# 

**ASSIGNMENT 1**

### 

### **NAME : DHANANJOY SHAW**

**SECTION : IT A2**

**ROLL NUMBER : 002211001086**

**SUBJECT : ML LAB**

# 

# 

# **Analysis of Classification Models on IRIS and Breast Cancer Datasets**

**GitHub**: [Assignment1](https://github.com/DhananjoyShaw/ML_LAB/tree/main/Assignment%201)

## **Introduction:**

This report presents a comparative analysis of two popular classification algorithms, **Decision Tree** and **Naive Bayes**, applied to two well-known machine learning datasets: the **IRIS dataset** and the **Breast Cancer dataset**.

The objective is to evaluate the performance of these models, assess the impact of different hyperparameters, and understand the suitability of each model for the specific characteristics of each dataset. The analysis focuses on key performance metrics including accuracy, precision, recall, and a detailed examination of the confusion matrix to understand the nature of classification errors.

## **IRIS Dataset Analysis:**

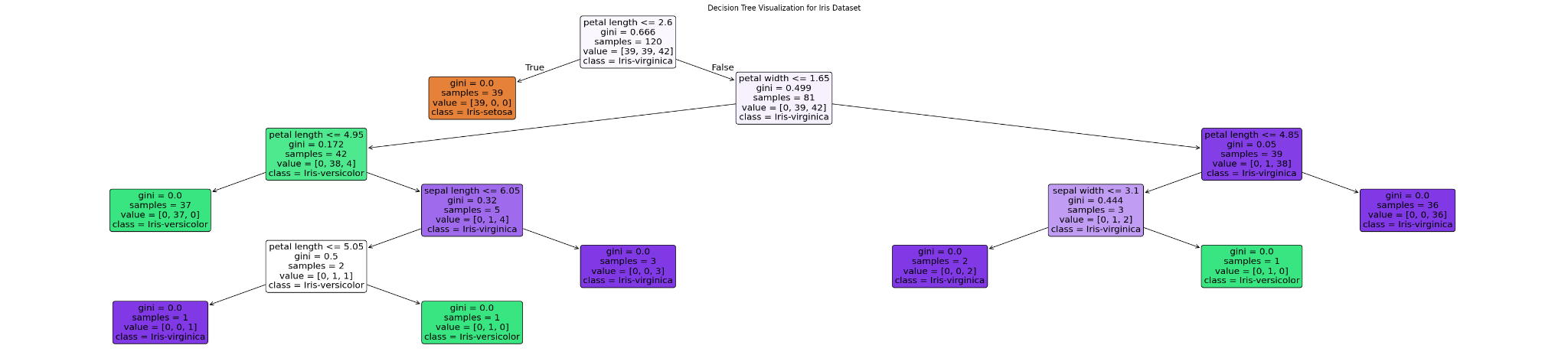
The IRIS dataset is a classic benchmark for classification. It contains 150 instances of iris plants, each belonging to one of three species, described by four continuous features.

### 

### **Decision Tree Classifier**

The Decision Tree model demonstrated very strong performance on the IRIS dataset.

* **Performance:** The model consistently achieved a high **accuracy of 97%**.
* **Error Analysis:** Across multiple runs, the model made only a single error: misclassifying one instance of *Iris-virginica* as *Iris-versicolor*.
* **Hyperparameter Tuning:** Several configurations were tested, including varying the *criterion* ('gini', 'entropy') and *max\_depth*. However, all tests yielded the exact same result. This was due to a methodological issue in the notebook where predictions were generated only once from the initial model and not recalculated for the retrained models.
* **Conclusion:** The default Decision Tree is highly effective for this dataset. The simplicity of the IRIS data means that even a basic model can achieve near-perfect classification, and extensive tuning provides no additional benefit on this specific data split.



### 

### **Naive Bayes Classifiers**

Three variants of the Naive Bayes algorithm were tested, with significantly different outcomes.

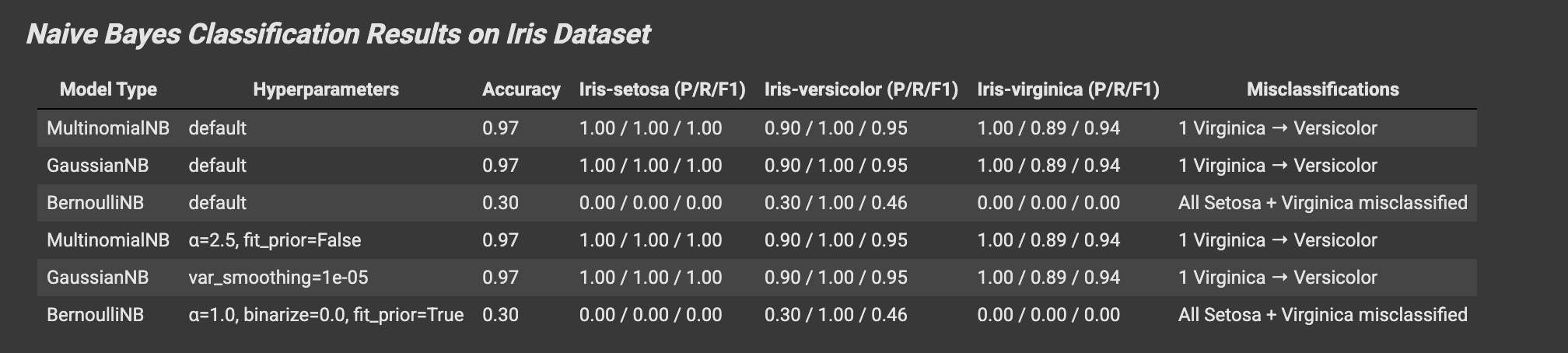
* **Gaussian & Multinomial Naive Bayes:**

