Analysis of Covid-19 data using Machine learning

*Submitted in partial fulfilment of the requirements*

*for the award of the degree of*

**Bachelor of Computer Applications**

To

Guru Gobind Singh Indraprastha University, Delhi

Guide: Submitted by:

Mr. Pramod Kumar Soni 1. Astha Jain

01990302018

2. Megha Sajwan

05690302018

****

**Nurturing Excellence**

**Institute of Innovation in Technology & Management,**

**New Delhi – 110058**

**Batch (2018-2021)**

Certificate

We, Astha Jain Roll No-01990302018 & Megha Sajwan Roll No-05690302018 certify that the Summer Training Project Report (BCA-355) entitled “Analysis of Covid-19 data using Machine learning” is done by us and it is an authentic work carried out by us at Institute of Innovation In Technology & Management. The matter embodied in this project work has not been submitted earlier for the award of any degree or diploma to the best of my knowledge and belief.

1. Signature of the Student 2. Signature of the Student

Astha Jain Megha Sajwan

Date:

Certified that the Project Report (BCA-355) entitled“ Analysis of Covid-19 data using Machine learning ” done by the above students is completed under my guidance.

Signature of the Guide

Date:

Name of the Guide:

Mr Pramod Kumar Soni

Designation:

Countersigned

Director

## Acknowledgement

I am highly indebted to Institute of Innovation in Technology & Management (IITM) for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I would like to express my gratitude towards my guide Mr. Pramod Kumar Soni for their kind co-operation and encouragement which help me in completion of this project.

My thanks and appreciations also go to my colleague in developing the project and people who have willingly helped me out with their abilities.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S No** | **Topic** | **Page No** |
| 1 | Certificate | i |
| 2 | Acknowledgements | ii |
| 3 | List of Tables/Figures | iv |
| 4 | Project Proposal | 1 |
| 5 | Chapter-1: Introduction | 4 |
| 6 | Chapter-2: System Analysis | 7 |
| 7 | Chapter-3: System Design | 15 |
| 8 | Appendix A | 22 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Title** | **Page No** |
| 1 | Interface/ Database | 9 |
| 2 | Methodology | 12 |
| 3 | Gang Chart | 14 |
| 4 | SVM | 16 |
| 5 | Linear Regression | 17 |
| 6 | Dataset Snapshot 1 | 20 |
| 7 | Dataset snapshot 2 | 20 |
| 8 | Dataset Snapshot 3 | 21 |

Proposal for Summer Training 2020

**COVID-19 Outbreak Predication Using Machine Learning**

**Project**

**1. Project Name**

Analysis of Covid-19 data using Machine learning

**2. Introduction of the Project**

Our society is in the era of unbelievable attempts to struggle upon the spread of this life-threatening condition in terms of infrastructure, finance, business, manufacturing, and several other resources. Artificial Intelligence (AI) researchers strengthen their proficiency in developing mathematical paradigms for investigating this pandemic using nationwide distributed data. This project intends to apply the machine learning models simultaneously with the forecast of expected reach ability of the COVID-19 over the nations by using the real-time data.

**3. Objectives of the Project**

In this project, we will analyse the outbreak of coronavirus around different regions of India, visualize them using charts and graphs, and predict analysis the number of cases. Using Linear Regression Model and Support Vector Machine using python. How you can use machine learning algorithms to make prediction on such sensitive issues, we will certainly cover the current scenario and the impact it has created in India.

**4. Project Category**

Machine Learning

**5. Requirements**

1. Python 3+
2. Anaconda Navigator
3. Jupyter Notebook
4. Some libraries like numpy, pandas, matplotlib.pyplot, random, time, etc.

**6. Are you doing this project for any Industry/college?**

Yes, I am doing this project for the Institute of Innovation in Technology and Management, D- 27&28, Institutional Area, D-Block, Janakpuri, New Delhi – 110058.

**7. Limitations**

The models predicting infection rates, total mortality, and intensive care capacity are simpler constructs. They are adjusted when the conditions on the ground materially change, such as when states reopen; otherwise, they remain static. The problem with such an approach lies partly in the complexity of COVID-19’s different variables. These variables mean the results of typical COVID-19 projections do not have linear relationships with the inputs used to create them. AI comes into play here, due to its ability to ignore assumptions about the ways the predictors building the models might assist or ultimately influence the prediction.

**8. Future scope and further enhancement of the project**

Currently, the data being labeled by humans, has a lot of uncertainty, and it’s hard to achieve good accuracy and we’re using ML and in the future, we could apply deep learning for more enhancement and could use it as a dashboard application for all the recent updates for a large amount of data overall world and could use the data for a comparison for further analysis. This project could be enhanced to some future predictions based on previous data for further 10 days using ML.

**9. Name of guide**

**Associate Professor**

**Shri Pramod Kumar Soni**

**10. Date of submission**

August 2020

**11. Student Details**

**Name:** Megha Sajwan

**Enroll:** 05690302018 (M1)

**Email: megha.sajwaan2000@gmail.com**

**Phone no:** +91-9818290827

**Name:** Astha Jain

**Enroll:** 01990302018 (M1)

**Email: jainastha0512@gmail.com**

**Phone no:** +91-8383077823

**Chapter 1.**

**Systems Introduction**

1. **Brief Description of the System under Study**:

Our society is in the era of unbelievable attempts to struggle upon the spread of this life-threatening condition in terms of infrastructure, finance, business, manufacturing, and several other resources. Artificial Intelligence (AI) researchers strengthen their proficiency in developing mathematical paradigms for investigating this pandemic using nationwide distributed data. In this study, we intend to apply the machine learning models simultaneously with the forecast of expected reachability of the COVID-19 over the nations by using the real-time data.

