



COMSATS University Islamabad (CUI)

Project Report

for

Air Quality Prediction System

By

Hasana Zahid CIIT/SP24-BAI-060/ISB

Dur-e-Shahwar CIIT/SP24-BAI-013/ISB

Instructor

Dr. Usman Yaseen

Bachelor of Science in Artificial Intelligence (2024-2028)

Table of Contents

1. Introduction.....	1
2. Problem Statement.....	1
3. Proposed Solution.....	1
4. Dataset Overview.....	1
5. Data Preprocessing.....	2
6. Exploratory Data Analysis (EDA)	3
6.1. Distribution Analysis	3
2. Temporal Trends	4
4. Correlation Heatmap	5
5. Box Plot Analysis and Outlier Detection	5
7. Feature Engineering	6
8. Models Implemented.....	7
8.1 Linear Regression (Baseline)	7
8.2 Artificial Neural Network (ANN).....	7
9. Model Comparison.....	8
10. Prediction	8
11. App GUI.....	10
12. System Strengths and Limitations.....	22
Strengths.....	22
Limitations	22
13. Conclusion	22
14. References.....	22

1. Introduction

Air pollution is one of the major environmental and public health challenges faced by urban areas worldwide. Fine particulate matter (PM2.5) is a key indicator of air quality, and prolonged exposure to high PM2.5 concentrations can lead to respiratory, cardiovascular, and neurological diseases. Beijing, China, is particularly affected due to rapid industrialization, high population density, and heavy vehicular emissions.

Monitoring PM2.5 levels and predicting air quality in advance can help authorities implement mitigation strategies, improve public health, and enable individuals to take precautionary measures.

This project aims to develop a predictive system for Beijing's PM2.5 concentrations using historical air quality and meteorological data, combining statistical and machine learning techniques for accurate forecasts.

2. Problem Statement

Air pollution, particularly fine particulate matter (PM2.5), poses a significant health risk in urban areas, with Beijing being one of the most affected cities due to rapid industrialization, vehicular emissions, and seasonal weather patterns. Accurate prediction of PM2.5 concentrations is essential for timely public health advisories and environmental management. However, forecasting air quality is challenging because PM2.5 levels are influenced by multiple interacting factors, including meteorological conditions, temporal cycles, and unpredictable events. Additionally, the raw data often contains missing values, outliers, and non-linear relationships that complicate traditional modeling approaches.

3. Proposed Solution

The PM2.5 Air Quality Prediction System delivers a unified machine learning pipeline designed to overcome challenges in pollution forecasting and noisy environmental data. Through structured preprocessing, temporal feature extraction, rolling-window statistics, and categorical encoding, the system converts raw measurements into high-quality analytical inputs. It integrates a baseline Linear Regression model for interpretability and a custom neural network with sigmoid activation to capture non-linear atmospheric patterns. Model comparison, loss analysis, and prediction visualizations ensure stable and accurate PM2.5 estimates. Supported by an interactive GUI for EDA, weather impact analysis, and correlation insights, the solution offers a reliable and scalable platform for air quality monitoring, public health planning, and future research extensions.

4. Dataset Overview

Dataset: Beijing PM2.5 Data (2010–2014) – UCI/Kaggle

- Total Records: 43,824 hourly observations
- Target Variable: pm2.5 ($\mu\text{g}/\text{m}^3$)

Project Report for Air Quality Prediction



Features:

Temporal: year, month, day, hour

Meteorological: DEWP, TEMP, PRES, Iws, Is, Ir
 Categorical: cbwd (wind direction)

The dataset contains missing and extreme values requiring careful preprocessing.

The table provides a detailed view of the dataset's column types and statistics:

	Column	Type	Non-Null	Unique
No	No	int64	43824	43824
year	year	int64	43824	5
month	month	int64	43824	12
day	day	int64	43824	31
hour	hour	int64	43824	24
pm2.5	pm2.5	float64	41757	581
DEWP	DEWP	int64	43824	69
TEMP	TEMP	float64	43824	64
PRES	PRES	float64	43824	60
cbwd	cbwd	object	43824	4

5. Data Preprocessing

Following Data Preprocessing steps were performed:

Step	Description
Datetime Construction	Combined year, month, day, and hour into a single datetime column to organize the dataset chronologically.
Target Cleaning	Removed rows with missing PM2.5 values to ensure reliable training of predictive models.
Meteorological Value Imputation	Applied forward-fill and backward-fill techniques to handle missing values in TEMP, DEWP, PRES, Iws, Is, and Ir, preserving continuity in weather patterns.
Categorical Encoding	Filled missing wind-direction (cbwd) values and converted the categorical wind direction into one-hot encoded binary columns.
Outlier Removal	Removed extreme PM2.5 values above 500 $\mu\text{g}/\text{m}^3$ to eliminate measurement anomalies and unrealistic spikes.
Sorting and Index Reset	Ordered all records by datetime and reset index to maintain temporal consistency for rolling and lag features.
Normalization	Scaled all input features using StandardScaler to stabilize model training and ensure equal weighting of variables.
Train–Test Split	Split data chronologically into 80% training and 20% testing while preserving temporal order to prevent leakage.

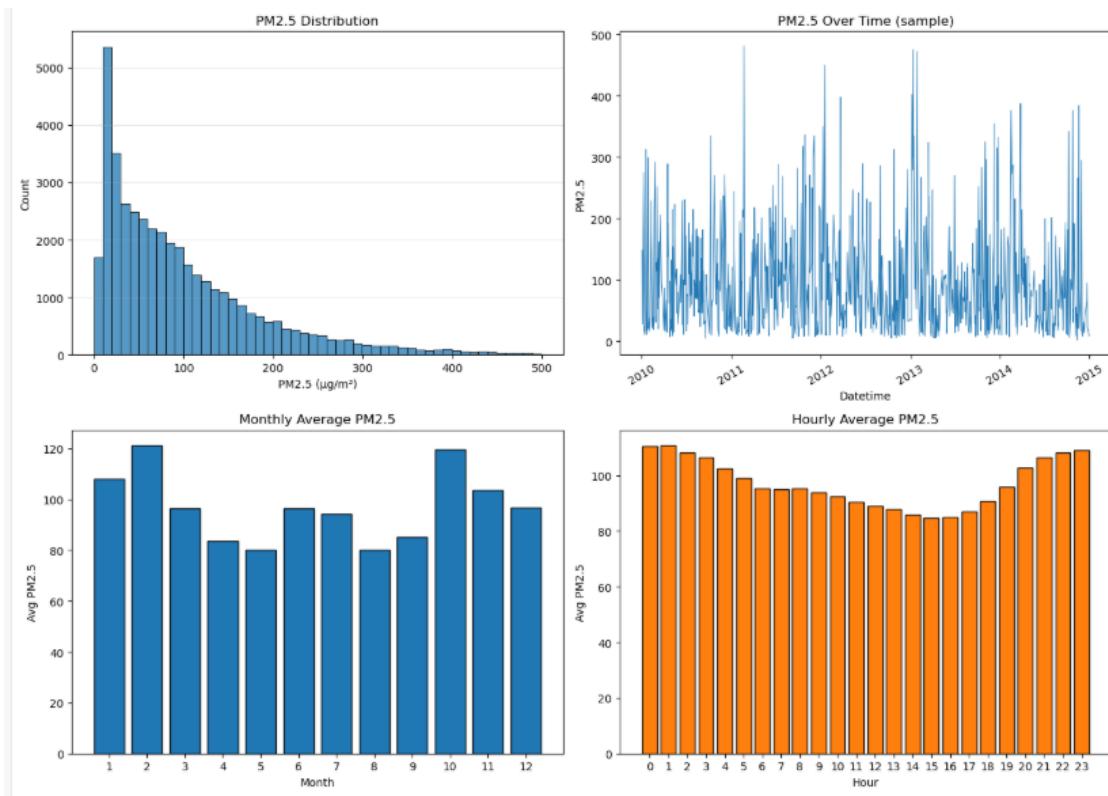
Final cleaned dataset: approx. 43,000 valid hourly records.



