DE-40 (DC&SE)

NUST COLLEGE OF

ELECTRICAL AND MECHANICAL ENGINEERING

A Generic Data Analytics Framework for Descriptive, Predictive and Prescriptive AI

A PROJECT REPORT

DE-40 (DC&SE)

Submitted by

NS AMNAH SIDDIQUA

NS SAJID MEHMOOD

BACHELORS

IN

COMPUTER ENGINEERING

YEAR

2022

PROJECT SUPERVISOR

DR.BRIG. SHOAB AHMED KHAN

DR. WASI HAIDER BUTT

COLLEGE OF

ELECTRICAL AND MECHANICAL ENGINEERING

PESHAWAR ROAD, RAWALPINDI

YEAR 2022

A Generic Data Analytics Framework for Descriptive, Predictive and Prescriptive AI

A PROJECT REPORT

DEGREE 40

Submitted by

NS AMNAH SIDDIQUA

NS SAJID MEHMOOD

BACHELORS

IN

COMPUTER ENGINEERING

Year

2022

PROJECT SUPERVISOR

DR. BRIG. SHOAB AHMED KHAN

DR. WASI HAIDER BUTT

COLLEGE OF

ELECTRICAL AND MECHANICAL ENGINEERING

PESHAWAR ROAD, RAWALPINDI

DECLARATION

We hereby declare that no portion of the work referred to in this Project Thesis has been submitted in support of an application for any other degree or qualification of this for any other university. If any act of plagiarism found, we are fully responsible for every disciplinary action taken against us depending upon the seriousness of the proven offence.

COPYRIGHT STATEMENT

ACKNOWLEDGEMENTS

Firstly, I would like to thank Allah Almighty for making us capable enough to execute this project as an FYP and help us at every phase of journey Alhamdulillah. It is due to His given strength that we can finish our proposed work.

Secondly, our profound gratitude goes to our supervisors, Dr. Shoab Ahmed khan and Dr. Wasi Haider Butt, for helping us with their professional experience, expertise, knowledge. Their guidance at every step always kept us motivated to move forward and accomplish our desired goals.

Also, heartiest appreciation to our parents and friends, who played key role in uplifting our morals and letting us focus on work, without their support we couldn't have reached this point. Their understanding attitude played vital role throughout our journey; we are eternally thankful to them. Their constant support, care and love made us do more and more when we were unable to do our best.

ABSTRACT

Currently there exist very limited approaches to handle public digital data with-in Pakistan. There is need to monitor public data (let say public health data from hospitals) to generate meaningful insights after performing analysis to make predictions about disease so that prescriptive steps should be taken out to control spread, we have witnessed devastating situations caused by COVID-19. Once this data is collected and analysed properly, we can save many lives. There is lack of data driven decision making in Pakistan. Though now Pakistan has started collecting digital data, but it is important that AI and data analytics tools should be built for decision makers to make effective decisions based on data.

Our proposed framework will in general be capable of continuously analysing data coming from different healthcare facilities to DataMart, using advanced AI and ML approaches primarily we want to do prescriptive AI accompanied by descriptive and predictive AI. Our customers would be ministries from health care for this domain. As it is generic platform, we are extending it to other sectors as well because underlying algorithms are same. So similar engines can be utilized for crime/supply chain analytics. We have taken assumption that data is already compiled from a data mart in csv format.

We shall be licensing it to ministries. They will be paying for license fee only. However, focus is to help developing countries to improve their processes and using IT to face calamities such as spread of diseases. Framework like this can help countries to make better decision for healthcare or related areas. Specially in Pakistan these tools are very critical, we encounter many disasters such as dengue, polio, which has been eliminated from world. So, data driven frameworks are required to monitor and mitigate such occurrences.

Contents

Table of Figures

Chapter 1: Introduction

1.1 Introduction

Data Analytics refers to extracting meaningful insights from raw data. It is process of analyzing and exploring large datasets to make sense of data, these conclusions then used to make predictions and see trends for data-driven decision making. Data Analytics allows to collect, organize, clean, and transform data for deriving meaningful insights. Most companies collect loads of raw data, this data means nothing until data analytics tools are utilized to get intelligible information that results in findings in form of recommendations and suggestion about what a firm should do next. A specific challenge is addressed by data analytics in an organization, it results in finding trend and patterns within relevant data. It ensures that some sense is derived from past, on basis of which future values are predicted and behavior is explored. Informed choices are made based on data. That is why data analytics has become crucial in running a business successfully as it leads to derive profit and growth of an organization.

Our project has made use of three type of analytics i.e., descriptive, predictive, and prescriptive. We have focused on spatial-temporal data having information of location and time of a transaction whether it be a disease, crime, or any other thing. Data analytics is important because it helps ministries optimize their performances and decision making. We can predict spread of diseases or crimes within certain localities then prescriptive measures can be taken out timey to mitigate such occurrences, becoming more innovative and forward-thinking in making logical decisions.[1]

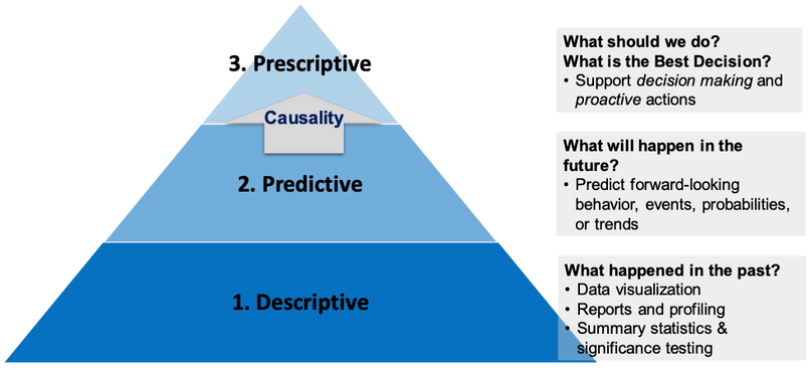

Figure 1:Types of Analytics

1.1.1 Descriptive Analytics

Descriptive Analytics refers to the surface level, it explains what has happened over a period. Historical data is interpreted and occurred changes are observed. It can be done using exploratory data analysis. Usually performed via two techniques that are aggregation and data mining. Aggregation refers to gathering data and summarizing it in presentable format and Mining refers to discovering patterns, it simply explains and determines "What".

1.1.2 Predictive Analytics

Predictive analysis as suggested by name refers to the predictions made taking into consideration trends and patterns visible in data. It can be achieved via predictive models and algorithms to carry out predictions about future outcomes and performances. Previous data is observed and determines if those occurrences are to exist again in near future. It can be utilized to improve already existing policies and reduce risks in relevant area of application. It tells what is likely going to happen soon. Below are the types and names that comes under Predictive Analytics.

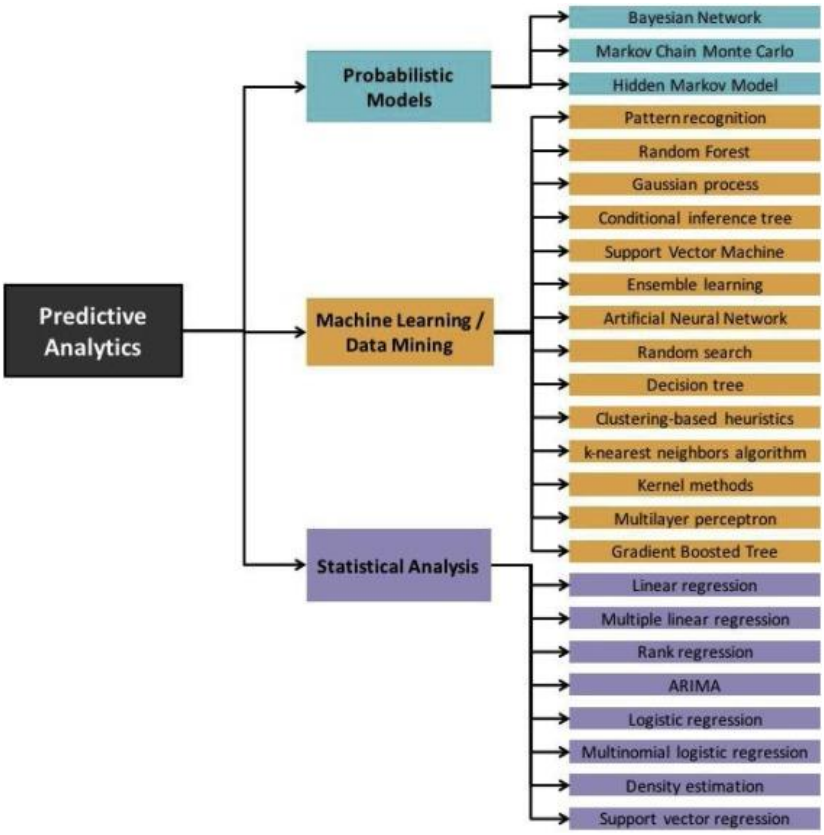

Figure 2: Predictive Analytics

1.1.3 Prescriptive Analytics

Once we are done with predictions, we focus on making prescriptive analysis that helps to make prescriptions based on observed predictions. It suggests measures and course of actions needed to overcome an issue. Let's take example of predictions made regarding increase in spread of certain disease in an area. Prescriptive analytics will help ministries to execute proper measures to stop that spread, whether it be arranging medicines, doctors, and hospitals in affected areas.

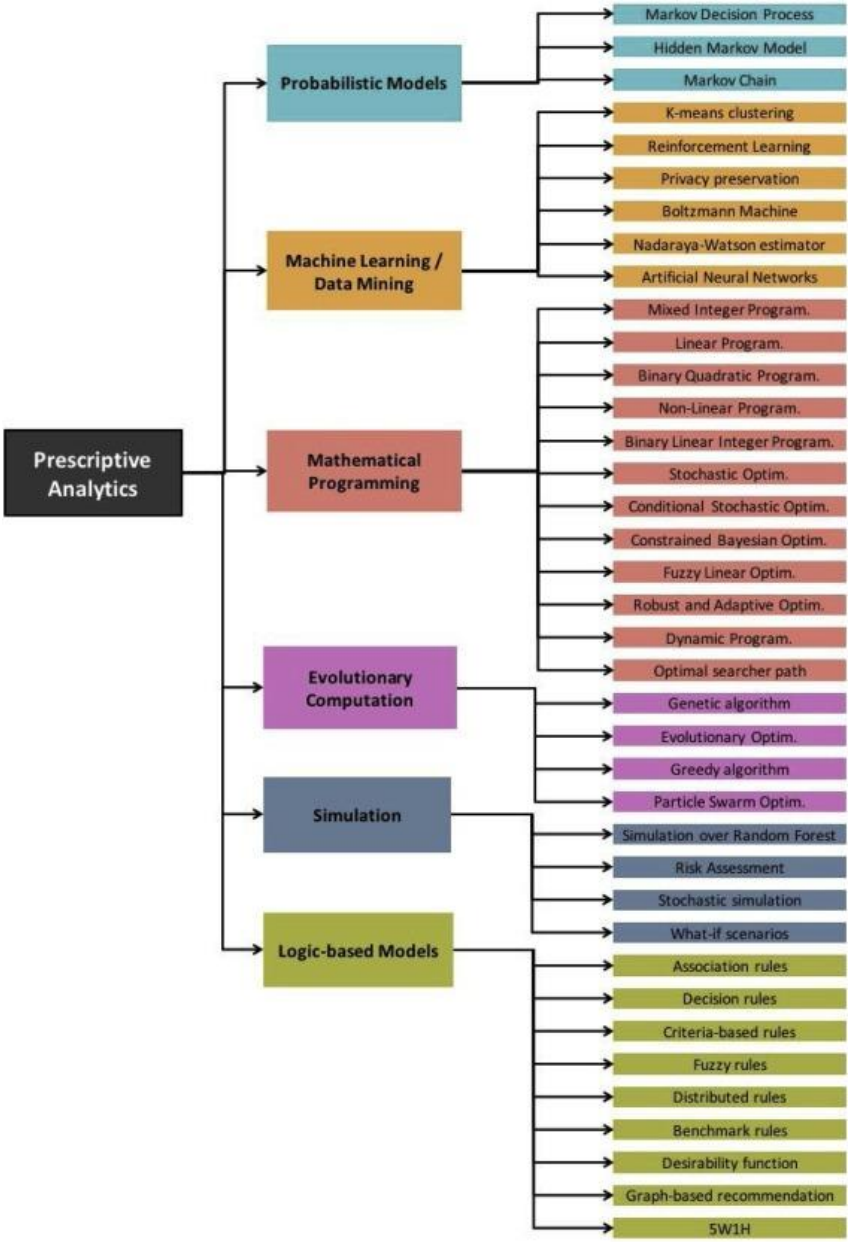

Figure 3: Prescriptive Analytics [2]

1.2 Motivation

Data Analytics can be used to overcome many problems existing in society with innovative and better approaches. Lack of data driven decision making results in fact less and illogical practices may it belong healthcare, crime, and much important aspects of society. It has been observed that Pakistan is deprived of AI based tools for analyzing data. Moreover, user-friendly platform that need no technical expertise regarding data analytics doesn't exist. We aim to develop generic platform that make use of same underlying algorithms for descriptive, predictive, and prescriptive analytics that make use of location and time information of relevant subject of application.[3]

There exists great need of time for Pakistan to lay focus on such tools and proper data collection channels at provincial and national levels. There are so many hurdles, cultural diversity and lack of knowledge about digital data collection and using it at its full potential. Pakistan should consider investing in the infrastructure that promotes data collection, free dissemination, and analysis. Setting up a national data bank as the nation's data archiving and dissemination body could be the first step towards evidence-based planning for economic and social development. Although Pakistan has institutes that has started collecting digital data such as NADRA and Federal Bureau of Statistics, our utilization of that collected data remains limited as data is withheld from researchers and academics.

