**MSBD566 – Predictive Modeling and Analysis**

Name: James Walton

Date: 10.20.2025

Assignment: Midterm Project

## Project Description

This project focuses on classifying breast cancer tumors as benign or malignant using diagnostic measurements collected from fine-needle aspirate (FNA) samples of breast tissue. The analysis is based on the Breast Cancer Wisconsin (Diagnostic) dataset, which includes 30 numerical features describing the characteristics of cell nuclei. A Random Forest Classifier was used to build the predictive model, chosen for its reliability and ability to handle complex data relationships. The goal of this work is to support early detection of breast cancer and provide insights that can assist in medical decision-making by identifying patterns associated with malignant cases.

## Data Description

The dataset used in this project is the **Breast Cancer Wisconsin (Diagnostic)** dataset, obtained from the [UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic). It contains diagnostic data from fine-needle aspirate (FNA) samples of breast tissue, analyzed at the University of Wisconsin Hospital. Each record includes 30 numerical features that describe the characteristics of cell nuclei—such as radius, texture, smoothness, compactness, and concavity—measured from digitized images of the samples. The dataset also includes an identifying case number and a categorical diagnosis label indicating whether the tumor is **benign** or **malignant**. The data are clean, balanced, and suitable for building and evaluating classification models for medical diagnostics.

**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 |

## Method and Analysis

The analysis began with loading and exploring the Breast Cancer Wisconsin (Diagnostic) dataset to understand its structure and check for missing values or inconsistencies. The categorical diagnosis variable was converted to a binary form, where malignant (M) was encoded as 1 and benign (B) as 0. All 30 numeric diagnostic features were used as predictors, representing various measurements of the cell nuclei. The data were split into training (80%) and testing (20%) sets using stratified sampling to preserve class balance. Features were standardized using a z-score scaler to ensure consistent scale across variables. A Random Forest Classifier was chosen as the modeling method due to its robustness, interpretability, and ability to handle non-linear relationships. The model was trained on the scaled training data and evaluated on the test set. Performance was assessed using metrics such as accuracy, precision, recall, and F1-score, along with a confusion matrix to visualize classification results. The confusion matrix helped illustrate how well the model distinguished between benign and malignant cases, highlighting the model’s reliability in predicting breast cancer outcomes.

A graph showing a number of forest classification

AI-generated content may be incorrect.

The feature importance plot highlights the variables that contributed most to the Random Forest model’s predictions. Features such as **radius\_worst**, **area\_mean**, and **concavity\_mean** ranked among the most influential in distinguishing between malignant and benign tumors. These attributes describe the size and shape irregularities of the cell nuclei, which are known diagnostic indicators of cancerous growth. The results suggest that the model correctly emphasizes features that align with medical understanding, larger and more irregular nuclei are typically associated with malignancy. This analysis helps confirm that the Random Forest model not only performs accurately but also bases its decisions on clinically meaningful patterns.

A graph with blue and white bars

AI-generated content may be incorrect.

## Evaluation

The Random Forest Classifier achieved strong and consistent performance on the test data, demonstrating high predictive accuracy and balanced results across both benign and malignant cases. The table below summarizes the key evaluation metrics obtained from the model’s classification report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **Benign (0)** | 0.97 | 0.99 | 0.98 | 71 |
| **Malignant (1)** | 0.99 | 0.96 | 0.98 | 43 |
| **Overall Accuracy** | **0.98** | – | – | 114 |

* **Precision** measures how many of the cases predicted as malignant were actually malignant. A score of **0.99** for malignant tumors means the model rarely misclassifies benign cases as malignant.
* **Recall (Sensitivity)** indicates how many of the actual malignant cases were correctly identified. A recall of **0.96** means the model successfully detected 96% of all true cancer cases.
* **F1-Score** is the mean of precision and recall, balancing both false positives and false negatives. An F1-score of **0.98** reflects excellent overall classification consistency.
* **Accuracy** represents the percentage of total correct predictions. The overall accuracy of **98%** shows that the model correctly classified nearly all cases in the test dataset.

The **confusion matrix** visualization confirmed these results, with only a few instances of misclassification between benign and malignant samples. This indicates that the Random Forest model performs reliably in distinguishing between cancerous and non-cancerous cases. In a medical context, such high precision and recall are critical for supporting early detection while minimizing diagnostic errors. Overall, the model demonstrates strong potential for use in predictive healthcare analytics and diagnostic support systems.