Recommended Databases and Datasets for Classification Using Decision Tree Methods (Sorted by Difficulty)

I. Classic Beginner-Level Datasets (Suitable for Foundation Consolidation)

1. **UCI Machine Learning Repository**

   Recommended Datasets:

   - **Iris Plants Database**
     - **Features:** 3 flower classes (150 samples), 4 numerical features (petal/sepal length and width), classic classification task with strong visualizability.
     - **Advantages:** No missing data, low feature dimension, suitable for understanding decision tree pruning and feature importance analysis for beginners.
   - **Wine Dataset**
     - **Features:** 3 wine classes (178 samples), 13 chemical features, requiring processing of continuous data to demonstrate the decision tree's logic for handling multi-feature scenarios.
     - **Usage Suggestion:** Suitable as a warm-up case for major assignments, combined with visualization (e.g., feature space partitioning) to help understand decision tree splitting logic.
2. **Scikit-learn Built-in Datasets**

   Recommended Dataset:

   - **Breast Cancer Wisconsin Dataset**
     - **Features:** 2 classes (benign/malignant tumors, 569 samples), 30 medical features with a small number of missing values (needing preprocessing), which can demonstrate the decision tree's noise resistance combined with feature selection.

II. Intermediate-Difficulty Datasets (Suitable for Comprehensive Application)

1. **Kaggle Public Datasets**

   Recommended Datasets:

   - **Titanic: Machine Learning from Disaster**
     - **Features:** Binary classification (survived/died, 891 samples), including numerical features (age, fare), categorical features (gender, cabin class), and text features (name titles), requiring a complete data preprocessing flow (feature encoding, missing value handling). Suitable for demonstrating decision trees in mixed-feature scenarios.
   - **Mushroom Classification**
     - **Features:** Binary classification (edible/poisonous, 8124 samples), 22 categorical features, requiring key demonstration of the impact of One-Hot Encoding on decision trees. It can compare the splitting effects of information gain and Gini index.
     - **Usage Suggestion:** As the core task of major assignments, students are required to complete the full process from data cleaning to model tuning, suitable for examining feature engineering capabilities.
2. **OpenML Dataset Platform**

   Recommended Dataset:

   - **Bank Marketing Data Set**
     - **Features:** Binary classification (whether customers subscribe to term deposits, 41,188 samples), including numerical features (age, balance), categorical features (occupation, education level), and time features (contact month). The data scale is moderate, suitable for demonstrating decision tree pruning strategies (such as CCP pruning) with large-scale data.

III. Advanced Application Datasets (Suitable for Expansion and Enhancement)

1. Simplified Datasets for Image/Text Classification

   Recommended Datasets:

   - **Simplified MNIST Handwritten Digit Dataset**
     - **Features:** 10 digit classes (0-9, 1000 samples), 28×28 pixel grayscale images (flattened into 784-dimensional features), suitable for demonstrating decision tree performance in high-dimensional feature scenarios (though not the optimal model, it can compare differences with neural networks and analyze feature importance).
   - **Amazon Review Sentiment Analysis (Short Text)**
     - **Features:** Binary classification (positive/negative reviews, 1000 samples), requiring conversion to feature vectors using TF-IDF or bag-of-words models first. It can demonstrate the basic process of text classification with decision trees (suitable for students with programming foundations).
2. Medical/Bioinformatics Datasets

   Recommended Dataset:

   - **Heart Disease Prediction Dataset**
     - **Features:** Binary classification (whether having heart disease, 303 samples), 13 medical features (age, blood pressure, cholesterol, etc.), with partial missing values. It can explain decision tree splitting rules combined with domain knowledge (e.g., the relationship between cholesterol thresholds and disease risk).

IV. Dataset Selection and Task Design Suggestions

1. **Difficulty Matching:**
   - For students with weak foundations, prioritize small datasets like Iris, Mushroom Classification, and Breast Cancer to focus on core decision tree principles (splitting criteria, pruning).
   - For demonstrating comprehensive capabilities, select datasets like Titanic and Bank Marketing, requiring completion of the full process: "data preprocessing → model training → hyperparameter tuning (max_depth, min_samples_split) → result visualization."
2. **Optional Additional Tasks (Not Mandatory):**
   - Compare the performance differences between decision trees and random forests, and analyze how ensemble learning optimizes single trees.
   - Use graphviz to visualize the decision tree structure and explain the splitting logic of key features (e.g., the impact of "age > 30 years" on survival prediction).
   - Try different splitting criteria (information gain vs. Gini index) and observe their effects on tree structure and accuracy.
3. **Extra Tips:**
   - Exercise caution when choosing overly complex datasets (e.g., CIFAR-10 images, high-dimensional gene data) to avoid deviating from core decision tree knowledge points due to high preprocessing difficulty.