**ARTIFICIAL INTELLIGENCE PROJECT REPORT**

**ON**

**COVID-19 IN US CITIES**



FACULTY NAME : DR. SHAVETA ARORA

STUDENT NAME: MUSKAAN RATRA(17CSU119)

MUSKAN NANDKANI(17CSU121)

SEMESTER 6-C

GROUP-C5

# Table of Contents:

1. ABSTRACT
2. INTRODUCTION
3. TRAINING DESCRIPTION
4. ANALYSIS
5. CONCLUSION
6. BIBLIOGRAPHY

**ABSTRACT**

Since the outbreak of the Coronaviruses (CoV) in Wuhan China, it continues to spread across countries like wildfire, defiling all the measures put in places by government and individuals to curb it. Corona Virus are a large family of viruses that cause illness ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV). Common symptoms of infection include respiratory difficulties, fever, cough, shortness of breath and breathing difficulties. In more complicated cases, infection may lead to pneumonia, severe acute respiratory syndrome, kidney failure and even death. The continuous rise in the spread of the virus has continue to pose great setback to Italy and the entire globe. The difficulties in containing the spread of the virus across localities is challenging and required drastic measures. It is in the light of this that this research proposed an urgent end-to-end data driven modeling approach that will utilize online survey questionnaire application to collect data that will be feed to a Supervised Machine learning model in order to ascertain individual's health status. Individuals exposed to the virus or noticed some symptoms of the infection and perhaps scared of reporting to local authorities can utilize this platform to join in curbing the spread of the virus. This will develop a more fast and efficient way for individuals to measure, calculate, classify, predicts their health status with regards to the activities and symptoms they developed over past two-three weeks. This information can assist healthcare officials to identify/report cases and ensure they access timely health care in their localities. This aimed to enhance citizens know their current and future health status in their respective locations before arrival or first contact with healthcare officials so as to avoid spreading the virus further among loved ones, neighbors, etc. and to make it easy for healthcare officials to know their conditions and locations. Creating a data driven model application that will bring the stakeholders and government closer to its' citizens at such a time as this with regards to curbing the spread of the COVID19 virus is paramount, in order to assist them in designing the best policies, timely response and quarantine measures across Italy and the globe.

**Training Description:**

We did programming in Python on **IDLE** (short for **integrated development environment** or **integrated development and learning environment**[]](https://en.wikipedia.org/wiki/IDLE#cite_note-3)) which is an [integrated development environment](https://en.wikipedia.org/wiki/Integrated_development_environment) for [Python](https://en.wikipedia.org/wiki/Python_(programming_language)), which has been bundled with the default implementation of the language. It is packaged as an optional part of the Python packaging with many [Linux distributions](https://en.wikipedia.org/wiki/Linux_distributions). It is completely written in Python and the [Tkinter](https://en.wikipedia.org/wiki/Tkinter) GUI toolkit ([wrapper](https://en.wikipedia.org/wiki/Wrapper_function) functions for [Tcl](https://en.wikipedia.org/wiki/Tcl)/[Tk](https://en.wikipedia.org/wiki/Tk_(framework))).

IDLE is intended to be a simple [IDE](https://en.wikipedia.org/wiki/Integrated_development_environment) and suitable for beginners, especially in an educational environment. To that end, it is cross-platform, and avoids feature clutter.

**Why Python ?**

The language's core philosophy is summarized in the document The Zen of Python (PEP 20), which includes aphorisms such as…

Beautiful is better than ugly

Simple is better than complex

Complex is better than complicated

Readability counts

Explicit is better than implicit

**Characteristics of Python:**

* Interpreted Language: Python is processed at runtime by Python Interpreter
* Easy to read: Python source-code is clearly defined and visible to the eyes.
* Portable: Python codes can be run on a wide variety of hardware platforms having the same interface.
* Extendable: Users can add low level-modules to Python interpreter.
* Scalable: Python provides an improved structure for supporting large programs than shell-scripts.
* Object-Oriented Language: It supports object-oriented features and techniques of programming.
* Interactive Programming Language: Users can interact with the python interpreter directly for writing programs.
* Easy language: Python is easy to learn language especially for beginners.
* Straight forward Syntax: The formation of python syntax is simple and straightforward which also makes it popular.

**Data Structures in Python**

* LISTS- Ordered collection of data.
* Supports similar slicing and indexing functionalities as in the case of Strings.
* They are mutable.
* Advantage of a list over a conventional array
* Lists have no size or type constraints (no setting restrictions beforehand).
  + They can contain different object types.
  + We can delete elements from a list by using Del list\_name[index\_val]
* Example-
  + my\_list = ['one', 'two','three',4,5] • len(my\_list) would output 5.

**Dictionaries :**

* Lists are sequences but the dictionaries are mappings.
* They are mappings between a unique key and a value pair
* These mappings may not retain order.
* Constructing a dictionary.
* Accessing object from a dictionary
* Nesting Dictionaries.
* Basic Dictionary Methods
* Basic Syntax o d={} empty dictionary will be generated and assign keys and values to it, like d[‘animal’] = ‘Dog’ o d = {'K1':'V1', 'K2’:’V2'} o d['K1'] outputs 'V1‘

**Sets :**

* A set contains unique and unordered elements and we can construct them by using a set() function.
* Convert a list into Set-

l=[1,2,3,4,1,1,2,3,6,7]

k = set(l)

k becomes {1,2,3,4,6,7}

* Basic Syntax

x=set()

x.add(1)

x = {1}

x.add(1) This would make no change in x now

**File Handling in Python**

Python too supports file handling and allows users to handle files i.e., to read and write files, along with many other file handling options, to operate on files.

