A Case Study Report

on

COVID-19 CASES PREDICTION AND VISUALISATION

Submitted in partial fulfilment for the requirements of

B.E. (CSE) VI Semester Case Study

in

COMPUTER SCIENCE AND ENGINEERING

By

Narsimha Sadula (160118733032)

D Santhosh Sai (160118733046)

Department of Computer Science and Engineering,

Chaitanya Bharathi Institute of Technology (Autonomous),

(Affiliated to Osmania University, Hyderabad)

Hyderabad, TELANGANA

(INDIA) – 500 075

CHAITANYA BHARATHI
INSTITUTE OF TECHNOLOGY (A)
Affliated to Osmania University

CERTIFICATE

This is to certify that the project entitled "COVID-19 CASES PREDICTION AND VISUALISATION", submitted to the Computer Science and Engineering Department, Chaitanya Bharathi Institute of Technology, in partial fulfilment of the requirement for the course Case Study, is a bonafide record of work done by **Narsimha Sadula (1601-18-733-032) and Santhosh Sai D(1601-18-733-046),** from February, 2021 to May, 2021 under our guidance and supervision.

Mentors,

Dr. E. Padmalatha

Assistant Professor

Department of CSE

CBIT

Faculty Coordinator,

Smt.T.Suvarna Kumari

Associate Professor

Department of CSE

CBIT

DECLARATION

I/we hereby declare that the case study entitled "COVID-19 CASES PREDICTION AND VISUALISATION" submitted for the B.E (CSE) degree is my original work and the project has not formed the basis for the award of any other degree, diploma, fellowship or any other similar titles.

Narsimha Sadula
D Santhosh Sai

Name(s) and Signature(s) of the

Student

Place: Hyderabad

Date: 29-05-2021

ACKNOWLEDGEMENT

Before we get into the thick of things, we would like to add a few words of appreciation for the people who have been a part of this project right from its inception. The writing of this project has been one of the significant academic challenges we have faced and without the support, patience, and guidance of the people involved, this task would not have been completed. It is to them we owe our deepest gratitude. We would like to express our heartfelt gratitude to Dr. E. Padmalatha, our project guides, for their invaluable guidance and constant support, along with their capable instruction and persistent encouragement.

It gives us immense pleasure in presenting this case study report on "COVID-19 CASES PREDICTION AND VISUALISATION". It has been our privilege to have a project guide who has assisted us from the commencement of this project. The success of the project is the result of sheer hard work, and determination put in by us with the help of my project guide. We hereby take this opportunity to add a special note of thanks for Smt.Suvarna Mam who undertook to act as our guide despite her many other academic and professional commitments. Her wisdom, knowledge, and commitment to the highest standards inspired and motivated us. Without her insight, support, and energy, this project wouldn't have kick-started and neither would have reached fruitfulness.

We would like to take this opportunity to thank our Principal, Dr. G.P.Saradhi Varma, as well as the management of the institute, for having designed an excellent learning atmosphere. Last but not least, we would like to express our gratitude to our friends and respondents for the support and willingness to spend some time with us to fill in the questionnaires.

Abstract

Coronavirus disease (COVID-19) is an inflammation disease from a new virus. The disease causes respiratory ailment (like influenza) with manifestations, for example, cold, cough and fever, and in progressively serious cases, the problem in breathing. COVID-2019 has been perceived as a worldwide pandemic and a few examinations are being conducted utilizing different numerical models to anticipate the likely advancement of this pestilence. These numerical models dependent on different factors and investigations are dependent upon potential inclination. Here, we presented a model that could be useful to predict the spread of COVID-2019. We have performed linear regression, Multilayer perceptron and Vector autoregression method for desire on the COVID-19 Kaggle data to anticipate the epidemiological example of the ailment and pace of COVID-2019 cases in India. Anticipated the potential patterns of COVID-19 effects in India dependent on data gathered from Kaggle. With the common data about confirmed, death and recovered cases across India for over the time length helps in anticipating and estimating the not so distant future. For extra assessment or future perspective, case definition and data combination must be kept up persistently.

Table of Contents

1. Introduction

The outbreak of CoronaVirus Disease 2019 (COVID-19) in Wuhan has significantly impacted the economy and society globally. Countries are in a strict state of prevention and control of this pandemic. Most public are being infected by this virus due to their negligence in taking precautions. One of the reasons for their negligence is lack of knowledge about the dangerous CoronaVirus. The main purpose of our project is to predict the Growth and Trend of COVID-19 Pandemic using Machine Learning and bring awareness about the dangerous infectious CoronaVirus.

This project can visualize and predict the growth and trend of covid cases in the world. The model we created uses different Machine Learning algorithms to predict covid cases in future. It uses different libraries to visualize the trend of covid cases in the world in the form of graphs and different diagrams. There is a great scope of using this model in any mobile applications in order to make an application and make it much easier for the public to understand and get to know about it. This model can not only visualize and predict trends of the world but also predicts and visualizes covid trends in specific countries.

The algorithms used in this project are very efficient in finding the correct predictions about the covid cases that may occur in the future of next 10 days. We used three different algorithms in which few give us better results than others. We want to show how algorithms can be used in order to find out the results and predictions of covid cases in the future. This model can not only visualize and predict trends of the world but also predicts and visualizes covid trends in specific countries. This can be very helpful to many people who want to know information about the covid cases in future so that they can take precautions.

2. Background Information

- The purpose of the software requirements specification document is to clearly define the system under development, namely the COVID-19 CASES PREDICTION AND VISUALISATION.
- The intended audience of this document includes the user who wants to use our program to find out the predictions of cases in their device.
- .This project is intended to bring awareness to society by showing them how cases will be in future and how we need to be prepared for such conditions . People tend to take precautions if they get to know about the danger that may cause in the future.
- The user can run this program on a desktop or laptop which has an internet connection and a jupyter notebook installed. Jupyter Notebook gives a very good visualization of the covid cases and predictions .
- Machine Learning Algorithms help us to get the predictions by studying a large dataset of cases and patterns of increasing and decreasing cases in the world. That is the reason we used machine learning algorithms to find the predictions of cases based on the cases dataset we already have.

3. Scope of Case Study

- The main driver of the shift towards voice user interfaces is the changing user demands. voice applications enable consumers to use natural language to eliminate or reduce manual effort, making it a lot faster to accomplish tasks. Once the voice capability is integrated into the application, users can analyse audio files and in return, receive a text file of the transcribed speech. Voice assistants will not only accurately understand what we are saying, but how we are saying it and the context in which the inquiry is made.
- The Importance of Voice assistants and Localization has been well understood over the years and many businesses have already been reaping the benefits. With Multilingual voice assistants, businesses now bring the power of localization and combine it with voice assistance to obtain even greater benefits.
- Multilingual voice assistants help to increase accessibility and reach, improves customer acquisition, build brand loyalty in customers, give a competitive edge over global competitors and level the playing field with local competitors, and finally, improves the revenue of your organization.