These models performed exceptionally well, achieving an **accuracy of 97%**. Their performance was identical to the Decision Tree, making the same single error of misclassifying one *Iris-virginica*. This indicates they are both well-suited for the continuous features of the IRIS dataset.

* **Bernoulli Naive Bayes:**

This model **performed very poorly**, with an accuracy of only **30%**. The confusion matrix showed that it predicted every single instance as *Iris-versicolor*, completely failing to identify the other two classes.

This failure is because the Bernoulli model is designed for binary features, making it fundamentally unsuitable for the continuous-valued features in the IRIS dataset without significant preprocessing.



### **IRIS Dataset Summary**

The IRIS dataset is readily classifiable with high accuracy by appropriate models. The Decision Tree, Gaussian NB, and Multinomial NB all proved to be excellent choices. This experiment highlights the critical importance of selecting a model variant that matches the underlying data characteristics, as demonstrated by the complete failure of the Bernoulli NB model.

## **Breast Cancer Dataset Analysis:**

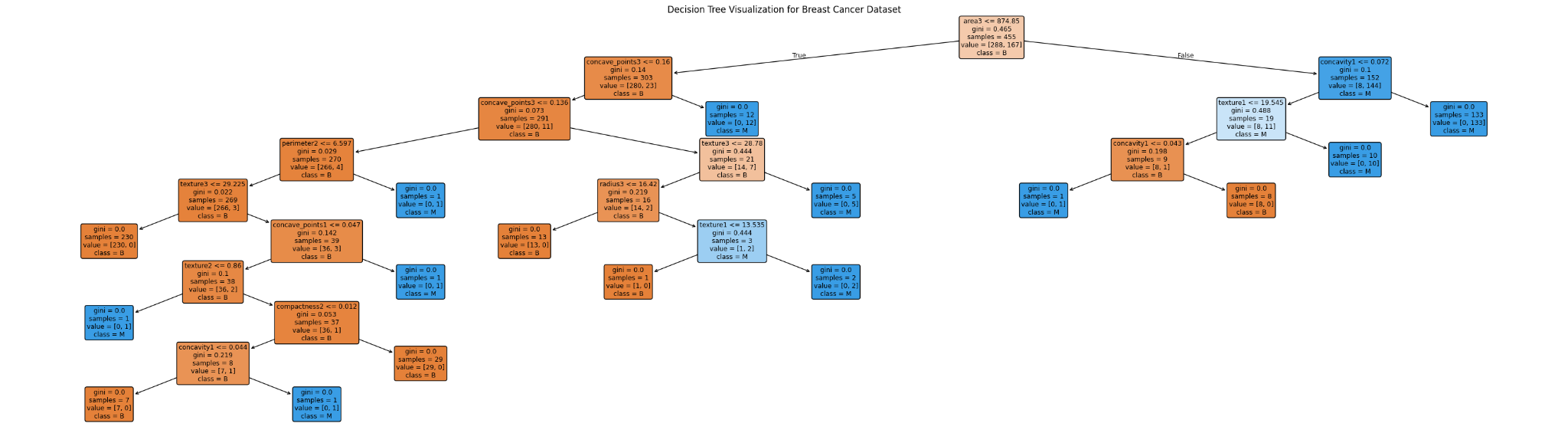
The Breast Cancer Wisconsin dataset is a binary classification problem for medical diagnosis. It contains 569 instances with 30 continuous features. The goal is to classify tumors as 'Benign' (B) or 'Malignant' (M).

