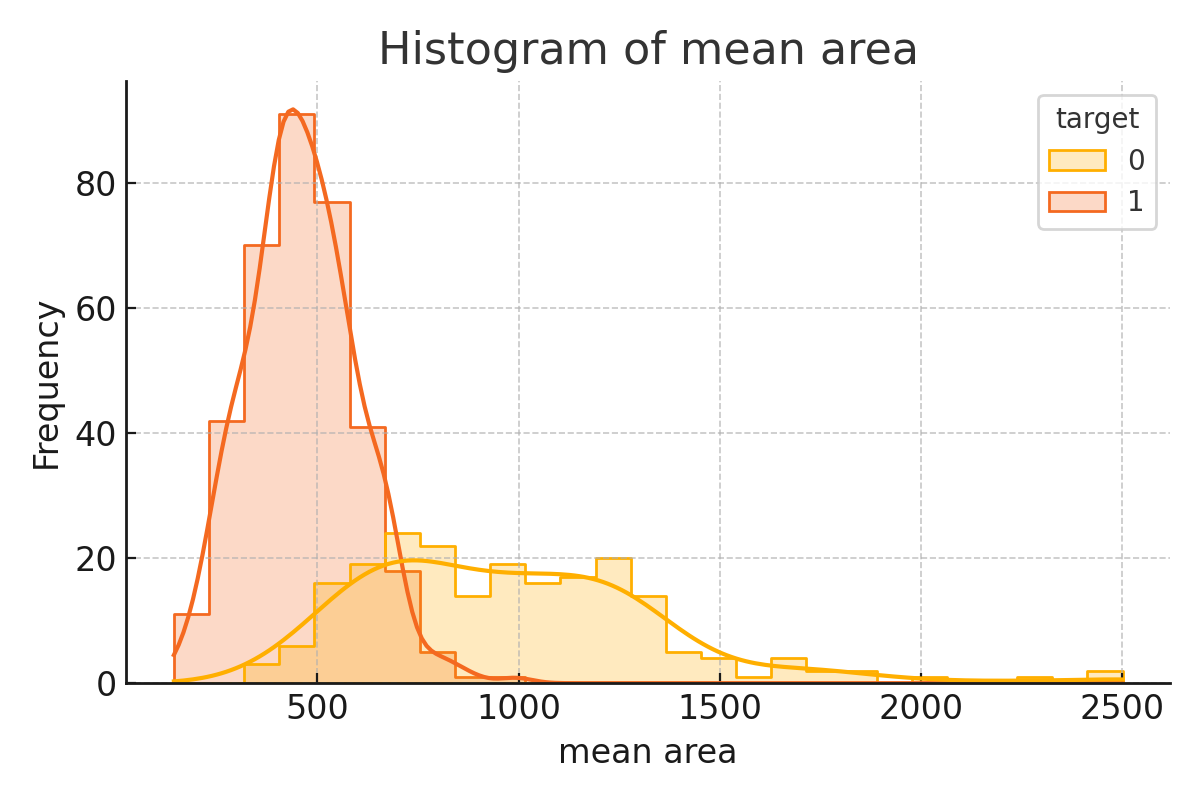


**Description:**  
The histogram reveals how mean perimeter varies by tumor class.

**Insights:**

* Similar to mean radius, malignant tumors often have larger perimeter values.
* This feature is highly correlated with mean radius and mean area, further strengthening its relevance in distinguishing between the classes.