* 1. **About the proposed System:**
* AIM:

To create a prediction model for COVID -19 using machine learning.

* Objectives:
* This model will automatically estimate the number of cases of weekly and bi-weekly data.
* This proposed study is beneficial for Indian doctors and the Indian government for managing the COVID-2019 outbreak for the next 7 days.
* In the future, we will develop a regression analysis based on artificial neural networks that can be developed to obtain data at regular intervals
  1. **The methodology used for Analysis, Design & Development:**

Artificial Intelligence (AI) researchers strengthen their proficiency in developing mathematical paradigms for investigating this pandemic using nationwide distributed data. This project intends to apply the machine learning models simultaneously to forecast the expected reachability of the COVID-19 over the nations by using the real-time data. We will analyse the outbreak of coronavirus around different regions of India, visualize them using charts and graphs, and predict the number of cases. They are using the Linear Regression Model and Support Vector Machine using python. And the software model used in this project is the iterative model. The iterative model is a particular implementation of a software development life cycle (SDLC) that focuses on an initial, simplified implementation, which then progressively gains more complexity and a broader feature set until the final system is complete.

* 1. **The methodology used for Data Collection:**

**(a) PRIMARY SOURCES** :

A primary source provides direct or first-hand evidence about an event, object, person, or work of art. Primary sources include historical and legal documents, eyewitness accounts, and results of experiments, statistical data, pieces of creative writing, audio and video recordings, speeches, and art objects.

Primary sources are: none

**(b) SECONDARY SOURCES :**

Secondary sources describe, discuss, interpret, comment upon, analyze, evaluate, summarize, and primary process sources.

Secondary sources are:

Information: Wikipedia

Covid-19 data of India which confirms the active cases, number of death, and confirmed cases. India state and population, which describes the sate wise adhere assigned and the land. Personal details in which traveling details and personal details are mentioned.

* 1. **System Requirement Tools:**

**Hardware Requirement Specifications**

* 4 GB RAM
* VGA monitor
* Intel 3.0 GHz or higher processor
* Keyboard and mouse
* Printer

**Software Requirement Specifications**

* Operating system - Windows (7/8/10)
* Visual Code editor/ Pycharm
* Jupyter Notebook

**Chapter 2**

**System Analysis**

**Software Requirement Specifications**:

The Iterative methodology takes the waterfall model and cycles through it several times in small increments. Rather than stretching the entire project across the phases of the SDLC, each step is turned into several mini-projects that can add value as the product evolves. The iterative approach shares many of the same goals as the agile model, except external customers are less involved, and the scope of each increment is usually fixed.

**2.1 Introduction**

The following subsections of the Software Requirement Specifications Document should provide the entire overview of the Information system “Analysis of Covid-19 data using Machine learning” under development. This document aims at defining the overall software requirements for data analysis. Efforts have been made to define the requirements of the Information system exhaustively and accurately

**2.1.1 Purpose**

The primary purpose of the Software Requirement Specifications Document is to describe in a precise manner all the capabilities that will be provided by the Software Program “Analysis of Covid-19 data using Machine learning”. It also states the various constraints which the system will abide by. This document further leads to a clear vision of the software requirements, specifications, and capabilities. These are to be exposed to the development, testing team, and end-users of the software.

**2.1.2 Scope**

Currently, the data being labeled by humans has much uncertainty, and it’s hard to achieve good accuracy, and we’re using ML and, in the future, we could apply deep learning for more enhancement. We could use it as a dashboard application for all the recent updates for a large amount of overall data world and could use the data to compare for further analysis. This project could be enhanced to some future predictions based on previous data using deep learning.

* + 1. **Definition, acronyms, abbreviations**

SDLC Software Development Life Cycle

SRS Software Requirement Specifications

* + 1. **References**
* [www.google.com](http://www.google.com)
* [www.kaggle.com](http://www.kaggle.com)
* [www.simplelearn.com](http://www.simplelearn.com)
* [www.greeksforgreeks](http://www.greeksforgreeks)
* [www.github.com](http://www.github.com)
  + 1. **Overview**

The rest of this SRS document describes the various system requirements, interfaces, features and functionality in detail.

**2.2 Overall description of proposed system**

2.2.1 Product Perspective

The code will be based on analysis, self-sufficient and independent.

Frontend

Jupyter Notebook Screen.

Output

Dataset.

Covid-19 data csv files

**Fig 1**

**2.2.1.1 System Interfaces:**

None

**2.2.1.2 Interfaces:**

The interface is Jupyter Notebook Screen , for every code of set there is an output screen.

**2.2.1.3 Hardware Interfaces:**

The hardware interface is an architecture used to interconnect two devices together. It includes the design of the plug and socket, the type, number and purpose of the wires and the electrical signals that are passed across them. In our project, simple working components are enough to implement the project.