4. Design and Implementation

4.1 Software Requirements Specification

4.1.1. INTRODUCTION

1.1 Purpose

The outbreak of CoronaVirus Disease 2019 (COVID-19) in Wuhan has significantly impacted the economy and society globally. Countries are in a strict state of prevention and control of this pandemic. Most public are being infected by this virus due to their negligence in taking precautions. One of the reasons for their negligence is lack of knowledge about the dangerous CoronaVirus. The main purpose of our project is to predict the Growth and Trend of COVID-19 Pandemic using Machine Learning and bring awareness about the dangerous infectious CoronaVirus.

1.2 Scope

- This project can visualize and predict the growth and trend of covid cases in the world. The model we created uses different Machine Learning algorithms to predict covid cases in future. It uses different libraries to visualize the trend of covid cases in the world in the form of graphs and different diagrams. There is a great scope of using this model in any mobile applications in order to make an application and make it much easier for the public to understand and get to know about it. This model can not only visualize and predict trends of the world but also predicts and visualizes covid trends in specific countries.

1.3 Definitions, Acronyms, and Abbreviations

- **Table**

COVID-19	CoronaVirus Disease 2019
SVM	Support Vector Machine
PR	Polynomial Regression
BRR	Bayesian Ridge Regression

- **Definitions :**

Support Vector Machine :

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Polynomial Regression :

Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. It is also called the special case of Multiple Linear Regression in COVID 19 CoronaVirus Disease 2019 SVM Support Vector Machine (Algorithm) PR Polynomial Regression (Algorithm) BRR Bayesian Ridge Regression (Algorithm) ML. Because we add some polynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression.If we apply a linear model on a linear dataset, then it provides us a good result as we have seen in Simple Linear Regression, but if we apply the same model without any modification on a non-linear dataset, then it will produce a drastic output. Due to which loss function will increase, the error rate will be high, and accuracy will be decreased.So for such cases, where data points are arranged in a non-linear fashion, we need the Polynomial Regression model. We can understand it in a better way using the below comparison diagram of the linear dataset and non-linear dataset.

Bayesian Ridge Regression :

Bayesian regression allows a natural mechanism to survive insufficient data or poorly distributed data by formulating linear regression using probability distributors rather than point estimates. The output or response ‘y’ is assumed to be drawn from a probability distribution rather than estimated as a single value.One of the most useful types of Bayesian regression is Bayesian Ridge regression which estimates a probabilistic model of the regression problem.

1.4 References

- **https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47**
- **https://towardsdatascience.com/machine-learning-basics-polynomial-regression-3f9dd30223d1**
- **https://towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7**
- **https://scikit-learn.org/stable/auto_examples/linear_model/plot_bayesian_ridge.html**
- **https://journals.plos.org/plosone/article?id=10.1371/ journal.pone.0239960**

1.5 Document Overview

- From now, the purpose of the project, its scope and the technologies that will be used in

the building of this project has been described. The rest of the SRS examines the specifications of the project in detail.

- Section 2 of the SRS presents the overall description and requirements, such as user characteristics, project constraints, architecture design, design etc.
- Section 3 outlines the detailed, specific functional, performance, system and other related.

requirements of the project.

4.1.2. GENERAL DESCRIPTION

2.1 Product Perspective

As it is a console application we can open it in any device in which Jupyter Notebook is installed. We just have to open the notebook inorder to see the trends and prediction of cases. Inorder to get the real time predictions and trends we have to install the required modules and run the notebook with the present day dataset.As it is an ML project there are many modules that are required to analyze , predict and visualize the data from the raw datasets.

2.2 Product Functions

As the project was made to bring awareness in public about the covid situation happening in our country and around the world, it can visualize the data into simple graphs which are very understandable. This model performs cleaning of raw data in order to transform it into understandable data. It has the ability to predict covid cases for upcoming 10 days using machine learning algorithms like SVM, PR and BRR. This model can predict and visualize data of many countries for which we provide the datasets.It has a great ability to predict future covid cases in any country with greater precision.

2.3 User Characteristics

It is considered that the user does have the basic knowledge of Computer and the console. One has to have the knowledge of Jupyter Notebook in order to open and run the project. It's an additional bonus if one has knowledge about installing modules and updating datasets to the model.

2.4 General Constraints

The most basic constraint is having a computer in which the Jupyter Notebook is installed. There should be internet connectivity with good speed in order to download the modules required to run the project. One important constraint is, we have to update the dataset everyday in order to get the predictions in real time.

2.5 Assumptions and Dependencies

It is assumed that the latest datasets will be made available for the model. As it requires internet connection to run the project, an internet connection with good speed must be provided. It is assumed that the user is familiar with Jupyter Notebook.

4.1.3. SPECIFIC REQUIREMENTS

.

Functional Requirement No:	Function Requirement Description
1	User should be able to import/retrieve Dataset
2	As the Dataset gets updated,the project should also be updated everyday.
3	Should be able to process data and predict future target data columns.

4.1.4 External Interface Requirements

3.2.1 Hardware Requirements

- Physical server or virtual machine.
- Internet access to download the files from Anaconda Cloud, or a USB drive containing all of the files you need with alternate instructions for air gapped installations.

3.2.2 Software Requirements

- Linux environment: Installations have been tested on Red Hat Enterprise Linux/CentOS 6.7, 7.3, 7.4, and 7.5, and Ubuntu 12.04+. Verify Linux version.
- Client environments may be Windows, macOS or Linux.
- Ubuntu users may need to install cURL. Verify cURL access

3.3 Performance Requirements

- CPU: 2 x 64-bit, 2.8 GHz, 8.00 GT/s CPUs or better. Verify machine architecture.

- Memory: minimum RAM size of 32 GB, or 16 GB RAM with 1600 MHz DDR3 installed, for a typical installation with 50 regular users. Verify memory requirements.

3.5 Security Requirements

- Root access or sudo capabilities. Verify root access and sudo privileges.
- OPTIONAL: Ability to make IPTables modifications.
- SELinux policy edit privileges.
  NOTE: SELinux does not have to be disabled for Repository operation.

3.6 Maintainability Requirements

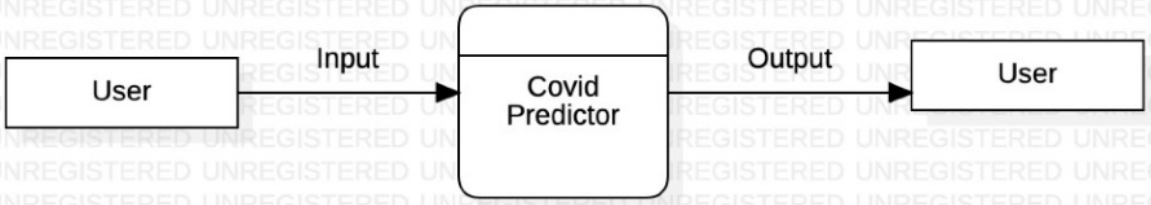
- The user learns from the interactions with the user and improves itself.
- The user should be consistent in its success and error responses to a user.
- New information should be updated and maintained in the database.

3.7 Reliability Requirements

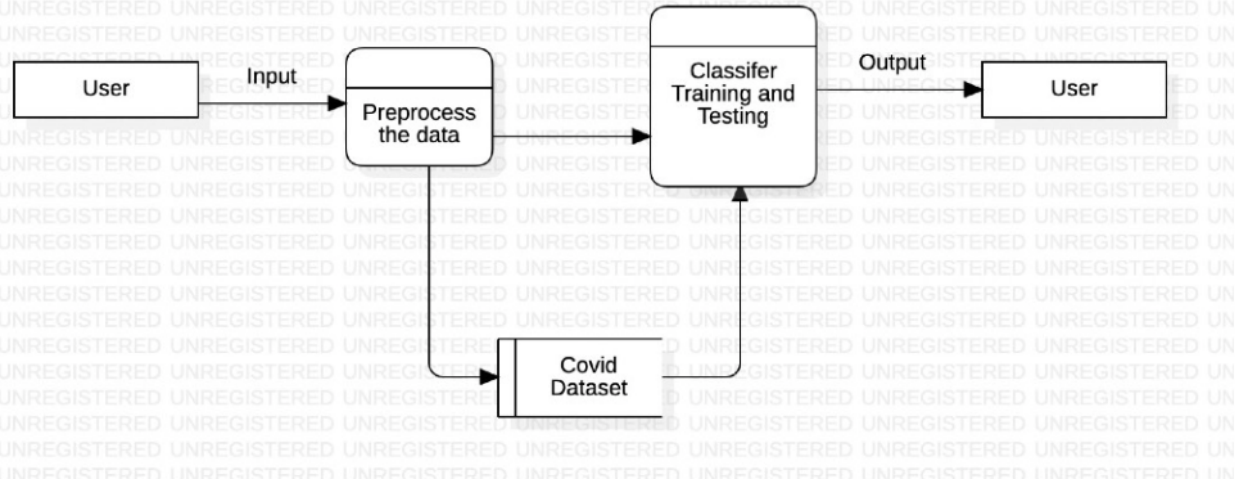
- The system should perform appropriate functions according to user input 95% of time only if the command is valid. The assistant may not respond to an invalid query.
- The assistant learns from the interactions with the user.
- Assistant may not respond if the cloud servers are down.
- Assistant may respond even in the absence of an internet connection.

4.2 System Design

4.2.1 Level – 0 DFD:

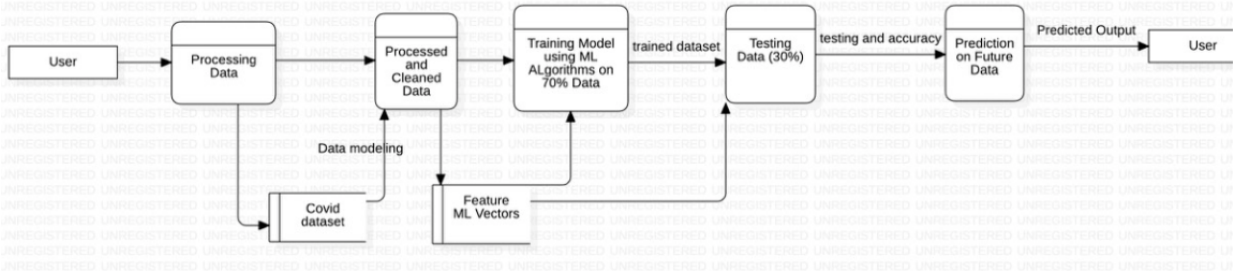


4.2.2 Level – 1 DFD:



The Data Flow diagram shows the flow of data within the application.

4.2.3 Level 2 DFD for Process 1:



4.2.5 Data dictionary:

Name of the Dataset : Covid Cases Dataset.

Covid Cases Dataset = **confirmed_df + deaths_df + latest_data**

It has 3 csv files with different data

confirmed_df - It has all the details of the total confirmed covid cases in different countries.

deaths_df - It has all the details of total confirmed covid cases in different countries.

latest_data - It has all the details of latest confirmed covid cases in different countries on a particular date.

cols = confirmed_df.keys() *All the column names present in confirmed_df* **dates** - All the dates from 22-01-2020 to today.

The above data are divided into sub groups in order to pass processable data to the ML model.

world_cases - It has the total covid cases in the world on a particular date. **total_deaths** - It has the total deaths in the world on a particular date.

mortality_rate - It has the mortality rate of covid patients in the world on a particular date.

mortality rate = deaths_sum/confirmed_sum

Inorder for the model to predict the patterns we have to calculate the daily increase and moving averages of the cases

For Confirmed Cases

world_daily_increases - It is the collection of daily increases of the cases in the world.

world_confirmed_avg - It is the collection of moving averages of the cases in the world.

world_daily_increase_avg - It has the moving averages of the daily increases of covid cases in the world.

For Deaths

world_daily_deaths - It is a collection of daily increases in deaths of covid affected patients.

world_death_avg - It is a collection of moving averages of deaths of covid affected patients.

world_daily_death_avg - It has the data of moving averages of the world daily deaths.

We have to pass some data to the machine learning model after reshaping it into processable data format

days_since_1_22 - It has all the dates, reshaped into a particular format of data.**future_forecast_dates** - It has the dates of next few days, of which we want to

predict cases for.

The days_since_1_22 and world_cases data is split into Training and Testing data using train_test_split() inbuilt function with test.size = 0.10

Training Data = **X_train_confirmed + y_train_confirmed X_train_confirmed** - It has the data of dates to train the model. **y_train_confirmed** - It has the world cases to train the model.

Testing Data = **X_test_confirmed + y_test_confirmed**

X_test_confirmed - It has the data of dates to test the model.

y_test_confirmed - It has the world cases to test the model.

Output predictions of machine learning algorithms are also stored inorder to visualize the results.

svm_pred - It stores the output predictions of the Support Vector Machine model from the training data.

svm_test_pred - It stores the output predictions of the Support Vector Machine from the testing data.

