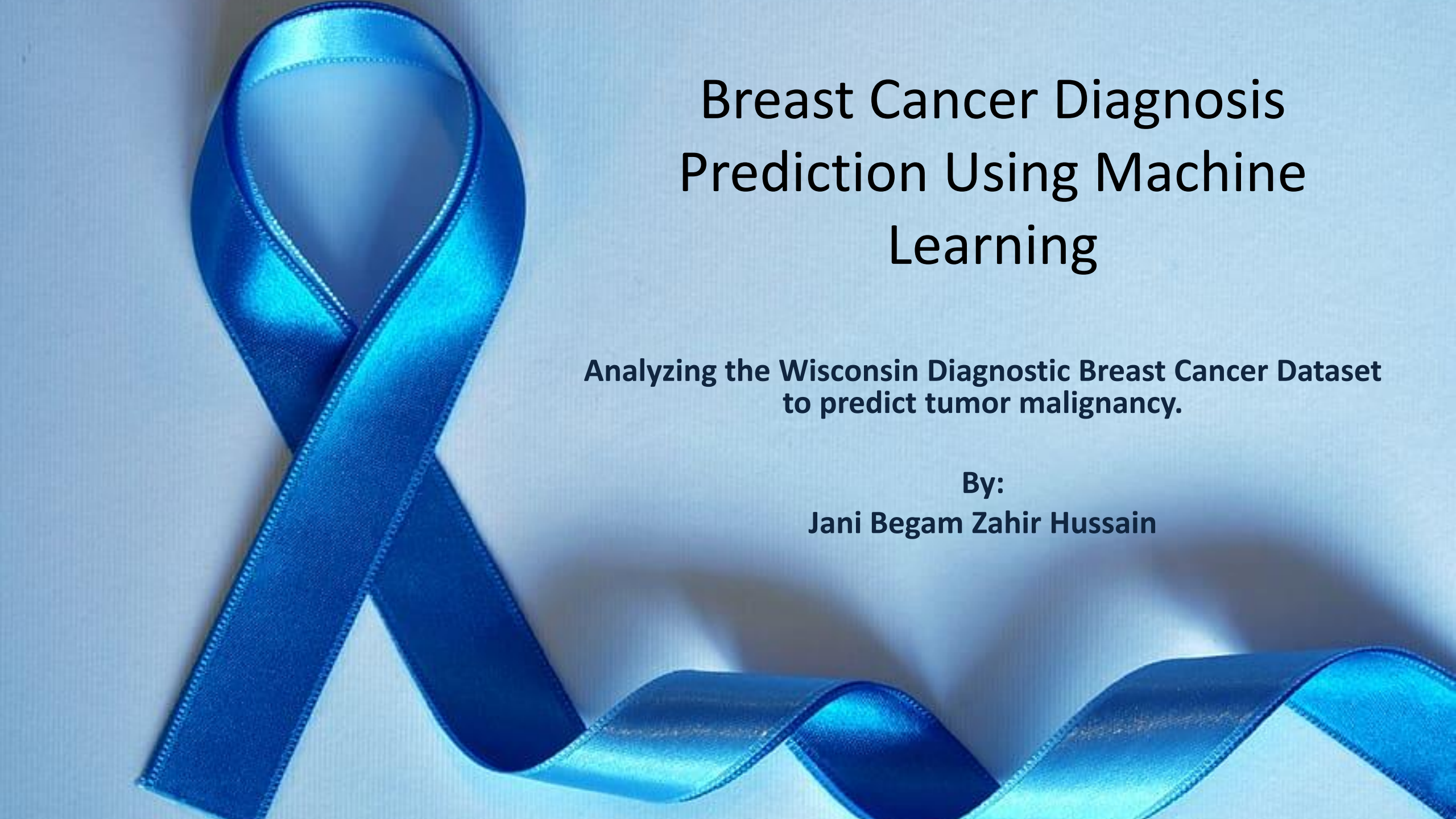


Breast Cancer Diagnosis Prediction Using Machine Learning

Analyzing the Wisconsin Diagnostic Breast Cancer Dataset to predict tumor malignancy.