**Hardware requirement:**

* GB RAM
* VGA monitor
* Intel 3.0 GHz or higher processor
* Keyboard and mouse
* Printer

**2.2.1.4 Software Interfaces:**

The software interfaces are the languages and codes that the applications use to communicate with each other and with the hardware.

**Software requirements:**

* Operating system - Windows (7/8/10)
* Visual Code editor/ Pycharm
* Jupyter Notebook

**2.2.1.5 Communication Interfaces:**

None

**3. Methodologies for Data Collection:**

**3.1. Primary Data Collection:**

A primary source provides direct or first-hand evidence about an event, object, person, or work of art. Primary sources include historical and legal documents, eyewitness accounts, and results of experiments, statistical data, pieces of creative writing, audio and video recordings, speeches, and art objects.

**3.2. Secondary Data Collection:**

Secondary sources describe, discuss, interpret, comment upon, analyse, evaluate, summarize, and primary process sources.

Secondary sources are:

Information: Wikipedia

Covid-19 data of India which confirms the active cases, number of deaths, and confirmed cases. India state and population, which describes the sate wise adhere assigned and the land. Personal details in which traveling details and personal details are mentioned.

**4. Methodology used for Analysis, Design and Development:**

**Iterative Process Model**

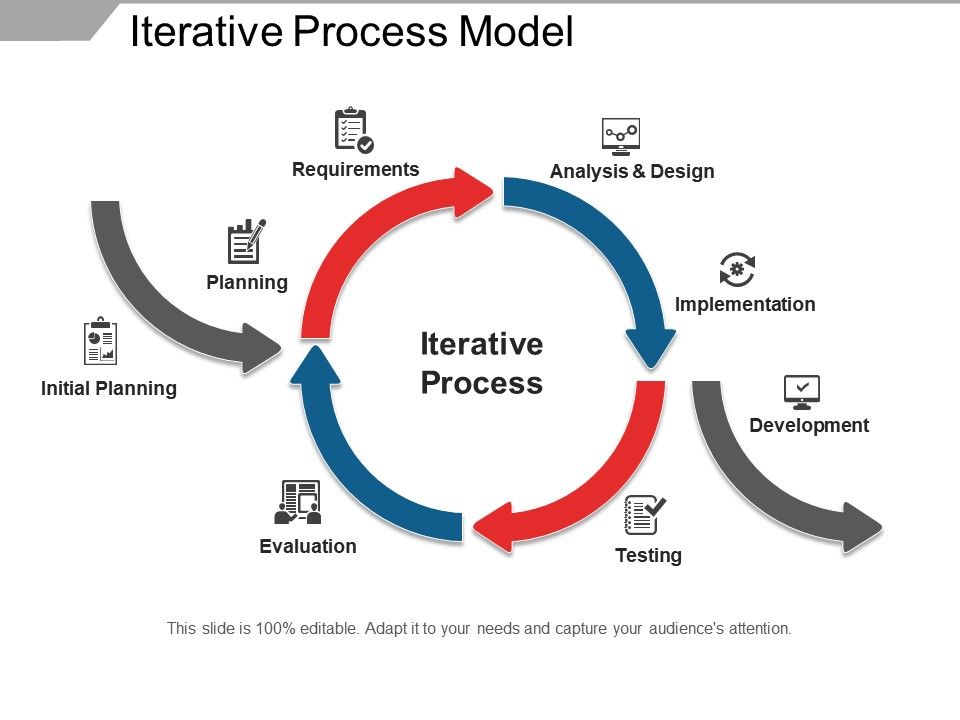
* The Iterative model is repetition incarnate. Instead of starting with fully known requirements, you implement a set of software requirements, then test, evaluate and pinpoint further requirements. A new version of the software is produced with each phase, or iteration. Rinse and repeat until the complete system are ready. This model gives you a working version early in the process and makes it less expensive to implement changes.
* Steps used in this project are:
* Requirement: check for dataset
* Analysis & Design: Analysis dataset of confirmed, death and active

cases using sum function. And design graphs for analysis.

* Implementation: Using seaborn graphs are made and then predict for

next week.

* Testing: then the accuracy of predictions is confirmed.



