Activity 4

Problem Specification: Telco Customer Churn Prediction

1. Overview

A telecommunications company (Telco) wants to predict whether a customer will churn (leave) or stay based on several customer-related features. The company collects data on various attributes of its customers and aims to predict if a customer will continue with their service or cancel their subscription. By analyzing these features, the company can implement targeted interventions to reduce churn and retain customers.

2. Objective

The goal is to build a **classification model** that predicts whether a customer will **churn** (leave) or **stay** based on the customer's attributes.

3. Features (Attributes)

Feature	Type	Description
CustomerID	Integer	Unique identifier for each customer.
Gender	Categorical	The gender of the customer (Male or Female).
SeniorCitizen	Binary	Whether the customer is a senior citizen (1: Yes, 0: No).
Partner	Categorical	Whether the customer has a partner (Yes or No).
Dependents	Categorical	Whether the customer has dependents (Yes or No).
Tenure	Integer	Number of months the customer has been with the company.
PhoneService	Categorical	Whether the customer has a phone service (Yes or No).
MultipleLines	Categorical	Whether the customer has multiple lines (Yes, No, or No phone service).
InternetService	Categorical	The type of internet service the customer subscribes to (DSL, Fiber optic, or No).
OnlineSecurity	Categorical	Whether the customer has online security (Yes or No).
OnlineBackup	Categorical	Whether the customer has online backup (Yes or No).
DeviceProtection	Categorical	Whether the customer has device protection (Yes or No).
TechSupport	Categorical	Whether the customer has tech support (Yes or No).
StreamingTV	Categorical	Whether the customer has streaming TV (Yes or No).
StreamingMovies	Categorical	Whether the customer has streaming movies (Yes or No).
Contract	Categorical	The type of contract the customer has (Month-to-month, One year, Two year).
PaperlessBilling	Categorical	Whether the customer is signed up for paperless billing (Yes or No).
PaymentMethod	Categorical	The payment method used by the customer (Electronic check, Mailed check, Bank transfer, Credit card).
MonthlyCharges	Float	The amount the customer pays for the service every month.
TotalCharges	Float	The total amount the customer has been charged.
Churn	Categorical	Target variable (Yes or No). Whether the customer has churned or stayed.

4. Target Variable

- **Churn**: A binary classification variable indicating if the customer has churned (left) or stayed with the company.
    - **Yes**: The customer has churned.
    - **No**: The customer has stayed.

5. Problem Type

This is a **binary classification** problem. The goal is to predict whether a customer will churn or stay based on the provided features.

Key Challenges

- **Imbalanced Dataset**: Churn prediction datasets often contain a higher number of customers who stay compared to those who churn, making the data imbalanced. This could affect model performance, especially when predicting the minority class (churn).
- **Non-linear Relationships**: There may be complex relationships between features that are non-linear, making it harder to classify customers based purely on a simple rule-based model.
- **Feature Engineering**: Some features, like **TotalCharges**, might need additional transformation to create more meaningful insights. For example, **TotalCharges** might be derived from multiplying **MonthlyCharges** by **Tenure**, but missing values could complicate this.

Evaluation Metrics

To evaluate the performance of the classification model, the following metrics will be considered:

- **Accuracy**: The proportion of correct predictions (both churn and stay).
- **Precision**: The proportion of true positive churn cases out of all predicted churn cases.
- **Recall**: The proportion of true positive churn cases out of all actual churn cases.
- **F1-Score**: The harmonic mean of precision and recall.
- **ROC Curve and AUC**: Plotting the receiver operating characteristic (ROC) curve and calculating the area under the curve (AUC) to evaluate how well the model distinguishes between churned and non-churned customers.

Approach

1. **Data Preprocessing**:
    - Handle missing values (e.g., use imputation or drop rows/columns).
    - Encode categorical variables using **Label Encoding** or **One-Hot Encoding**.
    - Scale numerical features using **StandardScaler** (e.g., **Tenure**, **MonthlyCharges**, **TotalCharges**).
    - Split the data into **training** and **test** sets.
2. **Model Selection**:
    - Use a **Decision Tree Classifier** to model the churn prediction problem.
    - Implement **Cross-Validation** to evaluate the performance of the model during training.
    - Perform **Hyperparameter Tuning** using **GridSearchCV** to optimize the model's performance.
    - Apply **Pruning** to reduce overfitting.
3. **Model Evaluation**:
    - Evaluate the trained model using various metrics like accuracy, precision, recall, F1-score, and AUC.
    - Visualize the **confusion matrix**, **ROC curve**, and other relevant performance metrics.
4. **Model Interpretation**:
    - Interpret the decision tree model to understand which features are most important for predicting churn.
    - Provide insights into the patterns associated with churn (e.g., customers with high monthly charges and no online security may have a higher likelihood of churning).

Considerations for Reducing Overfitting and Underfitting

- **Overfitting**: A decision tree that grows too deep could overfit the training data. To avoid this, we can apply **pruning** (both pre-pruning and post-pruning), set a maximum depth for the tree, or specify minimum samples per leaf or split.
- **Underfitting**: A shallow decision tree may not capture enough complexity in the data. To solve this, we could allow deeper trees or tune additional hyperparameters.
