National College of Ireland

Configuration Manual

MSc Research Project
Data Analytics

Vasanthi Badami
Student ID: 20186690

School of Computing
National College of Ireland

Supervisor: Prof. Aaloka Anant

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Vasanthi Badami
Student ID:	20186690
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Prof. Aaloka Anant
Submission Due Date:	31/01/2022
Project Title:	Configuration Manual
Word Count:	900
Page Count:	10

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:	
Date:	31st January 2022

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

Vasanthi Badami
20186690

1 Introduction

The environmental setup and configurations required to conduct the various experiments to support the research are detailed in this configuration handbook. This document specifies the requirements for performing an influenza transmission rate investigation and projecting future values in detail.

2 Hardware Configuration and Environment Specification

2.1 Hardware configuration

The following fig 1 are the device specifications, which are used to configure all of the software needed for this study.

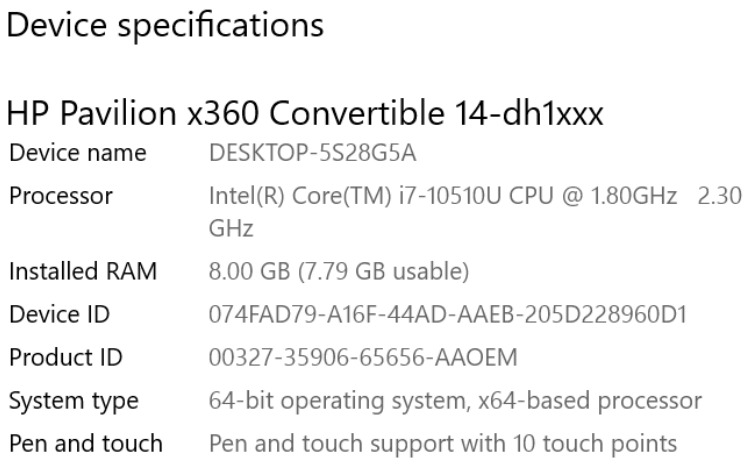


Figure 1: Devide Specification

3 Software Specifications

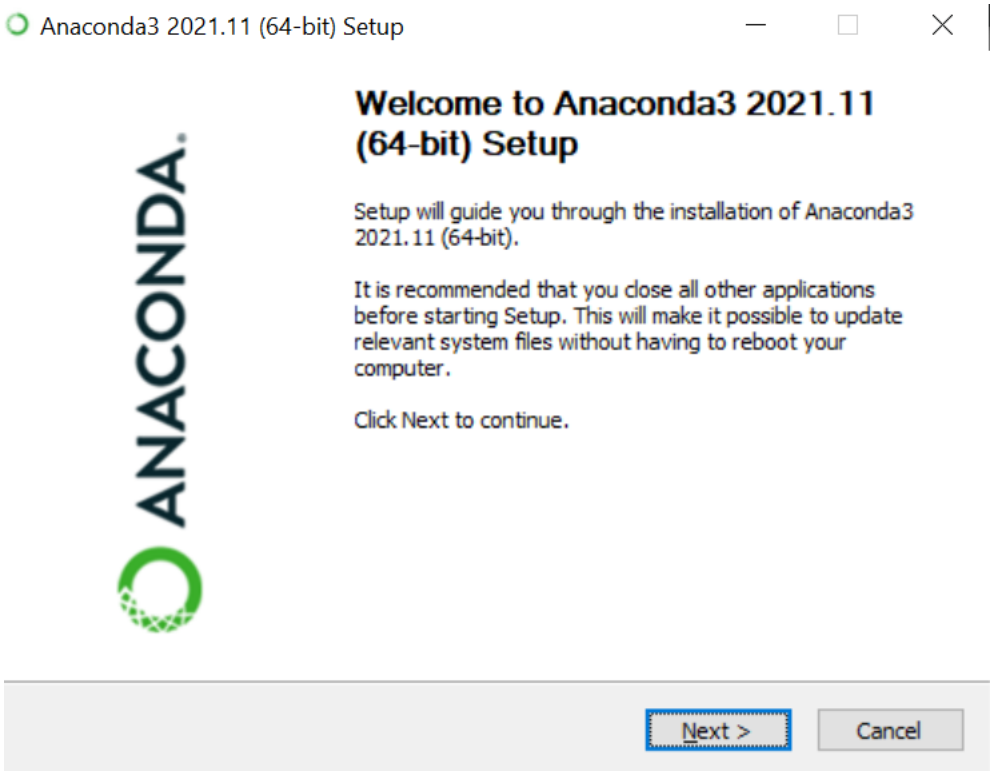
The following requirements were employed to conduct the experiments in this study:

- Anaconda (Jupyter notebook)
- Libraries

3.1 Anaconda (Jupyter notebook)

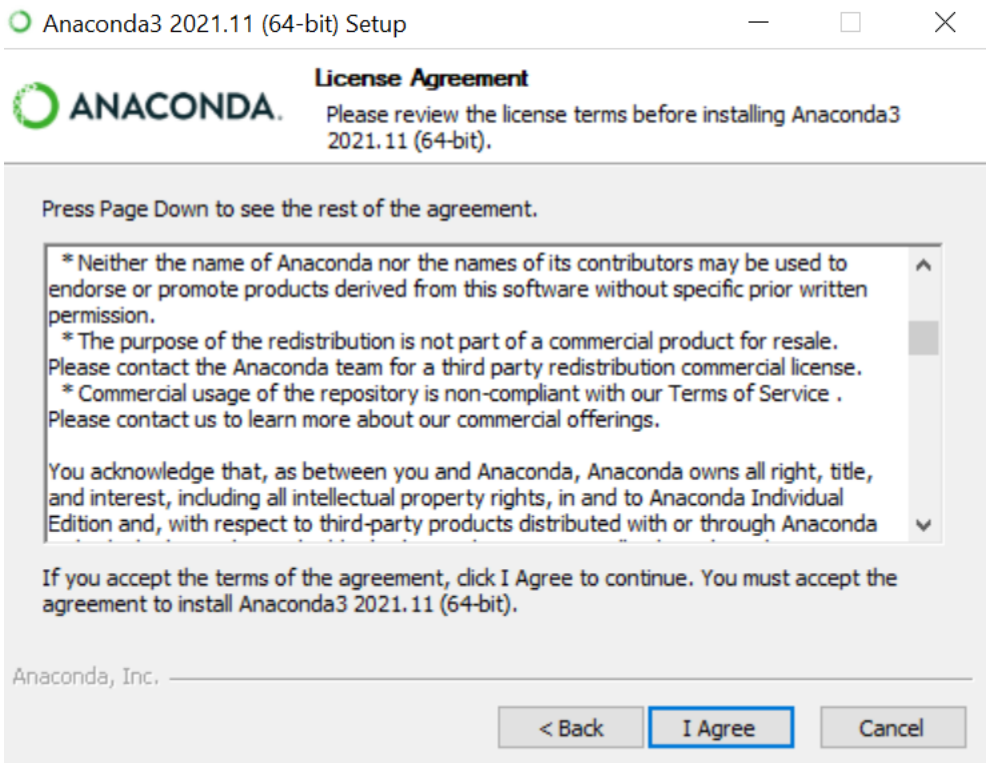
The Anaconda installer can be downloaded from anaconda.com[1]. and then proceeded with the following installation process.

- When we double-click the.exe file after installation, the dialogue as shown in fig ?? appears.