Output predictions of Polynomial regression model

linear_pred - It has the output prediction of the Polynomial Regression model from the training dataset.

test_linear_pred - It has the output prediction of the Polynomial Regression model from the testing dataset.

Output predictions of Bayesian ridge regression model

bayesian_pred - It has the output prediction of the Bayesian Ridge Regression model from the training dataset.

test_bayesian_pred - It has the output prediction of the Bayesian Ridge Regression model from the testing dataset.

4.2.6 ER Diagram :

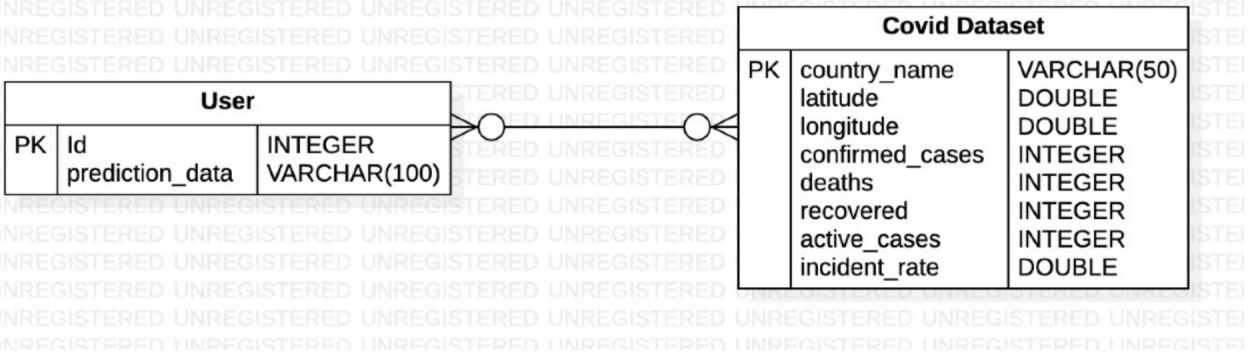
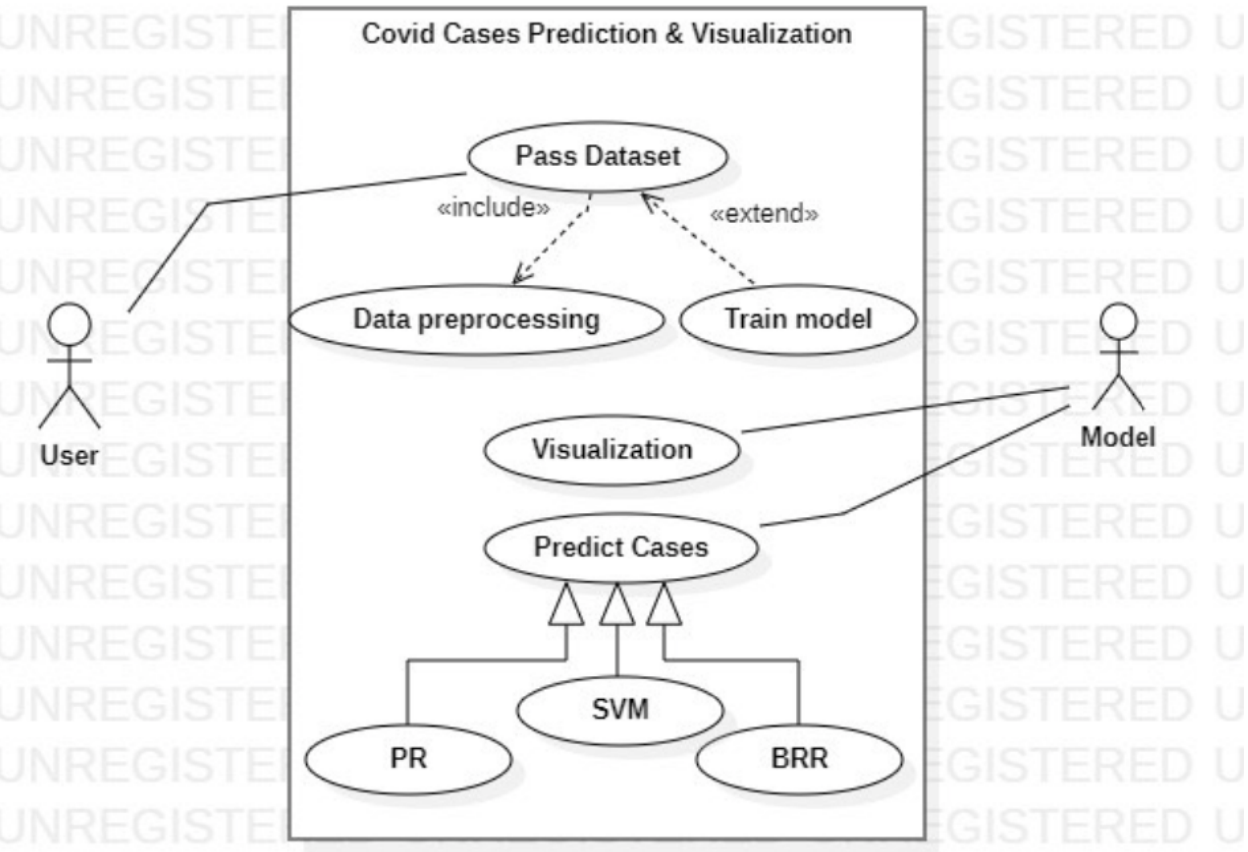


Fig. 4 Entity Relationship Diagram

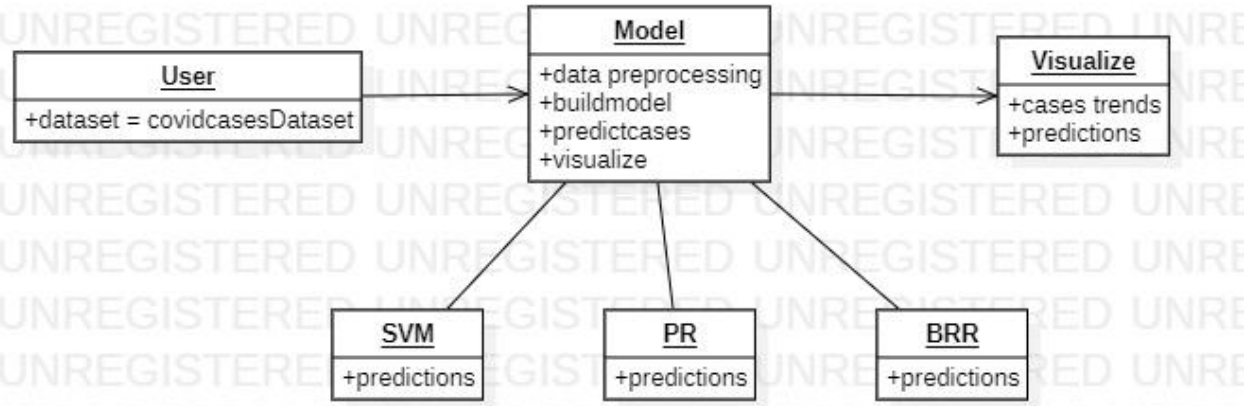
The ER-diagram shows the various entities present and the relationship between them. We have many entities: User, Question, Various tasks. They are connected by the relationships. The various tasks have many attributes and one primary key.