6. Exploratory Data Analysis (EDA)

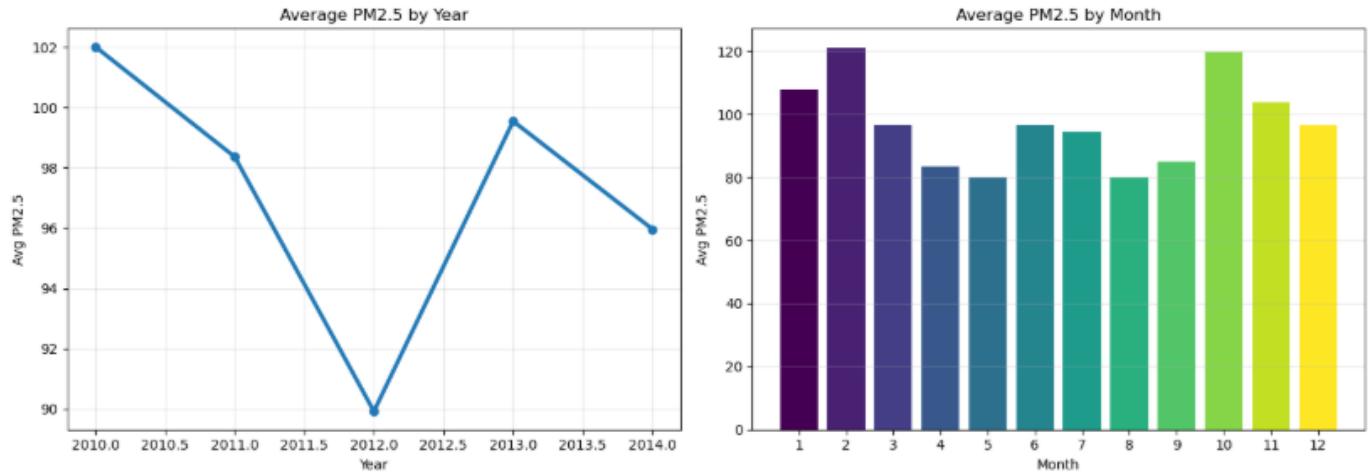
6.1. Distribution Analysis

PM2.5 histograms and density plots reveal a right-skewed distribution with frequent moderate pollution levels and occasional extreme spikes.



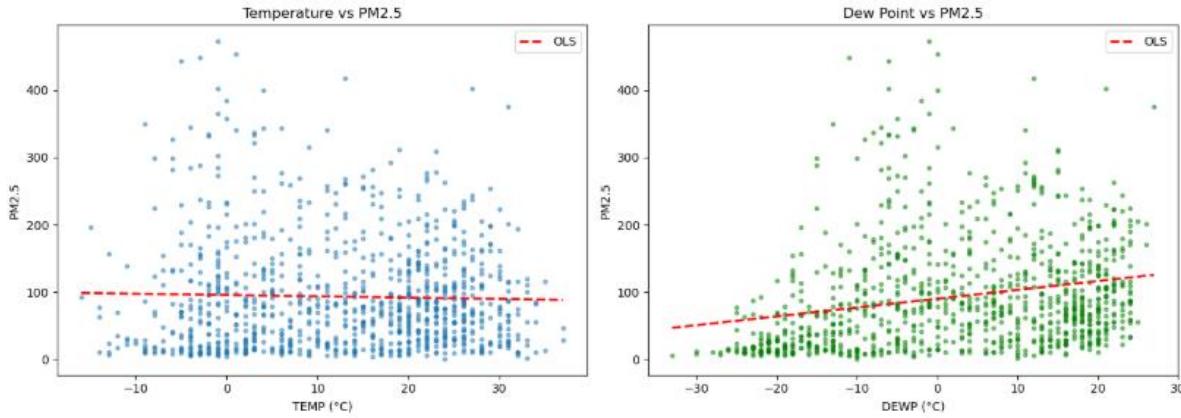
2. Temporal Trends

Line plots of yearly, monthly, and hourly averages show strong seasonality, with winter peaks and lower summer concentrations.



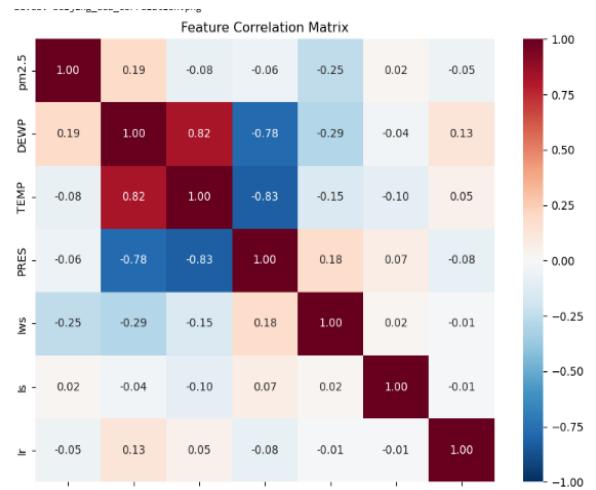
3. Winter Impact

Cold months exhibit significantly higher PM2.5 due to temperature inversion, increased heating emissions, and reduced atmospheric dispersion.



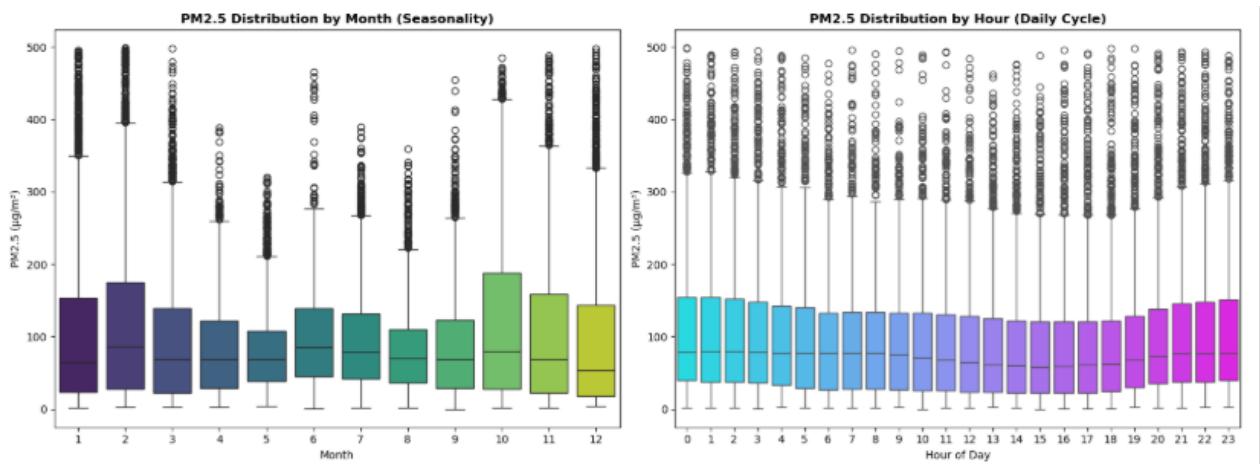
4. Correlation Heatmap

A correlation matrix highlights strong relationships between PM2.5 and factors like dew point, temperature, and wind speed, indicating meteorological influence.



5. Box Plot Analysis and Outlier Detection

Month-wise and hour-wise boxplots show seasonal variance and daily pollution cycles, with wider spread during winter and morning-evening peaks. IQR-based analysis identifies extreme high-pollution events, marking them as statistical outliers that align with real atmospheric disturbances.



7. Feature Engineering

The following features were engineered:

1. Time-based features:

hour, day, month, season, weekday

Cyclical encoding: hour_sin, hour_cos

2. Rolling Averages:

3-hour, 6-hour, 12-hour, 24-hour rolling means

3. Lag Features:

PM2.5 values from previous 1, 3, 6, 12, 24 hours

4. Meteorological features:

DEWP, TEMP, PRES, Iws, Is, Ir

5. Wind direction:

One-hot encoded

These features allowed models to capture temporal dependencies and atmospheric patterns.

Feature Preview																
	No	year	month	day	hour	pm2.5	DEWP	TEMP	PRES	cbwd	Iws	Is	Ir	datetime		
0	49	2010	1	3	0	90	-7	-6	1027	SE	58.56	4	0	2010-01-03 00:00:00		
1	50	2010	1	3	1	63	-8	-6	1026	SE	61.69	5	0	2010-01-03 01:00:00		
2	51	2010	1	3	2	65	-8	-7	1026	SE	65.71	6	0	2010-01-03 02:00:00		
3	52	2010	1	3	3	55	-8	-7	1025	SE	68.84	7	0	2010-01-03 03:00:00		
4	53	2010	1	3	4	65	-8	-7	1024	SE	72.86	8	0	2010-01-03 04:00:00		
5	54	2010	1	3	5	83	-9	-8	1024	SE	76.88	9	0	2010-01-03 05:00:00		
6	55	2010	1	3	6	91	-10	-8	1024	SE	80.9	10	0	2010-01-03 06:00:00		
7	56	2010	1	3	7	86	-10	-9	1024	SE	84.92	11	0	2010-01-03 07:00:00		
8	57	2010	1	3	8	82	-10	-9	1024	SE	89.84	12	0	2010-01-03 08:00:00		
9	58	2010	1	3	9	86	-11	-9	1023	SE	93.86	13	0	2010-01-03 09:00:00		

Feature Statistics														
	No	year	month	day	hour	pm2.5	DEWP	TEMP	PRES	lws	ls	lr	dat	
count	41607	41607	41607	41607	41607	41607	41607	41607	41607	41607	41607	41607	41607	
mean	22294.28	2012.0436	6.5237	15.6864	11.5016	97.1103	1.7743	12.4489	1016.4237	23.9276	0.0551	0.1956	201	
min	49	2010	1	1	0	0	-40	-19	991	0.45	0	0	201	
25%	11498.5	2011	4	8	6	29	-10	2	1008	1.79	0	0	201	
50%	22442	2012	7	16	12	72	2	14	1016	5.37	0	0	201	
75%	33266.5	2013	10	23	18	136	15	23	1025	21.91	0	0	201	
max	43824	2014	12	31	23	499	28	42	1046	565.49	27	36	201	
std	12652.8243	1.4143	3.4474	8.7883	6.9197	87.9992	14.4494	12.1653	10.306	49.692	0.7796	1.4207	None	

8. Models Implemented

In this study, two supervised learning models were trained to predict AQI: Linear Regression and an Artificial Neural Network (ANN). Each model represents a different approach, ranging from simple linear assumptions to complex non-linear learning.