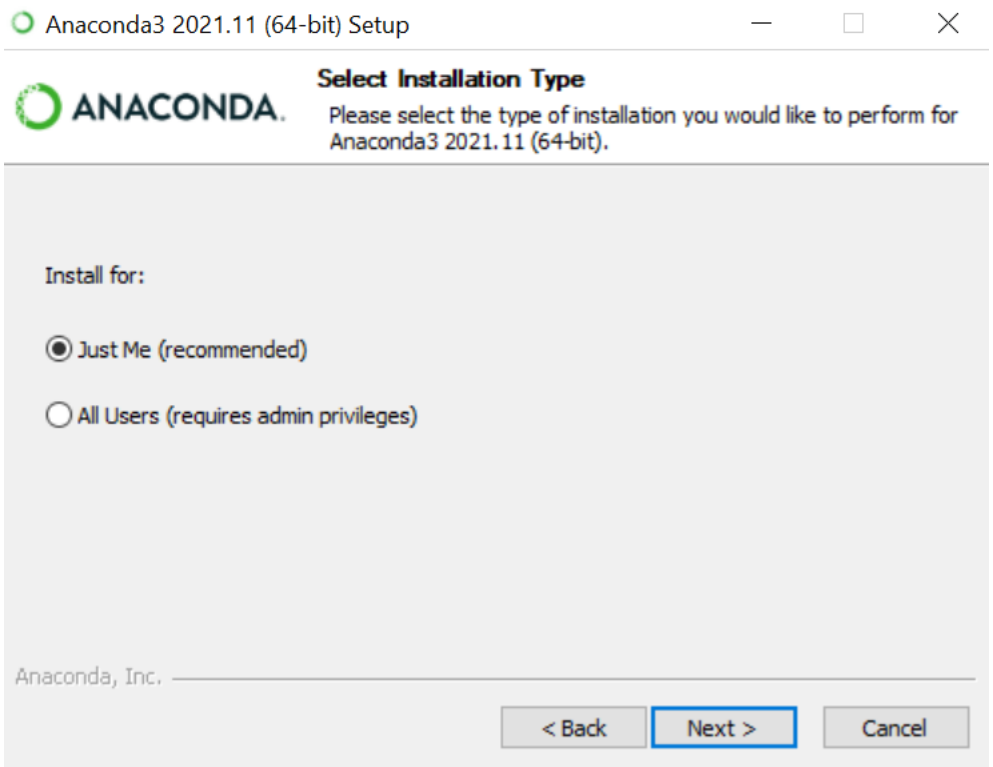


[1]anaconda installer `https://www.anaconda.com/products/individual#windows`

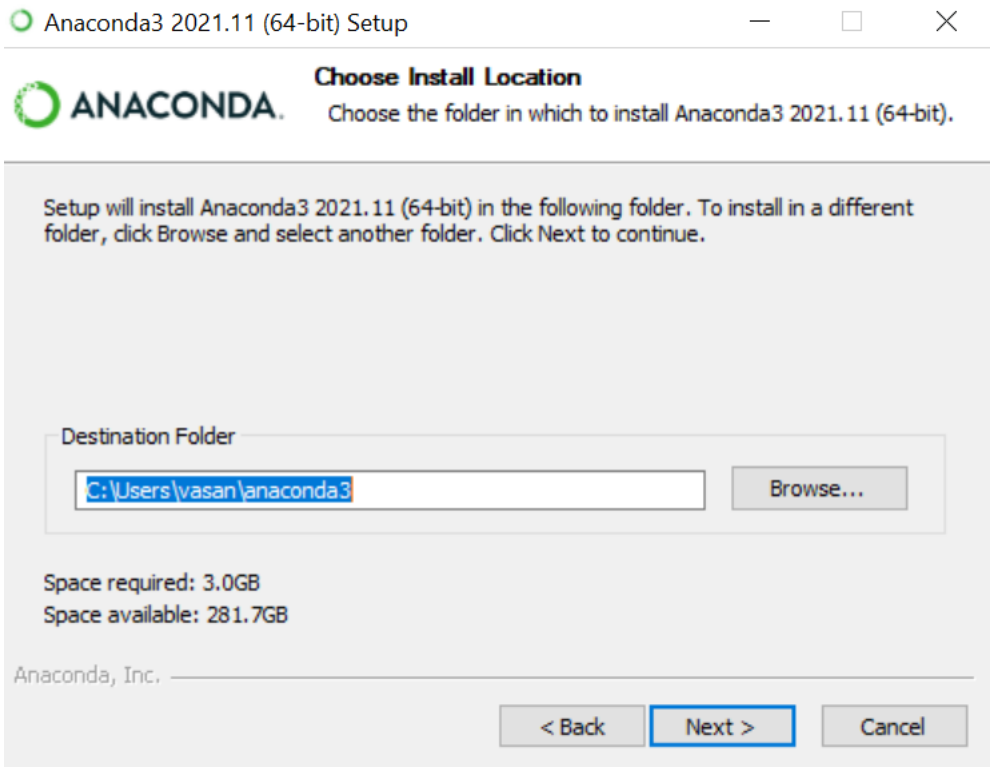
- The next window will display the licence agreement, which we must read and accept before clicking the I accept button.