The concept of file handling has stretched over various other languages, but the implementation is either complicated or lengthy, but alike other concepts of Python, this

concept here is also easy and short. Python treats file differently as text or binary and this is important. Each line of code includes a sequence of characters and they form text file. Each line of a file is terminated with a special character, called the EOL or End of Line characters like comma {,} or newline character. It ends the current line and tells the interpreter a new one has begun. Let’s start with Reading and Writing files. We use open () function in Python to open a file in read or write mode. As explained above, open ( ) will return a file object. To return a file object we use open () function along with two arguments, that accepts file name and the mode, whether to read or write. So, the syntax being: open(filename, mode).

There are three kinds of mode, that Python provides and how files can be opened:

• “ r “, for reading.

• “ w “, for writing.

• “ a “, for appending.

• “ r+ “, for both reading and writing

**Python Libraries that would be need to achieve the task:  
1. Numpy  
2. Pandas  
3. Sci-kit Learn  
4. Matplotlib**

**First we have to check which version of pip is installed in our system**

1. $ pip --version
2. pip 18.1 from /home/zhaosong/anaconda3/lib/python3.7/site-packages/**pip** (python 3.7)

**Following commands are used to install the above libraries**

1. pip install numpy
2. pip install pandas
3. pip install scipy
4. pip install matplotlib

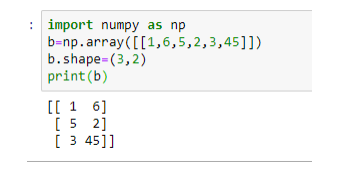
**Use of Numpy**

NumPy is a Python package. It stands for 'Numerical Python'. It is a library consisting of multidimensional array objects and a collection of routines for processing of array. Numeric, the ancestor of NumPy, was developed by Jim Hugunin. Another package Numarray was also developed, having some additional functionalities. In 2005, Travis Oliphant created NumPy package by incorporating the features of Numarray into Numeric package. There are many contributors to this open source project. Operations using NumPy Using NumPy, a developer can perform the following operations − Mathematical and logical operations on arrays.

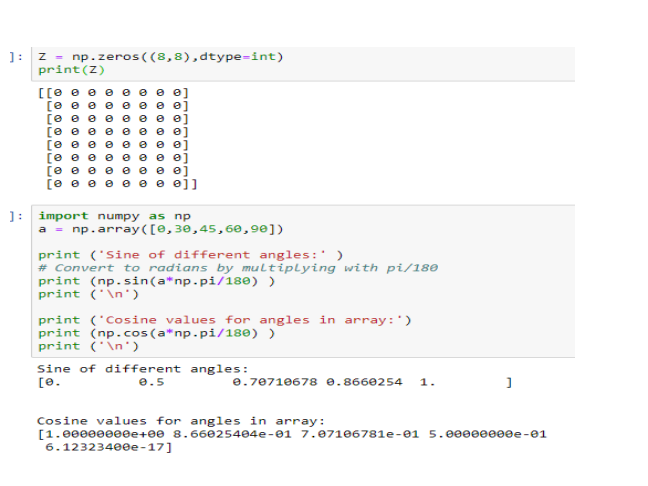
• Fourier transforms and routines for shape manipulation.

• Operations related to linear algebra. NumPy has in-built functions for linear

• algebra and random number generation. Simple program to create a matrixFirst of all we import numpy package then using this we take input in numpy function as a list then we create a matrix



There is many more function can be perform by using this like that take sin value of the given value ,print a zero matrix etc. we also take any image in the form of array.



**Use of Matplotlib-** Matplotlib is a library for making 2D plots of arrays in Python. Although it has itsorigins in emulating the MATLAB graphics commands, it is independent ofMATLAB, and can be used in a Pythonic, object oriented way. Although Matplotlibis written primarily in pure Python, it makes heavy use of NumPy and other

extension code to provide good performance even for large arrays.

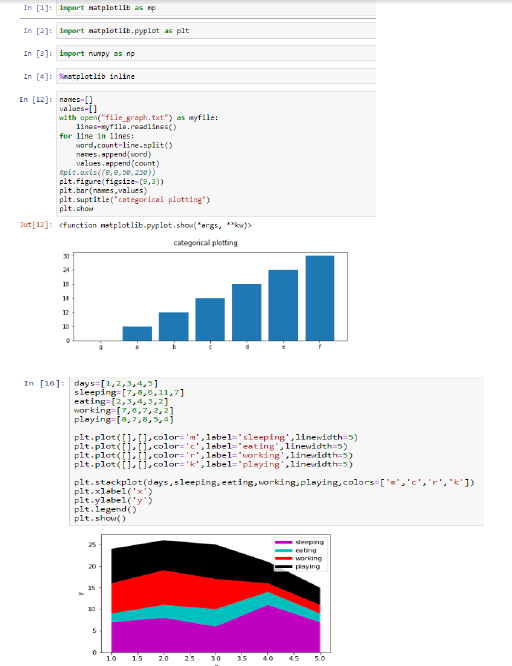
Matplotlib is designed with the philosophy that you should be able to create