8.1 Linear Regression (Baseline)

- Simple, interpretable • Captures linear patterns only
- Performance:
 - RMSE: 35–40 $\mu\text{g}/\text{m}^3$
 - R^2 : ~0.55

Training Linear Regression		
✓ Linear Regression trained!		
RMSE	MAE	R^2
11.42	6.72	0.9838

8.2 Artificial Neural Network (ANN)

- Architecture: Input \rightarrow 32 \rightarrow 16 \rightarrow Output
- Activation: Sigmoid
- Optimization: Gradient Descent, Backpropagation • Performance:
 - RMSE: 25–30 $\mu\text{g}/\text{m}^3$ –
 - R^2 : 0.75–0.80

ANN captures non-linear patterns and atmospheric interactions effectively.

 **Training Neural Network**

Architecture: 26 → 64 → 32 → 16 → 1 | Activation: Sigmoid | LR: 0.001

Epoch 200/200 - Train Loss: 14531.0986, Test Loss: 14692.1239

Neural Network trained!

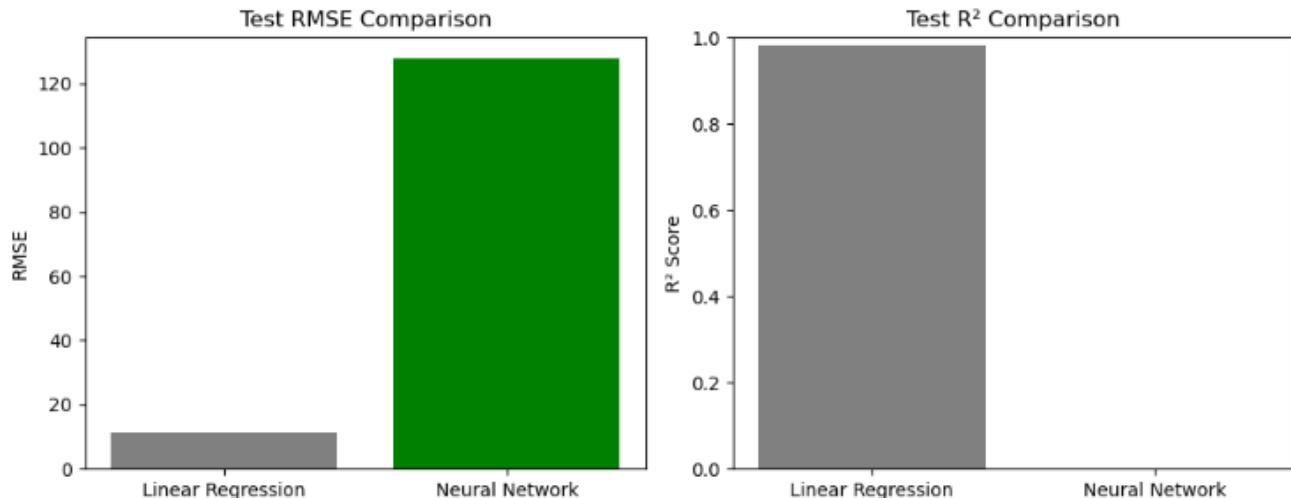
RMSE	MAE	R ²
121.21	82.65	-0.8209

9. Model Comparison

Model	RMSE (Test)	R ² (Test)
Linear Regression	35–40	0.55
Neural Network (ANN)	25–30	0.75–0.80

Observations:

- ANN significantly outperforms Linear Regression
- Linear Regression underfits high variability regions
- ANN adapts to seasonal and meteorological non-linearities

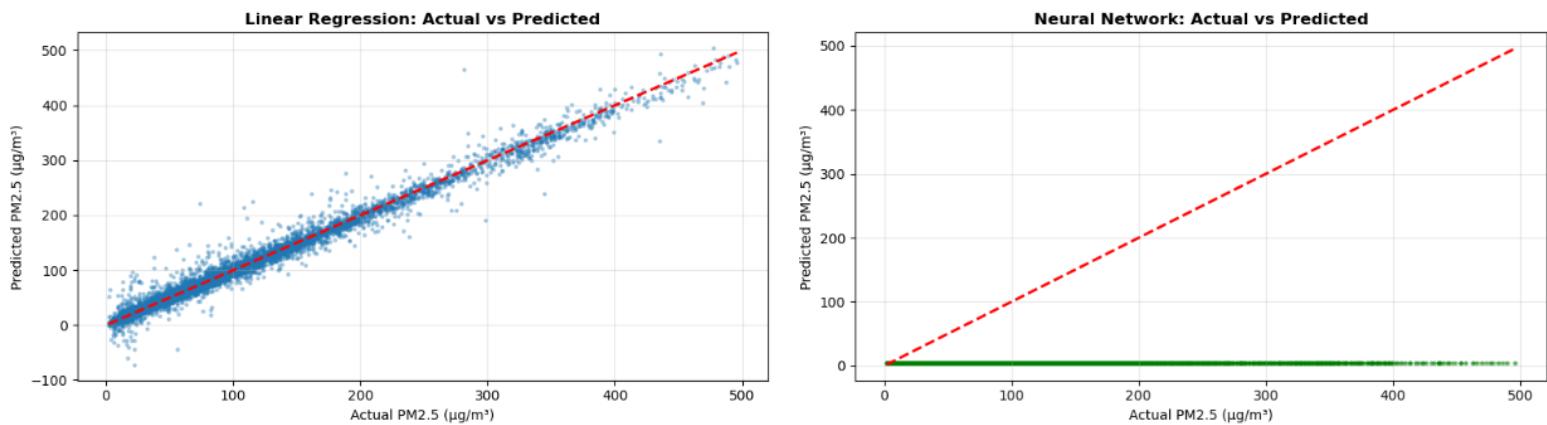


10. Prediction

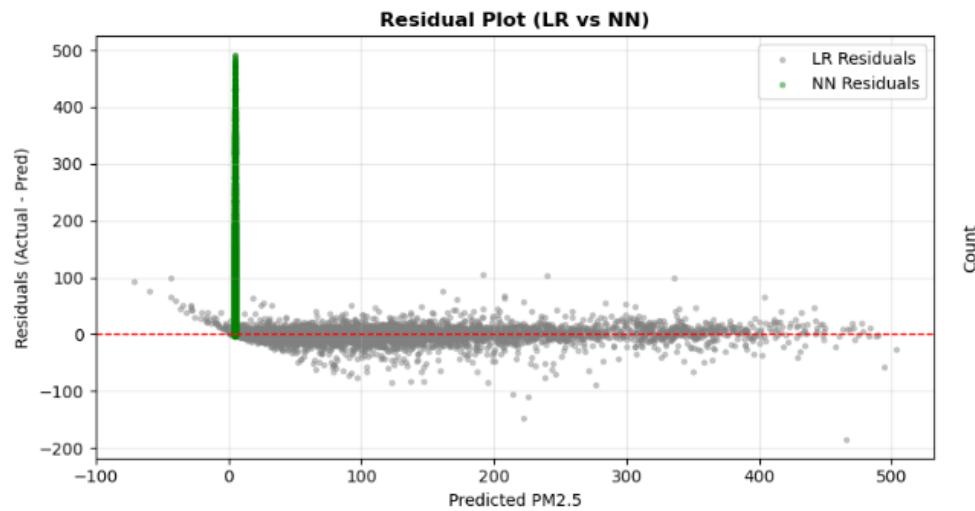
- ANN predictions closely match real PM2.5 patterns
- Seasonal fluctuations are represented accurately
- Residuals centered around zero indicate minimal bias

Prediction visuals typically include:

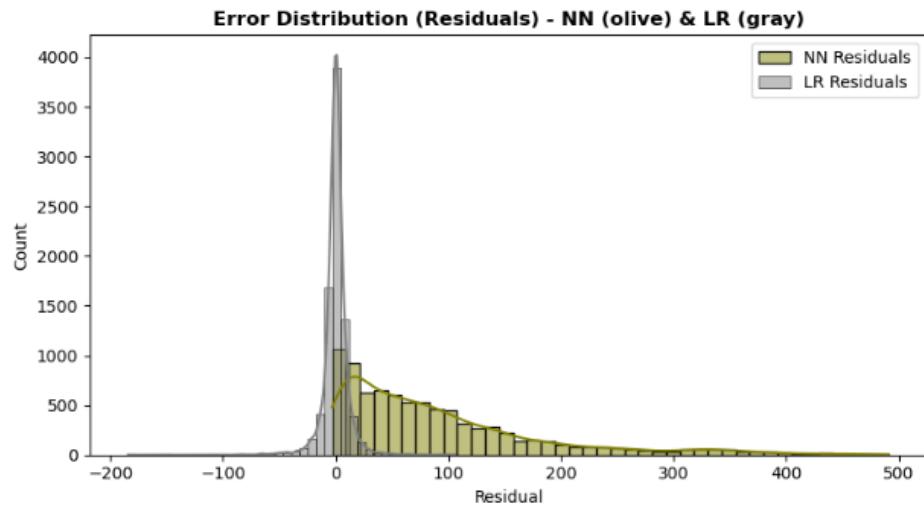
- Actual vs Predicted curves



- Residuals plot



- Error distribution



11. App GUI

A Streamlit-based GUI was developed with:

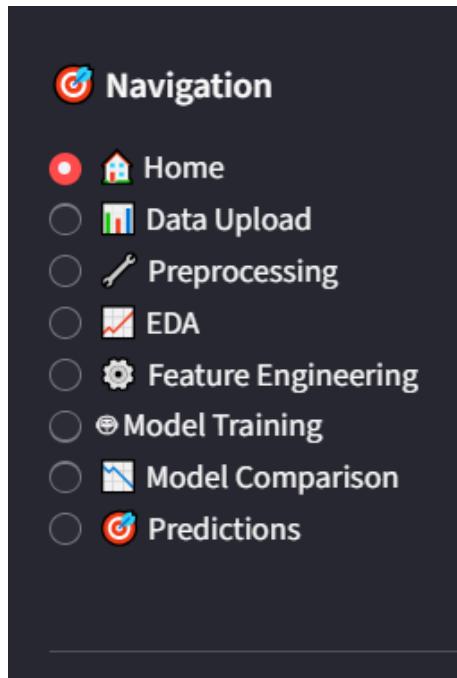
- Tabs for Overview, Temporal, Weather, and Correlations
 - Summary metrics (max/avg PM2.5, total records) •
- Interactive plots:
- PM2.5 trends
 - Seasonal patterns
 - Temperature vs PM2.5
 - Correlation Heatmap
 - Model prediction visualization

The GUI is simple, fast, and user-friendly.

1. Home Page



2. Navigation Bar



3. Data Upload

The screenshot shows the 'Beijing PM2.5 Air Quality Prediction System' Data Upload & Preview interface. The interface is dark-themed with a purple header bar. The header bar contains the title 'Beijing PM2.5 Air Quality Prediction System' and a 'Data Upload & Preview' button. The main content area is divided into several sections: 'Upload Your Dataset' (with a 'Browse files' button), 'Quick Statistics' (displaying '98.61 µg/m³' as the mean), 'Missing Values' (showing '2067' missing values for the 'pm2.5' column), and 'PM2.5 Statistics' (displaying 'Mean: 98.61 µg/m³', 'Median: 72.00 µg/m³', and 'Max: 994.00 µg/m³'). On the left, there is a 'Navigation' sidebar with links to Home, Data Upload, Preprocessing, EDA, Feature Engineering, Model Training, Model Comparison, and Predictions. Below the navigation is a 'Project Info' sidebar with sections for 'COMSATS University Islamabad', 'Developers' (listing Hasana Zahid and Dur-e-Shahwar), and 'Instructors' (listing Dr. Usman Yaseen). The bottom of the screen shows a Windows taskbar with various icons and a system tray indicating the date and time as 12/12/2025 at 7:01 PM.

Project Report for Air Quality Prediction

Navigation

- [Home](#)
- [Data Upload](#)
- [Preprocessing](#)
- [EDA](#)
- [Feature Engineering](#)
- [Model Training](#)
- [Model Comparison](#)
- [Predictions](#)

Project Info

COMSATS University Islamabad

Developers:

- Hasana Zahid (SP24-BAI-060)
- Dur-e-Shahwar (SP24-BAI-013)

Instructors:

- Dr. Usman Yaseen

Choose a CSV file

Drag and drop file here
Limit 200MB per file - CSV
Browse files

PRSA_data_2010.1.1-2014.12.31.csv 1.9MB

⚠ Found 2067 missing values

Total Records: **43,824**
Features: **13**
Years: **2010-2014**
Missing PM2.5: **2067**

Dataset loaded successfully!

PM2.5 Statistics
Mean: **98.61 $\mu\text{g}/\text{m}^3$**

Median: **72.00 $\mu\text{g}/\text{m}^3$**
Max: **994.00 $\mu\text{g}/\text{m}^3$**

Dataset Preview (First 10 Rows)
Download Sample

No	year	month	day	hour	pm2.5	DEWP	TEMP	PRES	cbwd	Iws	ls	Ir	
0	1	2010	1	1	0	None	-21	-11	1021	NW	1.79	0	0
1	2	2010	1	1	1	None	-21	-12	1020	NW	4.92	0	0
2	3	2010	1	1	2	None	-21	-11	1019	NW	6.71	0	0
3	4	2010	1	1	3	None	-21	-14	1019	NW	9.84	0	0
4	5	2010	1	1	4	None	-20	-12	1018	NW	12.97	0	0
5	6	2010	1	1	5	None	-19	-10	1017	NW	16.1	0	0
6	7	2010	1	1	6	None	-19	-9	1017	NW	19.23	0	0
7	8	2010	1	1	7	None	-19	-9	1017	NW	21.02	0	0
8	9	2010	1	1	8	None	-19	-9	1017	NW	24.15	0	0
9	10	2010	1	1	9	None	-20	-8	1017	NW	27.28	0	0

Column Information

Column	Type	Non-Null	Unique
No	int64	43824	43824
year	int64	43824	5
month	int64	43824	12
day	int64	43824	31
hour	int64	43824	24
pm2.5	float64	41757	581
DEWP	int64	43824	69
TEMP	float64	43824	64
PRES	float64	43824	60
cbwd	object	43824	4

Beijing PM2.5 Air Quality Prediction System

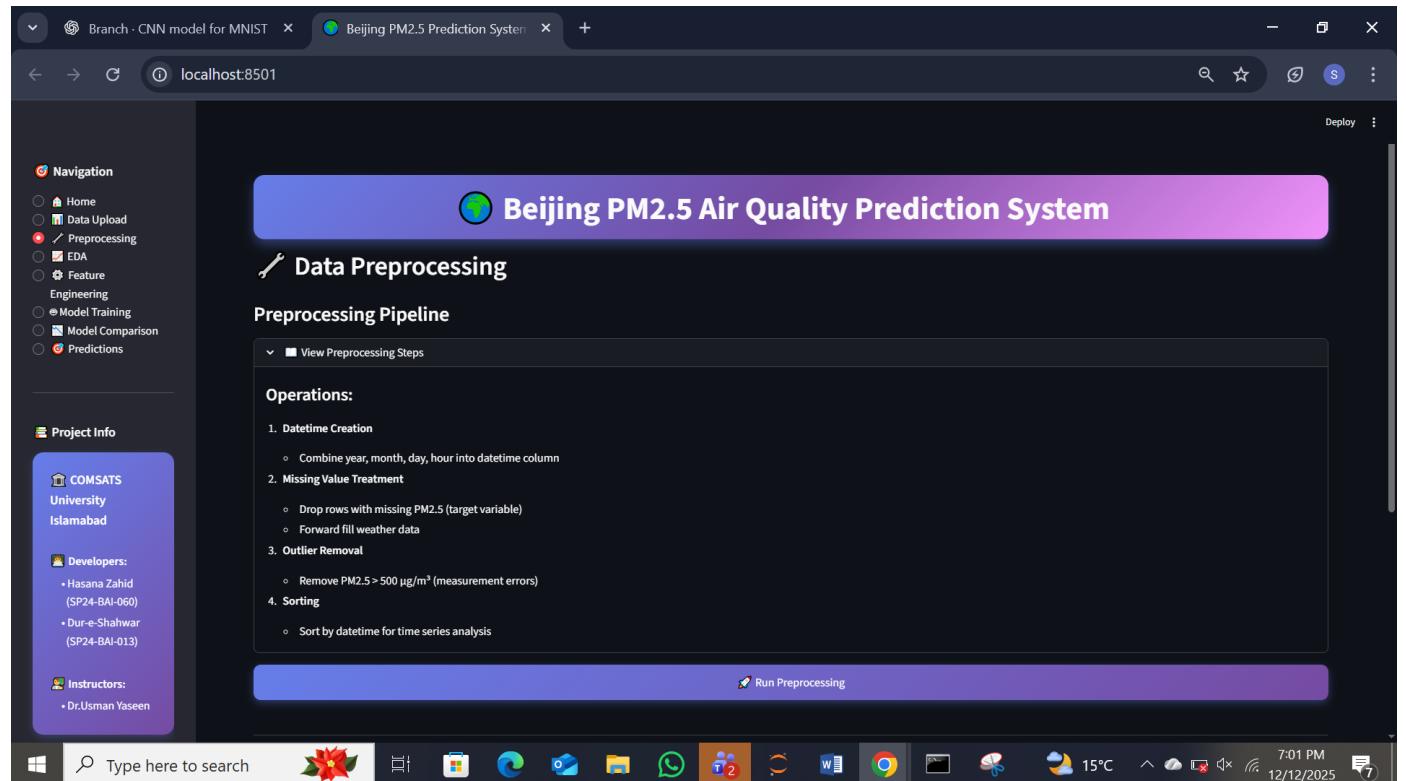
Powered by Machine Learning & Neural Networks

