Air Quality & Health Models - Final Summary

✅ All 4 Models Complete & Documented

Model 1: AQI Forecasting ⭐ ⭐ ⭐ ⭐

Purpose: Predict AQI 7 days ahead
Performance: R^2=0.52, Error ±16 AQI (±19%)
Best Model: Lasso Regression
Use Case: Weekly planning, trend detection
Status: ✅ Complete

Model 2: Severe Day Prediction ⭐ ⭐ ⭐ ⭐ ⭐

Purpose: Alert for tomorrow's severe pollution day (AQI ≥300)
Performance: Recall=98.7%, Precision=51.7%, F1=0.678
Best Model: RandomForest (threshold=0.20)
Trade-off: Catches 99% of severe days, 48% false alarms
Use Case: Emergency public health alerts
Status: ✅ Complete

Confusion Matrix:

- Caught: 77/78 severe days (missed only 1!)
- False alarms: 72 out of 2,409 days (~3%)

Model 3: Disease Burden Estimation ⭐ ⭐ ⭐ ⭐

Purpose: India state-level disease rates (per 100k)
Performance: R^2=0.81, Errors ±38-60 per 100k
Best Model: ElasticNet Strong (α=1.0)
Overfitting: Fixed! Gap=-0.07 (was 0.0)
Use Case: State comparisons, policy scenarios
Status: ✅ Complete (Improved from R^2=1.0 overfitted)

Key Improvement:

> Original: R^2=1.00, Error ±2 (fake perfection)
> Improved: R^2=0.81, Error ±38-60 (realistic)

Targets:

- Cardiovascular: ±38 per 100k
- Respiratory: ±28 per 100k
- All Diseases: ±60 per 100k

Model 4: Pollutant Synergy ⭐ ⭐ ⭐

Purpose: Global disease-pollution relationships

Performance: R^2=0.50, Error ±10,000-18,600 deaths

Best Model: RandomForest

Coverage: 156 countries, 2015-2019 (17,767 rows)

Use Case: Global comparisons, pollutant interactions

Status: ✅ Complete (Improved from R^2=0.10)

Note: Predictions in log-space, use `np.expm1()` to convert back

📊 Performance Ranking

Rank	Model	R^2/Metric	Error	Reliability
🥇	Model 3	0.81	±38-60	Excellent (India)
🥈	Model 2	Recall 0.99	48% FA	Excellent (alerts)
🥉	Model 1	0.52	±19%	Good (forecasting)
4️⃣	Model 4	0.50	±75%	Fair (global)

Note: Rankings by use case appropriateness, not just R^2

🎯 Quick Decision Tree

What do you need?

```
├─ 🔮 Forecast AQI for next week?
│  └─ Use Model 1 (±16 AQI, 7 days ahead)
│
├─ 🚨 Alert for tomorrow's severe day?
│  └─ Use Model 2 (99% catch rate)
│
├─ 📊 Compare India states by disease burden?
│  └─ Use Model 3 (R²=0.81, ±38-60)
│
└─ 🌍 Global pollution-health analysis?
   └─ Use Model 4 (156 countries, ±75%)
```

📁 Files Structure

Model 1 - AQI Forecasting

```
model1_best_Lasso_R2-0.523.pkl
model1_usage_improved.md
model1_predictions.csv
```

model1_comparison.csv
model1_actual_vs_predicted.png
model1_residuals.png
model1_time_series.png

Model 2 - Severe Day Prediction

model2_best_RandomForest_Recall-0.987_F1-0.678.pkl
model2_usage_improved.md
model2_threshold.txt (0.20)
model2_predictions.csv
model2_comparison.csv
model2_confusion_matrix.png
model2_roc_curve.png
model2_pr_curve.png
model2_classification_report.txt

Model 3 - Disease Burden (Improved)

improved_best_Cardiovascular_per_100k_ElasticNet_Strong_R2-0.805_gap--0.067.pkl
improved_best_Respiratory_per_100k_ElasticNet_Strong_R2-0.803_gap--0.074.pkl
improved_best_All_Key_Diseases_per_100k_ElasticNet_Strong_R2-0.814_gap--0.070.pkl
model3_usage_improved.md
improved_*_predictions.csv (3 files)
improved_*_comparison.csv (3 files)
improved_*_feature_importance.csv (3 files)
improved_*_actual_vs_pred.png (3 files)
comprehensive_model_comparison.png
model3_summary_improved.csv

