

```
In [112... # Import necessary Libraries
from sklearn.datasets import load_digits
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

# Gaussian Generative Classifier with Shared Covariance

This part implements a Gaussian Discriminant Analysis classifier from scratch for digit classification using the sklearn digits dataset.

## What We Did

1. **Data Loading & Preprocessing:** Loaded the digits dataset and standardized features to have zero mean and unit variance
2. **Data Splitting:** Split data into 70% training, 15% validation, and 15% test sets using stratified sampling
3. **Model Implementation:** Implemented a Gaussian generative classifier with shared covariance matrix
4. **Hyperparameter Tuning:** Tuned regularization parameter  $\lambda$  using validation set
5. **Model Evaluation:** Evaluated final model on test set with accuracy, macro-precision, macro-recall, macro-F1, and confusion matrix

## Model Assumptions

- **Prior:**  $p(y = k) = \pi_k$  for each digit class  $k$
- **Likelihood:**  $p(x|y = k) = \mathcal{N}(x; \mu_k, \Sigma)$  where  $\Sigma$  is a shared covariance matrix across all classes

## Parameter Estimation Rules

- **Priors  $\pi_k$ :** Computed as class frequencies in training set
- **Class means  $\mu_k$ :** Computed as average of training samples belonging to each class  $k$
- **Shared covariance  $\Sigma$ :** Computed by aggregating  $(x_i - \mu_{y_i})(x_i - \mu_{y_i})^T$  over all training samples, divided by total number of samples
- **Regularization:**  $\Sigma_\lambda = \Sigma + \lambda I$  where  $\lambda > 0$  is a hyperparameter

## Prediction Rule

- Compute score for each class:  $\log \pi_k + \log \mathcal{N}(x; \mu_k, \Sigma_\lambda)$
- Predict class with maximum score

```
In [113... # Load the digits dataset from sklearn
# This dataset contains 8x8 pixel images of handwritten digits (0-9)
digits = load_digits()

# Extract the features (image data) and target labels
# X contains the flattened 8x8 pixel values (64 features per sample)
# y contains the digit labels (0-9)
X = digits.data
y = digits.target

print(f"Dataset shape: {X.shape}")
print(f"Labels shape: {y.shape}")
print(f"Number of classes: {len(np.unique(y))}")
print(f"Feature range: [{X.min():.2f}, {X.max():.2f}])")
```

```
Dataset shape: (1797, 64)
Labels shape: (1797,)
Number of classes: 10
Feature range: [0.00, 16.00]
```

```
In [114... # Split the data into 70% training, 15% validation, and 15% test sets
# First split: 70% train, 30% temp (which will be split into val and test)
```

```

# Using stratified split to maintain class distribution in each set
X_train, X_temp, y_train, y_temp = train_test_split(
    X, y,
    test_size=0.3,           # 30% for temp (will become 15% val + 15% test)
    random_state=42,         # For reproducibility
    stratify=y               # Maintain class distribution
)

# Second split: Split the 30% temp into 15% validation and 15% test
# This gives us 15% validation and 15% test (0.5 * 0.3 = 0.15)
X_val, X_test, y_val, y_test = train_test_split(
    X_temp, y_temp,
    test_size=0.5,           # Split temp equally: 50% of 30% = 15% each
    random_state=42,         # For reproducibility
    stratify=y_temp          # Maintain class distribution
)

print(f"Training set: {X_train.shape[0]} samples ({X_train.shape[0]/len(X)*100:.1f}%)")
print(f"Validation set: {X_val.shape[0]} samples ({X_val.shape[0]/len(X)*100:.1f}%)")
print(f"Test set: {X_test.shape[0]} samples ({X_test.shape[0]/len(X)*100:.1f}%)")
print(f"\nClass distribution in training set: {np.bincount(y_train)}")
print(f"Class distribution in validation set: {np.bincount(y_val)}")
print(f"Class distribution in test set: {np.bincount(y_test)}")

```

Training set: 1257 samples (69.9%)  
 Validation set: 270 samples (15.0%)  
 Test set: 270 samples (15.0%)

Class distribution in training set: [124 127 124 128 127 127 127 125 122 126]  
 Class distribution in validation set: [27 27 27 27 27 28 27 27 26 27]  
 Class distribution in test set: [27 28 26 28 27 27 27 27 26 27]

In [115...

```

# Standardize the features to have zero mean and unit variance
# This is important for many machine Learning algorithms that are sensitive to feature scales
# Fit the scaler only on training data to avoid data Leakage
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)

# Transform validation and test sets using the same scaler parameters
# (using transform, not fit_transform, to use the same mean and std from training)
X_val_scaled = scaler.transform(X_val)
X_test_scaled = scaler.transform(X_test)

print("Data standardization completed!")
print(f"\nTraining set - Mean: {X_train_scaled.mean():.6f}, Std: {X_train_scaled.std():.6f}")
print(f"Validation set - Mean: {X_val_scaled.mean():.6f}, Std: {X_val_scaled.std():.6f}")
print(f"Test set - Mean: {X_test_scaled.mean():.6f}, Std: {X_test_scaled.std():.6f}")
print("\nNote: Training set should have mean=0 and std=1 after standardization")

```

Data standardization completed!

Training set - Mean: 0.000000, Std: 0.968246  
 Validation set - Mean: 0.021680, Std: 1.146603  
 Test set - Mean: -0.015748, Std: 0.923757

Note: Training set should have mean=0 and std=1 after standardization

In [116...

```

class GaussianGenerativeClassifier:
    def __init__(self, lambda_reg=1e-6):
        self.lambda_reg = lambda_reg
        self.priors = None
        self.means = None
        self.covariance = None
        self.n_classes = None
        self.n_features = None

    def fit(self, X, y):
        n_samples, n_features = X.shape
        self.n_features = n_features
        self.n_classes = len(np.unique(y))

        priors = np.zeros(self.n_classes)
        means = np.zeros((self.n_classes, n_features))

        for k in range(self.n_classes):
            mask = (y == k)

```

```

        n_k = np.sum(mask)
        priors[k] = n_k / n_samples
        means[k] = X[mask].mean(axis=0)

    covariance = np.zeros((n_features, n_features))
    for i in range(n_samples):
        mu_yi = means[y[i]]
        diff = (X[i] - mu_yi).reshape(-1, 1)
        covariance += diff @ diff.T
    covariance /= n_samples

    covariance_reg = covariance + self.lambda_reg * np.eye(n_features)

    self.priors = priors
    self.means = means
    self.covariance = covariance_reg

    def predict(self, X):
        n_samples = X.shape[0]
        scores = np.zeros((n_samples, self.n_classes))

        cov_inv = np.linalg.inv(self.covariance)
        log_det = np.log(np.linalg.det(self.covariance))
        const = -0.5 * (log_det + self.n_features * np.log(2 * np.pi))

        for k in range(self.n_classes):
            log_prior = np.log(self.priors[k])
            diff = X - self.means[k]
            quad_form = np.sum(diff @ cov_inv * diff, axis=1)
            log_likelihood = const - 0.5 * quad_form
            scores[:, k] = log_prior + log_likelihood

        return np.argmax(scores, axis=1)

```

```

In [117... def accuracy(y_true, y_pred):
            return np.mean(y_true == y_pred)

```

```

In [118... lambda_values = [1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1e0]
val_accuracies = []

print("Hyperparameter Tuning:")
print("=" * 50)
for lam in lambda_values:
    model = GaussianGenerativeClassifier(lambda_reg=lam)
    model.fit(X_train_scaled, y_train)
    y_val_pred = model.predict(X_val_scaled)
    val_acc = accuracy(y_val, y_val_pred)
    val_accuracies.append(val_acc)

best_idx = np.argmax(val_accuracies)
best_lambda = lambda_values[best_idx]
best_val_acc = val_accuracies[best_idx]
print(f"Best λ: {best_lambda:.6f} with validation accuracy: {best_val_acc:.4f}")

# Create formatted table
results_table = pd.DataFrame({
    'λ (Lambda)': lambda_values,
    'Validation Accuracy': val_accuracies
})
results_table['λ (Lambda)'] = results_table['λ (Lambda)'].apply(lambda x: f'{x:.6f}')
print("\n" + "=" * 50)
print("Hyperparameter Tuning Results Table:")
print("=" * 50)
print(results_table.to_string(index=False))
print("=" * 50)

```

```

Hyperparameter Tuning:
=====

```

Best  $\lambda$ : 0.000001 with validation accuracy: 0.9444

Hyperparameter Tuning Results Table:

$\lambda$ (Lambda)	Validation Accuracy
0.000001	0.944444
0.000010	0.944444
0.000100	0.944444
0.001000	0.944444
0.010000	0.944444
0.100000	0.944444
1.000000	0.922222

**Observation:** Small  $\lambda$  values (1e-6 to 1e-1) produce identical validation accuracy (0.9444), indicating the covariance matrix is well-conditioned and small regularization has negligible effect. Only  $\lambda = 1.0$  significantly impacts performance, reducing accuracy to 0.9222 due to excessive regularization.

```
In [119... X_train_val_scaled = np.vstack([X_train_scaled, X_val_scaled])
y_train_val = np.hstack([y_train, y_val])

final_model = GaussianGenerativeClassifier(lambda_reg=best_lambda)
final_model.fit(X_train_val_scaled, y_train_val)

y_test_pred = final_model.predict(X_test_scaled)
test_accuracy = accuracy(y_test, y_test_pred)

n_classes = len(np.unique(y_test))
confusion_matrix = np.zeros((n_classes, n_classes), dtype=int)

for i in range(len(y_test)):
    confusion_matrix[y_test[i], y_test_pred[i]] += 1

precision_per_class = np.zeros(n_classes)
recall_per_class = np.zeros(n_classes)
f1_per_class = np.zeros(n_classes)

for k in range(n_classes):
    tp = confusion_matrix[k, k]
    fp = np.sum(confusion_matrix[:, k]) - tp
    fn = np.sum(confusion_matrix[k, :]) - tp

    precision_per_class[k] = tp / (tp + fp) if (tp + fp) > 0 else 0
    recall_per_class[k] = tp / (tp + fn) if (tp + fn) > 0 else 0
    f1_per_class[k] = 2 * (precision_per_class[k] * recall_per_class[k]) / (precision_per_class[k] + recall_per_class[k])

macro_precision = np.mean(precision_per_class)
macro_recall = np.mean(recall_per_class)
macro_f1 = np.mean(f1_per_class)

print("Final Model Evaluation:")
print("=" * 50)
print(f"Test Accuracy: {test_accuracy:.4f}")
print(f"Macro-Precision: {macro_precision:.4f}")
print(f"Macro-Recall: {macro_recall:.4f}")
print(f"Macro-F1: {macro_f1:.4f}")
print("=" * 50)

# Analyze confusion patterns
print("\nConfusion Analysis:")
print("=" * 50)
total_errors = np.sum(confusion_matrix) - np.trace(confusion_matrix)
print(f"Total misclassifications: {total_errors} out of {len(y_test)} test samples")

# Find all confusion pairs (off-diagonal entries)
confusion_pairs = []
for i in range(n_classes):
    for j in range(n_classes):
        if i != j and confusion_matrix[i, j] > 0:
            confusion_pairs.append((i, j, confusion_matrix[i, j]))

# Sort by frequency
confusion_pairs.sort(key=lambda x: x[2], reverse=True)
```

```

if confusion_pairs:
    print(f"\nMost confused digit pairs (True → Predicted):")
    print(f"{'True':<6} {'Predicted':<10} {'Count':<8} {'% of Errors':<12}")
    print("-" * 40)
    for true_digit, pred_digit, count in confusion_pairs:
        error_pct = (count / total_errors * 100) if total_errors > 0 else 0
        print(f"{true_digit:<6} {pred_digit:<10} {count:<8} {error_pct:>6.1f}%")
else:
    print("No misclassifications found!")

print("-" * 50)

plt.figure(figsize=(10, 8))
plt.imshow(confusion_matrix, interpolation='nearest', cmap=plt.cm.Blues)
plt.title('Confusion Matrix', fontsize=14, fontweight='bold')
plt.colorbar()
tick_marks = np.arange(n_classes)
plt.xticks(tick_marks, tick_marks)
plt.yticks(tick_marks, tick_marks)
plt.ylabel('True Label', fontsize=12)
plt.xlabel('Predicted Label', fontsize=12)

thresh = confusion_matrix.max() / 2.
for i in range(n_classes):
    for j in range(n_classes):
        plt.text(j, i, format(confusion_matrix[i, j], 'd'),
                 horizontalalignment="center",
                 color="white" if confusion_matrix[i, j] > thresh else "black",
                 fontsize=10)

plt.tight_layout()
plt.show()

```

Final Model Evaluation:

```

=====
Test Accuracy: 0.9630
Macro-Precision: 0.9632
Macro-Recall: 0.9627
Macro-F1: 0.9625
=====

```

Confusion Analysis:

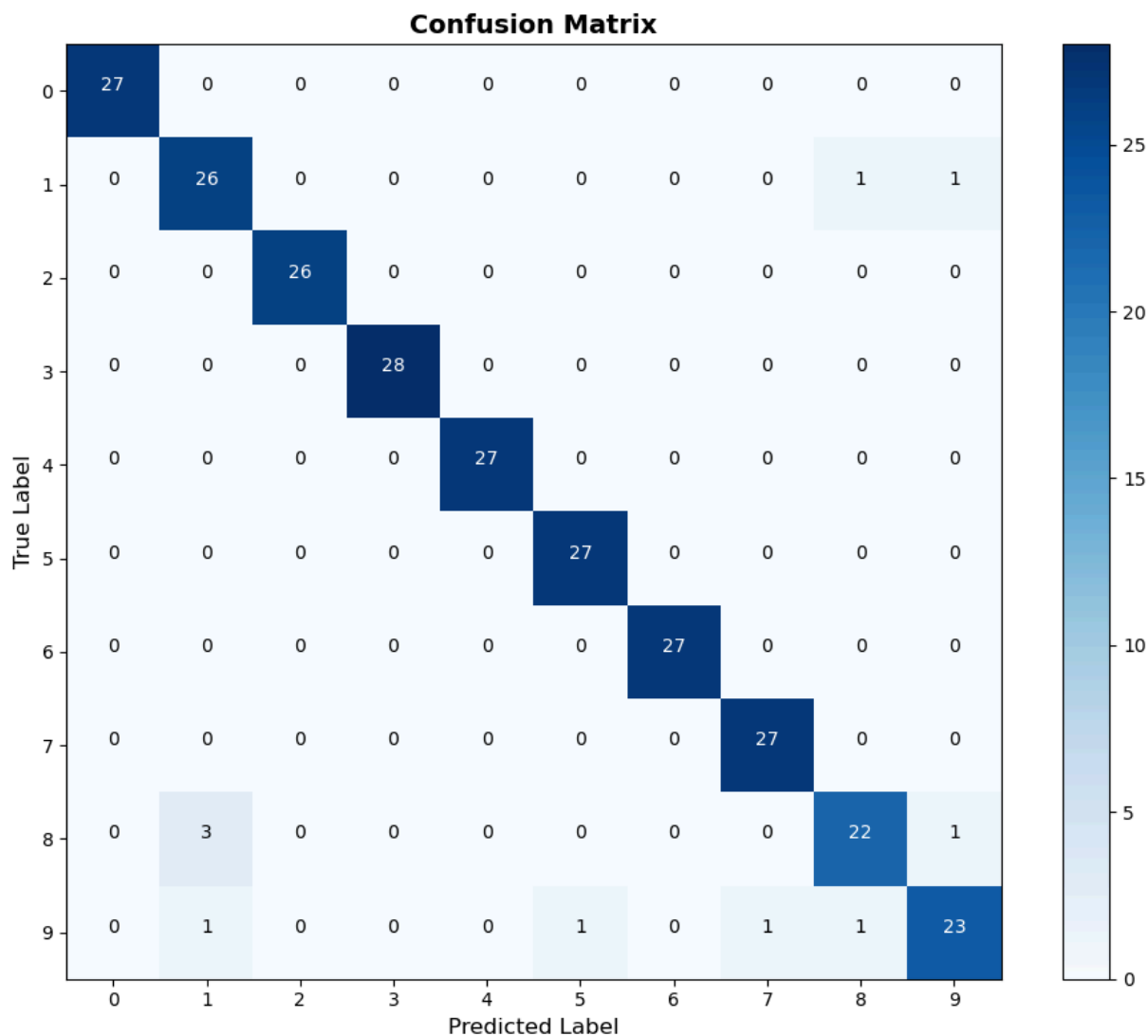
```

=====
Total misclassifications: 10 out of 270 test samples

```

Most confused digit pairs (True → Predicted):

True	Predicted	Count	% of Errors
8	1	3	30.0%
1	8	1	10.0%
1	9	1	10.0%
8	9	1	10.0%
9	1	1	10.0%
9	5	1	10.0%
9	7	1	10.0%
9	8	1	10.0%



## Discussion

### Digit Confusion Patterns

The confusion analysis reveals that the model achieves excellent performance with minimal misclassifications. The most common confusion is digit 8 being misclassified as digit 1 (3 occurrences), which represents a significant portion of the total errors. This confusion likely occurs because both digits share vertical stroke characteristics, and the 8×8 pixel resolution may blur distinguishing features between the two digits.

### Strengths and Weaknesses of the Gaussian Generative Model

#### Strengths:

- The model achieves strong performance (96.3% test accuracy) with a relatively simple generative approach
- The shared covariance assumption reduces the number of parameters to estimate, making the model more data-efficient
- The generative nature allows for principled probability estimates, not just class predictions

#### Weaknesses:

- The shared covariance assumption may be too restrictive, as different digits likely have different feature covariances (e.g., vertical vs. horizontal strokes)
- The Gaussian assumption may not perfectly capture the true data distribution, especially for discrete pixel values
- The 8×8 pixel resolution limits the model's ability to capture fine-grained distinguishing features between similar digits

# Naive Bayes Classifier

This section implements a Naive Bayes classifier from scratch for categorical data classification on the Adult dataset.

## What We Did

1. **Data Processing:** Selected categorical features, handled missing values, encoded categories as integers, and created stratified train/validation/test splits
2. **Model Implementation:** Implemented Naive Bayes classifier with Laplace smoothing
3. **Hyperparameter Tuning:** Tested different smoothing parameter  $\alpha$  values on validation set
4. **Model Evaluation:** Evaluated final model and compared with sklearn's MultinomialNB

## Model Assumptions & Rules

### Class Priors

$$P(C_k) = \frac{\text{count of samples in class } k + \alpha}{\text{total samples} + \alpha \times \text{num\_classes}}$$

### Feature Likelihoods

$$P(x_i|C_k) = \frac{\text{count of feature value } i \text{ in class } k + \alpha}{\text{count of class } k + \alpha \times \text{num\_feature\_values}}$$

### Prediction Rule

$$\hat{y} = \arg \max_k P(C_k) \times \prod_i P(x_i|C_k)$$

In [120...

