Assignment 1

Predicting Customer Churn in a Subscription-Based Business

Scenario:

Imagine you are a data scientist working for **StreamFlex**, a subscription-based streaming service that provides movies, TV shows, and live sports content. The company is concerned about **customer churn**, where users cancel their subscriptions. Your task is to build a **predictive model using Decision Trees and Random Forests** to help the business identify which customers are likely to churn.

StreamFlex has provided you with a dataset containing various customer attributes such as:

- **Demographics**: Age, Location, Subscription Length
- **Usage Behavior**: Watch Time, Number of Logins, Preferred Content Type
- **Subscription Details**: Membership Type (Basic, Standard, Premium), Payment Method, Payment Issues
- **Customer Support Interactions**: Number of Complaints, Resolution Time

Your goal is to analyze and model customer churn using **Decision Trees and Random Forests** and evaluate model performance using appropriate classification metrics.

Assignment Deliverables

You are required to complete the following sections:

Section 1: Application of Decision Trees in Business (Theoretical Analysis)

- Why are decision trees useful in customer churn prediction?
- What business actions can be taken based on the predictions of a decision tree model?

Section 2: Python Implementation – Building the Model

Task 1: Data Preparation and Exploration

- Load the provided dataset (customer_churn.csv) into Python.
- Perform exploratory data analysis (EDA):

- Display summary statistics.
- Identify missing values and handle them appropriately.
- Visualize data distributions using histograms and box plots.
- Check for correlations between variables.

Task 2: Building a Decision Tree Classifier

- Split the dataset into training and testing sets.
- Train a **Decision Tree Classifier** using `scikit-learn`.
- Use `GridSearchCV` to optimize hyperparameters (e.g., max depth, min samples split).
- Visualize the decision tree.
- Evaluate model performance using:
  - Accuracy
  - Precision
  - Recall
  - F1 Score
  - Confusion Matrix

Task 3: Improving Performance with Random Forests

- Train a **Random Forest Classifier**.
- Compare its performance against the Decision Tree model.
- Analyze feature importance.
- Explain why the Random Forest model performed better (if applicable).

Task 4: Business Insights and Recommendations

- Based on the model's predictions, what characteristics contribute the most to customer churn?
- What actionable insights can StreamFlex use to **reduce customer churn**?
- Suggest **three concrete business strategies** based on your findings.

Submission Requirements

- A **written report** (5–7 pages) covering theoretical analysis, Python implementation, results, and business recommendations.
- Python **code files** (Jupyter Notebook or .py script) with appropriate documentation.
- Visualizations and figures to support your analysis.

Evaluation Criteria:

Criteria	Weight
Theoretical understanding	10%
Data Exploration & Cleaning	20%
Model Implementation	30%
Performance Evaluation	20%
Business Insights	20%

Due Date: check Moodle
Submission Format: PDF for the report, `.ipynb` or `.py` for the code
Academic Integrity: Ensure all work is original. Any plagiarism will result in penalties.