As mentioned earlier the major problem lies within the fact that no substantiate solution exists at the moment in Pakistan for digital data driven analytics tools. Moreover, the lack of software and due to technological barrier in the country leaves a gap between the digital data collection and usage of such tools. Our aim is to bridge this underlying gap data driven tools and ministries using such tools for shaping better health, crime, disasters, and economics related policies. We have designed a web-based dashboard that will help in creating an innovative and factual decision making by authorities. Some issues in Pakistan related to data collection and analyzing are mentioned below [4]:

- **Frequent Leadership Changes:**

Top leadership play important role in departments. They are responsible for keeping technological transformation working else it goes to waste. As there comes new top

leadership, they don't continue the projects from previous tenure which causes panic, unrealistic demands from the delivery team, and unnecessary urgency.

- **No National Policy on Data Governance:**

There exists unsolved issue of data collaboration even for the internal usage of government departments. For example, for a justice sector, why authentic digital copy of a FIR is not shared to courts or relevant departments? No courts orders are digitally sent to respondents, police, prisons, prosecution, and related departments. The reason is that we do not have a national policy on data governance which clearly shows polices for data storage, data backup, data anonymization, internal/external data sharing protocols, and data monetization. Due to lack of such system, there will always be delays in handling some urgent situation by authorities for both policy making and decision making. Hence services towards nation can be hampered.

- **Lack of Secure and Reliable Digital Access:**

Lastly, there exists great challenge in accessing a secure, faster, and reliable communication network, particularly in remote areas, farms, motorways, and highways. As it is fundamental requirement to adapt any technological advancements. Due to this we will never have timely information regarding on-ground situations that is comprehensive, reliable also. This is critically required for good decision-making at the top.

As we can observe Pakistan is still facing ailments such as dengue and polio that are eliminated from globe. Today globally there is 99% cut down to cases. However, Pakistan due to inefficient technological implementations and lack of analytical tools still unable to get rid of polio. Below are some stats shown regarding cases occurrence in Pakistan [5].

WPV Polio Cases Across Pakistan's Provinces

PROVINCE	2015	2016	2017	2018	2019	2020
PUNJAB	2	0	1	0	12	14
SINDH	12	8	2	1	30	22
KHYBER PAKHTUNKHWA	33	10	1	8	93	22
BALOCHISTAN	7	2	3	3	12	26
GILGIT-BALTISTAN	0	0	1	0	0	0
AZAD JAMMU & KASHMIR	0	0	0	0	0	0
ICT	0	0	0	0	0	0
TOTAL POLIO CASES	54	20	8	12	147	84

Figure 4: Polio cases in Pakistan

Similarly, we as a nation facing many issues and hence unable to come up with effective policies and timely preventive measures. We are interested in making platform that can be used in this regard so that we can overcome barriers.

1.3 Scope

The project`s aim is to develop a generic web-based data analytics application that will enable user to use data belonging from different application areas such as healthcare, crime, and supply chain having space-temporal information of respective incidents. Our proposed work can perform Descriptive, Predictive, Prescriptive Analytics on provided data, making it easier to visualize and analyze recent changes going in society according to relevant domain.

The scope of the project can be defined in terms of the following objectives:

- User friendly interface, a lay man will be capable pf using it.
- Performing Descriptive, Predictive, Prescriptive Analytics and Visualizing outcomes.
- Development of an accretive and interactive dashboard.

Analytics competitors make expert use of statistics and modeling to improve a wide variety of functions. Here are some common applications:

FUNCTION	DESCRIPTION	EXEMPLARS
Supply chain	Simulate and optimize supply chain flows; reduce inventory and stock-outs.	Dell, Wal-Mart, Amazon
Customer selection, loyalty, and service	Identify customers with the greatest profit potential; increase likelihood that they will want the product or service offering; retain their loyalty.	Harrah's, Capital One, Barclays
Pricing	Identify the price that will maximize yield, or profit.	Progressive, Marriott
Human capital	Select the best employees for particular tasks or jobs, at particular compensation levels.	New England Patriots, Oakland A's, Boston Red Sox
Product and service quality	Detect quality problems early and minimize them.	Honda, Intel
Financial performance	Better understand the drivers of financial performance and the effects of nonfinancial factors.	MCI, Verizon
Research and development	Improve quality, efficacy, and, where applicable, safety of products and services.	Novartis, Amazon, Yahoo

Figure 5: Scope [6]

We can observe how broadly the insights from Data Analytics Tools are been utilized for Business Intelligence. We can extend this use in healthcare and crime domains. Pakistan is way behind the world, so our proposed project is just a small step towards implementation of such AI based tools in Pakistan.

1.4 Structure

Following is the structure of the report ahead:

- Chapter 2, it deals with literature Review and will lay light upon existing technologies related to our work.
- Chapter 3, it deals with steps involved to conduct our project and its methodologies.
- Chapter 4, it deals with tools and technologies utilized.
- Chapter 5, it deals with data cleaning and data visualization.
- Chapter 6, it consists implemented algorithms.
- Chapter 7, it deals with future directions and conclusions/Results
- Chapter 8, it contains references used for the purpose of compilation of the project report.

Chapter 2: Literature Review/ Related technologies

This chapter include literature review and will focus existing tools and technologies in field of AI based Data Analytics tools. There are companies that are working in this field and are famous for their work in related area. They have very high profile in this field and have tools with a lot of features. Our project is also inspired by these analytics tools that proved to be very beneficial towards society and data driven decision making.

2.1 Rapid Miner Studio:

Rapid Miner Studio is a powerful tool for data that is capable of performing all required operations such as data mining, model deployment, and model operations. This is an end-to-end data science software and offers all preparations regarding data and further applying machine learning for getting real insights that impact organizations [7].

It provides complete environment that is integrated and it provides services such as Data Analysis, Machine Learning Predictive Analysis, Text Minning, Deep Learning.It is applicable for vast applications and in variety of domainsmay it be educational, research, training, prototyping and application development.

Figure 6:Rapid Miner Logo

It provides unified approach that will enable businesses to boost efficiency, accuracy, and productivity through highly improved standards and accelerated learning. Below we will briefly be looking at features of Rapid Miner [8]:

Features:

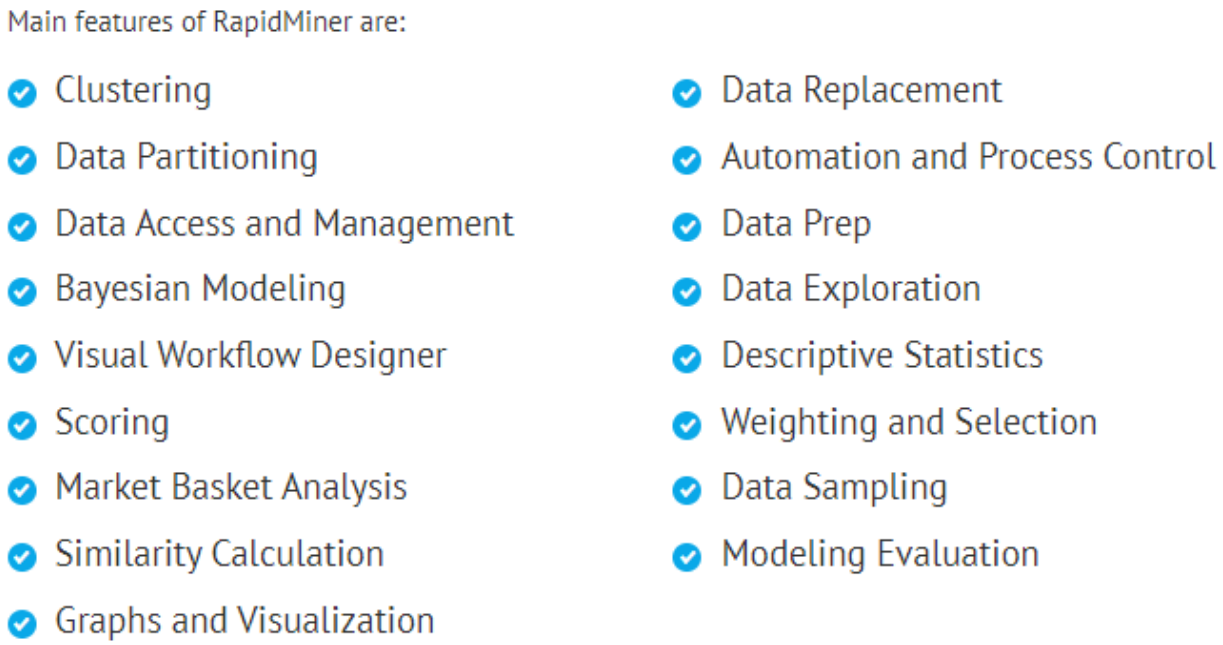

Figure 7:Features of Rapid Minor [8]

Benefits:

- **User-friendly graphical user interface:**

These tools provide very powerful features having capabilities for users with very user-friendly interactive interface which increase the efficiency and workflow of users, keeping them highly interested.

- **Robust features:**

It contains tools which have robust components that are easy to be operated. It gives platform to create, design, and deploy analytics process. It enables its users to have visual presentations. Also, to make and process models.

- **Maximize data usage:**

It has capacity to organize the unstructured, unflustered, disorganized data. For this Rapid Miner provides users ease of data access and management empowering users to handle different types of data like images and texts. Users can deal with data any way they want.

Pricing Plan:

- Free trial/free $0
- Small $2,500/user/year
- Medium $5,000/user/year
- Large $10,000/user/year

Technical Specifications:

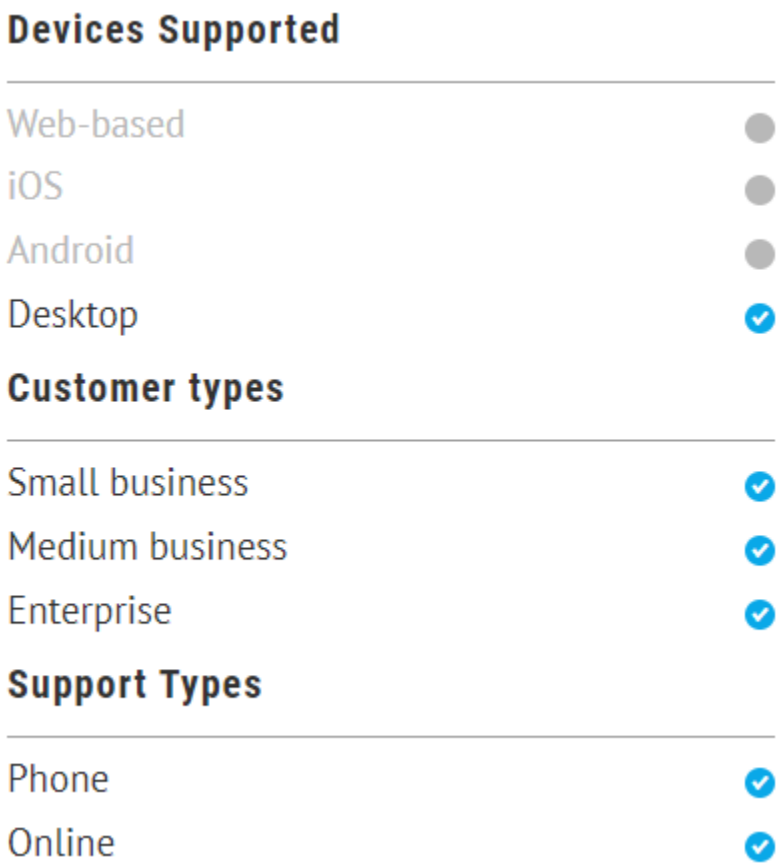

Figure 8:Technical Specs of Rapid Minor [8]

2.2 WEKA:

WEKA (Waikato Environment for Knowledge Analysis) is used for data mining. It comprises of many machine learning algorithms, tools for preparing data, classification, clustering, regression, association rules mining and data visualization. It is introduced by Waikato University, New Zealand. It is applicable in vast domains such as research, education, and projects. It is an open-source software that provides an environment for different learning techniques, evaluation methods and processing tools. It fulfilled the need of unified work bench

which offers researchers state-of the-art Machine Learning techniques. At the time of development in 1992, there exists many different algorithms in variety of languages to be implemented on different platforms [9]. It was most appreciative work to collect all learning schemes and made available on single platform for comparative study. Now it is considered to be a landmark in relevant field of study.

Figure 9: WEKA Logo

WEKA proved to be widely acceptable in academics, business intelligence, and data mining. It is accompanied by a book that is considered as a popular textbook for Machine Learning and Data Mining. Now we will briefly talk about version of WEKA over the time.

Versions:

- Version 3.0 called “book version” and is compatible with description in data mining book and is run on command-line.
- Version 3.2 called “GUI version” provides users with graphical interface for smooth flow of work.
- Version 3.3 called “development version” that comes with many improved features.

Version name	Most recent base number	Associated with book edition
Book 1st ed. version	3.0.x	1st edition
Old GUI version	3.2.x	none
Book 2nd ed. version	3.4.x	2nd edition
Book 3rd ed. version	3.6.x	3rd edition
Book 4th ed. version	3.8.x	4th edition
Development version	3.9.x	none

Figure 10: WEKA Versions

Testimonials:

Here in this part, we will be presenting some of feedbacks or reviews about the WEKA and user experiences that value this product [11].