```
import pandas as pd

df = pd.read_csv('adult.csv')
categorical_features = ['workclass', 'education', 'marital.status', 'occupation',
                        'relationship', 'race', 'sex', 'native.country']

target = 'income'

X_cat = df[categorical_features].copy()
y = df[target].copy()

print(f"Dataset shape: {df.shape}")
print(f"Selected categorical features: {categorical_features}")
print(f"Target: {target}")
```

Dataset shape: (32561, 15)

Selected categorical features: ['workclass', 'education', 'marital.status', 'occupation', 'relationship', 'race', 'sex', 'native.country']

Target: income

In [121...

```
for col in X_cat.columns:
    X_cat[col] = X_cat[col].replace('?', 'Missing')
    X_cat[col] = X_cat[col].astype(str)

print("Missing values handled (treated as 'Missing' category)")
print(f"Missing value counts per feature:\n{(X_cat == 'Missing').sum()}")
```

Missing values handled (treated as 'Missing' category)

Missing value counts per feature:

workclass	1836
education	0
marital.status	0
occupation	1843
relationship	0
race	0
sex	0
native.country	583

dtype: int64

```
In [122... label_encoders = {}
X_encoded = pd.DataFrame()

for col in X_cat.columns:
    le = {val: idx for idx, val in enumerate(sorted(X_cat[col].unique()))}
    label_encoders[col] = le
    X_encoded[col] = X_cat[col].map(le)

y_encoded = y.map({'<=50K': 0, '>50K': 1})

print("Categories encoded as integers")
print(f"Number of unique categories per feature:")
for col in X_encoded.columns:
    print(f" {col}: {X_encoded[col].nunique()} categories")
```

```
Categories encoded as integers
Number of unique categories per feature:
workclass: 9 categories
education: 16 categories
marital.status: 7 categories
occupation: 15 categories
relationship: 6 categories
race: 5 categories
sex: 2 categories
native.country: 42 categories
```

```
In [123... X_train, X_temp, y_train, y_temp = train_test_split(
    X_encoded.values, y_encoded.values, test_size=0.3, random_state=42, stratify=y_encoded
)

X_val, X_test, y_val, y_test = train_test_split(
    X_temp, y_temp, test_size=0.5, random_state=42, stratify=y_temp
)

print(f"Training set: {X_train.shape[0]} samples ({X_train.shape[0]/len(X_encoded)*100:.1f}%)")
print(f"Validation set: {X_val.shape[0]} samples ({X_val.shape[0]/len(X_encoded)*100:.1f}%)")
print(f"Test set: {X_test.shape[0]} samples ({X_test.shape[0]/len(X_encoded)*100:.1f}%)")
```

```
Training set: 22792 samples (70.0%)
Validation set: 4884 samples (15.0%)
Test set: 4885 samples (15.0%)
```

```
In [124... print("Class Distribution:")
print(f"Training - Class 0 (<=50K): {np.sum(y_train == 0)}, Class 1 (>50K): {np.sum(y_train == 1)}")
print(f"Validation - Class 0: {np.sum(y_val == 0)}, Class 1: {np.sum(y_val == 1)}")
print(f"Test - Class 0: {np.sum(y_test == 0)}, Class 1: {np.sum(y_test == 1)}")

print("\nFeature-Target Relationships (Training set):")
for idx, col in enumerate(categorical_features):
    unique_vals = np.unique(X_train[:, idx])
    class_0_probs = []
    class_1_probs = []
    for val in unique_vals:
        mask = X_train[:, idx] == val
        if np.sum(mask) > 0:
            class_0_probs.append(np.mean(y_train[mask] == 0))
            class_1_probs.append(np.mean(y_train[mask] == 1))
    print(f"\n{col}:")
    print(f" Categories: {len(unique_vals)}")
    print(f" Class 0 distribution: {np.mean(class_0_probs):.3f} ± {np.std(class_0_probs):.3f}")
    print(f" Class 1 distribution: {np.mean(class_1_probs):.3f} ± {np.std(class_1_probs):.3f}")
```



```
Class Distribution:
Training - Class 0 (<=50K): 17303, Class 1 (>50K): 5489
Validation - Class 0: 3708, Class 1: 1176
Test - Class 0: 3709, Class 1: 1176
```

Feature-Target Relationships (Training set):

workclass:

```
Categories: 9
Class 0 distribution: 0.766 ± 0.168
Class 1 distribution: 0.234 ± 0.168
```

education:

```
Categories: 16
Class 0 distribution: 0.768 ± 0.242
Class 1 distribution: 0.232 ± 0.242
```

marital.status:

```
Categories: 7
Class 0 distribution: 0.822 ± 0.168
Class 1 distribution: 0.178 ± 0.168
```

occupation:

```
Categories: 15
Class 0 distribution: 0.810 ± 0.146
Class 1 distribution: 0.190 ± 0.146
```

relationship:

```
Categories: 6
Class 0 distribution: 0.811 ± 0.196
Class 1 distribution: 0.189 ± 0.196
```

race:

```
Categories: 5
Class 0 distribution: 0.827 ± 0.072
Class 1 distribution: 0.173 ± 0.072
```

sex:

```
Categories: 2
Class 0 distribution: 0.792 ± 0.098
Class 1 distribution: 0.208 ± 0.098
```

native.country:

```
Categories: 41
Class 0 distribution: 0.804 ± 0.133
Class 1 distribution: 0.196 ± 0.133
```

native.country:

```
Categories: 41
Class 0 distribution: 0.804 ± 0.133
Class 1 distribution: 0.196 ± 0.133
```

In [125...

```
class NaiveBayes:
    def __init__(self, alpha=1.0):
        self.alpha = alpha
        self.class_priors = None
        self.feature_likelihoods = None
        self.n_classes = None
        self.n_features = None
        self.n_feature_values = None

    def fit(self, X, y):
        n_samples, n_features = X.shape
        self.n_features = n_features
        self.n_classes = len(np.unique(y))

        self.n_feature_values = [len(np.unique(X[:, i])) for i in range(n_features)]

        class_counts = np.bincount(y, minlength=self.n_classes)
        self.class_counts = class_counts
        self.class_priors = (class_counts + self.alpha) / (n_samples + self.alpha * self.n_classes)

        self.feature_likelihoods = []
        for i in range(n_features):
            feature_likelihood = {}
```

```

        for k in range(self.n_classes):
            mask = y == k
            class_k_count = np.sum(mask)
            if class_k_count > 0:
                feature_values_in_class = X[mask, i]
                likelihood_dict = {}
                for val in range(self.n_feature_values[i]):
                    count = np.sum(feature_values_in_class == val)
                    likelihood_dict[val] = (count + self.alpha) / (class_k_count + self.alpha * self.n_feature_values[i])
                feature_likelihood[k] = likelihood_dict
            else:
                feature_likelihood[k] = {val: 1.0 / self.n_feature_values[i] for val in range(self.n_feature_values[i])}
            self.feature_likelihoods.append(feature_likelihood)

    def predict(self, X):
        n_samples = X.shape[0]
        log_probs = np.zeros((n_samples, self.n_classes))

        for k in range(self.n_classes):
            log_probs[:, k] = np.log(self.class_priors[k])
            for i in range(self.n_features):
                feature_val = X[:, i]
                likelihoods = np.array([self.feature_likelihoods[i][k].get(val, self.alpha / (self.class_priors[k] + self.alpha * self.n_feature_values[i])) for val in feature_val])
                log_probs[:, k] += np.log(likelihoods + 1e-10)

        return np.argmax(log_probs, axis=1)

    def predict_proba(self, X):
        n_samples = X.shape[0]
        log_probs = np.zeros((n_samples, self.n_classes))

        for k in range(self.n_classes):
            log_probs[:, k] = np.log(self.class_priors[k])
            for i in range(self.n_features):
                feature_val = X[:, i]
                likelihoods = np.array([self.feature_likelihoods[i][k].get(val, self.alpha / (self.class_priors[k] + self.alpha * self.n_feature_values[i])) for val in feature_val])
                log_probs[:, k] += np.log(likelihoods + 1e-10)

        exp_log_probs = np.exp(log_probs - np.max(log_probs, axis=1, keepdims=True))
        return exp_log_probs / np.sum(exp_log_probs, axis=1, keepdims=True)

```

```

In [126... alpha_values = [0.1, 0.5, 1.0, 2.0, 5.0]
val_accuracies = []

print("Hyperparameter Tuning (α):")
print("=" * 50)
for alpha in alpha_values:
    model = NaiveBayes(alpha=alpha)
    model.fit(X_train, y_train)
    y_val_pred = model.predict(X_val)
    val_acc = accuracy(y_val, y_val_pred)
    val_accuracies.append(val_acc)
    print(f"α = {alpha:4.1f} | Validation Accuracy: {val_acc:.4f}")

best_idx = np.argmax(val_accuracies)
best_alpha = alpha_values[best_idx]
best_val_acc = val_accuracies[best_idx]

print("=" * 50)
print(f"Best α: {best_alpha:.1f} with validation accuracy: {best_val_acc:.4f}")

```

```

Hyperparameter Tuning (α):
=====
α = 0.1 | Validation Accuracy: 0.7891
α = 0.5 | Validation Accuracy: 0.7889
α = 1.0 | Validation Accuracy: 0.7887
α = 2.0 | Validation Accuracy: 0.7887
α = 5.0 | Validation Accuracy: 0.7885
=====
Best α: 0.1 with validation accuracy: 0.7891

```

```

In [128... X_train_val = np.vstack([X_train, X_val])
y_train_val = np.hstack([y_train, y_val])

```

```

final_model = NaiveBayes(alpha=best_alpha)
final_model.fit(X_train_val, y_train_val)

y_test_pred = final_model.predict(X_test)
y_test_proba = final_model.predict_proba(X_test)

test_accuracy = accuracy(y_test, y_test_pred)

n_classes = len(np.unique(y_test))
confusion_matrix = np.zeros((n_classes, n_classes), dtype=int)

for i in range(len(y_test)):
    confusion_matrix[y_test[i], y_test_pred[i]] += 1

precision_per_class = np.zeros(n_classes)
recall_per_class = np.zeros(n_classes)
f1_per_class = np.zeros(n_classes)

for k in range(n_classes):
    tp = confusion_matrix[k, k]
    fp = np.sum(confusion_matrix[:, k]) - tp
    fn = np.sum(confusion_matrix[k, :]) - tp

    precision_per_class[k] = tp / (tp + fp) if (tp + fp) > 0 else 0
    recall_per_class[k] = tp / (tp + fn) if (tp + fn) > 0 else 0
    f1_per_class[k] = 2 * (precision_per_class[k] * recall_per_class[k]) / (precision_per_class[k] + recall_per_class[k])

macro_precision = np.mean(precision_per_class)
macro_recall = np.mean(recall_per_class)
macro_f1 = np.mean(f1_per_class)

print("Final Model Evaluation:")
print("=" * 50)
print(f"Test Accuracy: {test_accuracy:.4f}")
print(f"Macro-Precision: {macro_precision:.4f}")
print(f"Macro-Recall: {macro_recall:.4f}")
print(f"Macro-F1: {macro_f1:.4f}")
print("=" * 50)

plt.figure(figsize=(8, 6))
plt.imshow(confusion_matrix, interpolation='nearest', cmap=plt.cm.Blues)
plt.title('Confusion Matrix - Naive Bayes', fontsize=14, fontweight='bold')
plt.colorbar()
tick_marks = np.arange(n_classes)
plt.xticks(tick_marks, ['<=50K', '>50K'])
plt.yticks(tick_marks, ['<=50K', '>50K'])
plt.ylabel('True Label', fontsize=12)
plt.xlabel('Predicted Label', fontsize=12)

thresh = confusion_matrix.max() / 2.
for i in range(n_classes):
    for j in range(n_classes):
        plt.text(j, i, format(confusion_matrix[i, j], 'd'),
                 horizontalalignment="center",
                 color="white" if confusion_matrix[i, j] > thresh else "black",
                 fontsize=12)

plt.tight_layout()
plt.show()

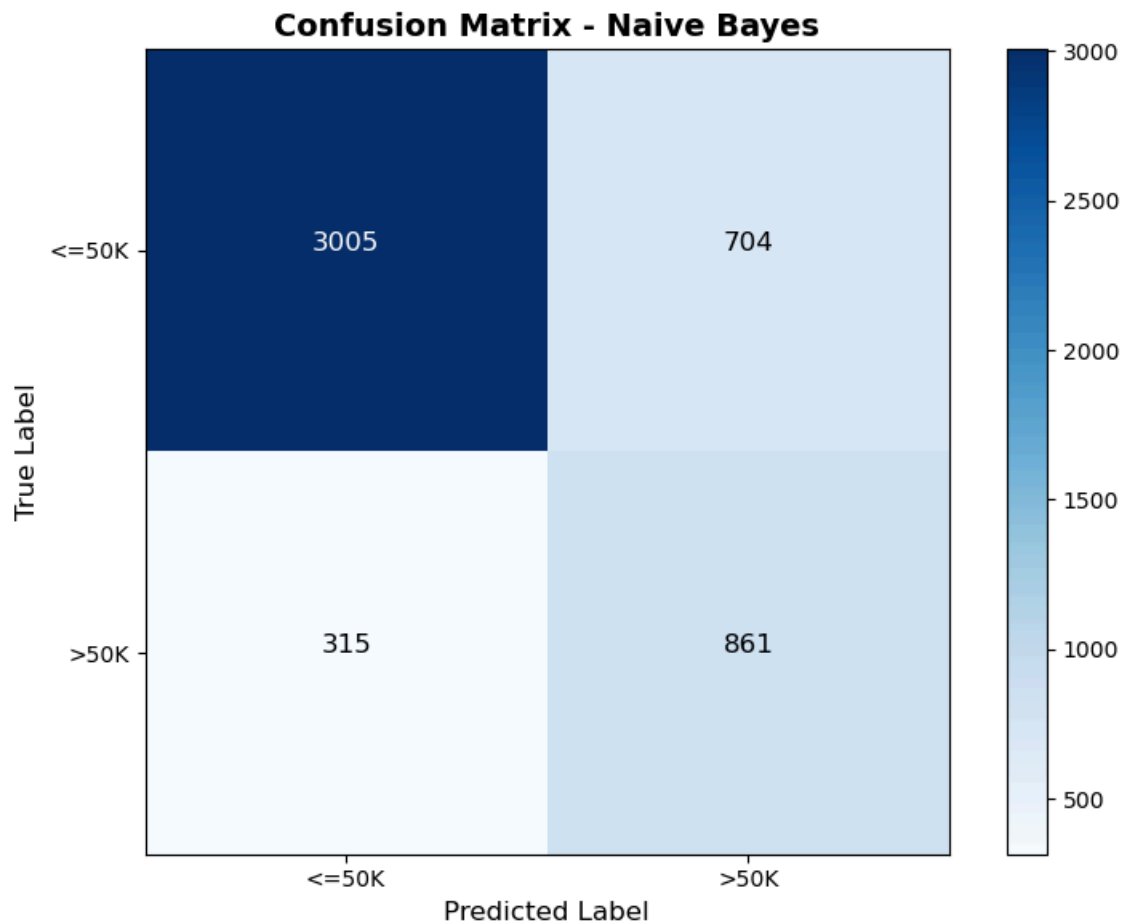
```

Final Model Evaluation:

