**Project Task: Predicting Customer Churn in a Subscription Service**

**Project Overview**

In this project, you will analyze customer data from a subscription-based service and develop a model to predict whether a customer will cancel their subscription. You’ll explore the dataset, process it, train predictive models, and evaluate their performance. Your goal is to identify patterns that indicate whether a customer is likely to "churn" (i.e., cancel the service) and provide actionable insights to reduce churn.

**Dataset Link:**

* [Telco Customer Churn Dataset on Kaggle](https://www.kaggle.com/datasets/yeanzc/telco-customer-churn-ibm-dataset/data)

**Week 1: Data Exploration and Preprocessing**

**Tasks:**

1. **Explore the Dataset:**
   * Download a publicly available dataset related to customer churn (or provide a curated dataset).
   * Examine the features (columns) and understand what each represents (e.g., customer demographics, usage patterns, subscription details, etc.).
   * Use summary statistics to understand the data distribution (e.g., mean, median, missing values).
2. **Data Cleaning:**
   * Handle missing values appropriately (imputation or removal).
   * Identify and remove irrelevant or redundant features that do not contribute to the prediction task.
3. **Data Visualization:**

* **Target Variable Analysis:**
  + - * + Visualize the distribution of the Churn variable using a **count plot** to see how many customers have churned vs. not churned.
        + Use a **pie chart** to get a proportionate view of the churn vs non-churn distribution.
* **Feature Distribution:**
  + - * + For **categorical features** (e.g., Contract, PaymentMethod), use **bar plots** or **count plots** to visualize how different categories are distributed.
        + For **numerical features** (e.g., MonthlyCharges, TotalCharges, Tenure), use **histograms** or **box plots** to understand the distribution, central tendency, and any outliers.
* **Churn Correlation:**
  + - * + Use **box plots** or **violin plots** to visualize the relationship between numerical features and churn
        + For categorical features, use **stacked bar plots** or **grouped bar charts** to compare churn rates across categories
* **Correlation Heatmap:**
  + - * + Create a **correlation matrix** and plot a **heatmap** to visualize relationships between numerical features and identify any strong correlations between them.
        + **Pair Plots (Optional):**
        + Use **pair plots** to visualize relationships between multiple numerical features and churn. This will give a deeper insight into the interactions between features.
* **Initial Insights:**
  + - * + Summarize key insights from your visualizations. For example:

Do churn customers have higher/lower charges?

Are certain contract types or payment methods associated with higher churn?

1. **Feature Engineering:**
   * Analyze categorical and numerical features.
   * Encode categorical features appropriately (One-Hot Encoding, Label Encoding).
   * Create new features, if relevant, based on customer behavior (e.g., average usage time, frequency of interactions).
2. **Data Transformation:**
   * Normalize or scale features (e.g., MinMaxScaler or StandardScaler) to ensure consistent scales for all numeric features.

**Deliverables:**

* Cleaned and preprocessed dataset.
* Jupyter notebook with steps and explanations for data exploration and preprocessing.

**Week 2: Problem Understanding and Model Selection**

**Tasks:**

1. **Identify the Problem Type:**
   * Analyze the data and determine what type of problem this is (e.g., regression, classification, etc.).
   * Justify your reasoning based on patterns in the target variable (whether the customer canceled or remained subscribed).
2. **Split the Data:**
   * Split the dataset into training and testing sets (e.g., 80% training, 20% testing).
3. **Select Algorithms:**
   * Based on your problem identification, choose at least three suitable algorithms (e.g., if it's a classification problem, consider Logistic Regression, Decision Trees, Random Forest, etc.).
4. **Model Training:**
   * Train the selected models on the training data.
5. **Hyperparameter Tuning:**
   * Tune the models' hyperparameters using GridSearchCV or RandomizedSearchCV to find the best configurations.

**Deliverables:**

* A Jupyter notebook showing the reasoning behind problem identification.
* Trained models with tuned hyperparameters.

**Week 3: Model Evaluation and Business Insights**

**Tasks:**

1. **Evaluate the Models:**
   * Use appropriate evaluation metrics based on the problem type (e.g., accuracy, precision, recall, F1-score, AUC-ROC for classification).
   * Compare the models based on their performance metrics and choose the best-performing model.
2. **Cross-Validation:**
   * Apply cross-validation (k-fold) to ensure the model generalizes well to unseen data.
3. **Customer Insights:**
   * Analyze the top features that influence customer churn (e.g., service usage patterns, contract type, payment methods).
   * Provide actionable insights that the business can implement to reduce churn (e.g., personalized offers for at-risk customers).
4. **Final Presentation:**
   * Summarize the project in a report or presentation, including:
     + Problem identification
     + Data preprocessing
     + Model performance comparison
     + Key insights and recommendations for the business.

**Deliverables:**

* Final model with evaluation metrics.
* A presentation or report summarizing the analysis, model performance, and recommendations.
* Well-commented Jupyter notebooks.