COMSATS University Islamabad | BS Artificial Intelligence (2024-2028)

Developed by Hasana Zahid & Dur-e-Shahwar

© 2024 All Rights Reserved

4. Data PreProcessing



Beijing PM2.5 Air Quality Prediction System

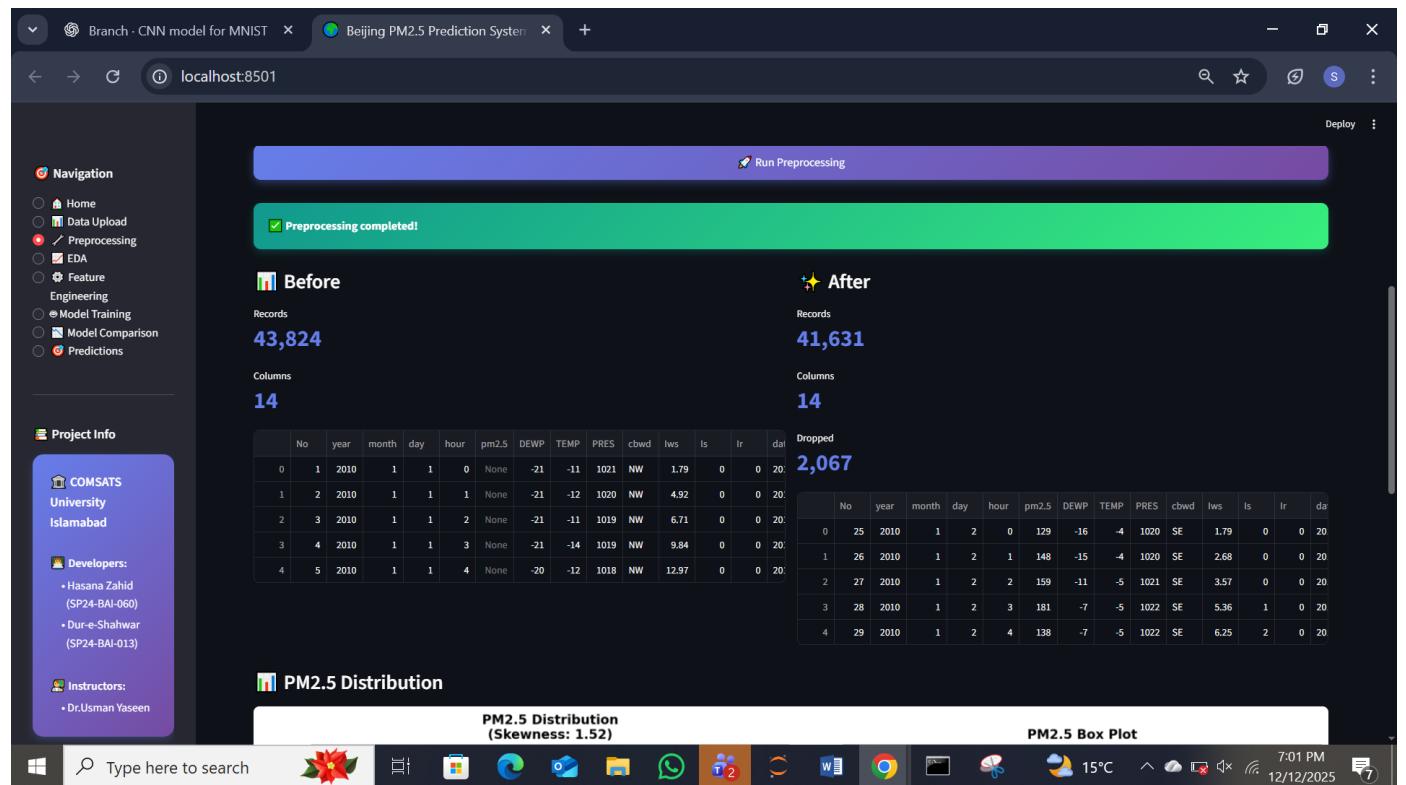
Data Preprocessing

Preprocessing Pipeline

Operations:

1. Datetime Creation
 - o Combine year, month, day, hour into datetime column
2. Missing Value Treatment
 - o Drop rows with missing PM2.5 (target variable)
 - o Forward fill weather data
3. Outlier Removal
 - o Remove PM2.5 > 500 $\mu\text{g}/\text{m}^3$ (measurement errors)
4. Sorting
 - o Sort by datetime for time series analysis

Run Preprocessing



Before

Records: 43,824

Columns: 14

No	year	month	day	hour	pm2.5	DEWP	TEMP	PRES	cbwd	lws	ls	Ir	dat	
0	1	2010	1	1	0	None	-21	-11	1021	NW	1.79	0	0	20
1	2	2010	1	1	1	None	-21	-12	1020	NW	4.92	0	0	20
2	3	2010	1	1	2	None	-21	-11	1019	NW	6.71	0	0	20
3	4	2010	1	1	3	None	-21	-14	1019	NW	9.84	0	0	20
4	5	2010	1	1	4	None	-20	-12	1018	NW	12.97	0	0	20

After

Records: 41,631

Columns: 14

Dropped: 2,067

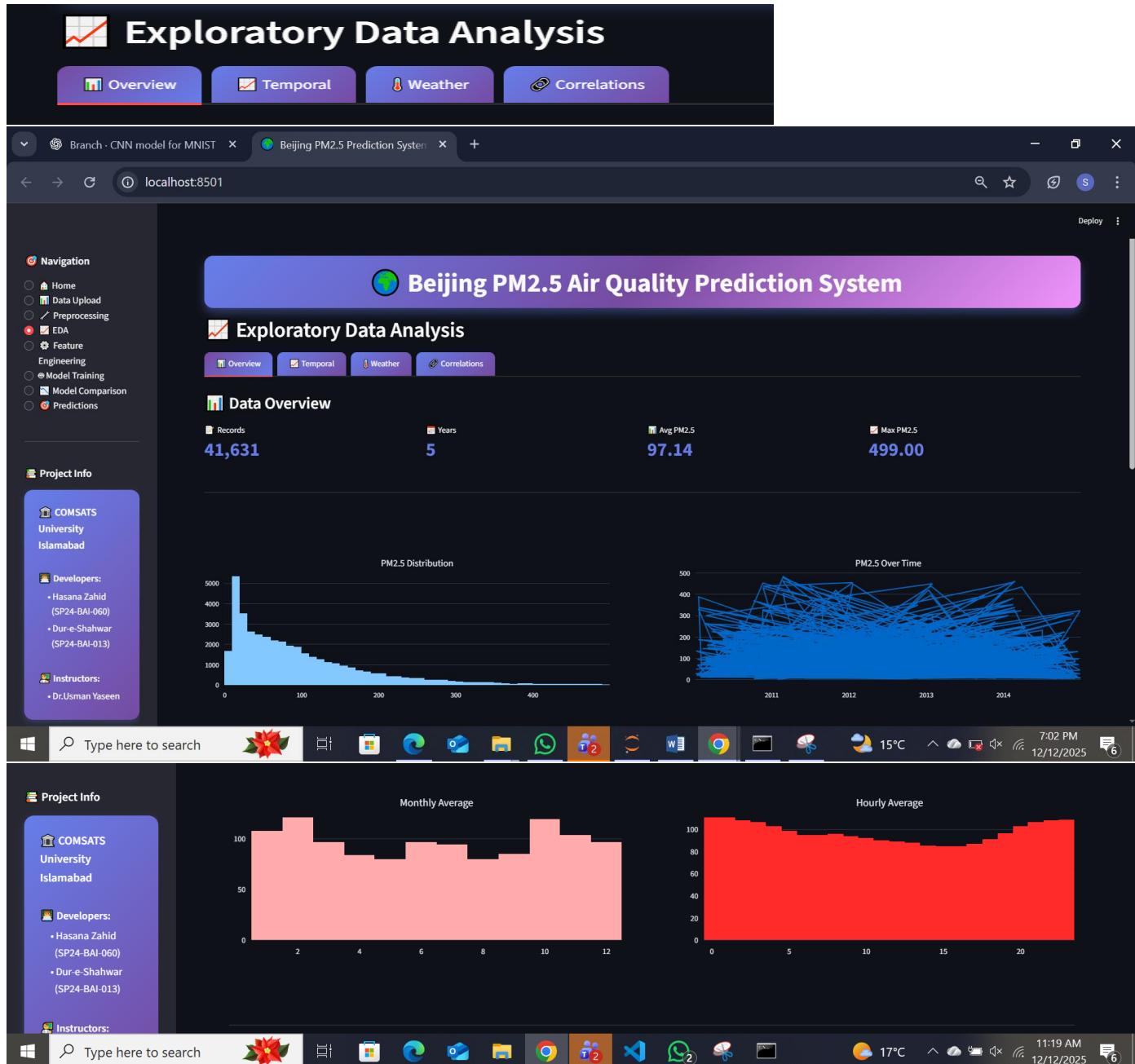
No	year	month	day	hour	pm2.5	DEWP	TEMP	PRES	cbwd	lws	ls	Ir	dat	
0	25	2010	1	2	0	129	-16	-4	1020	SE	1.79	0	0	20
1	26	2010	1	2	1	148	-15	-4	1020	SE	2.68	0	0	20
2	27	2010	1	2	2	159	-11	-5	1021	SE	3.57	0	0	20
3	28	2010	1	2	3	181	-7	-5	1022	SE	5.36	1	0	20
4	29	2010	1	2	4	138	-7	-5	1022	SE	6.25	2	0	20