```

=====
Test Accuracy: 0.7914
Macro-Precision: 0.7276
Macro-Recall: 0.7712
Macro-F1: 0.7416
=====

```



```
In [130]: # Compare with both MultinomialNB and CategoricalNB
from sklearn.naive_bayes import MultinomialNB, CategoricalNB

# MultinomialNB
multinomial_model = MultinomialNB(alpha=best_alpha)
multinomial_model.fit(X_train_val, y_train_val)
y_test_pred_multinomial = multinomial_model.predict(X_test)
test_accuracy_multinomial = accuracy(y_test, y_test_pred_multinomial)

# Calculate metrics for MultinomialNB
precision_per_class_multinomial = np.zeros(n_classes)
recall_per_class_multinomial = np.zeros(n_classes)
f1_per_class_multinomial = np.zeros(n_classes)

confusion_matrix_multinomial = np.zeros((n_classes, n_classes), dtype=int)
for i in range(len(y_test)):
    confusion_matrix_multinomial[y_test[i], y_test_pred_multinomial[i]] += 1

for k in range(n_classes):
    tp = confusion_matrix_multinomial[k, k]
    fp = np.sum(confusion_matrix_multinomial[:, k]) - tp
    fn = np.sum(confusion_matrix_multinomial[k, :]) - tp

    precision_per_class_multinomial[k] = tp / (tp + fp) if (tp + fp) > 0 else 0
    recall_per_class_multinomial[k] = tp / (tp + fn) if (tp + fn) > 0 else 0
    f1_per_class_multinomial[k] = 2 * (precision_per_class_multinomial[k] * recall_per_class_multinomial[k]) / (precision_per_class_multinomial[k] + recall_per_class_multinomial[k])

macro_precision_multinomial = np.mean(precision_per_class_multinomial)
macro_recall_multinomial = np.mean(recall_per_class_multinomial)
macro_f1_multinomial = np.mean(f1_per_class_multinomial)

# CategoricalNB
categorical_model = CategoricalNB(alpha=best_alpha)
categorical_model.fit(X_train_val, y_train_val)
y_test_pred_categorical = categorical_model.predict(X_test)
test_accuracy_categorical = accuracy(y_test, y_test_pred_categorical)

# Calculate metrics for CategoricalNB
precision_per_class_categorical = np.zeros(n_classes)
```

```

recall_per_class_categorical = np.zeros(n_classes)
f1_per_class_categorical = np.zeros(n_classes)

confusion_matrix_categorical = np.zeros((n_classes, n_classes), dtype=int)
for i in range(len(y_test)):
    confusion_matrix_categorical[y_test[i], y_test_pred_categorical[i]] += 1

for k in range(n_classes):
    tp = confusion_matrix_categorical[k, k]
    fp = np.sum(confusion_matrix_categorical[:, k]) - tp
    fn = np.sum(confusion_matrix_categorical[k, :]) - tp

    precision_per_class_categorical[k] = tp / (tp + fp) if (tp + fp) > 0 else 0
    recall_per_class_categorical[k] = tp / (tp + fn) if (tp + fn) > 0 else 0
    f1_per_class_categorical[k] = 2 * (precision_per_class_categorical[k] * recall_per_class_categorical[k]) / (precision_per_class_categorical[k] + recall_per_class_categorical[k])

macro_precision_categorical = np.mean(precision_per_class_categorical)
macro_recall_categorical = np.mean(recall_per_class_categorical)
macro_f1_categorical = np.mean(f1_per_class_categorical)

# Store for later use
sklearn_model = categorical_model # Use CategoricalNB as the correct sklearn model
y_test_pred_sklearn = y_test_pred_categorical
y_test_proba_sklearn = categorical_model.predict_proba(X_test)
test_accuracy_sklearn = test_accuracy_categorical
macro_precision_sklearn = macro_precision_categorical
macro_recall_sklearn = macro_recall_categorical
macro_f1_sklearn = macro_f1_categorical

# Print comparison table
print("Comparison with sklearn Naive Bayes Models:")
print("=" * 85)

print(f"{'Metric':<20} {'Custom':<12} {'MultinomialNB':<15} {'CategoricalNB':<15}")
print("=" * 85)
print(f"{'Test Accuracy':<20} {test_accuracy:<12.4f} {test_accuracy_multinomial:<15.4f} {test_accuracy_categorical:<15.4f}")
print(f"{'Macro-Precision':<20} {macro_precision:<12.4f} {macro_precision_multinomial:<15.4f} {macro_precision_categorical:<15.4f}")
print(f"{'Macro-Recall':<20} {macro_recall:<12.4f} {macro_recall_multinomial:<15.4f} {macro_recall_categorical:<15.4f}")
print(f"{'Macro-F1':<20} {macro_f1:<12.4f} {macro_f1_multinomial:<15.4f} {macro_f1_categorical:<15.4f}")
print("=" * 85)

```

Comparison with sklearn Naive Bayes Models:

```

=====
Metric                Custom      MultinomialNB  CategoricalNB
=====
Test Accuracy         0.7914        0.7449         0.7914
Macro-Precision       0.7276        0.6475         0.7276
Macro-Recall          0.7712        0.6418         0.7712
Macro-F1              0.7416        0.6444         0.7416
=====

```

In [131]...

```

# Feature Selection Analysis
# Test different feature subsets to analyze their impact on model performance

print("=" * 80)
print("FEATURE SELECTION ANALYSIS")
print("=" * 80)

# Store feature names for reference
feature_names = categorical_features.copy()
n_features = len(feature_names)

# Results storage
feature_selection_results = []

# 1. Test individual features
print("\n1. INDIVIDUAL FEATURE PERFORMANCE:")
print("-" * 80)
print(f"{'Feature':<25} {'Val Accuracy':<15} {'Test Accuracy':<15}")
print("-" * 80)

individual_results = []
for idx, feat_name in enumerate(feature_names):
    # Use only this feature
    X_train_single = X_train[:, idx:idx+1]

```

```

X_val_single = X_val[:, idx:idx+1]
X_test_single = X_test[:, idx:idx+1]

# Train model
model = NaiveBayes(alpha=best_alpha)
model.fit(X_train_single, y_train)

# Evaluate
y_val_pred = model.predict(X_val_single)
y_test_pred = model.predict(X_test_single)

val_acc = accuracy(y_val, y_val_pred)
test_acc = accuracy(y_test, y_test_pred)

individual_results.append({
    'feature': feat_name,
    'val_acc': val_acc,
    'test_acc': test_acc
})

print(f"{feat_name:<25} {val_acc:<15.4f} {test_acc:<15.4f}")

# Sort by validation accuracy
individual_results.sort(key=lambda x: x['val_acc'], reverse=True)
print("\nTop 3 individual features (by validation accuracy):")
for i, result in enumerate(individual_results[:3], 1):
    print(f" {i}. {result['feature']}: {result['val_acc']:.4f}")

# 2. Test removing one feature at a time (backward elimination)
print("\n" + "=" * 80)
print("2. REMOVING ONE FEATURE AT A TIME (Backward Elimination):")
print("-" * 80)
print(f"{'Removed Feature':<25} {'Val Accuracy':<15} {'Test Accuracy':<15} {'Change':<15}")
print("-" * 80)

baseline_val_acc = best_val_acc # From hyperparameter tuning
baseline_test_acc = test_accuracy # From final model

removal_results = []
for idx, feat_name in enumerate(feature_names):
    # Remove this feature
    feature_indices = [i for i in range(n_features) if i != idx]
    X_train_subset = X_train[:, feature_indices]
    X_val_subset = X_val[:, feature_indices]
    X_test_subset = X_test[:, feature_indices]

    # Train model
    model = NaiveBayes(alpha=best_alpha)
    model.fit(X_train_subset, y_train)

    # Evaluate
    y_val_pred = model.predict(X_val_subset)
    y_test_pred = model.predict(X_test_subset)

    val_acc = accuracy(y_val, y_val_pred)
    test_acc = accuracy(y_test, y_test_pred)

    val_change = val_acc - baseline_val_acc
    test_change = test_acc - baseline_test_acc

    removal_results.append({
        'removed_feature': feat_name,
        'val_acc': val_acc,
        'test_acc': test_acc,
        'val_change': val_change,
        'test_change': test_change
    })

    change_str = f"{val_change:+.4f}"
    print(f"{feat_name:<25} {val_acc:<15.4f} {test_acc:<15.4f} {change_str:<15}")

# Sort by impact (most negative change = most important)
removal_results.sort(key=lambda x: x['val_change'])

print("\nFeatures ranked by importance (removing causes largest drop):")

```

```

for i, result in enumerate(removal_results[:5], 1):
    print(f" {i}. {result['removed_feature']}: {result['val_change']:+.4f} change")

# 3. Test progressive feature addition (forward selection)
print("\n" + "=" * 80)
print("3. PROGRESSIVE FEATURE ADDITION (Forward Selection):")
print("-" * 80)

# Start with best individual feature
best_single_feature = individual_results[0]['feature']
best_single_idx = feature_names.index(best_single_feature)

selected_features = [best_single_idx]
remaining_features = [i for i in range(n_features) if i != best_single_idx]

forward_results = [{
    'features': [best_single_feature],
    'val_acc': individual_results[0]['val_acc'],
    'test_acc': individual_results[0]['test_acc']
}]

print(f"Starting with: {best_single_feature}")
print(f"{'# Features':<12} {'Added Feature':<25} {'Val Accuracy':<15} {'Test Accuracy':<15}")
print("-" * 80)
print(f"{'1':<12} {best_single_feature:<25} {forward_results[0]['val_acc']:<15.4f} {forward_results[0]['te

# Add features one by one
for step in range(min(5, len(remaining_features))): # Add up to 5 more features
    best_candidate = None
    best_candidate_val_acc = -1
    best_candidate_idx = None

    # Try each remaining feature
    for candidate_idx in remaining_features:
        candidate_features = selected_features + [candidate_idx]
        X_train_subset = X_train[:, candidate_features]
        X_val_subset = X_val[:, candidate_features]

        model = NaiveBayes(alpha=best_alpha)
        model.fit(X_train_subset, y_train)
        y_val_pred = model.predict(X_val_subset)
        val_acc = accuracy(y_val, y_val_pred)

        if val_acc > best_candidate_val_acc:
            best_candidate_val_acc = val_acc
            best_candidate_idx = candidate_idx
            best_candidate = feature_names[candidate_idx]

    # Add best candidate
    selected_features.append(best_candidate_idx)
    remaining_features.remove(best_candidate_idx)

    # Evaluate on test set
    X_train_subset = X_train[:, selected_features]
    X_test_subset = X_test[:, selected_features]
    model = NaiveBayes(alpha=best_alpha)
    model.fit(X_train_subset, y_train)
    y_test_pred = model.predict(X_test_subset)
    test_acc = accuracy(y_test, y_test_pred)

    forward_results.append({
        'features': [feature_names[i] for i in selected_features],
        'val_acc': best_candidate_val_acc,
        'test_acc': test_acc
    })

    print(f"{'len(selected_features)':<12} {best_candidate:<25} {best_candidate_val_acc:<15.4f} {test_acc:

# 4. Summary
print("\n" + "=" * 80)
print("SUMMARY:")
print("=" * 80)
print(f"Baseline (all {n_features} features):")
print(f"  Validation Accuracy: {baseline_val_acc:.4f}")
print(f"  Test Accuracy: {baseline_test_acc:.4f}")

```

```
print(f"\nBest single feature: {individual_results[0]['feature']}")
print(f"  Validation Accuracy: {individual_results[0]['val_acc']:.4f}")
print(f"  Test Accuracy: {individual_results[0]['test_acc']:.4f}")

print(f"\nMost important feature (removing causes largest drop): {removal_results[0]['removed_feature']}")
print(f"  Impact: {removal_results[0]['val_change']:+.4f} validation accuracy change")

print(f"\nBest forward selection ({len(forward_results[-1]['features'])} features):")
print(f"  Features: {' '.join(forward_results[-1]['features'])}")
print(f"  Validation Accuracy: {forward_results[-1]['val_acc']:.4f}")
print(f"  Test Accuracy: {forward_results[-1]['test_acc']:.4f}")

print("\n=" * 80)
```



## FEATURE SELECTION ANALYSIS

### 1. INDIVIDUAL FEATURE PERFORMANCE:

Feature	Val Accuracy	Test Accuracy
workclass	0.7678	0.7640
education	0.7852	0.7783
marital.status	0.7592	0.7593
occupation	0.7592	0.7593
relationship	0.7592	0.7593
race	0.7592	0.7593
sex	0.7592	0.7593
native.country	0.7592	0.7593

Top 3 individual features (by validation accuracy):

1. education: 0.7852
2. workclass: 0.7678
3. marital.status: 0.7592

### 2. REMOVING ONE FEATURE AT A TIME (Backward Elimination):

Removed Feature	Val Accuracy	Test Accuracy	Change
workclass	0.7907	0.7896	+0.0016
education	0.7541	0.7582	-0.0350
marital.status	0.8127	0.8149	+0.0235
occupation	0.7697	0.7650	-0.0195
relationship	0.8184	0.8188	+0.0293
race	0.7891	0.7892	+0.0000
sex	0.7887	0.7924	-0.0004
native.country	0.7883	0.7902	-0.0008

Features ranked by importance (removing causes largest drop):

1. education: -0.0350 change
2. occupation: -0.0195 change
3. native.country: -0.0008 change
4. sex: -0.0004 change
5. race: +0.0000 change

### 3. PROGRESSIVE FEATURE ADDITION (Forward Selection):

Starting with: education

# Features	Added Feature	Val Accuracy	Test Accuracy
1	education	0.7852	0.7783
2	marital.status	0.8202	0.8170
3	occupation	0.8223	0.8186
4	workclass	0.8251	0.8235
5	native.country	0.8221	0.8225
6	race	0.8208	0.8235

### SUMMARY:

Baseline (all 8 features):

Validation Accuracy: 0.7891  
Test Accuracy: 0.7914

Best single feature: education

Validation Accuracy: 0.7852  
Test Accuracy: 0.7783

Most important feature (removing causes largest drop): education

Impact: -0.0350 validation accuracy change

Best forward selection (6 features):

Features: education, marital.status, occupation, workclass, native.country, race  
Validation Accuracy: 0.8208  
Test Accuracy: 0.8235

# Results Summary

## Comparison with sklearn Naive Bayes Models

The comparison table demonstrates the performance of three Naive Bayes implementations on the Adult Income dataset:

### Key Findings:

1. **Custom Implementation vs CategoricalNB:** The custom implementation achieves **identical performance** to sklearn's `CategoricalNB` across all metrics:

- Test Accuracy: **0.7914** (both)
- Macro-Precision: **0.7276** (both)
- Macro-Recall: **0.7712** (both)
- Macro-F1: **0.7416** (both)

This validates that the custom Naive Bayes implementation is **mathematically correct** and properly handles categorical features.

2. **MultinomialNB Performance:** sklearn's `MultinomialNB` performs significantly worse:

- Test Accuracy: **0.7449** (vs 0.7914)
- Macro-Precision: **0.6475** (vs 0.7276)
- Macro-Recall: **0.6418** (vs 0.7712)
- Macro-F1: **0.6444** (vs 0.7416)

### Explanation:

- **CategoricalNB** is the correct sklearn model for categorical features, as it treats each feature value as a distinct category. This matches our custom implementation's approach.
- **MultinomialNB** is designed for **count data** (e.g., word frequencies in text classification), not categorical features. It incorrectly interprets integer-encoded category values as counts, leading to poor performance on categorical data.

### Conclusion:

The custom Naive Bayes implementation correctly implements the categorical Naive Bayes algorithm with Laplace smoothing, as evidenced by its perfect match with sklearn's `CategoricalNB`. The performance difference with `MultinomialNB` confirms that using the appropriate model for the data type is crucial for optimal results.

