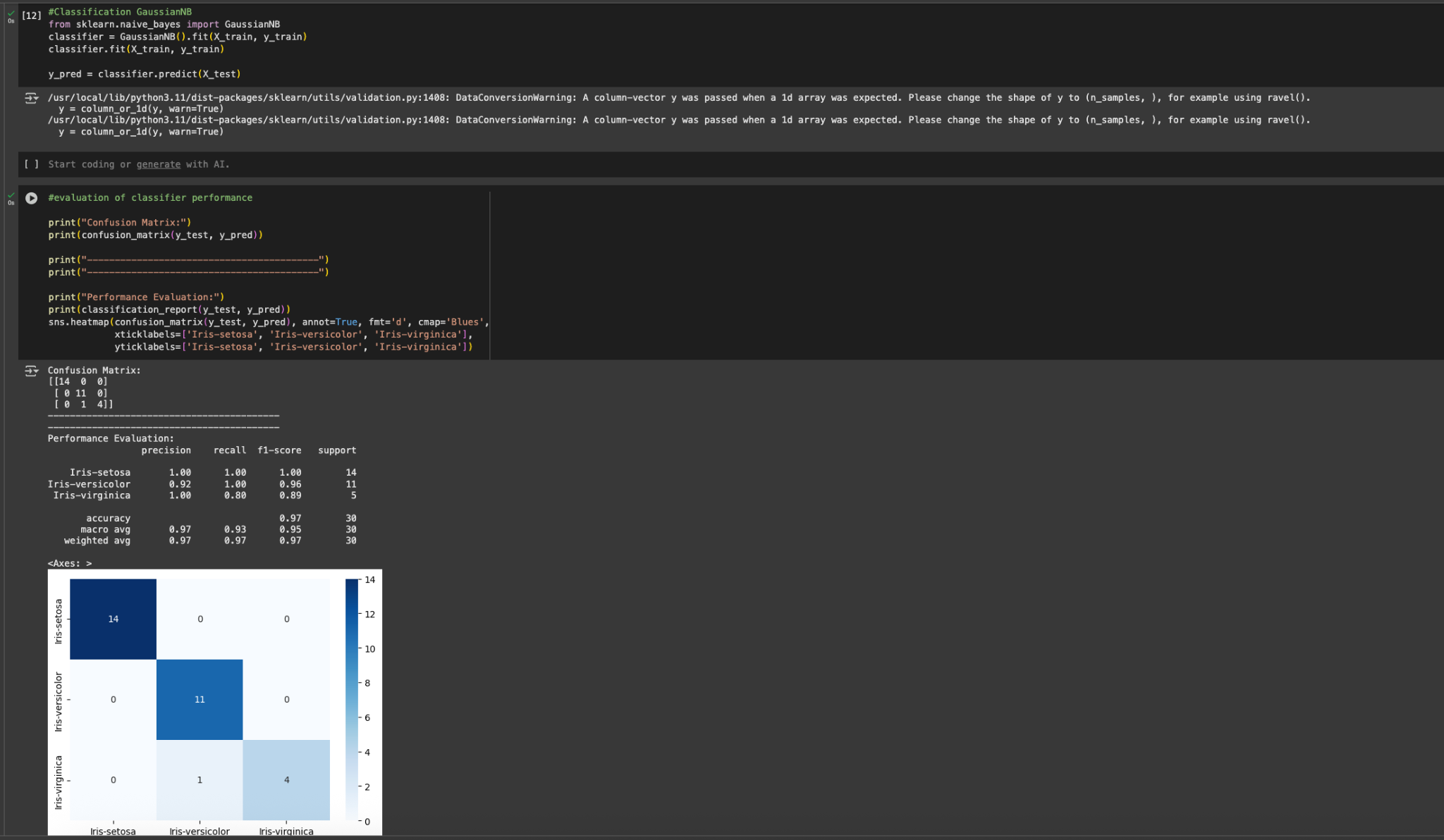
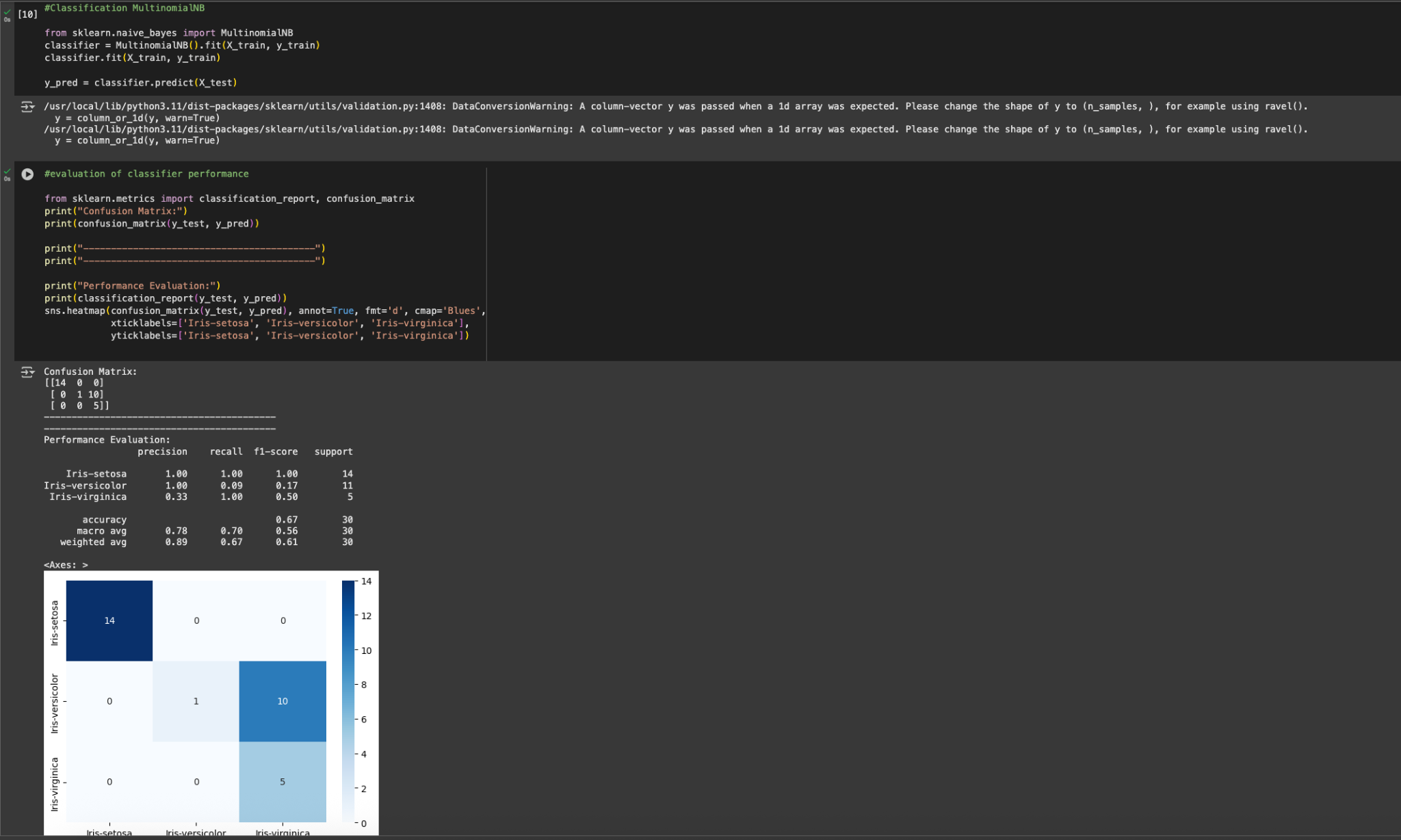
