Decision Tree Model Analysis Report

1. Introduction

A **Decision Tree** is a simple but powerful machine learning model that makes predictions by splitting data into smaller groups based on feature conditions (like a flowchart). This report explains how different settings affect the Decision Tree's performance and which works best for the given dataset.

---

2. Methodology

2.1 How Decision Trees Work

- The tree **splits data** based on features (e.g., "Is Age > 30?").
- It uses **Gini Impurity** or **Information Gain** to decide the best splits.
- Stops splitting when it hits a limit (e.g., max depth or minimum samples per leaf).

2.2 Steps Taken

Step 1: Data Preprocessing

- **Checked for missing values** (no empty data).
- **Exploratory Data Analysis (EDA):** Studied how features relate to the target.
- **Encoded categorical data** (converted text into numbers).
- **Split data:**
    - **70% Training set** (to train the model).
    - **30% Testing set** (to check performance).

Step 2: Model Training

- Tested different **split criteria** (Gini, Entropy, Friedman MSE, Poisson).
- Adjusted **hyperparameters**:
    - max_depth (how deep the tree can grow).
    - min_samples_split (minimum samples needed to split a node).
    - max_features (how many features to consider for splitting).

Step 3: Performance Evaluation

Measured using:

- **Accuracy** (how often predictions are correct).
- **Precision, Recall, F1-Score** (for classification tasks).

---

3. Model Comparison & Results

Model Type	Key Settings	Accuracy	Performance
Default Decision Tree	No restrictions	**0.9333**	Best model (no tuning needed).
Limited Depth (max_depth=5)	Prevents overfitting	0.9271	Slightly worse but more stable.
Minimum Samples per Split (min_samples_split=10)	Fewer splits	0.9112	Reduces overfitting but lowers accuracy.
Friedman MSE Criterion	criterion="friedman_mse"	0.9154	Works but not better than default.
Poisson Criterion	criterion="poisson"	0.9159	Similar to Friedman MSE.
Restricted Features (max_features="sqrt")	Uses fewer features	**-0.7754**	**Failed** (needs all features).
Random State (random_state=42)	Ensures same results	0.9122	Good for reproducibility.
Limited Leaf Nodes (max_leaf_nodes=20)	Controls tree size	0.9057	Less overfitting but lower accuracy.

Key Findings:

1. **Default Decision Tree works best (93.3% accuracy)** → No tuning needed for this dataset.
2. **Limiting depth (**max_depth=5**) helps prevent overfitting** (slightly lower accuracy but more reliable).
3. **Restricting features (**max_features="sqrt"**) ruins performance** → All features are important.
4. **Other criteria (Friedman MSE, Poisson) don't improve results** → Stick with Gini/Entropy.

4. Recommendations

Best Model Choice:

Use Default Decision Tree (highest accuracy, no extra tuning needed).

If Overfitting is a Concern:

- **Limit tree depth (**max_depth=5**)** → Balances accuracy and stability.
- **Set** min_samples_split=10 → Prevents tiny, unreliable splits.

Avoid These Settings:

- **Don't restrict features (**max_features="sqrt"**)** → Causes model failure.
- **Don't use Friedman/Poisson criteria** → No improvement over default.

General Tips:

✓ **Always check feature importance** (see which features matter most).
✓ **Use** random_state=42 for reproducible results.
✓ **Try pruning (limiting leaf nodes)** if the tree is too complex.

---

5. Conclusion

- **Best Model: Default Decision Tree (93.3% accuracy)** → Simple and effective.
- **For Stability:** Limit depth (max_depth=5) or adjust min_samples_split.
- **Avoid:** Feature restriction (max_features="sqrt") → It breaks the model.

Final Decision: Use the **Default Decision Tree** for best results. If the model overfits, limit depth or adjust split samples.