By:
Jani Begam Zahir Hussain

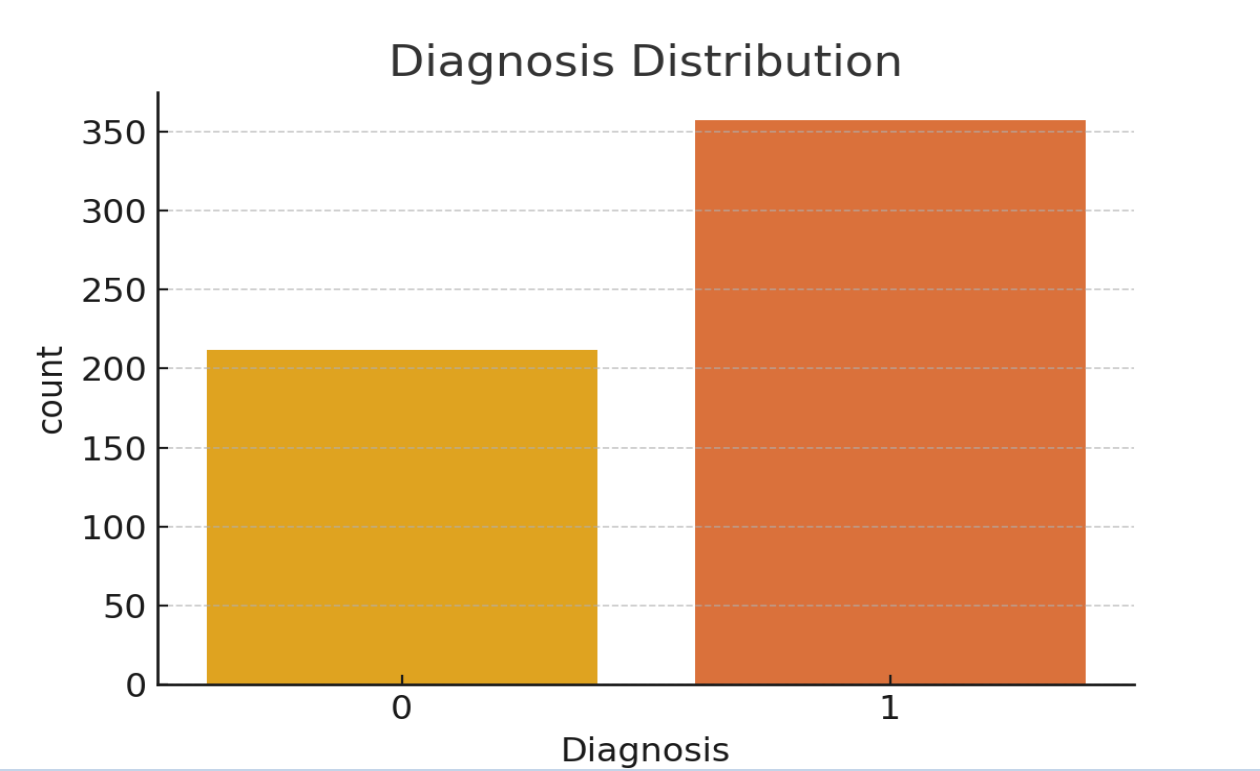
Project Overview

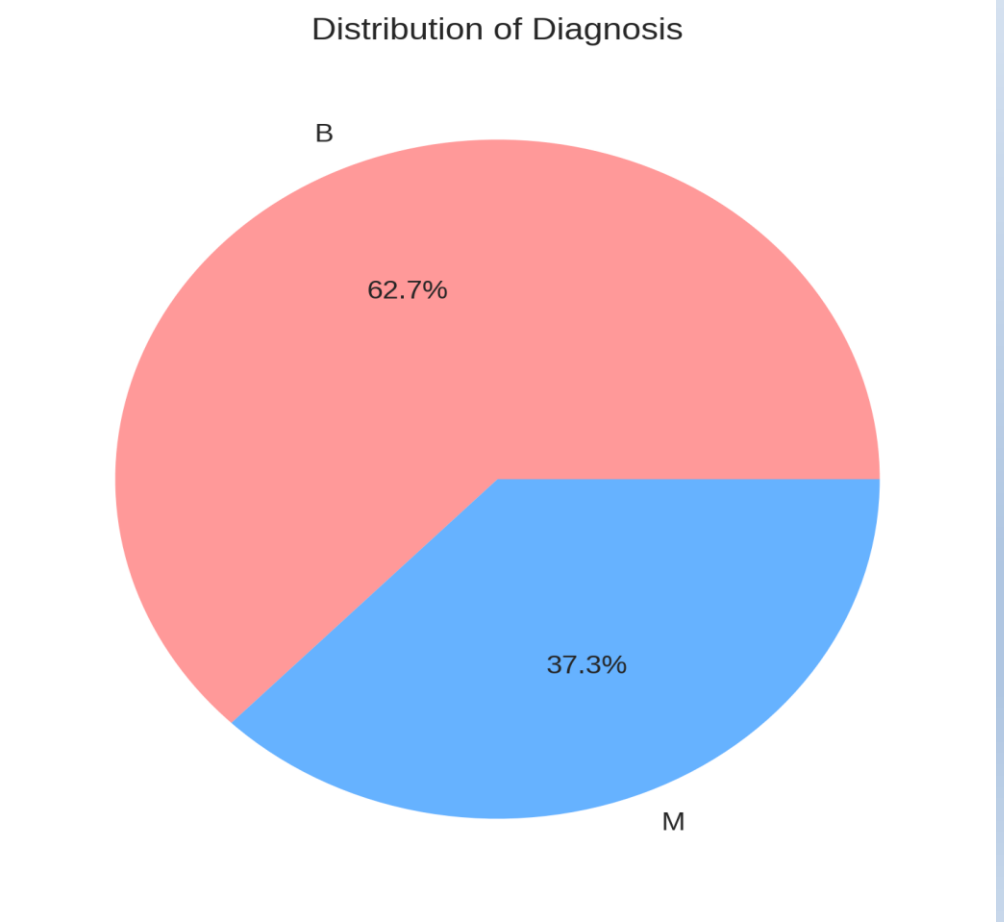
Contents:

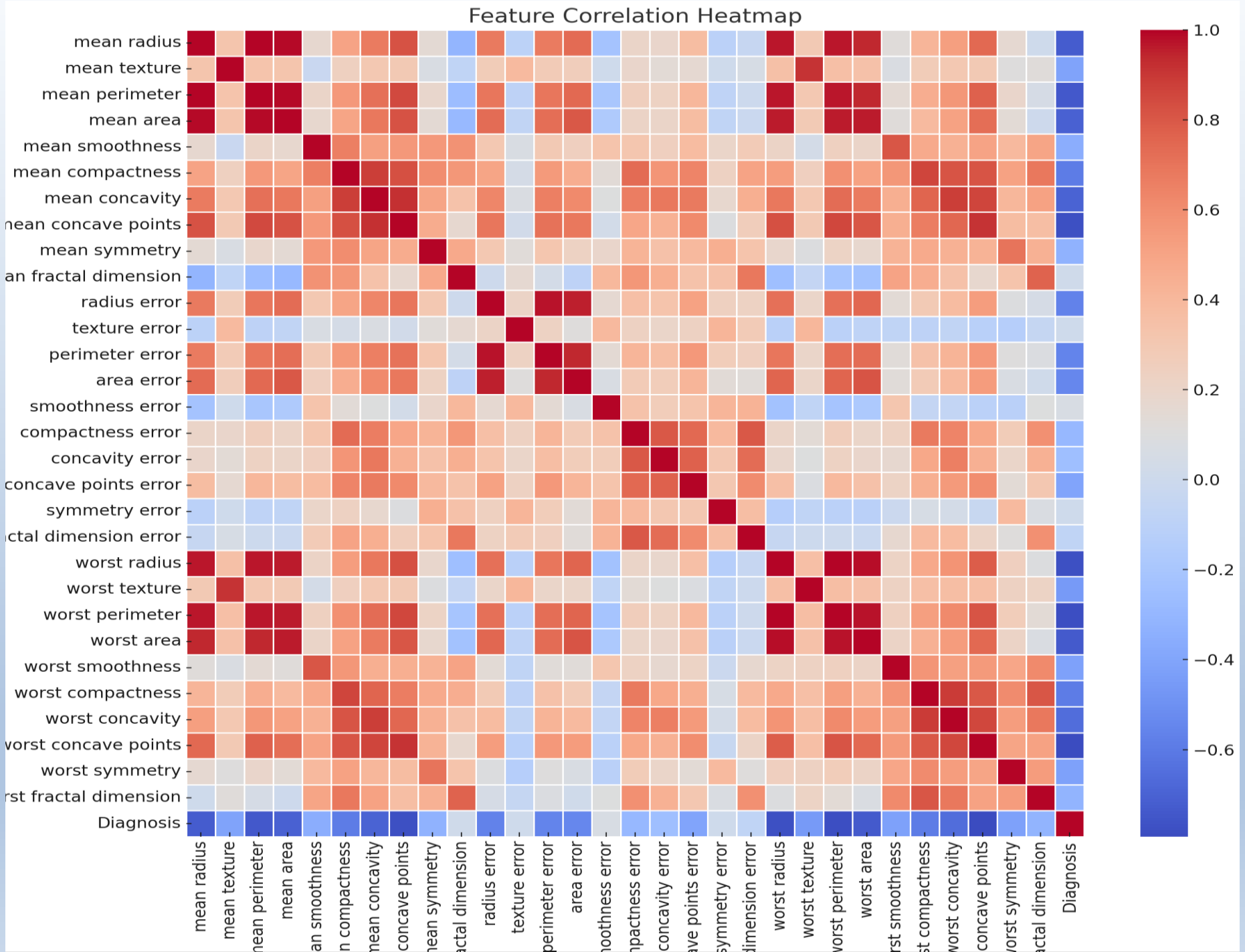
- **Dataset Source**: UCI Machine Learning Repository
- **Total Samples**: 569
- **Features**: 30 real-valued measurements of cell nuclei
- **Target Classes**:
  - 0 = Benign (non-cancerous tumor)
  - 1 = Malignant (cancerous tumor)
- **Objective**: Classify tumor types based on input features.

Diagnosis Distribution

- The dataset has more benign cases than malignant.
- Benign cases: ~62%, Malignant cases: ~38%.
- Important to ensure model generalizes well for minority class (malignant tumors).