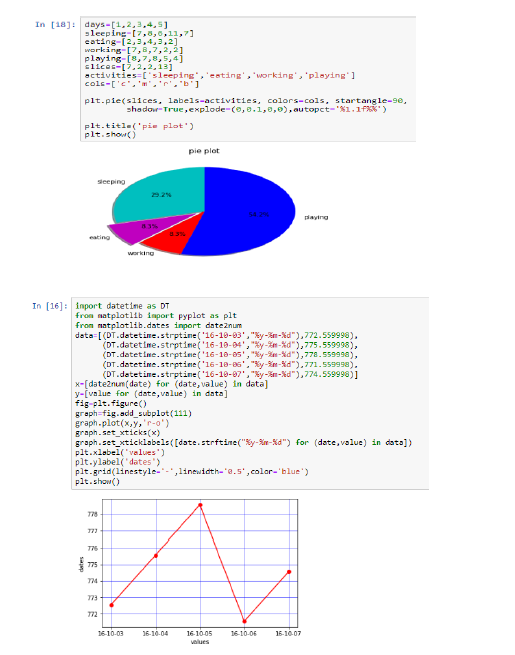
simple plots with just a few commands, or just one! If you want to see a

histogram of your data, you shouldn’t need to instantiate objects, call methods,

set properties, and so on; it should just work.

These are the some example of matplotlib..





**Use of Pandas –**

Pandas is an open-source, BSD-licensed Python library providing high performance,

easy-to-use data structures and data analysis tools for the Python

programming language. Python with Pandas is used in a wide range of fields

including academic and commercial domains including finance, economics,

Statistics, analytics, etc.

Pandas is an open-source Python Library providing high-performance data

manipulation and analysis tool using its powerful data structures. The name

Pandas is derived from the word Panel Data – an Econometrics from

Multidimensional data.

**Key Features of Pandas-**

• Fast and efficient Data Frame object with default and customized indexing**.**

• Tools for loading data into in-memory data objects from different file

formats.

• Data alignment and integrated handling of missing data.

• Reshaping and pivoting of date sets.

• Label-based slicing, indexing and subsetting of large data sets.

• Columns from a data structure can be deleted or inserted.

• Group by data for aggregation and transformations.

Pandas deals with the following three data structures −

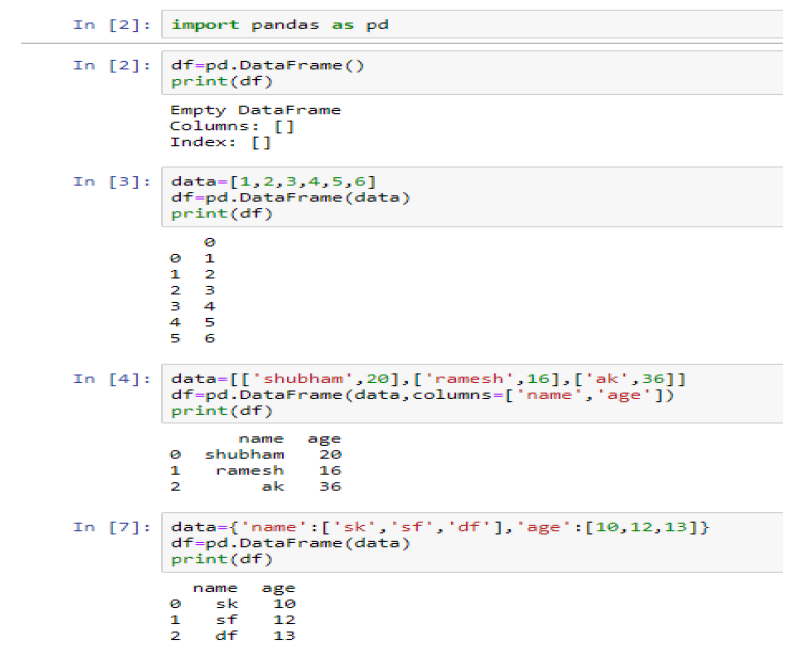
Series

DataFrame

Panel

These data structures are built on top of Numpy array, which means they are

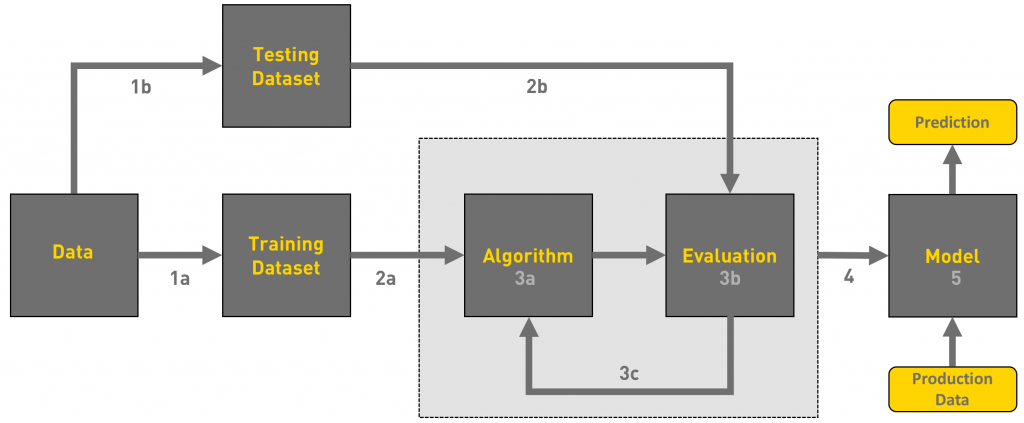
fast.

