Week 4, Lab 3: Data Preprocessing-II

Duration: Until 3 PM today

Objective:

Advance your data preprocessing skills by applying a range of techniques to prepare the "Air_county.csv" dataset for complex modeling.

Instructions for Submission:

- **Python Script (.ipynb):**
  - Write and execute your code in a Jupyter Notebook environment.
  - Include comments within your script to explain the purpose and functionality of each code block.
- **PDF Conversion:**
  - After completing the lab in a Jupyter Notebook, convert the file to a PDF that includes all code and output.
  - Ensure that the PDF is well-formatted, readable, and includes your name and student ID at the top.

Questions:

1. Normalization and Standardization

- **Q1:** Normalize the dataset using Min-Max normalization.
- **Q2:** Standardize the dataset using Z-score standardization.
- **Q3:** Compare the results of normalization and standardization on the Max AQI column. When would you prefer one technique over the other?

2. Encoding Categorical Variables

- **Q4:** Convert the categorical variable State into a numerical form using one-hot encoding. How many new columns are created, and what are their names?
- **Q5:** Use label encoding for the County column. What are the numerical labels assigned to each county?

- **Q6:** Discuss the impact of using one-hot encoding vs label encoding on the model performance when the categorical variables are used in regression analysis.

3. Discretization and Binning

- **Q7:** Bin the Max AQI column into 4 categories: "Low", "Moderate", "High", and "Very High" using equal-width binning. How many rows fall into each category?

4. Handling Skewed Data

- **Q8:** Check the skewness of the Max AQI and Median AQI columns. Which of these columns is highly skewed?
- **Q9:** Apply a log transformation to the Max AQI column to handle its skewness. How does this affect the skewness?
- **Q10:** After applying the log transformation, what changes do you observe in the distribution of the Max AQI data?

5. Feature Engineering

- **Q11:** Create a new feature called Bad_Air_Days, which is the sum of Unhealthy for Sensitive Groups Days, Unhealthy Days, Very Unhealthy Days, and Hazardous Days. How many counties have more than 10 Bad_Air_Days?

6. Multicollinearity

- **Q12:** Calculate the correlation matrix for all numeric variables. Which variables are highly correlated with each other?
- **Q13:** Based on the correlation matrix, identify pairs of features that might cause multicollinearity. How could you address this issue?
- **Q14:** Remove one of the highly correlated features from the dataset. Recompute the correlation matrix and compare it with the original one. What changes do you observe?

Deliverables:

- Upload your .ipynb and .pdf files to the designated Canvas submission area labeled “Week4-Lab 4: DataPreprocessing-II” before the deadline.

- Ensure both files are named appropriately (e.g., LastName.Week4-Lab3) and include necessary comments and documentation.