**Fig 2**

**5. CHART IN TABULAR FORM**

COVID-19 Outbreak Predication Using Machine Learning

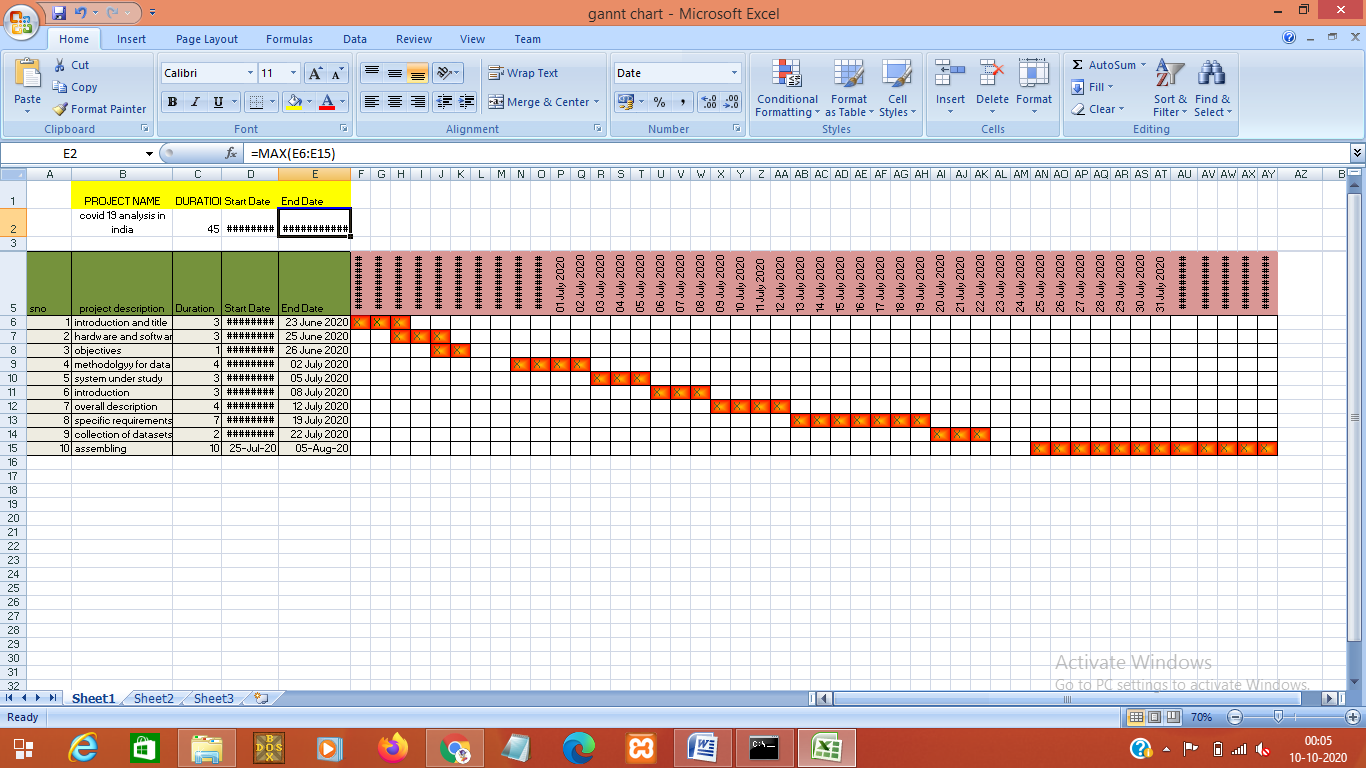
Duration: 55 days

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SNO | Task Name | No of Days | Start Date | Stop Date | Members/tasks Assigned |
| 1. | **SYNOPSIS** | 7 Days | 21 June 2020 | 26 June 2020 | Megha: Introduction, Objective, Title  Astha: hardware and software required |
| 2. | **Chapter 1**: System Introduction | 7 Days | 29 June 2020 | 05 July 2020 | Megha: methodology used for data collection  Astha: system under study and proposed system |
|  | **Chapter2**: System Analysis |  |  |  |  |
|  | Introduction | 3 Days | 06 July 2020 | 08 July 2020 | Megha |
|  | Overall description of proposed system | 4 Days | 09 July 2020 | 12 July 2020 | Astha |
| 3. | Specific requirements | 7 Days | 13 July 2020 | 19 July 2020 | Megha |
| 4. | **Chapter 3 :**System Design |  |  |  |  |
|  | Collection of datasets | 2 Days | 20 July 2020 | 22 July 2020 | Megha and Astha |
|  | Assemble all datasets | 10 Days | 25 July 2020 | 05 Aug 2020 | Megha and Astha |
| 5. | Assemble and Compilation | 5 Days | 7 Aug 2020 | 12 Aug 2020 | Megha and Astha |
| 6 . | Report | 2 Days | 13 Aug 2020 | 15 Aug 2020 | Megha and Astha |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Start Date: 21 June 2020

End Date: 15 August 2020

Team Members: Astha Jain, Megha Sajwan



**Chapter 3**

**System Design**

In this project, we have used Machine Learning, which is used to analyze and predict the data by datasets. A dataset is a structured collection of data generally associated with a unique body of work. It serves as the primary form of data storage and as a standard container for results returned by most algorithms. By using some libraries in the code, it reads the data and analysis the records, and also predicts it for future references.

**3.1. Algorithms Used**

Machine learning algorithms are about different math equations. An algorithm is just a customizable math function. That's why most algorithms have things like cost functions, weight values, and parameter functions that you can interchange based on the data you're working with. At its core, machine learning is just a bunch of math equations that need to be solved really fast.

That's why there are so many different algorithms to handle different kinds of data.