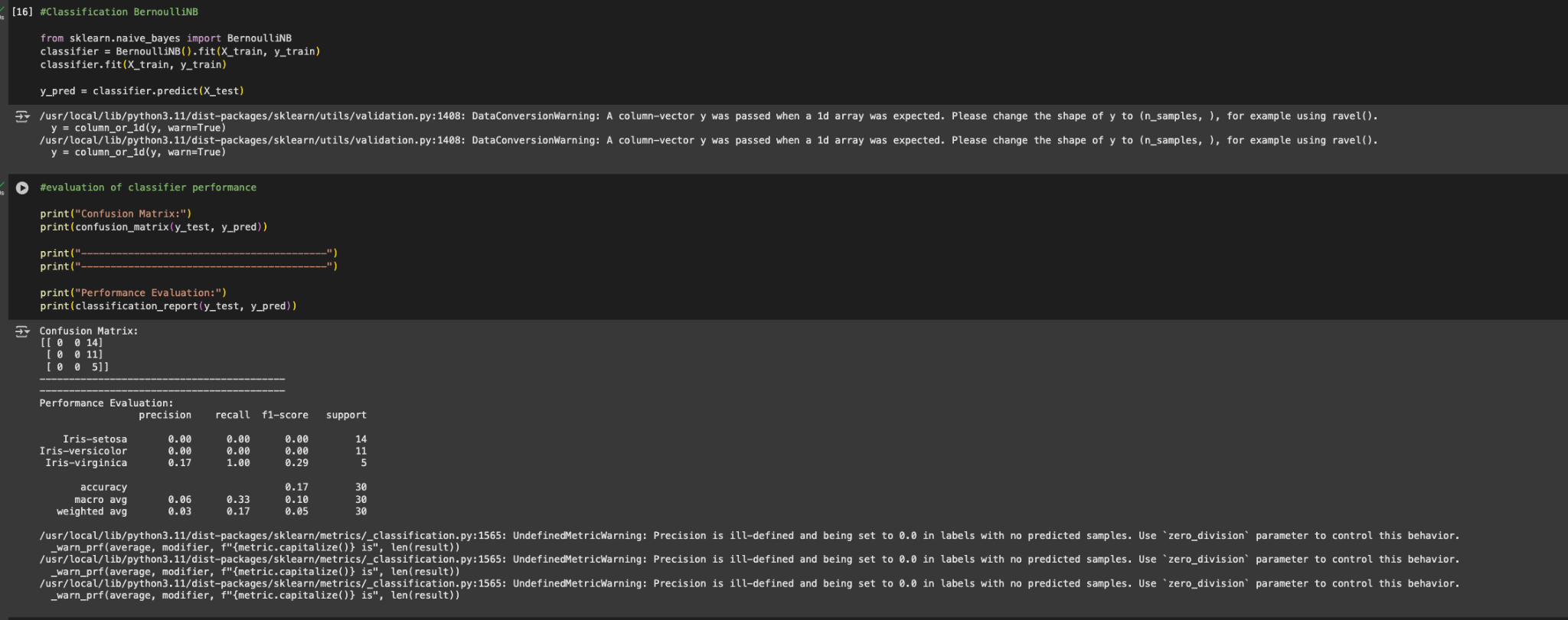
# **Name -** Swapnadeep Mishra

