#### A Major Project Report on

**PREDICTION OF CORONA CASES USING MACHINE LEARNING**

#### Submitted to

**Jawaharlal Nehru Technological University, Hyderabad *in partial fulfillment of requirements for the award of the degree of* BACHELOR OF TECHNOLOGY**

**IN**

### COMPUTER SCIENCE ANDENGINEERING

#### By

#### 

#### NEELIMA RAMPALLI (18E41A0520)

#### NIKITHA PEDOLLA (18E41A0510)

#### NITHISH VELDHANDA (18E41A0517)

#### ANKITHA MANDA (18E41A0519)

#### Under the guidance of

Department of CSE

### Department of Computer Science and Engineering SREE DATTHA INSTITUTE OF ENGINEERING & SCIENCE

Approved by AICTE, Affiliated to JNTUH

**Sheriguda, Ibrahimpatnam, R.R. Dist. Hyderabad**.

**2(f) recognition by UGC Accredited by NAAC for 6 years**

### SREE DATTHA INSTITUTE OF ENGINEERING & SCIENCE

**(Accredited by NAAC, Approved By A.I.C.T.E, Affiliated to JNTU, Hyderabad)**

**2(f) recognition by UGC Accredited by NAAC for 6 years**

### DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

**CERTIFICATE**

This is to certify that the project entitled “**PREDICTION OF CORONA CASES USING MACHINE LEARNING”** being submitted by

#### NEELIMA RAMPALLI (18E41A0520)

**NITHISH VELDHANDA (18E41A0517)**

#### NIKITHA PEDOLLA (18E41A0510)

**ANKITHA MANDA (18E41A0519)**

In partial fulfillment for theaward of the degree of **Bachelor of Technology** in **Computer Science and Engineering** affiliated to the **Jawaharlal Nehru Technological University, Hyderabad.**

#### Internal Guide HOD-Incharge CSE

(**Prof Dr. Md. SAMEERUDDIN KHAN**) (**Ms. RUHITHSULTANA)**

**External Examiner**

We hereby declare that the project report entitled “**PREDICTION OF HEART DISEASES USING MACHINE LEARNING”** is done

in the partial fulfillment for the award of the Degree in Bachelor of Technology in Computer Science and Engineering from Sree Dattha Institute of Engineering & Science affiliated to Jawaharlal Nehru Technological University, Hyderabad. This project has not been submitted anywhere else.

**NEELIMA RAMPALLI (18E41A0520)**

**NIKITHA PEDOLLA (18E41A0510)**

**NITHISH VELDHANDA (18E41A0517)**

**ANKITHA MANDA (18E41A0519)**

We are extremely grateful to **Shri G. Pandu Ranga Reddy**, Chairman and

**Prof Dr. Md. Sameeruddin Khan,** Principal and **MS. Ruhiat Sultana** Incharge Head of the Department of CSE, Sree Dattha Institute of Engineering &Science.

We are extremely thankful to **Prof Dr. Md. Sameeruddin Khan,** internal guide, Department of CSE, for his constant guidance, encouragement, and moral support throughout the project.

We will be failing in duty if we do not acknowledge with grateful thanks to the authors of the references and other literatures referred in this Project.

We express our thanks to all staff members and friends for all the help and co-ordination extended in bringing out this Project successfully in time.

Finally, we are very much thankful to our parents who guided us for every step.

### NEELIMA RAMPALLI (18E41A0520)

### NIKITHA PEDOLLA (18E41A0510)

### NITHISHVELDHANDA (18E41A0517) ANKITHA MANDA (18E41A0519)

**Place: Date:**

**LIST OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **CONTENTS** | **PAGE NO** |
|  | **MAIN PAGE** | i |
|  | **CERTIFICATION** | ii |
|  | **DECLARATION** | iii |
|  | **ACKNOWLEDGMENT** | iv |
|  | **ABSTRACT** | 1 |
| **1** | **INTRODUCTION** | 3 |
| **2** | **LITERATURE SURVEY** | 5  5  8  10  13  15 |
|  | 2.1 MACHINE LEARNING |
|  | 2.2 REGRESSION |
|  | 2.3 RANDOM FOREST |
|  | 2.4 RANDOM FORESTREGRESSOR |
|  | 2.5 DATA SET |
| **3** | **SYSTEM ANALYSIS** | 19 |
| **4** | **SYSTEM OVERVIEW ANDREQUIREMENT**  4.1 FUNCTIONAL REQUIREMENTS | 21  21 |
|  | 4.2 NON FUNCTIONAL REQUIREMENTS | 21 |
|  | * 1. SOFTWARE REQUIREMENTS   2. HARDWARE REQUIREMENTS | 22  22 |
| **5** | **SOFTWARE TOOLS AND DESCRIPTION** | 24 |
|  | 5.1 PYTHON | 24 |
|  | 5.2 LIBRARIES | 25 |
| **6** | **OBJECTIVES** | 29 |
| **7** | **SYSTEM DESIGN**  7.1 SYSTEM ARCHITECTURE | 31  31  33  34  36  38 |
|  | 7.2 INTRODUCTION TO UML |
|  | 7.2.1 USE CASE DIAGRAM |
|  | 7.2.2 SEQUENCE DIAGRAM |
|  | 7.2.3 CLASS DIAGRAM |

|  |  |  |
| --- | --- | --- |
| **8** | **IMPLEMENTATION**   * 1. PSEUDO CODE   2. CODE SNIPPETS | 41  41  42 |
| **9** | **TESTING** | 47 |
| **10** | **RESULTS** | 51 |
| **11** | **CONCLUSION** | 59 |
| **12** | **FUTURE SCOPE** | 61 |
| **13** | **REFERENCES** | 63 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **LIST OF FIGURES** | **PAGE NO** |
| 2.1 | Machine Learning Categories | 6 |
| 2.1.1 | Traditional Programming vs. Machine Learning | 7 |
| 2.2 | Regression | 8 |
| 2.3 | Random Forest | 11 |
| 2.4 | Random Forest Regressor | 13 |
| 7.1 | System Architecture | 31 |
| 7.2.1 | Use Case Diagram | 35 |
| 7.2.2 | Sequence Diagram | 37 |
| 7.2.3 | Class Diagram | 39 |
| 10.1 | Importing Packages and read Dataset | 51 |
| 10.2.1 | Data Analysis | 51 |
| 10.2.2 | Data Analysis (Cont.) | 52 |
| 10.3.1 | Data Pre-Processing | 52 |
| 10.3.2 | Data Pre-Processing(Cont.) | 53 |
| 10.3.3 | Data Pre-Processing(Cont.) | 53 |
| 10.3.4 | Data Pre-Processing(Cont.) | 54 |
| 10.4.1 | Correlation Matrix | 54 |
| 10.4.2 | Correlation Matrix | 55 |
| 10.4.3 | Data Pre-Processing(Cont.) | 55 |
| 10.4.4 | Data Classification | 56 |
| 10.5 | K Nearest Neighbors(KNN) | 56 |
| 10.6 | Decision Tree | 57 |
| 10.7 | Random Forest | 57 |

## ABSTRACT

Effective screening of SARS-CoV-2 enables quick and efficient diagnosis of COVID-19 and can mitigate the burden on healthcare systems. Prediction models that combine several features to estimate the risk of infection have been developed. These aim to assist medical staff worldwide in triaging patients, especially in the context of limited healthcare resources. We established a machine-learning approach that trained on records from 51,831 tested individuals (of whom 4769 were confirmed to have COVID-19). The test set contained data from the subsequent week (47,401 tested individuals of whom 3624 were confirmed to have COVID-19). Our model predicted COVID-19 test results with high accuracy using only eight binary features: sex, age ≥60 years, known contact with an infected individual, and the appearance of five initial clinical symptoms. Overall, based on the nationwide data publicly reported by the Israeli Ministry of Health, we developed a model that detects COVID-19 cases by simple features accessed by asking basic questions. Our framework can be used, among other considerations, to prioritize testing for COVID-19 when testing resources are limited.