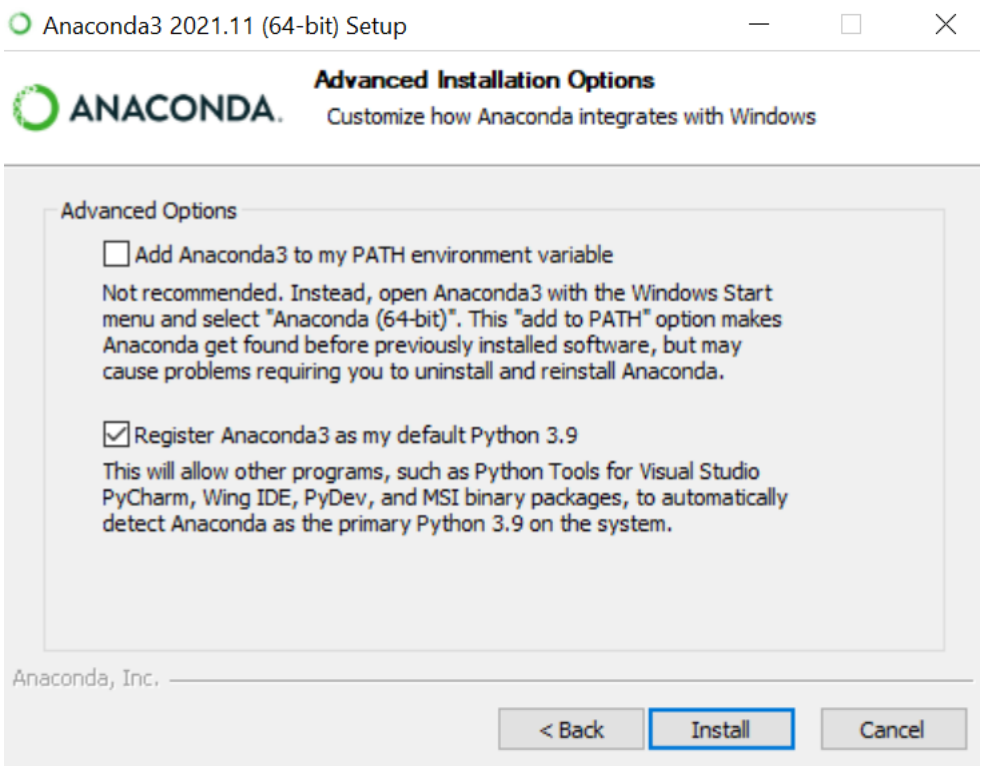
- We must select the installation type: either just me or all users who require administrative access.



