Configuration Manual

MSc Research Project
Data Analytics 2023-2024

Abhijit Singh
Student ID: x22157271

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Abhijit Singh
Student ID:	x22157271
Programme:	MSc Data Analytics
Year:	2023-2024
Module:	MSc Research Project
Supervisor:	Jorge Basilio
Submission Due Date:	12/08/2024
Project Title:	Optimizing placement of new EV charging stations in India using machine learning methodology
Word Count:	1357
Page Count:	6

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:	Abhijit Singh
Date:	12/08/2024

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):	

Optimizing placement of new EV charging stations in India using machine learning methodology

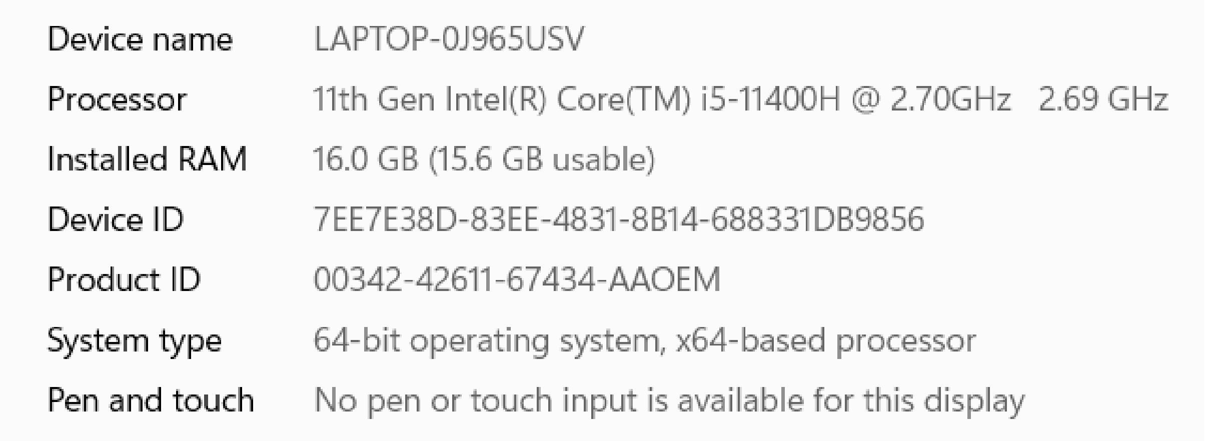
Abhijit Singh
x22157271

1 Introduction

This is the Configuration Manual document that provides a detailed information about the system requirements, environment setup, tools, and algorithms used in the research project on optimizing the placement of new EV charging stations in India using machine learning techniques. The project uses the clustering algorithms to identify optimal locations for the upcoming new EV charging stations based on various socio-economic factors. The manual outlines the process followed during the development phase and documents the final research findings.

2 System Specification

The system specifications that was used for the implementation of this research project are as below:



Device name	LAPTOP-0J965USV
Processor	11th Gen Intel(R) Core(TM) i5-11400H @ 2.70GHz 2.69 GHz
Installed RAM	16.0 GB (15.6 GB usable)
Device ID	7EE7E38D-83EE-4831-8B14-688331DB9856
Product ID	00342-42611-67434-AAOEM
System type	64-bit operating system, x64-based processor
Pen and touch	No pen or touch input is available for this display

Figure 1: System Specification

3 Software Requirements

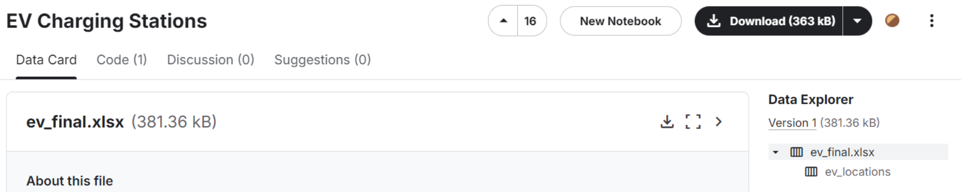
For the analysis and research work many different software, tools and libraries were used to implement machine learning models. The primary programming language used was Python and the development was carried out in a Jupyter Notebook environment using VS code IDE. Below are the key software requirements:

- Python 3.8
- Jupyter Notebook
- VS Code IDE
- Microsoft SQL Server
- SQL Server Integration Services (SSIS)
- Libraries: pandas, numpy, scikit-learn, hdbscan, folium, matplotlib, seaborn, pyodbc

3.1 Data Source

The data used for this research was gathered from multiple public repositories and government databases. The have below link: Links:

- **https://www.kaggle.com/datasets/nezukokamaado/e-v-charging-stations**
- **https://www.kaggle.com/datasets/saketpradhan/electric-vehicle-charging-stations-in-india**
- **https://www.data.gov.in/**
  - **Kaggle**: Provided the existing location data of EV charging stations in India with longitude and latitude details.



EV Charging Stations

ev_final.xlsx (381.36 kB)

About this file

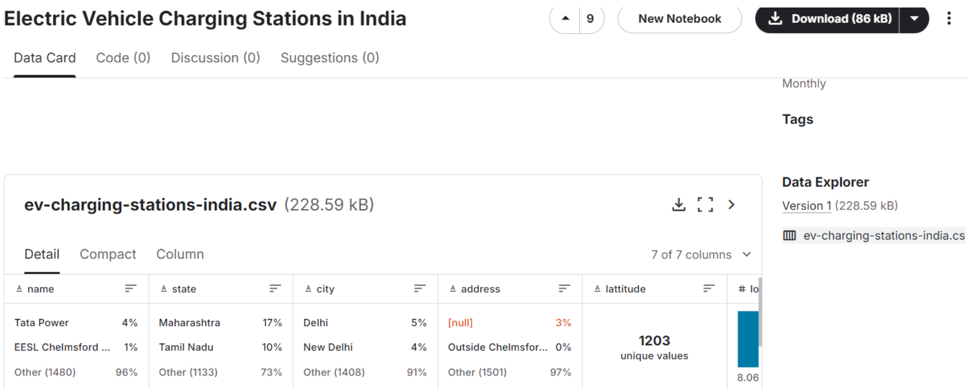
Data Explorer

Version 1 (381.36 kB)

ev_final.xlsx

ev_locations

Figure 2: EV charger data



Electric Vehicle Charging Stations in India

Monthly

Tags

ev-charging-stations-india.csv (228.59 kB)

Data Explorer

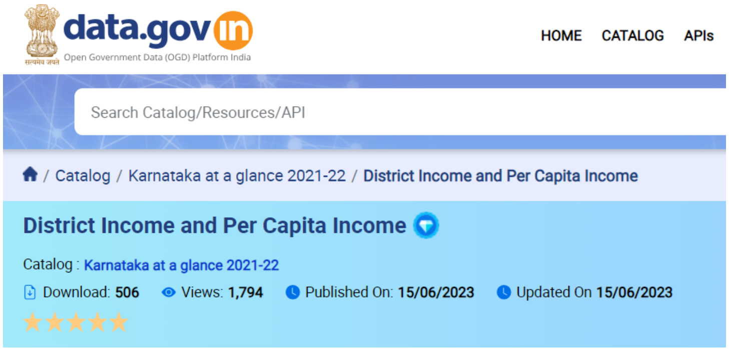
Version 1 (228.59 kB)

ev-charging-stations-india.cs

name		state		city		address		lattitude	# lo
Tata Power	4%	Maharashtra	17%	Delhi	5%	[null]	3%	1203 unique values	8.06
EESL Chelmsford ...	1%	Tamil Nadu	10%	New Delhi	4%	Outside Chelmsfor...	0%		
Other (1480)	96%	Other (1133)	73%	Other (1408)	91%	Other (1501)	97%		

Figure 3: EV charger data

  - **Open Government Data Platform, Government of India**:Government of India portal OGDP provided data on national highways, per capita income, and population density in different region in India for giving a socio-economic point of view in the research .


data.gov.in
Open Government Data (OGD) Platform India
HOME
CATALOG
APIs
Search Catalog/Resources/API

/ Catalog / Karnataka at a glance 2021-22 / District Income and Per Capita Income