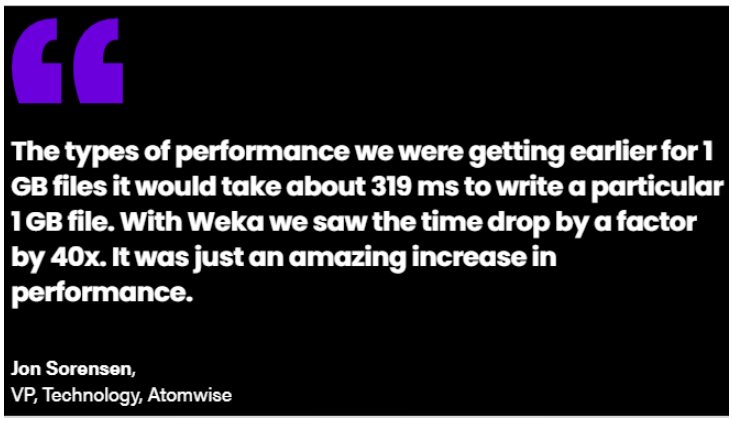

Figure 11: WEKA Testimonial 1

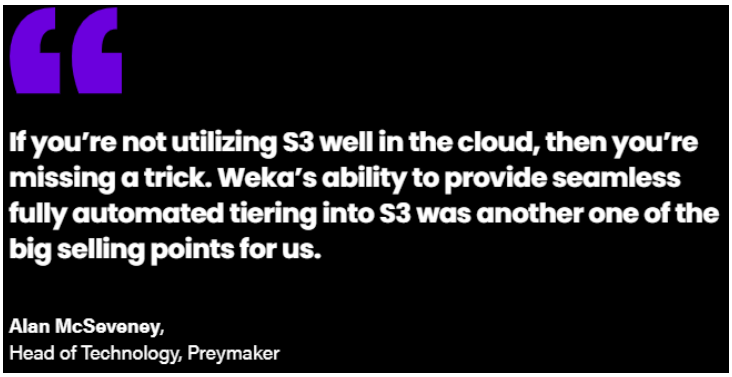

Figure 12: WEKA Testimonial 2

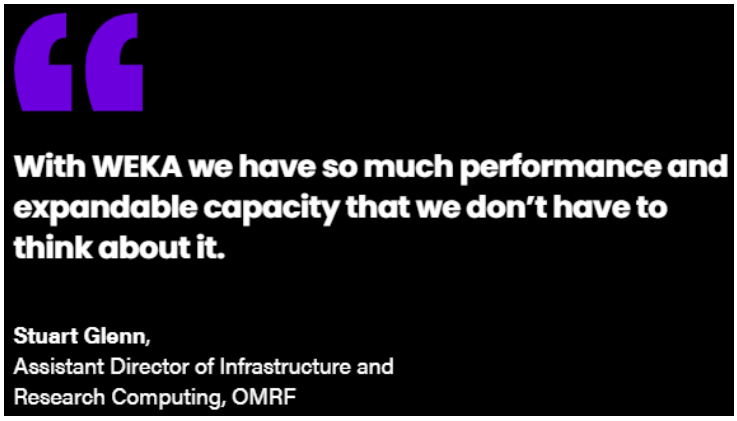

Figure 13: WEKA Testimonial 3

2.3 Power BI:

Power Bi is a Business Intelligence tool. It is collection of software services and applications that convert your data to meaningful insights and visually impressive interactive insights. It is used to discover what your data is about and what can be extracted from data for data driven decision making. It deals with different sort of data types that may be a csv, excel, cloud-based hybrid data warehouse. It enables its users to easily connect their data sources and perform Machine Learning on that. Below we will mention different components that work together [12].

- Power BI Desktop:

  It is a windows desktop application.
- Power BI Services:

  It is an online SAAS (software as a service).
- Mobile Apps:

  It works on mobiles for Windows, iOS, and Android devices.

Figure 14: Power BI Logo

Benefits:

- It provides users with amazing data experience.
- It will make decisions based on facts and data after proper analyzing it.
- It provides security and end-to-end encryption for data.
- Provides user with smart and efficient tools [14].

Pricing Plan:

Here is pricing plan for this product [15]:

- Power BI PRO $13.70/user
- Power BI premium $27.50/user
- Power BI premium $6,858.10/capacity

It is great platform for working, connecting teams, data connectors and it is applicable in vast domain and variety of areas. It offers great user experience, facility to create models and visually enhanced display of results. It gives a complete environment to enhance data management, data acquisition, and end-user delivery experiences [16].

Testimonials:

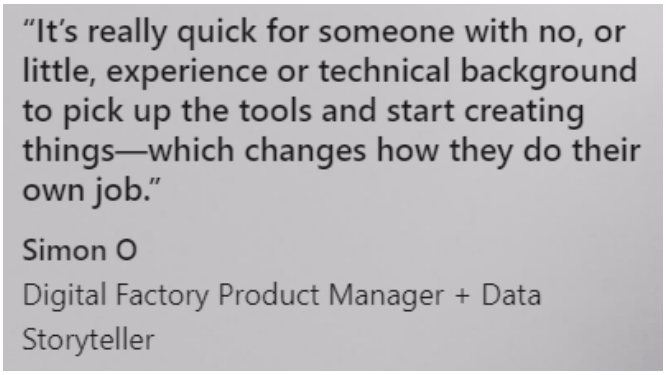

Figure 15: PBI Testimonial 1

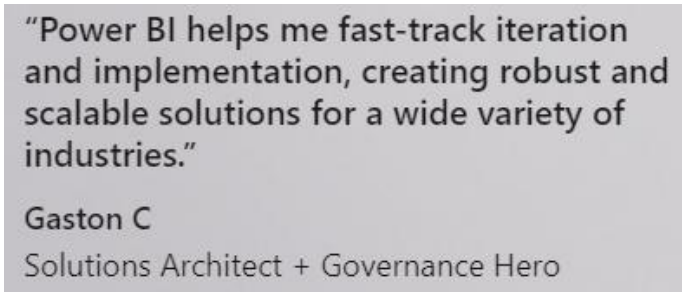

Figure 16: PBI Testimonial 2

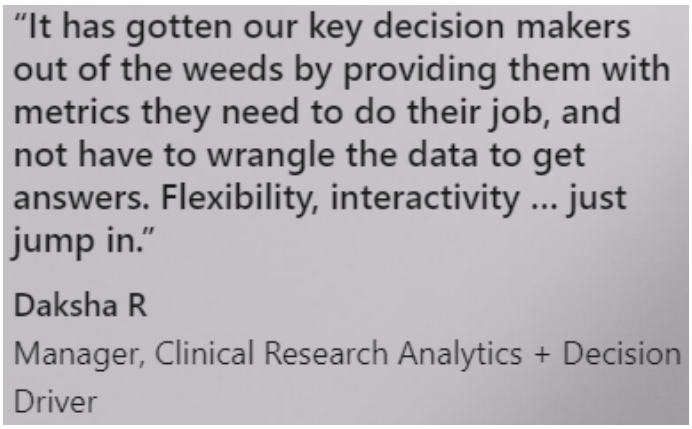

Figure 17: PBI Testimonial 3

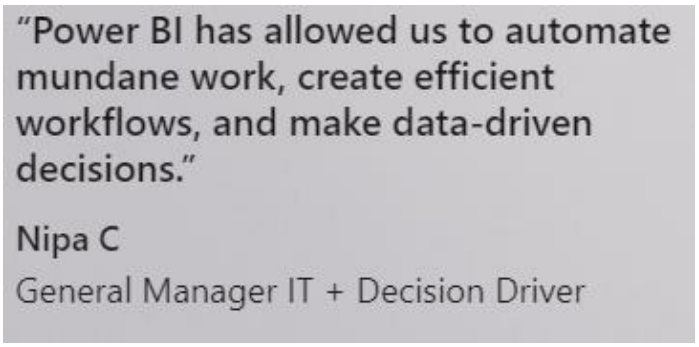

Figure 18: PBI Testimonial 4

Chapter 3: DATA ANALYTICS STEPS/PROCESS

In this Chapter we will be discussing brief overview of step that are followed throughout the project.

3.1 Steps Involved

Below are the steps that we followed in our proposed project that is web-based data analytics dashboard [17].

- **Data Collection:**

The first step associated with data analytics is data collection. There can be many ways to gather data, data can be taken from user in different file formats such as csv, excel or any other source such as establishes data marts, data warehouses, social media, web servers etc.

- **Data Preparation/Cleaning:**

After data is collecting, there comes the next step that is concerned with making good use of data, which is possible only if data is prepared well. For data preparation we focus ion cleaning data, making it sure that data types are in accordance with our use and there are no null values. To make data correct, complete and accurate it is necessary.

- **Data Exploration/Data Visualization:**

Once we are done with data collection and data cleaning, we move ahead towards exploring our data. For this purpose, we visualize our data to see information and patterns it shows. We can use these trends further for making predictions and prescriptions.

- **Analytics Algorithms Implementation:**

Next step is accompanied by the data analytics algorithms. This involves implementation of ML/AI based algorithms. We are focusing on Descriptive, Predictive, and then prescriptive analysis. Then on basis of this we will be predicting future trends for results so that we can take measures timely, and data driven decision making can be enhanced.

- **Results Interpretation:**

To clearly get some meaningful insights from data we must get correct interpretations of results. It is very important to check that if your results are in accordance with expected results.

- **Making web-based dashboard:**

Once we have gone through all steps, there lies an important practice to be followed that is concerned with how to show your work. It is important that your results are displayed in very attractive way. To do so, we have designed a dashboard that provides users with very user-friendly interactive interface having ease of access to different algorithm implementations and viewing their results on single click of button.

There exist several different steps involved in data analytics that are described below [18]:

- Firstly, we focus on data determination. We see how data is organized and grouped. It is important to keep in view data types and categories that it belongs to, it may be gender, income, demographics etc.
- Then comes data collection, which can be performed by different methods and variety of sources that can be online, files, data warehouses, cameras etc.
- After collecting it is necessary to clean up data so that we can make it ready for performing analysis on it. It is made sure that there remain no errors and there exists no Null values. It makes analysis to be performed smoothly.
- After performing analysis, we make our results, predictions, and prescriptions necessary for data driven decision making. Here after getting useful insights, we are done with our analytics.

Here we are assuming that data is collected and is taken from a source, whatever source it is, whether a data mart or by user file upload system We have focused on creation of dashboard that provides with different algorithms to be performed on variety of data having location and time information about occurrences that maybe of diseases, crimes, or anything in locality. Once getting our analytics done, we can further get insights about environmental aspects that may be a reason for spread of disease in a locality and further get information about features of area that may cause spread.

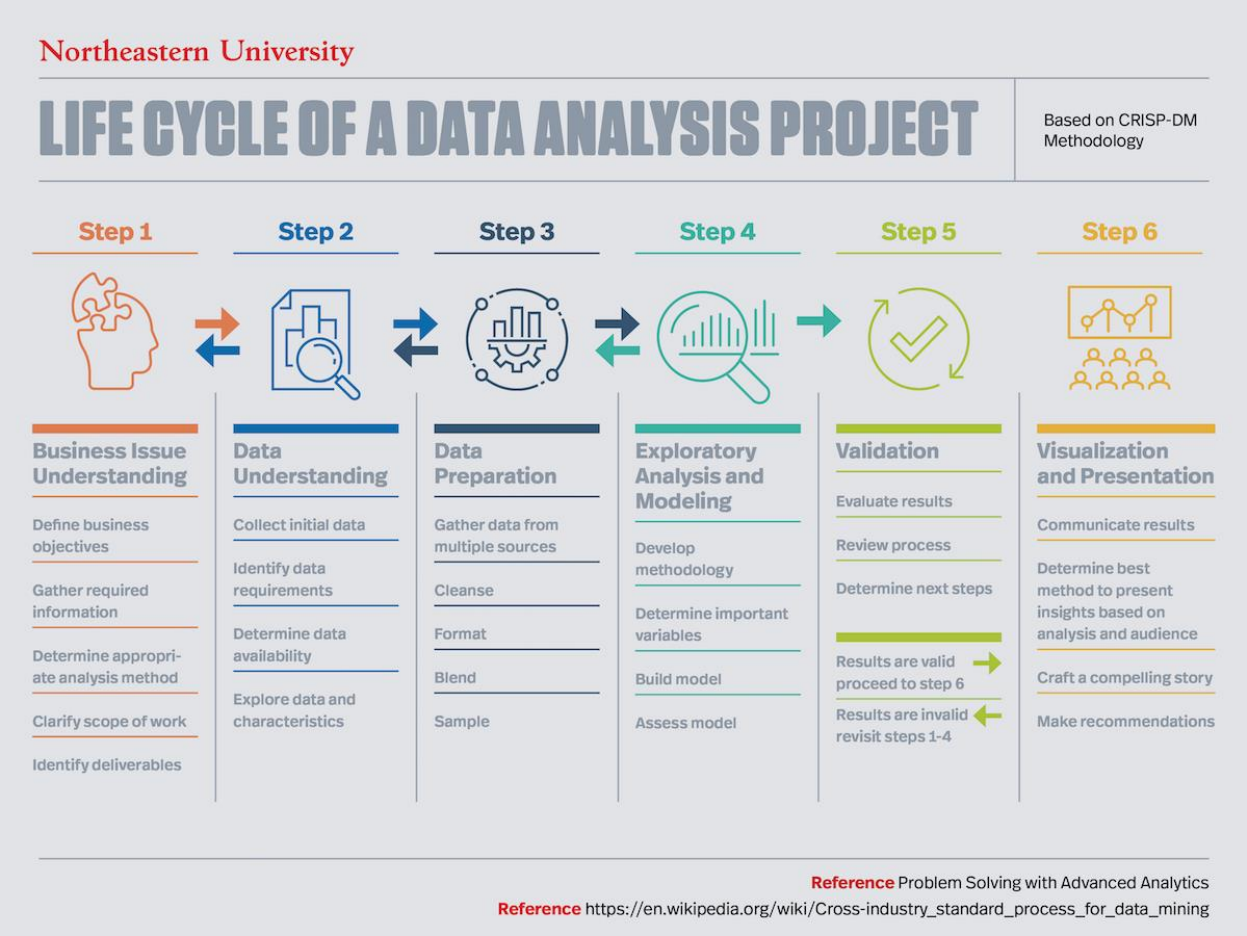

Figure 19: Data Analytics Life Cycle [19]

Chapter 4: TOOLS USED

This chapter deals with the tools and technologies we have used to perform our project. We will be mentioning names in first section, then elaborating them one by one in coming sections of chapter. This chapter will enlighten us about technologies used and why are they chosen to serve the purpose.

4.1 Tools

To accomplish our goals and objectives that are included in our project we need tools and libraries that help us during our working.

- Python
- IDE: PyCharm
- Anaconda
- Django for web
- HTML
- CSS

Libraries used:

We have utilized fully many libraries that could assist us in making our workflow smooth and hindrance free. Some of them are mentioned below:

- Pandas
- NumPy
- Matplotlib
- Folium
- Bokeh
- Django
- Scikit-learn

4.2 Elaboration of tools and Libraries

Python

Python is a very widely used programming language. It is almost used for everything nowadays whether it is Machine learning, website development, automated tasks, or software testing. It is used by both developers and non-developers. It can also conduct data analysis, for this reason we have used it to conduct our project. Python has become staple in data science, data analytics, Machine Learning, data visualization or complex statistical calculations [20].