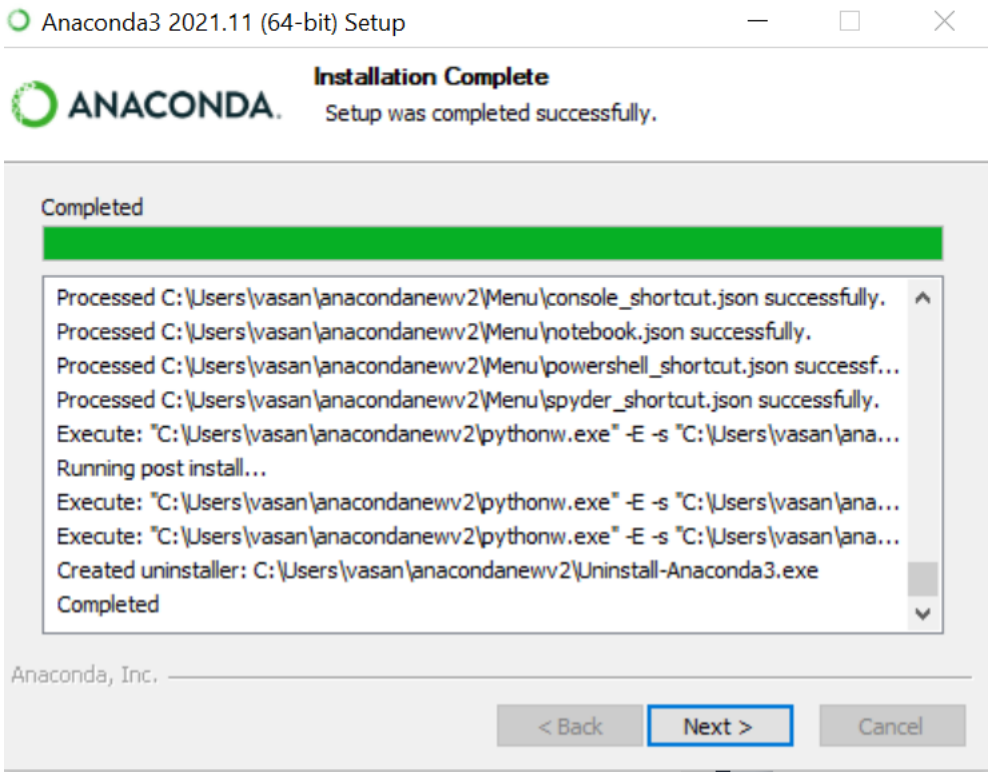
- To install the file, selecting a destination folder. Which will install all the supporting software in the same folder.



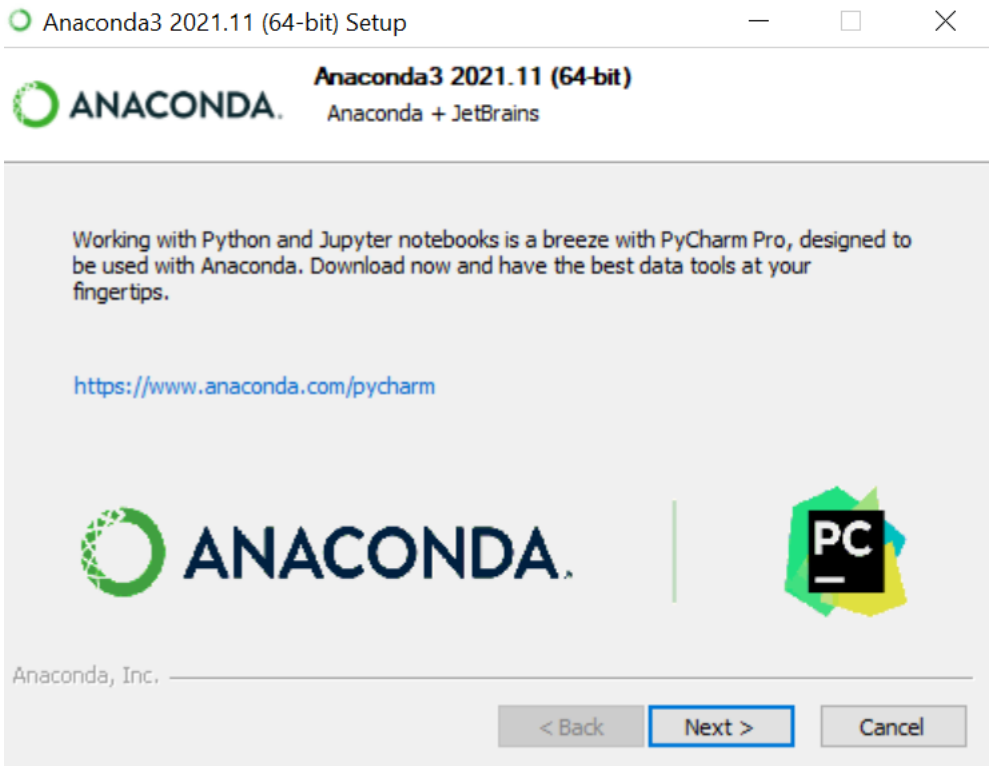
- Customize the way Anaconda works with Windows. We can either add anaconda to the environment variable path or set it as the default Python.