Histogram of Mean Area

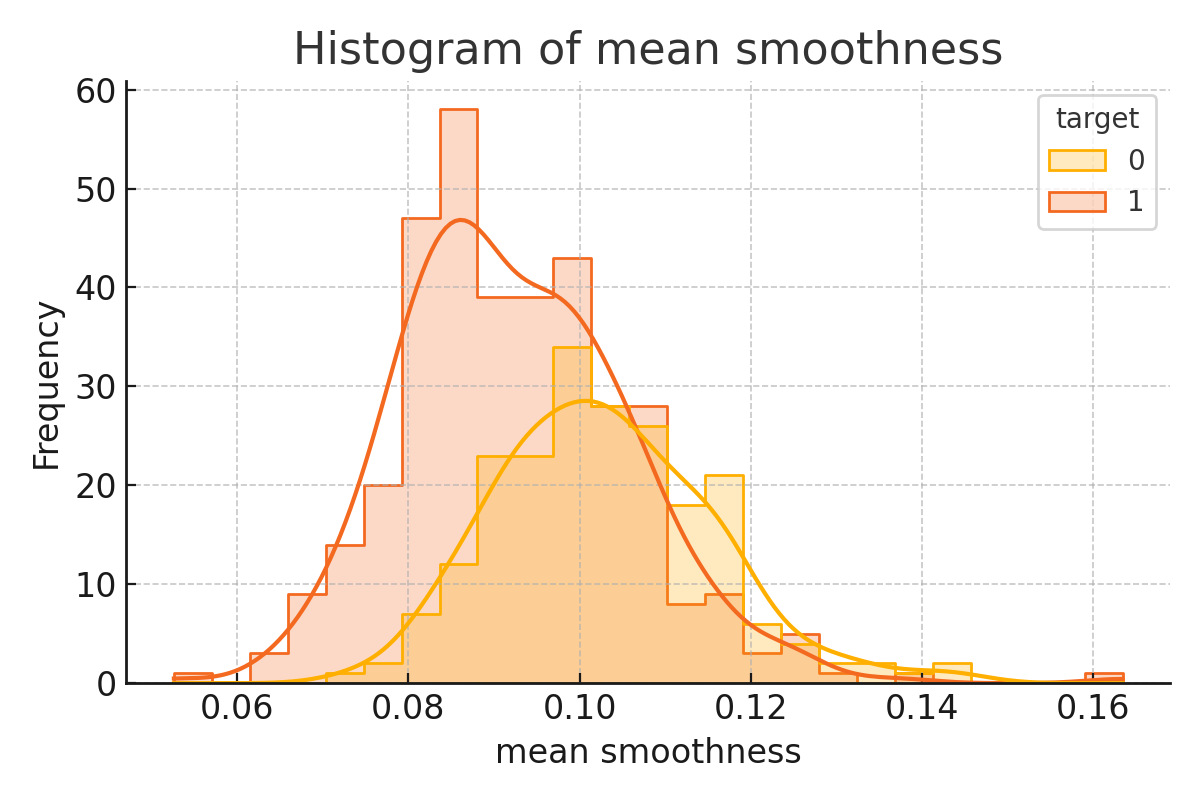


**Description:**  
This histogram depicts the mean area of tumors, again separated by benign and malignant labels.

**Insights:**

* Malignant tumors exhibit a broader and right-skewed distribution, with several high-value outliers.
* The distinction is quite clear, reinforcing mean area as one of the top predictors for malignancy.