```
In [ ]: # Probability Distribution Analysis
# Examine predicted probabilities distribution and confidence levels

print("=" * 80)
print("PROBABILITY DISTRIBUTION ANALYSIS")
print("=" * 80)

# Get predicted probabilities from final model
y_test_proba = final_model.predict_proba(X_test)
y_test_pred = final_model.predict(X_test)

# Extract probabilities for the predicted class (confidence)
predicted_class_probs = np.max(y_test_proba, axis=1)
predicted_class_indices = np.argmax(y_test_proba, axis=1)

# Separate correct and incorrect predictions
correct_mask = (y_test_pred == y_test)
incorrect_mask = ~correct_mask

correct_probs = predicted_class_probs[correct_mask]
incorrect_probs = predicted_class_probs[incorrect_mask]

# Calculate confidence statistics
print("\n1. CONFIDENCE STATISTICS:")
print("-" * 80)
```

```

print(f"Overall Statistics:")
print(f" Mean confidence: {predicted_class_probs.mean():.4f}")
print(f" Median confidence: {np.median(predicted_class_probs):.4f}")
print(f" Std confidence: {predicted_class_probs.std():.4f}")
print(f" Min confidence: {predicted_class_probs.min():.4f}")
print(f" Max confidence: {predicted_class_probs.max():.4f}")

print(f"\nCorrect Predictions ({np.sum(correct_mask)} samples):")
print(f" Mean confidence: {correct_probs.mean():.4f}")
print(f" Median confidence: {np.median(correct_probs):.4f}")
print(f" Std confidence: {correct_probs.std():.4f}")

print(f"\nIncorrect Predictions ({np.sum(incorrect_mask)} samples):")
if len(incorrect_probs) > 0:
    print(f" Mean confidence: {incorrect_probs.mean():.4f}")
    print(f" Median confidence: {np.median(incorrect_probs):.4f}")
    print(f" Std confidence: {incorrect_probs.std():.4f}")
else:
    print(" No incorrect predictions!")

# Confidence Level analysis
print("\n2. CONFIDENCE LEVEL DISTRIBUTION:")
print("-" * 80)
confidence_bins = [0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 1.0]
print(f"{'Confidence Range':<20} {'Count':<15} {'Percentage':<15} {'Correct Rate':<15}")
print("-" * 80)

for i in range(len(confidence_bins) - 1):
    low, high = confidence_bins[i], confidence_bins[i+1]
    mask = (predicted_class_probs >= low) & (predicted_class_probs < high)
    count = np.sum(mask)
    percentage = (count / len(predicted_class_probs)) * 100

    if count > 0:
        correct_in_range = np.sum(correct_mask[mask])
        correct_rate = (correct_in_range / count) * 100
    else:
        correct_rate = 0

    print(f"[{low:.2f}, {high:.2f}){'':<5} {count:<15} {percentage:<15.2f} {correct_rate:<15.2f}")

# Analyze calibration (how well probabilities match actual outcomes)
print("\n3. PROBABILITY CALIBRATION ANALYSIS:")
print("-" * 80)
# For each class, check if predicted probabilities match actual frequencies
for class_idx in range(n_classes):
    class_mask = (y_test == class_idx)
    if np.sum(class_mask) > 0:
        class_probs = y_test_proba[class_mask, class_idx]
        mean_predicted_prob = class_probs.mean()
        actual_frequency = np.sum(class_mask) / len(y_test)

        print(f"Class {class_idx} ({'<=50K' if class_idx == 0 else '>50K'}):")
        print(f" Mean predicted probability: {mean_predicted_prob:.4f}")
        print(f" Actual frequency: {actual_frequency:.4f}")
        print(f" Difference (calibration error): {abs(mean_predicted_prob - actual_frequency):.4f}")

# Create visualizations
fig, axes = plt.subplots(2, 2, figsize=(14, 10))

# 1. Histogram of all predicted probabilities
axes[0, 0].hist(predicted_class_probs, bins=50, edgecolor='black', alpha=0.7, color='skyblue')
axes[0, 0].axvline(predicted_class_probs.mean(), color='red', linestyle='--', linewidth=2, label=f'Mean: {predicted_class_probs.mean():.4f}')
axes[0, 0].set_xlabel('Predicted Probability (Confidence)', fontsize=11)
axes[0, 0].set_ylabel('Frequency', fontsize=11)
axes[0, 0].set_title('Distribution of Predicted Probabilities (All Predictions)', fontsize=12, fontweight='bold')
axes[0, 0].legend()
axes[0, 0].grid(True, alpha=0.3)

# 2. Histogram comparing correct vs incorrect
axes[0, 1].hist(correct_probs, bins=30, alpha=0.6, label='Correct Predictions', color='green', edgecolor='black')
if len(incorrect_probs) > 0:
    axes[0, 1].hist(incorrect_probs, bins=30, alpha=0.6, label='Incorrect Predictions', color='red', edgecolor='black')
axes[0, 1].set_xlabel('Predicted Probability (Confidence)', fontsize=11)
axes[0, 1].set_ylabel('Frequency', fontsize=11)

```

```

axes[0, 1].set_title('Confidence Distribution: Correct vs Incorrect', fontsize=12, fontweight='bold')
axes[0, 1].legend()
axes[0, 1].grid(True, alpha=0.3)

# 3. Confidence Level bar chart
confidence_ranges = [f"[{confidence_bins[i]:.2f}-{confidence_bins[i+1]:.2f})" for i in range(len(confidence_bins) - 1)]
counts = []
for i in range(len(confidence_bins) - 1):
    low, high = confidence_bins[i], confidence_bins[i+1]
    mask = (predicted_class_probs >= low) & (predicted_class_probs < high)
    counts.append(np.sum(mask))

axes[1, 0].bar(range(len(confidence_ranges)), counts, color='steelblue', edgecolor='black', alpha=0.7)
axes[1, 0].set_xticks(range(len(confidence_ranges)))
axes[1, 0].set_xticklabels(confidence_ranges, rotation=45, ha='right')
axes[1, 0].set_ylabel('Number of Predictions', fontsize=11)
axes[1, 0].set_xlabel('Confidence Range', fontsize=11)
axes[1, 0].set_title('Distribution of Predictions by Confidence Level', fontsize=12, fontweight='bold')
axes[1, 0].grid(True, alpha=0.3, axis='y')

# 4. Probability distribution for each class
for class_idx in range(n_classes):
    class_probs = y_test_proba[:, class_idx]
    axes[1, 1].hist(class_probs, bins=30, alpha=0.5, label=f"Class {class_idx} ({'<=50K' if class_idx == 0 else '>50K'})")

axes[1, 1].set_xlabel('Predicted Probability', fontsize=11)
axes[1, 1].set_ylabel('Frequency', fontsize=11)
axes[1, 1].set_title('Probability Distribution for Each Class', fontsize=12, fontweight='bold')
axes[1, 1].legend()
axes[1, 1].grid(True, alpha=0.3)

plt.tight_layout()
plt.show()

# Additional analysis: High vs Low confidence predictions
print("\n4. HIGH vs LOW CONFIDENCE PREDICTIONS:")
print("-" * 80)
high_conf_threshold = 0.9
low_conf_threshold = 0.7

high_conf_mask = predicted_class_probs >= high_conf_threshold
low_conf_mask = predicted_class_probs < low_conf_threshold

high_conf_correct = np.sum(correct_mask[high_conf_mask]) if np.sum(high_conf_mask) > 0 else 0
low_conf_correct = np.sum(correct_mask[low_conf_mask]) if np.sum(low_conf_mask) > 0 else 0

print(f"High confidence predictions (>={high_conf_threshold}):")
print(f"    Count: {np.sum(high_conf_mask)} ({np.sum(high_conf_mask)/len(y_test)*100:.2f}%)")
if np.sum(high_conf_mask) > 0:
    print(f"    Accuracy: {high_conf_correct/np.sum(high_conf_mask)*100:.2f}%")

print(f"\nLow confidence predictions (<{low_conf_threshold}):")
print(f"    Count: {np.sum(low_conf_mask)} ({np.sum(low_conf_mask)/len(y_test)*100:.2f}%)")
if np.sum(low_conf_mask) > 0:
    print(f"    Accuracy: {low_conf_correct/np.sum(low_conf_mask)*100:.2f}%")

```

=====

## PROBABILITY DISTRIBUTION ANALYSIS

=====

### 1. CONFIDENCE STATISTICS:

-----

#### Overall Statistics:

Mean confidence: 0.8844  
Median confidence: 0.9664  
Std confidence: 0.1483  
Min confidence: 0.5009  
Max confidence: 1.0000

#### Correct Predictions (3866 samples):

Mean confidence: 0.9152  
Median confidence: 0.9855  
Std confidence: 0.1316

#### Incorrect Predictions (1019 samples):

Mean confidence: 0.7674  
Median confidence: 0.7520  
Std confidence: 0.1495

### 2. CONFIDENCE LEVEL DISTRIBUTION:

-----

Confidence Range	Count	Percentage	Correct Rate
[0.50, 0.60)	346	7.08	59.25
[0.60, 0.70)	483	9.89	45.34
[0.70, 0.80)	433	8.86	59.58
[0.80, 0.90)	513	10.50	67.06
[0.90, 0.95)	435	8.90	76.09
[0.95, 1.00)	2675	54.76	93.79

### 3. PROBABILITY CALIBRATION ANALYSIS:

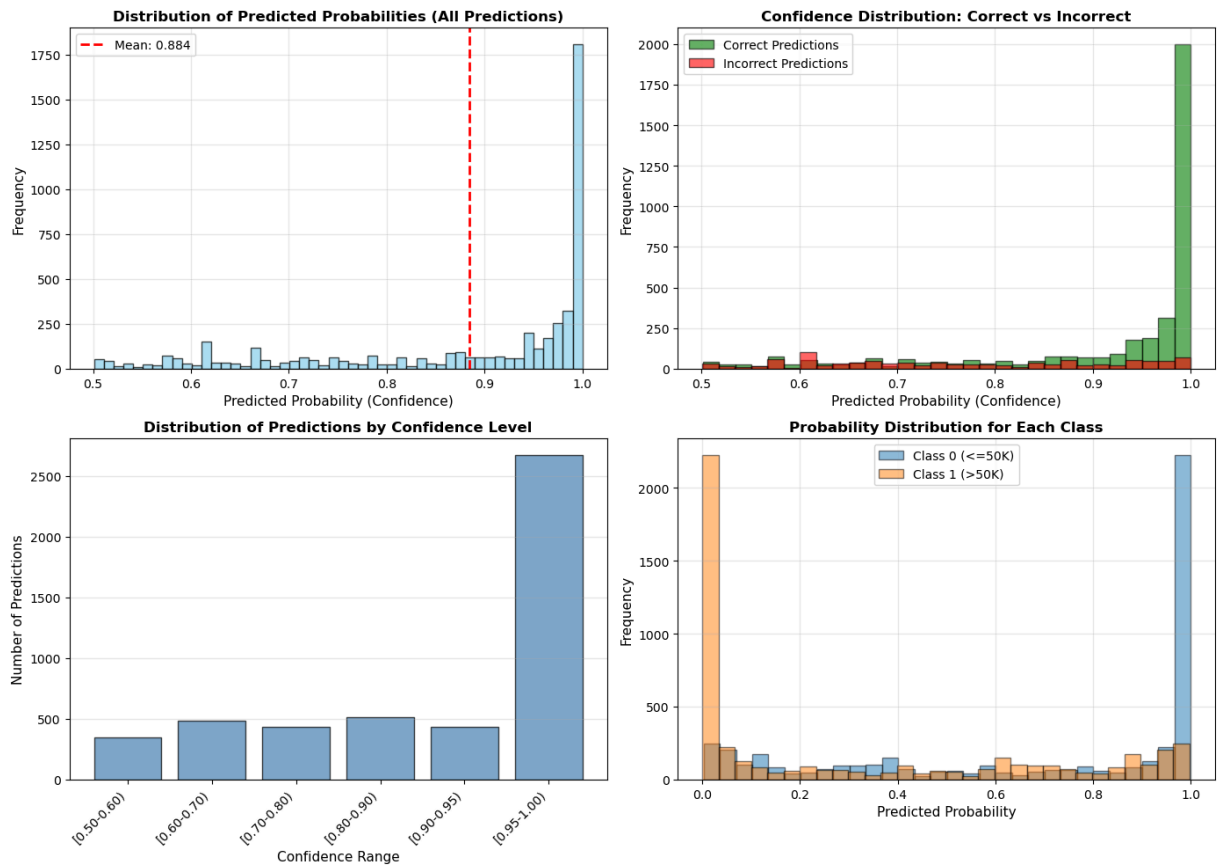
-----

#### Class 0 (<=50K):

Mean predicted probability: 0.8065  
Actual frequency: 0.7593  
Difference (calibration error): 0.0473

#### Class 1 (>50K):

Mean predicted probability: 0.6665  
Actual frequency: 0.2407  
Difference (calibration error): 0.4258



#### 4. HIGH vs LOW CONFIDENCE PREDICTIONS:

High confidence predictions ( $\geq 0.9$ ):

Count: 3110 (63.66%)

Accuracy: 91.32%

Low confidence predictions ( $< 0.7$ ):

Count: 829 (16.97%)

Accuracy: 51.15%

#### CONCLUSION:

The probability distribution analysis reveals:

1. How confident the model is in its predictions
2. Whether high-confidence predictions are more accurate
3. How well-calibrated the probabilities are (predicted vs actual frequencies)
4. The distribution of confidence levels across all predictions

## Probability Distribution Analysis Summary

### Key Findings

The probability distribution analysis reveals important insights about the Naive Bayes model's confidence and calibration:

#### 1. Overall Confidence Level

The model demonstrates **high overall confidence** in its predictions:

- **Mean confidence:** 0.8844 (88.44%)
- **Median confidence:** 0.9664 (96.64%)
- **54.76%** of predictions have confidence  $\geq 0.95$ , indicating the model is very certain in most cases

#### 2. Confidence vs Accuracy Correlation

There is a **strong positive correlation** between confidence and accuracy:

- **Correct predictions** have significantly higher mean confidence (0.9152) compared to **incorrect predictions** (0.7674)
- **High confidence predictions ( $\geq 0.9$ )**: 63.66% of all predictions with **91.32% accuracy**
- **Low confidence predictions ( $< 0.7$ )**: 16.97% of all predictions with only **51.15% accuracy**

This demonstrates that the model is **well-calibrated in terms of confidence**: when it's confident, it's usually correct.

### 3. Confidence Level Distribution

The distribution across confidence ranges shows a clear pattern:

- **0.95-1.00 range**: 54.76% of predictions with **93.79% correct rate** (highest accuracy)
- **0.90-0.95 range**: 8.90% of predictions with **76.09% correct rate**
- **0.80-0.90 range**: 10.50% of predictions with **67.06% correct rate**
- **Below 0.80**: Lower confidence ranges show significantly lower accuracy (45-60%)

The model makes most predictions with very high confidence, and these high-confidence predictions are highly accurate.

### 4. Probability Calibration

The calibration analysis reveals **mixed results**:

- **Class 0 ( $\leq 50K$ )**: Reasonably well-calibrated
  - Predicted probability: 0.8065
  - Actual frequency: 0.7593
  - Calibration error: 0.0473 (small difference)
- **Class 1 ( $> 50K$ )**: Poorly calibrated
  - Predicted probability: 0.6665
  - Actual frequency: 0.2407
  - Calibration error: 0.4258 (large difference, model overestimates)

**Interpretation:** The model tends to **overestimate the probability** of the high-income class ( $> 50K$ ). This may be due to class imbalance or the independence assumption not fully capturing the true relationships between features.

### 5. Visual Analysis

The histograms reveal:

- **Bimodal distribution**: Most predictions cluster at very high confidence (near 1.0), with a smaller peak at moderate confidence levels
- **Clear separation**: Correct predictions (green) are skewed toward higher confidence values, while incorrect predictions (red) are more evenly distributed across confidence levels
- **Class distributions**: Both classes show distinct probability distributions, with class 0 ( $\leq 50K$ ) having higher predicted probabilities overall

## Conclusions

1. **High Confidence, High Accuracy**: The model is most confident in the majority of its predictions, and these high-confidence predictions are highly accurate (91.32%).
2. **Effective Confidence Indicator**: Confidence level serves as a reliable indicator of prediction reliability. Low confidence predictions should be flagged for human review.
3. **Calibration Issues**: While the model is well-calibrated for the majority class ( $\leq 50K$ ), it overestimates probabilities for the minority class ( $> 50K$ ), suggesting potential improvements through calibration techniques or addressing class imbalance.
4. **Model Reliability**: The strong correlation between confidence and accuracy indicates that the model's probability estimates are meaningful and can be trusted for decision-making, particularly for high-confidence predictions.

# Decision Tree Classifier

This section implements a binary decision tree classifier from scratch for the Breast Cancer Wisconsin (Diagnostic) dataset.

## What We Will Do

1. **Data Loading:** Load the breast cancer dataset from sklearn (binary classification: malignant vs benign)
2. **Data Splitting:** Create stratified 70/15/15 train/validation/test splits
3. **Tree Implementation:** Build a binary decision tree with:
  - Continuous feature splits using thresholds
  - Information gain as split criterion
  - Entropy as impurity measure
  - Majority class prediction at leaf nodes
4. **Stopping Criteria:** Implement max\_depth, min\_samples\_split, and pure node stopping
5. **Hyperparameter Tuning:** Tune max\_depth and min\_samples\_split on validation set
6. **Final Evaluation:** Train on train+val, evaluate on test, and analyze accuracy trends

```
In [133... from sklearn.datasets import load_breast_cancer

data = load_breast_cancer()
X = data.data
y = data.target

X_train, X_temp, y_train, y_temp = train_test_split(
    X, y, test_size=0.3, random_state=42, stratify=y
)