# INTRODUCTION

## INTRODUCTION

The novel coronavirus disease 2019 (COVID-19) pandemic caused by the SARS-CoV-2 continues to pose a critical and urgent threat to global health. The outbreak in early December 2019 in the Hubei province of the People’s Republic of China has spread worldwide. As of October 2020, the overall number of patients confirmed to have the disease has exceeded 39,500,000, in >180 countries, though the number of people infected is probably much higher. More than 1,110,000 people have died from COVID-19.

This pandemic continues to challenge medical systems worldwide in many aspects, including sharp increases in demands for hospital beds and critical shortages in medical equipment, while many healthcare workers have themselves been infected. Thus, the capacity for immediate clinical decisions and effective usage of healthcare resources is crucial. The most validated diagnosis test for COVID-19, using reverse transcriptase polymerase chain reaction (RT-PCR), has long been in shortage in developing countries. This contributes to increased infection rates and delays critical preventive measures.

Effective screening enables quick and efficient diagnosis of COVID-19 and can mitigate the burden on healthcare systems. Prediction models that combine several features to estimate the risk of infection have been developed, in the hope of assisting medical staff worldwide in triaging patients, especially in the context of limited healthcare resources. These models use features such as computer tomography (CT) scans, clinical symptoms, laboratory tests, and an integration of these features. However, most previous models were based on data from hospitalized patients, thus are not effective in screening for SARS-CoV-2 in the general population.

The Israeli Ministry of Health publicly released data of all individuals who were tested for SARS-CoV-2 via RT-PCR assay of a nasopharyngeal swab. During the first months of the COVID-19 pandemic in Israel, all diagnostic laboratory tests for COVID-19 were performed according to criteria determined by the Israeli Ministry of Health. While subject to change, the criteria implemented during the study period included the presence and severity of clinical symptoms, possible exposure to individuals confirmed to have COVID-19, certain geographical areas, and the risk of complications if infected. Except for a small minority who were tested under surveys among healthcare workers, all the individuals tested had indications for testing. Thus, there was no apparent referral bias regarding the vast majority of the subjects in the dataset used in this study; this contrasts with previous studies, for which such bias was a drawback. In addition, all negative and positive COVID-19 cases this dataset were confirmed via RT-PCR assay.

# LITERATURE SURVEY

## 2. LITERATURESURVEY

### Machine learning

Tom Mitchell states machine learning as “A computer program is said to learn from experience and from some tasks and some performance on, as measured by, improves with experience”. Machine Learning is combination of correlations and relationships, most machine learning algorithms in existence are concerned with finding and/or exploiting relationship between datasets. Once Machine Learning Algorithms can pinpoint on certain correlations, the model can either use these relationships to predict future observations or generalize the data to reveal interesting patterns. In Machine Learning there are various types of algorithms such as Regression, Linear Regression, Logistic Regression, Naive Bayes Classifier, Bayes theorem, KNN (K-Nearest Neighbor Classifier),Decision Tress, Entropy, ID3, SVM (Support Vector Machines), K-means Algorithm, Random Forest and etc.,

The name machine learning was coined in 1959 by Arthur Samuel. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data Machine learning is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making through the use of computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. Machine learning is sometimes conflated with data mining, where the latter subfield focuses more on exploratory data analysis and is known as unsupervised learning.

Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as predictive analytics. These analytical models allow researchers, data scientists, engineers, and analysts to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data.

An external file that holds a picture, illustration, etc.
Object name is f20-01-9780128245361.jpg

Machine learning tasks Machine learning tasks are typically classified into several broad categories:

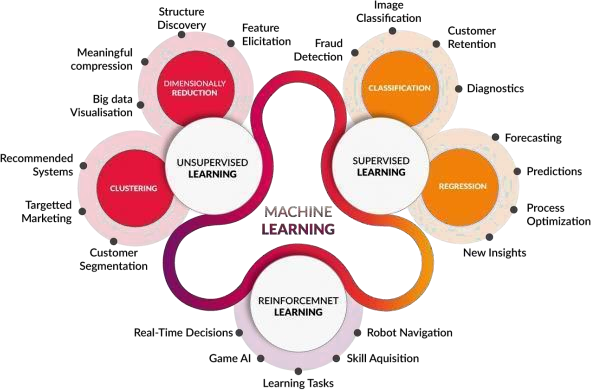
**Supervised learning**: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs. As special cases, the input signal can be only partially available, or restricted to special feedback.

**Semi-supervised learning**: The computer is given only an incomplete training signal, a training set with some (often many) of the target outputs missing.

**Active learning**: The computer can only obtain training labels for a limited set of instances (based on a budget), and also has to optimize its choice of objects to acquire labels for. When used interactively, these can be presented to the user for labelling.

**Reinforcement learning**: Data (in form of rewards and punishments) are given only as feedback to the program's actions in a dynamic environment, such as driving a vehicle or playing a game against an opponent.

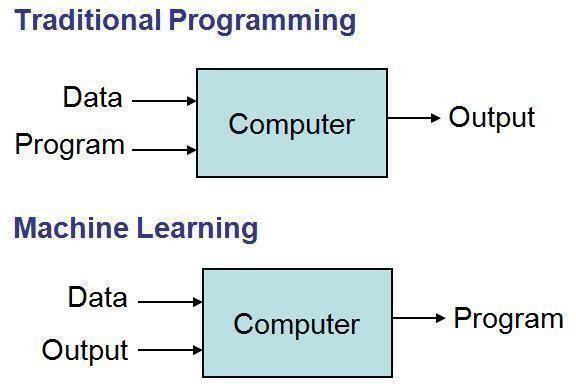
**Unsupervised learning**: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).



**Fig2.1-MachineLearningCategories**

### Features of machine learning

* It is nothing but automating the Automation.
* Getting computers to program themselves.
* Writing Software is bottleneck.
* Machineleaningmodelsinvolvesmachineslearningfromdatawithoutthehelpof humans or any kind of human intervention.
* Machine Learning is the science of making the computers learn and act like humans by feeding data and information without being explicitly programmed.
* Machine Learning is a combination of Algorithms, Datasets, and Programs.
* Machine Learning is totally different from traditionally programming, here data and output is given to the computer and in return it gives us the program which provides solution to the various problems. Below is the figure.