District Income and Per Capita Income

Catalog : Karnataka at a glance 2021-22

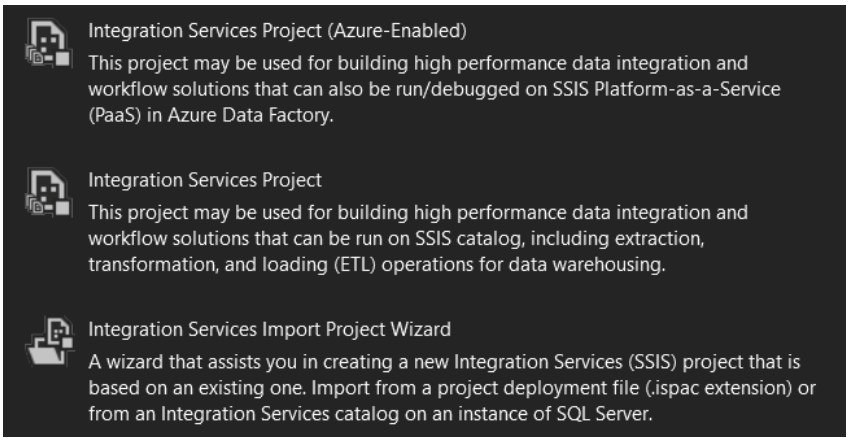
Download: 506 Views: 1,794 Published On: 15/06/2023 Updated On 15/06/2023

Figure 4: Open Government Data

3.2 Data Load and Analysis

The data loading and analysis process involved several key steps including the extract,transform and load step (ETL):

- Data Extraction: SQL Server Integration Services (SSIS) is a tool from Microsoft for flexible warehousing and ETL. It is a free tool and its extension can be downloaded from official microsoft website and after installation can be run with Visual Studio IDE. SSIS was used in the research project to extract data from raw CSV files then transform and load it into the MS SQL data base.
- Data Storage: The SSIS moved the transform data from raw CSV to MS SQL database server after transforming. MS SQL is also free to use from Microsoft and can be downloaded from the microsoft website as well.
- Data Transformation: The transformations of data was done to handle null values, standardize column names, and prepare the data for analysis.



Integration Services Project (Azure-Enabled)
This project may be used for building high performance data integration and workflow solutions that can also be run/debugged on SSIS Platform-as-a-Service (PaaS) in Azure Data Factory.

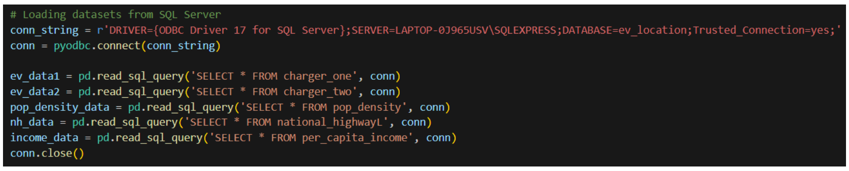
Integration Services Project
This project may be used for building high performance data integration and workflow solutions that can be run on SSIS catalog, including extraction, transformation, and loading (ETL) operations for data warehousing.

Integration Services Import Project Wizard
A wizard that assists you in creating a new Integration Services (SSIS) project that is based on an existing one. Import from a project deployment file (.ispac extension) or from an Integration Services catalog on an instance of SQL Server.

Figure 5: SSIS

3.2.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed to gain insights into the data and identify patterns and know charateristicks of the data. Key aspects of the EDA in the research are:



```
# Loading datasets from SQL Server
conn_string = r'DRIVER={ODBC Driver 17 for SQL Server};SERVER=LAPTOP-0J965USV\SQLEXPRESS;DATABASE=ev_location;Trusted_Connection=yes;'
conn = pyodbc.connect(conn_string)

ev_data1 = pd.read_sql_query('SELECT * FROM charger_one', conn)
ev_data2 = pd.read_sql_query('SELECT * FROM charger_two', conn)
pop_density_data = pd.read_sql_query('SELECT * FROM pop_density', conn)
nh_data = pd.read_sql_query('SELECT * FROM national_highwayL', conn)
income_data = pd.read_sql_query('SELECT * FROM per_capita_income', conn)
conn.close()
```