4.2.7 Use Case Diagram:

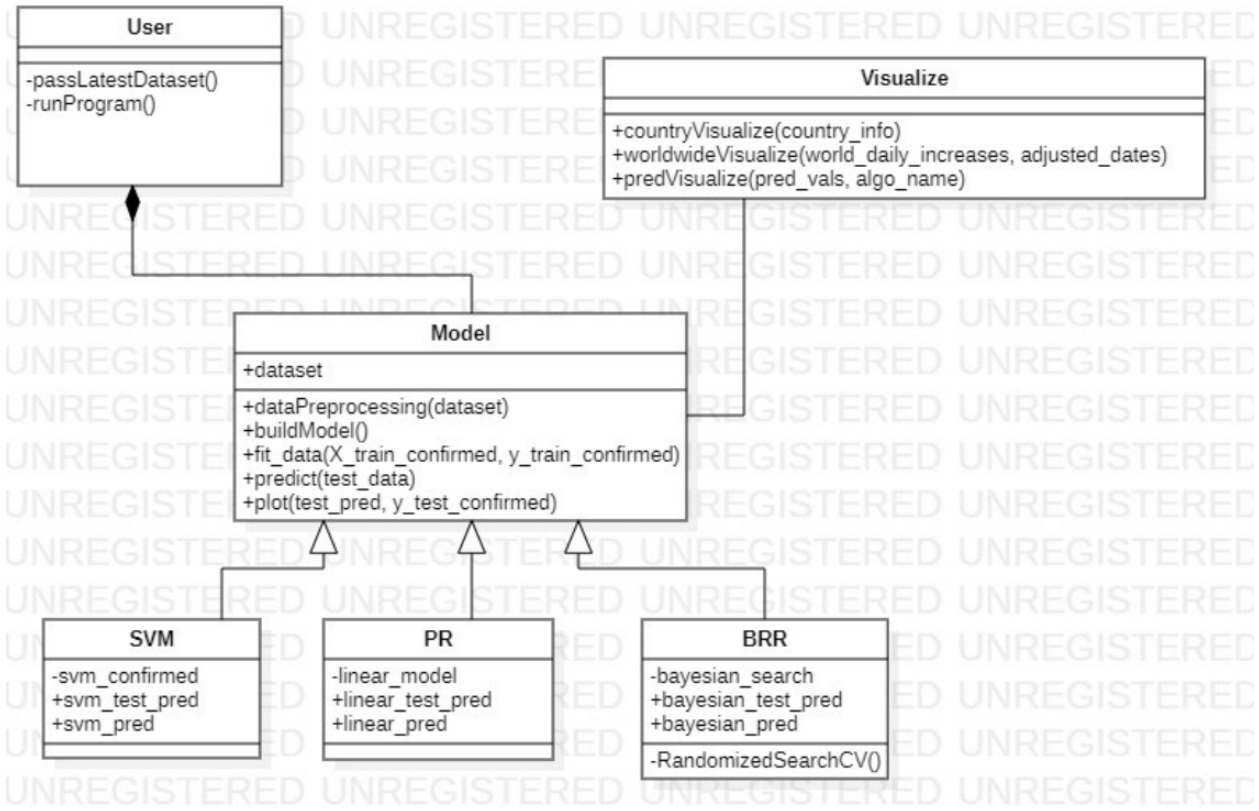


Use case diagram shows the graphical interface of the user's interaction with the system. The user can tell assistant to perform many actions as shown.