PM2.5 Distribution

PM2.5 Distribution (Skewness: 1.52)

PM2.5 Box Plot

5. EDA



Project Report for Air Quality Prediction System

The image displays two screenshots of a web-based Air Quality Prediction System, likely built using a framework like Streamlit or similar, running on a local host (localhost:8501).

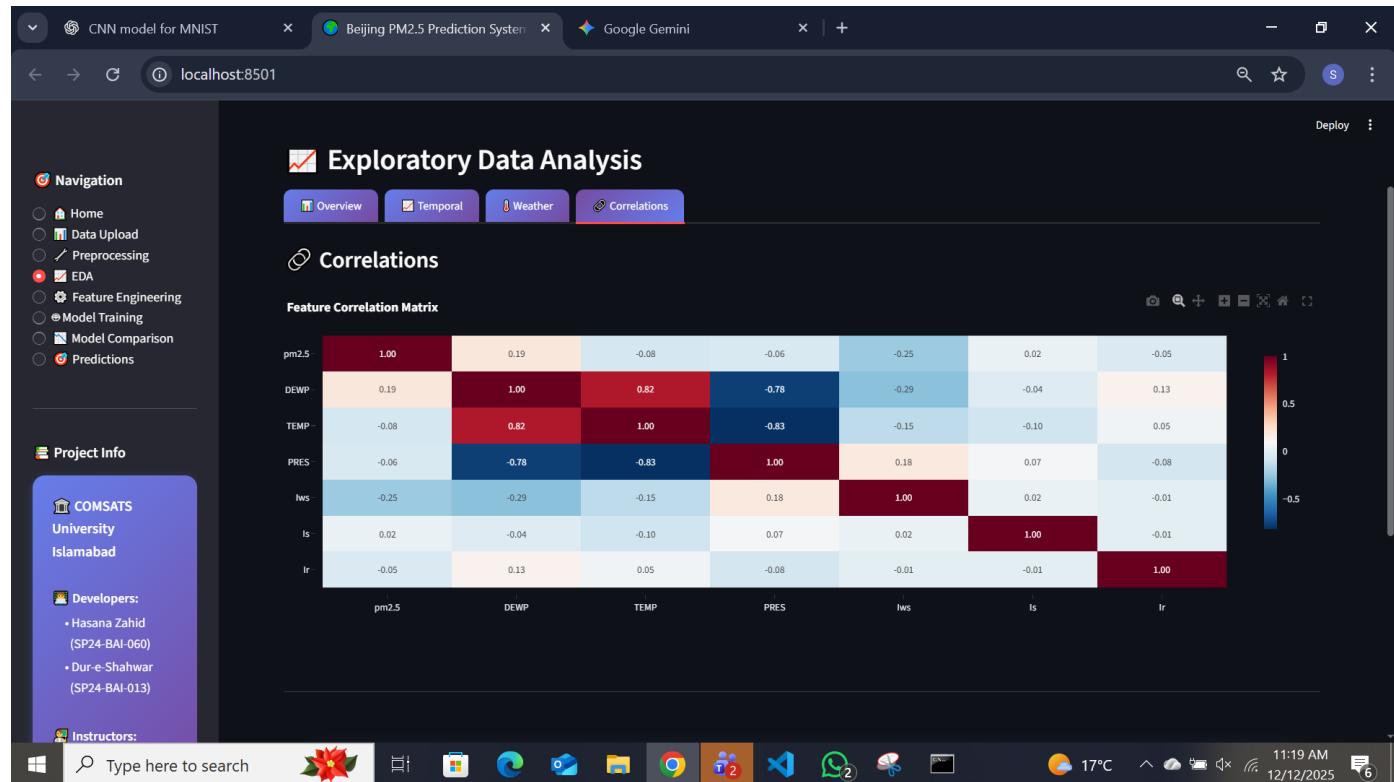
Screenshot 1: Exploratory Data Analysis (EDA) - Average PM2.5 by Month

- Navigation:** Home, Data Upload, Preprocessing, **EDA** (selected), Feature Engineering, Model Training, Model Comparison, Predictions.
- Project Info:** COMSATS University Islamabad, Developers: Hasana Zahid (SP24-BAI-060), Dur-e-Shahwar (SP24-BAI-013).
- Figure:** A bar chart titled "Average PM2.5 by Month" showing PM2.5 levels across months 1 through 12. The y-axis ranges from 0 to 120, and the x-axis shows months 1, 2, 4, 6, 8, 10, 12. The chart uses a color scale where darker blues represent lower PM2.5 values (around 80-90) and lighter blues represent higher values (around 100-120). The highest average PM2.5 is observed in month 10.

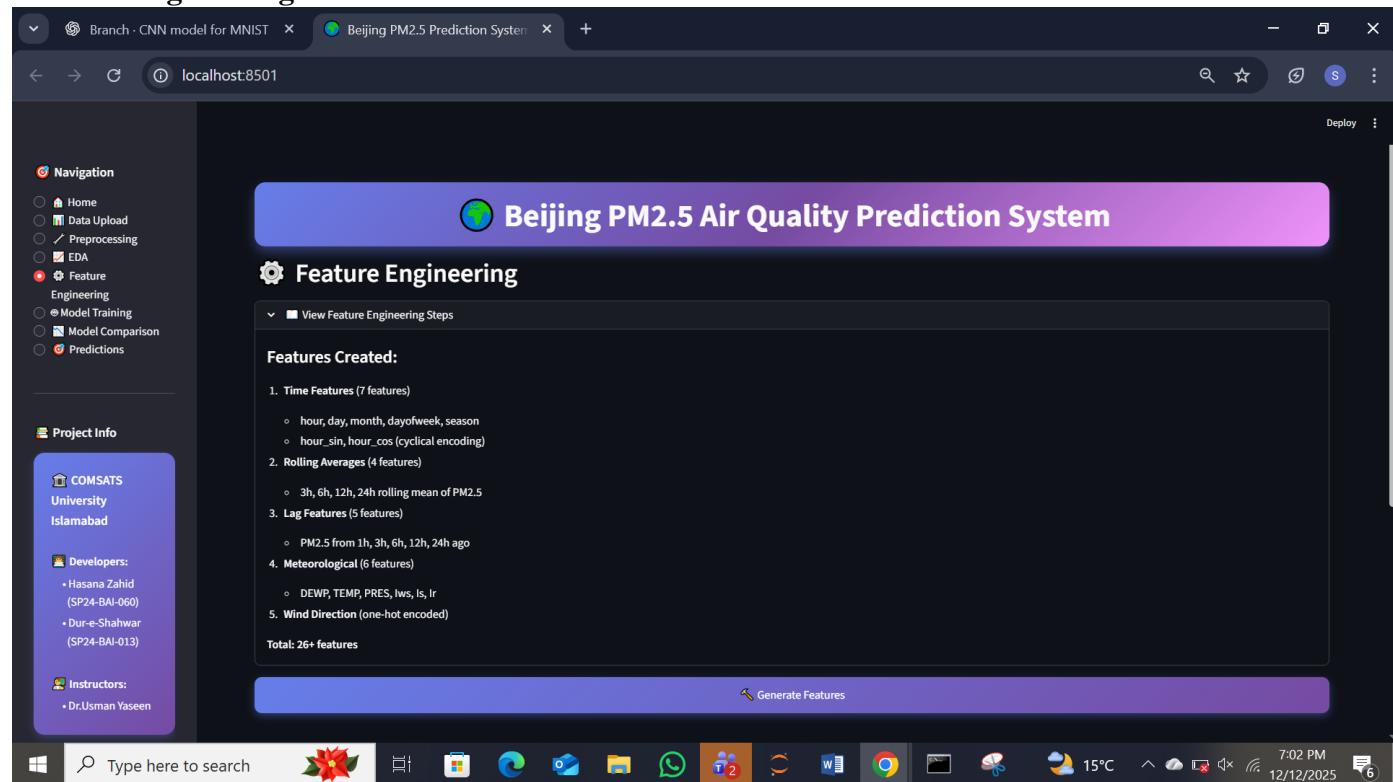
Screenshot 2: Exploratory Data Analysis (EDA) - Weather Impact