**Fig 2.1.1 Traditional Programming vs. Machine Learning**

### RGRESSION

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as **temperature, age, salary, price,** etc.

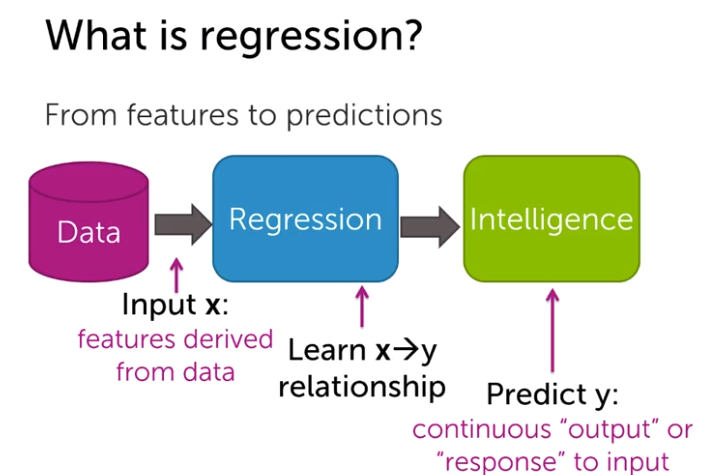
Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for **prediction, forecasting, time series modelling, and determining the causal-effect relationship between variables**.

In Regression, we plot a graph between the variables which best fits the given data-points, using this plot, the machine learning model can make predictions about the data. In simple words, **"Regression shows a line or curve that passes through all the data-points on target-predictor graph in such a way that the vertical distance between the data-points and the regression line is minimum."** The distance between data-points and line tells whether a model has captured a strong relationship or not.

**2.2.1 Types of Regression**

There are various types of regressions which are used in data science and machine learning. Each type has its own importance on different scenarios, but at the core, all the regression methods analyze the effect of the independent variable on dependent variables. Here we are discussing some important types of regression which are given below:

* Linear Regression
* Logistic Regression
* Polynomial Regression
* Support Vector Regression
* Decision Tree Regression
* Random Forest Regression
* Ridge Regression
* Lasso Regression

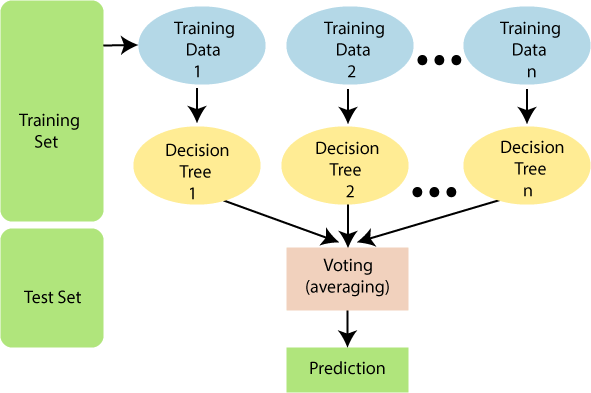


### Random Forest

Random Forest can be used for both classification and regression problems. Random Forest algorithm is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach.

The decision tree is a decision support tool. It uses a tree-like graph to show the possible consequences. If you input a training dataset with targets and features into the decision tree, it will formulate some set of rules. These rules can be used to perform predictions. Through the decision tree algorithm, you can generate the rules. You can then input the features of this movie and see whether it will be liked by your daughter.

The process of calculating these nodes and forming the rules is using information gain and Gini index calculations. The difference between Random Forest algorithm and the decision tree algorithm is that in Random Forest, the processes of finding the root node and splitting the feature nodes will run randomly.

Over fitting is one critical problem that may make the results worse, but for Random Forest algorithm, if there are enough trees in the forest, the classifier won’t over fit the model. The third advantage is the classifier of Random Forest can handle missing values, and the last advantage is that the Random Forest classifier can be modeled for categorical values.

.

#### Fig2.4- Random Forest

**Random Forest creation pseudocode**:

1. Randomly select “K” features from total “m” features where k <<m
2. Among the “K” features ,calculate the node “d” using the best split point
3. Split the node into daughter nodes using the best split
4. Repeat the a to c steps until “l” number of nodes has been reached
5. Build forest by repeating steps a to d for “n” number times to create “n” number of trees.

In the next stage, with the random forest classifier created, we will make the prediction. The random forest prediction pseudocode is shown below:

Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target). Calculate the votes for each predicted target

Consider the high voted predicted target as the final prediction from the random forest algorithm

### Advantages:

* + Compared with other classification techniques, there are three advantages as the author mentioned.
  + For applications in classification problems, Random Forest algorithm will avoid the Over fitting problem.
  + For both classification and regression task, the same random forest algorithm can be used.
  + The Random Forest algorithm can be used for identifying the most important features from the training dataset, in other words, feature engineering’s.

### Random Forest Regressor

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as **bagging**. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.  
Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

We need to approach the Random Forest regression technique like any other machine learning technique

* Design a specific question or data and get the source to determine the required data.
* Make sure the data is in an accessible format else convert it to the required format.
* Specify all noticeable anomalies and missing data points that may be required to achieve the required data.
* Create a machine learning model
* Set the baseline model that you want to achieve
* Train the data machine learning model.
* Provide an insight into the model with test data
* Now compare the performance metrics of both the test data and the predicted data from the model.
* If it doesn’t satisfy your expectations, you can try improving your model accordingly or dating your data or use another data modeling technique.
* At this stage you interpret the data you have gained and report accordingly.

**Below is a step by step sample implementation of Rando Forest Regression.**

* Step 1 : Import the required libraries.
* Step 2 : Import and print the dataset.
* Step 3 : Select all rows and column 1 from dataset to x and all rows and column 2 as y.
* Step 4 : Fit Random forest regressor to the dataset.
* **Step 5 :** Predicting a new result
* **Step 6 :** Visualising the result

### Data set

The dataset consists of 50 individual data. There are 7 columns in the dataset. Our dataset contains the following attributes:

**Age**:

It displays the age of the individual.

**Weather**:

It displays the weather of the individual region using the following format: 7-10 = winter

4-6= Rainy

0-3= summer

#### Vaccinated:

It displays whether a person is vaccinated or not

#### Crowdy:

Crowdy is a measure of population or number of people staying in a particular region. It is indicated basis on the 7-10 = More crowded place

4-6 = Moderately crowded place

0-3 = Less crowded place

#### Lockdown:

This shows whether there is a lockdown imposed or not.

#### Symptoms:

It shows that what kind of symptoms does a particular person is having.

#### Cases Reported:

#### 

#### Finally it shows the chances of cases being reported in a particular area.

# SYSTEM ANALYSIS

## SYSTEMANALYSIS

### Existing System

Existing system consists of identifying corona by performing tests such as RTPCR and Rapid tests which take long time and where there be more spread of corona by that time gap.