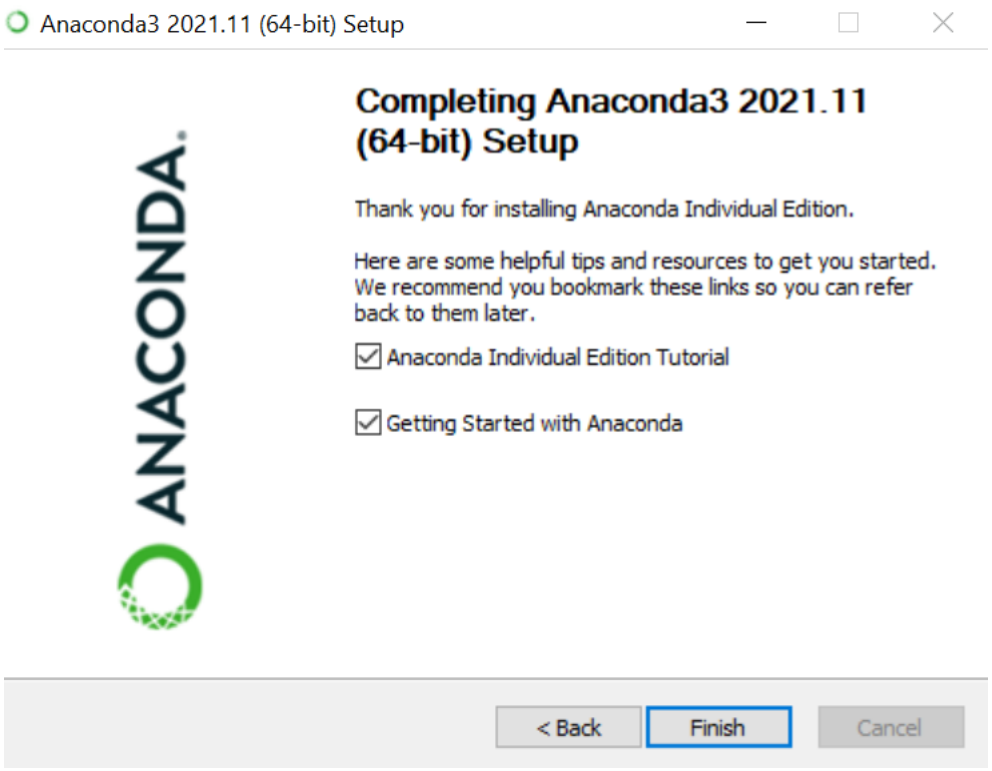
- After successfully completing the installation click on next button



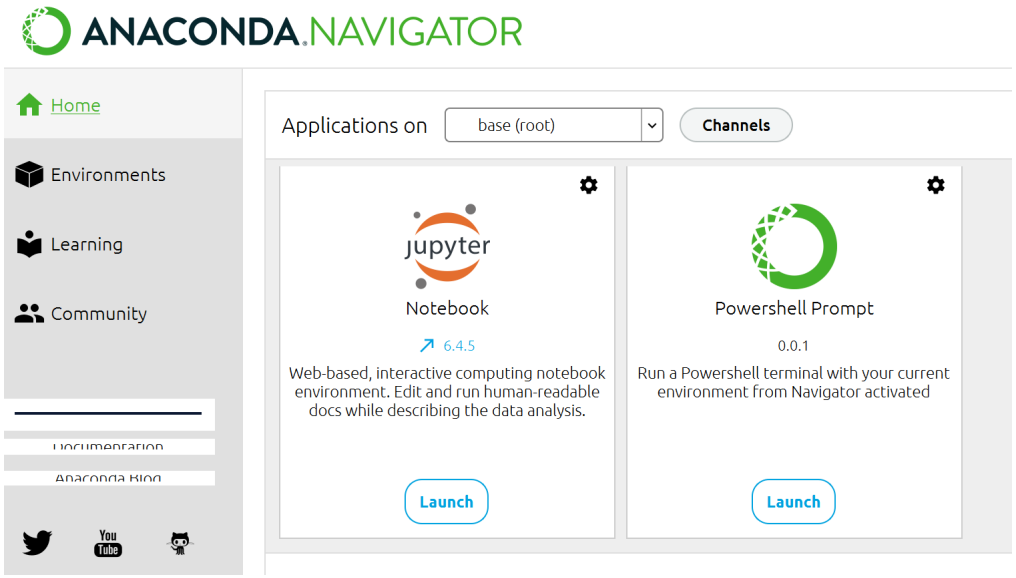
- This installation method includes a pycharm installation setup by default, which is useful for data analytics projects.



