National College of Ireland

Configuration Manual

MSc Research Project
Data Analytics

Deepika Kancharla
Student ID: 23406739

School of Computing
National College of Ireland

Supervisor: Hamilton Niculescu

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Deepika Kancharla

Student ID: 23406739

Programme: Master's in Data Analytics **Year**:2025

Module: Research Practicum

Lecturer: Hamilton Niculescu

Submission Due Date: 11/12/2025

Project Title: Evaluating the Long-Term Impact and Predictive Analytics of COVID-19 Booster Vaccination Uptake

Word Count: 577 **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: Deepika Kancharla

Date: 10/12/2025

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):	

1 Configuration Manual

1. System Requirements & Setup

All experiments in this research were executed using **Google Colab**, which provides a cloud-based environment suitable for large-scale data analysis and machine-learning experiments. The following hardware and software configurations were used to ensure full reproducibility of the results.

1.1 Hardware Configuration (Google Colab Runtime)

Google Colab provides virtualised hardware. During all modelling and analysis tasks, the following configuration was used:

- CPU: 2 × Intel(R) Xeon® Virtual CPU @ 2.20 GHz
- RAM: 12.7 GB
- Disk Storage: ~68 GB temporary runtime storage
- Runtime Type: Standard CPU Runtime

This configuration supports fast processing for large tabular datasets, scaling of feature-engineering pipelines, and efficient execution of machine-learning models such as Linear Regression and Random Forest.

1.2 Software Configuration

- Platform: Google Colaboratory (browser-based Jupyter environment)
- Operating System: Linux (Ubuntu 22.04 LTS backend)
- Programming Language: Python 3.10 (default Colab Python version)
- Development Environment: Google Colab Notebook

2. Core Libraries Used

Data Handling & Processing

- pandas
- numpy

Visualisation

- matplotlib
- seaborn

Machine Learning Models

- scikit-learn
    - LabelEncoder
    - StandardScaler

- train_test_split
- LinearRegression
- RandomForestRegressor

Model Evaluation

- r2_score
- mean_absolute_error
- mean_squared_error

These libraries support data collection through APIs, cleaning, preprocessing, modelling, prediction, and evaluation of booster uptake across countries.

3. Dataset Description

This study uses publicly available global COVID-19 datasets from multiple trusted sources, representing vaccination behaviour, booster uptake, hospitalisation indicators, and daily epidemiological updates.

3.1 Dataset Sources

3.1.1 Our World in Data (OWID)

Booster vaccination data & COVID-19 hospitalisation indicators

- COVID-19 hospitalisation data: https://github.com/owid/covid-19-data/blob/master/public/data/hospitalizations/covid-hospitalizations.csv
- Booster vaccination data: https://ourworldindata.org/grapher/cumulative-covid-vaccine-booster-doses

3.1.2 World Health Organization (WHO)

Global daily epidemiological data

- https://data.who.int/dashboards/covid19/data

3.1.3 Kaggle Dataset

COVID-19 global vaccination progress

- https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress

3.2 Dataset Usage in This Project

The following steps were applied to each dataset:

- Imported using pandas
- Merged on country and date
- Cleaned for missing values, duplicates, and inconsistent date formats
- Feature transformations applied:
  - Label encoding for categorical variables

- Scaling for numeric variables

- Target variable: daily booster uptake
- Predictor variables include:
  - Historical vaccination metrics
  - Epidemiological indicators
  - Hospitalisation & ICU occupancy rates
  - WHO daily case/death counts

The final merged dataset contained 4,991 country-date observations used for modelling.

4. How to Run the Project

This section explains how to reproduce the entire analysis pipeline on Google Colab.

4.1 Uploading the Notebook and Loading the Dataset

Step 1: Visit
https://colab.research.google.com
Step 2: Upload the provided notebook file – x23406739.ipynb
Step 3: Upload downloaded datasets (OWID, WHO, Kaggle CSV files)
Step 4: Load the datasets in the notebook using Pandas library.

4.2 Setting Up the Environment in Colab

All dependencies used in this project are available in Colab by default.
Import all required libraries:

```
import pandas as pd
import requests
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression, LogisticRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared
```

No additional installations are required.

4.3 Running the Notebook

Execute cells in sequence:

1. Load OWID, WHO, and Kaggle datasets
2. Merge datasets on countryand date

3. Clean and preprocess data
4. Encode categorical variables
5. Split into training/test sets
6. Train Linear Regression model
7. Train Random Forest model
8. Evaluate each model using:
    - R^2
    - RMSE
    - MAE
9. Generate plots:
    - Correlation heatmap
    - Feature importance
    - Booster uptake prediction plots
    - Hospital/ICU vs predicted uptake scatterplots

4.4 Accessing Results

Each execution cell outputs visual and numerical results including:

- Model performance metrics
- Predicted vs actual booster uptake graphs
- Feature importance ranking
- Scatterplots showing hospital/ICU occupancy trends

All outputs can be downloaded from Colab for reporting.

5. References

NumPy Developers (2025) *NumPy: Scientific computing for Python.* Available at: https://numpy.org
Scikit-learn Developers (2025) *Scikit-learn: Machine Learning in Python.* Available at: https://scikit-learn.org
SciPy Developers (2025) *SciPy: Scientific computing tools for Python.* Available at: https://scipy.org
Hunter, J.D. et al. (2025) *Matplotlib: Visualization library for Python.* Available at: https://matplotlib.org
OpenCV Team (2025) *OpenCV: Open-source computer vision library.* Available at: https://opencv.org
Seaborn Developers (2025) *Seaborn statistical visualisation library.* Available at:

https://seaborn.pydata.org
Google Research (2025) *Google Colaboratory: Cloud-based Python environment.* Available at: https://colab.research.google.com
OWID (2025) *Our World in Data – COVID-19 Vaccination and Hospitalisation Data.* Available at: https://ourworldindata.org
WHO (2025) *WHO COVID-19 Global Data.* Available at: https://data.who.int/dashboards/covid19/data
Kaggle (2025) *COVID-19 World Vaccination Progress.* Available at: https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress