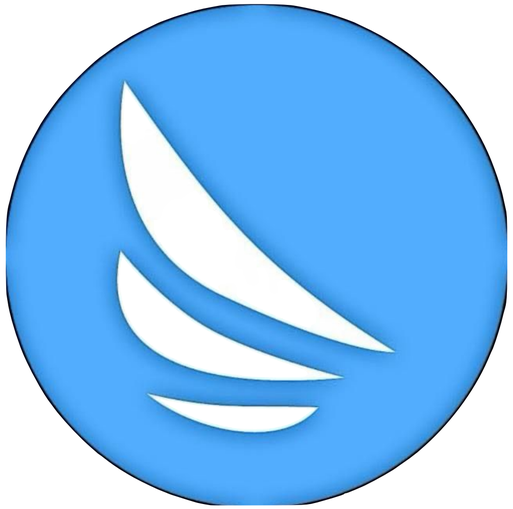
Major Project



Presented by : Om Nanda
(MST03-0110)
Mentor : Ms. Urooj Khan
Trainer
Meta Scifor Technologies

Project Details

- Title : Air Quality Forecasting &Analysis dashboard
- Intern : Om Nanda
- Mentor : Urooj Khan
- Organization : Meta Scifor Technologies

Problem Statement

- Air pollution is a critical issue affecting health and the environment.
- WHO reports millions of premature deaths due to poor air quality.
- Forecasting pollution helps governments, citizens, and healthcare systems.
- Visualizing trends makes the data actionable.
- 


Project Objectives

- Explore and visualize air pollution data (PM2.5, PM10, NO_2, etc.)
- Analyze pollution trends by city, date, and season
- Forecast pollutant levels using ML models
- Create an interactive dashboard for exploration
- Deploy the app for public use via Streamlit

Dataset Description

- Source: Kaggle (US Pollution Data) and OpenAQ (for live data)
- Total Records: ~70,000+
- Key Features:
- Date, City, State
- PM2.5, PM10, CO, SO_2, NO_2, O_3 levels

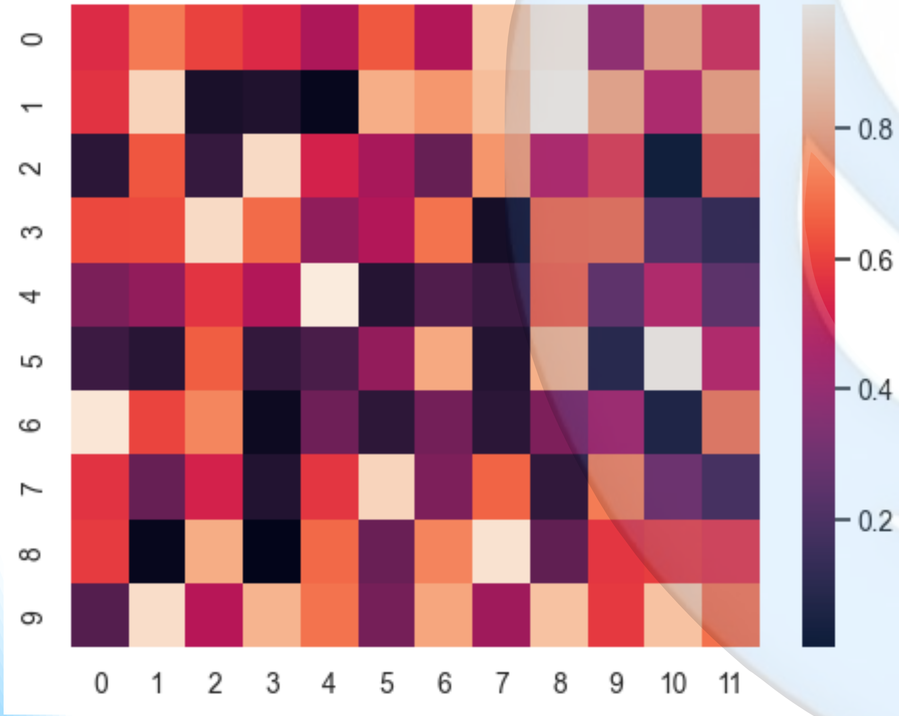
	PM2.5	PM10	NO_2	CO	O_3	SO_2
count	290621	227747	292359	268197	290589	292462
mean	58.78	88.05	45.79	0.96	55.69	8.98
std	66.11	89.29	32.06	1.00	53.82	11.70
min	2.0	5.0	1.0	0.1	1.0	1.0
25%	16.0	37.0	20.0	0.4	2912.0	2.0
50%	39.0	70.0	39.0	0.7	45.0	5.0
75%	77.0	113.0	66.0	1.2	79.0	11.0
max	1004	3000	300	15	504	307

