National College of Ireland

Configuration Manual

MSc Research Project
Data Analytics

Prachi Mahajan
Student ID: x22158511

School of Computing
National College of Ireland

Supervisor: Prof.Cristina Hava Muntean

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Prachi Mahajan
Student ID:	22158511
Programme:	MSc in Data Analytics
Year:	2023-2024
Module:	MSc Research Project
Supervisor:	Prof. Cristina Hava Muntean
Submission Due Date:	14 December 2023
Project Title:	Configuration Manual
Word Count:	746
Page Count:	5

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:	Prachi Mahajan
Date:	14th December 2023

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

Prachi Mahajan
x22158511

1 Introduction

The objective of this configuration handbook is to provide users with clear instructions on how to set up and carry out the thesis project. The project encompasses data processing, conducting exploratory data analysis (EDA), developing models using multiple machine learning algorithms, and creating a dashboard using PowerBI.

2 Hardware Specifications

Processor: 12th Gen Intel(R) Core(TM) i5-1240P
RAM: 16.0 GB
Operating System: Windows 11
Architecture: 64-bit
Processor Architecture: x64-based
GPU: Intel(R) Iris(R) Xe Graphics

3 Software Specifications

3.1 Integrated Development Environment(IDE)

The Jupyter Notebook was used as the Integrated Development Environment (IDE) for this project, with Python being the chosen programming language. The specific version of:
Jupyter Notebook : 6.4.12

The Anaconda Navigator platform has been installed, which encompasses Jupyter Notebook, and has ability to open a Python 3 file for launching and executing code. The installation of Anaconda's 64-bit version for Windows 11 is required. Once the installation is complete, open Anaconda Navigator and proceed to start Jupyter notebook. Upon clicking the launch button, the application will automatically open in the Browser.

Link for downloading Anaconda: `https://www.anaconda.com/download`

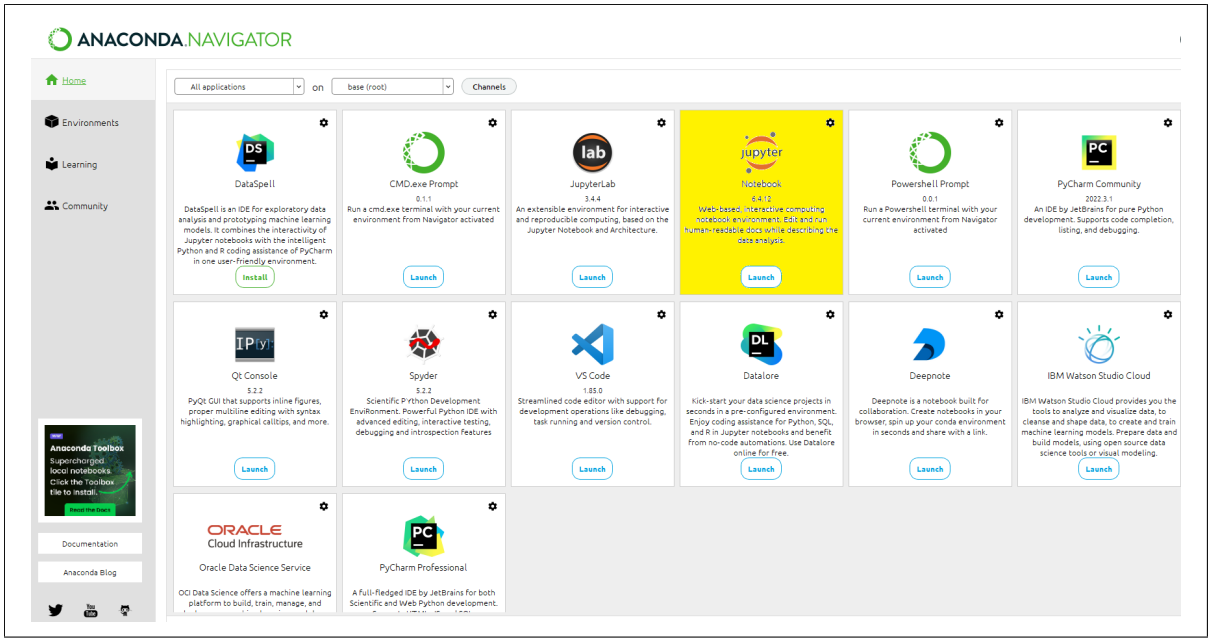

Figure 1: Anaconda Navigator

3.2 Softwares Used

3.2.1 Python:

The programming language selected for coding in this project was Python. The code can be found in the artefact zip file saved as EDA_&_ ModelTraining.ipynb, which includes the implementation of the Jupyter Notebook.