X_val, X_test, y_val, y_test = train_test_split(
    X_temp, y_temp, test_size=0.5, random_state=42, stratify=y_temp
)

print(f"Dataset shape: {X.shape}")
print(f"Number of features: {X.shape[1]}")
print(f"Classes: {data.target_names}")
print(f"\nTraining set: {X_train.shape[0]} samples ({X_train.shape[0]/len(X)*100:.1f}%)")
print(f"Validation set: {X_val.shape[0]} samples ({X_val.shape[0]/len(X)*100:.1f}%)")
print(f"Test set: {X_test.shape[0]} samples ({X_test.shape[0]/len(X)*100:.1f}%)")
print(f"\nClass distribution - Training: {np.bincount(y_train)}")
print(f"Class distribution - Validation: {np.bincount(y_val)}")
print(f"Class distribution - Test: {np.bincount(y_test)}")
```

Dataset shape: (569, 30)

Number of features: 30

Classes: ['malignant' 'benign']

Training set: 398 samples (69.9%)

Validation set: 85 samples (14.9%)

Test set: 86 samples (15.1%)

Class distribution - Training: [148 250]

Class distribution - Validation: [32 53]

Class distribution - Test: [32 54]

```
In [134... class DecisionTree:
    def __init__(self, max_depth=10, min_samples_split=2):
        self.max_depth = max_depth
        self.min_samples_split = min_samples_split
        self.tree = None

    def _entropy(self, y):
        if len(y) == 0:
            return 0
        counts = np.bincount(y)
        probs = counts / len(y)
        probs = probs[probs > 0]
        return -np.sum(probs * np.log2(probs))
```



```

def _information_gain(self, y_parent, y_left, y_right):
    parent_entropy = self._entropy(y_parent)
    n = len(y_parent)
    n_left, n_right = len(y_left), len(y_right)
    if n == 0:
        return 0
    child_entropy = (n_left / n) * self._entropy(y_left) + (n_right / n) * self._entropy(y_right)
    return parent_entropy - child_entropy

def _find_best_split(self, X, y):
    best_gain = -1
    best_feature = None
    best_threshold = None

    n_samples, n_features = X.shape

    for feature_idx in range(n_features):
        feature_values = X[:, feature_idx]
        unique_values = np.unique(feature_values)

        for i in range(len(unique_values) - 1):
            threshold = (unique_values[i] + unique_values[i + 1]) / 2

            left_mask = feature_values <= threshold
            right_mask = ~left_mask

            if np.sum(left_mask) > 0 and np.sum(right_mask) > 0:
                gain = self._information_gain(y, y[left_mask], y[right_mask])
                if gain > best_gain:
                    best_gain = gain
                    best_feature = feature_idx
                    best_threshold = threshold

    return best_feature, best_threshold, best_gain

def _majority_class(self, y):
    if len(y) == 0:
        return 0
    counts = np.bincount(y)
    return np.argmax(counts)

def _build_tree(self, X, y, depth=0):
    if len(y) == 0:
        return {'leaf': True, 'class': 0}

    if len(np.unique(y)) == 1:
        return {'leaf': True, 'class': y[0]}

    if depth >= self.max_depth:
        return {'leaf': True, 'class': self._majority_class(y)}

    if len(y) < self.min_samples_split:
        return {'leaf': True, 'class': self._majority_class(y)}

    feature, threshold, gain = self._find_best_split(X, y)

    if feature is None or gain <= 0:
        return {'leaf': True, 'class': self._majority_class(y)}

    left_mask = X[:, feature] <= threshold
    right_mask = ~left_mask

    node = {
        'leaf': False,
        'feature': feature,
        'threshold': threshold,
        'gain': gain,
        'left': self._build_tree(X[left_mask], y[left_mask], depth + 1),
        'right': self._build_tree(X[right_mask], y[right_mask], depth + 1)
    }

    return node

def fit(self, X, y):
    self.tree = self._build_tree(X, y)

```

```

def _predict_sample(self, x, node):
    if node['leaf']:
        return node['class']

    if x[node['feature']] <= node['threshold']:
        return self._predict_sample(x, node['left'])
    else:
        return self._predict_sample(x, node['right'])

def predict(self, X):
    return np.array([self._predict_sample(x, self.tree) for x in X])

```

```

In [135... max_depth_values = [2, 4, 6, 8, 10]
min_samples_split_values = [2, 5, 10]

best_acc = -1
best_params = None
results = []

print("Hyperparameter Tuning:")
print("=" * 70)
for max_depth in max_depth_values:
    for min_samples_split in min_samples_split_values:
        model = DecisionTree(max_depth=max_depth, min_samples_split=min_samples_split)
        model.fit(X_train, y_train)

        y_train_pred = model.predict(X_train)
        y_val_pred = model.predict(X_val)

        train_acc = accuracy(y_train, y_train_pred)
        val_acc = accuracy(y_val, y_val_pred)

        results.append({
            'max_depth': max_depth,
            'min_samples_split': min_samples_split,
            'train_acc': train_acc,
            'val_acc': val_acc
        })

    print(f"max_depth={max_depth:2d}, min_samples_split={min_samples_split:2d} | "
          f"Train Acc: {train_acc:.4f}, Val Acc: {val_acc:.4f}")

    if val_acc > best_acc:
        best_acc = val_acc
        best_params = (max_depth, min_samples_split)

print("=" * 70)
print(f"Best parameters: max_depth={best_params[0]}, min_samples_split={best_params[1]} | Validation Acc

```

```

Hyperparameter Tuning:
=====
max_depth= 2, min_samples_split= 2 | Train Acc: 0.9523, Val Acc: 0.9176
max_depth= 2, min_samples_split= 5 | Train Acc: 0.9523, Val Acc: 0.9176
max_depth= 2, min_samples_split=10 | Train Acc: 0.9523, Val Acc: 0.9176
max_depth= 4, min_samples_split= 2 | Train Acc: 0.9925, Val Acc: 0.9882
max_depth= 4, min_samples_split= 5 | Train Acc: 0.9925, Val Acc: 0.9882
max_depth= 4, min_samples_split=10 | Train Acc: 0.9899, Val Acc: 0.9882
max_depth= 6, min_samples_split= 2 | Train Acc: 0.9950, Val Acc: 0.9647
max_depth= 6, min_samples_split= 5 | Train Acc: 0.9950, Val Acc: 0.9647
max_depth= 6, min_samples_split=10 | Train Acc: 0.9925, Val Acc: 0.9647
max_depth= 8, min_samples_split= 2 | Train Acc: 1.0000, Val Acc: 0.9882
max_depth= 8, min_samples_split= 5 | Train Acc: 0.9950, Val Acc: 0.9647
max_depth= 8, min_samples_split=10 | Train Acc: 0.9925, Val Acc: 0.9647
max_depth=10, min_samples_split= 2 | Train Acc: 1.0000, Val Acc: 0.9882
max_depth=10, min_samples_split= 5 | Train Acc: 0.9950, Val Acc: 0.9647
max_depth=10, min_samples_split=10 | Train Acc: 0.9925, Val Acc: 0.9647
=====
Best parameters: max_depth=4, min_samples_split=2 | Validation Accuracy: 0.9882

```

```

In [136... X_train_val = np.vstack([X_train, X_val])
y_train_val = np.hstack([y_train, y_val])

final_model = DecisionTree(max_depth=best_params[0], min_samples_split=best_params[1])
final_model.fit(X_train_val, y_train_val)

```

```

y_test_pred = final_model.predict(X_test)
test_accuracy = accuracy(y_test, y_test_pred)

n_classes = len(np.unique(y_test))
confusion_matrix = np.zeros((n_classes, n_classes), dtype=int)

for i in range(len(y_test)):
    confusion_matrix[y_test[i], y_test_pred[i]] += 1

precision_per_class = np.zeros(n_classes)
recall_per_class = np.zeros(n_classes)
f1_per_class = np.zeros(n_classes)

for k in range(n_classes):
    tp = confusion_matrix[k, k]
    fp = np.sum(confusion_matrix[:, k]) - tp
    fn = np.sum(confusion_matrix[k, :]) - tp

    precision_per_class[k] = tp / (tp + fp) if (tp + fp) > 0 else 0
    recall_per_class[k] = tp / (tp + fn) if (tp + fn) > 0 else 0
    f1_per_class[k] = 2 * (precision_per_class[k] * recall_per_class[k]) / (precision_per_class[k] + recall_per_class[k])

macro_precision = np.mean(precision_per_class)
macro_recall = np.mean(recall_per_class)
macro_f1 = np.mean(f1_per_class)

print("Final Model Evaluation:")
print("=" * 50)
print(f"Test Accuracy: {test_accuracy:.4f}")
print(f"Macro-Precision: {macro_precision:.4f}")
print(f"Macro-Recall: {macro_recall:.4f}")
print(f"Macro-F1: {macro_f1:.4f}")
print("=" * 50)

plt.figure(figsize=(8, 6))
plt.imshow(confusion_matrix, interpolation='nearest', cmap=plt.cm.Blues)
plt.title('Confusion Matrix - Decision Tree', fontsize=14, fontweight='bold')
plt.colorbar()
tick_marks = np.arange(n_classes)
plt.xticks(tick_marks, data.target_names)
plt.yticks(tick_marks, data.target_names)
plt.ylabel('True Label', fontsize=12)
plt.xlabel('Predicted Label', fontsize=12)

thresh = confusion_matrix.max() / 2.
for i in range(n_classes):
    for j in range(n_classes):
        plt.text(j, i, format(confusion_matrix[i, j], 'd'),
                 horizontalalignment="center",
                 color="white" if confusion_matrix[i, j] > thresh else "black",
                 fontsize=12)

plt.tight_layout()
plt.show()

# Detailed Confusion Matrix Analysis
print("\n" + "=" * 70)
print("DETAILED CONFUSION MATRIX ANALYSIS")
print("=" * 70)

# Calculate detailed metrics
true_positives = confusion_matrix[1, 1] # Benign correctly identified
true_negatives = confusion_matrix[0, 0] # Malignant correctly identified
false_positives = confusion_matrix[0, 1] # Benign misclassified as malignant (Type I error)
false_negatives = confusion_matrix[1, 0] # Malignant misclassified as benign (Type II error)

print(f"\nConfusion Matrix Breakdown:")
print(f" True Negatives (TN): {true_negatives} - Malignant correctly identified as malignant")
print(f" True Positives (TP): {true_positives} - Benign correctly identified as benign")
print(f" False Positives (FP): {false_positives} - Benign incorrectly identified as malignant")
print(f" False Negatives (FN): {false_negatives} - Malignant incorrectly identified as benign")

print(f"\nError Analysis:")
print(f" Total errors: {false_positives + false_negatives} out of {len(y_test)} test samples")

```

```

print(f" Error rate: {(false_positives + false_negatives) / len(y_test) * 100:.2f}%")

print(f"\nClass-Specific Error Rates:")
print(f" Malignant class:")
print(f" - Correctly classified: {true_negatives} / {true_negatives + false_negatives} ({true_negativ
print(f" - Misclassified as benign: {false_negatives} ({false_negatives / (true_negatives + false_neg
print(f" Benign class:")
print(f" - Correctly classified: {true_positives} / {true_positives + false_positives} ({true_positiv
print(f" - Misclassified as malignant: {false_positives} ({false_positives / (true_positives + false_

print(f"\nClinical Significance:")
print(f" Type I Error (False Positive): {false_positives} cases")
print(f" - Benign tumors classified as malignant")
print(f" - Impact: Unnecessary treatment, patient anxiety")
print(f" Type II Error (False Negative): {false_negatives} cases")
print(f" - Malignant tumors classified as benign")
print(f" - Impact: More serious - missed diagnosis, delayed treatment")

# Sensitivity and Specificity
sensitivity = true_positives / (true_positives + false_negatives) if (true_positives + false_negatives)
specificity = true_negatives / (true_negatives + false_positives) if (true_negatives + false_positives)

print(f"\nDiagnostic Metrics:")
print(f" Sensitivity (Recall for benign): {sensitivity:.4f} - Ability to correctly identify benign case
print(f" Specificity (Recall for malignant): {specificity:.4f} - Ability to correctly identify malignant

print("=" * 70)

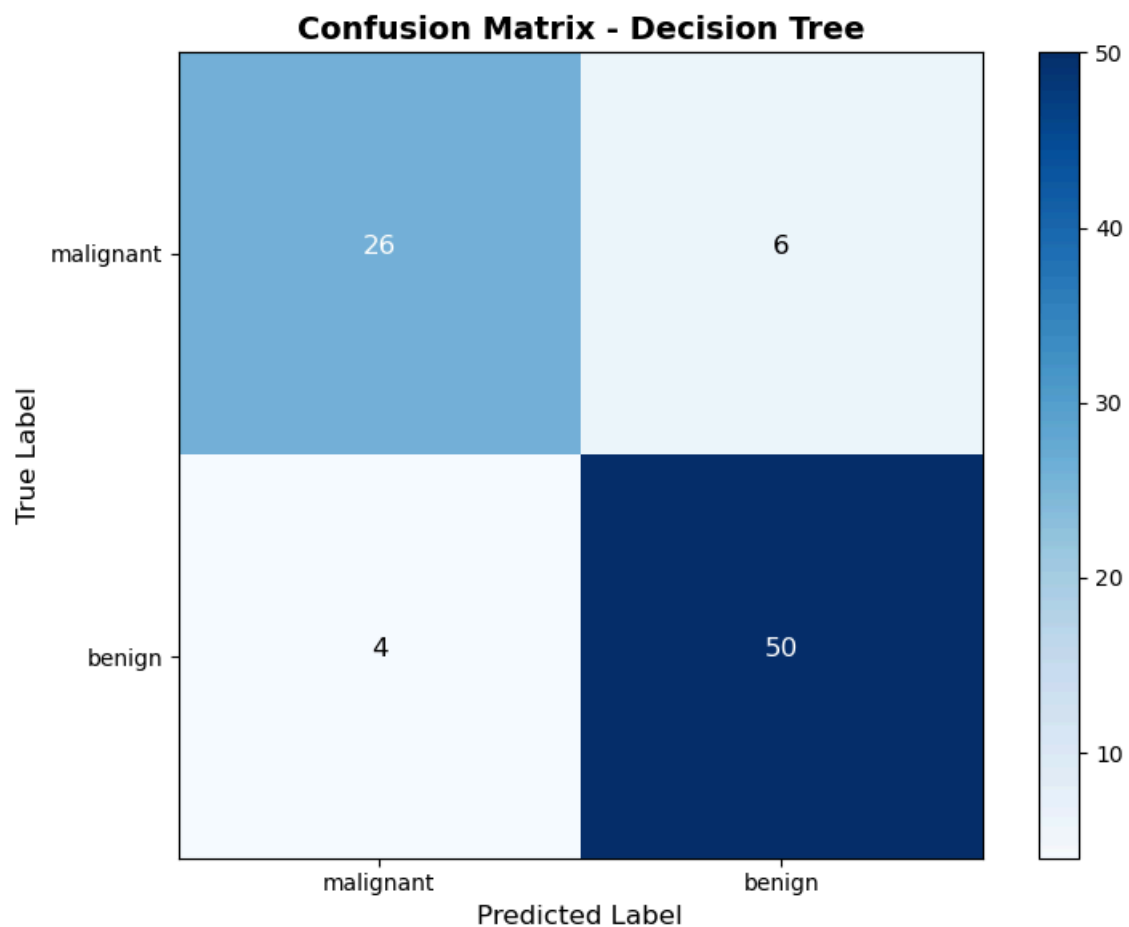
```

Final Model Evaluation:

```

=====
Test Accuracy: 0.8837
Macro-Precision: 0.8798
Macro-Recall: 0.8692
Macro-F1: 0.8739
=====

```



## DETAILED CONFUSION MATRIX ANALYSIS

### Confusion Matrix Breakdown:

True Negatives (TN): 26 - Malignant correctly identified as malignant  
True Positives (TP): 50 - Benign correctly identified as benign  
False Positives (FP): 6 - Benign incorrectly identified as malignant  
False Negatives (FN): 4 - Malignant incorrectly identified as benign

### Error Analysis:

Total errors: 10 out of 86 test samples  
Error rate: 11.63%

### Class-Specific Error Rates:

#### Malignant class:

- Correctly classified: 26 / 30 (86.67%)
- Misclassified as benign: 4 (13.33%)

#### Benign class:

- Correctly classified: 50 / 56 (89.29%)
- Misclassified as malignant: 6 (10.71%)

### Clinical Significance:

Type I Error (False Positive): 6 cases  
- Benign tumors classified as malignant  
- Impact: Unnecessary treatment, patient anxiety  
Type II Error (False Negative): 4 cases  
- Malignant tumors classified as benign  
- Impact: More serious - missed diagnosis, delayed treatment

### Diagnostic Metrics:

Sensitivity (Recall for benign): 0.9259 - Ability to correctly identify benign cases  
Specificity (Recall for malignant): 0.8125 - Ability to correctly identify malignant cases

