James Walton

MSBD 566 – Predictive Modeling and Analytics

10/8/2025

Assignment #2 – Pre-Midterm Project

**Data Title:** Breast Cancer Wisconsin (Diagnostic) Dataset

**Source:** UCI Machine Learning Repository

**Source Link:**  <https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>

**Data Description**: The Breast Cancer Wisconsin (Diagnostic) dataset contains records from fine-needle aspirate (FNA) samples of breast tissue analyzed at the University of Wisconsin Hospital. Each sample is described by 30 numeric measurements that characterize the size, texture, and shape of cell nuclei observed under the microscope.

The data also include an identifying case number and a categorical diagnosis, **benign** or **malignant,** as determined by histopathologic review. The variables capture both central tendency (e.g., mean radius, mean texture) and variation (e.g., standard error, worst values) for each morphological feature.

**Data Dictionary**:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| ID | Categorical | Sample identifier |
| Diagnosis | Categorical | M = Malignant, B = Benign |
| Radius\_Mean | Numeric | Average radius of cell nuclei |
| Texture\_Mean | Numeric | Standard deviation of gray-scale values |
| Smoothness\_Mean | Numeric | Local variation in radius lengths |
| Concavity\_Mean | Numeric | Severity of concave portions |