### Proposed System

After evaluating the results from the existing methodologies, we have used python operations to perform corona prediction for the data obtained from the dataset. It provides an easy-to-use visual representation of the dataset, working environment and building the predictive analytics.ML process starts from a preprocessing data phase followed by feature selection based on data cleaning, classification of modelling performance evaluation.

# SYSTEM OVERVIEW AND REQUREMENTS

## SYSTEM OVERVIEW AND RERUIREMENTS

### Functional Requirements

For documenting the functional requirements, the set of functionalities supported by the system are to be specified. A function can be specified by identifying the state at which data is to be input to the system, its input data domain, the output domain, and the type of processing to be carried on the input data to obtain the output data. Functional requirements define specific behavior or function of the application. Following are the functional requirements:

FR1) the system should allow administrator to monitor and remove inappropriate datasets and code.

FR2) the system allows the users to predict corona cases.

FR3) the system allows users to input attributes. FR4) Predict chances with the given symptoms.

FR5) Compare the given symptoms with the input datasets.

### Non-Functional Requirements

A non-functional requirement is a requirement that specifies criteria that can be used to judge the operation of a system, rather than specific behaviors. Especially these are the constraints the system must work within. Following are the non- functional requirements:

NFR 1) the website should provide values used during prediction to the user.

NFR 2) the website should be responsive and consistent.

NFR 3) the model should display the correct chance of getting corona.

### Software Requirements

Operating System : Windows 10 or MAC OS. Platform : Jupyter-Lab Programming Language : Python

|  |  |  |
| --- | --- | --- |
| **4.4 Hardware** | **Requirements** |  |
| Processor | : | Intel core i3 and above. |
| Hard Disk | : | 100 GB or above. |
| RAM | : | 1 GB or above. |
| Internet | : | 4 Mbps or above (Wireless). |

# SOFTWARE TOOLS AND DESCRIPTION

## SOFTWARE TOOLS AND DESCRIPTION

### PYTHON

PythonisoneofthemostpopularprogramminglanguagesinboththecodingandData Science communities. Guido Van Rossum created it in 1991 and ever since its inception has been one of the most widely used languages along with C++, Java, etc. Python is an open- source, high-level, general-purpose programming language that incorporates the features of object-oriented,structural,andfunctionalprogramming.WhilePython’ssimplesyntaxallows for writing readable code, which can be further applied to complex software development processes to facilitate test-driven software application development, machine learning, and data analytics. Python can run on all the major operating systems, including Windows, Linux, and iOS.

Since it functions on cross-platform operating systems, Python can be used to develop a host of applications, including web apps, gaming apps, enterprise-level applications, ML apps, image processing, text processing, and so much more.

Since it functions on cross-platform operating systems, Python can be used to develop a host of applications, including web apps, gaming apps, enterprise-level applications, ML apps, image processing, text processing, and so much more.

But beyond its innate simplicity and versatility, what makes Python stand out are its vast assortments of libraries and packages that can cater to a wide range of development as well as Data Science requirements.

* Python has Prebuilt Libraries like Numpy for scientific computation, Scipy for advanced computing and Pybrain for machine learning (Python Machine Learning) making it one of the best languages for AI.
* Python developers around the world provide comprehensive support and assistance via forums and tutorials making the job of the coder easier than any other popular languages.
* Python is platform Independent and is hence one of the most flexible and popular choices for use across different platforms and technologies with the least tweaks in basic coding.
* Python is the most flexible of all others with options to choose between OOPs approach and scripting. You can also use IDE itself to check for most codes and is a boon for developers struggling with different algorithms.

Python also supports data analyzation and visualization, there by further simplifying the process of creating custom solutions minus the extra effort and time investment.

### LIBRARIES

#### NUMPY:

NumPy stands for ‘Numerical Python’. It is an open-source Python library used to perform various mathematical and scientific tasks. It contains multi-dimensional arrays and matrices, along with many high-level mathematical functions that operate on these arrays and matrices. **1.Installing Numpy:**

You can install NumPy with:

##### conda install numpy

**or**

***pip install numpy***

#### How to importNumpy:

After installing NumPy,you can now use this library by importing it. To import NumPy use:

**import numpy asnp**

### Pandas:

Pandas is Python Data Analysis Library, pandas is an open source, BSD-licensed libraryprovidinghigh-performance,easy-to-usedatastructuresanddataanalysistools.Pandas is an open-source python package built on top of Numpy developed by Wes McKinney. It is used as one of the most important data cleaning and analysis tool. It provodies fast, flexible, and expressive datastructures.

Primary object types: int, float, object type

Data Frame: rows and columns (like a spreadsheet) Series: a single column

#### Installing pandas:

You can install Pandas by using the following commands:

# To install pandas in terminal or command line use one of the commands.

##### pip install pandas

or

##### conda install pandas

# To install pandas in jupyter note book use this commands.

##### !pip install pandas

* 1. **How to import Pandas: import pandas as pd**

By using the above command, you can easily import pandas library.

Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data — load, prepare, manipulate, model, and analyze. Python with Pandas is used in a wide range offields including academic and commercial domains including finance, economics, Statistics, analytics,etc.

### SCIKITLEARN

Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports Python numerical and scientific libraries like NumPy and SciPy. Scikit-learn is probably the most useful library for machine learning in Python.

#### Installing ScikitLearn:

For pip installation-

Run the following command in the terminal:

##### pip install scikit-learn

If you like conda, you can also use the conda for package installation Run the following command:

***conda install scikit-learn***

#### How to import Scikit-learn:

Once you are done with the installation, you can use scikit-learn easily in your Python code by importing it as:

***Import sklearn***

### MATPLOTLIB:

Matplotlib is an open-source plotting library in Python introduced in the year 2003. It is a very comprehensive library and designed in such a way that most of the functions for plotting in MATLAB can be used in Python.

It consists of several plots like the Line Plot, Bar Plot, Scatter Plot, Histogram etc. through which we can visualize various types of data.

1. **InstallingMatplotlib:**

#Windows, Linus, MacOS users can install this library using the following command:

***python -mpip install -U matplotlib***

#To install Matplotlib in Jupyter Notebook run the following command:

***pip install matplotlib***

#To install Matplotlib in Anaconda Prompt use the following command:

***conda install matplotlib***

1. **ImportingMatplotlib:**

#importing pyplot module from matplotlib

***from matplotlib import pyplot as plt***

or

***import matplotlib.pyplot as plt***

# OBJECTIVE

## OBJECTIVES

The novel coronavirus disease 2019 (COVID-19) pandemic caused by the SARS-CoV-2 continues to pose a critical and urgent threat to global health. The outbreak in early December 2019 in the Hubei province of the People’s Republic of China has spread worldwide. As of October 2020, the overall number of patients confirmed to have the disease has exceeded 39,500,000, in >180 countries, though the number of people infected is probably much higher. More than 1,110,000 people have died from COVID-19.

This pandemic continues to challenge medical systems worldwide in many aspects, including sharp increases in demands for hospital beds and critical shortages in medical equipment, while many healthcare workers have themselves been infected. Thus, the capacity for immediate clinical decisions and effective usage of healthcare resources is crucial. The most validated diagnosis test for COVID-19, using reverse transcriptase polymerase chain reaction (RT-PCR), has long been in shortage in developing countries. This contributes to increased infection rates and delays critical preventive measures.