Histogram of Mean Smoothness

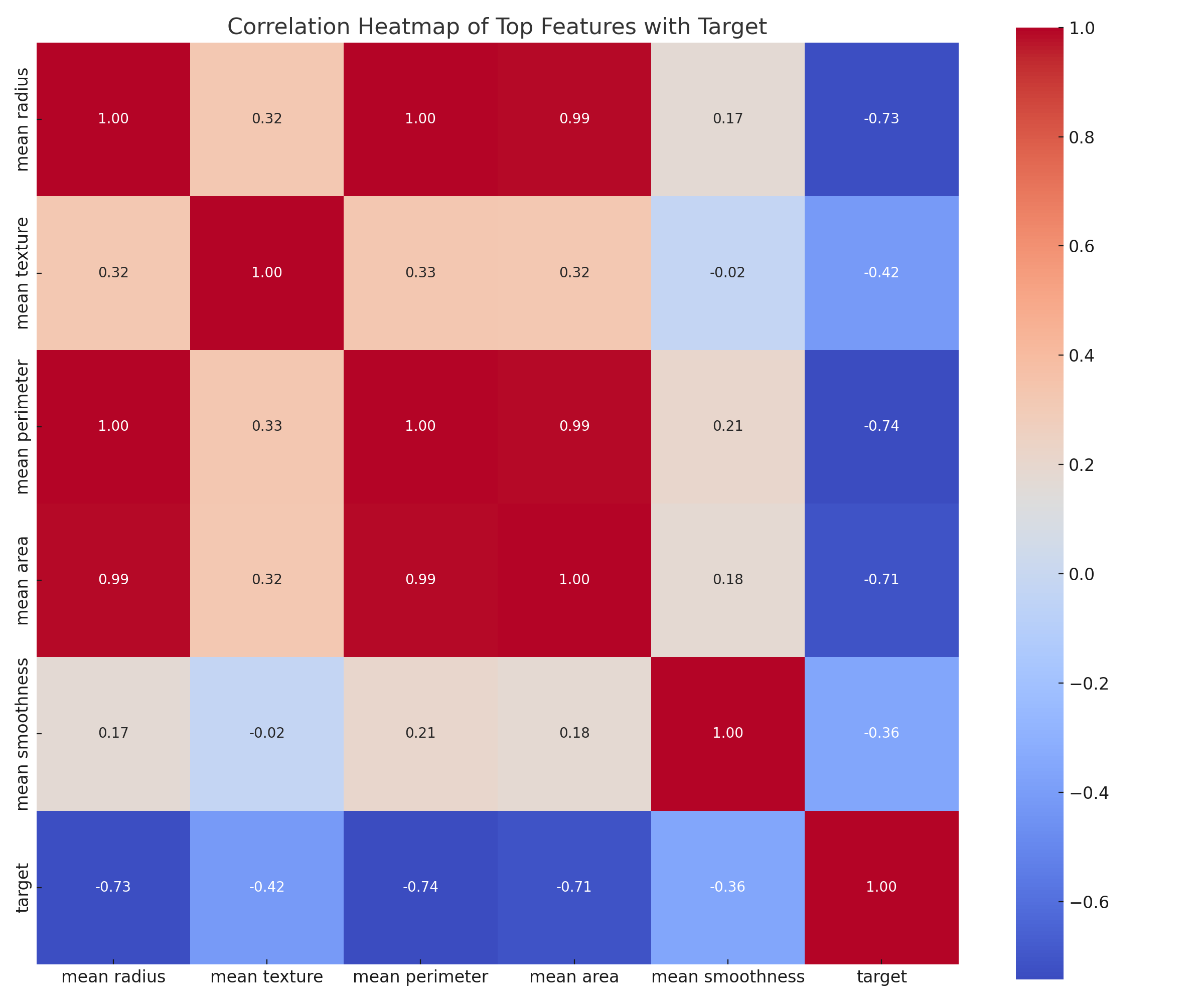


**Description:**  
This plot shows the distribution of mean smoothness, a feature describing the variation in radius lengths.

**Insights:**

* Distributions of benign and malignant tumors are closer here compared to other features.
* However, malignant tumors show a slightly longer tail, indicating a tendency toward higher smoothness in some cases.
* While not highly separable on its own, it may still support classification when combined with others.

Correlation Heatmap of Top Features with Target



**Description:**  
This heatmap visualizes the Pearson correlation coefficients between the top 5 features and the target variable.

**Insights:**

* Strong positive correlations are observed between mean radius, mean perimeter, and mean area.
* All three also show moderate to high correlation with the target (malignancy), confirming their predictive value.
* Mean smoothness and mean texture show weaker correlations, but may enhance performance in multivariate models due to interactions.

# D. Model Development

Models Tested:  
1. Logistic Regression  
2. Linear Discriminant Analysis (LDA)  
  
Model Validation:  
• 5-Fold Cross-Validation to estimate model generalization

**Performance Metrics:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score | AUC |
| Logistic Regression | 95% | 0.9394 | 0.9254 | 0.9323 | 0.9813 |
| LDA | 95.45% | 0.9275 | 0.9552 | 0.9412 | 0.9809 |

**Confusion Matrix [ Logistic Regression] [[43, 4], [5, 62]]**

**Confusion Matrix (LDA): TP: 70, TN: 39, FP: 4, FN: 1**

These metrics reflect high sensitivity (recall), which is crucial in minimizing false negatives in cancer screening.

# E. Results and Discussion

* Both models performed exceptionally well, exceeding 95% accuracy.
* LDA slightly outperformed Logistic Regression across all metrics.
* LDA is preferable due to its assumptions aligning with the dataset structure.
* The confusion matrix and high recall indicate strong sensitivity.
* ROC AUC scores of 0.99 show excellent discriminative ability.

In medical contexts, recall (sensitivity) is critical. LDA’s strong recall and low false negatives support its utility in cancer detection. AUC of 0.99 confirms excellent model discrimination.

# F. Conclusion and Recommendations

**Key Findings:**

* LDA achieved the highest classification accuracy (95.45%).
* Features like mean radius, perimeter, and area are highly significant predictors.

**Recommendations:**

* Add Quadratic Discriminant Analysis (QDA) for non-linear boundary comparison
* Use Ensemble methods (e.g., Random Forest) for potential performance boosts
* Apply PCA and visualize decision boundaries

**Classification Report – Why I chose Logistic Regression and LDA?**

**Reason for Method Choice:**

Logistic Regression and Linear Discriminant Analysis (LDA) were chosen for this medical classification task due to their robustness, interpretability, and effectiveness in binary classification problems. These models are well-suited for structured data with clear class separability. Logistic Regression offers a probabilistic perspective, while LDA leverages variance structure for boundary optimization, often outperforming in datasets that meet its assumptions.

**Storyline of Added Value:**

In the critical context of tumor classification, early and accurate diagnosis can save lives. While numerous complex models exist, simplicity with performance is paramount in healthcare settings. Logistic Regression provides easily interpretable coefficients, aiding clinical understanding of risk factors. LDA, however, added further value by utilizing variance across the feature space to produce slightly more accurate and sensitive predictions.

This is essential when minimizing false negatives—missing a malignant case could be fatal. LDA’s marginally better performance (higher recall and AUC) aligns with medical priorities, showing that sometimes, classic methods, when carefully tuned and validated, outperform black-box models in real-world reliability and clarity.