****

.

**What is the machine learning Model?**

The machine learning model is nothing but a piece of code; an engineer or data scientist makes it smart through training with data. So, if we give garbage to the model, we will get garbage in return, i.e. the trained model will provide false or wrong predictions.

****

**We can define the machine learning workflow stages as:**

**1.** Gathering data

**2.** Data pre-processing

**3.** Researching the model that will be best for the type of data

**4.** Training and testing the model

**5.** Evaluation

**1. Gathering Data**

The process of gathering data depends on the type of project we desire to make, if we want to make an ML project that uses real-time data, then we can build an IoT system that using different sensors data. The data set can be collected from various sources such as a file, database, sensor and many other such sources but the collected data cannot be used directly for performing the analysis process as there might be a lot of missing data, extremely large values, unorganized text data or noisy data. Therefore, to solve this problem Data Preparation is done.

We can also use some free data sets which are present on the

internet. [**Kaggle**](http://www.kaggle.com/)and [**UCI Machine learning Repository**](https://archive.ics.uci.edu/ml/datasets.html) are the repositories that are used the most for making Machine learning models. Kaggle is one of the most visited websites that is used for practicing machine learning algorithms, they also host competitions in which people can participate and get to test their knowledge of machine learning.

**2. Data pre-processing**

Data pre-processing is one of the most important steps in machine learning. It is the most important step that helps in building machine learning models more accurately. In machine learning, there is an 80/20 rule. Every data scientist should spend 80% time for data pre-processing and 20% time to actually perform the analysis.

Data pre-processing is a process of cleaning the raw data i.e. the data is collected in the real world and is converted to a clean data set. In other words, whenever the data is gathered from different sources it is collected in a raw format and this data isn’t feasible for the analysis.  
Therefore, certain steps are executed to convert the data into a small clean data set, this part of the process is called as data pre-processing.

As we know that data pre-processing is a process of cleaning the raw data into clean data, so that can be used to train the model. So, we definitely need data pre-processing to achieve good results from the applied model in machine learning and deep learning projects.

Most of the real-world data is messy, some of these types of data are:

1. **Missing data:** Missing data can be found when it is not continuously created or due to technical issues in the application (IOT system).

2. **Noisy data:** This type of data is also called outliners, this can occur due to human errors (human manually gathering the data) or some technical problem of the device at the time of collection of data.

3. **Inconsistent data:** This type of data might be collected due to human errors (mistakes with the name or values) or duplication of data.

Three Types of Data

1. Numeric e.g. income, age

2. Categorical e.g. gender, nationality

3. Ordinal e.g. low/medium/high

## How can data pre-processing be performed?

These are some of the basic pre — processing techniques that can be used to convert raw data.

1. **Conversion of data:** As we know that Machine Learning models can only handle numeric features, hence categorical and ordinal data must be somehow converted into numeric features.

2. **Ignoring the missing values:** Whenever we encounter missing data in the data set then we can remove the row or column of data depending on our need. This method is known to be efficient but it shouldn’t be performed if there are a lot of missing values in the dataset.

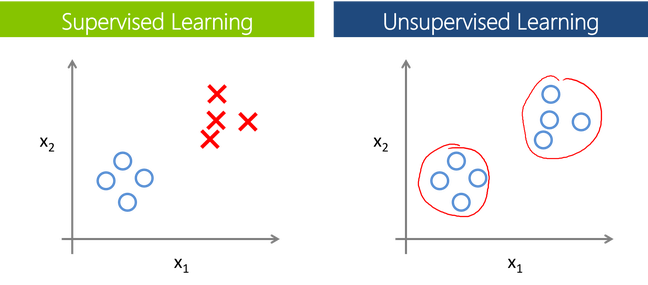
3. **Filling the missing values:** Whenever we encounter missing data in the data set then we can fill the missing data manually, most commonly the mean, median or highest frequency value is used.

4.**Machine learning:** If we have some missing data then we can predict what data shall be present at the empty position by using the existing data

5. **Outliers detection:** There are some error data that might be present in our data set that deviates drastically from other observations in a data set. [Example: human weight = 800 Kg; due to mistyping of extra 0]

**3. Researching the model that will be best for the type of data**

Our main goal is to train the best performing model possible, using the pre-processed data.

