National College of Ireland

Configuration Manual

MSc Research Project
Data Analytics

Sanket Patne
Student ID: x22171746

School of Computing
National College of Ireland

Supervisor: Prof. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sanket Patne
Student ID:	22171746
Programme:	MSc in Data Analytics
Year:	2023-2024
Module:	MSc Research Project
Supervisor:	Prof. Catherine Mulwa
Submission Due Date:	14 December 2023
Project Title:	Configuration Manual
Word Count:	569
Page Count:	7

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:	Sanket Patne
Date:	31st January 2024

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

Sanket Patne
x22171746

1 Introduction

This configuration manual is made with an end user in mind that would want to replicate this study from scratch. It is designed to minimize any confusions or misdirections so that the users can follow a set of easy step by step instructions to get to the findings of the study implemented. The Configuration Manual consists the following six sections:
• Environmental Setup
• Libraries Required
• Data-set
• User Interface
• Implementation
• Code Repository

2 Environment Setup

2.1 Hardware Specifications

Processor: 8th Gen Intel(R) Core(TM) i5-8300H CPU @ 2.30GHz
RAM: 8.0 GB
Operating System: Windows 11
Architecture: 64-bit
Processor Architecture: x64-based
GPU: Intel UHD Graphics 630, NVIDIA GeForce GTX 1050TI

2.2 Software Specifications

2.2.1 Integrated Development Environment(IDE)

Google Collab 22.04 with Python version 3.10+. Google Collab is an online Python IDE similar to Jupyter notebook which allows the user to have separate environments than the existing ones on local systems to avoid conflicts in library requirements. The main reason behind using Collab was also the GPU and TPU cores it offers for deep learning models and neural network executions.
Link to access Google Collab online: https://colab.google/

3 Libraries required

Figure 1 shows which libraries were involved in the development of the study.