Model 4 - Pollutant Synergy (Improved)

model4_best_Cardiovascular_deaths_per_100k_RandomForest_R2-0.480.pkl
model4_best_Respiratory_deaths_per_100k_RandomForest_R2-0.504.pkl
model4_best_Combined_disease_risk_score_RandomForest_R2-0.504.pkl
model4_usage_improved.md
model4_*_predictions.csv (3 files)
model4_*_comparison.csv (3 files)
model4_*_actual_vs_pred.png (3 files)
model4_pollutant_synergy.csv

Master Guides

ALL_MODELS_GUIDE.md (comprehensive comparison)

🔑 Key Insights

Model 1 (AQI Forecasting)

- ✅ Good for planning (7-day horizon)
- ⚠️ Under-predicts extremes (AQI > 400)
- 💡 Best use: Trend detection, not exact values

Model 2 (Severe Day Alert)

- ✅ Excellent recall (catches 99% of severe days)
- ⚠️ High false alarm rate (48%)
- 💡 Trade-off favors safety (better safe than sorry)
- 🎯 Perfect for emergency response

Model 3 (Disease Burden)

- ✅ Excellent R^2 for small dataset (77 obs)
- ✅ Honest error estimates (±38-60)
- ⚠️ India-specific only
- 💡 Use for state comparisons, not absolutes

Model 4 (Pollutant Synergy)

- ✅ Broad coverage (156 countries)
- ✅ 5x improvement over original (R^2 0.10→0.50)
- ⚠️ High error (±75%)
- 💡 Use for exploratory analysis
- 🔧 Remember: Predictions in log-space!

⚠️ Critical Reminders

Model 2 Threshold

```python
# Default sklearn threshold = 0.50
# Optimized threshold = 0.20 (for max recall)
proba = model.predict_proba(X)[:, 1]
predictions = (proba >= 0.20).astype(int)  # Use 0.20, not 0.50!
```

Model 3 Overfitting Fix

```
Problem: R²=1.00 with 77 observations = memorization
```

Solution: Removed State encoding, strong regularization
Result: R^2=0.81 (realistic), Gap=-0.07 (healthy)

Model 4 Log Transform

```python
# Model predicts in log-space
log_preds = model.predict(X)
actual_preds = np.expm1(log_preds)  # Convert back!
```

📈 Comparison: Model 3 vs Model 4

Aspect	Model 3	Model 4
Region	India states	156 countries
R^2	0.81	0.50
Error	±38-60 per 100k	±10k-18k deaths
Dataset	77 rows	17,767 rows
Targets	Per 100k rates	Absolute deaths
Scale	Normalized	Log-transformed
Use	India analysis	Global trends

When to use each:

- India-focused → Model 3 (better accuracy)
- Global scope → Model 4 (broader coverage)

🎓 Lessons Learned

1. Small Datasets Require Simplicity

- 77 observations → max 8-10 features
- Strong regularization essential
- R^2=0.81 is excellent (not 1.0!)

2. Trade-offs Matter

- Model 2: Recall > Precision (public health)
- False alarms acceptable to catch severe days

3. Honest Uncertainty

- Report ± error ranges

- Lower R^2 often means better generalization
- Perfect scores = red flag for overfitting

4. Context-Specific Metrics

- Classification: Recall/Precision/F1
- Regression: R^2, RMSE, Gap
- Always check overfitting gap!

🚀 Next Steps

For Production Use:

1. **Validate** on independent test data
2. **Monitor** performance over time
3. **Retrain** quarterly with new data
4. **Update** thresholds as needed

For Research:

1. Test on other regions (Model 1, 2)
2. Expand feature sets cautiously
3. Compare with domain models
4. Publish uncertainty ranges

For Policy:

1. Use Model 3 for state comparisons
2. Use Model 2 for daily alerts
3. Test intervention scenarios
4. Report with confidence intervals

📞 Support

Documentation:

- Individual model guides: `model*_usage_improved.md`
- Master comparison: `ALL_MODELS_GUIDE.md`
- This summary: `FINAL_SUMMARY.md`

Data Prep Scripts:

- Model 1: `model1_data_prep.py`
- Model 2: `model2_data_prep.py`
- Model 3: `model3_data_prep.py`
- Model 4: `model4_data_prep.py`

Key Decisions:

- All models use train/test split (70/30 or 80/20)
- Model 2 optimizes for recall (not F1)
- Model 3/4 prefer simple models (avoid overfitting)
- Model 4 uses log-transform for stability

Version: Final (all 4 models complete)
Date: 2025
Dataset: Indian air quality 2015-2020 + Global health data
Total Models: 10 files (M1=1, M2=1, M3=3, M4=3, all +improved versions)