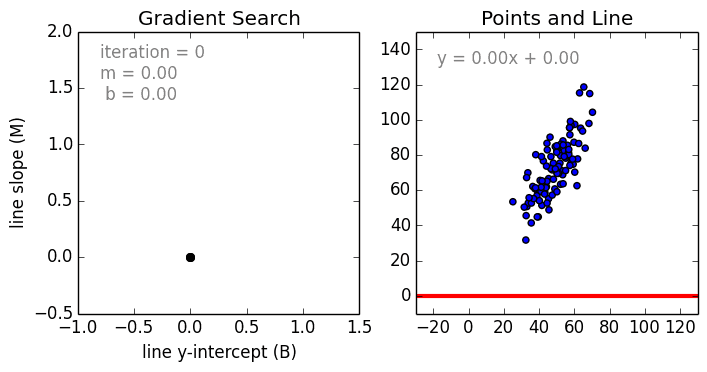
****

The supervised learning is categorized into 2 other categories which are “**Classification**” and “**Regression**”.

**Classification**: Classifcation problem is when the target variable is **categorical**(i.e. the output could be classified into classes — it belongs to either Class A or B or something else).A classification problem is when the output variable is a category, such as “red” or “blue” , “disease” or “no disease” or “spam” or “not spam”.As shown in the above representation, we have 2 classes which are plotted on the graph i.e. red and blue which can be represented as ‘setosa flower’ and ‘versicolor flower’, we can image the X-axis as

the ‘Sepal Width’ and the Y-axis as the ‘Sepal Length’, so we try to create the [best fit line](https://mathbits.com/MathBits/TISection/Statistics1/LineFit.htm) that separates both classes of flowers.These some most used classification algorithms.

* **K-Nearest Neighbor**
* **Naive Bayes**
* **Decision Trees/Random Forest**
* **Support Vector Machine**
* **Logistic Regression**

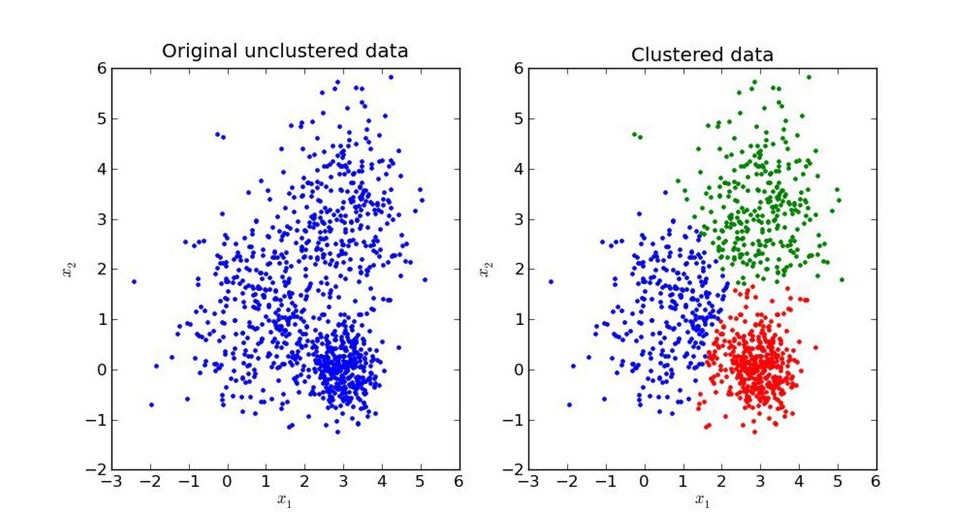
**Regression**: While a **Regression**problem is when the target variable is **continuous**(i.e. the output is numeric).

As shown in the above representation, we can imagine that the graph’s X-axis is the ‘Test scores’ and the Y-axis represents ‘IQ’. So we try to create the [best fit line](https://mathbits.com/MathBits/TISection/Statistics1/LineFit.htm) in the given graph so that we can use that line to predict any approximate IQ that isn’t present in the given data.

These some most used regression algorithms.

* **Linear Regression**
* **Logistic Regression**
* **Decision Tress/Random Forest**
* **Gaussian Progresses Regression**
* **Ensemble Methods**

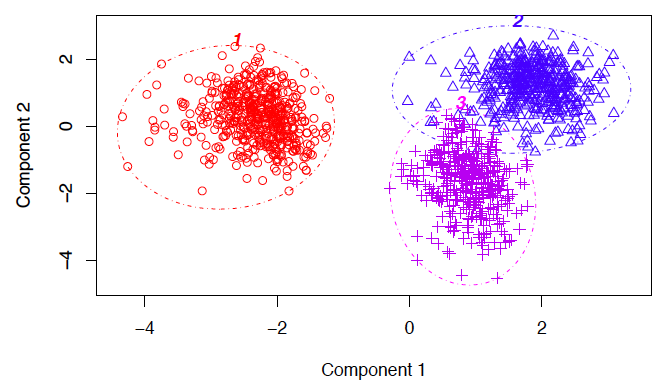
## Unsupervised Learning:



In unsupervised learning, an AI system is presented with unlabeled, un-categorized data and the system’s algorithms act on the data without prior training. The output is dependent upon the coded algorithms. Subjecting a system to unsupervised learning is one way of testing AI. The unsupervised learning is categorized into 2 other categories which are “**Clustering**” and “**Association**”

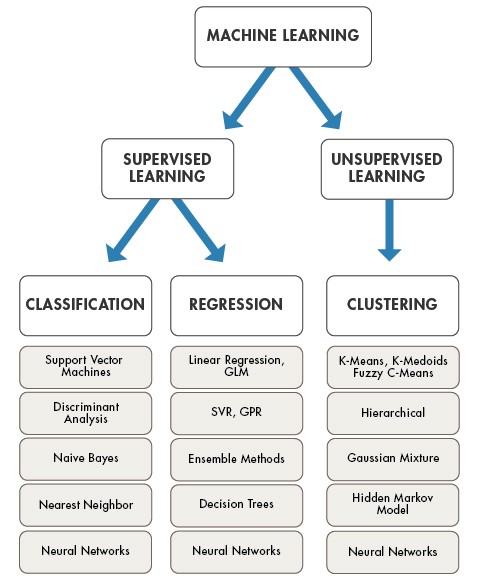
## Clustering:

A set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.