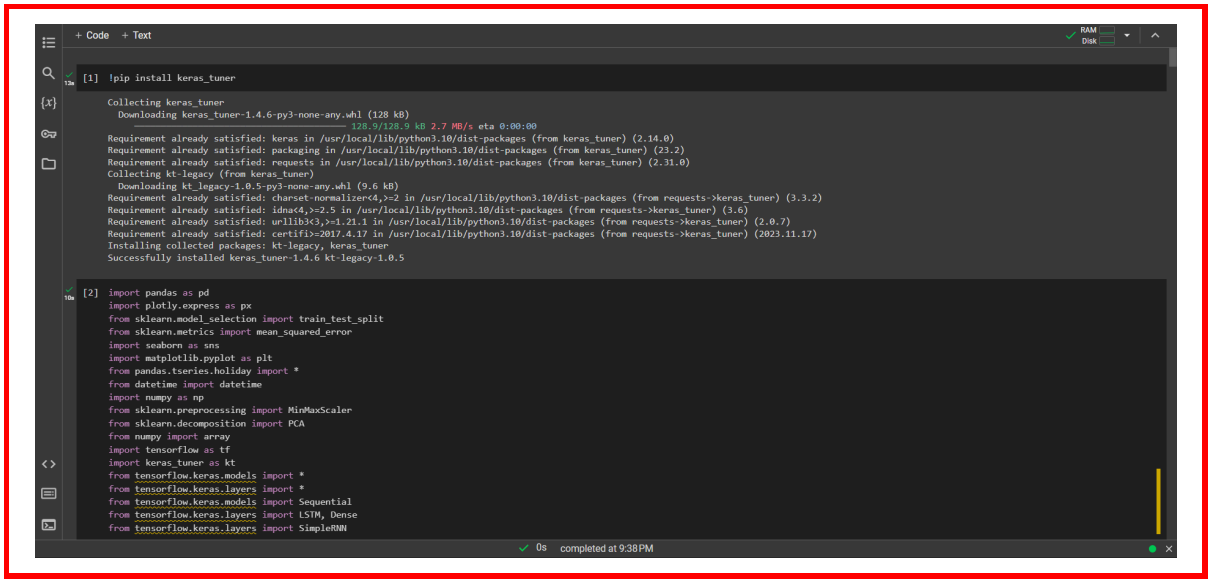


Figure 1: List of libraries required

4 Dataset

We were able to use the open datasetNetworks (2023) provided and maintained by the Government of London at link[1]. This dataset consisted of half hourly energy consumption readings of various households catered to by the UK Government and the energy providers. The data was obtained from https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households . When the project needs to be run, the datasets can be downloaded from the google drive link and the link variable in the code would need to be updated with the current value.

[1]https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households

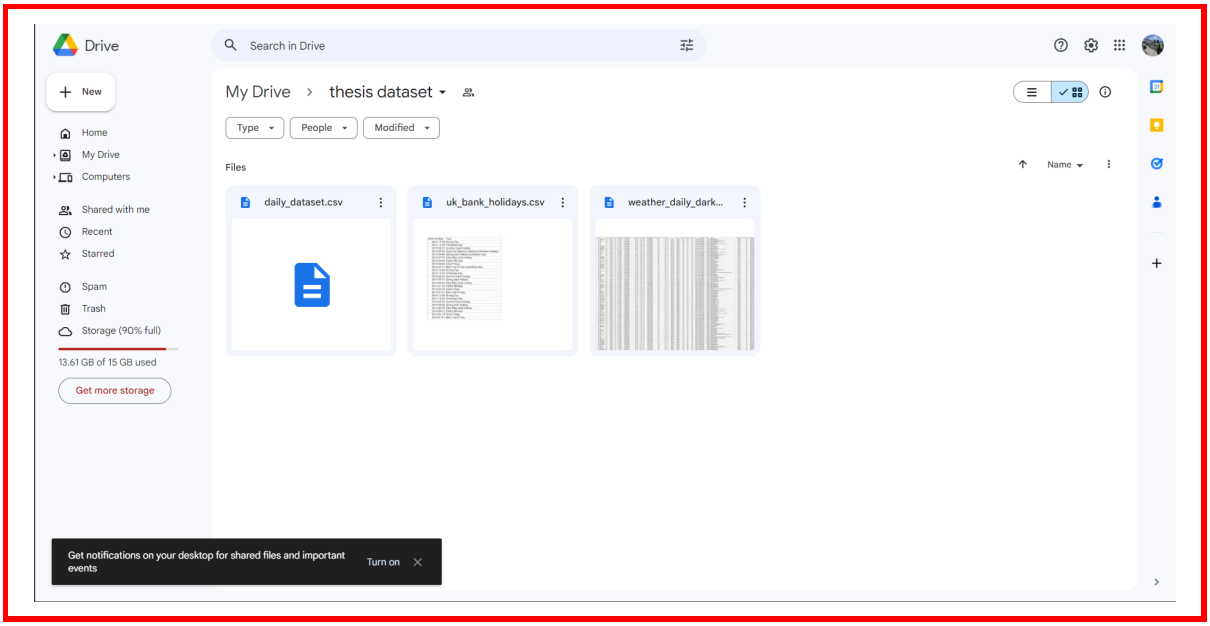


Figure 2: Google Drive Location of datasets compiled

Apart from this we also integrated weather data from the darkskySupport (2023) api available in the form of a csv available on Kaggle. To pair with this time series data we also created an instance of the abstract class UKHolidaysColman-Goff (2023). The csv file with weather data and the energy consumption data are stored in a personal Google Drive at link: https://drive.google.com/drive/folders/1W0LbpgCJ1a4vVD-BFiplhSN1z7gkg-dF?usp=sharing