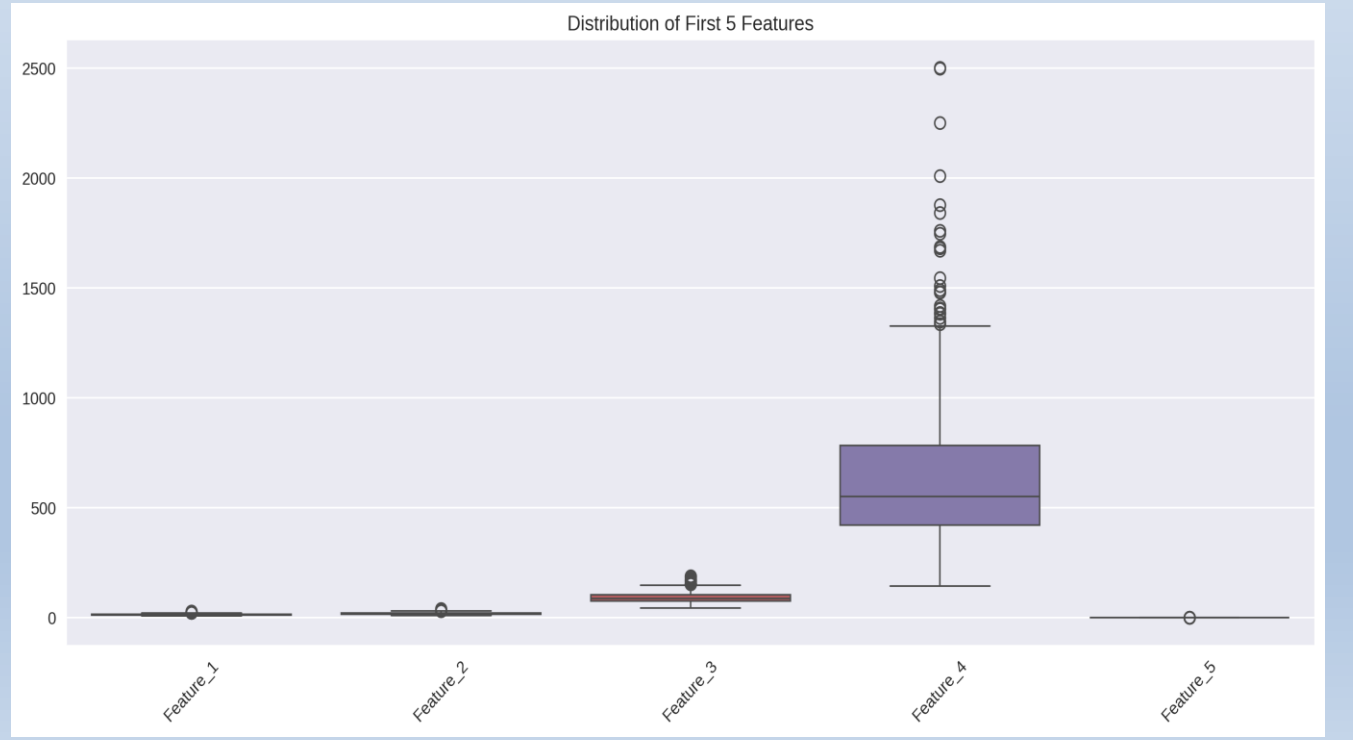


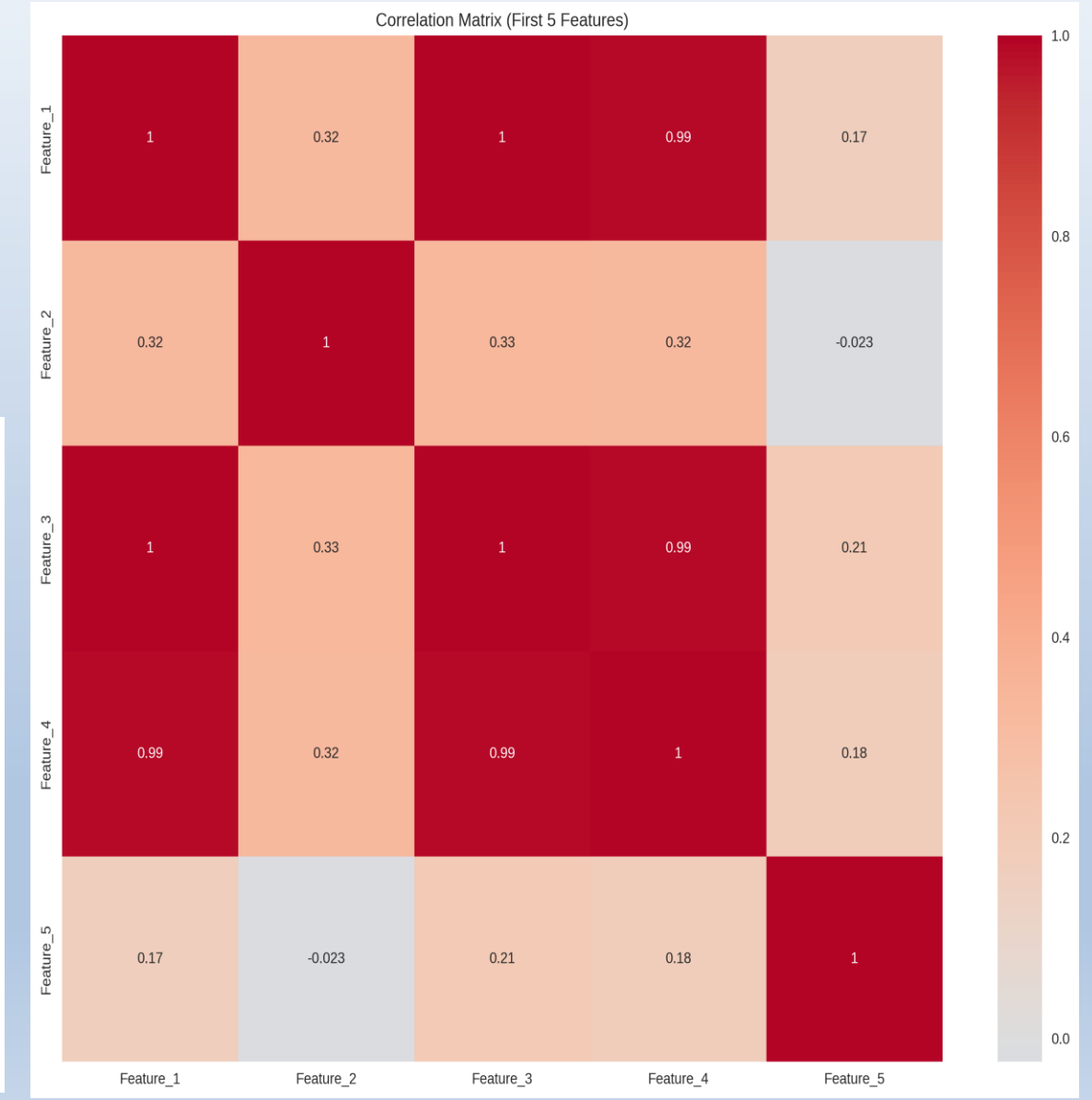



Feature Correlation Heatmap
mean radius
mean texture
mean perimeter
mean area
mean smoothness
mean compactness
mean concavity
nean concave points
mean symmetry
an fractal dimension
radius error
texture error
perimeter error
area error
smoothness error
compactness error
concavity error
concave points error
symmetry error
ctal dimension error
worst radius
worst texture
worst perimeter
worst area
worst smoothness
worst compactness
worst concavity
orst concave points
worst symmetry
rst fractal dimension
Diagnosis
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Feature Correlation Analysis

