# Project 2: Breast Cancer Prediction Documentation

## Data Preprocessing

1. **Data Loading**:
   - Loaded the dataset from data.csv.
   - Inspected the first and last few rows, dataset information, and summary statistics.
2. **Handling Missing Values**:
   - Removed the column Unnamed: 32 as it contained no useful data.
   - Dropped the id column since it was not useful for analysis.
3. **Outliers Removal**:
   - Calculated the Interquartile Range (IQR) for each numerical feature.
   - Filtered out rows where feature values were outside the bounds defined by the IQR.
4. **Data Type Conversion and Scaling**:
   - Converted the categorical diagnosis column into numerical values (Malignant: 1, Benign: 0).
   - Standardized numerical features using StandardScaler to normalize their scales.

## Feature Engineering

1. **New Features Created**:
   - **radius_texture_mean**: Ratio of radius_mean to texture_mean.
   - **area_perimeter_mean**: Ratio of area_mean to perimeter_mean.
   - **smoothness_compactness_mean**: Product of smoothness_mean and compactness_mean.
   - **radius_worst_area_worst**: Ratio of radius_worst to area_worst.
   - **compactness_concavity_mean**: Product of compactness_mean and concavity_mean.

## Feature Selection

1. **Correlation Analysis**:
   - Selected features based on their correlation with the diagnosis column.
   - Used a correlation threshold of 0.5 to determine relevance.
2. **Selected Features**:
   - Included features with an absolute correlation greater than the threshold.

## Machine Learning Model

1. **Model Training and Tuning**:

   - Used Support Vector Machine (SVM) with hyperparameter tuning via GridSearchCV.
   - Optimized hyperparameters: C, kernel, and gamma.
2. **Performance Metrics**:
   - **Accuracy**: 91%
   - **Confusion Matrix**:
     - True Positives (Malignant): 27
     - True Negatives (Benign): 82
     - False Positives: 4
     - False Negatives: 7
   - **Classification Report**:
     - Precision, Recall, and F1-score for both classes (Benign and Malignant).

## Challenges and Observations

1. **Challenges**:
   - Handling outliers effectively while maintaining data integrity.
   - Deciding on the appropriate correlation threshold for feature selection.
2. **Observations**:
   - The model achieved a high accuracy of 91%, indicating strong performance.
   - The confusion matrix and classification report highlighted good precision and recall, with minor improvements needed in predicting malignant cases.