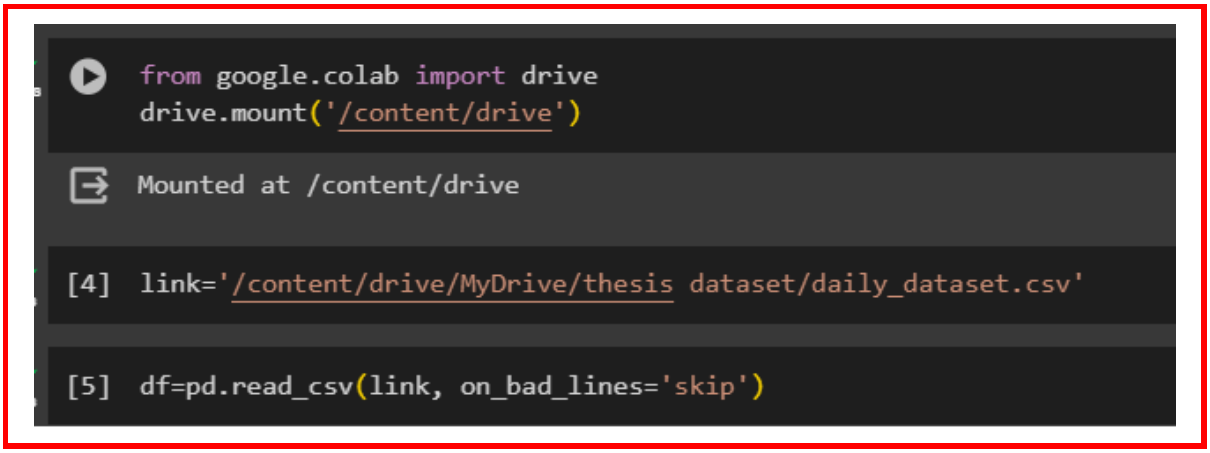


Figure 3: Python code to use dataset from Google Drive in Google Collab

5 Exploratory Data Analysis

EDA was carried out to get relevant features for the model creation on 3 fronts.

```
df.head()
```

	LCLid	day	energy_median	energy_mean	energy_max	energy_count	energy_std	energy_sum	energy_min
0	MAC000131	2011-12-15	0.4850	0.432045	0.868	22	0.239146	9.505	0.072
1	MAC000131	2011-12-16	0.1415	0.296167	1.116	48	0.281471	14.216	0.031
2	MAC000131	2011-12-17	0.1015	0.189812	0.685	48	0.188405	9.111	0.064
3	MAC000131	2011-12-18	0.1140	0.218979	0.676	48	0.202919	10.511	0.065
4	MAC000131	2011-12-19	0.1910	0.325979	0.788	48	0.259205	15.647	0.066