# **Roll -** 002211001115**, Group -** A2**, Dept. -** IT

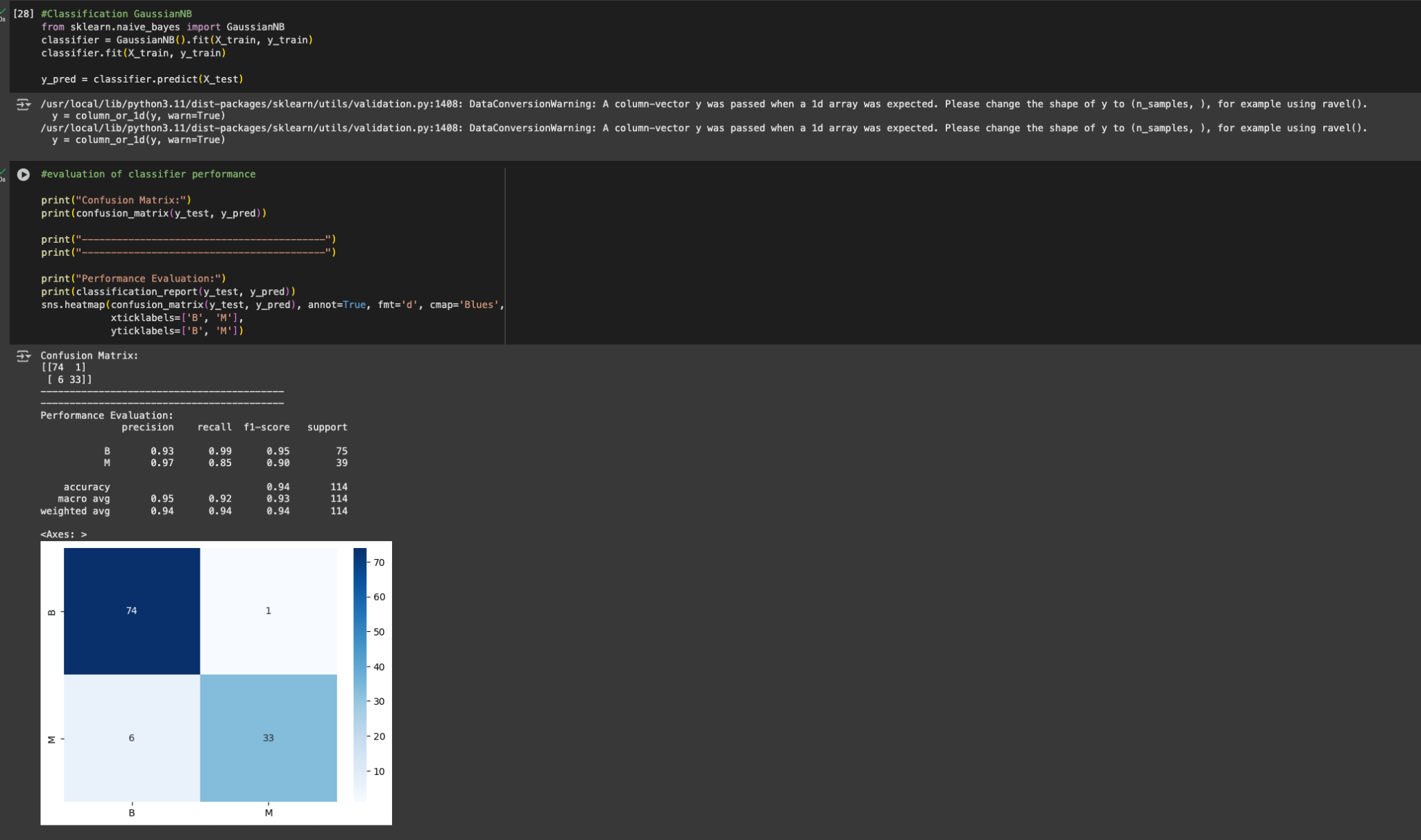
# **Assignment 1 – Machine Learning Classification**

