**Project Overview:**

This project applies supervised machine learning algorithms to classify breast cancer tumors as **malignant (M)** or **benign (B)** using the **Breast Cancer Wisconsin Dataset**. The models used include:

* Decision Tree
* Naive Bayes
* Neural Network (MLPClassifier)

**Tools and Technologies Used:**

| **Tool/Library** | **Purpose** |
| --- | --- |
| Python | Programming language used for the project |
| pandas | Data manipulation and loading CSV files |
| scikit-learn | Building ML models, preprocessing, and evaluation |
| StandardScaler | Feature normalization (standard scaling) |
| LabelEncoder | Encoding categorical target labels |
| train\_test\_split | Dividing data into training and testing sets |
| DecisionTreeClassifier | Tree-based classification model |
| GaussianNB | Naive Bayes model assuming Gaussian distribution |
| MLPClassifier | Multi-layer perceptron (neural network) |
| accuracy\_score | To evaluate the performance of each classifier |

**Dataset Description:**

* **Name:** breast-cancer.csv
* **Target Column:** diagnosis
  + M = Malignant (encoded as 1)
  + B = Benign (encoded as 0)

**Preprocessing Steps:**

1. Removed id column – It's a unique identifier, not useful for prediction.
2. Encoded diagnosis – Converted 'M' and 'B' into numeric values.
3. Separated Features and Target:
   * X: All columns except diagnosis
   * y: Only diagnosis
4. Feature Normalization:
   * Applied StandardScaler to bring all feature values to the same scale (mean = 0, std = 1).
5. Train-Test Split:
   * 80% training, 20% testing using train\_test\_split.

**Machine Learning Models Used:**

**1. Decision Tree Classifier**

* A flowchart-like structure used for decision making.
* Trained using entropy or Gini impurity.
* Result:  
  Accuracy = 95%

**2. Naive Bayes Classifier (GaussianNB)**

* Based on Bayes' Theorem.
* Assumes features are independent and normally distributed.
* Result:  
  Accuracy = 96%

**3. Neural Network (MLPClassifier)**

* Multi-layer Perceptron (MLP) used for classification.
* One hidden layer with 100 neurons.
* Activation function: ReLU (by default).
* Result:  
  Accuracy = 97%

**Results Summary:**

| **Model** | **Type** | **Accuracy** |
| --- | --- | --- |
| Decision Tree | Supervised | 95% |
| Naive Bayes | Supervised | 96% |
| Neural Network | Supervised | 97% |

**Conclusion:**

* Neural Network (MLPClassifier) performed the best with 97% accuracy.
* Naive Bayes was the simplest and still performed excellently.
* Decision Tree is interpretable and easy to visualize.
* This project shows that preprocessing and choosing the right model significantly affects the classification accuracy in medical data.