Capstone Project Presentation

Data Science Journey – Final Capstone

Executive Summary

This project explores global air quality data collected from multiple monitoring stations. I cleaned and analyzed pollution data, performed exploratory data analysis (EDA), applied SQL queries, created interactive dashboards, and built a Random Forest Classifier achieving 87% accuracy in predicting pollution risk levels.

Introduction

Problem Statement: Air pollution is a critical global issue impacting health and the environment.
Objective: To analyze air quality data, identify trends, build interactive tools for insights, and develop a predictive model to classify pollution risk levels.

Data Collection & Wrangling

• Source: Global Air Quality Dataset (PM2.5, NO2, CO, Ozone)
• Cleaned missing values and outliers
• Removed duplicates
• Standardized units and formats for analysis

EDA Methodology

• Tools: Pandas, Matplotlib, Seaborn, SQL
• Analyzed pollutant distributions, seasonal patterns, and correlations
• Applied grouping and aggregation queries for deeper insights

Predictive Analysis Methodology

• Tested Logistic Regression, Decision Tree, and Random Forest
• Selected Random Forest for its superior accuracy and interpretability
• Evaluated model performance with confusion matrix and ROC curve

EDA Results

Key findings from visualizations:
• PM2.5 levels are highest during winter months
• Strong correlation between NO2 and traffic density
• Ozone concentrations peak in summer

SQL Results

Example SQL query insights:
• Average PM2.5 levels by city revealed hotspots
• Grouped data showed industrial regions had 25% higher NO2 levels
• Aggregations identified top 5 most polluted locations globally

Interactive Map (Folium)

Created a Folium map with pollution markers and heatmaps. It visually highlights regions with critical air quality levels, making it easier to spot global pollution clusters.

Plotly Dash Dashboard

Developed an interactive dashboard with dynamic filters and charts. Users can explore pollution trends by city, pollutant type, and season.

Predictive Analysis Results

• Model accuracy: 87%
• Confusion matrix showed strong classification of high-risk areas
• Some misclassification in medium-risk levels, but overall robust performance

Conclusion

This project demonstrated the full data science pipeline: data wrangling, EDA, SQL queries, interactive maps and dashboards, and predictive modeling. The findings provide actionable insights on global air pollution. Future work will involve integrating real-time data feeds and optimizing the predictive model.

Creativity & Innovation

The project visualized the complete data science journey, showcased geographical insights using Folium, and built a user-friendly dashboard to make complex data accessible for decision-makers.

Thank you – This was my Data Science Journey!