4.2.8 Object Diagram:

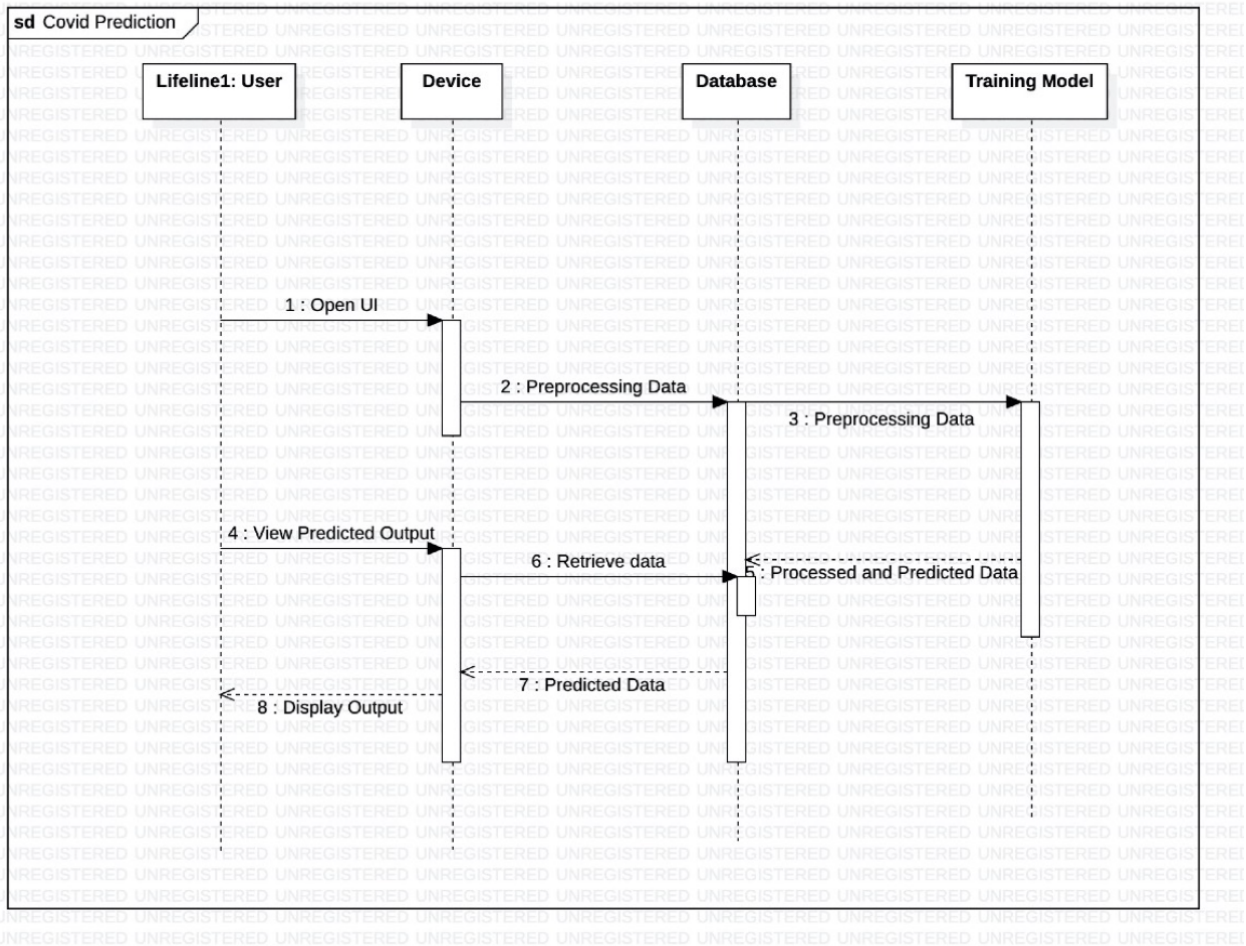


4.2.9 Class Diagram



Class diagram shows the various classes involved in the application. We have many classes: User, Task, Question . They have different attributes and methods. Task classes are having generalization relationship to the parent class Task.

4.2.10 Sequence Diagram:



Sequence diagram shows the interaction between the objects of the system. We have four objects: User, Assistant, Device, Cloud. The user sends the voice input to the assistant which selects the required object for further help in answering the query

4.2.11 Collaboration Diagram:

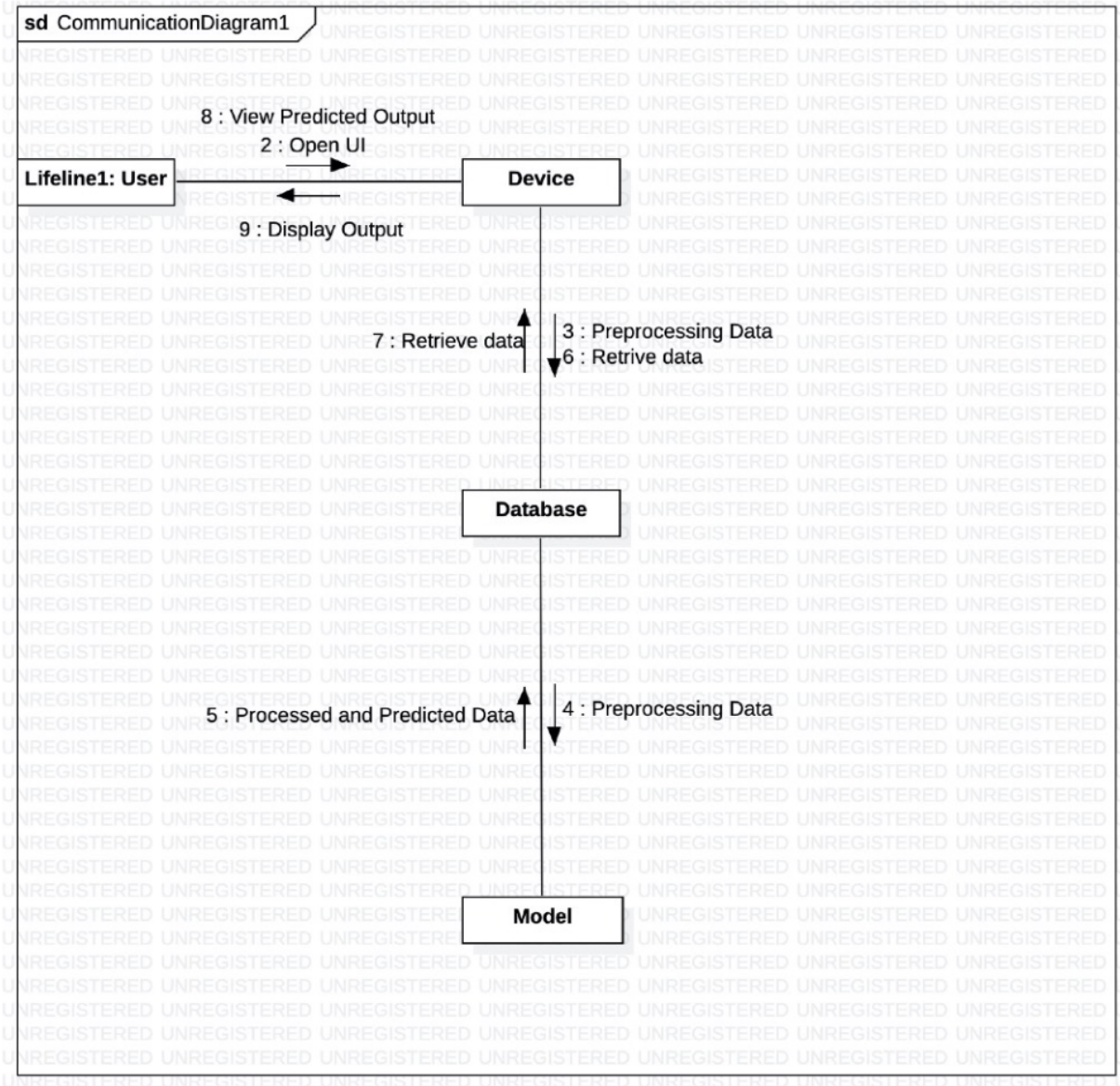


Fig. 8 Collaboration Diagram

Collaboration diagram focuses on the message passing between objects.