Effective screening enables quick and efficient diagnosis of COVID-19 and can mitigate the burden on healthcare systems. Prediction models that combine several features to estimate the risk of infection have been developed, in the hope of assisting medical staff worldwide in triaging patients, especially in the context of limited healthcare resources. These models use features such as computer tomography (CT) scans, clinical symptoms, laboratory tests, and an integration of these features. However, most previous models were based on data from hospitalized patients, thus are not effective in screening for SARS-CoV-2 in the general population.

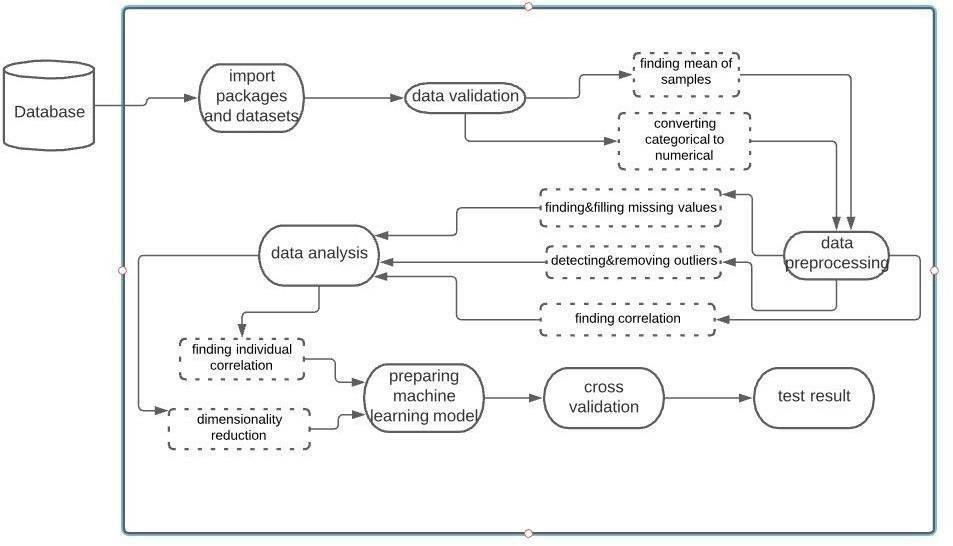
The Israeli Ministry of Health publicly released data of all individuals who were tested for SARS-CoV-2 via RT-PCR assay of a nasopharyngeal swab. During the first months of the COVID-19 pandemic in Israel, all diagnostic laboratory tests for COVID-19 were performed according to criteria determined by the Israeli Ministry of Health..

# SYSTEM DESIGN

## SYSTEMDESIGN

### System Architecture

The architecture of the proposed system is as displayed in the figure below. The major components of the architecture are as follows: patient database, preprocessing, tokenization, training the model, test the model, design fitness function, application of genetic algorithm, results collection and prediction of heart disease.



#### Fig 7.1 – System Architecture

**Data Validation**

* While data validation is a critical step in any data workflow, it’s often skipped over. It may seem as if data validation is a step that slows down your pace of work, however, it is essential because it will help you create the best results possible. These days data validation can be a much quicker process than you might’ve thought.

**Data**

**Preprocessing**

* In data preprocessing, we use different graphs and plots to visualize complex data to ease the discovery of data patterns. It involves the creation and study of the visual representation of data. To communicate information clearly and efficiently, data visualization uses statistical graphics, plots, information graphics and other tools.

**Data Analysis**

* + Data analysis is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision- making. The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis. It consists of correlation and dimensionality reduction.

**Data Prepration**

* Data preparation is the process of transforming raw data so that data scientists and analysts can run it through machine learning algorithms to uncover insights or make predictions.

**Cross Validation**

* Cross Validation is a technique to assess the performance of a statistical prediction model on an independent data set. Cross validation is conducted during the training phase where the user will assess whether the model is prone to underfitting or overfitting to the data.

**Test Result**

* Test Result is a technique using which we use to finally analyse the accuracy result and classification report obtained by predicting using the classification algorithms and select one specific algorithm which gives high accuracy rate, which means compared to the remaining algorithms the selected algorithm should work more effectively to get the correct and high accuracy rate.

### Introduction to UML

The Unified Modeling Language allows the software engineer to express an analysis model using the modeling notation that is governed by a set of syntactic, semantic and pragmatic rules. A UML system is represented using five different views that describe the system from distinctly different perspective. Each view is defined by a set of diagram, which is as follows:

* + 1. User Model View

This view represents the system from the users’ perspective. The analysis representation describes a usage scenario from the end-users’ perspective.

* + 1. Structural ModelView

In this model, the data and functionality are arrived from inside the system. This model view models the static structures.

* + 1. Behavioral ModelView

It represents the dynamic of behavioral as parts of the system, depicting he interactions of collection between various structural elements described in the user model and structural model view.

* + 1. Implementation Model View

In this view, the structural and behavioral as parts of the system are represented as they are to be built.

* + 1. Environmental ModelView

In this view, the structural and behavioral aspects of the environment in which the system is to be implemented are represented.

### UML DIAGRAMS

#### Use-Case Diagram

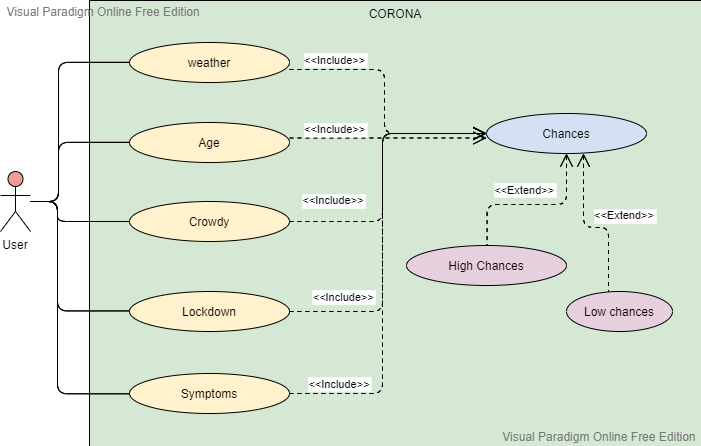
To model a system, the most important aspect is to capture the dynamic behavior. To clarify a bit in details, dynamic behavior means the behavior of the system when it is running/operating.

So only static behavior is not sufficient to model a system rather dynamic behavior is more important than static behavior. In UML there are five diagrams available to model dynamic nature and use case diagram is one of them. Now as we have to discuss that the use case diagram is dynamic in nature there should be some internal or external factors for making the interaction.

These internal and external agents are known as actors. So use case diagrams are consisting of actors, use cases and their relationships. The diagram is used to model the system/subsystem of an application. A single use case diagram captures a particular functionality of a system. So to model the entire system numbers of use case diagrams are used.