Figure 6: Connection to SQL

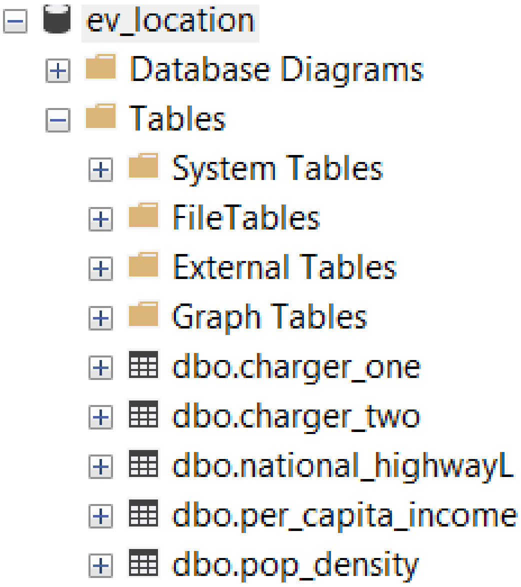


Figure 7: SQL DB Tables

- **Distribution of Charging Stations**: Analyzing the distribution of EV charging stations across different regions in India.
- **Charging Station Types**: Identifying the types of charging stations for example whether the type is conventional charging or battery swapping type.
- **Staffed vs. Unstaffed Stations:** Analyzing the proportion of staffed versus unstaffed charging stations.
- **Power Capacity and Open Duration**: Understanding the power capacity and operational hours of the charging stations.
- **Supported Vehicle Type**: Identifying the types of vehicles supported by the charging stations.

3.3 Data Preprocessing

Data preprocessing involved several steps and libraries to ensure the data was clean and suitable for modeling:

LIBRARIES:

- Pandas: It is a powerful data manipulation library in Python that was used for reading the data from SQL query , data cleaning, integration and feature engineering.

- NumPy: Numpy is a library of python with is used to perform numerical operations in it. In the research project it was utilized for numerical calculation like log transforms for EDA and handling arrays.
- Matplotlib and Seaborn: These libraries in Python that are used for data visualization which helps in the creation of various plots such as histograms, box plots, and line plots to understand the data distribution and trends during the EDA process.
- PyODBC: This is a library that facilitates a connection to the SQL Server database which allows efficient way of data fetching and integration from the database into the Python environment so that it can be further analysed, manipulated and modelled.

STEPS

- **Data Cleaning**: The data was cleaning by handling the missing values, correcting inconsistencies in the data, and standardizing column names.
- **Data Integration**: The data from all sources is merged using the columns of state and city from the multiple datasets.
- **Feature Engineering**: Two of the new feature EV per capita and income per person is created to results of analysis
- **Data Standardization**: The features was scaled to make sure that they contributes equally to clustering process.

3.4 Model Building

Model building was done using the clustering algorithms to identify optimal locations for new EV charging stations:

- Feature Selection: The new features and using the existing ones such as latitude, longitude, population density, and income per person was created
- K-Means Clustering: The K-Means algorithm was applied to divide the data into clusters and identify optimal locations.
- HDBSCAN Clustering: The HDBSCAN algorithm which is a advance clustering method that keeps noise and changing density in consideration is used to apply on the features for location prediction analysis as it considers noise and varying densities.

3.5 Model Evaluation and Comparison

Model performance was checked by using metric like:

- Silhouette Score: It is measure of how similar an object is to its own cluster when comparing to other clusters. The high value close to 1 shows well defined strong clusters.

- Davies-Bouldin Index: Its the measure of the average similarity of each cluster with the cluster most similar to it where a lower value indicate better clustering.

The results of the both K-Means and HDBSCAN clustering models was compared to know what is the most effective approach for identifying optimal locations for new EV charging stations in India.

3.6 Content of the zip file

- SSIS file EV dataflow.sln
- Data Folder
- test2.ipynp
- ev_charging cluster map

  To run the application install SSIS extension and MS SQL . Run the EV dataflow.sln to import data to SQL. Open the test2.ipynb file in Vs Code environment. Install required libraries and run test2.ipynb file.