🌍 AI-Based Urban Environmental Risk Prediction System

(City-Specific | Time-Series Driven | Multi-Hazard Intelligence Platform)

📌 1. Problem Statement (In Detail)

Modern cities face increasing environmental threats such as:

• rising air pollution
• contaminated water sources
• floods due to heavy rainfall
• poor urban drainage systems
• illegal industrial dumping
• overpopulation stress
• socially vulnerable communities

Currently, most systems are **reactive** — action happens after damage occurs.

🎯 Our goal:

Build an **AI-powered backend system** that predicts future environmental risks using historical data so authorities can act **before problems become disasters**.

✅ The system will provide:

1. **Pollution Forecasts**
   Predict future PM2.5 and PM10 levels for upcoming days.
2. **Environmental Risk Score**
   A numeric risk value (0–100) showing how dangerous an area will become.
3. **High-Risk Zone Identification**
   Detect which city zones are repeatedly becoming unsafe.

🏙️ 2. Modeling Scope – City-Specific Intelligence

Instead of one global model for all cities, we use:

✔ City-Specific Models

Each city will have:

• its own dataset
• its own trained ML models

💡 Why this is better:

Different cities have:

• different industries
• different climate
• different population density
• different governance patterns

So their risk patterns are unique.

🎯 Hackathon Approach:

We will:

➡ Select ONE city
➡ Train model using its past data
➡ Demonstrate accurate future prediction

Later, the same system can scale to multiple cities.

📊 3. Data Structure – Time Series Based

The system works on **daily historical records**.

Each row represents:

📍 One zone of the city for one specific day

Example row:

Zone Date PM2.5 Rainfall Water Quality Violations Population Density Risk

This allows:

✔ trend detection
✔ seasonality learning
✔ cause-effect understanding

🎯 4. Final Feature Set (Core Inputs)

We divide features into logical categories.

🌫 AIR QUALITY

• PM2.5 – fine particulate pollution
• PM10 – coarse particulate pollution
• NO2 – nitrogen dioxide level
• pollution_trend_3days – recent rise/fall pattern

🌧 WEATHER & CLIMATE

• humidity
• wind_speed
• rainfall_last_3_days

These influence how pollution spreads and flood formation.

💧 WATER & RESOURCES

• water_quality_index – contamination level
• reservoir_level – available water storage

Helps detect water crisis risks.

⚠ ENVIRONMENTAL VIOLATIONS

• violations_last_7_days – recent incidents
• avg_violation_severity – seriousness of violations
• repeat_offender_rate – recurring pollution sources

These show human-caused environmental stress.

🏙 URBAN STRUCTURE

• population_density
• industrial_density
• green_cover_percentage
• drainage_quality_index

These define city resilience and pollution pressure.

👥 SOCIAL RISK

• social_vulnerability_index

Represents poor living conditions and exposure risk.

(Optional: average_temperature)

⚙ 5. Feature Engineering Pipeline (How raw data becomes ML-ready)

Raw data is not directly useful — we process it.

📈 Pollution trend:

Calculate 3-day moving slope:

Shows if pollution is increasing or decreasing.

⚠ Violation metrics:

• total violations in last 7 days
• weighted severity score
• frequency of same offenders

🌧 Rainfall processing:

Sum of rainfall over recent days for flood stress.

📏 Normalization:

All values scaled between 0–1 for ML stability.

🧠 6. Model Architecture (Multi-Model System)

Instead of one confused model, we use focused models.

🌫 Model A – Pollution Forecast Model

🎯 Purpose:

Predict future PM2.5 & PM10 for next 3–7 days.

📥 Inputs:

Air data
Weather
Violations
Industrial density

📤 Output:

Future pollution levels.

🧠 Suggested Algorithms:

✔ Random Forest Regressor (stable & fast)
✔ Optional LSTM for deep time series

🟢 Model B – Environmental Risk Score Model (Main Model)

🎯 Purpose:

Predict overall environmental danger.

📥 Inputs:

All 16 engineered features.

📤 Output:

risk_score between 0 and 100.

🚦 Risk categories:

Low Risk → below 30
Medium Risk → 30 to 70
High Risk → above 70

🧠 Suggested Models:

✔ Gradient Boosting
✔ Random Forest

□ Model C – Hotspot Detection (Optional)

🎯 Purpose:

Find zones that are structurally dangerous over time.

📥 Inputs:

Pollution + violations + infrastructure + vulnerability

🧠 Algorithm:

K-Means or DBSCAN clustering

📤 Output:

High-risk area clusters.

📈 7. Target Variable Design – Risk Score

Initially we compute a logical risk score:

```
risk_score =
 0.4 × air_pollution_index
+0.2 × water_risk_index
+0.2 × violation_risk
+0.1 × flood_risk
+0.1 × social_vulnerability
```

This creates training labels.

Later the ML model automatically learns better relationships.

🗄 8. Backend Data Storage (MongoDB)

🗀 sensors collection

Stores daily air & water readings.

🗀 weather collection

Stores daily weather data.

🗀 violations collection

Stores incident-level pollution cases.

🗀 engineered_features collection

Stores ML-ready dataset.

🗀 predictions collection

Stores daily risk predictions.

🔁 9. Backend Workflow

```
Collect data
↓
Clean & merge datasets
↓
Feature engineering
↓
Train city-specific models
↓
Generate predictions
↓
Store results
↓
Serve via APIs/dashboard
```

📊 10. Training Strategy

Per City:

✔ Minimum 6 months data
✔ Ideal 1–3 years

Validation:

Rolling time-window split (no data leakage)

Metrics:

Pollution forecast → RMSE
Risk classification → Accuracy, F1-score

🔨 12. Hackathon Execution Plan

Phase 1 – Data Preparation

Collect or simulate city data

Phase 2 – Feature Pipeline

Build trend & aggregation logic

Phase 3 – Model Training

Train pollution + risk models

Phase 4 – Prediction Engine

Generate daily forecasts & risk scores

🏆 Final Vision

This project becomes:

A smart city environmental intelligence system that predicts pollution, water stress, flood risk, and urban vulnerability using AI.