- After the installation is complete, it will display a completion window, where we must click the Finish button to complete the installation.



- The Jupyter notebook can be accessed from the anaconda navigator home after installation is complete.



3.2 Required Libraries

In order to perform this research progressively, significant libraries must be installed. The following libraries for this study.

- Pandas: Pandas is very useful library in python which allows data analysis project to flexibility to analyse and manipulate the data. and also we can perform various functionalities integrated in the pandas to explore data. In this study pandas used to read the data from .csv file and also used various group by functions to visualize the data.
- Numpy: Numpy is third party library which employs numerical computing in python which provides flexible mathematical functionalities to access arrays of number.
- Matplotlib: Matplotlib is a graphing package for Python with NumPy, the Python numerical mathematics extension. These library will also help in data visualization
- Seaborm : Seaborn library also employed in this study for better visualization of dataset. Seaborn is a matplotlib-based Python data visualisation package.
- Plotly: The plotly Python library is a free and open-source interactive charting library.
- Sklearn : Scikit-learn is a Python-based machine learning library that is available for free. in this study various functionalities of Sklearn library have been imported.

SVM (sklearn.svm)and Random Forest Regressor (sklearn.ensemble import RandomForestRegressor) are imported from this library. Performance metrics such as mean absolute error, mean squared error are also evaluated using sklearn library.

- Statsmodels: Statsmodels is a Python tool for exploring data, estimating statistical models, and performing statistical tests. In this study, the adfuller and coint johansen test models were imported from statmodels to check for stationery and non-stationery effects. Vector Autoregression Moving-Average with Exogenous Regressors (VARMAX) is also imported from the statmodels library.

- Pmdarima: Pmdarima is a statistical package that was utilised in this work to obtain an auto arima function that is equivalent to the auto arima functionality in R.

- Tensorflow library: TensorFlow is a machine learning and artificial intelligence software library that is free and open-source. It can be used for a variety of applications, but it focuses on deep neural network training and inference.In this study tensorflow.Keras is used to import layers for the model where, keras is a Python-based deep learning API that runs on top of TensorFlow.

4 Implementation

The steps taken to carry out this research are represented in the diagram below.

- Data is collected and altered as appropriate for the research, and then all of the data is integrated into a csv file.

- During the data visualisation phase of the project, data preprocessing and analysis were taken care of. Models have been implemented after gaining a thorough understanding of the data.

- On the basis of performance metrics, these models were assessed to discover the best fit model.

- Finally, the attention LSTM model was chosen since it has a lower $RMSE$ value. As a result, the attention LSTM model is employed to forecast future values.

Note: Merged and tweaked t he influenzareport.csv file has been uploaded to the jupyter notebook. The identical csv file, together with the code artefacts, will be uploaded and will be instantly accessible.