Figure 4: Descriptive Stats of Data Frame

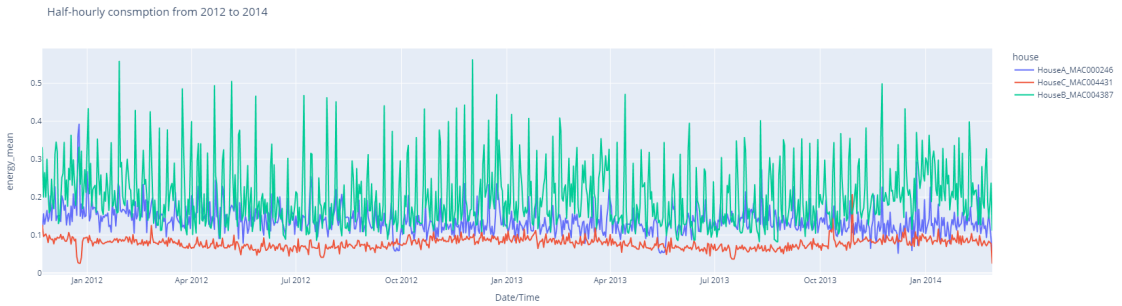


Figure 5: EDA

6 Implementation and Code repository

This section is used to display the code snippets from model development, evaluation and results sections.

```
# House A
X_traina,X_testa,Y_traina,Y_testa=train_test_split(house_a_new.drop('energy_mean',axis=1),house_a_new['energy_mean'].values,shuffle=False,test_size=0.2)
# House B
X_trainb,X_testb,Y_trainb,Y_testb=train_test_split(house_b_new.drop('energy_mean',axis=1),house_b_new['energy_mean'].values,shuffle=False,test_size=0.2)
# House C
X_trainc,X_testc,Y_trainc,Y_testc=train_test_split(house_c_new.drop('energy_mean',axis=1),house_c_new['energy_mean'].values,shuffle=False,test_size=0.2)
```

Figure 6: List of features being used to create the models

Figure 7: CNN Model

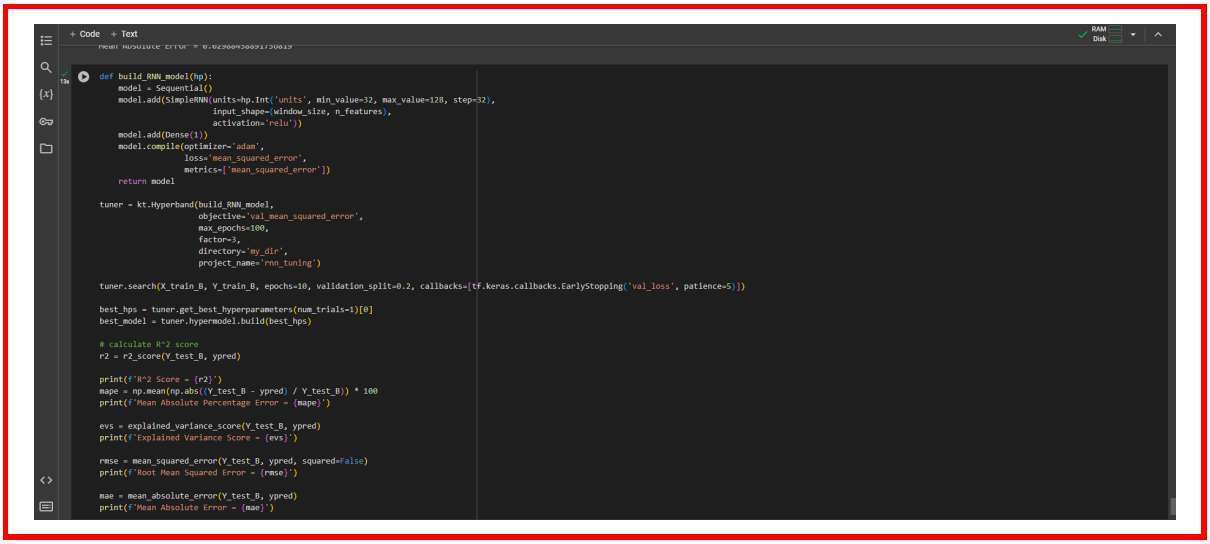
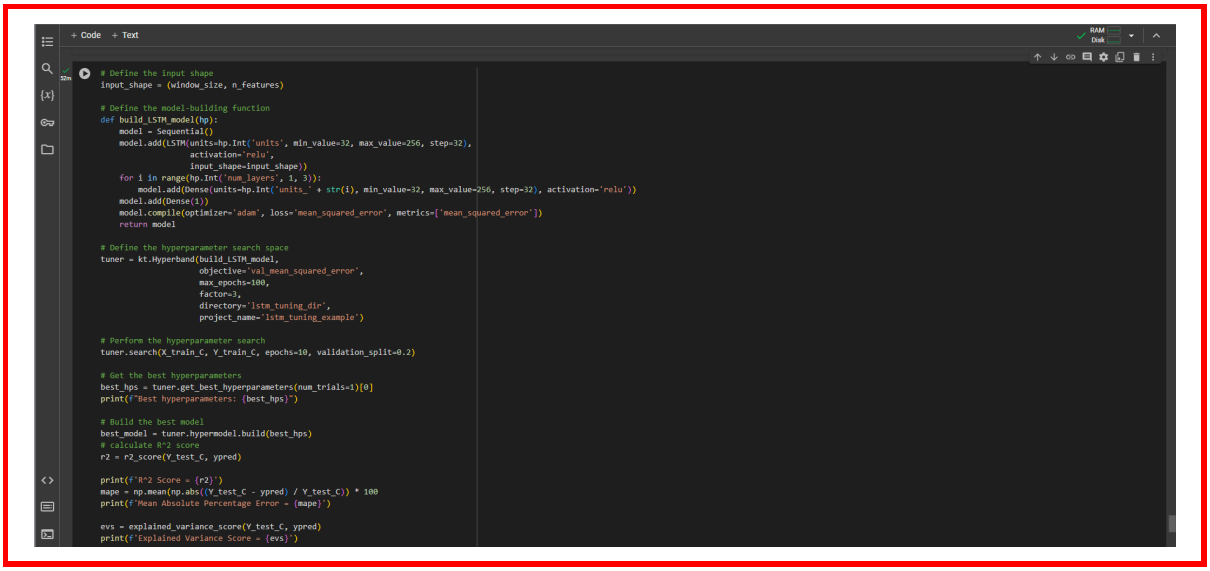