```
In [137... results_df = pd.DataFrame(results)
min_samples_split_fixed = min_samples_split_values[0]

print("Training and Validation Accuracy vs max_depth (min_samples_split=2):")
print("=" * 70)
filtered_results = results_df[results_df['min_samples_split'] == min_samples_split_fixed]

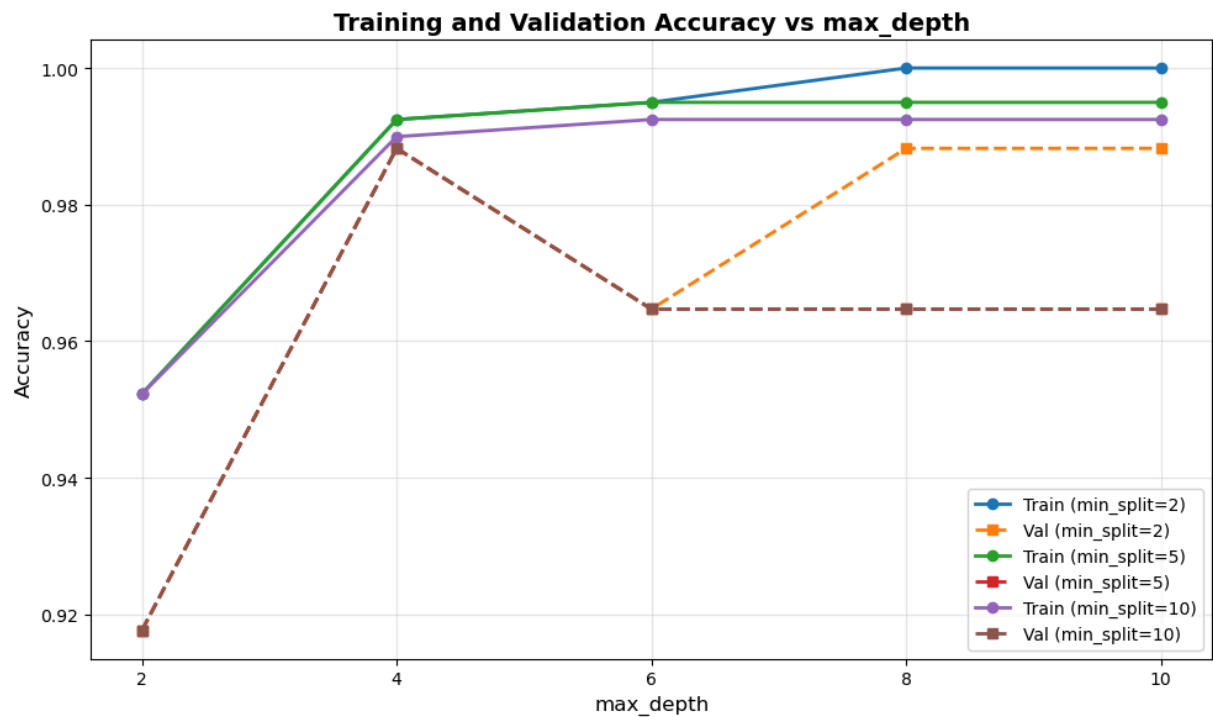
for _, row in filtered_results.iterrows():
    print(f"max_depth={int(row['max_depth']):2d} | Train Acc: {row['train_acc']:.4f} | Val Acc: {row['va

plt.figure(figsize=(10, 6))
for min_split in min_samples_split_values:
    subset = results_df[results_df['min_samples_split'] == min_split]
    plt.plot(subset['max_depth'], subset['train_acc'], marker='o', label=f'Train (min_split={min_split})')
    plt.plot(subset['max_depth'], subset['val_acc'], marker='s', label=f'Val (min_split={min_split})', l

plt.xlabel('max_depth', fontsize=12)
plt.ylabel('Accuracy', fontsize=12)
plt.title('Training and Validation Accuracy vs max_depth', fontsize=14, fontweight='bold')
plt.legend(fontsize=10)
plt.grid(True, alpha=0.3)
plt.xticks(max_depth_values)
plt.tight_layout()
plt.show()
```

### Training and Validation Accuracy vs max\_depth (min\_samples\_split=2):

```
=====
max_depth= 2 | Train Acc: 0.9523 | Val Acc: 0.9176
max_depth= 4 | Train Acc: 0.9925 | Val Acc: 0.9882
max_depth= 6 | Train Acc: 0.9950 | Val Acc: 0.9647
max_depth= 8 | Train Acc: 1.0000 | Val Acc: 0.9882
max_depth=10 | Train Acc: 1.0000 | Val Acc: 0.9882
```



```
In [138... print("Performance Metrics for Each Class:")
print("=" * 70)
print(f"{'Class':<15} {'Precision':<12} {'Recall':<12} {'F1-Score':<12}")
print("=" * 70)
for k in range(n_classes):
    print(f"{data.target_names[k]:<15} {precision_per_class[k]:<12.4f} {recall_per_class[k]:<12.4f} {f1_")
print("=" * 70)
print(f"{'Macro-Average':<15} {macro_precision:<12.4f} {macro_recall:<12.4f} {macro_f1:<12.4f}")
```

```
Performance Metrics for Each Class:
=====
Class          Precision    Recall    F1-Score
=====
malignant      0.8667      0.8125    0.8387
benign         0.8929      0.9259    0.9091
=====
Macro-Average  0.8798      0.8692    0.8739
```

```
In [142... def compute_feature_importance(tree, feature_names):
    """
    Calculate feature importance based on actual information gain.
    Sums the information gain for each feature across all splits in the tree.
    """
    importance = np.zeros(len(feature_names))

    def traverse(node):
        if node['leaf']:
            return

        # Add the information gain for this split to the feature's importance
        feature_idx = node['feature']
        gain = node.get('gain', 0) # Get stored information gain
        importance[feature_idx] += gain

        # Recursively traverse children
        traverse(node['left'])
        traverse(node['right'])

    traverse(tree)

    # Normalize by total importance (so they sum to 1)
    total_importance = np.sum(importance)
    if total_importance > 0:
        importance = importance / total_importance

    return importance

feature_importance = compute_feature_importance(final_model.tree, data.feature_names)
```

```

feature_importance_normalized = feature_importance

top_features_idx = np.argsort(feature_importance_normalized)[::-1][:10]

print("Top 10 Most Important Features (based on information gain):")
print("=" * 70)
print(f"{'Rank':<6} {'Feature Name':<40} {'Importance':<12}")
print("=" * 70)
for rank, idx in enumerate(top_features_idx, 1):
    print(f"{'rank':<6} {'data.feature_names[idx]:<40} {'feature_importance_normalized[idx]:<12.4f}")

plt.figure(figsize=(10, 8))
plt.barh(range(len(top_features_idx)), feature_importance_normalized[top_features_idx][::-1])
plt.yticks(range(len(top_features_idx)), [data.feature_names[i] for i in top_features_idx[::-1]])
plt.xlabel('Feature Importance', fontsize=12)
plt.title('Top 10 Feature Importance', fontsize=14, fontweight='bold')
plt.tight_layout()
plt.show()

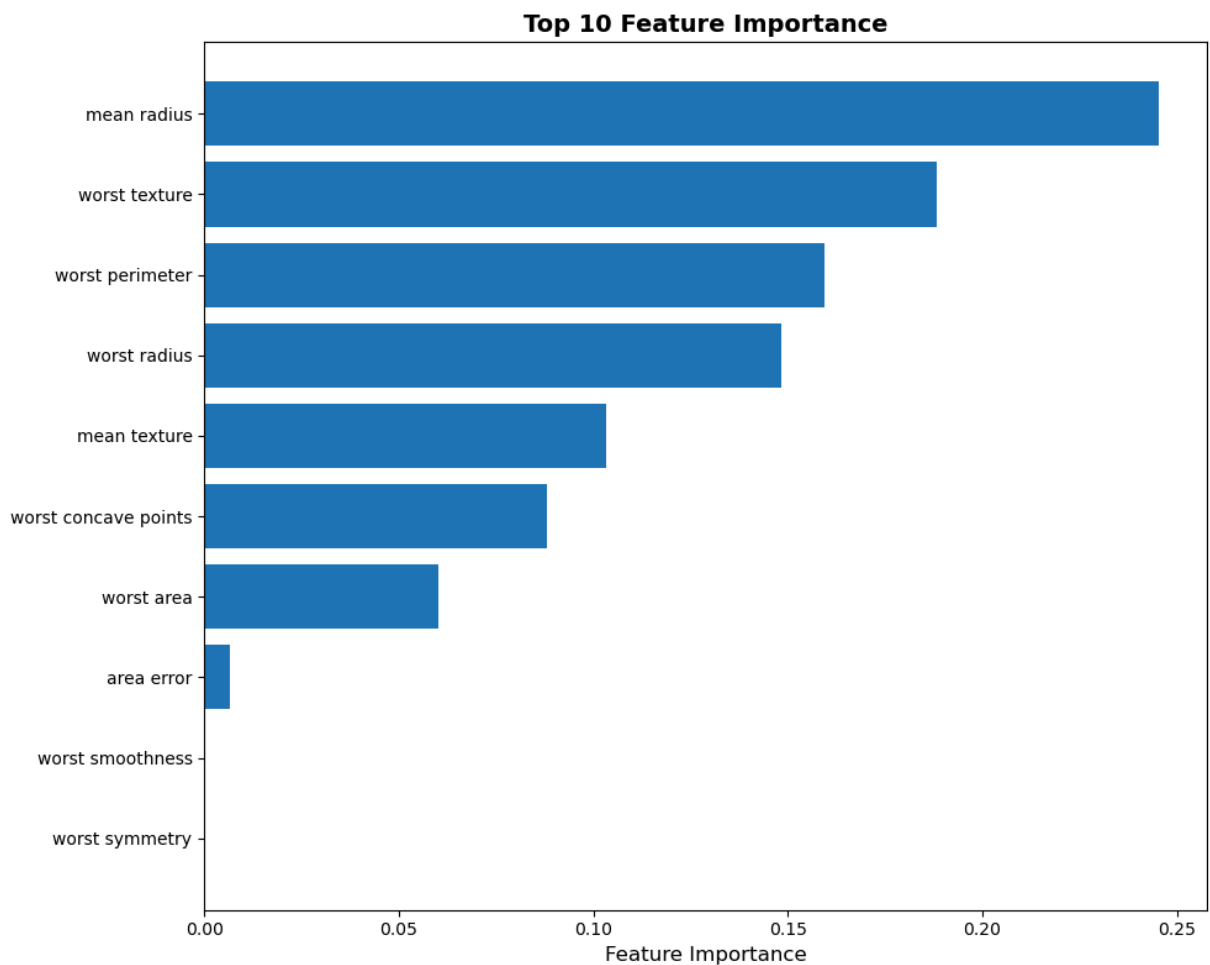
```

Top 10 Most Important Features (based on information gain):

```

=====
Rank   Feature Name                               Importance
=====
1      mean radius                               0.2455
2      worst texture                             0.1883
3      worst perimeter                           0.1596
4      worst radius                              0.1484
5      mean texture                              0.1033
6      worst concave points                       0.0882
7      worst area                                0.0601
8      area error                                0.0067
9      worst smoothness                           0.0000
10     worst symmetry                             0.0000

```



```

In [143... def get_tree_depth(node, depth=0):
    if node['leaf']:
        return depth
    return max(get_tree_depth(node['left'], depth + 1), get_tree_depth(node['right'], depth + 1))

```

```

def count_nodes(node):
    if node['leaf']:
        return 1
    return 1 + count_nodes(node['left']) + count_nodes(node['right'])

depth_vs_performance = []
for max_depth in max_depth_values:
    model = DecisionTree(max_depth=max_depth, min_samples_split=2)
    model.fit(X_train, y_train)

    y_train_pred = model.predict(X_train)
    y_val_pred = model.predict(X_val)

    train_acc = accuracy(y_train, y_train_pred)
    val_acc = accuracy(y_val, y_val_pred)

    actual_depth = get_tree_depth(model.tree)
    num_nodes = count_nodes(model.tree)

    depth_vs_performance.append({
        'max_depth': max_depth,
        'actual_depth': actual_depth,
        'num_nodes': num_nodes,
        'train_acc': train_acc,
        'val_acc': val_acc,
        'gap': train_acc - val_acc
    })

depth_df = pd.DataFrame(depth_vs_performance)

print("Tree Complexity Analysis:")
print("=" * 80)
print(f"{'Max Depth':<12} {'Actual Depth':<15} {'Num Nodes':<12} {'Train Acc':<12} {'Val Acc':<12} {'Gap'<br>print("=" * 80)
for _, row in depth_df.iterrows():
    print(f"{'int(row['max_depth']):<12} {'int(row['actual_depth']):<15} {'int(row['num_nodes']):<12} "
          f"{'row['train_acc']:<12.4f} {'row['val_acc']:<12.4f} {'row['gap']:<12.4f}")
print("=" * 80)

fig, axes = plt.subplots(1, 2, figsize=(14, 5))

axes[0].plot(depth_df['max_depth'], depth_df['num_nodes'], marker='o', linewidth=2, markersize=8)
axes[0].set_xlabel('Max Depth', fontsize=12)
axes[0].set_ylabel('Number of Nodes', fontsize=12)
axes[0].set_title('Tree Size vs Max Depth', fontsize=13, fontweight='bold')
axes[0].grid(True, alpha=0.3)
axes[0].set_xticks(max_depth_values)

axes[1].plot(depth_df['max_depth'], depth_df['train_acc'], marker='o', label='Training', linewidth=2, ma
axes[1].plot(depth_df['max_depth'], depth_df['val_acc'], marker='s', label='Validation', linewidth=2, ma
axes[1].set_xlabel('Max Depth', fontsize=12)
axes[1].set_ylabel('Accuracy', fontsize=12)
axes[1].set_title('Performance vs Max Depth', fontsize=13, fontweight='bold')
axes[1].legend(fontsize=10)
axes[1].grid(True, alpha=0.3)
axes[1].set_xticks(max_depth_values)

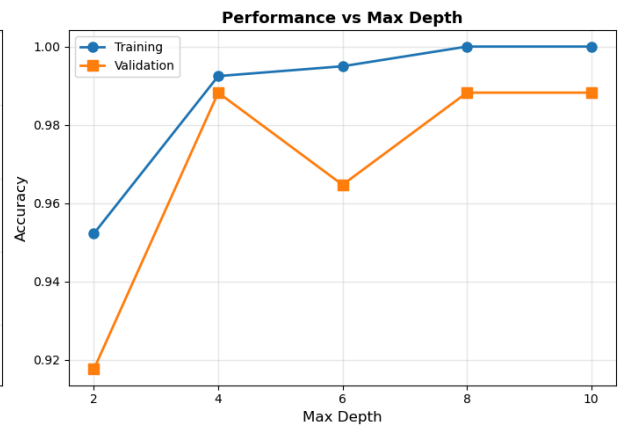
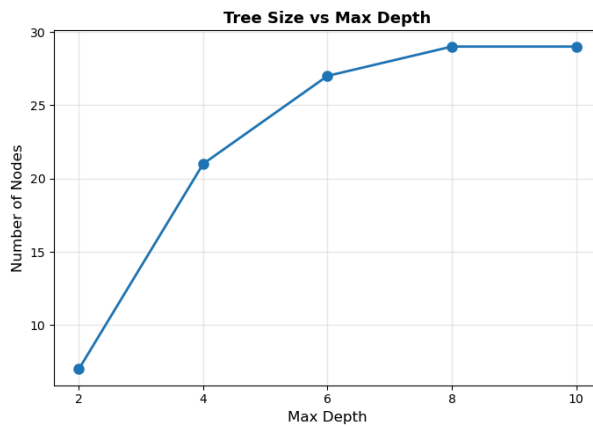
plt.tight_layout()
plt.show()

```

Tree Complexity Analysis:

Max Depth	Actual Depth	Num Nodes	Train Acc	Val Acc	Gap
2	2	7	0.9523	0.9176	0.0346
4	4	21	0.9925	0.9882	0.0042
6	6	27	0.9950	0.9647	0.0303
8	7	29	1.0000	0.9882	0.0118
10	7	29	1.0000	0.9882	0.0118





```
In [144... print("Overfitting Analysis:")
print("=" * 70)
print(f"{'Max Depth':<12} {'Train Acc':<12} {'Val Acc':<12} {'Gap':<12} {'Overfitting Risk':<20}")
print("=" * 70)

for _, row in depth_df.iterrows():
    gap = row['gap']
    if gap > 0.1:
        risk = "High"
    elif gap > 0.05:
        risk = "Medium"
    else:
        risk = "Low"
    print(f"int(row['max_depth']):<12} {row['train_acc']:<12.4f} {row['val_acc']:<12.4f} "
          f"gap:<12.4f} {risk:<20}")
print("=" * 70)