Use case diagrams are used to gather the requirements of a system including internal and external influences. These requirements are mostly design requirements. So when a system is analyzed to gather its functionalities use cases are prepared and actors are identified. In brief, the purposes of use case diagrams can be as follows:

1. Used to gather requirements of a system.
2. Used to get an outside view of a system.
3. Identify external and internal factors influencing the system.
4. Show the interacting among the requirements areactors.



#### Fig 7.2.1 – Use Case Diagram

* + 1. **Sequence Diagram**

Sequence diagrams describe interactions among classes in terms of an exchange of messages over time. They're also called event diagrams. A sequence diagram is a good way to visualize and validate various runtime scenarios. These can help to predict how a system will behave and to discover responsibilities a class may need to have in the process of modelling a new system.

The aim of a sequence diagram is to define event sequences, which would have a desired outcome. The focus is more on the order in which messages occur than on the message per se. However, the majority of sequence diagrams will communicate what messages are sent and the order in which they tend to occur.

#### Class Roles or Participants

Class roles describe the way an object will behave in context. Use the UML object symbol to illustrate class roles, but don't list object attributes.

#### Activation or Execution Occurrence

Activation boxes represent the time an object needs to complete a task. When an object is busy executing a process or waiting for a reply message, use a thin grey rectangle placed vertically on its lifeline.

#### Messages

Messages are arrows that represent communication between objects. Use half- arrowed lines to represent asynchronous messages.

Asynchronous messages are sent from an object that will not wait for are response from the receiver before continuing its tasks.

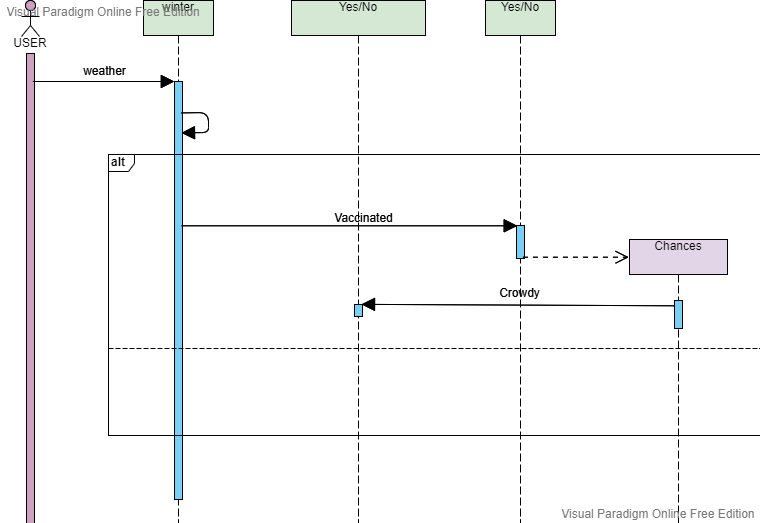
#### Lifelines

Lifelines are vertical dashed lines that indicate the object's presence over time.

#### Loops

A repetition or loop within a sequence diagram is depicted as a rectangle. Place the condition for exiting the loop at the bottom left corner in square brackets [].

When modelling object interactions, there will be times when a condition must be met for a message to be sent to an object. Guards are conditions that need to be used throughout UML diagrams to control flow.



#### Fig 7.2.2 – Sequence Diagram

* + 1. **Class Diagram**

Class diagrams are the main building blocks of every object oriented methods. The class diagram can be used to show the classes, relationships, interface, association, and collaboration. UML is standardized in class diagrams. Since classes are the building block of an application that is based on OOPs, so as the class diagram has appropriate structure to represent the classes, inheritance, relationships, and everything that OOPs have in its context. It describes various kinds of objects and the static relationship in between them.

The main purpose to use class diagrams are:

* + - 1. This is the only UML which can appropriately depict various aspects of OOPs concept.
      2. Proper design and analysis of application can be faster and efficient.
      3. It is base for deployment and component diagram.

Each class is represented by a rectangle having a subdivision of three compartments name, attributes and operation.

**Dependency:** A dependency is a semantic relationship between two or more classes where a change in one class cause changes in another class. It forms a weaker relationship.

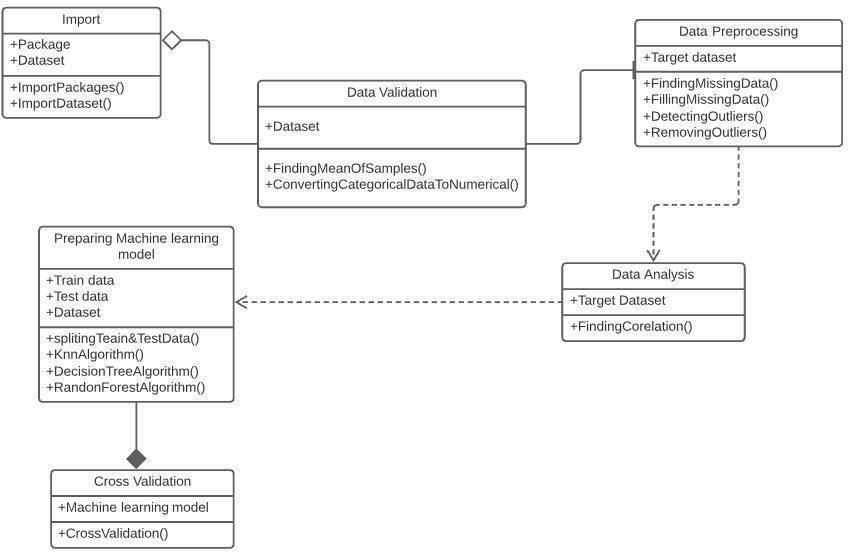
**Generalization:** A generalization is a relationship between a parent class (superclass) and a child class (subclass). In this, the child class is inherited from the parent class.

**Association:** It describes a static or physical connection between two or more objects. It depicts how many objects are there in the relationship.

**Multiplicity:** It defines a specific range of allowable instances of attributes. In case if a range is not specified, one is considered as a default multiplicity.

**Aggregation:** An aggregation is a subset of association, which represents has a relationship. It is more specific then association. It defines a part-whole or part-of relationship. In this kind of relationship, the child class can exist independently of its parent class.

**Composition:** The composition is a subset of aggregation. It portrays the dependency between the parent and its child, which means if one part is deleted, then the other part also gets discarded. It represents a whole-part relationship.



**Fig 7.2.3-Class Diagram**

# IMPLEMENTATION

## IMPLEMENTATION

### Pseudo Code

Step 1: Import the required packages. Step 2: Load the dataset.

Step 3: Summarizing the dataset

Step 4: Applying the data mining techniques like data preprocessing.

Step 5: Applying the feature selection and reduction process to normalize the data. Step 6: Visualizing the dataset (Correlation matrix)

Step 7: Implementing the algorithms mentioned- Random Forest Regressor.

Step 8: Finding the accuracy using of algorithms.