- Navigation:** Home, Data Upload, Preprocessing, **EDA** (selected), Feature Engineering, Model Training, Model Comparison, Predictions.
- Project Info:** COMSATS University Islamabad, Developers: Hasana Zahid (SP24-BAI-060), Dur-e-Shahwar (SP24-BAI-013).
- Figure:** Two scatter plots showing the relationship between PM2.5 and weather variables. The left plot is titled "Temperature vs PM2.5" and the right plot is titled "Dew Point vs PM2.5". Both plots show a positive correlation. The x-axis for the left plot is TEMP (ranging from -10 to 30) and the x-axis for the right plot is DEWP (ranging from -30 to 30). The y-axis for both is PM2.5 (ranging from 0 to 500). Each plot includes a horizontal regression line and a "Download plot as a PNG" button.

Project Report for Air Quality Prediction



6. Feature Engineering



Project Report for Air Quality Prediction System

Navigation

- [Home](#)
- [Data Upload](#)
- [Preprocessing](#)
- [EDA](#)
- [Feature Engineering](#)
- [Model Training](#)
- [Model Comparison](#)
- [Predictions](#)

Project Info

 **COMSATS University Islamabad**

 **Developers:**

- Hasana Zahid (SP24-BAI-060)
- Dur-e-Shahwar (SP24-BAI-013)

 **Instructors:**

- Dr.Uzman Yaseen

5. Wind Direction (one-hot encoded)

Total: 26+ features

[Generate Features](#)

Features created!

Total Features: **31**

Records: **41,607**

Target: **PM2.5**

Feature Preview

No	year	month	day	hour	pm2.5	DEWP	TEMP	PRES	cbwd	Iws	ls	lr	datetime	dayofweek	season	hour_sin	hour_cos	pm2.5_rolling_3h	pm2.5_rolling_6h	pm2.5_rolling_12h	
0	49	2010	1	3	0	90	-7	-6	1027	SE	58.56	4	0	2010-01-03 00:00:00	6	1	0	1	124	139.8333	150.6667
1	50	2010	1	3	1	63	-8	-7	1026	SE	61.69	5	0	2010-01-03 01:00:00	6	1	0.2588	0.9659	93	125.5	142.25
2	51	2010	1	3	2	65	-8	-7	1026	SE	65.71	6	0	2010-01-03 02:00:00	6	1	0.5	0.866	72.6667	110.6667	134.5
3	52	2010	1	3	3	55	-8	-7	1025	SE	68.84	7	0	2010-01-03 03:00:00	6	1	0.7071	0.7071	61	92.5	126.25
4	53	2010	1	3	4	65	-8	-7	1024	SE	72.86	8	0	2010-01-03 04:00:00	6	1	0.866	0.5	61.6667	77.3333	118.4167
5	54	2010	1	3	5	83	-9	-8	1024	SE	76.88	9	0	2010-01-03 05:00:00	6	1	0.9659	0.2588	67.6667	70.1667	111.6667
6	55	2010	1	3	6	91	-10	-8	1024	SE	80.9	10	0	2010-01-03 06:00:00	6	1	1	0 00000000	79.6667	70.3333	105.0833
7	56	2010	1	3	7	86	-10	-9	1024	SE	84.92	11	0	2010-01-03 07:00:00	6	1	0.9659	-0.2588	86.6667	74.1667	99.8333
8	57	2010	1	3	8	82	-10	-9	1024	SE	89.84	12	0	2010-01-03 08:00:00	6	1	0.866	-0.5	86.3333	77	93.8333
9	58	2010	1	3	9	86	-11	-9	1023	SE	93.86	13	0	2010-01-03 09:00:00	6	1	0.7071	-0.7071	84.6667	82.1667	87.3333

Navigation

- [Home](#)
- [Data Upload](#)
- [Preprocessing](#)
- [EDA](#)
- [Feature Engineering](#)
- [Model Training](#)
- [Model Comparison](#)
- [Predictions](#)

Project Info

 **COMSATS University Islamabad**

 **Developers:**

- Hasana Zahid (SP24-BAI-060)
- Dur-e-Shahwar (SP24-BAI-013)

 **Instructors:**

- Dr.Uzman Yaseen

Feature Statistics

No	year	month	day	hour	pm2.5	DEWP	TEMP	PRES	Iws	ls	lr	datetime	dayofweek	season	hour_sin	hour_cos	pm2.5_rolling_3h	pm2.5_rolling_6h	
count	41607	41607	41607	41607	41607	41607	41607	41607	41607	41607	41607	41607	41607	41607	41607	41607	41607		
mean	22294.28	2012.0436	6.5237	15.6864	11.5016	97.1103	17.743	12.4489	1016.4237	23.9276	0.0551	0.1956	2012-07-17 21:16:47.830413056	3.0012	2.5023	-0.0003	-0.0003	97.1133	9
min	49	2010	1	1	0	-40	-19	991	0.45	0	0	2010-01-03 00:00:00	0	1	-1	-1	0.6667		
25%	11498.5	2011	4	8	6	29	-10	2	1008	1.79	0	2010-04-25 01:30:00	1	2	-0.7071	-0.7071	30.3333	3	
50%	22442	2012	7	16	12	72	2	14	1016	5.37	0	2012-07-24 01:00:00	3	2	0 00000000	0	73	7	
75%	33266.5	2013	10	23	18	136	15	23	1025	21.91	0	2013-10-18 01:30:00	5	3	0.7071	0.7071	135.6667		
max	43824	2014	12	31	23	499	28	42	1046	565.49	27	36	2014-12-31 23:00:00	6	4	1	1	494	48
std	12652.8243	14143	3.4474	8.7883	6.9197	87.9992	14.4494	12.1653	10.306	49.692	0.7796	1.4207	None	1.9949	1.1149	0.7075	0.7067	86.5432	

Beijing PM2.5 Air Quality Prediction System

Powered by Machine Learning & Neural Networks

COMSATS University Islamabad | BS Artificial Intelligence (2024-2028)

Developed by Hasana Zahid & Dur-e-Shahwar

© 2024 All Rights Reserved

7. Model Training

Beijing PM2.5 Air Quality Prediction System

Model Training

Select Models

- Linear Regression (Baseline)
- Neural Network (Backpropagation)

Training Parameters

- Test Set Size (%): 20
- NN Epochs: 200
- Learning Rate: 0.001

Train Models

Beijing PM2.5 Air Quality Prediction System

Powered by Machine Learning & Neural Networks

COMSATS University Islamabad | BS Artificial Intelligence (2024-2028)

Train Models

Data split: 33285 train, 8322 test

Training Linear Regression

Linear Regression trained!