Data Preprocessing

- Converted date fields into datetime object
- Extracted features: year, month, day, weekday
- Dropped/replaced missing values
- Normalized pollutant levels for ML compatibility
- Created lag features to improve forecasting

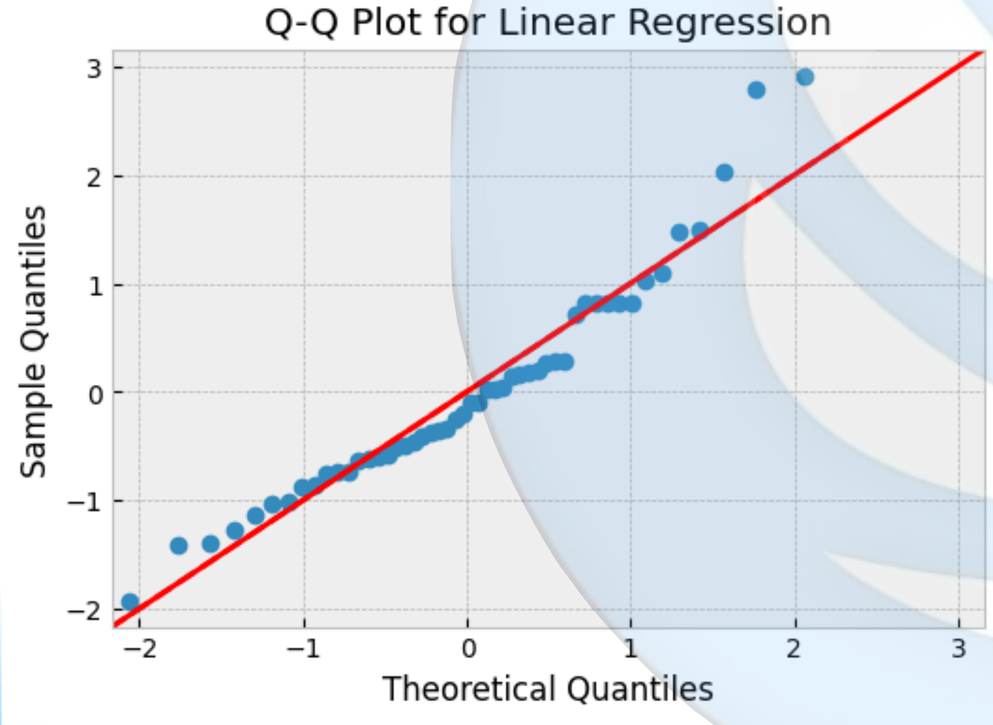
Exploratory Data Analysis

- Line plots show PM2.5 trends over time
- Box plots reveal weekday & monthly variations
- Heatmaps highlight pollutant correlations
- Bar charts compare average pollution across cities
- Key insight: pollution spikes during winter months



Machine Learning Models

- Linear Regression – simple trend modeling
- Random Forest Regressor – handles non-linearity, low error
- XGBoost Regressor – best accuracy, fast training
- Features used: city, month, weekday, previous PM2.5 values