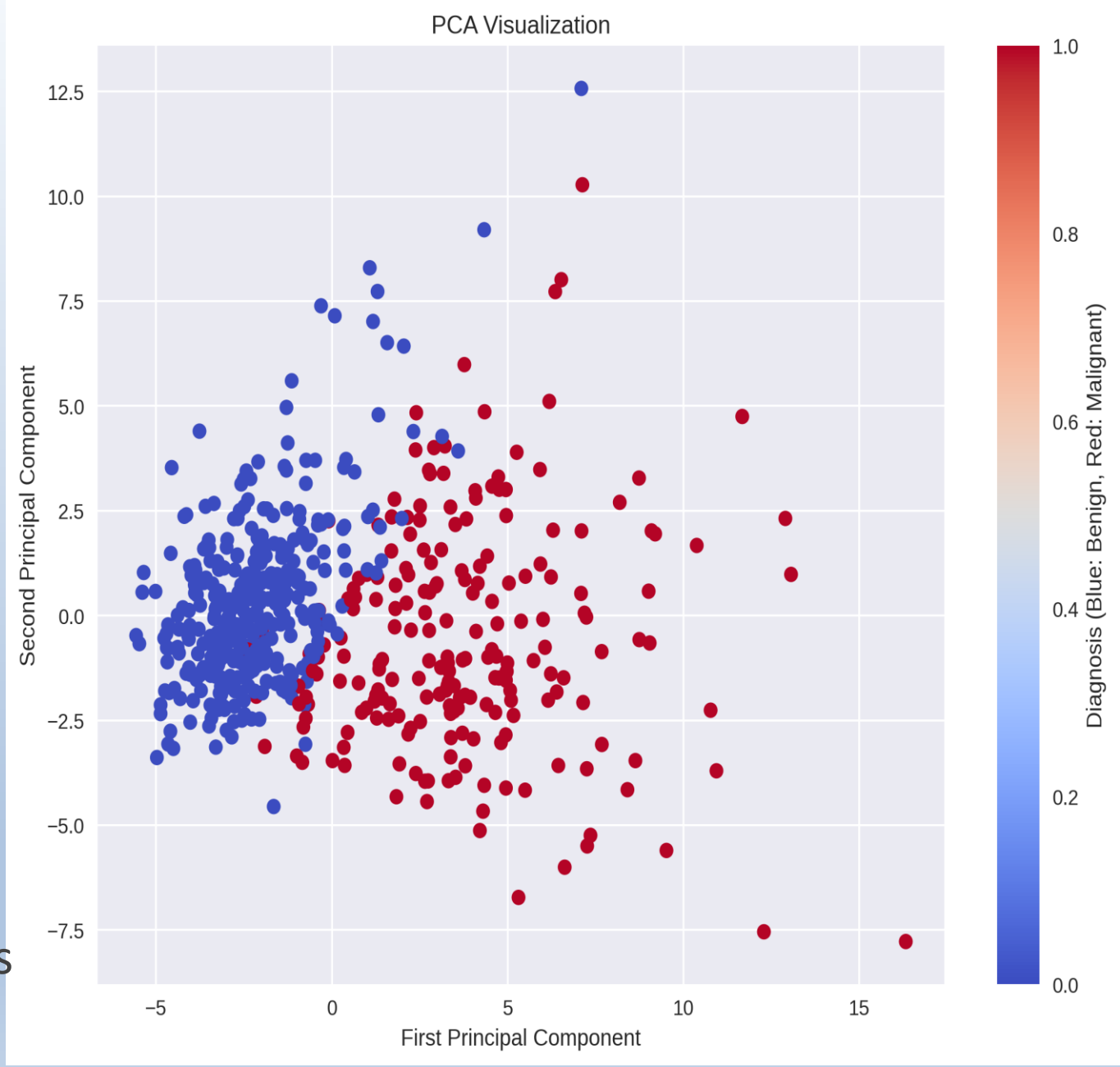
- Strong correlations between Radius, Perimeter, and Area features.
- Some features are redundant; feature selection could enhance model simplicity.
- Multicollinearity observed — especially for size-related measurements.





PCA Visualization

- A scatter plot created by reducing the dimensionality of the features via Principal Components Analysis (PCA), where points are colored by diagnosis.
- **Interpretation:** The PCA visualization reveals how well the complex, high-dimensional data can be separated in a two-dimensional plot. A clear separation between malignant and benign clusters implies that the features have discriminatory power.