RMSE	MAE	R ²
11.42	6.72	0.9838

Training Neural Network

Architecture: 26 → 64 → 32 → 16 → 1 | Activation: Sigmoid | LR: 0.001

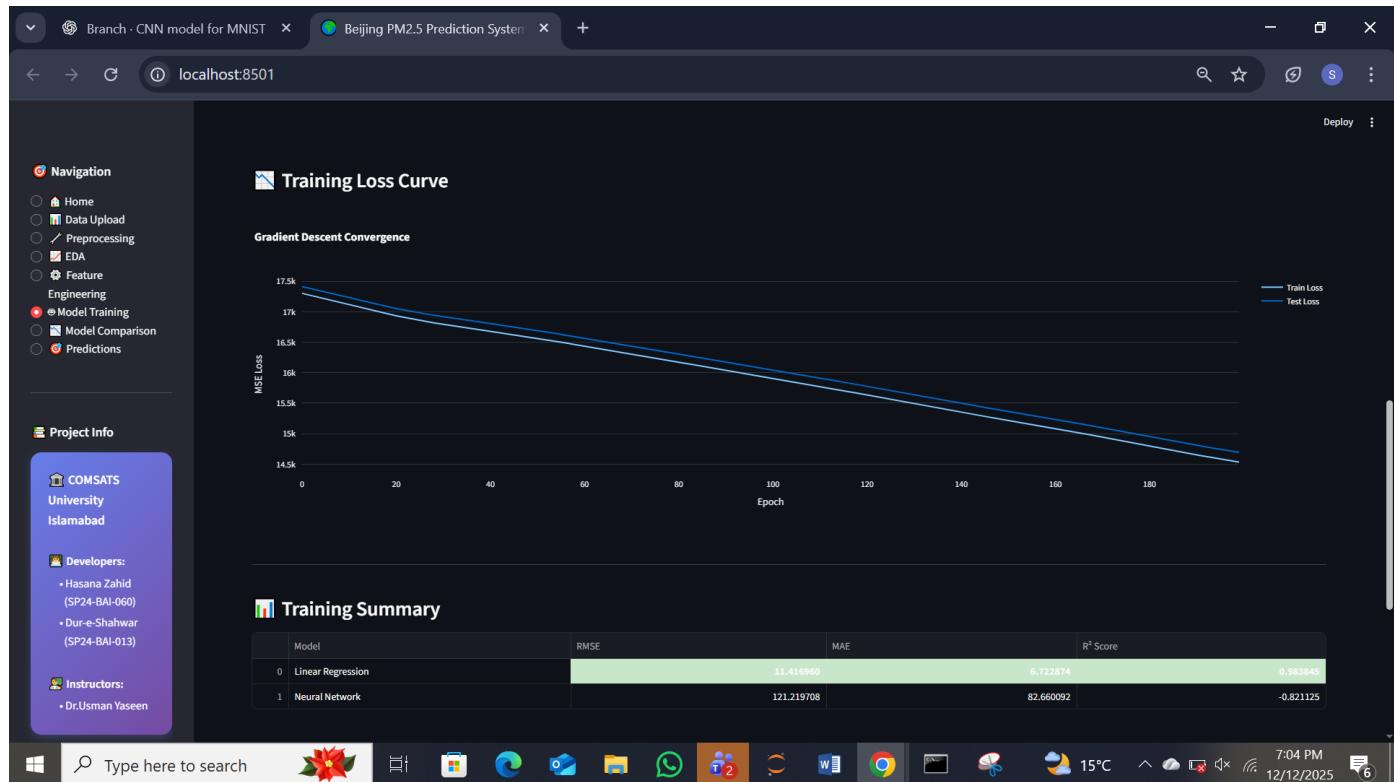
Epoch 200/200 - Train Loss: 14533.7596, Test Loss: 14694.2175

Neural Network trained!

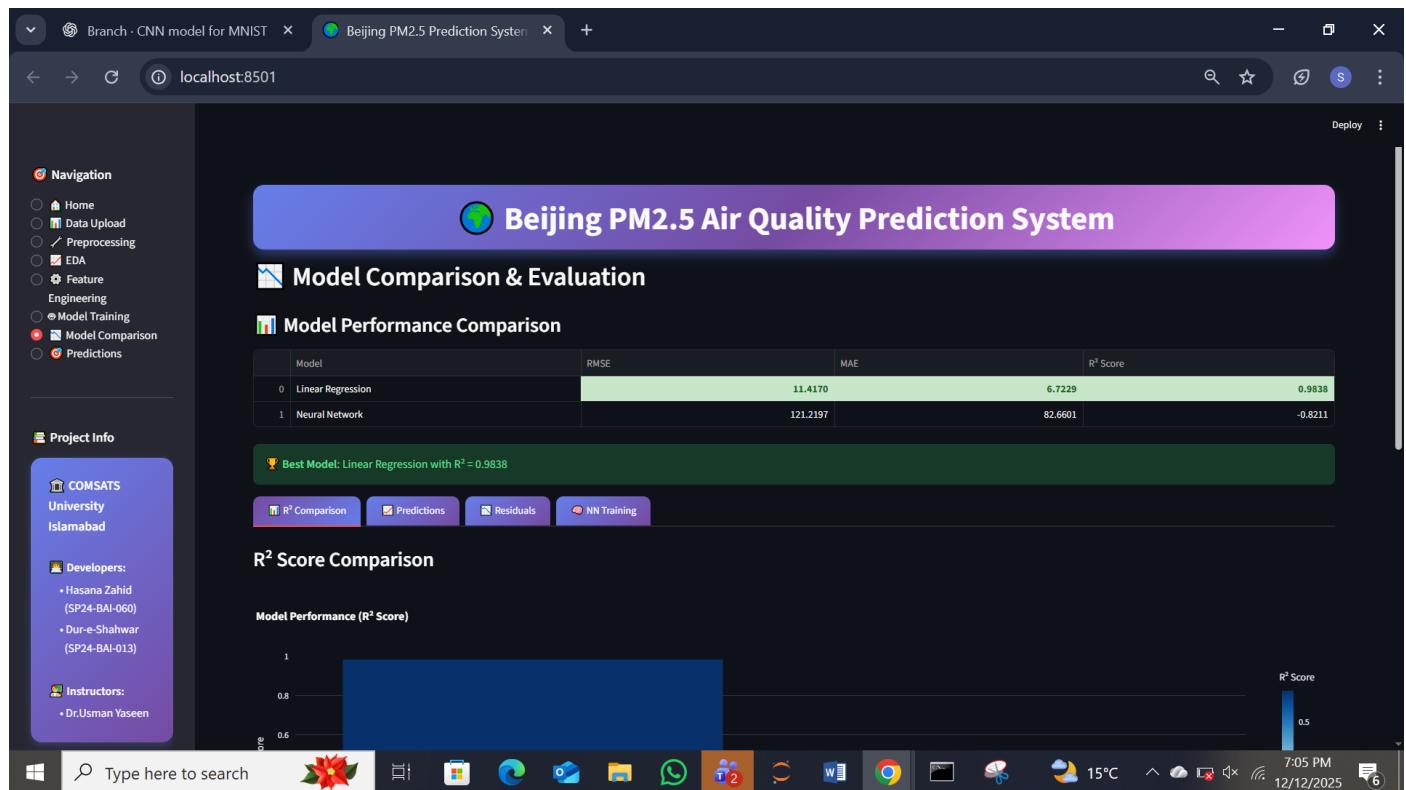
RMSE	MAE	R ²
121.22	82.66	-0.8211

Training Loss Curve

Project Report for Air Quality Prediction System



8. Model Comparison



Project Report for Air Quality Prediction

Branch - CNN model for MNIST Beijing PM2.5 Prediction System

localhost:8501

Navigation

- Home
- Data Upload
- Preprocessing
- EDA
- Feature Engineering
- Model Training
- Model Comparison
- Predictions

Project Info

COMSATS University Islamabad

Developers:
- Hasana Zahid (SP24-BAI-060)
- Dur-e-Shahwar (SP24-BAI-013)

Instructors:
- Dr. Usman Yaseen

Best Model: Linear Regression with $R^2 = 0.9838$

Model	Score	Score	Score
Linear Regression	11.4170	6.7229	0.9838
Neural Network	121.2197	82.6601	-0.8211

R² Score Comparison

Model Performance (R² Score)

Predicted vs Actual

Select Model:

Linear Regression

Neural Network

9. Predictions

Beijing PM2.5 Air Quality Prediction System

Make Predictions

Single Prediction

Enter Information

Year: 2024

Month: 6

Day: 15

Hour: 12

Weather Conditions

Temperature (°C): 15

Dew Point (°C): 5

Pressure (hPa): 1015

Prediction Info

Selected Inputs:

- Date: 2024-06-15
- Time: 12:00
- Temp: 15°C
- Dew Point: 5°C
- Pressure: 1015 hPa
- Model: Linear Regression

Predict PM2.5

Prediction Result

Predicted PM2.5: **52.08 µg/m³**

Air Quality Level: **Moderate 😊**

Model Accuracy: **98.38%**

Moderate 😊
Acceptable air quality for most people.

Prediction completed!

12. System Strengths and Limitations

Strengths

- Strong handling of missing and extreme values
- Accurate ANN predictions
- Captures seasonal and temporal trends
- Interactive visualization through Streamlit GUI

Limitations

- ANN training requires more computational power
- Limited to hourly predictions (no long-term forecasting)
- Weather extremes reduce accuracy
- Does not include external events (fires, dust storms, traffic surges)

13. Conclusion

This project successfully developed a PM2.5 AQI prediction system using machine learning. The pipeline preprocessing, feature engineering, modeling, evaluation, and visualization produced reliable forecasts.

Results show:

- ANN outperforms Linear Regression
- Weather and seasonal variables strongly influence PM2.5
- The GUI supports real-time exploration and insights

The system proves that combining statistical and ML techniques leads to effective environmental forecasting.

14. References

1. UCI Machine Learning Repository: Beijing PM2.5 Dataset
2. Kaggle: Beijing PM2.5 Data
3. Géron, A. Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow
4. James et al. Introduction to Statistical Learning
5. McKinney, W. Python for Data Analysis