Figure 20: Python Logo

Python contains many libraries to perform many tasks on its own. That is why we have used python to do data analytics.

Why Python:

- It is very simple and has syntax like natural language.
- It is easy to learn and user friendly.
- It is an open source, which means it is free to use and distribute.
- It is easy to expand capabilities of python.
- It has largest growing and active community.
- It supports code reusability and modularity.

IDE: PyCharm

The IDE Has an editer and compiler we use to write and compile programs.It shpws digferent clors to rigth and wrong keywords.We can diffrentiaire between class and methids.It suggests words that are to be used in code. PyCharm is a hybrid platform by JetBrains as an

IDE for python.Some organizatios such as Twitter, Facebook, Amazon and printerest used pyChram as their Python IDE.It can run on Windows, Kinux, MacOS.

Why PyCharm [21]:

- **Better code navigation:**

  It allows coders to navigate functions, class, or file easily. It makes finding a variable, symbol, element very fast. It enables users to enhance code with less efforts and time.

- **Intelligent Code Editor:**

  It allows coders to write high quality codes that consists of different colors for representing different things such as variables, symbols, class, functions etc. It makes it easy to read code, allows to find errors and identify mistakes faster.

- **Refactoring:**

  Using PyCharm IDE we can make very efficient changes to code in both local and global variables. Refactoring helps enhancing internal structure.

- **Support popular web frameworks for python:**

  Frameworks built for pythons a such as Django are fully supported by autocomplete features that suggest parameters of Django. It also allows debugging efficiently. Also, other frameworks that are web2py and pyramid.

- **Python Scientific Libraries:**

  The main reason to use this is the fact that it supports many built-in libraries that make it easy to perform algorithms associated with data analytics, data science and machine learning. It contains graphs that are interactive and easily understandable.

- **Supports other web technologies:**

  It also supports web technologies such as CSS, HTML, and JavaScript. There is choice for editing live, at same time it can be previewed and updated on web page. It supports NodeJS and AngularJS for web development.

Figure 21: PyCharm Logo

Anaconda

It is open-source distribution for R and python, also applications extend towards data science, Machine Learning, and deep learning etc. It consists of more than 300 libraries to fulfil this purpose, mainly it assists in package managing and development. It comes it carious tools that helps to collect data from different sources using AI and ML algorithms. It offers users with ease of access to management of environment where a user can deploy project with single click [22].

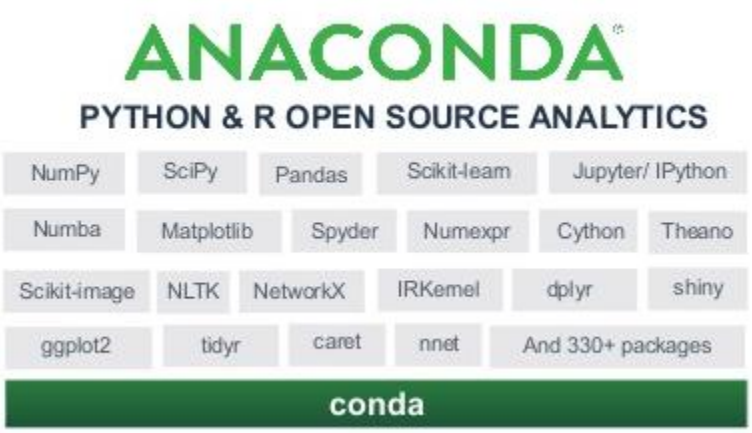

Figure 22: Anaconda

First of All, what's anaconda & Why ought to I trouble regarding it?[23]

Firstly, since anaconda comes with a bunch of data science packages, we'll be set to start out operating with data. Secondly, utilizing conda to manage your packages and environments can scale back future problems coping with the varied libraries you'll be using. In most of the real-world data Science tasks, conda primarily based package and environments are wide used.

We sometimes used conda to form environments for uninflected our projects that use totally different versions of Python and/or different version of packages. we tend to additionally use it to put in, uninstall, and update packages in our project environments. once we download anaconda initial time it comes with conda, Python, and over a hundred and fifty scientific packages and their dependencies. Anaconda may be a large download (~500 MB) because of it comes with the foremost common information science packages in Python, for those that

are conservative regarding space, there's additionally Miniconda, a smaller distribution that features solely conda and Python.

Django

Django is Python framework that creates it easier to make internet sites made on Python. Django takes care of the tough stuff, so you'll think about building your web applications. Django emphasizes reusability of elements, additionally referred to as DRY (Don't Repeat Yourself) and comes with ready-to-use options like login system, database affiliation and CRUD operations (Create read Update Delete) [24].

Django, a free and open-source internet application framework, written in Python. A web framework is a set of elements that helps you to develop websites quicker and easier. When you are building a site, you mostly would like an identical set of components: the simplest way to handle user authentication (signing up, signing in, signing out), a management panel for your web site, forms, the simplest way to upload files, etc. Luckily for you, others way back noticed that web developers face similar issues once building a brand-new website, so they teamed up and created frameworks (Django being one in every of them) that provide you with ready-made elements to use. Frameworks exist to avoid wasting you from having to reinvent the wheel and to assist alleviate several the overhead once you’re building a brand-new website [25].

Figure 23: Django Logo

NumPy

NumPy, a Python library used for operating with arrays. It conjointly has functions for operating in domain of algebra, Fourier transform, and matrices. NumPy stands for Numerical Python.

Why Use NumPy?

In Python we've lists that serve the aim of arrays, however they're slow to method. NumPy aims to supply an array object that's up to 50x quicker than ancient Python lists. The array object in NumPy is termed ND array, it provides a great deal of supporting functions that build operating with ND array extremely straightforward. Arrays are very often employed in data science, wherever speed and resources are vital [26].

Figure 24: NumPy Logo

Pandas

The Pandas is associated as open-source Python package that's most generally used for data science/data analysis and machine learning tasks. it's designed on over of another package named NumPy, that provides support for multi-dimensional arrays. It has functions for analyzing, cleaning, exploring, and manipulating knowledge.

Why Use Pandas? Pandas permits us to investigate massive data and build conclusions supported by applied math theories. Pandas will clean untidy data sets and build them legible and relevant. Relevant knowledge is extremely vital in data science.

What will Pandas Do? Pandas makes it easy to try and do several of the time intense, repetitive tasks related to operating with data, including Data cleansing, Data fill, Data normalization, Merges and joins, Data visualization, Statistical analysis, Data inspection Loading and saving data. And much a lot of in fact, with Pandas, you'll be able to do everything that creates world-leading data scientists vote Pandas because the best data analysis and manipulation tool obtainable [27].

Figure 25: Pandas Logo

Matplotlib

Matplotlib is a very popular cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy. It is one amongst the foremost powerful plotting libraries in Python. It is a cross-platform library that has numerous tools to make 2D plots from the data in lists or arrays in python. It provides a user to visualize data employing a type of differing kinds of plots to form data comprehensible . It enables us to use, these differing kinds of plots (scatterplots, histograms, bar charts, errorcharts, boxplots, etc.) by writing few lines of code in python.You can use the matplotlib package in any Python shell, IPython shell, Jupyter notebook, jupyter lab, cloud (IBM Watson studio, Google collab, etc), and web application servers (flask, Django utilizing pycharm or anaconda).Matplotlib is open source so that users use it freely [28].

Figure 26: Matplotlib Logo

Folium

Folium is considered to be a very powerful Python library that helps you produce many varieties of Leaflet maps. By default, folium creates a map in a separate hypertext mark-up language file. Since folium results are interactive, this library is extremely helpful for dashboard building. Folium allows you to come up with a base map of specific breadth and height with either default tilesets (i.e., map styles) or a custom tileset URL [29].

Figure 27: Folium Logo

Bokeh

Bokeh is considered to be widely used data visualization library in Python that gives superior interactive charts and plots. Bokeh output which are obtained in numerous mediums like notebook, markup language and server. it's possible to implant bokeh plots in Django and flask apps [30].

Figure 28: Bokeh Logo

Scikit-learn

Scikit-learn (sclera), most helpful and strong library for machine learning in Python. It provides a range of efficient tools for machine learning and applied math modeling together with classification, regression, clustering, and spatial property reduction via a consistence interface in Python. This library, that is basically written in Python, is constructed upon NumPy, SciPy and Matplotlib [31].

Figure 29: SciKit Learn Logo

HTML

- HTML stands for Hyper Text Markup Language
- HTML is the standard markup language for creating Web pages
- HTML describes the structure of a Web page
- HTML consists of a series of elements
- HTML elements tell the browser how to display the content
- HTML elements label pieces of content such as "this is a heading", "this is a paragraph", "this is a link", etc [32].

Figure 30: HTML Logo

CSS

- CSS stands for Cascading Style Sheets
- CSS describes how HTML elements are to be displayed on screen, paper, or in other media
- CSS saves a lot of work. It can control the layout of multiple web pages all at once
- External stylesheets are stored in CSS files

Figure 31: CSS Logo

Chapter 5: Data Cleaning and Data Visualization

This chapter covers all aspects of data cleaning and data visualizations. It is very necessary to have great interactive user-friendly data visualization to understand the data in better way and retrieve best outcomes from analytics.

5.1 Data Cleaning

Data cleanup is that the method of editing, correcting, and structuring data within a data set in order that it's usually uniform and ready for analysis. This includes removing corrupt or extraneous data and formatting it into a language that computers will perceive for optimum analysis [33].

There is very important expression in data analysis: "Garbage in, garbage out," which implies that, if you begin with unhealthy data (garbage), you'll solely get "garbage" results. Data cleanup is commonly a tedious method, however it's completely essential to induce prime results and powerful insights from your knowledge. So, it's vital that we perform proper data cleaning to make sure the best possible results [34].

In machine learning, data scientists agree that accurate data is even a lot of vital than the foremost powerful algorithms. This is often as a result of machine learning models only perform as well as the data they're trained on. If model is trained with wrong dataset, the final analysis results won't only be unreliable, however often typically be fully harmful to your organization. Proper data cleanup can save time and cash and create your organization a lot of economical, assist you higher target distinct markets and teams, and permit you to use a similar dataset for multiple analyses and downstream functions.

Following are the steps followed to get our data cleaned and preprocessed to get clear results finally.

- **Dealing with Inconsistent column:**

When DataFrame contains columns that are irrelevant, or if they are never used then these columns can be dropped so that more focus is laid on columns. Here is python code line to perform this in python

```
df=df.drop(columns='Column_Name')
```

- **Handling missing data**:

It is very common that we get missing values in real world data. While dealing with real world data, we get to observe that there occur many missing values. Removing such values is essential for health of our algorithms implemented. So, you must ensure that there exist no null values and they must be dealt properly if they exist.

```
df = df.dropna()
```

- **Handling Duplicate rows:**

There occur many duplicate rows within dataset, so it is vital to remove repeating rows. It is most easy to remove the rows which are similar using a single piece of code line in python.

```
dataset_name.drop_duplicates()
```

- **Dealing with conversion of data types:**

Data can be in many types; some are written below:

  - Categorical data
  - Object data
  - Numeric data
  - Boolean data

Within dataset we can face inconsistent data types and values can be changed due to some reasons. To convert from one data type to another we use pandas. DataFrame_astype.Below is code of python to do this task [35]:

```
for i in range(len(df['lat'])):
    try:
        float(df['lat'][i])
    except:
        df['lat'][i] = None
```

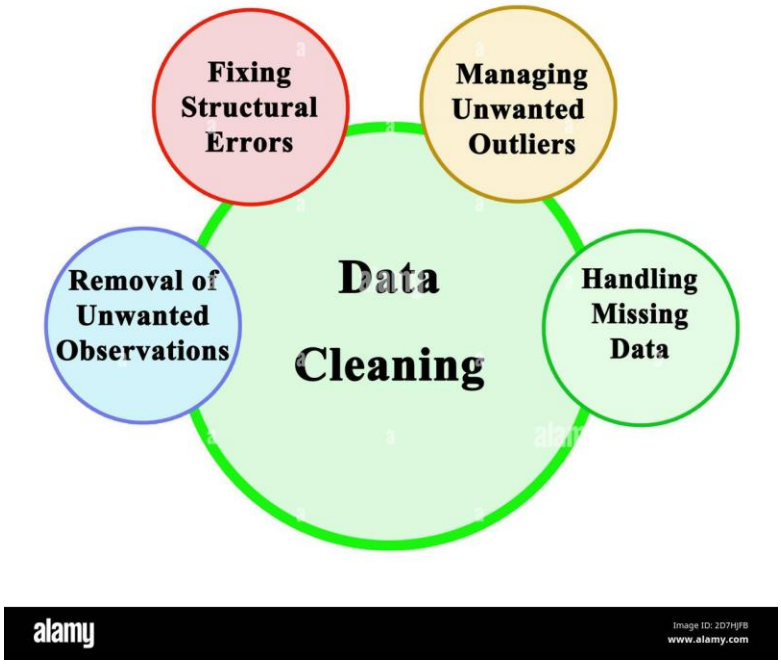

Figure 32: Data Cleaning [36]

5.2 Data Visualization

The Data visualization is referred as graphical illustration of information and data. By using visual components like charts, graphs, and maps, data visualization tools give accessible methods to see and perceive trends, outliers, and patterns in information. In the world of data Analytics, information visualization tools and technologies are essential to research large amounts of knowledge and build data-driven choices [37].

Human eyes are drawn to colours and patterns. we are able to quickly determine red from blue, square from circle. Our culture is visual, as well as everything from art and advertisements to TV and films. data visualization is another kind of visual art that grabs our interest and keeps our eyes on the message. once we see a chart, we have a tendency to quickly see trends and outliers. If we are able to see something, we have a tendency to internalize it quickly. It's storytelling with a purpose. If you've ever stared at a huge spreadsheet of data and are unable to see a trend, you recognize how much more effective a visualization may be.

A good visualization removes noise, highlights the useful information and it tells a story. Below we will see some of common techniques that are followed to visualize the data.