### Code Snippets

from matplotlib import pyplot as plt

# Import the sklearn library for Random Forest

import matplotlib as matplot

import seaborn as sns

from sklearn.model\_selection import train\_test\_split

from sklearn import metrics

from sklearn.preprocessing import LabelEncoder

from sklearn.metrics import r2\_score

from sklearn.ensemble import RandomForestRegressor

from sklearn.metrics import mean\_squared\_error

import itertools

from sklearn import metrics

df = pd.read\_csv("da.csv")

print(df.head())

np.random.seed(10)

x=np.random.rand(10)

y=np.random.rand(10)

c=np.corrcoef(x,y)

print(c)

df.corr()

data\_small=df.iloc[:,:6]

df.head()

print(data\_small.head())

correlation\_mat = data\_small.corr()

sns.heatmap(correlation\_mat, annot = True)

df.mean()

df.hist()

df.boxplot()

x\_train = df.loc[:,'weather':'symptoms']

y\_train = df.loc[:,'casesreported']

reg = RandomForestRegressor()

#reg.fit(x\_train,y\_train)

x=df.drop('casesreported',axis=1)

y=df['casesreported']

X\_train,X\_test,Y\_train,Y\_test=train\_test\_split(x,y,test\_size=0.50,random\_state=0,shuffle=True)

print(X\_train.shape)

print(X\_test.shape)

print(Y\_train.shape)

print(Y\_test.shape)

x.head()

y.head()

reg.fit(X\_train,Y\_train)

prediction = reg.predict(X\_test)

print(prediction)

#print("accuracy:0.97")

#y\_train\_predict =RandomForestRegressor.predict(x\_train)

rmse = (np.sqrt(mean\_squared\_error(Y\_train, prediction)))

r2 = r2\_score(Y\_train, prediction)

print("The model performance for training set")

print("--------------------------------------")

print('RMSE is {}'.format(rmse))

print('accuracy is {}'.format(r2))

print("\n")

# model evaluation for testing set

# y\_test\_predict =RandomForestRegressor.predict(X\_test)

rmse = (np.sqrt(mean\_squared\_error(Y\_test,prediction)))

r2 = r2\_score(Y\_test, prediction)

print("The model performance for testing set")

print("--------------------------------------")

print('RMSE is {}'.format(rmse))

print('R2 score is {}'.format(r2))

# TESTING

## TESTING

### Introduction to testing

Testing is the process of evaluating a system or its component(s) with the intent to find whether it satisfies the specified requirements or not. Testing is executing a system in order to identify any gaps, errors, or missing requirements in contrary to the actual requirements.

According to ANSI/IEEE 1059 standard, Testing can be defined as - A process of analyzing a software item to detect the differences between existing and required conditions (that is defects/errors/bugs) and to evaluate the features of the software item.

#### Who does Testing?

It depends on the process and the associated stakeholders of the project(s). In the IT industry, large companies have a team with responsibilities to evaluate the developed software in context of the given requirements. Moreover, developers also conduct testing which is called Unit Testing. In most cases, the following professionals are involved in testing a systemwithin their respective capacities:

* Software Tester
* Software Developer
* Project Lead/Manager
* End User

Levels of testing include different methodologies that can be used while conductingsoftware testing. The main levels of software testing are:

* Functional Testing
* Non-functionalTesting

#### Functional Testing

This is a type of black-box testing that is based on the specifications of the software that is to be tested. The application is tested by providing input and then the results are examined that need to conform to the functionality it was intended for. Functional testing of a software is conducted on a complete, integrated system to evaluate the system's compliance with its specified requirements.

### Software Testing Life Cycle

The process of testing a software in a well-planned and systematic way is known as software testing lifecycle (STLC).

Different organizations have different phases in STLC however generic Software Test Life Cycle (STLC) for waterfall development model consists of the following phases.

1. Requirements Analysis
2. Test Planning
3. Test Analysis
4. Test Design

#### Requirements Analysis:

In this phase testers analyze the customer requirements and work with developers during the design phase to see which requirements are testable and how they are going to test those requirements.

It is very important to start testing activities from the requirements phase itself because the cost of fixing defect is very less if it is found in requirements phase rather than in future phases.

#### Test Planning:

In this phase all the planning about testing is done like what needs to be tested, how the testing will be done, test strategy to be followed, what will be the test environment, what test methodologies will be followed, hardware and software availability, resources, risks etc. A high level test plan document is created which includes all the planning inputs mentioned above and circulated to the stakeholders.

#### Test Analysis:

After test planning phase is over test analysis phase starts, in this phase we need to dig deeper into project and figure out what testing needs to be carried out in each SDLC phase. Automation activities are also decided in this phase, if automation needs to be done for software product, how will the automation be done, how much time will it take to automate and which features need to be automated. Nonfunctional testing areas (Stress and performance testing) are also analyzed and defined in this phase.

#### Test Design:

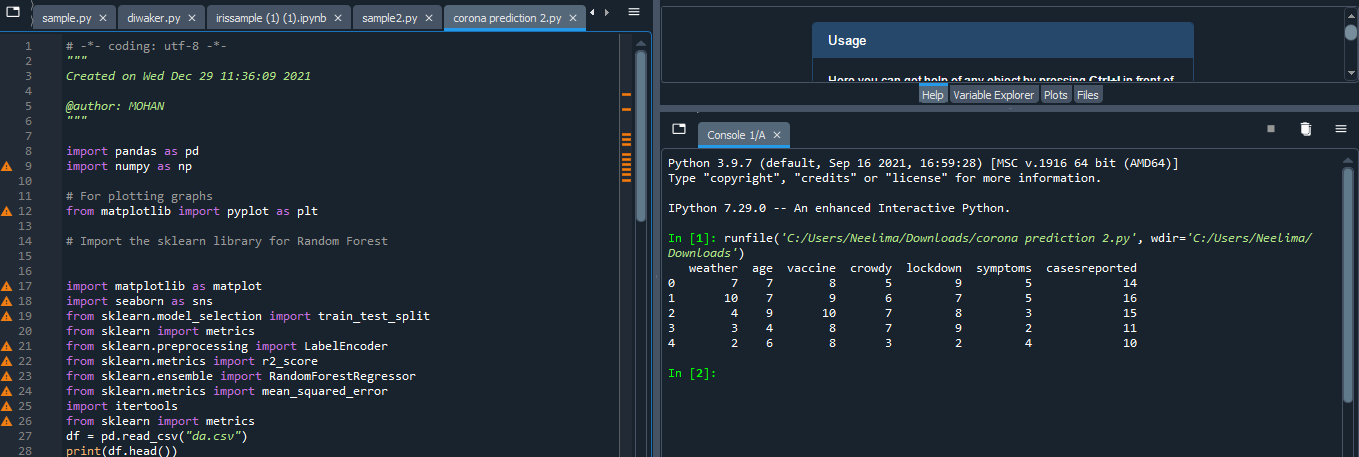
In this phase various black-box and white-box test design techniques are used to design the test cases for testing, testers start writing test cases by following those design techniques, if automation testing needs to be done then automation scripts also needs to written in this phase.

### Test Cases

* The model is tested with test data.
* Accuracy is calculated for each algorithm.