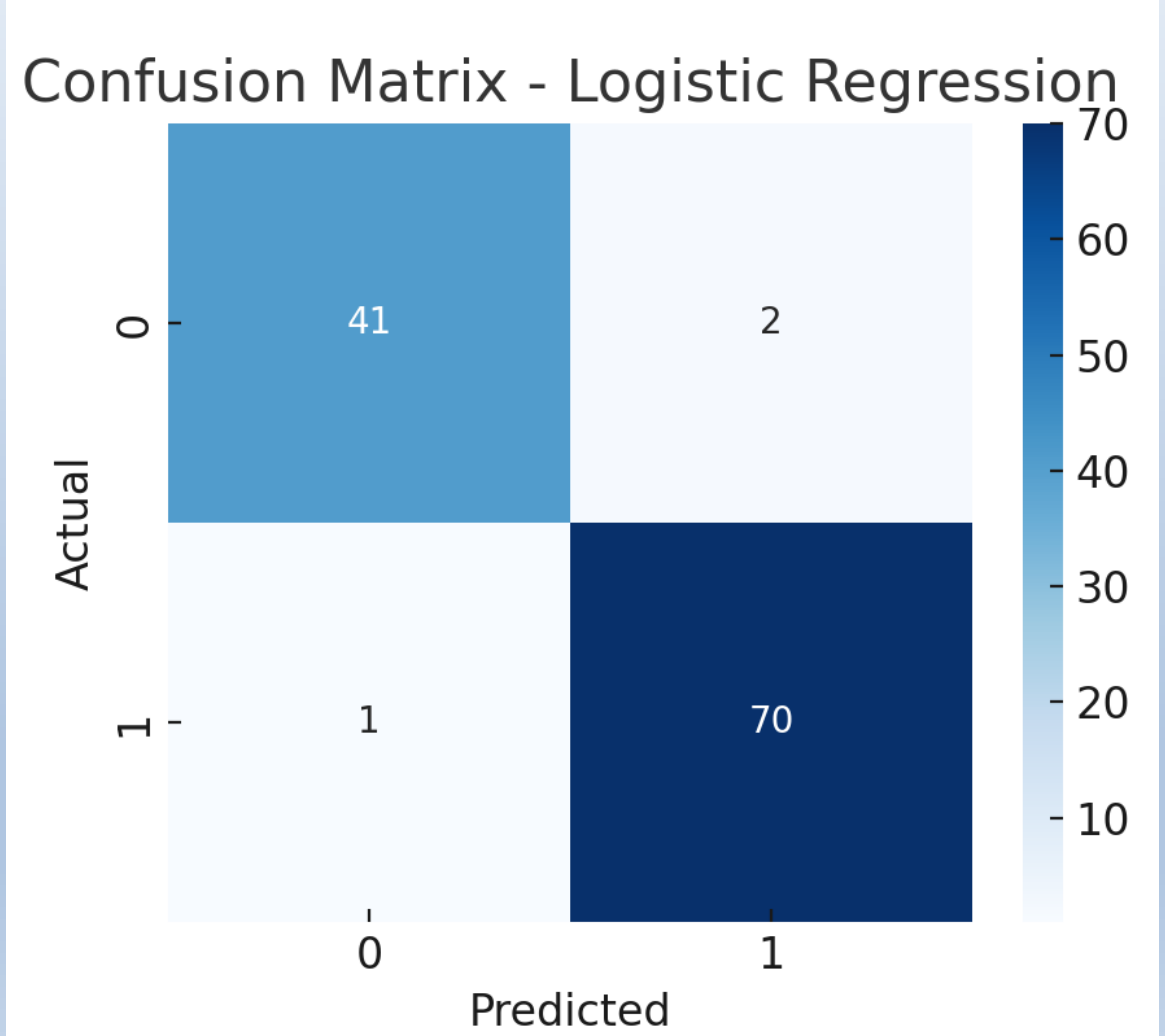


Model Building

- **Algorithms Used**:
- Logistic Regression
- Random Forest Classifier
- Support Vector Machine (SVM)
- **Data Processing**:
- Standardized features using StandardScaler.
- Train-Test split: 80% training, 20% testing.
- **Goal**: Maximize classification performance while minimizing errors.

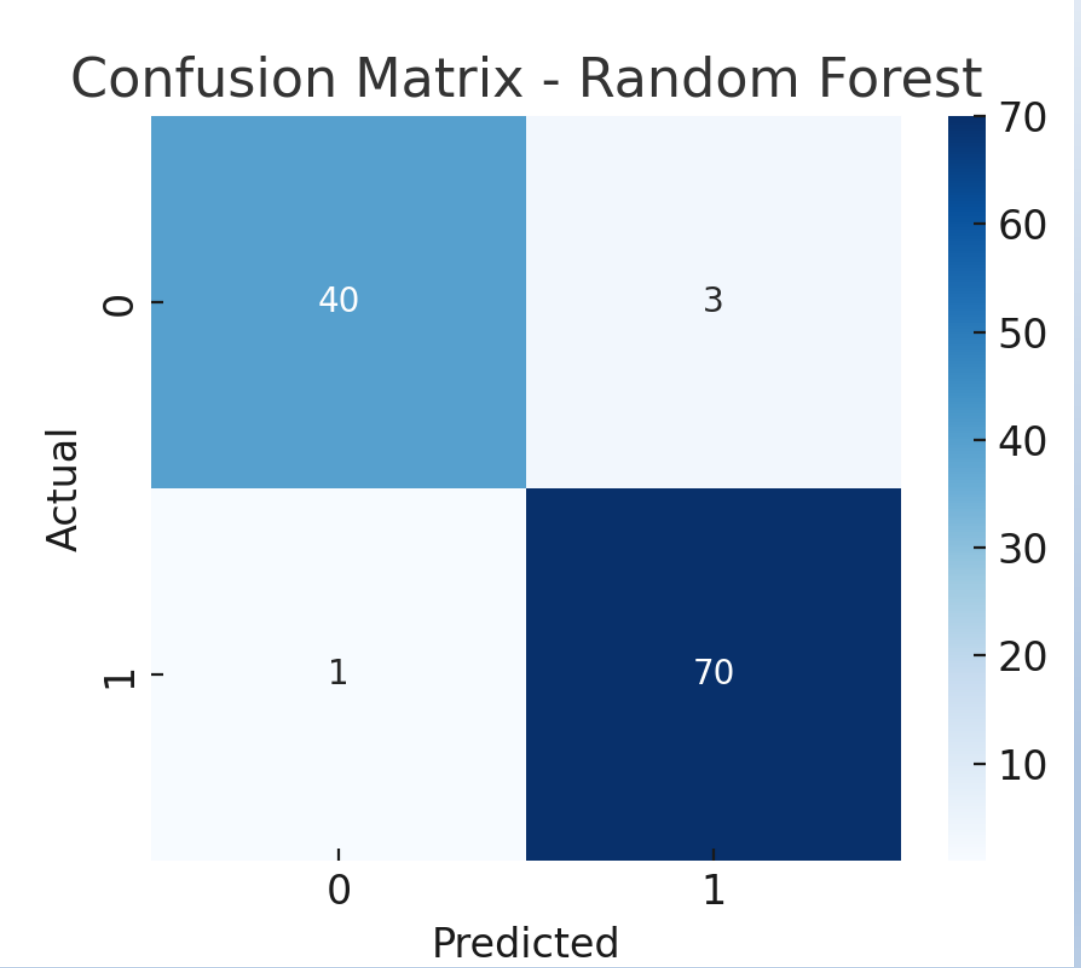
Confusion Matrix - Logistic Regression

- Logistic Regression achieved high accuracy.
- A few malignant cases misclassified as benign (needs careful monitoring).
- Simple model with fast training and good interpretability.