- Charts

- Tables
- Graphs
- Maps
- Infographics
- Dashboards

Data visualization tools:

There are dozens of tools that can be used for data visualization and data analysis. These vary from straightforward to complicated, from intuitive to obtuse. Not each tool is correct for each person wanting to find out visualization techniques, and not each tool will scale to business or enterprise functions. However, we used Matplotlib, Bokeh and Folium in our project explained below are the results we get from these tools on our project.

- **Matplotlib**

We have used Matplotlib. Pyplot module for data visualization in Matplotlib. Pyplot is a Matplotlib module that has a MATLAB-like interface. Matplotlib is intended to be as usable as MATLAB, with the power to use Python and the advantage of being free and open source. every Pyplot operate makes some amendment to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc. the assorted plots we will be able to utilize Pyplot are Line Plot, Histogram, Image, Contour, Polar Scatter, and 3D Plot [38].

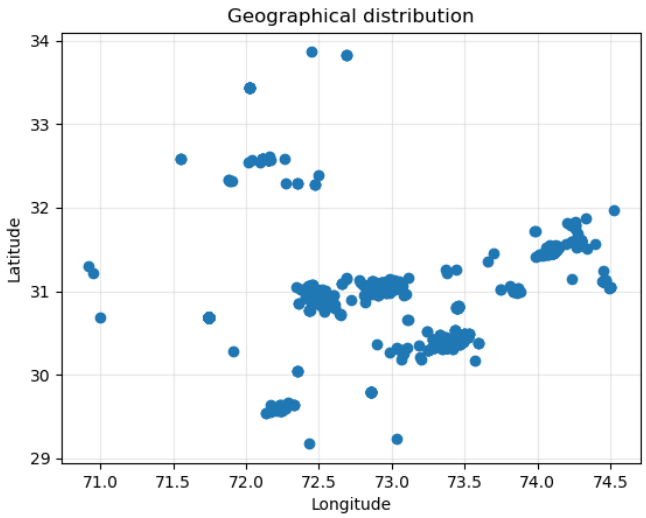

Figure 33: Plot using Matplotlib

- **Bokeh**

Bokeh offers it users with two visualization interfaces that are mentioned below [39]:

bokeh. Models: An interface that is low level but provides high flexibility to application developers.

bokeh. Plotting: An interface that is high level and is for creating visual glyphs.

Bokeh plots are interactive, provides great user experience and better understanding towards data. Here are the results we have obtained from our project using this tool. Our plot using bokeh can be observed below showing relevant information to our project.

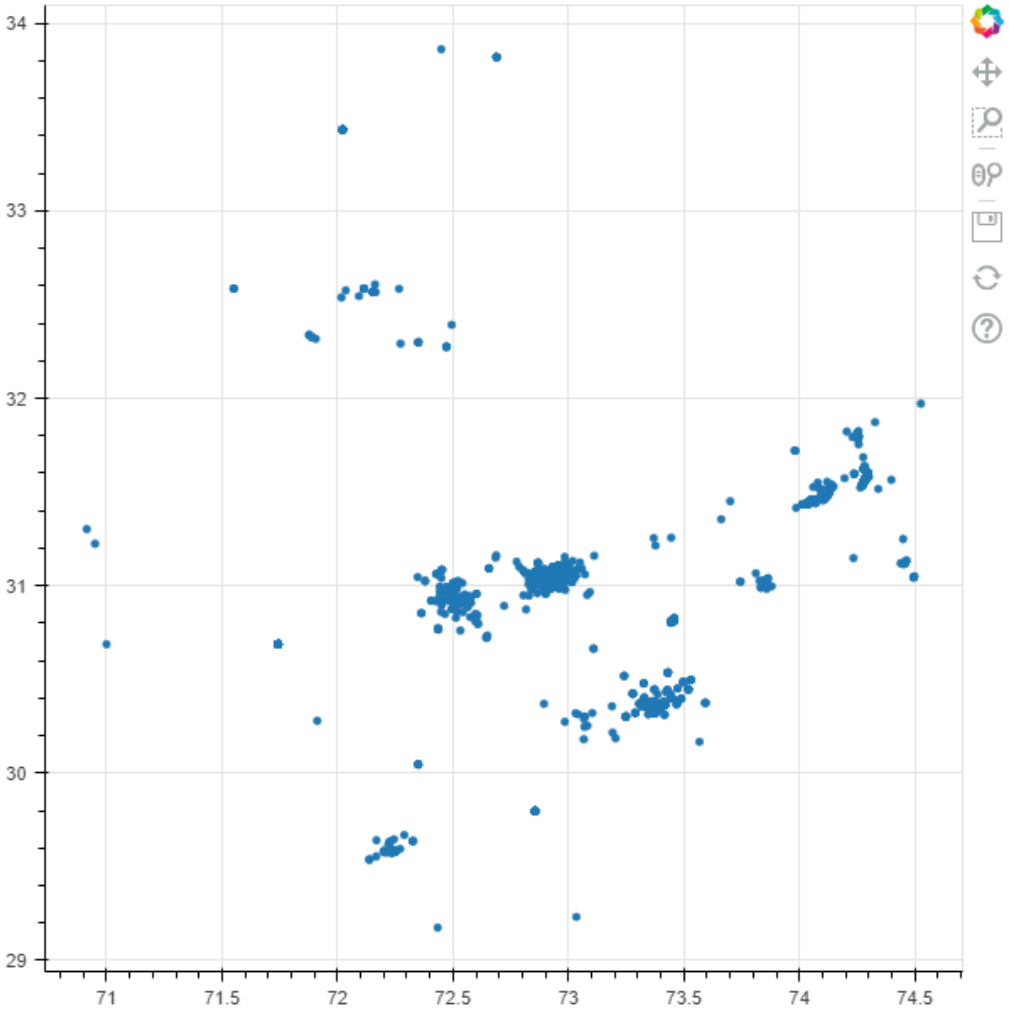

Figure 34: Plot using Bokeh

- **Folium**

Data that's been manipulated in Python is easily visualized by Folium, on an interactive Leaflet map. This library has several built-in tilesets from OpenStreetMap, Map box etc.

Here is data visualization result of dataset that is cholera disease dataset from Punjab, in Folium [40]:

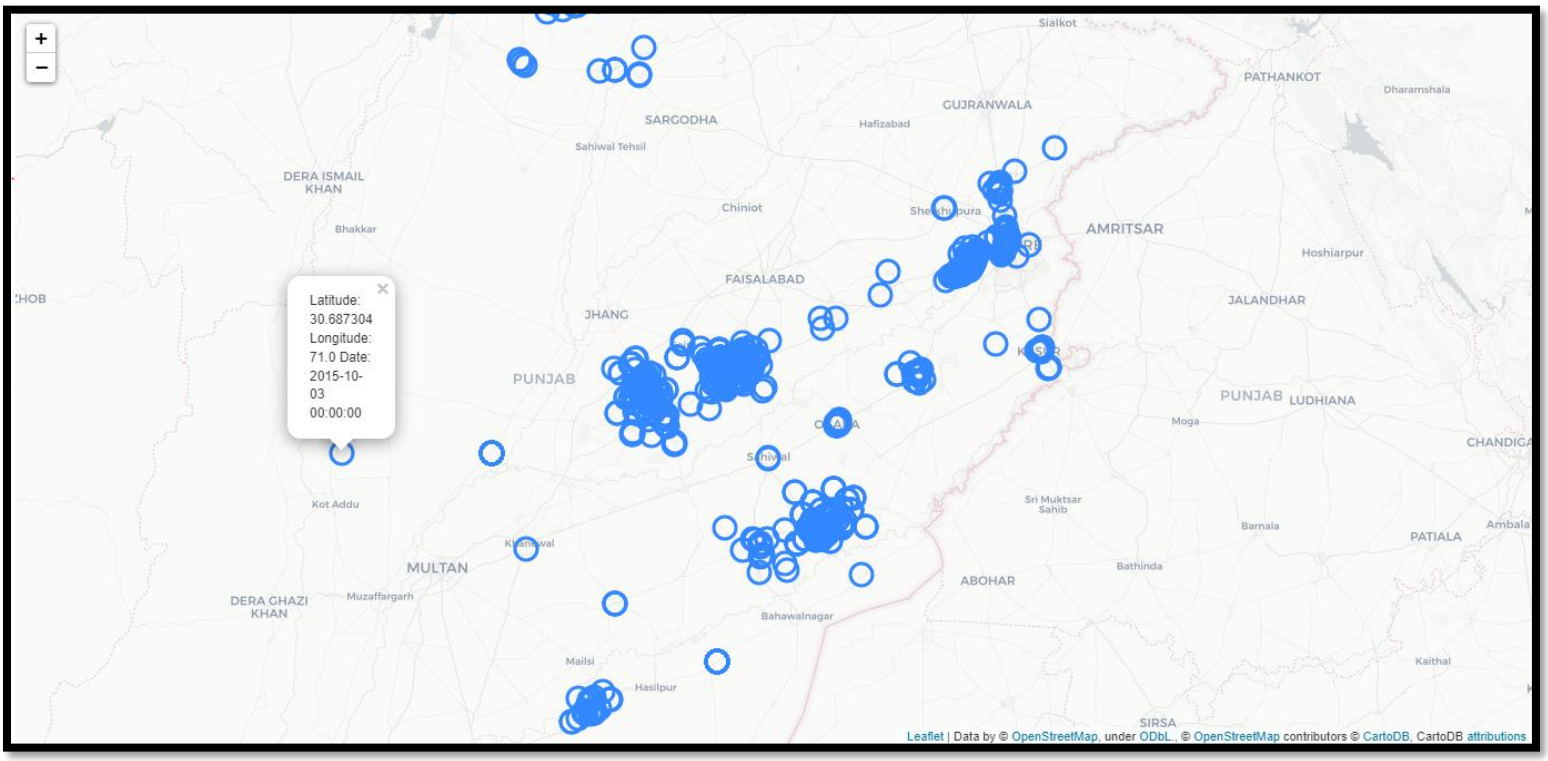

Figure 35: Plot using Folium

Chapter 6: Algorithms Implemented

This chapter puts light on the algorithms we have implemented in our project. There are many algorithms that are associated with predictive, prescriptive, and descriptive analytics. However, we haven't performed all these due to limited time.

6.1 Time series Analysis

The Time series analysis is a method utilized for analyzing time series data to obtain usefull statistical information from the data

Let's have a look into our dataset, it contains three columns namely lat for latitude, long for longitude and date for date.

	A	B	C
1	lat	long	date
2	30.99652	73.87993	07/13/2015
3	31.02123	73.74298	07/14/2015
4		72.89407	07/13/2015
5	32.27616	72.47222	07/14/2015
6	32.27616	72.47222	07/14/2015
7		73.07217	07/13/2015
8	31.02131	73.74595	07/14/2015
9	31.05249	72.93935	07/15/2015
10	31.08543	72.89678	07/15/2015
11	31.01748	72.89815	07/15/2015
12	29.63747	72.32648	07/14/2015
13	31.06573	72.92047	07/15/2015
14	29.57941	72.24034	07/14/2015
15	31.06514	72.98192	07/15/2015
16	29.64148	72.16895	07/14/2015
17	29.58251	72.2389	07/14/2015
18	29.57449	72.20927	07/14/2015
19	29.57968	72.23872	07/16/2015
20	29.63495	72.23	07/14/2015
21	31.35452	73.66142	07/15/2015
22	33.43345	72.02223	07/15/2015
23	29.58261	72.24001	07/16/2015

We performed time series analysis on our data using python code that is given below and then results are shown after plotting grouped_df in matplotlib next to code:

```
data_f = pd.DataFrame([])

data_f.insert(0, "Date", date)

grouped_df =
data_f.groupby(['Date']).size().reset_index(name="Count")
```

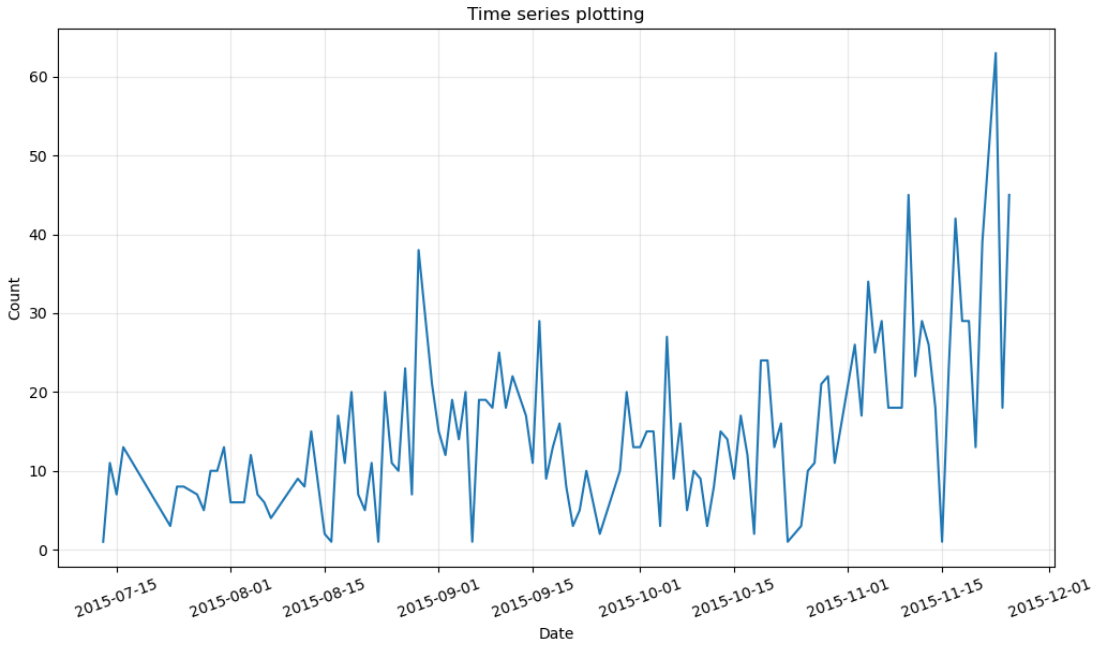

Figure 36: Time series Analysis

6.2 Cluster Analysis

Cluster analysis or cluster is the task of clustering a collection of objects in such how that object within the same group (called a cluster) are a lot of similar (in some sense) to every different to those in remaining teams (clusters). Thus, we can refer clusters to a collection of data points aggregated together because of certain similarities. There are many types of clustering analysis, we have performed the following techniques on our dataset [41].