Figure 8: RNN Model



```
# Define the input shape
input_shape = (window_size, n_features)

# Define the model-building function
def build_LSTM_model(hp):
    model = Sequential()
    model.add(LSTM(units=hp.Int('units', min_value=32, max_value=256, step=32),
                   activation='relu',
                   input_shape=input_shape))
    for i in range(hp.Int('num_layers', 1, 3)):
        model.add(Dense(units=hp.Int('units_' + str(i), min_value=32, max_value=256, step=32), activation='relu'))
    model.add(Dense(1))
    model.compile(optimizer='adam', loss='mean_squared_error', metrics=['mean_squared_error'])
    return model

# Define the hyperparameter search space
tuner = kt.Hyperband(build_LSTM_model,
                     objective='val_mean_squared_error',
                     max_epochs=100,
                     factor=3,
                     directory='lstm_tuning_dir',
                     project_name='lstm_tuning_example')

# Perform the hyperparameter search
tuner.search(X_train_C, Y_train_C, epochs=10, validation_split=0.2)

# Get the best hyperparameters
best_hps = tuner.get_best_hyperparameters(num_trials=1)[0]
print(f"Best hyperparameters: {best_hps}")

# Build the best model
best_model = tuner.hypermodel.build(best_hps)
# calculate R^2 score
r2 = r2_score(Y_test_C, ypred)

print(f'R^2 Score = {r2}')
mape = np.mean(np.abs((Y_test_C - ypred) / Y_test_C)) * 100
print(f'Mean Absolute Percentage Error = {mape}')

evs = explained_variance_score(Y_test_C, ypred)
print(f'Explained Variance Score = {evs}')
```

Figure 9: LSTM Model

7 Conclusion

House	MAPE (%)	RMSE	MAE	Explained Variance Score	R squared
House A with CNN	28.0559	0.0370	0.02940	0.0389	-0.0911
House B with RNN	31.8411	0.0945	0.0657	0.0835	-0.209
House C with LSTM	108.438	0.0813	0.0766	-3.0634	-35.305

Table 1: Performance Metrics for Different Houses

Figure 10: Results

7.1 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.

7.2 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

Colman-Goff, L. (2023). How to automate bank holidays using python, https://medium.com/analytics-vidhya/python-how-to-automate-bank-holidays-767de0d42d44. [Online; accessed December 14th, 2023].

Networks, U. P. (2023). London data store, https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households. [Online; accessed 14th December, 2023].

Support, A. (2023). Darksky users apple support, https://support.apple.com/en-us/102594. [Online; accessed 14th December, 2023].