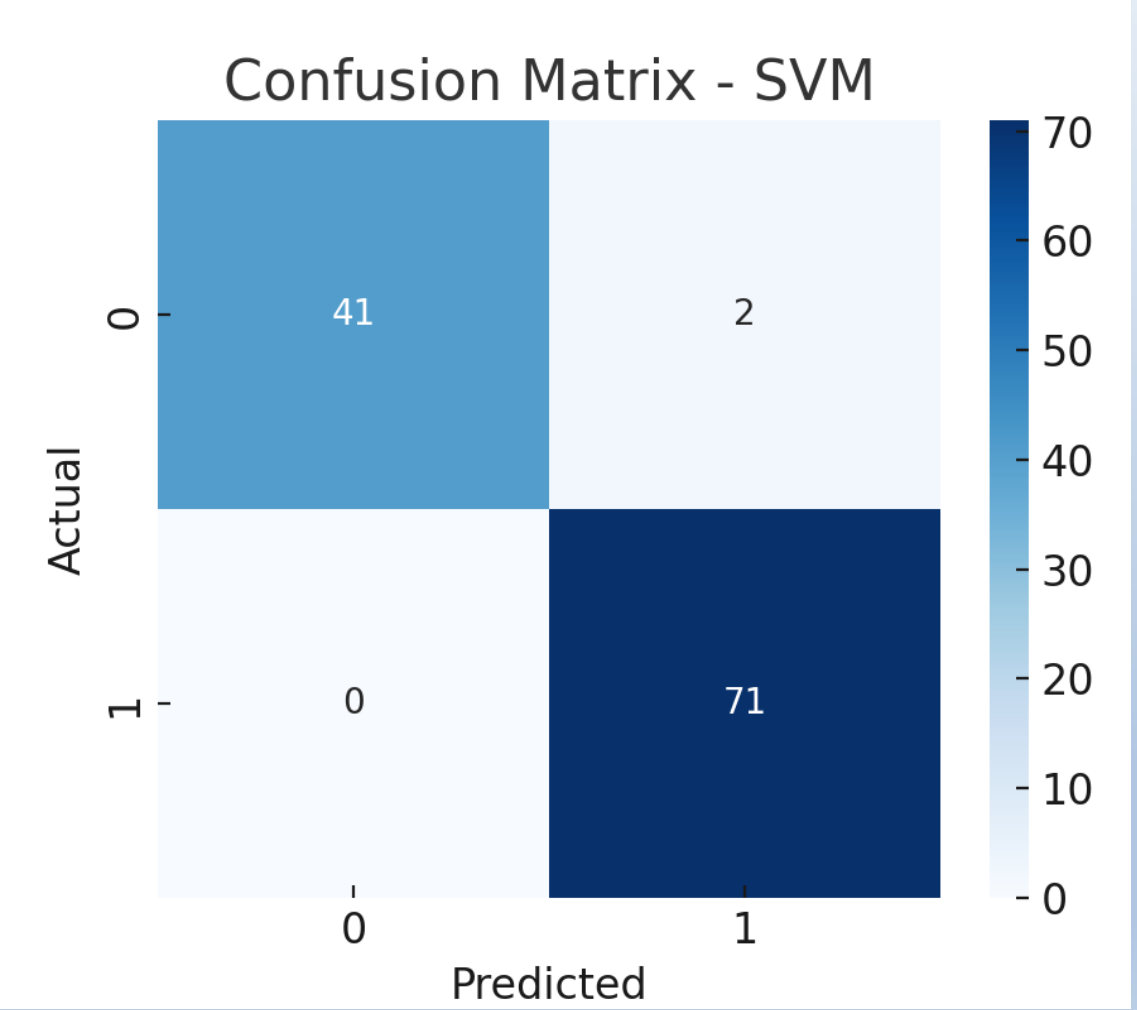
Confusion Matrix - Random Forest

- Random Forest achieved near-perfect classification.
- Very few errors — extremely good generalization on test data.
- Handles feature interactions and non-linearity automatically.



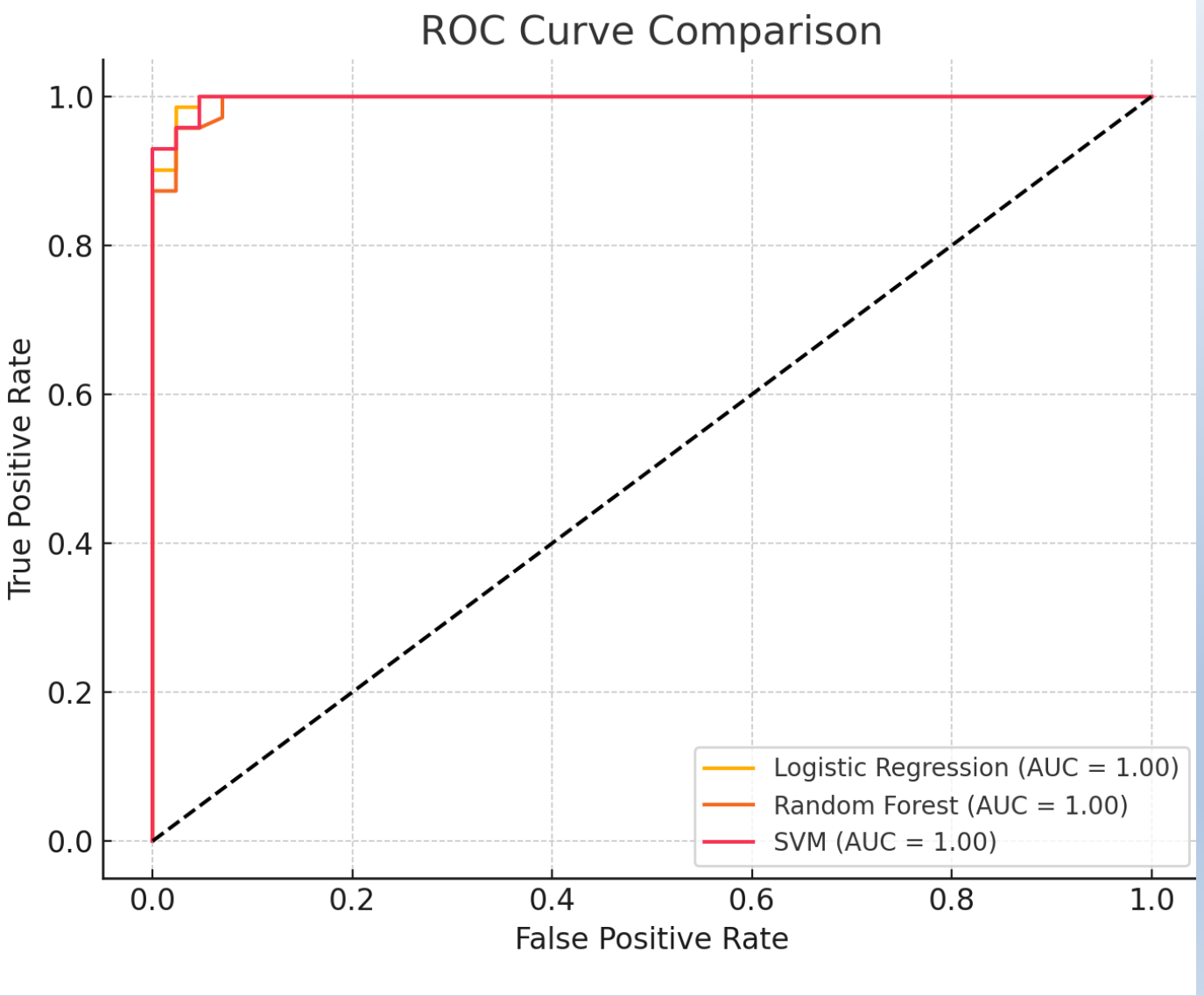
Confusion Matrix - SVM

- SVM model delivered almost perfect separation.
- Only minimal false positives and false negatives.
- Particularly effective after scaling due to sensitivity to feature magnitude.