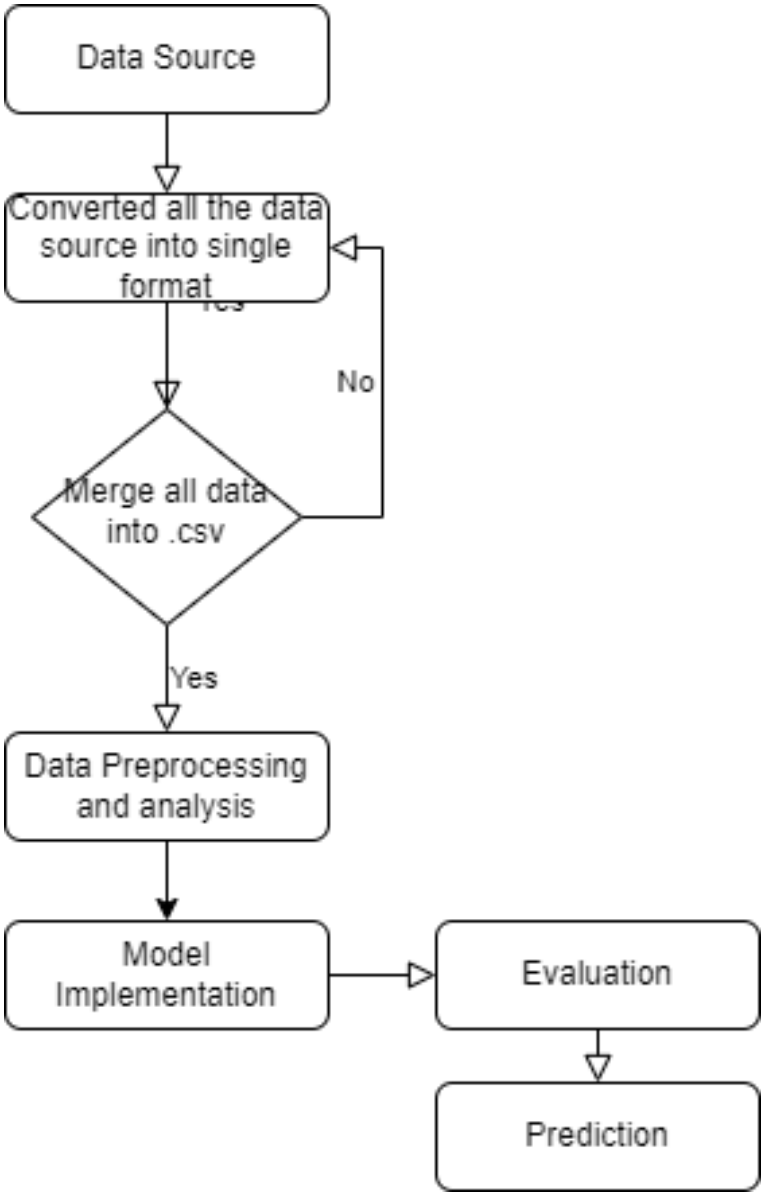


Figure 2: Project flow chart

5 Configuration Details

Following the installation of essential software and libraries, the project's different milestones were carried out. The table below describes various phases of this research as well as the experiments that were performed.

Project Phase	Experiment Carried Out
Phase 1: Data Collection	• Flunet data collected from WHO database • Daily temperature data collected from Power Access NASA • Average monthly weather data (1991-2020) collected from ClimatologyKnowledge portal • Tourist arrival data collected from CEICData website • Population data collected from https://population.un.org/
Phase 2: Modifying data collected	All the collected data converted into single format and merged into a .csv file
Phase 3: Data Pre-processing	• Categorical variable converted into numerical values • Data is converted into Timestep data
Phase 4: Data Analysis	Using various visualization libraries (seaborn, pandas, matplotlib, plotly) data is displayed, and valuable insights are extracted.
Phase 5: Model Implementation	ARIMA, RF, SVM, attention LSTM, ARIMA+LSTM models were implemented
Phase 6: Evaluation	RMSE, MAE and MAPE score were analysed to evaluate the implemented models
Phase 7: Predictions	Using the best fit model, future values were predicted.

6 Conclusion

All of the essential software and hardware are discussed, as well as the procedures to install them. The importance of supporting libraries is also discussed in this document. This configuration manual can be used to implement and thus validate research that has been performed.