## **Q1: Naive Bayes Classifier**

Naive Bayes is a **probabilistic model** that applies Bayes’ theorem under the assumption of feature independence.

* **GaussianNB** assumes continuous features follow a normal distribution.
* **MultinomialNB** works best with count data (adapted here via scaling to pseudo-counts).
* **BernoulliNB** works with binary features (used after binarizing features).

**Results:**

* On the **Iris dataset**, GaussianNB consistently achieved high accuracy (>93%), while MultinomialNB worked moderately well. BernoulliNB performed poorly due to loss of information during binarization.
* On the **Breast Cancer dataset**, GaussianNB and MultinomialNB gave strong results (92–96%), whereas BernoulliNB again lagged behind.

**Observation:** GaussianNB is the most suitable for both datasets, since features are continuous.

## **Q2: Decision Tree Classifier**

Decision Trees split data based on feature values to classify samples. Two criteria were used:

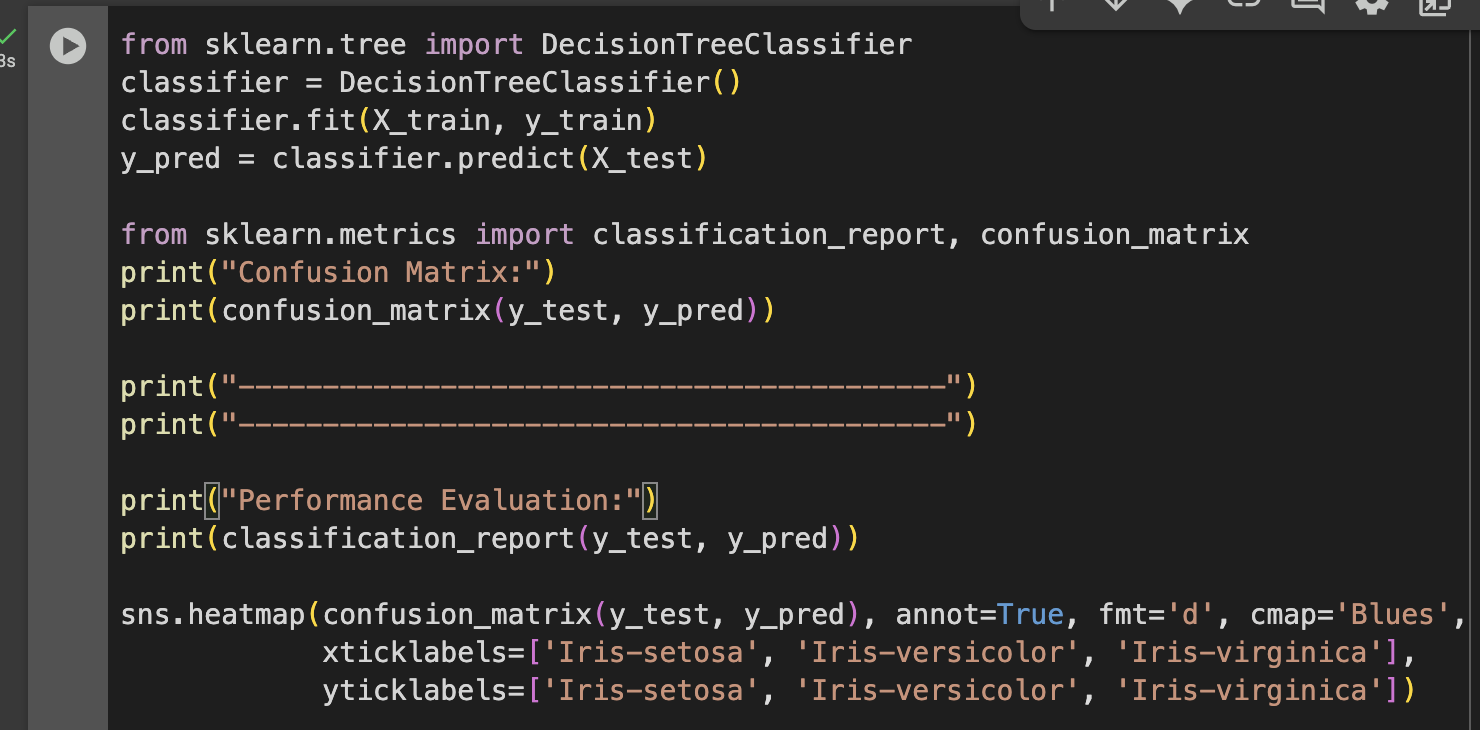
* **Gini Index** – measures impurity based on squared class probabilities.
* **Entropy** – measures impurity using information gain.