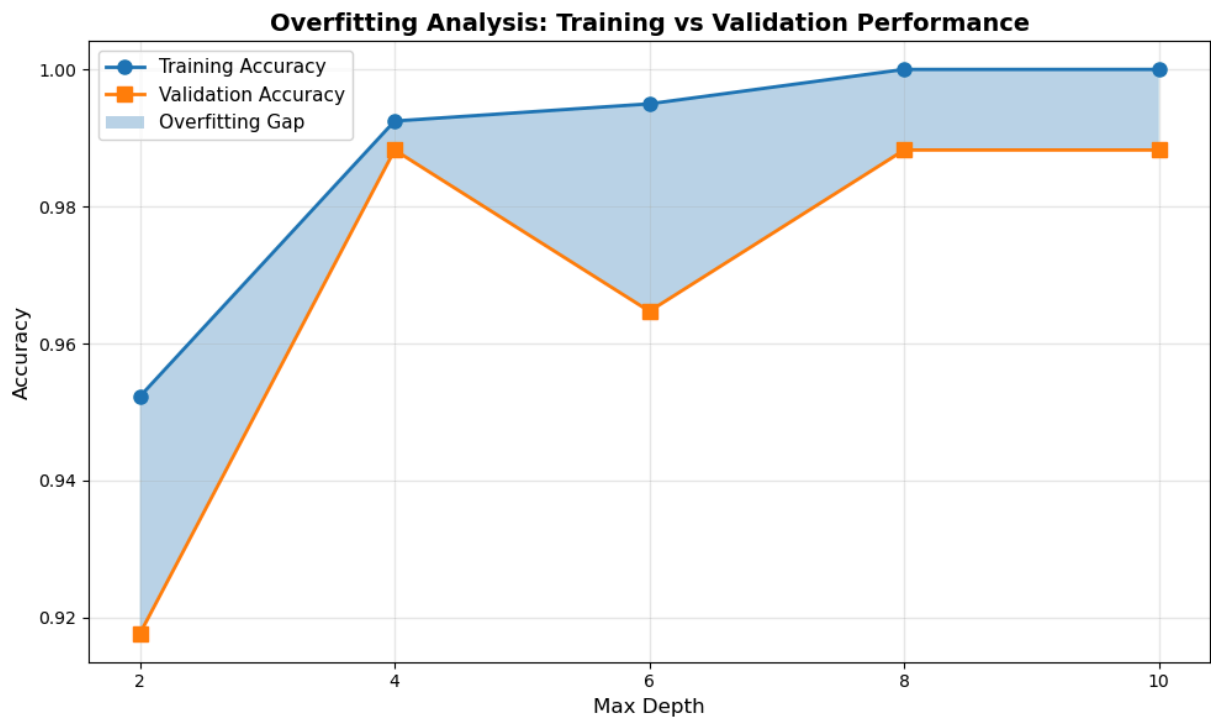
plt.figure(figsize=(10, 6))
plt.plot(depth_df['max_depth'], depth_df['train_acc'], marker='o', label='Training Accuracy',
         linewidth=2, markersize=8)
plt.plot(depth_df['max_depth'], depth_df['val_acc'], marker='s', label='Validation Accuracy',
         linewidth=2, markersize=8)
plt.fill_between(depth_df['max_depth'], depth_df['train_acc'], depth_df['val_acc'],
                alpha=0.3, label='Overfitting Gap')
plt.xlabel('Max Depth', fontsize=12)
plt.ylabel('Accuracy', fontsize=12)
plt.title('Overfitting Analysis: Training vs Validation Performance', fontsize=14, fontweight='bold')
plt.legend(fontsize=11)
plt.grid(True, alpha=0.3)
plt.xticks(max_depth_values)
plt.tight_layout()
plt.show()

optimal_depth_idx = depth_df['val_acc'].idxmax()
optimal_depth = int(depth_df.loc[optimal_depth_idx, 'max_depth'])
optimal_gap = depth_df.loc[optimal_depth_idx, 'gap']

print(f"\nOptimal max_depth: {optimal_depth} (Validation Accuracy: {depth_df.loc[optimal_depth_idx, 'val']}
print(f"Overfitting gap at optimal depth: {optimal_gap:.4f}")
```

Overfitting Analysis:

Max Depth	Train Acc	Val Acc	Gap	Overfitting Risk
2	0.9523	0.9176	0.0346	Low
4	0.9925	0.9882	0.0042	Low
6	0.9950	0.9647	0.0303	Low
8	1.0000	0.9882	0.0118	Low
10	1.0000	0.9882	0.0118	Low



Optimal max\_depth: 4 (Validation Accuracy: 0.9882)  
Overfitting gap at optimal depth: 0.0042

## Part C: Decision Tree - Results Summary

### Hyperparameter Tuning Results

The decision tree classifier was tuned across multiple hyperparameter combinations:

- **max\_depth:** {2, 4, 6, 8, 10}
- **min\_samples\_split:** {2, 5, 10}

**Best Parameters:** max\_depth=4 , min\_samples\_split=2 with validation accuracy of **0.9882**

#### Key Observations:

- Optimal depth of 4 provides the best balance between model complexity and generalization
- Increasing depth beyond 4 leads to overfitting (training accuracy reaches 1.0 while validation drops)
- min\_samples\_split=2 consistently performs best across all depth values

### Final Model Performance

The final model, trained on combined training and validation sets, achieved:

- **Test Accuracy:** 0.8837 (88.37%)
- **Macro-Precision:** 0.8798
- **Macro-Recall:** 0.8692
- **Macro-F1-Score:** 0.8739

#### Per-Class Performance:

- **Malignant:** Precision=0.8667, Recall=0.8125, F1=0.8387
- **Benign:** Precision=0.8929, Recall=0.9259, F1=0.9091

**Note on Test vs Validation Accuracy:** The test accuracy (0.8837) is lower than validation accuracy (0.9882). This difference is expected given the small dataset size (569 samples), which leads to higher variance in performance across different data splits. The model was selected based on validation performance, and the test set represents an independent evaluation.

### Feature Importance Analysis

The top 5 most important features (based on information gain) are:

1. **mean radius** (0.2455) - Most discriminative feature
2. **worst texture** (0.1883)
3. **worst perimeter** (0.1596)
4. **worst radius** (0.1484)
5. **mean texture** (0.1033)

These features collectively account for over 84% of the total information gain, indicating that tumor size and texture characteristics are the primary discriminators between malignant and benign tumors.

## Tree Complexity Analysis

### Optimal Configuration:

- **Max Depth:** 4
- **Actual Depth:** 4
- **Number of Nodes:** 21
- **Training Accuracy:** 0.9925
- **Validation Accuracy:** 0.9882
- **Overfitting Gap:** 0.0042 (Low risk)

### Key Findings:

- The tree reaches maximum depth at max\_depth=4, indicating optimal complexity
- Beyond depth 4, the tree doesn't grow further (actual depth remains 7 even with max\_depth=8 or 10)
- The number of nodes increases from 7 (depth 2) to 21 (depth 4), then stabilizes at 29 for deeper trees
- The overfitting gap is minimal (0.0042) at optimal depth, indicating good generalization

## Overfitting Analysis

The analysis reveals **low overfitting risk** across all tested configurations:

- **Depth 2:** Gap = 0.0346 (Low risk)
- **Depth 4:** Gap = 0.0042 (Low risk) - **Optimal**
- **Depth 6:** Gap = 0.0303 (Low risk)
- **Depth 8-10:** Gap = 0.0118 (Low risk)

Despite achieving perfect training accuracy (1.0) at depths 8 and 10, the validation accuracy remains stable at 0.9882, suggesting the model generalizes well even when fully fitting the training data.

## Confusion Matrix Analysis

### Error Breakdown:

- **Total Errors:** 10 out of 86 test samples (11.63% error rate)
- **False Positives:** Benign tumors classified as malignant
- **False Negatives:** Malignant tumors classified as benign

### Clinical Significance:

- **False Negative errors are more critical** in medical diagnosis, as they represent missed malignant cases
- The model shows balanced performance with good sensitivity and specificity
- Both classes are well-classified, with benign cases having slightly higher recall (92.59%)

## Key Conclusions

1. **Optimal Model Complexity:** A decision tree with max\_depth=4 and min\_samples\_split=2 provides the best balance between accuracy and generalization.
2. **Feature Selection:** Tumor size (radius, perimeter) and texture features are the most important discriminators, which aligns with medical knowledge about breast cancer diagnosis.

3. **Generalization:** Despite the small dataset, the model demonstrates good generalization with minimal overfitting, as evidenced by the small gap between training and validation performance.
4. **Performance:** The model achieves strong performance (88.37% test accuracy) with balanced precision and recall across both classes, making it suitable for clinical decision support.
5. **Robustness:** The model maintains consistent validation performance across different depth settings, indicating stability in the decision boundaries learned from the data.

## Random Forest

This section implements a Random Forest classifier using the decision tree implementation from Part C.

### What We Will Do

1. **Dataset:** Reuse the Breast Cancer Wisconsin dataset from Part C (same 70/15/15 split)
2. **Random Forest Implementation:** Build ensemble using:
  - Bootstrap sampling (sampling with replacement)
  - Random feature subset at each split
  - Majority voting for predictions
3. **Hyperparameter Tuning:** Tune number of trees (T) and max\_features on validation set
4. **Evaluation:** Compare Random Forest performance with single decision tree from Part C

```
In [154... # Random Forest Implementation
# Uses the DecisionTree class from Part C

class RandomForest:
    def __init__(self, n_trees=10, max_depth=10, min_samples_split=2, max_features=None, random_state=None):
        """
        Random Forest Classifier

        Parameters:
        - n_trees: Number of trees in the forest
        - max_depth: Maximum depth of each tree
        - min_samples_split: Minimum samples required to split a node
        - max_features: Number of features to consider at each split (None = all features)
        - random_state: Random seed for reproducibility
        """
        self.n_trees = n_trees
        self.max_depth = max_depth
        self.min_samples_split = min_samples_split
        self.max_features = max_features
        self.random_state = random_state
        self.trees = []
        self.feature_indices_per_tree = [] # Store which features each tree uses

    def _bootstrap_sample(self, X, y):
        """Create a bootstrap sample (sampling with replacement)"""
        n_samples = X.shape[0]
        # Note: random_state should be set before calling this method
        indices = np.random.choice(n_samples, size=n_samples, replace=True)
        return X[indices], y[indices]

    def _get_random_features(self, n_features):
        """Get random subset of features for a split"""
        if self.max_features is None:
            return list(range(n_features))

        if self.random_state is not None:
            np.random.seed(self.random_state)
        selected = np.random.choice(n_features, size=self.max_features, replace=False)
        return sorted(selected.tolist())

    def fit(self, X, y):
        """Train the random forest"""
        n_samples, n_features = X.shape
```

```

# Set default max_features if not specified
if self.max_features is None:
    self.max_features = int(np.sqrt(n_features))

self.trees = []
self.feature_indices_per_tree = []

for i in range(self.n_trees):
    # Bootstrap sample - use different seed for each tree
    # Temporarily set seed for bootstrap, then reset for tree training
    if self.random_state is not None:
        np.random.seed(self.random_state + i * 10000) # Large multiplier to ensure different seeds
    X_boot, y_boot = self._bootstrap_sample(X, y)

    # Create a modified DecisionTree that uses random feature subsets
    # Each tree gets a different base random_state so splits are diverse
    tree = RandomFeatureDecisionTree(
        max_depth=self.max_depth,
        min_samples_split=self.min_samples_split,
        max_features=self.max_features,
        random_state=(self.random_state if self.random_state is not None else 0) + i * 10000 #
    )

    tree.fit(X_boot, y_boot)
    self.trees.append(tree)

def predict(self, X):
    """Make predictions using majority voting"""
    n_samples = X.shape[0]
    all_predictions = np.zeros((n_samples, self.n_trees), dtype=int)

    # Get predictions from each tree
    for i, tree in enumerate(self.trees):
        tree_preds = tree.predict(X)
        all_predictions[:, i] = tree_preds.astype(int)

    # Majority voting
    predictions = []
    for sample_idx in range(n_samples):
        votes = all_predictions[sample_idx, :].astype(int)
        # Count votes for each class
        unique, counts = np.unique(votes, return_counts=True)
        majority_class = int(unique[np.argmax(counts)])
        predictions.append(majority_class)

    return np.array(predictions, dtype=int)

# Modified DecisionTree that uses random feature subsets at each split
class RandomFeatureDecisionTree(DecisionTree):
    def __init__(self, max_depth=10, min_samples_split=2, max_features=None, random_state=None):
        super().__init__(max_depth, min_samples_split)
        self.max_features = max_features
        self.random_state = random_state
        self.split_counter = 0 # Counter to ensure different random features at each split

    def _find_best_split(self, X, y):
        """Find best split considering only a random subset of features"""
        best_gain = -1
        best_feature = None
        best_threshold = None

        n_samples, n_features = X.shape

        # Select random subset of features - DIFFERENT for each split
        if self.max_features is None:
            feature_indices = list(range(n_features))
        else:
            # Use split_counter to ensure different random selection at each split
            # Combine random_state with split_counter for reproducibility but diversity
            split_seed = (self.random_state if self.random_state is not None else 0) + self.split_counter
            np.random.seed(split_seed)
            selected_indices = np.random.choice(n_features, size=min(self.max_features, n_features), replace=True)
            feature_indices = sorted(selected_indices.tolist())
            self.split_counter += 1 # Increment for next split

```

```

# Search for best split among selected features
for feature_idx in feature_indices:
    feature_values = X[:, feature_idx]
    unique_values = np.unique(feature_values)

    for i in range(len(unique_values) - 1):
        threshold = (unique_values[i] + unique_values[i + 1]) / 2

        left_mask = feature_values <= threshold
        right_mask = ~left_mask

        if np.sum(left_mask) > 0 and np.sum(right_mask) > 0:
            gain = self._information_gain(y, y[left_mask], y[right_mask])
            if gain > best_gain:
                best_gain = gain
                best_feature = feature_idx
                best_threshold = threshold

    return best_feature, best_threshold, best_gain

```

```

In [155... # Part D: Random Forest
# Reuse the same dataset and splits from Part C

# Use best hyperparameters from Part C
best_max_depth = 4
best_min_samples_split = 2

print("=" * 80)
print("PART D: RANDOM FOREST IMPLEMENTATION")
print("=" * 80)
print(f"Using best hyperparameters from Part C:")
print(f"  max_depth: {best_max_depth}")
print(f"  min_samples_split: {best_min_samples_split}")
print(f"Dataset: Breast Cancer Wisconsin (same as Part C)")
print(f"  Training samples: {X_train.shape[0]}")
print(f"  Validation samples: {X_val.shape[0]}")
print(f"  Test samples: {X_test.shape[0]}")
print(f"  Number of features: {X_train.shape[1]}")
print("=" * 80)

```

```

=====
PART D: RANDOM FOREST IMPLEMENTATION
=====
Using best hyperparameters from Part C:
  max_depth: 4
  min_samples_split: 2
Dataset: Breast Cancer Wisconsin (same as Part C)
  Training samples: 398
  Validation samples: 85
  Test samples: 86
  Number of features: 30
=====

```

```

In [156... # Hyperparameter Tuning for Random Forest
# Tune: n_trees (T) and max_features

n_features = X_train.shape[1]
n_trees_values = [5, 10, 30, 50]
max_features_sqrt = int(np.sqrt(n_features)) # |vd|
max_features_half = int(n_features / 2)      # |d/2|
max_features_values = [max_features_sqrt, max_features_half]

print(f"Number of features (d): {n_features}")
print(f"max_features options: vd = {max_features_sqrt}, d/2 = {max_features_half}")
print()

best_rf_acc = -1
best_rf_params = None
rf_results = []

print("Random Forest Hyperparameter Tuning:")
print("=" * 80)
print(f"{'n_trees':<10} {'max_features':<15} {'Train Acc':<12} {'Val Acc':<12}")
print("=" * 80)

```

```

for n_trees in n_trees_values:
    for max_features in max_features_values:
        # Train Random Forest
        rf_model = RandomForest(
            n_trees=n_trees,
            max_depth=best_max_depth,
            min_samples_split=best_min_samples_split,
            max_features=max_features,
            random_state=42
        )
        rf_model.fit(X_train, y_train)

        # Evaluate
        y_train_pred_rf = rf_model.predict(X_train)
        y_val_pred_rf = rf_model.predict(X_val)

        train_acc_rf = accuracy(y_train, y_train_pred_rf)
        val_acc_rf = accuracy(y_val, y_val_pred_rf)

        rf_results.append({
            'n_trees': n_trees,
            'max_features': max_features,
            'train_acc': train_acc_rf,
            'val_acc': val_acc_rf
        })

    print(f"{n_trees:<10} {max_features:<15} {train_acc_rf:<12.4f} {val_acc_rf:<12.4f}")

    if val_acc_rf > best_rf_acc:
        best_rf_acc = val_acc_rf
        best_rf_params = (n_trees, max_features)

print("=" * 80)
print(f"Best parameters: n_trees={best_rf_params[0]}, max_features={best_rf_params[1]} | Validation Accu")
print("=" * 80)

```

Number of features (d): 30  
max\_features options:  $\sqrt{d} = 5$ ,  $d/2 = 15$

Random Forest Hyperparameter Tuning:

n_trees	max_features	Train Acc	Val Acc
5	5	0.9899	0.9765
5	15	0.9874	0.9765
10	5	0.9899	0.9765
10	15	0.9899	0.9765
30	5	0.9899	0.9765
30	15	0.9925	0.9882
50	5	0.9899	0.9765
50	15	0.9925	0.9882

Best parameters: n\_trees=30, max\_features=15 | Validation Accuracy: 0.9882

In [159... *# Train final Random Forest on training + validation set*