**3.1.1. Support Vector Machines (SVM)**

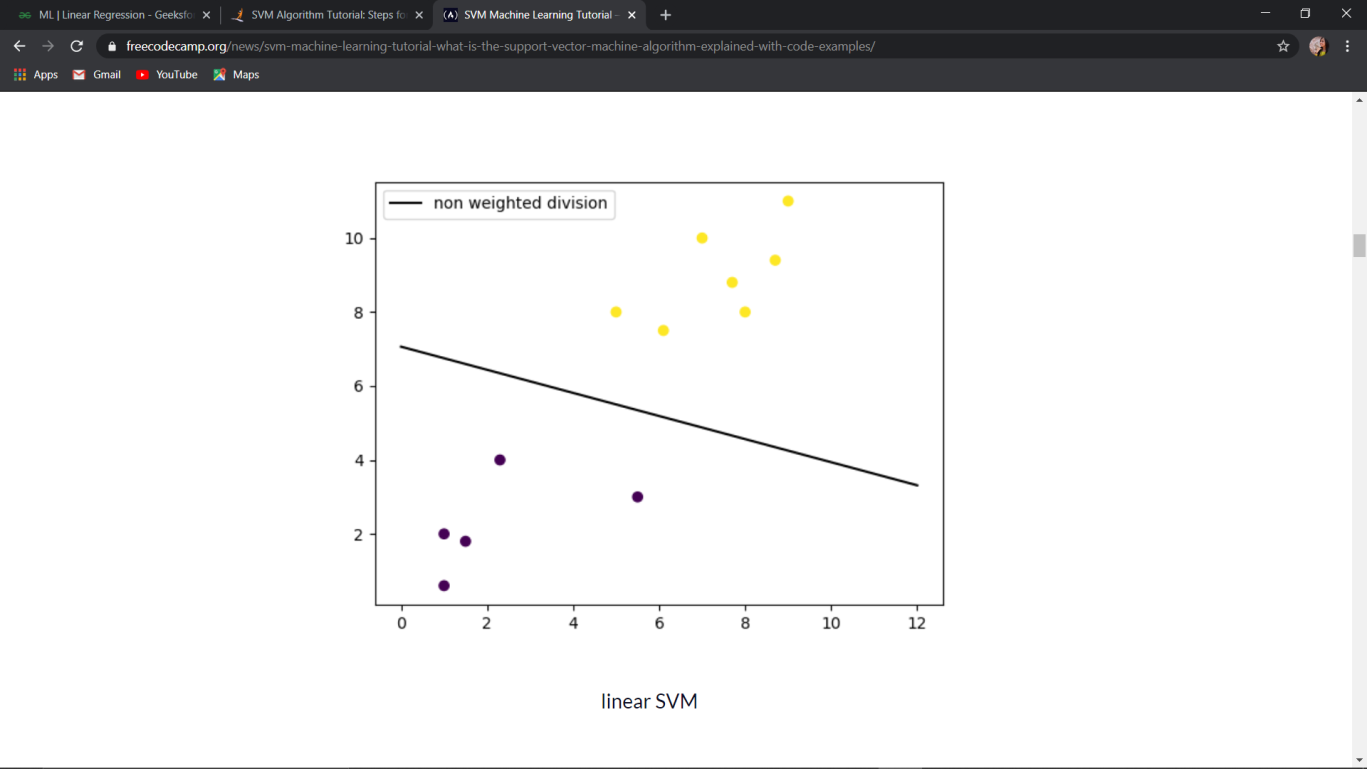
Support vector machines are a set of supervised learning methods used for classification, regression, and outlier’s detection. All of these are common tasks in machine learning.

You can use them to detect cancerous cells based on millions of images or you can use them to predict future driving routes with a well-fitted regression model.

There are specific types of SVMs you can use for particular machine learning problems, like support vector regression (SVR) which is an extension of support vector classification (SVC).

**3.1.2. How SVM Works**

A simple linear SVM classifier works by making a straight line between two classes. That means all of the data points on one side of the line will represent a category and the data points on the other side of the line will be put into a different category. This means there can be an infinite number of lines to choose from, that it chooses the best line to classify your data points. It chooses the line that separates the data and is the furthest away from the closet data points as possible.



**Fig 3**

**3.1.3. Types of SVMs**

There are two different types of SVMs, each used for different things:

a). **Simple SVM**: Typically used for linear regression and classification problems.

b). **Kernel SVM**: Has more flexibility for non-linear data because you can add more features to fit a hyperplane instead of a two-dimensional space.

**3.1.4. Advantages of Support Vector Machine Algorithm**

* Accuracy
* Works very well with limited datasets
* Kernel SVM contains a non-linear transformation function to convert the complicated non-linearly separable data into linearly separable data.

**3.1.5. Disadvantages of Support Vector Machine Algorithm**

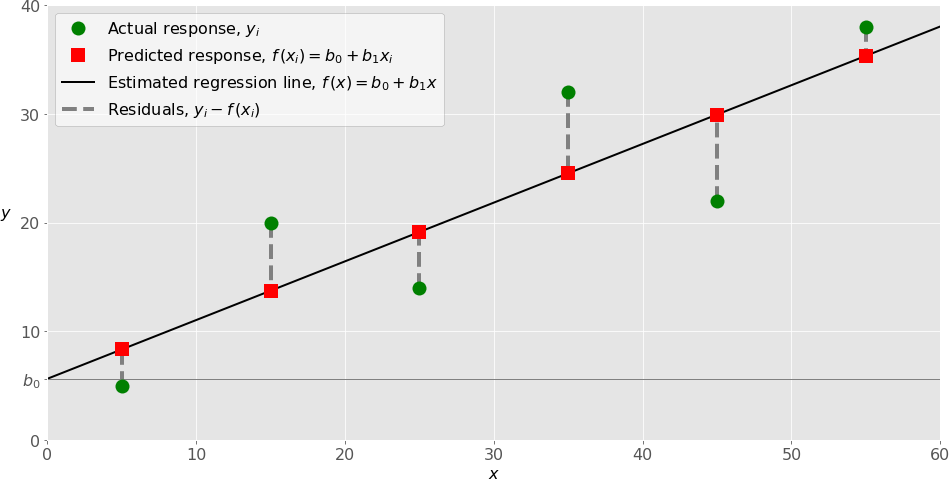
* Does not work well with larger datasets.
* Sometimes, training time with SVMs can be high.