- **KMeans**

K-means bunch is one of the simplest and standard unsupervised machine learning algorithms.

How the K-means algorithm works:

To process the learning data, the K-means algorithmic program in data mining starts with a primary cluster of at random elect centroids, which are used as the starting points for each

cluster, so performs iterative (repetitive) calculations to optimize the positions of the centroids. It halts making and optimizing clusters when either:

- The centroids have stabilized — there's no modification in their values as a result of the clustering has been successful.

- The defined number of iterations has been achieved.

Below are the results we have obtained on our dataset after performing KMeans. We have used sklearn.cluster to perform the KMeans algorithm on our dataset [42].

code and figure:

```
from sklearn.cluster import KMeans

df = pd.read_csv('fyp.csv')

X = df.iloc[:, [0,1]].values

kmeans = KMeans(n_clusters=i, init='random',
random_state=42).fit(X)
```

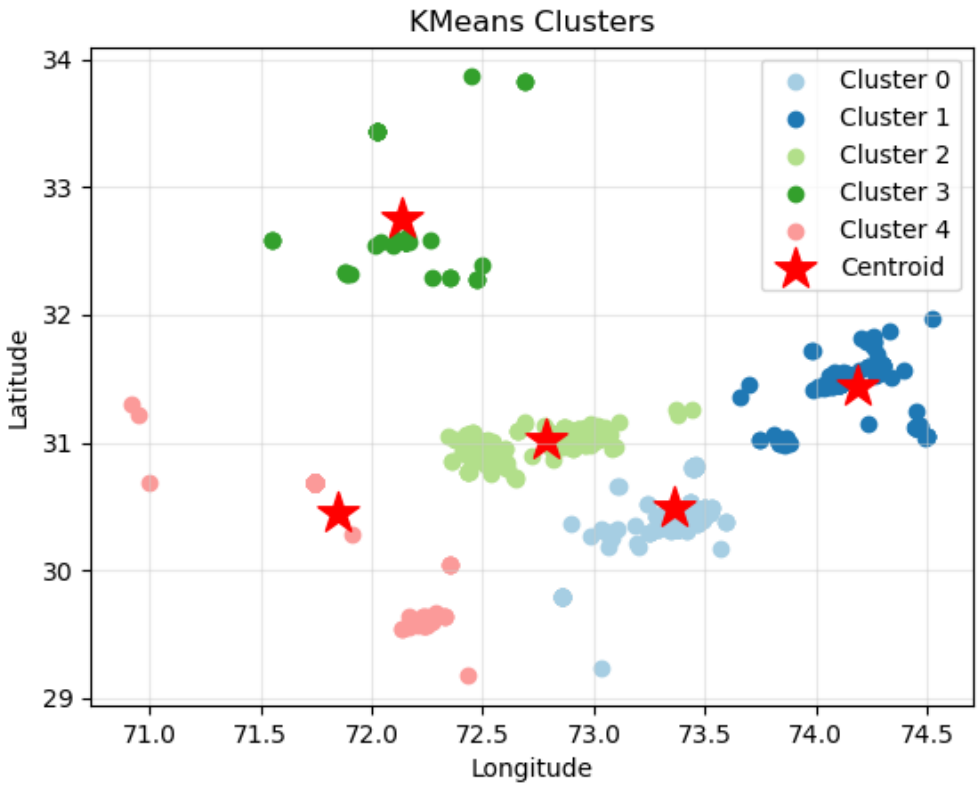

Figure 37: KMeans Plot

- **DBSCAN**

Clusters are dense regions within the data space, separated by regions of the lower density of points. The DBSCAN algorithmic program relies on this intuitive notion of “clusters” and “noise”. The key plan is that for every point of a cluster, the neighborhood of a given radius must contain at least a minimum variety of points.

DBSCAN algorithm needs two parameters:

- ✓ **eps:** It defines the neighborhood around a data point i.e., if the gap between 2 points is lower or adequate to ‘eps’ then they're thought of as neighbors.
- ✓ **MinPts:** Minimum number of neighbors (data points) inside eps radius. Larger the dataset, the larger value of MinPts should be chosen.

Below are the results we have obtained on our dataset after performing DBSCAN. We have used sklearn.cluster to perform the DBSCAN algorithm on our dataset and different colors here represent different cluster based on dense region [43].

code:

```
from sklearn.cluster import DBSCAN
from sklearn import preprocessing as p
df = pd.read_csv('fyp.csv')
dbscan_data = df[['long','lat']]
dbscan_data = dbscan_data.values.astype('float32', copy=False)

dbscan_data_scaler = p.StandardScaler().fit(dbscan_data)
dbscan_data=dbscan_data_scaler.transform(dbscan_data)
model=DBSCAN(eps=0.2,min_samples=20,metric="euclidean").fit(dbscan_data)
```

Figure:

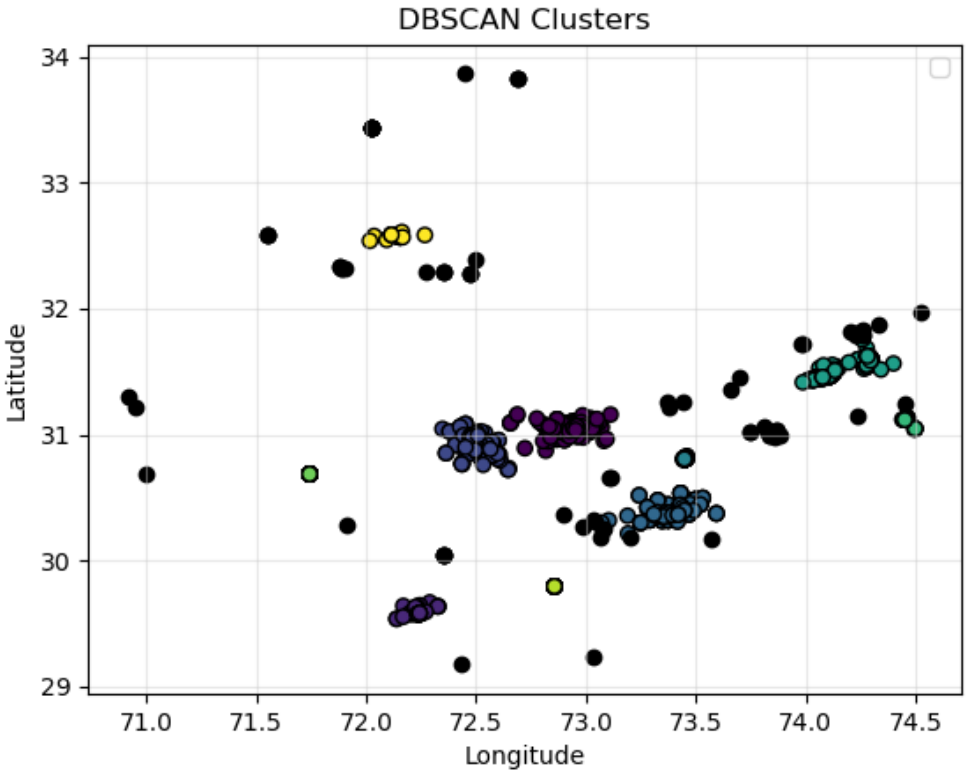

Figure 38:DBSCAN Plot

- **OPTICS**

Ordering points to spot the cluster structure (OPTICS) is associate algorithmic program for locating density-based clusters in special data. Its basic plan is analogous to DBSCAN; however, it addresses one of DBSCAN's major weaknesses: the matter of detection meaningful clusters in data of varied density. To do so, the points of the database are (linearly) ordered such that spatially nearest points become neighbors within the ordering. Moreover, a special distance is stored for every point that represents the density that must be accepted for a cluster, so each point belongs to the same cluster [44].

Below are the results we have obtained on our dataset after performing OPTICS. We have used sklearn.cluster to perform the OPTICS algorithm.

Code and Figure:

```
from sklearn.cluster import OPTICS

from sklearn.preprocessing import normalize, StandardScaler

df = pd.read_csv('fyp.csv')

df = df.drop("date",axis=1,)

df = df.dropna()

scaler = StandardScaler()

df_scaled = scaler.fit_transform(df)

df_normalized = normalize(df_scaled)

# Converting the numpy array into a pandas DataFrame

df_normalized = pd.DataFrame(df_normalized)

# Renaming the columns

df_normalized.columns = df.columns

optics_model = OPTICS(min_samples = 10, xi = 0.05, min_cluster_size =
0.05).fit(df_normalized)
```

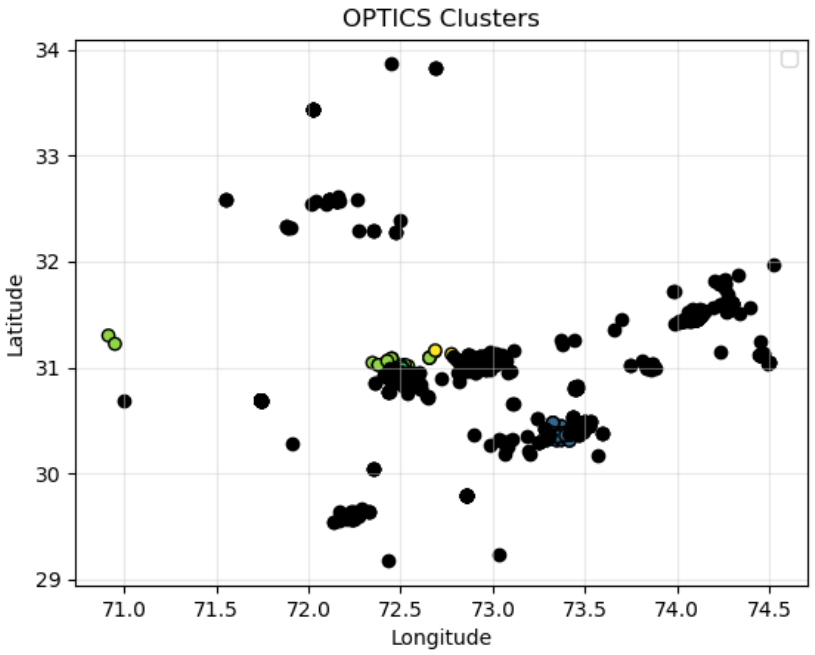

Figure 39: Optics Clustering

Each algorithm offers a different approach to the challenge of discovering natural groups in data. There is no best clustering algorithm, and no easy way to find the best algorithm for your data without using controlled experiments.

6.3 Statistical models

When it involves time series prediction using statistical models, there are quite few well-liked and well-accepted algorithms. each of them has different mathematical modalities and that they escort a distinct set of assumptions that must be satisfied. We will not go in-depth on the mathematical ideas, can simply offer an intuition that you simply can hopefully find useful. We have implemented the following algorithms, we have divided our data in test and train represented by red and blue colors respectively, we will predict test values and then compare the accuracy [45].

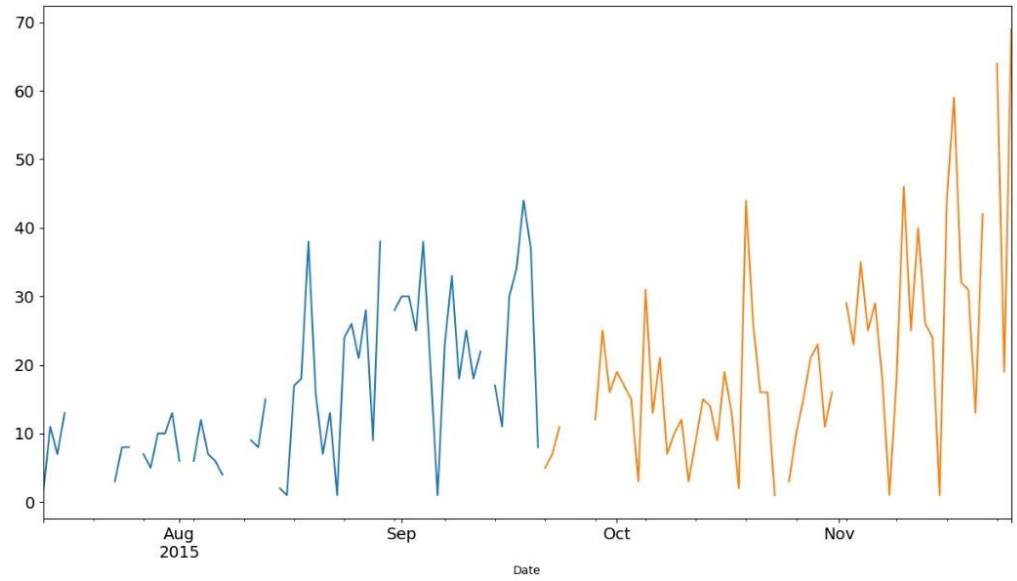

Figure 40: Test and Train Data

- **ARIMA:**

ARIMA is one among the foremost common classical strategies for time series forecasting. It stands for autoregressive integrated moving average and could be a model that forecasts given statistic supported its own past values, that is, its own lags and therefore the lagged forecast errors. ARIMA consists of three components:

- ✓ Autoregression (AR): is a model that shows a dynamical variable that regresses on its own lagged, or prior, values.
- ✓ Integrated (I): shows the differencing of raw observations to permit for the time series to become stationary (i.e., data values are replaced by the distinction between the data values and therefore their own previous values).
- ✓ Moving average (MA): incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

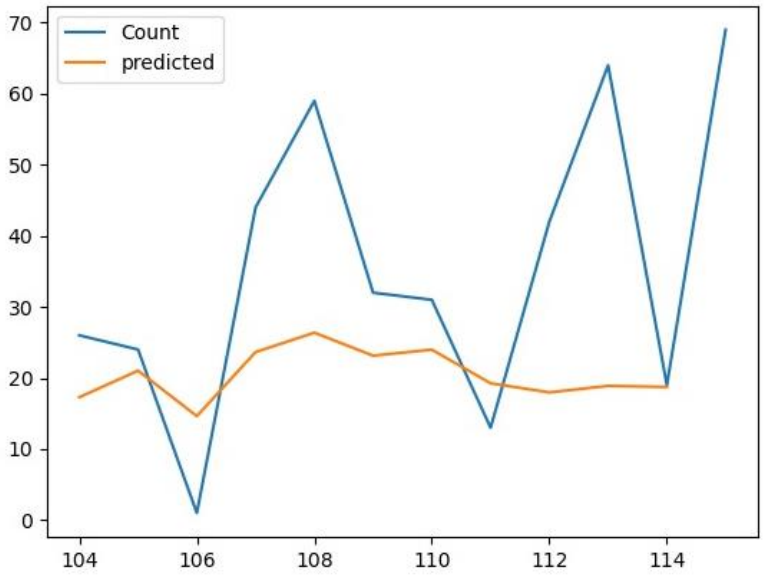

Figure 41: ARIMA

- **SARIMA:**

An extended implementation to ARIMA that enables us to direct modeling of the seasonal part of the series is named SARIMA. A haul with the ARIMA model is that it does not support seasonal data, that's a statistic with a repeating cycle. ARIMA expects data that is either not seasonal or has the seasonal part removed, e.g., seasonally adjusted via ways like seasonal differencing. SARIMA adds 3 new hyperparameters to specify the autoregression (AR), differencing (I), and moving average (MA) for the seasonal part of the series.