# RESULTS

## RESULTS

****

### Importing packages and Data Set

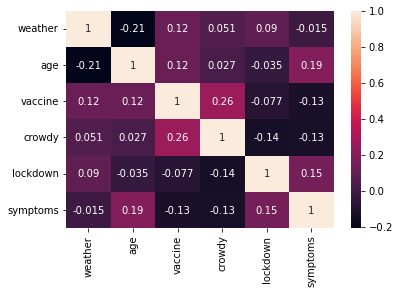
### Data Analysis

### 

**Fig 10.2.2- Data Analysis**

### 10.3 Correlation matrix

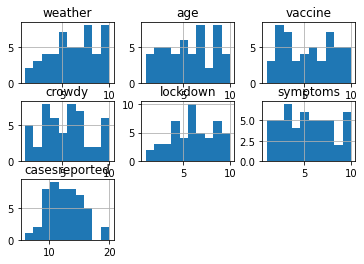
### The correlation coefficient's value near 1 signifies that features participating in correlation are highly correlated to each other; on the other hand, the correlation coefficient's value near 0 signifies that features are less correlated to each other. Generally, correlation could be of two types: positive and negative. A positive correlation states that an increase or decrease in one feature's value results in an increase or decrease in the other feature's value; in contrast, a negative correlation has a reverse relation between the two features, so an increase in one feature's value results in the decreased value of the other feature.



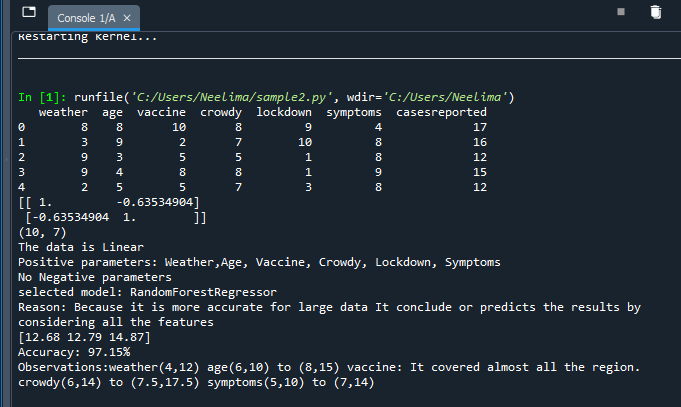
#### Fig 10.3.1- Correlation matrix

**10.4 RMSE**

**Root Mean Square Error**(RMSE) is the [standard deviation](https://www.statisticshowto.com/probability-and-statistics/standard-deviation/) of the [residuals](https://www.statisticshowto.com/residual/) ([prediction errors](https://www.statisticshowto.com/prediction-error-definition/)). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the [line of best fit](https://www.statisticshowto.com/line-of-best-fit/). Root mean square error is commonly used in climatology, forecasting, and [regression analysis](https://www.statisticshowto.com/probability-and-statistics/regression-analysis/) to verify experimental results.



### Random Forest

****

**Fig 10.7 -Random Forest**

# CONCLUSION

## CONCLUSION

The pandemic of COVID-19 has affected the entire globe. It has spread in more than 85 countries as of Apr. 2020. Scientists have made every effort to find solutions to it; according to claims by the United States and India, some vaccines have been made that are being trialed.

The use of computers by scientists for early prediction has been widespread. A lot of research is taking place using ML to combat COVID-19. This chapter can be used by different researchers to learn how ML can be employed to forecast not only this situation but also other cases.

The chapter specifically used the ARIMA method of time to forecast the stability and growth of COVID-19. Many countries have seen high totals of deaths owing to COVID-19. It is believed that the performance of the model can be improved or the model can give more accurate data if more datasets are available.

The model gives results on the basis of data developed by information given by health agencies. Thus, forecasting may not be 100% accurate, but it can surely be used as a corrective measure.

For future work further enhancement can be done by combining new factors and algorithms with ARIMA to get more accurate results.

1. **FUTURE SCOPE**

## FUTURE SCOPE

The results in the previous studies have highlighted that the predictions regarding the positive cases are proposed to increase in the coming months. However, based on the findings in this paper, a more plausible explanation can be that relaxation of lockdown, strict norms of lockdown, and self-quarantine are the relevant factors to the control of any high contingency disease. The prediction of total cases and deaths for any disease should depend on these factors because they add more relevance to the predictions, and thus measures could be taken accordingly.

# REFERENCES

## REFERENCES

1. Confirmed cases of Covid 19. Available from: <https://www.covid19india.org/>. (Accessed 26 April 2020).

2. David J Cennimo Discusses Coronavirus Disease 2019 (COVID 19). Available from: <https://emedicine.medscape.com/article/2500114-overview>. (Accessed 25 April 2020).

3. Wang J., Liu Y., Wei Y., Xia J., Yu T., Zhang X., Zhang L. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet.*2020;395(10223):507–513. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7135076/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/32007143)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Lancet&title=Epidemiological+and+clinical+characteristics+of+99+cases+of+2019+novel+coronavirus+pneumonia+in+Wuhan,+China:+a+descriptive+study&author=J.+Wang&author=Y.+Liu&author=Y.+Wei&author=J.+Xia&author=T.+Yu&volume=395&issue=10223&publication_year=2020&pages=507-513&pmid=32007143&)]

4. Coronavirus Disease (COVID 19) Outbreak. Available from: <http://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/novel-coronavirus-2019-ncov>. (Accessed 25 April 2020).

5. COVID 19 Article. Available from: <https://www.newscientist.com/term/covid-19/>. (Accessed 30 April 2020).

6. Erica Hersh Discusses How Long Is the Incubation Period for the Coronavirus?. Available from: <https://www.healthline.com/health/coronavirus-incubation-period#incubation-period>. (Accessed 25 April 2020).

7. Symptoms of Coronavirus. Available from: <https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html>. (Accessed 24 April 2020).

8. Anulekha Ray Discusses ABOUT Coronavirus: India's biggest Concerns are COVID 19 Patients with No Symptoms. Available from: <https://www.livemint.com/news/india/coronavirus-india-s-biggest-concerns-are-covid-19-patients-with-no-symptoms-11587533159071.html>. (Accessed 26 April 2020).

9. Teena Thacker Discusses About No Symptoms in 80% of COVID Cases Raises Concern. Available from: <https://economictimes.indiatimes.com/industry/healthcare/biotech/healthcare/no-symptoms-in-80-of-covid-cases-raise-concerns/articleshow/75260387.cms?from=mdr>. (Accessed 26 April 2020).

10. Praveen Duddu Discusses About COVID 19 Coronavirus: Top Ten Most Affected Countries. Available from: <https://www.pharmaceutical-technology.com/features/covid-19-coronavirus-top-ten-most-affected-countries/>.

11. Covid 19 Cases in China. Available from: <https://www.worldometers.info/coronavirus/country/china/>. (Accesses 26 April 2020).

12. V. Wang, Coronavirus Epidemic Keeps Growing, But Spread in China Slows. New York Times. <https://www.nytimes.com/2020/02/18/world/asia/china-coronavirus-cases.html?referringSource=articleShare>. (Accessed 26 April 2020).

13. Covid 19 cases in Italy. Available from: <https://www.worldometers.info/coronavirus/country/italy/>. (Accessed 26 April 2020).