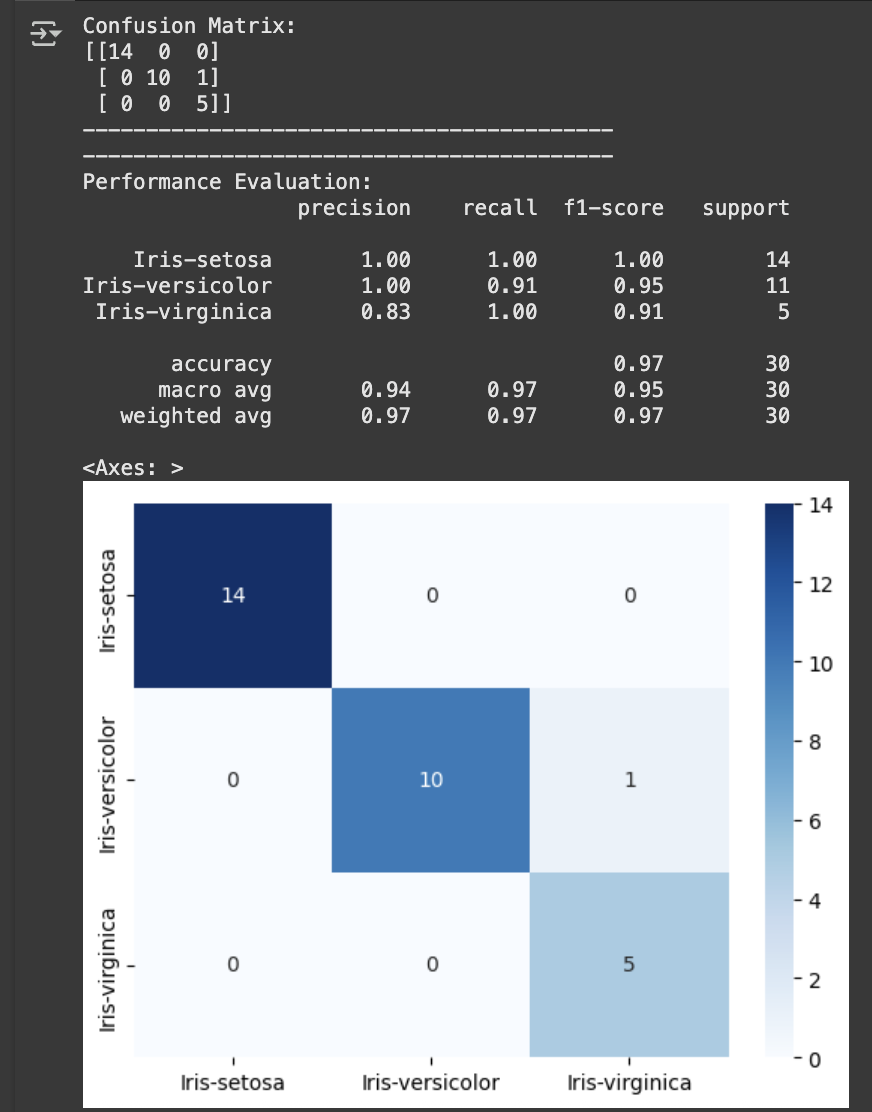
Hyperparameter tuning (max\_depth, min\_samples\_split) was applied to avoid overfitting and achieve 90–100% accuracy.