6.4 Machine Learning Algorithms

If we are not willing to use statistical models or they are not providing best result, there can be other methods that can be implemented. Machine learning is considered to be an alternative way of modeling time-series data for purpose of forecasting. Here, we extract features from the date to add to our "X variable" and the value of the time-series is "Y variable". We can use the alternative ways to perform analytics on data.

	Model
lar	Least Angle Regression
lr	Linear Regression
huber	Huber Regressor
br	Bayesian Ridge
ridge	Ridge Regression
lasso	Lasso Regression
en	Elastic Net
omp	Orthogonal Matching Pursuit
xgboost	Extreme Gradient Boosting
gbr	Gradient Boosting Regressor
rf	Random Forest Regressor
catboost	CatBoost Regressor
et	Extra Trees Regressor
ada	AdaBoost Regressor
dt	Decision Tree Regressor
knn	K Neighbors Regressor
lightgbm	Light Gradient Boosting Machine
llar	Lasso Least Angle Regression
par	Passive Aggressive Regressor

Figure 42: ML Algorithms

PyCaret, Python library is used to perform these machine learning algorithms. PyCaret is an open-source, low-code machine learning library in Python that automates machine learning workflows. With PyCaret, you spend less time coding and more time on analysis. You can train your model, analyze it, iterate faster than ever before, and deploy it instantaneously

6.5 Python Frameworks for Forecasting

There are few more open-source frameworks that are best resources if we want to build and scale time series solutions [46]:

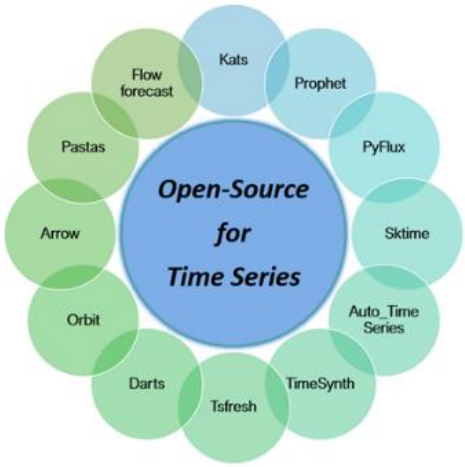

Figure 43:Time Series Forecasting

- **Prophet:**

  Prophet is open-source software released by Facebook's Core Data Science team. It is available for download on CRAN and PyPI. Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well [47].

Results:

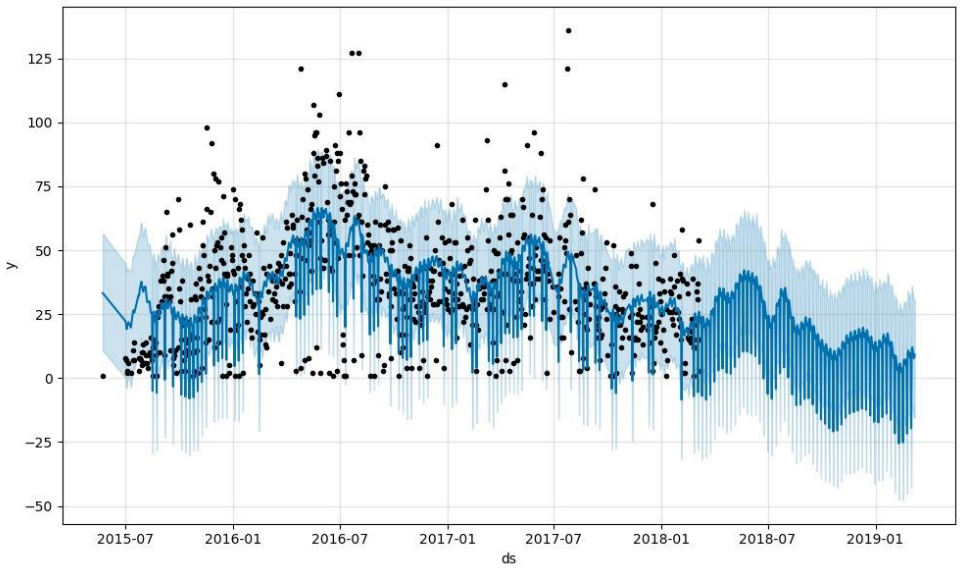

Figure 44: Prophet Algo Results

- **Pmdarima:**

Pmdarima is a statistical library designed to fill the void in Python's time series analysis capabilities. This includes [48]:

- ✓ The equivalent of R's auto.arima functionality
- ✓ A collection of statistical tests of stationarity and seasonality
- ✓ Seasonal time series decompositions
- ✓ Cross-validation utilities
- ✓ A rich collection of built-in time series datasets for prototyping and examples

Results:

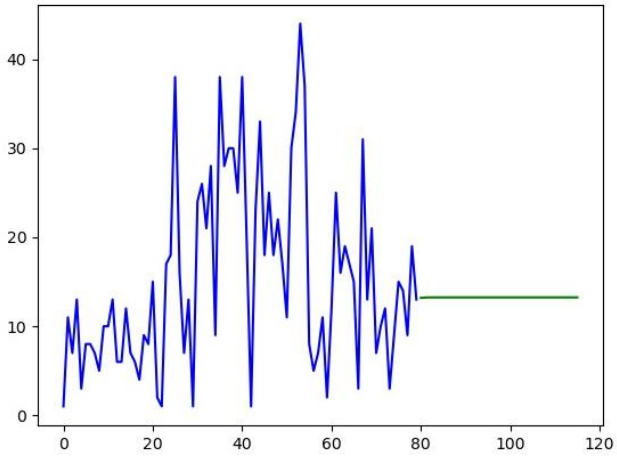

Figure 45: Pmdarima Results

We can clearly observe that this is not working well on our dataset, so we can choose other algorithms too to complete our analysis.

- **Kats**

Kats is another amazing open-source project by Facebook, released by their Infrastructure Data Science team. It is available for download on PyPI. Kats is a toolkit to analyze time-series data; a lightweight, easy-to-use, and generalizable framework to perform time series analysis. Kats aims to provide a one-stop shop for time series analysis, including detection, forecasting, feature extraction/embedding, and multivariate analysis, etc. [49]

- **Orbit**

Orbit is an amazing open-source project by Uber. It is a Python library for Bayesian time series forecasting. It provides a familiar and intuitive initialize-fit-predict interface for time series tasks, while utilizing probabilistic programming languages under the hood [50].

- **Sktime:**

Sktime is an open-source, unified framework for machine learning with time series. It provides an easy-to-use, flexible, and modular platform for a wide range of time series machine learning tasks. It offers scikit-learn compatible interfaces and model composition tools, with the goal to make the ecosystem more usable and interoperable [51].

Results:

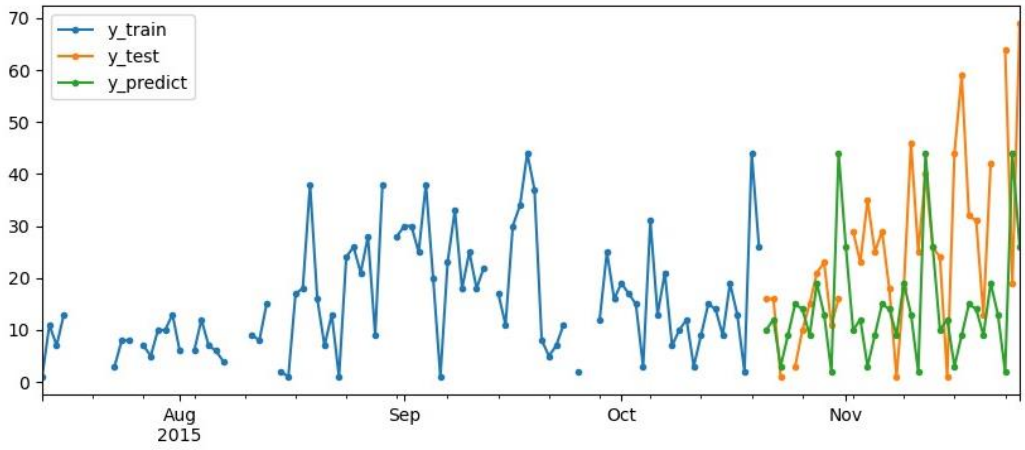

Figure 46: Sktime Results

6.6 Prescriptive Analytics

Prescriptive Analytics is a form of advanced analytics which examines data or content to answer the question “What should be done?” or “What can we do to make our required thing happen?” Here are few prescriptive analytics discussed:

- **Markov Decision Process:**

Markov Decision Processes are used to describe complex models or situations where each event depends on the previous event only. This attribute is called the Markov Property. It is Probabilistic model for prescriptive analytics.

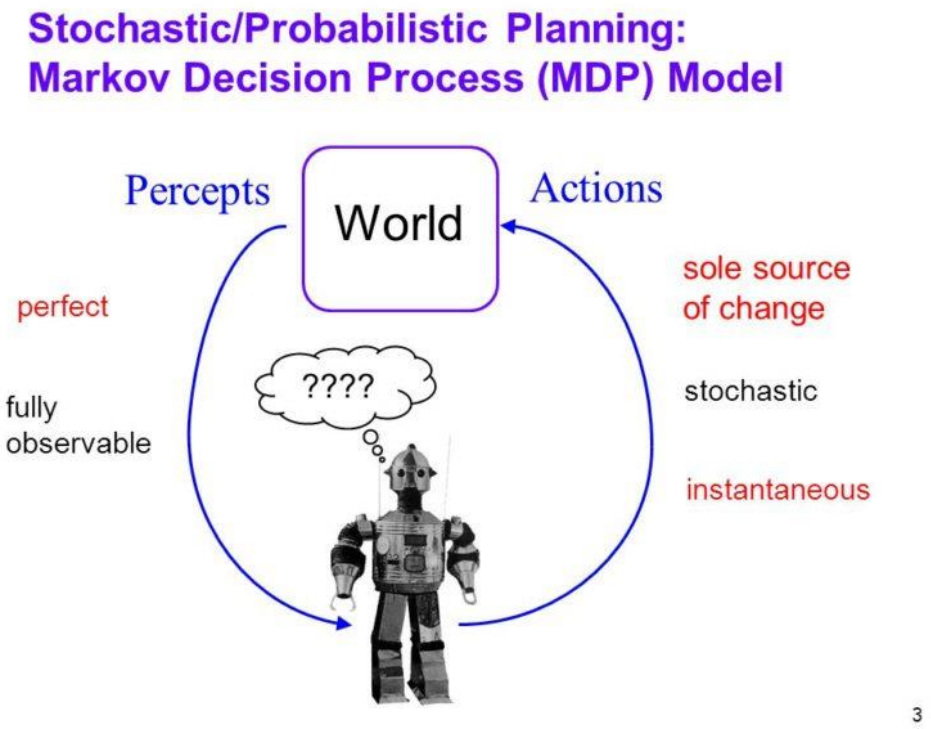

Figure 47: Markov Decision Process

A Markov process typically consists of an Agent, a State and an Environment with restrictions and rewards. We will implement each of this sequentially to get our final Markov model.

The State: We will proceed to create a state class that defines the state of each event of the model. This class will consist of many functions. The state class to construct a state object which in this instance represents a position in the grid defined by its x and y co-ordinates.

- **Risk Assessment:**

It categorizes in simulation prescriptive analytics. If we tend to take the time to gather data concerning the risks our organization could face within the future, do not simply leave that data sitting in a table. Analyze it! Risk assessment analytics is that the practice of applying data analytics to our risk assessment to spot threats to our organization and develop a

transparent risk assessment set up for remediating and responding to those threats. data analytics for risk assessments will vary from basic applied math modeling throughout the design phases of a project to sophisticated machine learning models that paint a dynamic risk picture supported by data sets.

A good risk management set up collects and collates all potential risks our organization would face. It ought to break down the risks into classes, assign risks to specific owners, and consolidate information that shows wherever risk mitigation efforts are worthy. Your approach to risk assessment analytics can depend upon wherever information is sourced, the kind of data is collected, the regulatory needs we are following, and our goals. For good planning we are going to forever consider worst case situation once planning with what-if situations.

6.6.1 Logic Based Models:

Here are few logic-based models are discussed.

- **Association rules:**

Association rules are "if-then" statements, which help to show the probability of relationships between data items, within large data sets in several types of databases. Like these association rules, there are many other rules which can be performed to make data-driven decision makings. Which are mentioned below:

- ✓ Decision rules
- ✓ Criteria Based rules
- ✓ Fuzzy rules
- ✓ Distribution rules
- ✓ Benchmark rules

Chapter 7: Conclusion/Results and Future Prospects

7.1 Conclusion

With the changing technology, as the data is increasing day by day, there is a lot more need to have such tools which can manipulate and manage this data very easily, very quickly, very efficiently. There are multiple operations being performed on data in our project, like data cleaning, data filtering, data Visualization and data analytics. The data which is being produced daily in tons of course keeps some meaning and is especially useful for industries. The industries use all the data to look at the past business trends and analyze it carefully and make such decisions which can make their business grow in future.