Python version : 3.9.13

3.2.2 PowerBI:

The interactive dashboard for this work was made using the business analytics tool PowerBI. The specific version of PowerBI used is 2.114.864.0 64-bit.

The Power BI dashboard is publicly hosted and can be accessed using NovyPro `https://www.novypro.com/` which serves as the platform for publishing and sharing the dashboard to a broader audience.
The dashboard can be accessible through the following link:
`https://www.novypro.com/project/french-railway-system-overview`

Additionally, the PowerBI dashboard file French Railway System Dashboard.pbix and the dataset used for visualization which is downloaded from Kaggle and cleaned before dashboard creation is available in the artefact zip folder.

3.3 Implementation Details

3.3.1 Dataset used

To import the dataset into the Python environment, the dataset public transport traffic data in france DUBUC (2021) needs to be downloaded first from the official website `https://www.kaggle.com/`.The data was downloaded and is now stored in the artefact zip folder, named Trains_France enabling easy access and the ability to reproduce the data.

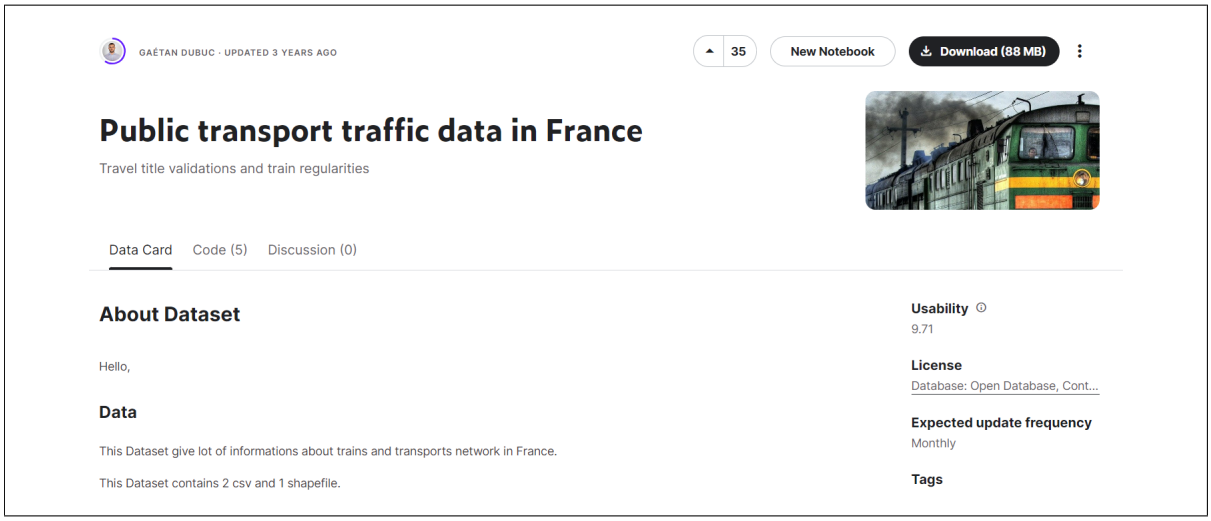

Figure 2: Public transport traffic data in France

3.3.2 Importing Libraries

Import the libraries specified below :

- Libraries for Loading Dataset and EDA

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import textwrap
```

Figure 3: Libraries for Loading Dataset and EDA

- Libraries for Train-Test Split and Standardization
- Libraries for Model Selection and Implementation

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
```

Figure 4: Libraries for Train-Test Split and Standardization

```
#pip install tensorflow
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import mean_squared_error,r2_score,make_scorer
from sklearn import tree
from sklearn.svm import SVR
from sklearn.multioutput import MultiOutputRegressor
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Activation
from tensorflow.keras.wrappers.scikit_learn import KerasRegressor
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import make_scorer, mean_squared_error, r2_score
```

Figure 5: Libraries for Model Selection and Implementation

3.3.3 Model development

This thesis encompasses the development of four models, including the Random Forest Regressor, Decision Tree, Support Vector Regression (SVR), and Artificial Neural Network(ANN).