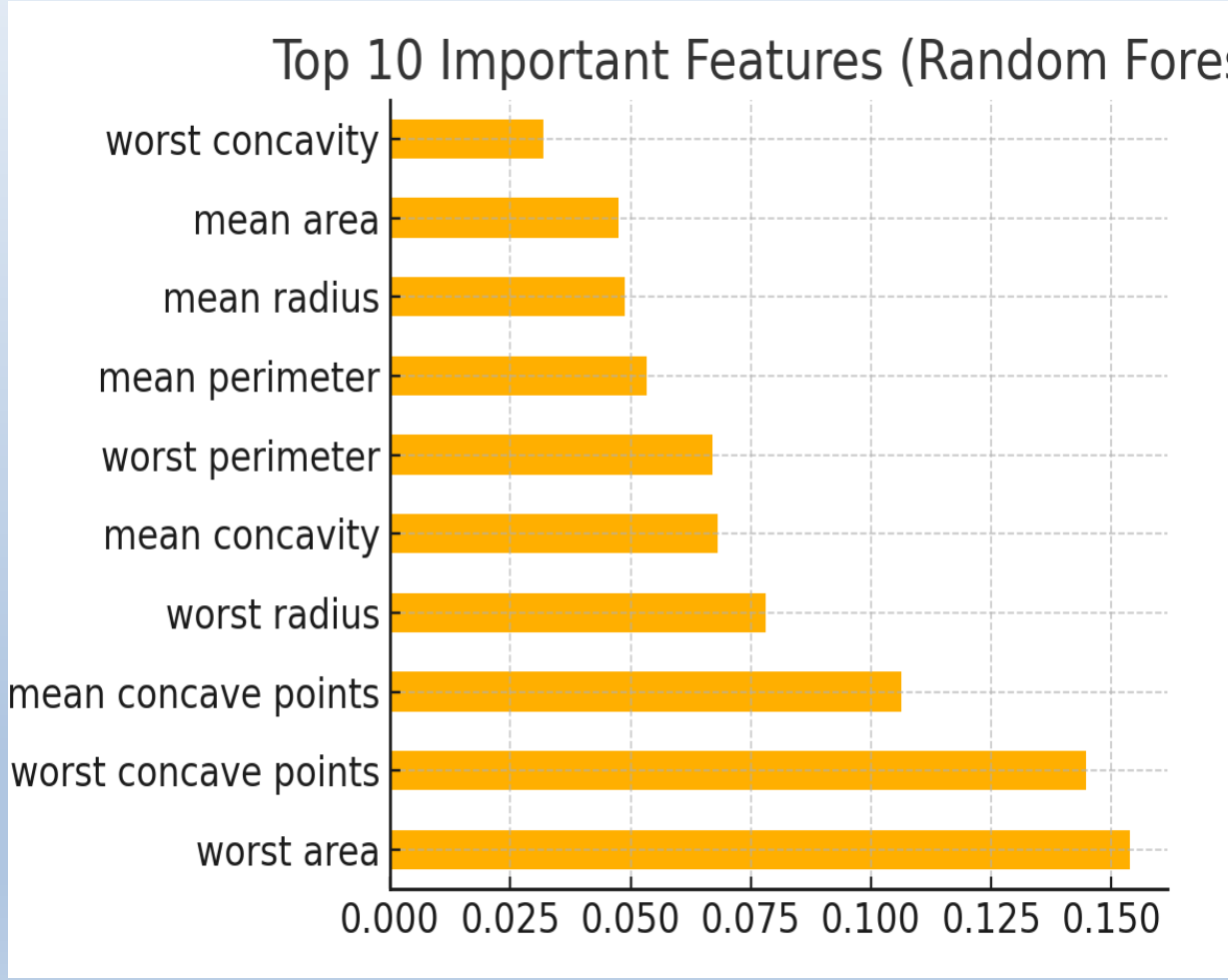
ROC Curve Comparison

- All models achieved high AUC scores (> 0.98).
- SVM and Random Forest slightly outperform Logistic Regression.
- ROC curves confirm excellent model discrimination between malignant and benign.

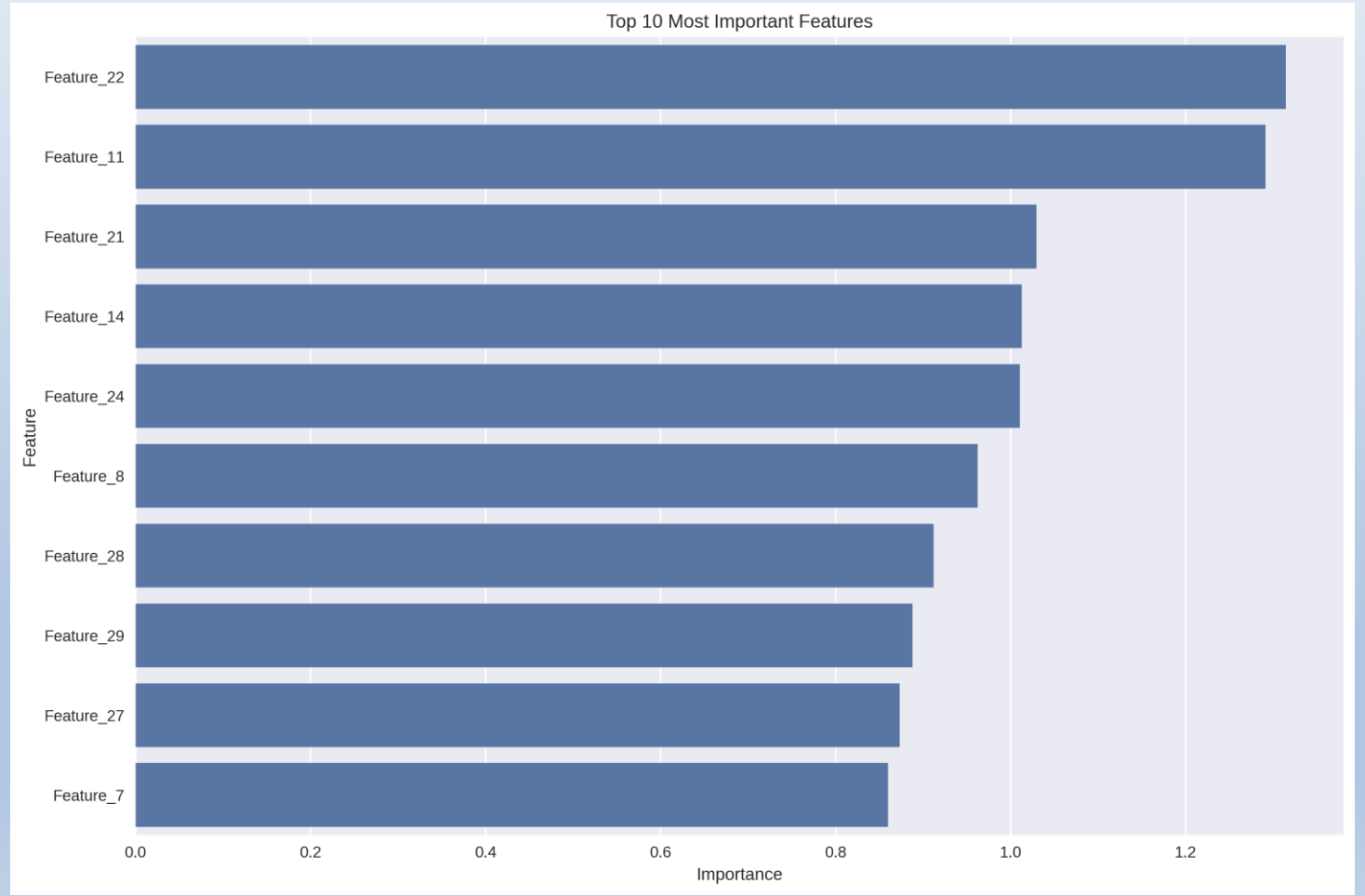


Feature Importance

- Top predictive features:
- Radius Worst
- Perimeter Worst
- Area Worst
- Concavity Worst
- Worst-case tumor measurements are more predictive than mean or standard error metrics.
- Feature importance can guide dimensionality reduction efforts.



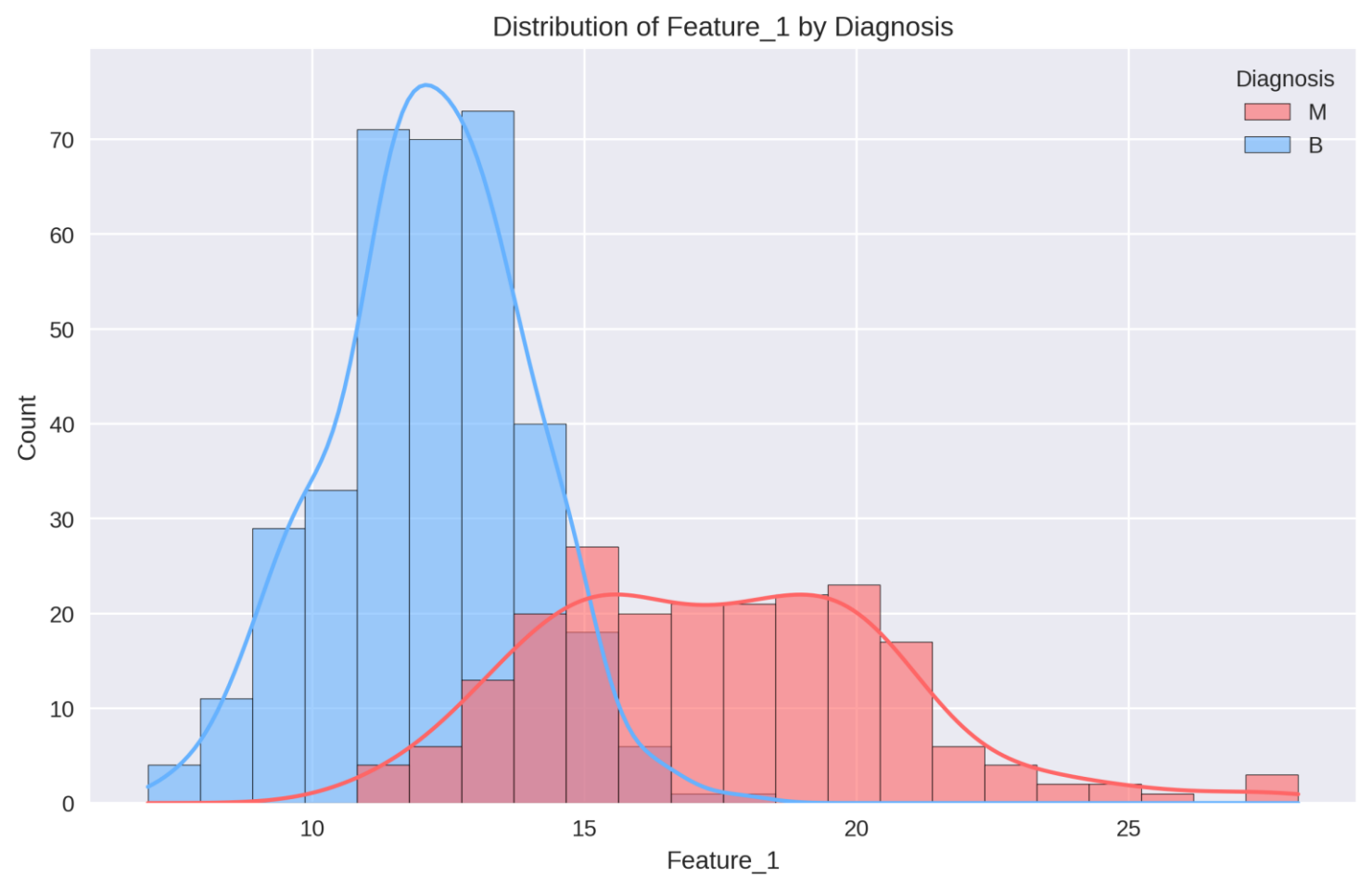
Top 10 Most Important Features



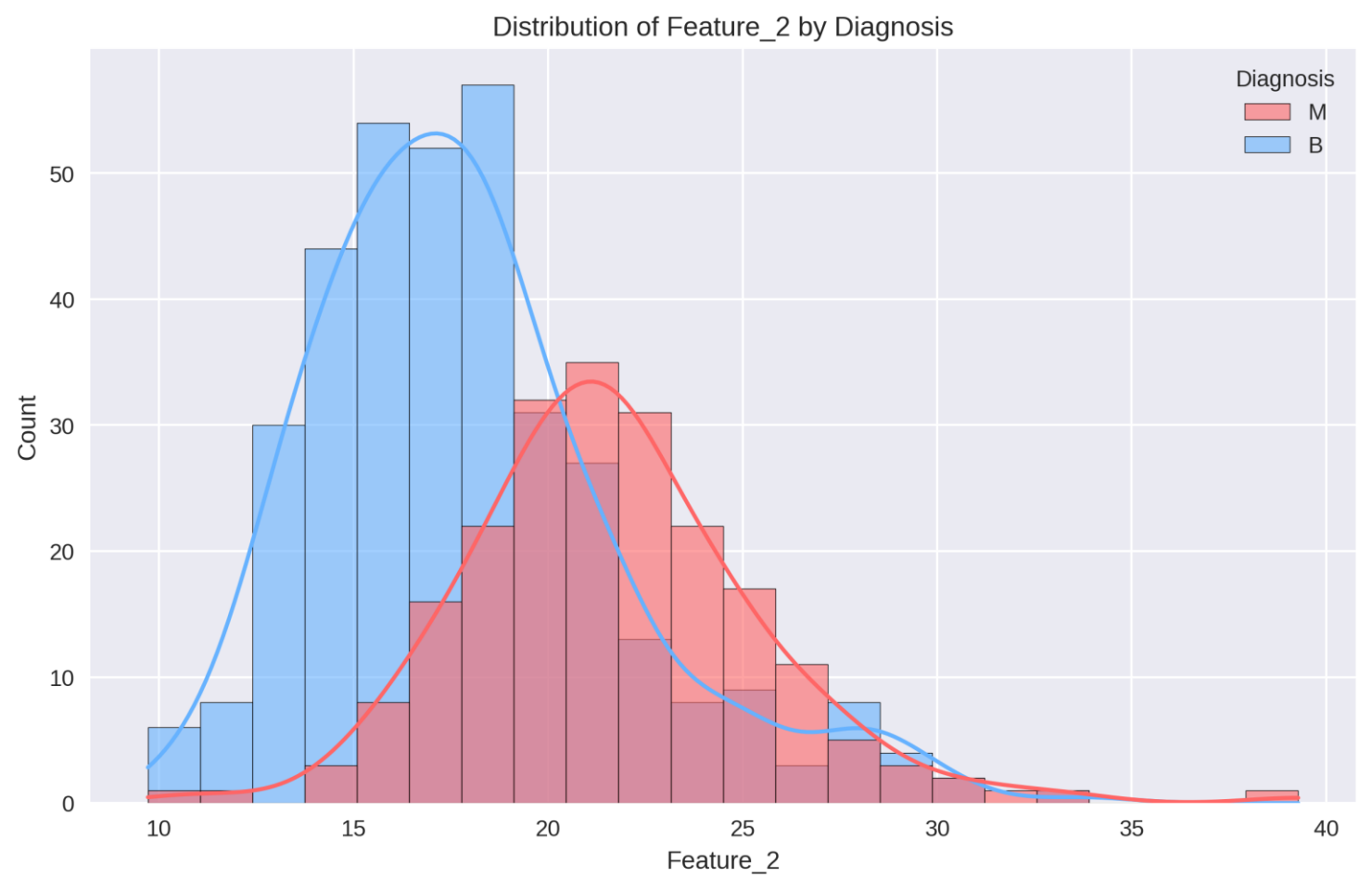
Conclusion

- Machine learning models successfully predict breast cancer diagnosis.
- SVM and Random Forest models are highly reliable.
- Critical features involve size and shape irregularities of tumors.
- Future directions:
- Hyperparameter tuning
- Cross-validation
- Further feature engineering to improve robustness

Distribution of Feature_1



Distribution of Feature_2



Distribution of Feature_3