National College of Ireland

Configuration Manual

Hybrid Predictive Modelling for Cargo Traffic Forecasting at Major and Non-Major Ports

MSc Research Project
Data Analytics

Pintoo Ramkis Baghel
Student ID: x23287501

School of Computing
National College of Ireland

Supervisor: Prof. Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Pintoo Ramkis Baghel
Student ID:	x23287501
Programme:	Data Analytics **Year:** 2025 – 26
Module:	Research Practicum
Lecturer:	Prof. Jorge Basilio
Submission Due Date:	11/08/2025
Project Title:	Hybrid Predictive Modelling for Cargo Traffic Forecasting at Major and Non-Major Ports
Word Count:	340 **Page Count:** 3

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: Pintoo Ramkis Baghel

Date: 11/08/2025

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

Pintoo Ramkis Baghel
x23287501

1. Environment System Requirements

To run the forecasting system and its components, the following software and hardware specifications are recommended:

System Requirements:

Operating System: Windows 10/11, macOS Monterey+, or Ubuntu 20.04+
Processor: Intel i5 or higher / AMD Ryzen 5 or higher
RAM: 8 GB minimum (16 GB recommended)
Storage: Minimum 5 GB free space
Python Version: 3.10 or later
IDE: Google Colab (Upload the dataset), Jupyter Notebook or any Python IDE (Change the path for accessing the dataset) (e.g., VS Code, PyCharm)

2. Python Libraries and Dependencies

Install the required packages using pip:

pip install the following libraries.

pandas 2.2.2, matplotlib 3.8.4, seaborn 0.13.2 – Stable modern stack for Python 3.10
scikit-learn 1.3.2 – It avoids some breaking changes in newer 1.4/1.5
xgboost 1.7.6 – Stable pre-2.0 API
statsmodels 0.14.1 - It is compatible with NumPy 1.26
tensorflow – It works nicely with Python 3.10 across various other OS.
numPy 1.26.4 – maximum compatibility across wheels and TF
scipy - 1.11.4 – Used for statistical testing

3. Dataset Requirements

The project uses two CSV datasets:

1. year-month-and-state-wise-total-cargo-handled-at-major-and-minor-ports-in-india.csv
2. cargo-traffic-handled-at-major-and-nonmajor-ports.csv
These should be placed in a `data/` directory relative to the project root folder.

4. Project File Structure

The project files are organized as follows:
- data/
  └── year-month-and-state-wise-total-cargo-handled-at-major-and-minor-ports-in-india.csv
    - 1_preprocessing.py

- 2.1 Pivoted_port_data.py
- 2_stationarity_arima.py
- 3_forecast_arima.py
- 4_xgboost_major.py
- 5_xgboost_minor.py
- 6_lstm_major.py
- 7_lstm_data_minor.py
- 8_lstm_minor.py
- 9_plot_arima_major.py
- 10_plot_xgboost_major.py
- 11_plot_hybrid_major.py
- 12_hybrid_forecast.py
- 13_visualize_summary.py

5. How to Run the Code

Follow these steps to execute the project files in sequence:

1. Preprocess the dataset:
   - Run "1_preprocessing.py" to clean and prepare data.
   - Run pivot_data.py (or the pivoting code) to generate pivoted_port_data.csv. This is required for LSTM scripts.

2. Generate forecasts:
   - Run ARIMA: "2_stationarity_arima.py" then "3_forecast_arima.py"
   - Run XGBoost: "4_xgboost_major.py" and "5_xgboost_minor.py"
   - Run LSTM: "6_lstm_major.py", "7_lstm_data_minor.py", "8_lstm_minor.py"

3. Create ensemble forecast:
   - Run "12_hybrid_forecast.py"

4. Generate visuals:

   - ARIMA Plot: "9_plot_arima_major.py"
   - XGBoost Plot: "10_plot_xgboost_major.py"
   - Hybrid Plot: "11_plot_hybrid_major.py"
   - Output Summary: "13_visualize_summary.py"

6. Reproducibility

Environment

Python 3.10–3.11

Libraries used: pandas, numpy, scikit-learn, statsmodels, xgboost, matplotlib, scipy

Random seed used: 42 (set for all models)

Data & Splits

Data sources: India Data Portal; Dataful.in

Granularity: monthly, by port type (Major/Minor)

Splits (In chronological order):

Train: (Jan 2016 – Jun 2024)

Validation (for selecting settings/weights): {Jul 2024 – Sep 2024}

Test/forecast horizon: Oct 2024 – Mar 2025

Preprocessing

Missing 1 month - linear interpolation

Multiple missing months - dropped

Outliers - winsorized (kept seasonality)

Scaling used only for LSTM (Min-Max)

ARIMA/SARIMA: (p,d,q) and seasonal (P,D,Q,12)

XGBoost: depth, learning rate, number of trees, subsample

LSTM: lookback window, units, dropout, batch size, epochs

Hybrid: equal weights (0.5/0.5).

How to Reproduce

a) Statistical test

Wilcoxon signed-rank on absolute errors (Oct 2024–Mar 2025).
Major: Hybrid vs ARIMA p=0.03; Hybrid vs XGBoost p=0.06; n=6.

Install the listed libraries.

Run the ARIMA, XGBoost, and LSTM notebooks/scripts.

Run the ensemble step to create the hybrid.

Run the evaluation step to save metrics and Figures from 6.1 - 6.6.

Ensemble Weights

If using weighted average: weight = (1 ÷ RMSE) for each model, then scale so weights add to 1.

Example (Major): ARIMA 4.59, XGBoost 4.10 → approx XGBoost 0.53, ARIMA 0.47.

7. Notes and Troubleshooting

- Ensure all required libraries are installed before running scripts.
- Use absolute file paths in Colab or set correct working directory if using Jupyter.
- Some forecasts may vary due to random initialization (especially in LSTM).

References

Dataful (Factly (2025). *Year-, Month- and State-wise Total Cargo Handled at Major and Minor Ports in India*. [online] Dataful.in. Available at: https://dataful.in/datasets/13/

Indiadataportal.com. (2025). *IDP | Cargo Traffic Handled at Major and Non-Major Ports*. [online] Available at: https://indiadataportal.com/p/maritime-trade/r/mopsw-cargo_handled-st-yr-aaa