# **Breast Cancer Prediction Project Documentation**

## **1. Data Preprocessing**

### **1.1 Dataset Overview**

The Breast Cancer Wisconsin (Diagnostic) dataset is used for this project. It contains features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. The goal is to predict whether the tumor is malignant or benign.

### **1.2 Data Cleaning**

* Handling Missing Values: Checked for missing values in the dataset. Imputed missing values using appropriate methods, such as mean or median imputation.
* Outlier Detection: Identified outliers and decided on appropriate strategies, such as removal or transformation, to handle them.
* Handling Inconsistencies: Ensured consistency in data types, resolved any inconsistencies, and verified the integrity of the dataset.

## **2. Feature Selection and Engineering**

### **2.1 Feature Identification**

* Relevant Features: Identified relevant features based on domain knowledge and exploratory data analysis. Features like radius, texture, and smoothness are considered important in breast cancer diagnosis.

### **2.2 Feature Engineering**

* Creation of New Features: Derived new features or applied transformations to existing ones. For instance, creating a feature representing the ratio of perimeter to area.

## **3. Machine Learning Model (SVM)**

### **3.1 Model Selection**

* Support Vector Machine (SVM): Chose SVM as the classification model due to its effectiveness in binary classification tasks and robustness in handling non-linear relationships.

### **3.2 Model Training and Evaluation**

* Data Splitting: Split the dataset into training and testing sets to train the model and evaluate its performance.
* Model Training: Trained the SVM model using the training data.
* Evaluation Metrics: Utilized metrics such as accuracy, precision, recall, and F1-score to assess the model's performance on the test set.
* Hyperparameter Tuning: Conducted hyperparameter tuning to optimize the SVM model for better performance.