Model Evaluation

- Metrics used:
- R^2 Score (explains variance)
- RMSE (error magnitude)
- MAE (average prediction error)
- Results:
- XGBoost: 93% accuracy, lowest RMSE
- Random Forest: 91% accuracy
- Linear Regression: lower accuracy (~78%)
-

Model	Train Score	Test Score	RMSE	MAE
Linear Regression	0.81	0.78	5.2	3.9
Random Forest	0.95	0.91	2.1	1.8
XGBoost	0.97	0.93	1.6	1.2

Dashboard&Deployment

- Built using Streamlit – fast Python framework
- Features:
- Sidebar filters: City, Pollutant, Date range
- Interactive charts and trend lines
- Model-based predictions for PM2.5
- Summary statistics and AQI-level display
- Tools used:
- GitHub (code repo)
- Streamlit Cloud / Hugging Face Spaces
- App files: app.py, model.pkl, requirements.txt
- Public URL generated for dashboard access
- Can be shared with users for real-time insights

Real-Time / Live Data Sc

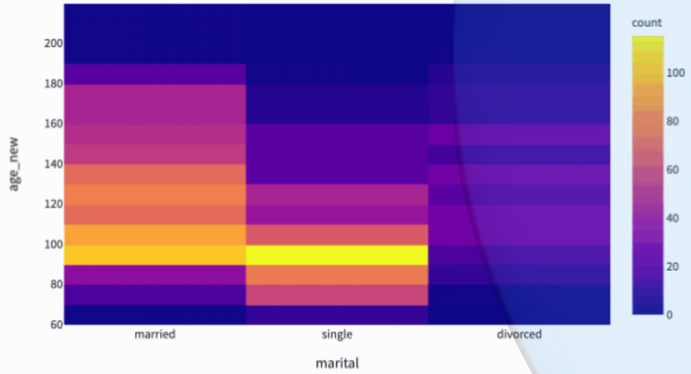
Real-Time / Live Data Science Dashboard

Select the Job

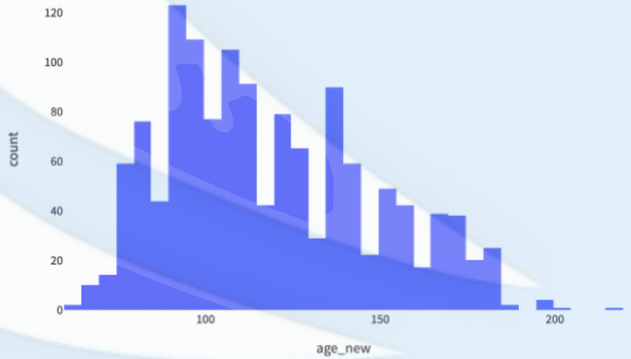
services

Age	Married Count	A/C Balance $
118	701	$ 3587.6
↑ 108	↑ 691	↓ -500

