ST 228: Data Analysis, ML and AI

Assignment # 3 (10 marks)

(Data Handling, Impuration, Time Series, PCA, Normalization and Regression)

28th of Jan 2025

Due on: 5th **of February 2025 before 5 PM**

Instructions :

- Clearly state all the assumptions made.
- Clearly quote the source of data used.
- Clearly show your work
- Reports exceeding with the word limit will not be considered.

1. **Data Collection and Preprocessing**

   a. How did you visualize the time-series data to identify trends in air quality before and after the lockdown? What insights did you gain from this visualization?

2. **Correlation Analysis**

   a. What are the key correlations observed between the Air Quality Index (AQI) and individual pollutants in both cities?

   b. How does the correlation between AQI and pollutants in Bangalore compare to that of Delhi? What potential environmental or industrial factors could explain these differences?

3. **Normalization and Its Effect**

   a. Explain the process and importance of applying normalization (Min-Max scaling or Standardization) to the dataset?

4. **Regression Analysis**

   a. Perform a regression analysis to predict AQI based on individual pollutants. What is the relationship between specific pollutants and AQI in both cities?

   b. Compare the results of linear regression models before and after normalization. Did normalization improve the regression model's performance or interpretability? How did you handle missing values in the dataset for Delhi (DL) and Bangalore (KA)? What imputation methods/removal techniques were applied, and why?

5. **Principal Component Analysis (PCA)**

   a. Perform PCA on the dataset for both Delhi and Bangalore. What are the principal components, and how do they explain the variance in the data?

   b. is PCA dependent on normalization?

6. **Impact of Lockdown on Air Quality**

   a. Using regression and correlation analysis, assess the impact of the lockdown period on air quality in both cities. Did the regression models show significant changes in pollutant levels during the lockdown?

   b. How did the normalization of data help identify changes in air quality during the lockdown period? Was there a notable difference in the distribution of AQI during this time?

7. **Model Evaluation**

   a. What statistical metrics (e.g., R-squared, RMSE) did you use to evaluate the performance of regression models? How did these metrics differ before and after normalization?

   b. Discuss the overall impact of PCA on model performance. Did reducing dimensionality result in better model accuracy or simplification without sacrificing important information?