So, for managing these loads of data, one should have knowledge of the data mining, data science and machine learning in order to handle that all data and make decisions. But all these things need programmers, coders to go for these tasks, and if a person has no any knowledge of programming and coding, he will badly lag behind and will not be able to accomplish these tasks, so due to all these reasons, we have come up with this to facilitate the machine learners, data scientists, and data miners to decrease the time of manipulating that data.

It really takes large amount of time to manage that data and use it for the betterment of society. If we go for mere coding so it becomes highly time consuming and burden to deal with those large chunks of data, so data mining tools come to play in these situations, what they do is basically, they provide you a complete interface to retrieve your data, perform any operation on your that is gained through these tools they make the work easier. In this way the data and then finally visualize your data on dashboards. So, the main benefit time is also saved, and the data is managed very easily. This is also very efficient and helpful tool even a new user and beginner can play with tool.

7.2 Future Prospects

Our future work is adding the other modules in this tool. Like we have put limited number of machine learning algorithms, we have limited data set CSV. In future we are hoping to extend our project towards multivariant analysis as currently we are able to perform it only

for single one. We will also extend it for other data types too. And we will go for also adding more user-friendly interface to manage this tool easily.

7.3 Final Product/Dashboard/Results

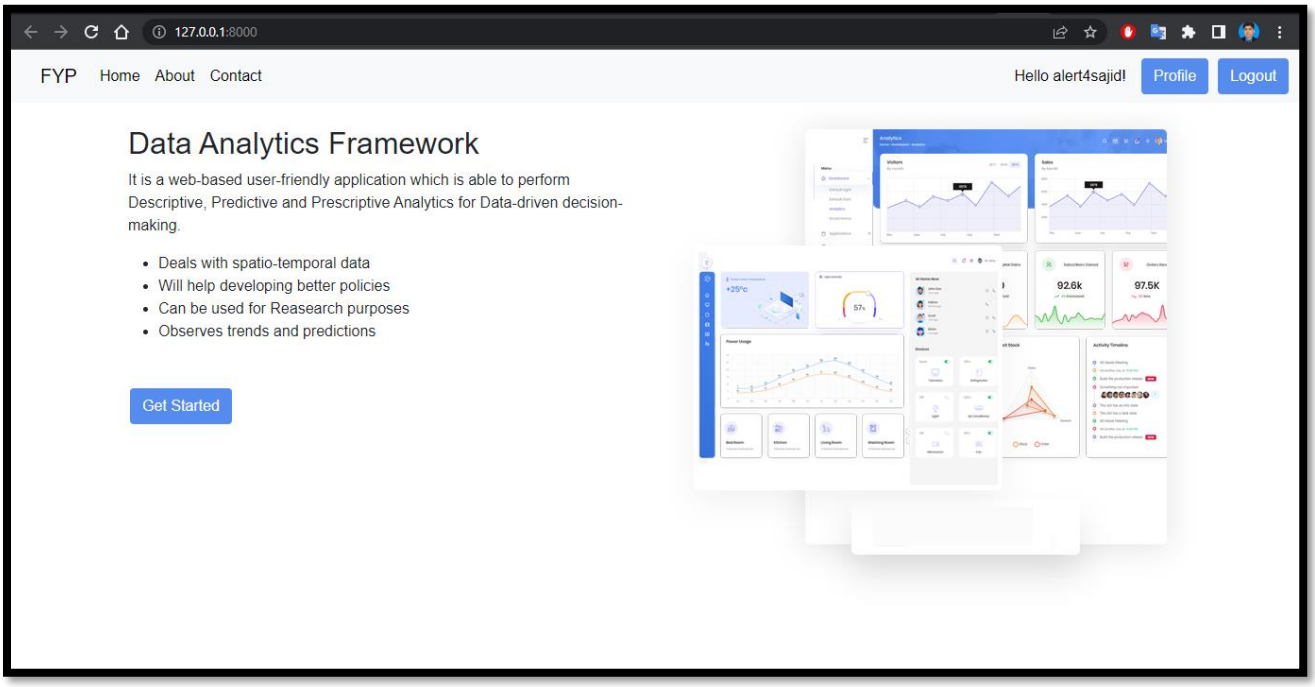

Figure 48: Our Front-end Web Interface

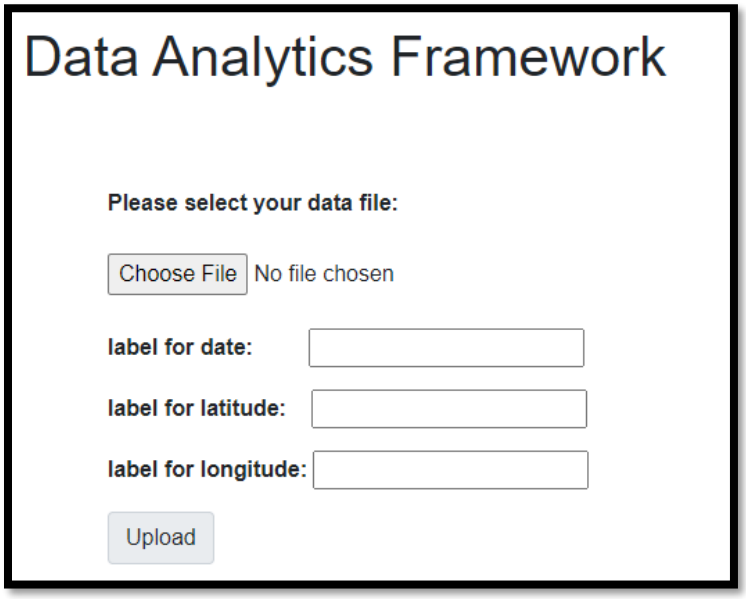

Figure 49: File Upload Interface for User

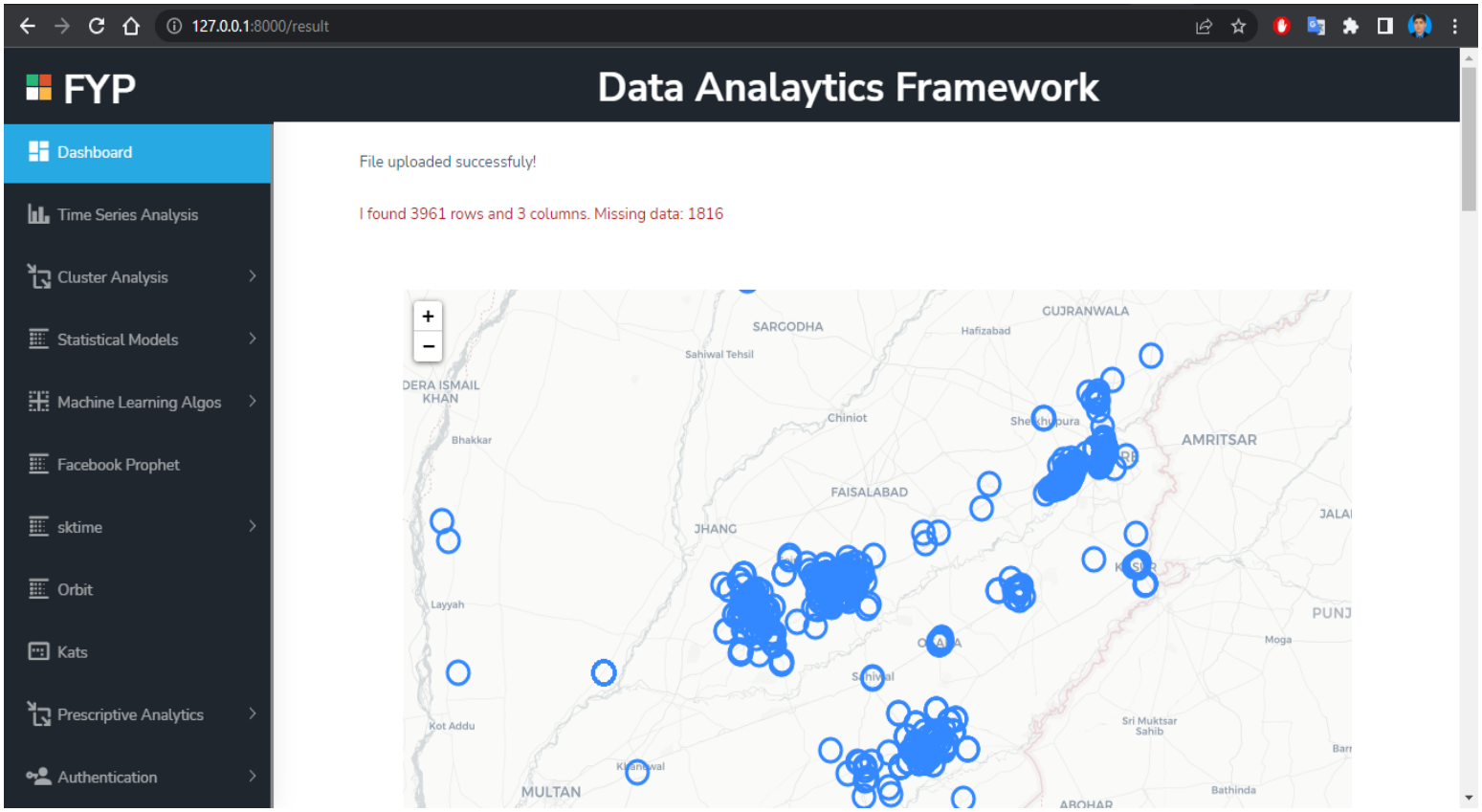

Figure 50: Our Dashboard for Data Analytics

Chapter 8: References

References

[1] https://hdsr.mitpress.mit.edu/pub/diub13so/release/5

[2] https://www.sciencedirect.com/science/article/pii/S0268401218309873#fig0005

[3] https://www.dawn.com/news/1110537

[4] https://mittalsouthasiainstitute.harvard.edu/2020/09/8-challenges-pakistans-digital-transformation-journey/

[5] https://www.endpolio.com.pk/global-polio-pakistan

[6] Davenport, T. H. (2006). Competing on analytics. *Harvard business review*, *84*(1), 98.

[7] https://rapidminer.com/glossary/data-mining-tools/

[8] https://comparecamp.com/rapidminer-review-pricing-pros-cons-features/

[9] Sharma, T. C., & Jain, M. (2013). WEKA approach for comparative study of classification algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*, *2*(4), 1925-1931.

[10] https://waikato.github.io/weka-wiki/faqs/different_versions/

[11] https://www.weka.io/

[12] https://github.com/MicrosoftDocs/powerbi-docs/blob/main/powerbi-docs/fundamentals/power-bi-overview.md

[13] https://www.freelogovectors.net/power-bi-logo-microsoft/

[14] https://powerbi.microsoft.com/en-au/what-is-power-bi/

[15] https://powerbi.microsoft.com/en-au/pricing/

[16] https://info.microsoft.com/ww-landing-gigaom-data-governance.html?lcid=en-us

[17] https://www.simplilearn.com/tutorials/data-analytics-tutorial/data-analytics-with-python

[18] https://www.investopedia.com/terms/d/data-analytics.asp

[19] https://www.northeastern.edu/graduate/blog/data-analysis-project-lifecycle/

[20] https://www.coursera.org/articles/what-is-python-used-for-a-beginners-guide-to-using-python

[21] https://intellipaat.com/blog/what-is-pycharm/

[22] Mathur, P. (2016). What is Anaconda and why should I bother about it?

[23] https://www.edureka.co/blog/python-anaconda-tutorial/

[24] https://www.w3schools.com/django/django_intro.php

[25] https://tutorial.djangogirls.org/en/django/

[26] https://www.w3schools.com/python/numpy/numpy_intro.asp

[27] https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to know/#:~:text=Pandas%20is%20an%20open%20source,support%20for%20multi%2Ddimensional%20arrays.

[28] Ari, N., & Ustazhanov, M. (2014, September). Matplotlib in python. In *2014 11th International Conference on Electronics, Computer and Computation (ICECCO)* (pp. 1-6). IEEE.

[29] https://www.dominodatalab.com/data-science-dictionary/folium#:~:text=Folium%20is%20a%20powerful%20Python,inline%20Jupyter%20maps%20in%20Folium

[30] https://www.geeksforgeeks.org/python-data-visualization-using-bokeh/

[31] https://www.tutorialspoint.com/scikit_learn/index.htm

[32] https://www.w3schools.com/html/html_intro.asp

[33] https://medium.com/bitgrit-data-science-publication/data-cleaning-with-python-f6bc3da64e45

[34] https://monkeylearn.com/blog/data-cleaning-steps/

[35] https://pandas.pydata.org/pandas-docs/stable/user_guide/

[36] https://www.alamy.com/four-components-of-data-cleaning-image383318415.html

[37] https://www.tableau.com/learn/articles/data-visualization

[38] https://www.geeksforgeeks.org/data-visualization-using-matplotlib/

[39] https://www.geeksforgeeks.org/python-data-visualization-using-bokeh/

[40] https://www.geeksforgeeks.org/python-plotting-google-map-using-folium-package/

[41]https://en.wikipedia.org/wiki/Cluster_analysis#:~:text=Cluster%20analysis%20or%20clustering%20is,in%20other%20groups%20(clusters)

[42] Davidson, I. (2002). Understanding K-means non-hierarchical clustering. *Computer Science Department of State University of New York (SUNY), Albany*.

[43] https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/

[44] https://en.wikipedia.org/wiki/OPTICS_algorithm

[45] https://www.datacamp.com/tutorial/tutorial-time-series-forecasting#arima

[46] Lazzeri, F. (2020). *Machine learning for time series forecasting with Python*. John Wiley & Sons.

[47] https://facebook.github.io/prophet/docs/quick_start.html#python-api

[48] https://github.com/alkaline-ml/pmdarima

[49] https://github.com/facebookresearch/Kats

[50] https://github.com/uber/orbit

[51] https://www.sktime.org/en/stable/examples/01_forecasting.html