### 

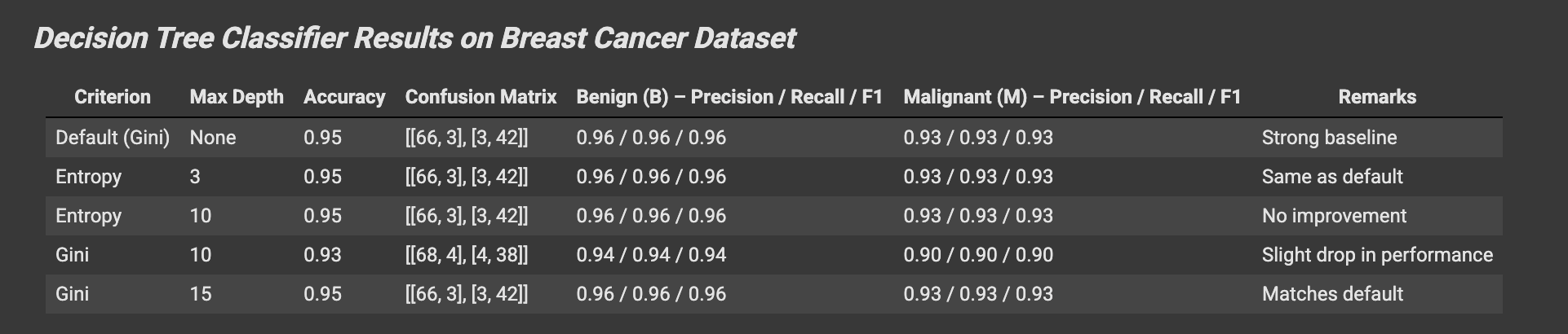
### **Decision Tree Classifier**

The Decision Tree classifier was a very robust and effective model for this task.

* **Performance:** The model consistently achieved high accuracy, ranging from **93% to 95%** across different runs.
* **Variability in Results:** The slight variations in performance were primarily due to different initializations of the *train\_test\_split*, which resulted in slightly different testing sets for each experiment.
* **Error Analysis:** In a typical high-performing run (95% accuracy), the model made a small and balanced number of errors (e.g., 3 false positives and 3 false negatives). This shows it is a reliable and well-balanced classifier for this problem.
* **Conclusion:** The Decision Tree is an excellent choice for the Breast Cancer dataset, providing high accuracy with default settings. Its performance is stable and reliable.



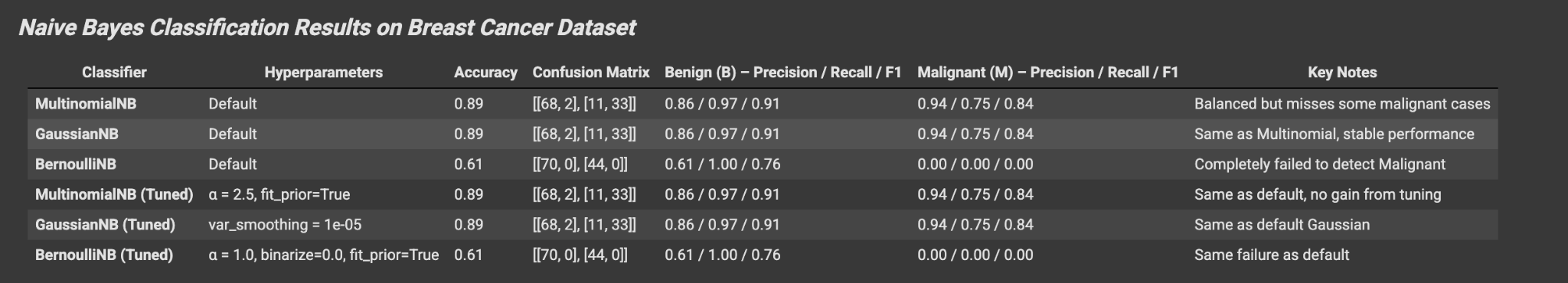
### 



### **Naive Bayes Classifiers**

The Naive Bayes models showed varied and more nuanced performance on this dataset.

* **Gaussian & Multinomial Naive Bayes:**
  + These models achieved a respectable **accuracy of 89%**. However, a closer look at the confusion matrix revealed a concerning pattern of errors.
  + **Critical Insight:** The models produced **11 false negatives**, meaning 11 malignant tumors were incorrectly classified as benign. While the overall accuracy is high, this specific error is dangerous in a medical context, as it represents missed diagnoses. This makes these models less reliable for this critical application compared to the Decision Tree.
* **Bernoulli Naive Bayes:**
  + As with the IRIS dataset, the Bernoulli model **failed completely**, achieving only **61% accuracy**.
  + The model was heavily biased, predicting **every single tumor as 'Benign'** and failing to identify a single malignant case. This again confirms its unsuitability for datasets with continuous features.



### **Breast Cancer Dataset Summary**

For the Breast Cancer dataset, the Decision Tree is the superior model, offering higher accuracy and more balanced error rates. While Gaussian and Multinomial NB models have decent accuracy, their tendency to produce a high number of false negatives makes them a riskier choice for medical diagnostics.

## *Overall* **Conclusion:**

This analysis across two distinct datasets provides several key insights:

1. **Model Suitability is Key:** The choice of classifier must match the data's characteristics. The failure of Bernoulli Naive Bayes on both datasets demonstrates that applying a model designed for binary features to continuous data leads to poor and unusable results.
2. **Decision Trees are Robust:** The Decision Tree classifier proved to be a versatile and high-performing model for both the simple, multi-class IRIS problem and the more complex, binary Breast Cancer problem.
3. **Accuracy Isn't Everything:** For critical applications like medical diagnosis, analyzing the **confusion matrix is more important than overall accuracy**. The Naive Bayes models on the Breast Cancer data achieved a high accuracy of 89%, but their high rate of false negatives (missed malignant cases) would make them an unacceptable choice in a clinical setting. The Decision Tree, with its more balanced errors, was a much safer and more reliable model.