Lecture 6: Review and Class Activity

Objective

The goal of this activity is to guide through data preparation, exploration, visualization, and modeling processes using a real-world climate dataset.

Data Engineering

1. **Handle Missing Values**:
   - Use mean imputation for numerical columns.
   - Forward-fill categorical columns.
   - Discuss: How does handling missing values affect the analysis process?
2. **Standardize or Normalize Numerical Columns**:
   - Standardize or normalize all numerical columns (e.g., temperature values).
   - Explain the difference between standardization and normalization and why it is important for machine learning.
3. **Feature Engineering**:
   - Create a new feature, `Temperature_Trend`, which calculates the average temperature change over the last 10 years (2012–2022) for each country.
   - Add this feature as a column in the dataset.
4. **Encode Categorical Columns**:
   - Encode the `Country` and `ISO2` columns using a suitable encoding technique (e.g., one-hot or label encoding).
   - Justify your choice of encoding.

Exploratory Data Analysis (EDA)

1. **Top 10 Countries with Highest Temperature Trends**:
   - Find the top 10 countries with the highest temperature trend over the last decade (2012–2022).
   - Visualize the results using a horizontal bar chart.

2. **Global Heatmap for Temperature Changes (2022)**:
    - Plot a global heatmap showing temperature changes for 2022 using a geospatial visualization library like `plotly` or `folium`.
3. **Global Average Temperature Change (Time Series)**:
    - Create a time series plot of the global average temperature change (1961–2022).
    - Highlight significant trends or patterns in the data.
4. **Distribution of Temperature Changes (2022)**:
    - Visualize the distribution of temperature changes for a selected year (e.g., 2022) using a histogram or KDE plot.
    - Describe the shape of the distribution.
5. **Country-wise Comparison of Temperature Changes**:
    - Compare temperature changes for 5 specific countries of your choice over the years using a grouped line chart.
    - Identify and discuss observed patterns.

Machine Learning Models

Supervised Learning

1. Train a simple linear regression model:
    - Predict temperature change for a given year (1961–2022) based on historical data (features: past temperature changes).
    - Evaluate the model using Mean Squared Error (MSE).

Unsupervised Learning

1. Apply k-means clustering:
    - Group countries based on their temperature trends over the last 20 years.
    - Use `k=3` clusters and visualize the clusters using a scatter plot.
2. Perform Principal Component Analysis (PCA):
    - Reduce the dimensionality of the dataset (focus on temperature columns).
    - Visualize the first two principal components for all countries and interpret the results.

Interpretation and Insights

- Write a summary of your findings:
  - What trends do you observe in the data?
  - How do the machine learning models help in understanding temperature change patterns?
  - Which regions are most impacted, and why might this be happening?