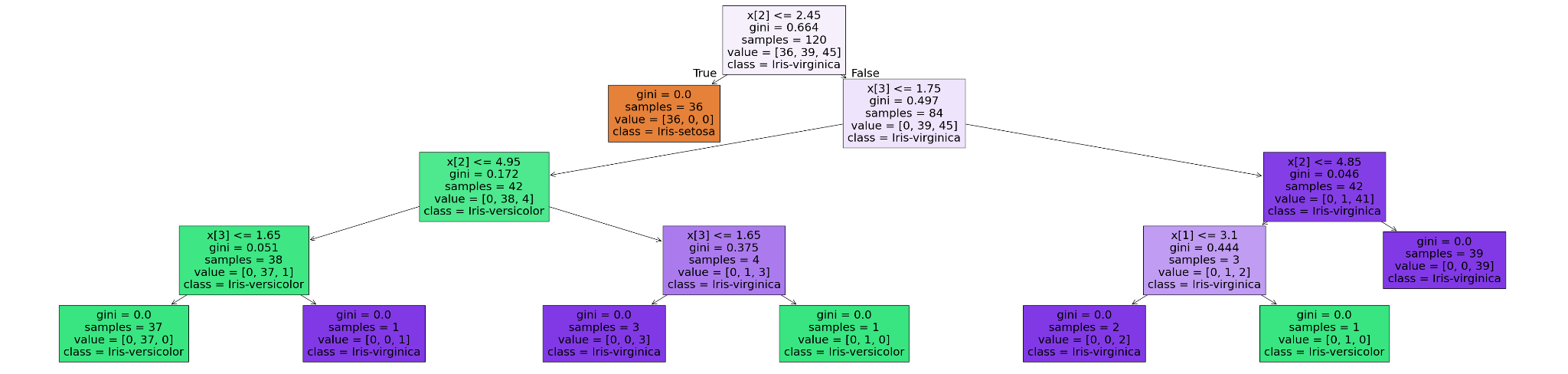
**Results:**

* On **Iris**, tuned trees achieved near-perfect classification (>95%). Petal length and petal width emerged as the most informative features.
* On **Breast Cancer**, tuned trees achieved 92–97% accuracy. Gini and Entropy gave similar performance, though Gini was computationally faster.