**3.2 Linear regression**

Linear Regression is probably one of the most important and widely used regression techniques. It’s among the simplest regression methods. One of its main advantages is the ease of interpreting results.

### **Simple Linear Regression**

Simple or single-variate linear regression is the simplest case of linear regression with a single independent variable, 𝐱 = 𝑥.

The following figure illustrates simple linear regression:[](https://files.realpython.com/media/fig-lin-reg.a506035b654a.png)

Example of simple linear regression

When implementing simple linear regression, you typically start with a given set of input-output (𝑥-𝑦) pairs (green circles). These pairs are your observations. For example, the leftmost observation (green circle) has the input 𝑥 = 5 and the actual output (response) 𝑦 = 5. The next one has 𝑥 = 15 and 𝑦 = 20, and so on.

The estimated regression function (black line) has the equation (𝑥) = 𝑏₀ + 𝑏₁𝑥. Your goal is to calculate the optimal values of the predicted weights 𝑏₀ and 𝑏₁ that minimize SSR and determine the estimated regression function. The value of 𝑏₀, also called the **intercept**, shows the point where the estimated regression line crosses the 𝑦 axis. It is the value of the estimated response (𝑥) for 𝑥 = 0. The value of 𝑏₁ determines the **slope** of the estimated regression line.

The predicted responses (red squares) are the points on the regression line that correspond to the input values. For example, for the input 𝑥 = 5, the predicted response is (5) = 8.33 (represented with the leftmost red square).

The residuals (vertical dashed grey lines) can be calculated as 𝑦ᵢ - (𝐱ᵢ) = 𝑦ᵢ - 𝑏₀ - 𝑏₁𝑥ᵢ for 𝑖 = 1, …, 𝑛. They are the distances between the green circles and red squares. When you implement linear regression, you are actually trying to minimize these distances and make the red squares as close to the predefined green circles as possible.

### **Multiple Linear Regression**

Multiple or multivariate linear regression is a case of linear regression with two or more independent variables.

If there are just two independent variables, the estimated regression function is (𝑥₁, 𝑥₂) = 𝑏₀ + 𝑏₁𝑥₁ + 𝑏₂𝑥₂. It represents a regression plane in a three-dimensional space. The goal of regression is to determine the values of the weights 𝑏₀, 𝑏₁, and 𝑏₂ such that this plane is as close as possible to the actual responses and yield the minimal SSR.

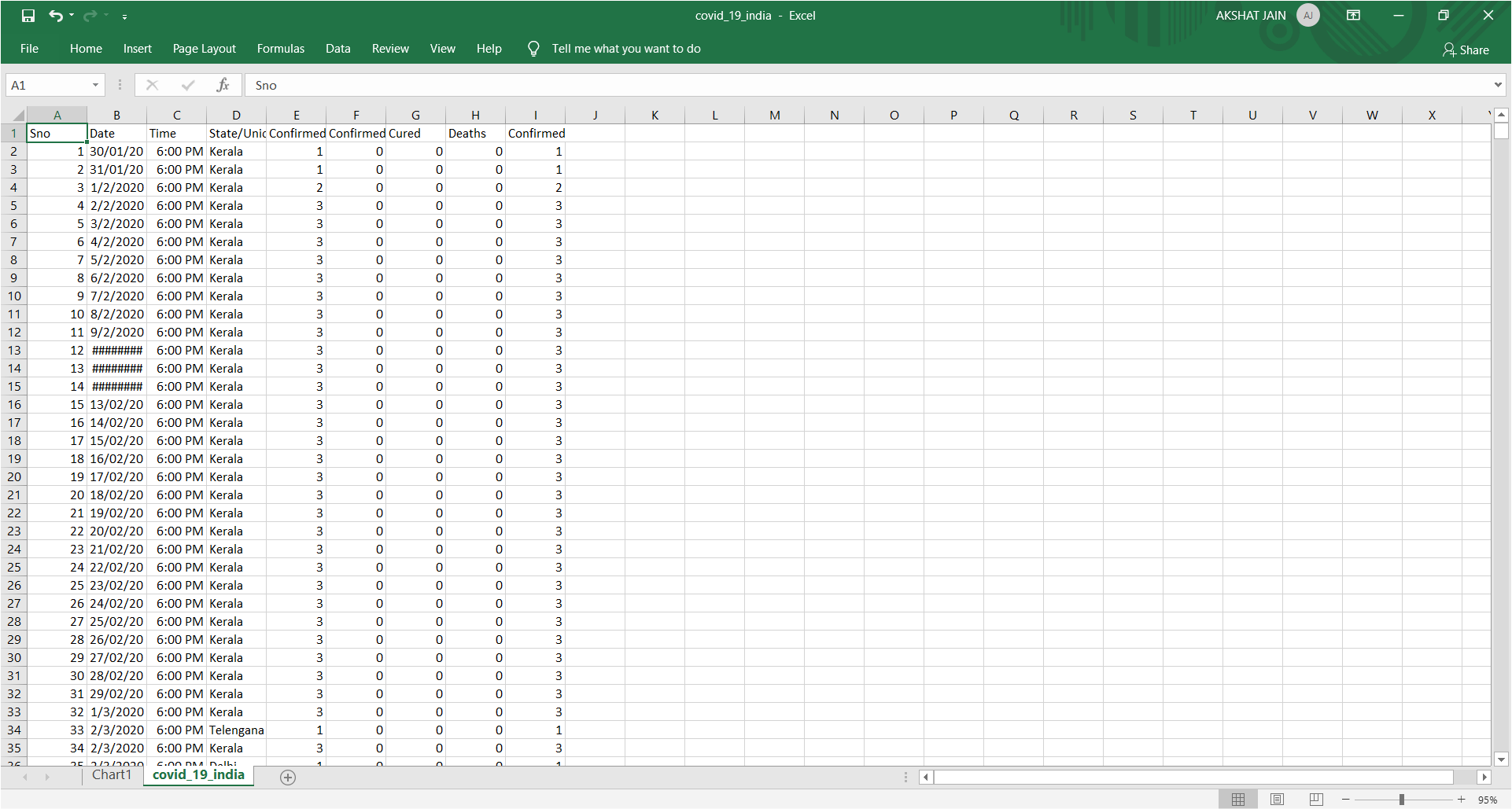
The case of more than two independent variables is similar, but more general. The estimated regression function is (𝑥₁, …, 𝑥ᵣ) = 𝑏₀ + 𝑏₁𝑥₁ + ⋯ +𝑏ᵣ𝑥ᵣ, and there are 𝑟 + 1 weights to be determined when the number of inputs is 𝑟.