3.3.4 Dashboard Implementation

The main objective of the Power BI dashboard within the scope of this thesis is to function as an interactive tool for acquiring insights into French railway data. The dashboard has been created to offer a user-friendly and visually attractive interface for the purpose of examining and analysing crucial elements of the railway dataset.

The dashboard employs a range of visualisations to improve the depiction of data:

- Donut chart: is a graphical representation that displays the proportions of categorical data in a circular format.
- Line chart: is a visual representation that depicts trends and patterns in data across time.
- Area chart: is a graphical representation that displays the total magnitude of data as it accumulates over time.
- Cards: are used to present essential measurements or condensed data.
- Slicer: Facilitating dynamic filtering based on given criteria

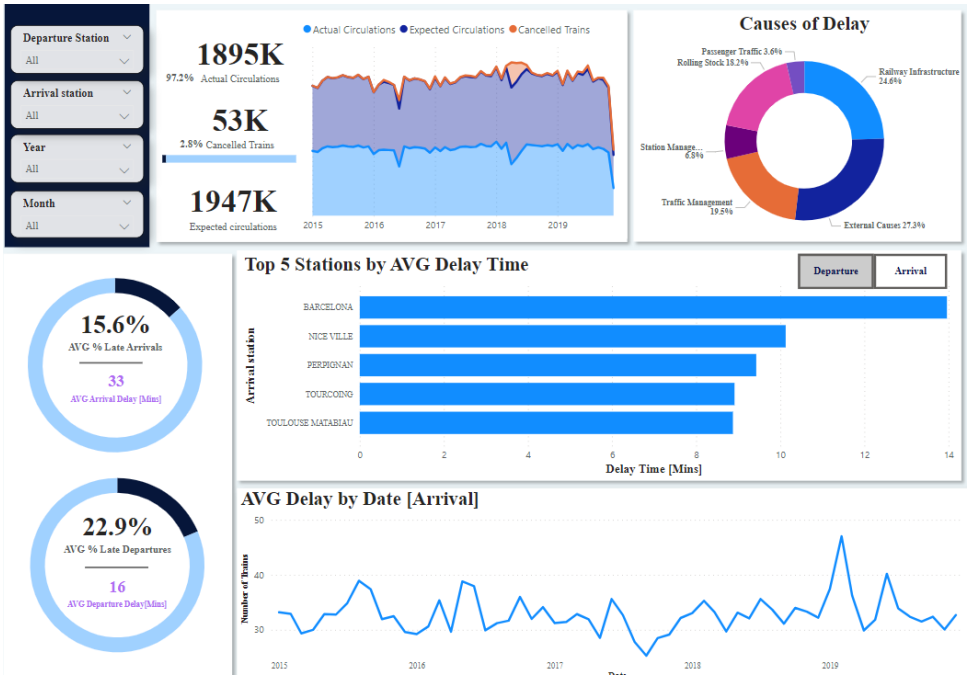

Figure 6: PowerBi dashboard for France Railway Network

4 Conclusion

4.1 Summary

This manual provides detailed instructions for configuring the thesis environment, covering essential aspects such as hardware requirements, software installation, model creation, and the development of a Power BI dashboard. The dataset, obtained from Kaggle, is comprehensive, while NovyPro is utilised for hosting the public dashboard.

4.2 Recommendation for best Practices

In order to guarantee the success of the project, it is crucial to thoroughly document the code, utilise version control for collaborative work, prioritise the ability to reproduce the project's environment, and regularly create backups of key files. Following these standards improves effectiveness and fosters the long-term viability of the project.

References

DUBUC, G. (2021). Public transport traffic data in france.
URL: *https://www.kaggle.com/datasets/gatandubuc/public-transport-traffic-data-in-france*