4.2.12 Activity Diagram:

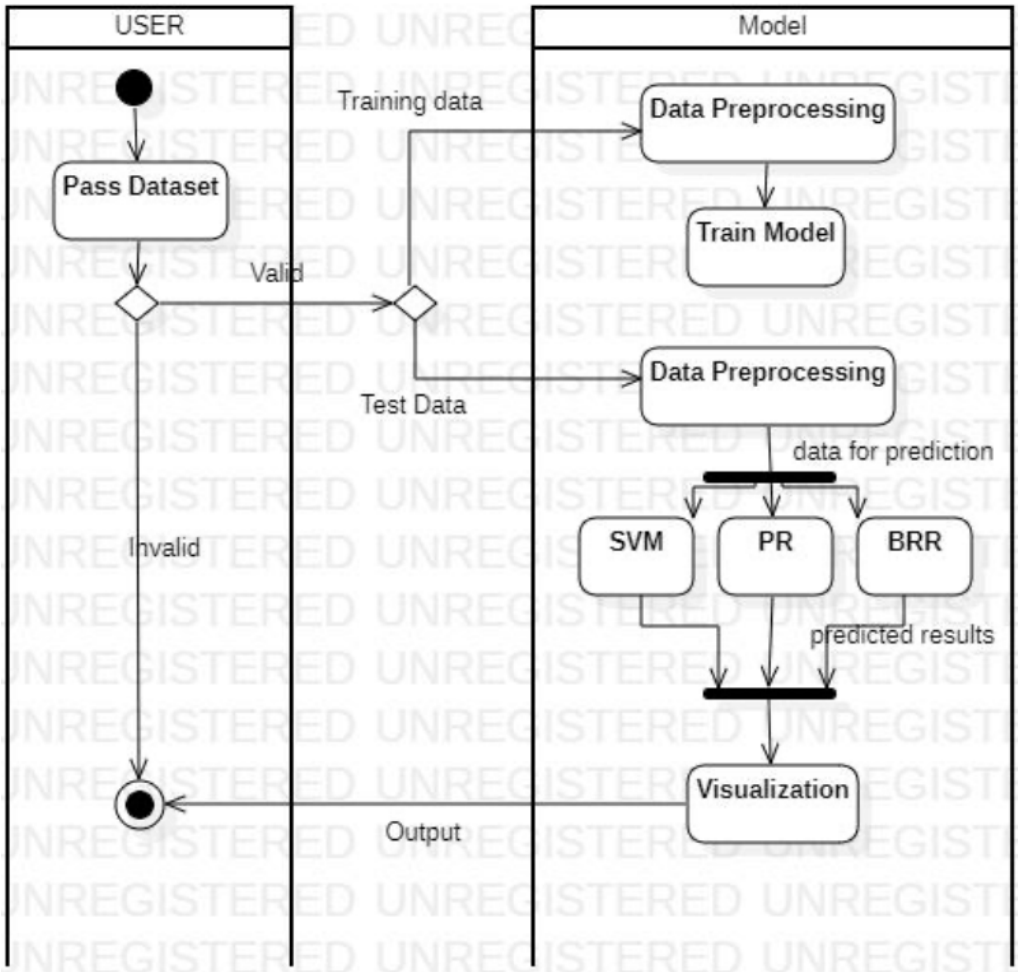


Fig. 9 Activity Diagram

Activity diagram represents the flow from one activity to another.

4.2.13 State Diagram

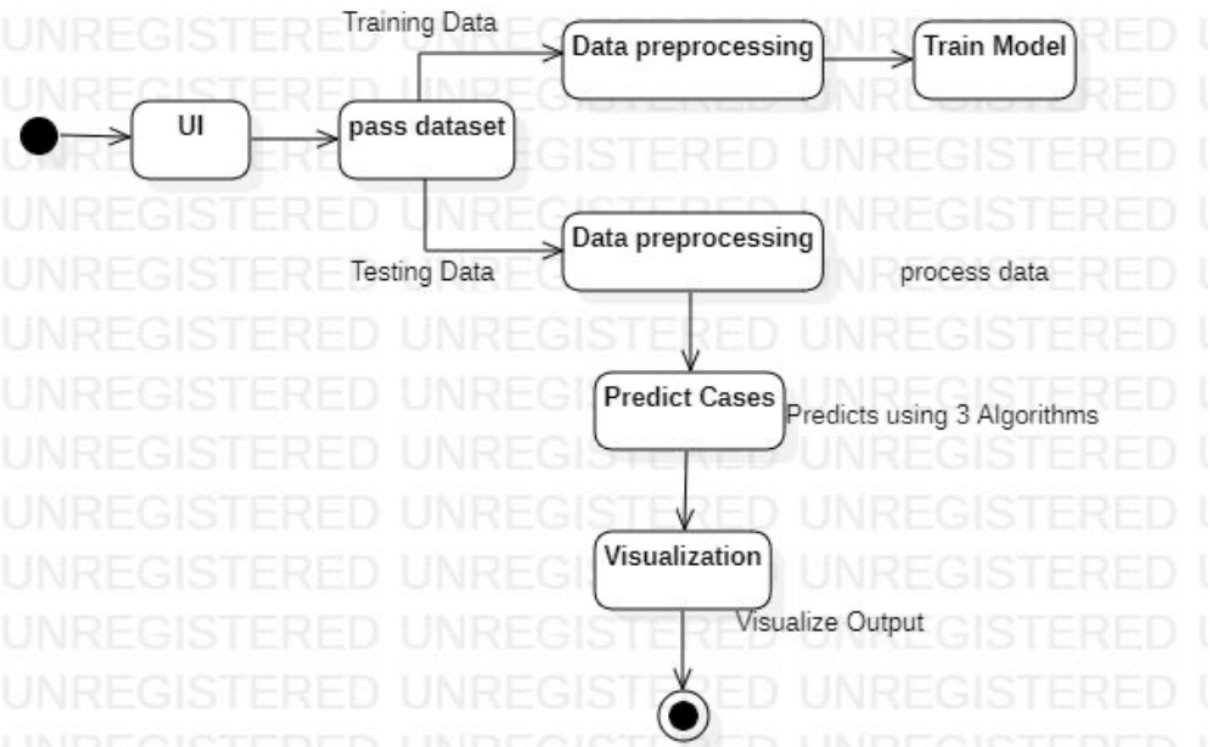


Fig. 10 State Diagram

State diagram shows how the process occurs in different states.