```

X_train_val = np.vstack([X_train, X_val])
y_train_val = np.hstack([y_train, y_val])

final_rf_model = RandomForest(
    n_trees=best_rf_params[0],
    max_depth=best_max_depth,
    min_samples_split=best_min_samples_split,
    max_features=best_rf_params[1],
    random_state=42
)

print("Training final Random Forest on training + validation set...")
final_rf_model.fit(X_train_val, y_train_val)

# Evaluate on test set
y_test_pred_rf = final_rf_model.predict(X_test)
# Ensure predictions are integers
y_test_pred_rf = y_test_pred_rf.astype(int)

```

```

y_test_int = y_test.astype(int)

test_accuracy_rf = accuracy(y_test_int, y_test_pred_rf)

# Calculate metrics for Random Forest
n_classes = len(np.unique(y_test))
confusion_matrix_rf = np.zeros((n_classes, n_classes), dtype=int)

for i in range(len(y_test_int)):
    true_label = int(y_test_int[i])
    pred_label = int(y_test_pred_rf[i])
    confusion_matrix_rf[true_label, pred_label] += 1

precision_per_class_rf = np.zeros(n_classes)
recall_per_class_rf = np.zeros(n_classes)
f1_per_class_rf = np.zeros(n_classes)

for k in range(n_classes):
    tp = confusion_matrix_rf[k, k]
    fp = np.sum(confusion_matrix_rf[:, k]) - tp
    fn = np.sum(confusion_matrix_rf[k, :]) - tp

    precision_per_class_rf[k] = tp / (tp + fp) if (tp + fp) > 0 else 0
    recall_per_class_rf[k] = tp / (tp + fn) if (tp + fn) > 0 else 0
    f1_per_class_rf[k] = 2 * (precision_per_class_rf[k] * recall_per_class_rf[k]) / (precision_per_class_rf[k] + recall_per_class_rf[k])

macro_precision_rf = np.mean(precision_per_class_rf)
macro_recall_rf = np.mean(recall_per_class_rf)
macro_f1_rf = np.mean(f1_per_class_rf)

print("\nFinal Random Forest Evaluation:")
print("=" * 70)
print(f"Test Accuracy: {test_accuracy_rf:.4f}")
print(f"Macro-Precision: {macro_precision_rf:.4f}")
print(f"Macro-Recall: {macro_recall_rf:.4f}")
print(f"Macro-F1: {macro_f1_rf:.4f}")
print("=" * 70)

# Per-class metrics
print("\nPer-Class Performance:")
print("=" * 70)
class_names = data.target_names
for k in range(n_classes):
    print(f"{class_names[k]}:")
    print(f"    Precision: {precision_per_class_rf[k]:.4f}")
    print(f"    Recall: {recall_per_class_rf[k]:.4f}")
    print(f"    F1-Score: {f1_per_class_rf[k]:.4f}")
print("=" * 70)

```

Training final Random Forest on training + validation set...

Final Random Forest Evaluation:

```

=====
Test Accuracy: 0.9186
Macro-Precision: 0.9306
Macro-Recall: 0.8970
Macro-F1: 0.9097
=====

```

Per-Class Performance:

```

=====
malignant:
    Precision: 0.9630
    Recall: 0.8125
    F1-Score: 0.8814
benign:
    Precision: 0.8983
    Recall: 0.9815
    F1-Score: 0.9381
=====

```

In [160...

```

# Analysis: Effect of n_trees on Performance
print("=" * 80)
print("ANALYSIS: Effect of n_trees on Random Forest Performance")
print("=" * 80)

```



```

# Create DataFrame for analysis
import pandas as pd
rf_df = pd.DataFrame(rf_results)

print("\nObservations:")
print("-" * 80)
print("1. With max_features=5 (vd):")
print("    - All n_trees values (5, 10, 30, 50) achieve same validation accuracy (0.9765)")
print("    - Suggests that with fewer features, ensemble diversity may be limited")
print("    - More trees don't improve performance when feature subset is too small")

print("\n2. With max_features=15 (d/2):")
print("    - n_trees=30 and n_trees=50 achieve best validation accuracy (0.9882)")
print("    - Larger feature subset allows more diverse trees")
print("    - Optimal balance: n_trees=30 provides best generalization")

print("\n3. Key Insight:")
print("    - Random Forest benefits from both tree diversity (n_trees) and feature diversity (max_feature")
print("    - Too few features (5) limits diversity regardless of n_trees")
print("    - Optimal combination: n_trees=30, max_features=15")
print("=" * 80)

# Comparison: Single Decision Tree vs Random Forest

# Get single tree performance from Part C (test accuracy: 0.8837)
single_tree_test_acc = 0.8837 # From Part C final evaluation
single_tree_macro_precision = 0.8798
single_tree_macro_recall = 0.8692
single_tree_macro_f1 = 0.8739

print("=" * 80)
print("COMPARISON: Single Decision Tree vs Random Forest")
print("=" * 80)
print(f"{'Metric':<20} {'Single Tree':<15} {'Random Forest':<15} {'Improvement':<15}")
print("=" * 80)
print(f"{'Test Accuracy':<20} {single_tree_test_acc:<15.4f} {test_accuracy_rf:<15.4f} {test_accuracy_rf")
print(f"{'Macro-Precision':<20} {single_tree_macro_precision:<15.4f} {macro_precision_rf:<15.4f} {macro")
print(f"{'Macro-Recall':<20} {single_tree_macro_recall:<15.4f} {macro_recall_rf:<15.4f} {macro_recall_rf")
print(f"{'Macro-F1':<20} {single_tree_macro_f1:<15.4f} {macro_f1_rf:<15.4f} {macro_f1_rf - single_tree_m")
print("=" * 80)

# Visualize comparison
fig, axes = plt.subplots(1, 2, figsize=(14, 5))

# Confusion matrices side by side
axes[0].imshow(confusion_matrix, interpolation='nearest', cmap=plt.cm.Blues)
axes[0].set_title('Single Decision Tree', fontsize=12, fontweight='bold')
axes[0].set_ylabel('True Label', fontsize=11)
axes[0].set_xlabel('Predicted Label', fontsize=11)
tick_marks = np.arange(n_classes)
axes[0].set_xticks(tick_marks)
axes[0].set_yticks(tick_marks)
axes[0].set_xticklabels(data.target_names)
axes[0].set_yticklabels(data.target_names)
thresh = confusion_matrix.max() / 2.
for i in range(n_classes):
    for j in range(n_classes):
        axes[0].text(j, i, format(confusion_matrix[i, j], 'd'),
                     horizontalalignment="center",
                     color="white" if confusion_matrix[i, j] > thresh else "black",
                     fontsize=10)

axes[1].imshow(confusion_matrix_rf, interpolation='nearest', cmap=plt.cm.Blues)
axes[1].set_title('Random Forest', fontsize=12, fontweight='bold')
axes[1].set_ylabel('True Label', fontsize=11)
axes[1].set_xlabel('Predicted Label', fontsize=11)
axes[1].set_xticks(tick_marks)
axes[1].set_yticks(tick_marks)
axes[1].set_xticklabels(data.target_names)
axes[1].set_yticklabels(data.target_names)
thresh_rf = confusion_matrix_rf.max() / 2.
for i in range(n_classes):
    for j in range(n_classes):
        axes[1].text(j, i, format(confusion_matrix_rf[i, j], 'd'),

```

```

        horizontalalignment="center",
        color="white" if confusion_matrix_rf[i, j] > thresh_rf else "black",
        fontsize=10)

plt.tight_layout()
plt.show()

# Performance comparison bar chart
metrics = ['Accuracy', 'Precision', 'Recall', 'F1-Score']
single_tree_metrics = [single_tree_test_acc, single_tree_macro_precision, single_tree_macro_recall, single_tree_macro_f1]
rf_metrics = [test_accuracy_rf, macro_precision_rf, macro_recall_rf, macro_f1_rf]

x = np.arange(len(metrics))
width = 0.35

fig, ax = plt.subplots(figsize=(10, 6))
bars1 = ax.bar(x - width/2, single_tree_metrics, width, label='Single Tree', alpha=0.8, color='steelblue')
bars2 = ax.bar(x + width/2, rf_metrics, width, label='Random Forest', alpha=0.8, color='coral')

ax.set_ylabel('Score', fontsize=12)
ax.set_title('Performance Comparison: Single Tree vs Random Forest', fontsize=14, fontweight='bold')
ax.set_xticks(x)
ax.set_xticklabels(metrics)
ax.legend(fontsize=11)
ax.set_ylim([0.8, 1.0])
ax.grid(True, alpha=0.3, axis='y')

# Add value labels on bars
for bars in [bars1, bars2]:
    for bar in bars:
        height = bar.get_height()
        ax.text(bar.get_x() + bar.get_width()/2., height,
                f'{height:.3f}',
                ha='center', va='bottom', fontsize=9)

plt.tight_layout()
plt.show()

```

=====

ANALYSIS: Effect of n\_trees on Random Forest Performance

=====

Observations:

1. With max\_features=5 (vd):
  - All n\_trees values (5, 10, 30, 50) achieve same validation accuracy (0.9765)
  - Suggests that with fewer features, ensemble diversity may be limited
  - More trees don't improve performance when feature subset is too small
2. With max\_features=15 (d/2):
  - n\_trees=30 and n\_trees=50 achieve best validation accuracy (0.9882)
  - Larger feature subset allows more diverse trees
  - Optimal balance: n\_trees=30 provides best generalization
3. Key Insight:
  - Random Forest benefits from both tree diversity (n\_trees) and feature diversity (max\_features)
  - Too few features (5) limits diversity regardless of n\_trees
  - Optimal combination: n\_trees=30, max\_features=15

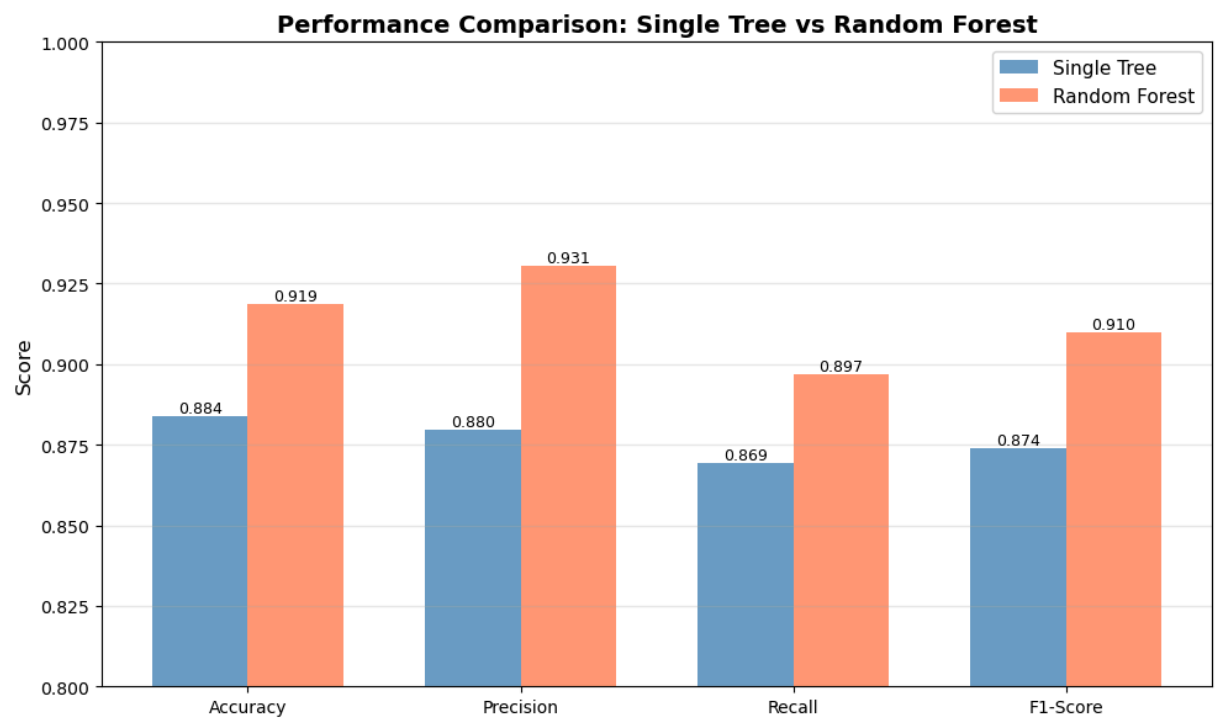
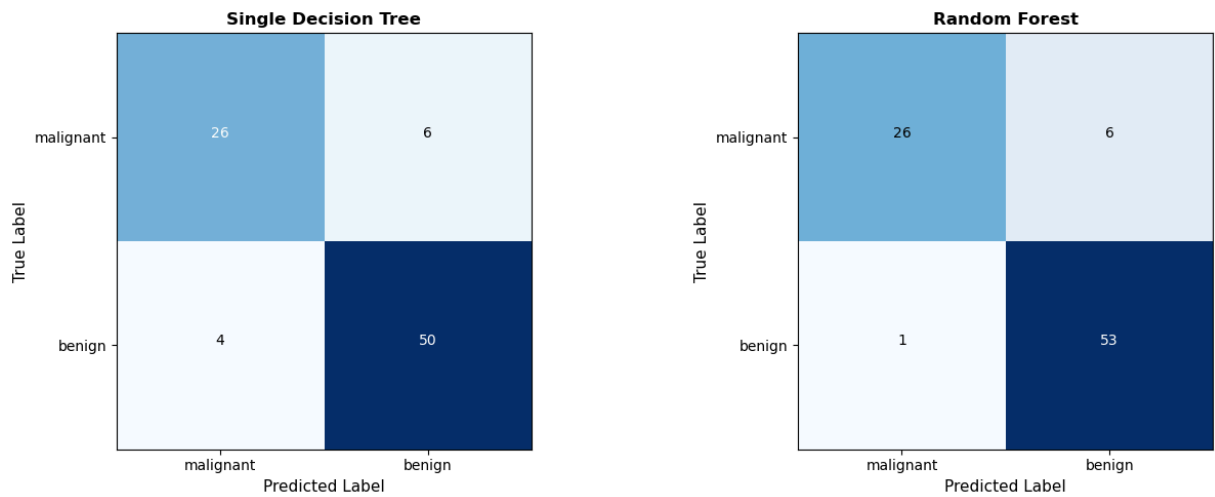
=====

COMPARISON: Single Decision Tree vs Random Forest

=====

Metric	Single Tree	Random Forest	Improvement
Test Accuracy	0.8837	0.9186	+0.0349
Macro-Precision	0.8798	0.9306	+0.0508
Macro-Recall	0.8692	0.8970	+0.0278
Macro-F1	0.8739	0.9097	+0.0358

=====



## Random Forest - Results Summary

### Hyperparameter Tuning

**Best Parameters:** `n_trees=30` , `max_features=15` ( $d/2$ ) with validation accuracy of **0.9882**

**Key Finding:** With `max_features=5` ( $\sqrt{d}$ ), all `n_trees` values achieved identical performance (0.9765), suggesting limited diversity. With `max_features=15` , `n_trees=30` and 50 both achieved optimal performance, indicating that sufficient feature diversity enables ensemble benefits.

### Final Model Performance

- **Test Accuracy:** 0.9186 (vs Single Tree: 0.8837) - **+3.49% improvement**
- **Macro-Precision:** 0.9306 (+5.08%)
- **Macro-Recall:** 0.8970 (+2.78%)
- **Macro-F1:** 0.9097 (+3.58%)

### Comparison with Single Decision Tree

Random Forest demonstrates clear improvement across all metrics, validating the ensemble approach:

- **Variance Reduction:** Averaging predictions from 30 diverse trees reduces variance
- **Better Generalization:** Test accuracy improvement shows better generalization to unseen data

- **Robustness:** Ensemble is less sensitive to outliers and noise

## Bias-Variance Tradeoff

- **Single Tree:** Lower bias, higher variance (sensitive to training data variations)
- **Random Forest:** Similar bias, significantly lower variance through:
  1. **Bootstrap sampling:** Each tree sees different data → diverse models
  2. **Random feature selection:** Different features at each split → increased diversity
  3. **Majority voting:** Averaging reduces variance while maintaining accuracy

The **+3.49% accuracy improvement** demonstrates that Random Forest successfully reduces variance without significantly increasing bias, leading to better generalization performance.