1. Introduction

This project predicts PM2.5 concentrations, a vital air quality metric, using time-series data to support environmental monitoring and public health. I employed a Long Short-Term Memory (LSTM) neural network, optimized with Optuna, to model temporal dependencies in meteorological and temporal features.

2. Data Exploration

Dataset Summary

The training dataset (30,676 rows) and test dataset (13,148 rows) .

Preprocessing Steps

- Converted datetime to pandas datetime for feature extraction.
- Dropped 1,921 rows with missing PM2.5 in training; interpolated other missing values linearly to preserve temporal continuity.
- Added temporal features (hour, day, month, dayofweek) to capture diurnal and seasonal patterns.
- Applied MinMaxScaler to normalize features (DEWP, TEMP, PRES, Iws, Is, Ir, cbwd_NW, cbwd_SE, cbwd_cv, hour, day, month, dayofweek) and PM2.5 to [0, 1].
- Generated sequences of length 36 with stride 2 for LSTM input, based on daily and seasonal patterns observed in PM2.5 trends.

3. Model Design

Architecture

The best-performing model has:

- **Input**: Sequences of length 36 with 13 features.
- **LSTM Layers**: 1 layer with 64 units (ReLU activation).
- **Dropout**: Rate of 0.2 after LSTM and dense layers.
- **Dense Layers**: 16-neuron layer (ReLU), followed by a single output neuron (float32).
- **Optimizer**: Adam with learning rate 0.000641.
- **Loss**: Mean Squared Error (MSE) with RMSE metric.

Rationale

- **LSTM**: Captures long-term dependencies in PM2.5 time-series data.

- **Optuna**: Optimized parameters (sequence length: 36, units: 64, layers: 1, dropout: 0.2, learning rate: 0.000641, batch size: 64, stride: 2) to achieve validation RMSE of 71.77.
- **Mixed Precision**: Enhanced training efficiency.
- **Early Stopping**: Prevented overfitting by restoring best weights after 5 epochs of no improvement.

4. Experiment Table

The table summarizes successful Kaggle submissions for "Assignment 1 - Time Series Forecasting May 2025."

Submission File	Seq Length	Units	Layers	Dropout	Learning Rate	Batch Size	Stride	Val RMSE	Public RMSE
submissiontest.csv	24	64	2	0.3	0.0005	64	2	107.64	6369.6387
subm_fixed.csv	36	96	2	0.3	0.0005	64	1	99.15	5868.0520
submission (1).csv	24	64	1	0.2	0.001	32	1	80.95	5659.5398
submission3.csv (Best)	36	64	1	0.2	0.000641	64	2	71..77	4740.7316

Failed Submissions

- Failed due to incorrect row count (expected 13,148 rows).
- Others failed due to mismatched row IDs or values, likely from incorrect datetime formatting

5. Results

Performance

- **Public RMSE**: achieved 4740.7316 on public score, outperforming others by optimizing sequence length and learning rate.
- **Validation Metrics** (last 100 samples):
    - **RMSE**: 71.77
    - **MAE**: 64.50
    - **R^2**: -0.09 (indicating limited variance explanation, suggesting room for improvement)

Key Findings

- **Temporal Features**: Hour, month, and (dayofweek) captured critical patterns.
- **Hyperparameter Tuning**: Optuna's optimization (e.g., longer sequence length, lower learning rate) reduced RMSE significantly.
- **Iterative Prediction**: Effectively handled test set by updating sequences with predicted PM2.5 values.

6. Conclusion

This project developed a robust LSTM model for PM2.5 prediction, achieving a public RMSE of 4740.7316 . Comprehensive preprocessing (interpolation, temporal features, scaling) and Optuna-based tuning drove improvements over earlier submissions

Proposed Improvements

- Feature Expansion
- Model Enhancements (like Exploring bidirectional LSTMs or attention mechanisms for complex patterns.)
- Error Reduction: Analyze residual outliers to improve prediction accuracy.
- Submission Robustness (Ensure consistent row counts and ID formats.)
- Run Optuna optimization on more trials=

GitHub Repository

The documented code, including visualizations, is available at: https://github.com/Jules-gatete/Assignment-1--Time-Series-Forecasting-.git.