4.2.14 Component Diagram

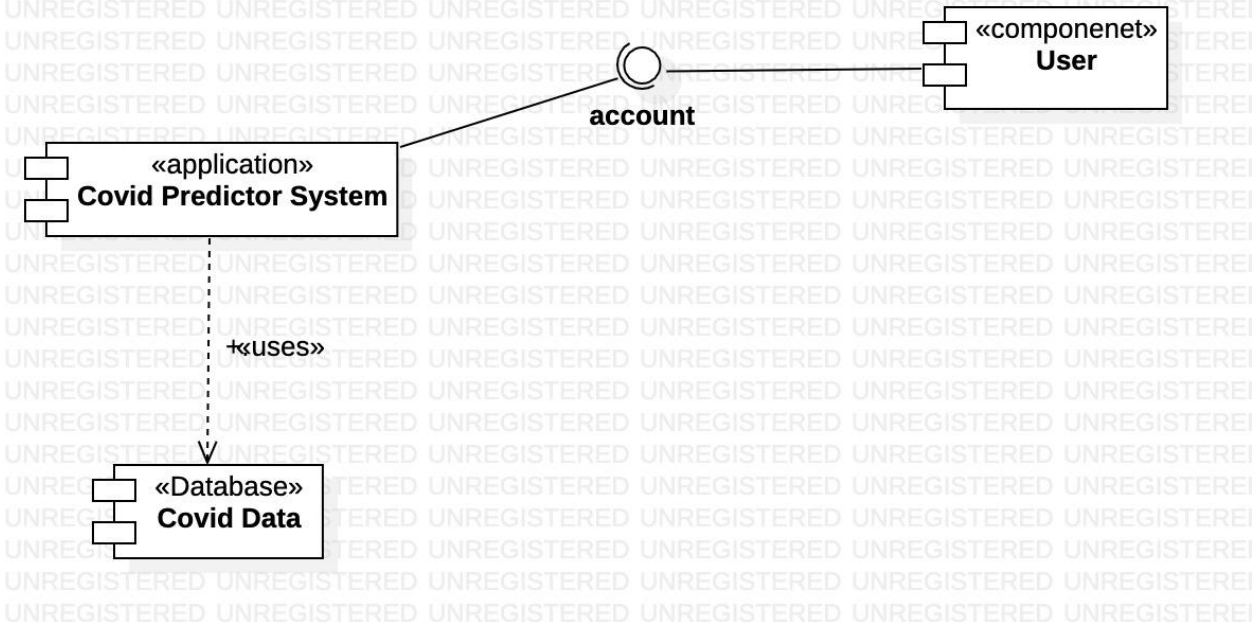


Fig. 11 Component Diagram

Component diagram shows the various components in the system. We have three different components: User, Assistant, Cloud Server.

4.2.15 Deployment Diagram

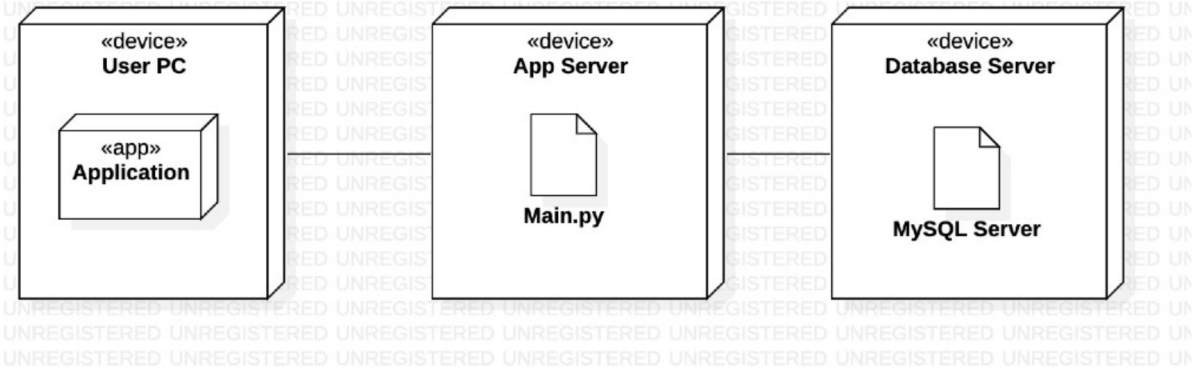


Fig. 12 Deployment Diagram

Deployment diagram shows how the application is deployed, both the hardware and software.

4.3 Implementation

Anaconda Jupiter:

Jupyter Notebook, an open-source, web-based IDE with deep cross-language integration that allows you to create and share documents containing live code, equations, visualizations, and narrative text.

Data:

The Covid - data is taken from the GitHub repository.

Coronavirus Case Data is provided by [Johns Hopkins University]

Outputs:

Future Prediction on following days worldwide(SVM)

	Date	SVM Predicted # of Confirmed Cases Worldwide
0	03/26/2021	157332643.000000
1	03/27/2021	158420364.000000
2	03/28/2021	159513157.000000
3	03/29/2021	160611033.000000
4	03/30/2021	161714003.000000

5	03/31/2021	162822079.000000
6	04/01/2021	163935274.000000
7	04/02/2021	165053599.000000
8	04/03/2021	166177065.000000
9	04/04/2021	167305685.000000

Future Prediction on following days worldwide(PolynomialRegression)

	Date	Polynomial Predicted # of Confirmed Cases Worldwide
0	03/26/2021	128809244.000000
1	03/27/2021	129219431.000000
2	03/28/2021	129624278.000000
3	03/29/2021	130023699.000000
4	03/30/2021	130417609.000000

5	03/31/2021	130805919.000000
6	04/01/2021	131188543.000000
7	04/02/2021	131565393.000000
8	04/03/2021	131936380.000000
9	04/04/2021	132301415.000000

Future Prediction on following days worldwide(BayesianRegression)

	Date	Bayesian Ridge Predicted # of Confirmed Cases Worldwide
0	03/26/2021	132088572.000000
1	03/27/2021	132629779.000000
2	03/28/2021	133168372.000000
3	03/29/2021	133704293.000000
4	03/30/2021	134237480.000000

5	03/31/2021	134767873.000000
6	04/01/2021	135295410.000000
7	04/02/2021	135820030.000000
8	04/03/2021	136341672.000000
9	04/04/2021	136860271.000000

5. Result Analysis and Alternative Solutions/Recommendations

- Covid cases visualization and predictions helps in bringing awareness to people about the covid situation and cases that are happening in the world. People will be able to know about the danger they may face in future and take precautions in advance in order to be safe and healthy in future.

- This can be used for free by any person with a device which has internet connection and a jupyter notebook installed in their system . It will be very easy to use and easy to get information about the covid situation in our world.

6. Conclusion and Discussion

- Information and communication technology help in the decision- making process based on the past data with the data analytics and data mining perspective. The size of data available is huge and gathering information and getting an interesting pattern out of the accumulated data is a challenging task. With the prevailing data about confirmed, recovered and death across India for over the time duration helps in predicting and forecasting the near future. The correctness of the model could be increased by introducing related attributes like several hospitals, the immune system of the infected person, age of the patient, gender of the patient, steps taken to combat the proliferation of the virus, and so on to make it completely informative. As of now, it's very prudent that yards to carry needs to be stringent and vigil in nature to handle this crucial situation by social distancing, lockdown, curfew, quarantine, and isolation to prevent the transmission.

- By seeing the predicted values and matching with cases from Covid India Cases we can conclude that the Bayesian Ridge Regression has better accuracy than both linear/polynomial regression and SVM.In future we can work with some deep learning methods for forecasting time series data for getting better predictions.

7.References:

https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

https://towardsdatascience.com/machine-learning-basics-polynomial-regression-3f9dd30223d1

https://towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7

https://scikit-learn.org/stable/auto_examples/linear_model/plot_bayesian_ridge.html

https://journals.plos.org/plosone/article?id=10.1371/ journal.pone.0239960