### **Polynomial Regression**

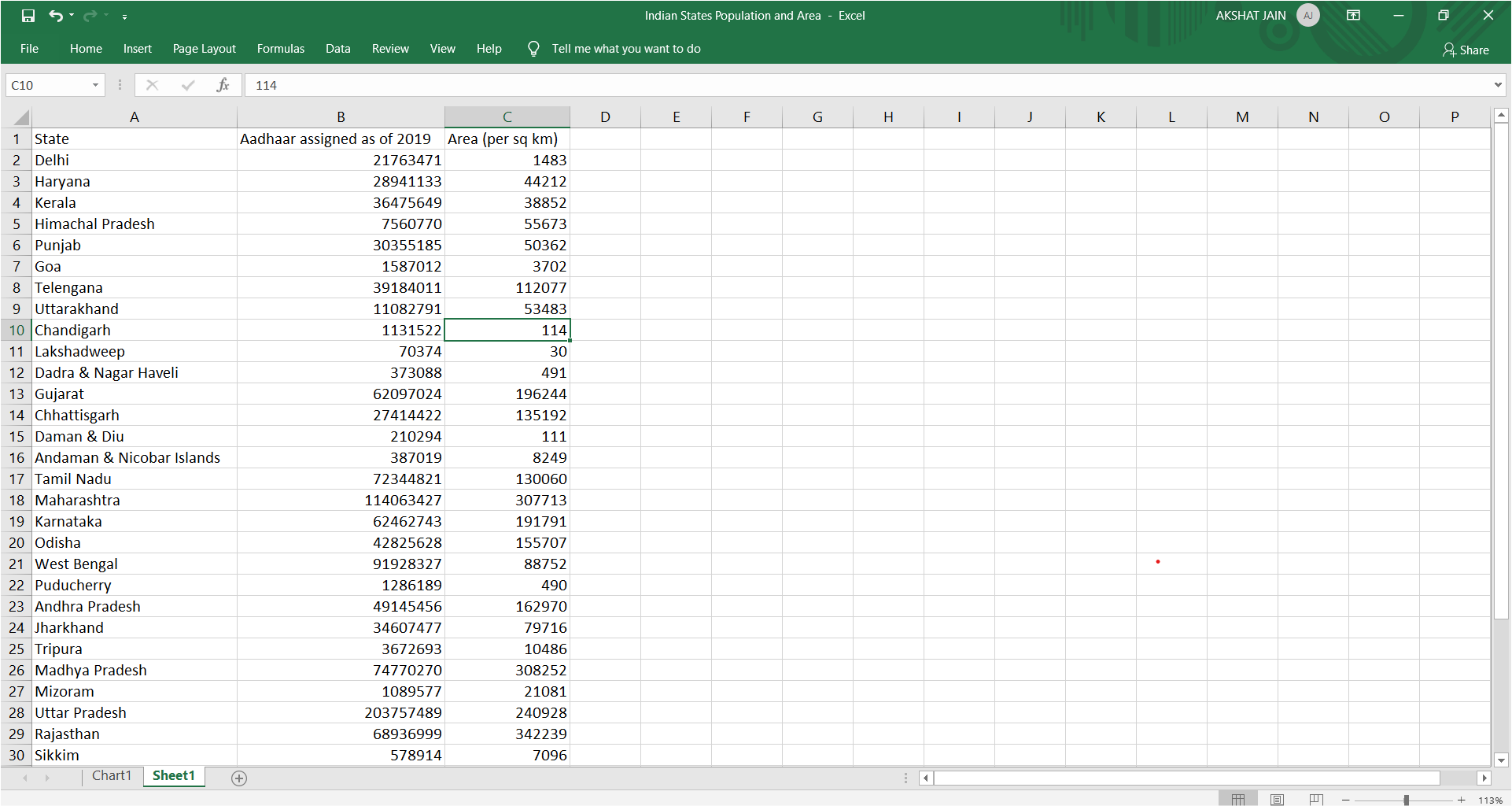
You can regard polynomial regression as a generalized case of linear regression. You assume the polynomial dependence between the output and inputs and, consequently, the polynomial estimated regression function.