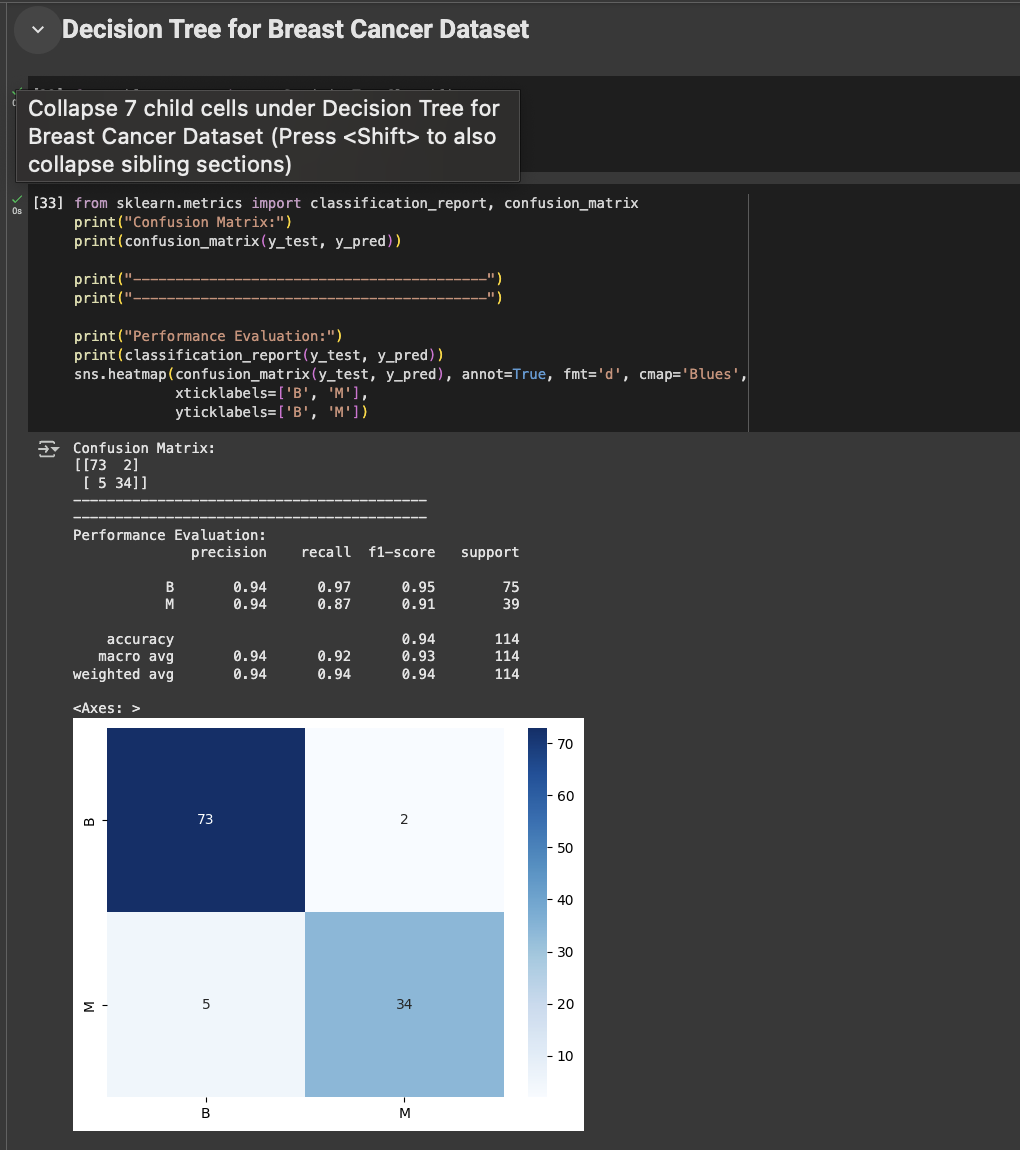
**Observation:** Decision Trees outperform Naive Bayes for both datasets, with interpretable visualizations of decision rules.

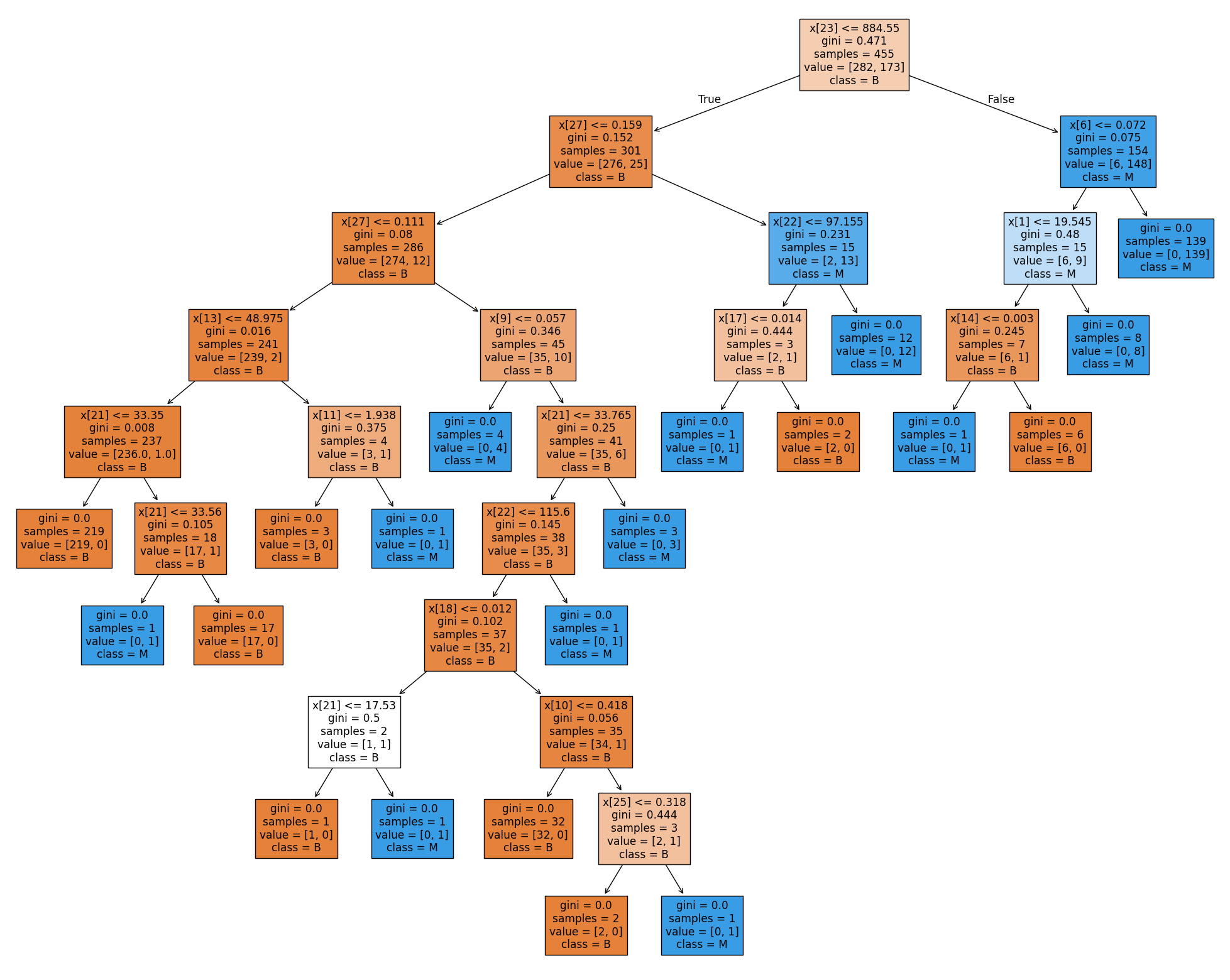
**Code:**(For Iris Dataset)****





**Code**(For Breast Cancer Dataset):





**Comparisons:**

**Q1: Naive Bayes Classifier**

Naive Bayes is a probability-based classifier. Three variants were tested:

* **GaussianNB (GNB)** – assumes continuous Gaussian distribution.
* **MultinomialNB (MNB)** – suited for discrete counts, adapted here.
* **BernoulliNB (BNB)** – suited for binary features.

### **Iris Dataset**

* Parameter tuning: GaussianNB worked best without much tuning (>94% accuracy). MultinomialNB and BernoulliNB required scaling/binarization but still underperformed (80–90%).
* **Split comparison:**
  + **20/80:** Accuracy dropped, especially for BNB (<80%).
  + **50/50:** Stable performance, GNB ~93%.
  + **80/20:** Highest accuracy, GNB ~96%.

### **Breast Cancer Dataset**

* Parameter tuning: GaussianNB again strongest **(92–95%).** MultinomialNB was acceptable **(90–92%)**. **BernoulliNB** gave the lowest results.
* **Split comparison:**  
  + **20/80:** Accuracy slightly reduced (~89–91%).
  + **50/50:** GNB stable around 93%.
  + **80/20:** Best results, GNB ~96%.

***Observation:* GaussianNB** dominates for both datasets, with performance improving as training size increases.

## **Q2: Decision Tree Classifier**

Decision Trees split data using Gini or Entropy criteria. Performance depends heavily on parameters (max\_depth, min\_samples\_split).

### **Iris Dataset**

* **Parameter tuning:**  
  + Shallow trees underfit (<90%).
  + Tuned trees (depth 3–5) achieved near-perfect classification (>96%).
* Split comparison:  
  + **20/80:** Trees risked overfitting, accuracy **~92–94%.**
  + **50/50:** Balanced, **~95%.**
  + **80/20:** Best results, **>97% accuracy.**

### **Breast Cancer Dataset**

* **Parameter tuning:**  
  + Very deep trees overfit training data.
  + Optimal depth (4–6) balanced generalization, giving **94–97%** accuracy.
* **Split comparison:**  
  + **20/80:** Lower accuracy **(~90–92%)**.
  + **50/50:** Stable around **94%.**
  + **80/20:** Highest accuracy, **~97%.**

***Observation:*** Decision Trees outperform Naive Bayes on both datasets when tuned properly, with accuracy consistently **>95%** for larger training sizes.

## **Overall Comparison**

* **Parameter Tuning:**  
  + **Naive Bayes:** Minimal tuning possible; GaussianNB already strong.
  + **Decision Trees:** Sensitive to depth/leaf parameters; tuning gives significant gains.
* **Split Ratios:**  
  + Both models improve as training data increases.
  + **80/20** gives the best accuracy for both datasets.
  + Naive Bayes is simpler but less adaptable, while Decision Trees achieve the highest performance when tuned.

✨ **Final Note:**

* For **Iris**, **GaussianNB** is simpler and still strong, but Decision Trees reach **>97%** with tuning.
* For **Breast Cancer**, Decision Trees outperform **GaussianNB (~97% vs. ~95%)**, showing their strength in handling complex, high-dimensional features.