First Chart



Second Chart

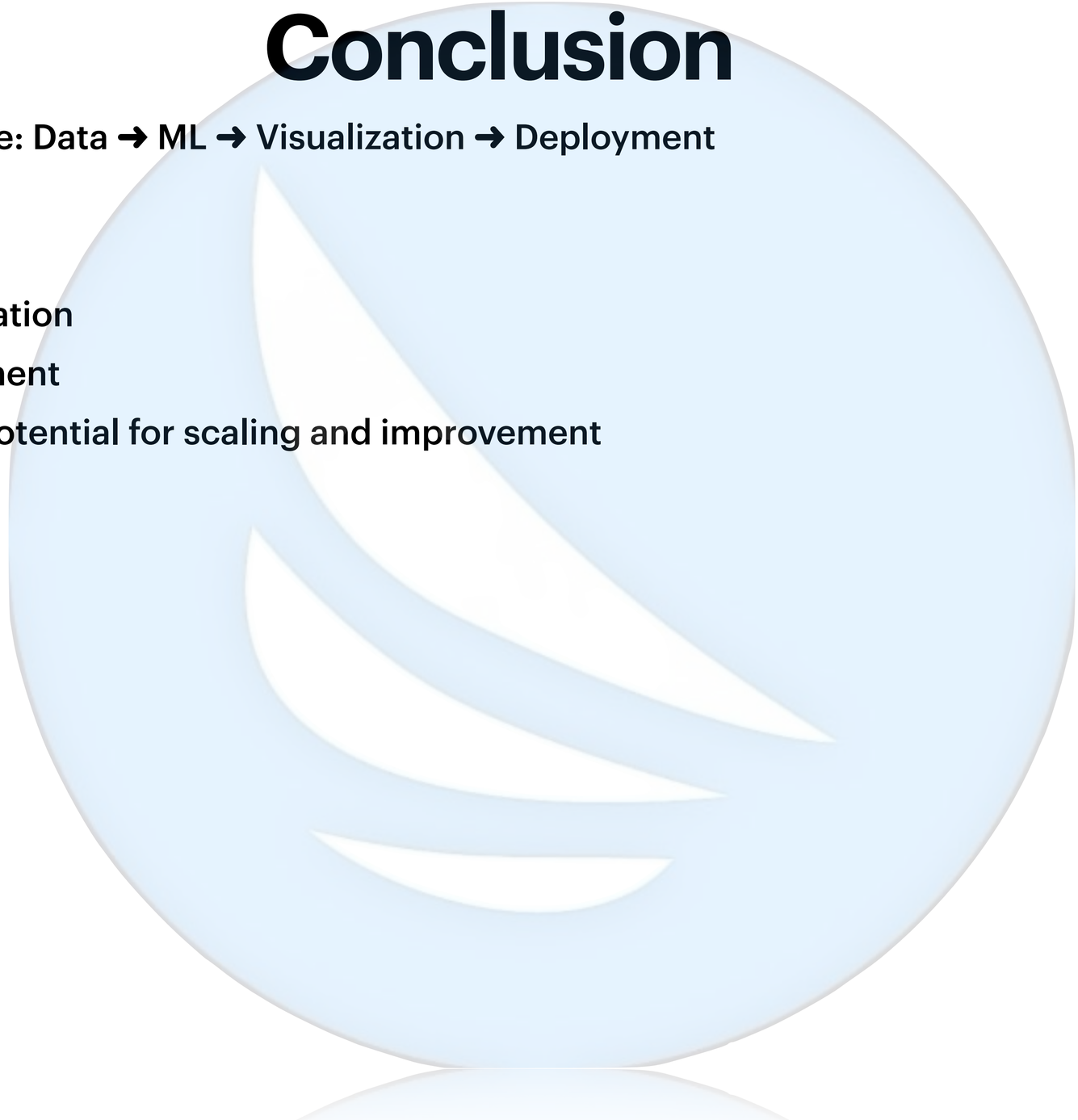


Detailed Data View

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	deposit	age_new	balance_new
0	59	admin.	married	secondary	no	2343	yes	no	unknown	5	may	1042	1	-1	0	unknown	yes	177	702
1	56	admin.	married	secondary	no	45	no	no	unknown	5	may	1467	1	-1	0	unknown	yes	168	13
4	54	admin.	married	tertiary	no	184	no	no	unknown	5	may	673	2	-1	0	unknown	yes	162	55
10	38	admin.	single	secondary	no	100	yes	no	unknown	7	may	786	1	-1	0	unknown	yes	114	30

Conclusion

- Built a complete pipeline: Data ➜ ML ➜ Visualization ➜ Deployment
- Gained experience in:
- Feature engineering
- Model tuning and evaluation
- Streamlit web development
- Project has real-world potential for scaling and improvement



References

- Kaggle: https://www.kaggle.com/datasets/sogun3/uspollution
- OpenAQ: https://openaq.org/
- Libraries: Scikit-learn, XGBoost, Pandas, Plotly, Streamlit
- Docs: seaborn.pydata.org | streamlit.io | xgboost.readthedocs.io

Thank you .

Meta
Scifor Technologies

Script. Sculpt. Socialize