In other words, in addition to linear terms like 𝑏₁𝑥₁, your regression function 𝑓 can include non-linear terms such as 𝑏₂𝑥₁², 𝑏₃𝑥₁³, or even 𝑏₄𝑥₁𝑥₂, 𝑏₅𝑥₁²𝑥₂, and so on.

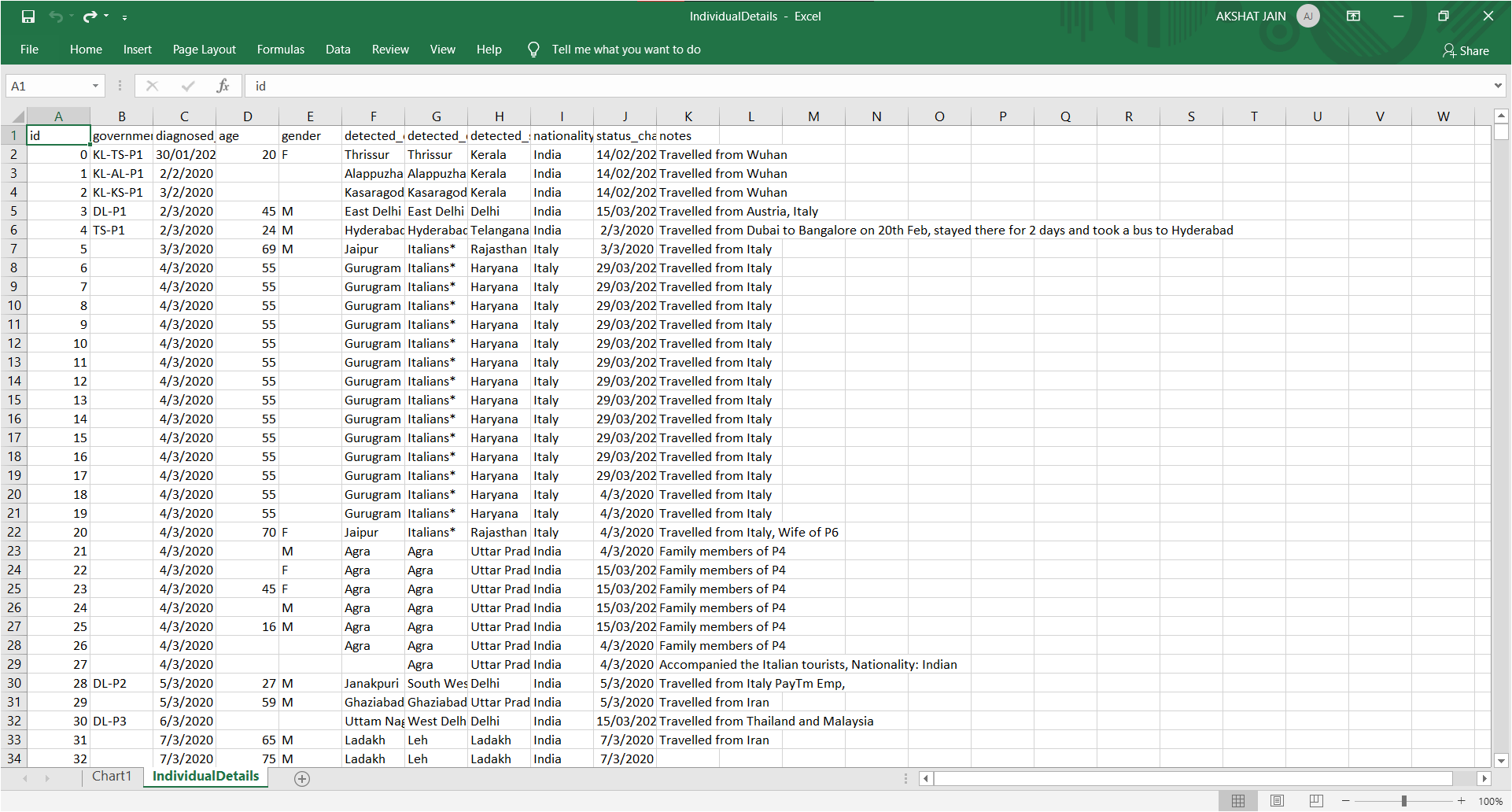
**3.3. Snapshot of dataset**



**Fig 4**



**Fig 5**



**Fig 6**

**APPENDIX A**

Snapshot of code:

