Decision Tree Classifier – Project Plan (Drugs A, B, C, X, Y Dataset)

This document outlines a complete plan for implementing a Decision Tree Classifier on the 'Drugs A, B, C, X, Y' dataset. The goal is to learn and apply all essential machine learning workflow steps: data cleaning, preprocessing, model training, and evaluation.

Step 1 – Import & Understand the Dataset

- Load the dataset using Pandas.
- Display first few rows (head).
- Check column names, data types, and basic statistics (info, describe).
- Identify categorical vs numerical features.

Step 2 – Data Cleaning

- Check for missing values and handle them (drop or fill).
- Check for duplicate rows and remove them.
- Ensure all data types are correct.

Step 3 – Exploratory Data Analysis (EDA)

- Explore relationships between features and target (Drug).
- Find counts for each Drug type.
- Look at distributions (Age, Na_to_K).
- Optional: visualize data using bar charts or histograms.

Step 4 – Data Preprocessing for Model

- Encode categorical variables (Sex, BP, Cholesterol).
- Split data into Features (X) and Target (y).

Step 5 – Split the Dataset

- Split data into training and testing sets (e.g., 70% train, 30% test).
- Use train_test_split from sklearn.

Step 6 – Train the Decision Tree Model

- Create a DecisionTreeClassifier object.
- Train the model on training data.
- Understand parameters like criterion (gini/entropy) and max_depth.

Step 7 – Make Predictions

- Use the trained model to predict on the test set.
- Compare predicted vs actual labels.

Step 8 – Evaluate Model Performance

- Calculate accuracy, precision, recall, and F1-score.
- Generate and interpret a confusion matrix.
- Summarize model performance.

Step 9 – Visualize the Decision Tree

- Plot the tree using sklearn.tree.plot_tree.
- Interpret branches and decision rules.

Step 10 – Make Conclusions

- Summarize key insights (important features, accuracy, etc.).
- Discuss possible improvements (like tree pruning or tuning).

This structured plan provides a clear roadmap for implementing a Decision Tree Classifier using the Drugs A, B, C, X, Y dataset. Follow each step carefully, implement code for each, and record your observations to strengthen your understanding of machine learning fundamentals.