National College of Ireland

Configuration Manual

MSc Research Project
Programme Name

ALFIN BIJU
X23278579

School of Computing
National College of Ireland

Supervisor: JASWINDER SINGH

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ALFIN BIJU

Student ID: X23278579

Programme: MSC DATA ANALYTICS **Year:** 2025

Module: RESEARCH PRACTICUM

Lecturer: JASWINDER SINGH

Submission Due Date: 11/08/25

Project Title: Air Quality Forecasting Using Transformer and LSTM Models: A Comparative Study on Single-City Data

Word Count: 4711 **Page Count:** 4

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ALFIN BIJU

Date: 11/08/25

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

ALFIN BIJU
X23278579

1. System Requirements

Component	Requirement
OS	Windows 10+, macOS 10.15+, or Ubuntu 18.04+
Python Version	Python 3.8 – 3.11
GPU Support	Optional but recommended (CUDA-enabled GPU)
RAM	Minimum 8GB (16GB recommended)
Disk Space	At least 5GB for datasets and models

2. Python Dependencies

Install all necessary packages via pip:

```
pip install pandas numpy matplotlib scikit-learn statsmodels seaborn torch torchvision torchaudio

pip install pandas>=2.0 numpy>=1.24 scikit-learn>=1.3 matplotlib>=3.7 statsmodels>=0.14 torch>=2.1 google-colab>=1.0
```

3. Folder Structure

```
project_root/
│
├── data/
│   └── Houston_AirQuality_AllPollutants_2015_2024.csv
│
```

```
├── notebooks/
│   ├── lstm_forecast.ipynb
│   ├── informer_transformer.ipynb
│   ├── sarimax_model.ipynb
│
├── models/
│   └── saved_lstm.pt
│   └── saved_informer.pt
│
├── outputs/
│   └── plots/
│   └── evaluation_metrics/
│
├── config_manual.txt
```

4. Model Input and Output Configuration

Model Type	Input Sequence	Forecast Horizon	Targets
LSTM	336 hours (14 days)	168 hours (7 days)	PM2.5, PM10, NO_2, SO_2, CO, O_3
Informer	336 hours (14 days)	168 hours (7 days)	Same as LSTM
SARIMAX	N/A (uses full daily history; seasonal=7)	7 days (daily steps)	PM2.5 only

5. Data Preprocessing Pipeline

1. **Datetime Resampling**: Hourly (resample('1H'))

2. **Interpolation**: Linear forward/backward filling
3. **Feature Engineering**:
    - Time features: hour, day of week, month, weekend
    - Lag features: lag24_PM2.5, lag168_PM2.5
4. **Normalization**:
    - Inputs and targets scaled via MinMaxScaler()

6. Training Configuration

Model	Epochs	Batch Size	Optimizer	Loss Function	Scheduler	Dropout
LSTM	100	128	Adam	MSE	ReduceLROnPlateau (opt.)	0.3
Informer	80	64	Adam	SmoothL1	ReduceLROnPlateau (opt.)	0.4
SARIMAX	N/A	N/A	MLE (internal)	N/A	N/A	N/A

7. Evaluation Metrics

All models evaluated using:

- **RMSE**: Root Mean Square Error
- **MAE**: Mean Absolute Error
- **R^2 Score**: Coefficient of Determination

8. Execution Instructions

1. Prepare the Dataset

• Place your CSV dataset in the data/ directory.

- Make sure the file is properly formatted (with the timestamps and the readings of different pollutants).

2. Open the Desired Notebook

Open either of the following Jupyter notebooks, depending on the model you would like to run:

- sarimax_model.ipynb – for SARIMAX forecasting.
- lstm_model.ipynb – for LSTM forecasting.
- transformer_model.ipynb – for Informer Transformer forecasting.

3. Run the Notebook

- Execute all cells sequentially from top to bottom.
- Each notebook will:
- Load and preprocess the data.
- Train the model.
- Generate forecasts and plots.

4. Check the Outputs

The model files which are trained will be stored in the directory models/.
The forecast plots and assessment measures will be saved under the outputs/ directory.

9. Output Samples

- Forecast vs Actual Plots (per pollutant)
- Final metrics table comparing SARIMAX, LSTM, and Informer

10. Known Limitations

SARIMAX models only a single pollutant (PM2.5) and has a problem with spikes.

- Missing values >30% may require robust imputation.
- Transformer-based models require more GPU memory and fine-tuning.