Methods used for clustering are:

* **Gaussian mixtures**
* **K-Means Clustering**
* **Boosting**
* **Hierarchical Clustering**
* **K-Means Clustering**
* **Spectral Clustering**



# 4. Training and testing the model on data

For training a model we initially split the model into 3 three sections which are ‘**Training data**’ ,‘**Validation data**’ and ‘**Testing data**’.

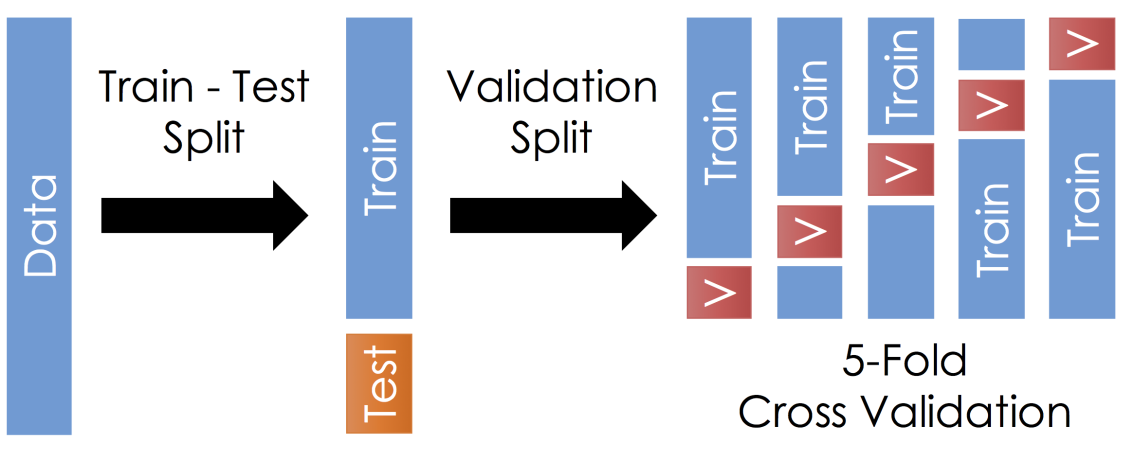
You train the classifier using ‘**training data set**’, tune the parameters using ‘**validation set**’ and then test the performance of your classifier on unseen ‘**test data set**’. An important point to note is that during training the classifier only the training and/or validation set is available. The test data set must not be used during training the classifier. The test set will only be available during testing the classifier.



**Training set:** The training set is the material through which the computer learns how to process information. Machine learning uses algorithms to perform the training part. A set of data used for learning , that is to fit the parameters of the classifier.

**Validation set:** Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. A set of unseen data is used from the training data to tune the parameters of a classifier.

**Test set:** A set of unseen data used only to assess the performance of a fully-specified classifier.



Once the data is divided into the 3 given segments we can start the training process.

In a data set, a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. Data points in the training set are excluded from the test (validation) set. Usually, a data set is divided into a training set, a validation set (some people use ‘test set’ instead) in each iteration, or divided into a training set, a validation set and a test set in each iteration.

The model uses any one of the models that we had chosen in step 3/ point 3. Once the model is trained we can use the same trained model to predict using the testing data i.e. the unseen data. Once this is done we can develop a confusion matrix, this tells us how well our model is trained. A confusion matrix has 4 parameters, which are ‘**True positives’**,**‘True Negatives’**, **‘False Positives’**and ‘**False Negative’**. We prefer that we get more values in the True negatives and true positives to get a more accurate model. The size of the Confusion matrix completely depends upon the number of classes.

Once the data is divided into the 3 given segments we can start the training process .In a data set, a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. Data points in the training set are excluded from the test (validation) set. Usually, a data set is divided into a training set, a validation set (some people use ‘test set’ instead) in each iteration, or divided into a training set, a validation set and a test set in each iteration.

The model uses any one of the models that we had chosen in step 3/ point 3. Once the model is trained we can use the same trained model to predict using the testing data i.e. the unseen data. Once this is done we can develop a confusion matrix, this tells us how well our model is trained. A confusion matrix has 4 parameters, which are ‘**True positives’**,**‘True Negatives’**, **‘False Positives’**and ‘**False Negative’**. We prefer that we get more values in the True negatives and true positives to get a more accurate model. The size of the Confusion matrix completely depends upon the number of classes.

* **True positives:** These are cases in which we predicted TRUE and our predicted output is correct.
* **True negatives:** We predicted FALSE and our predicted output is correct.
* **False positives:** We predicted TRUE, but the actual predicted output is FALSE.
* **False negatives:** We predicted FALSE, but the actual predicted output is TRUE.

We can also find out the accuracy of the model using the confusion matrix.

*Accuracy = (True Positives +True Negatives) / (Total number of classes)*

I .e. for the above example:

Accuracy = (100 + 50) / 165 = 0.9090 (90.9% accuracy)

**5. Evaluation**

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future.

To improve the model we might tune the hyper-parameters of the model and try to improve the accuracy and also looking at the confusion matrix to try to increase the number of true positives and true negatives.

**Project Procedure**

**Gathering Data:**

US.csv

Covid-19.csv

**Data Pre Processing:**

As the data present in testing.csv and training.csv have values in the form of 0’s and 1’s

And also no missing value is formed so data pre processing is done

**Choosing the Correct Model for Analysis:**

* 1. As we want to predict the future of Us against covid-19,classification algorithm is the best suited for this type
  2. Classification Algorithm has many algorithms providing various levels of accuracy
  3. The best models namely Logistic Regression and Linear Regression will be used here

**BRIEF INTRODUCTION**

In February, the Centers for Disease Control and Prevention warned that the United States should brace for a domestic coronavirus outbreak. Throughout the month of March, the number of confirmed cases of COVID-19 in the U.S. rose dramatically, and the U.S. now has the highest confirmed case count in the world. As of Friday morning, at least 463,619 people across every state, Washington, D.C., and four territories have tested positive for the disease.Globally, more than 1.5 million confirmed cases of COVID-19 have been reported in 177 countries, with at least 96,000 deaths so far. In March, the World Health Organization officially declared COVID-19 a pandemic, and President Trump declared a national state of emergency. As of Friday morning, more than 16,000 patients with the virus have died in the U.S.

How many new coronavirus cases have been confirmed in the U.S.?

As of April 10, there are at least 463,619 confirmed cases of COVID-19 in the U.S., and at least 16,695 patients with the virus have died. Thousands of new cases and well over 1,000 deaths are being reported each day. Yet the New York Times reports that there’s still wide variation in the rates of testing among states, and that the total number of infected individuals, as well as the true death toll from the virus, is likely much higher than the official count.

New York has had by far the largest outbreak in the country, with at least 159,937 confirmed cases and 7,067 deaths as of Friday morning. New Jersey is also among the states that have been hardest hit by the virus, along with California and Michigan. Washington State had some of the country’s earliest confirmed cases, with 37 deaths from the virus linked to an outbreak of COVID-19 at a long-term-care facility, the Life Care Center, in Kirkland, Washington. Louisiana recently experienced a sudden spike in confirmed COVID-19 cases, which experts suspect could be linked to Mardi Gras celebrations, which concluded on February 25. Though the vast majority of confirmed cases and deaths from COVID-19 have been in cities and suburbs, the New York Times reported this week that coronavirus has officially reached more than two-thirds of the country’s rural counties, where cases are growing fast.

While death tolls continue to rise sharply, the number of hospitalizations appears to be plateauing in places like New York City, suggesting that social distancing may be working. In late March, Dr. Anthony Fauci, the nation’s leading infectious-disease expert, warned that even with aggressive measures to slow the spread of the virus, between 100,000 and 200,000 people in the U.S. would die; earlier this week, he said the total toll currently “looks more like 60,000.” However, he cautioned against “claiming victory” prematurely, and stressed the importance of continuing to adhere to social-distancing mandates.

What measures are being taken to limit the spread of coronavirus?

As the number of confirmed cases of COVID-19 has continued to rise, a growing number of states have announced drastic measures to slow the spread of the virus. As of April 7, officials in 42 states have urged residents to stay at home, including Alabama, Alaska, Arizona, California, Colorado, Connecticut, Delaware, Florida, Georgia, Hawaii, Idaho, Illinois, Indiana, Kansas, Kentucky, Louisiana, Maine, Maryland, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, Montana, Nevada, New Hampshire, New Jersey, New Mexico, New York, North Carolina, Ohio, Oregon, Pennsylvania, Rhode Island, South Carolina, Tennessee, Texas, Vermont, Virginia, Washington, West Virginia, and Wisconsin, as well as Washington, D.C., and Puerto Rico. Fauci advocated last week for every state to issue a stay-at-home order.

These “stay-at-home” orders require residents to stay in their homes except for essential activities, which include buying food, seeking medical treatment, and exercising outdoors, provided they stay six feet away from anyone not part of their household. Medical professionals, caregivers, public-safety officials, sanitation workers, and other essential workers, such as those who work in grocery stores and pharmacies, are exempt.

The widespread lockdowns have already had serious economic consequences: the Labor Department reported on Wednesday that more than 16 million Americans have lost their jobs in the past three weeks. Though President Trump had previously questioned whether the economic toll of stay-at-home orders was worth it, on March 29, he announced that Americans must continue to avoid nonessential travel, going to work, eating at bars and restaurants, and gathering in groups of more than 10 for at least another month, and possibly until June.

As of this week, 43 states have ordered schools closed, and seven states have recommended schools close. In New York City, the largest school district in the country, public schools will remain closed until at least April 20, though Mayor Bill de Blasio said there was a strong chance they would not reopen before the end of the school year. The mayor has said that some schools will reopen as “enrichment centers” to provide services to vulnerable children, including homeless students and those with special needs, and that the city would open centers to provide child care for the children of health-care and emergency workers.

On Monday, Governor Andrew Cuomo expressed cautious optimism that the virus appears to be nearing its peak in New York, after the state reported lower death tolls on Sunday and Monday for the first time since the outbreak began. However, a record number of people in the state died from the virus in the following three days, with Cuomo reporting another 799 deaths on Wednesday. Still, Cuomo pointed to the decreasing rate of hospitalizations as a sign that the spread of the virus may be plateauing, though he has warned that the situation remains dire and that residents should stay vigilant about following social-distancing rules.

Cuomo said on Monday that New York is no longer currently in need of ventilators after receiving additional medical devices from California, Oregon, and elsewhere. Previously, both Cuomo and de Blasio have criticized the federal government for not taking sufficient action to provide needed medical equipment and supplies, though on Monday de Blasio thanked the Trump administration for sending the city 600,000 N95 masks. Cuomo also said this week that President Trump had approved his request to use the USNS Comfort, a naval hospital ship docked in Manhattan, to treat coronavirus patients.

What is the federal government doing to fight coronavirus?

In February, President Donald Trump put Vice-President Mike Pence — a man with a frankly dismal track record in public health — in charge of the coronavirus response, assuring the public that the White House is “very, very ready for this.”

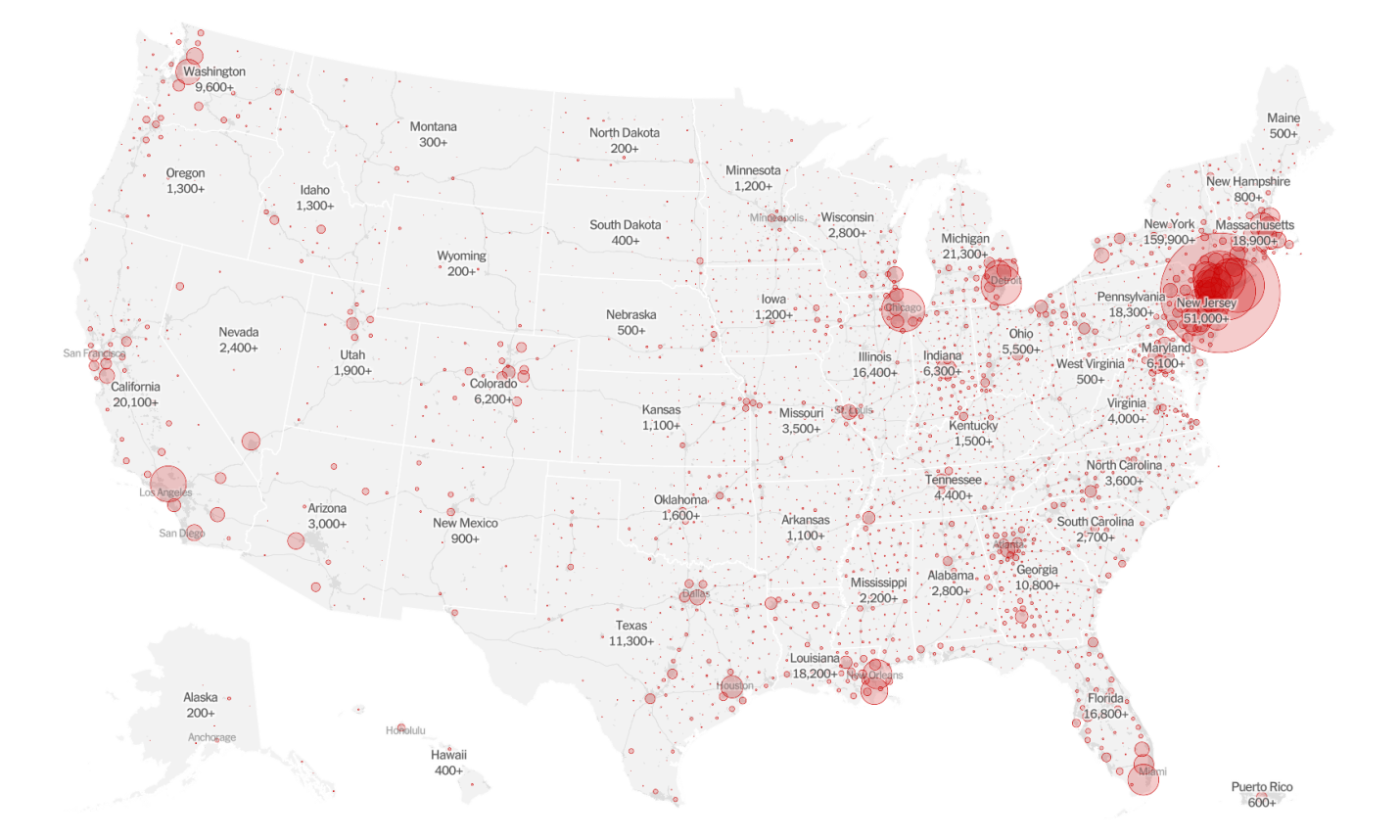
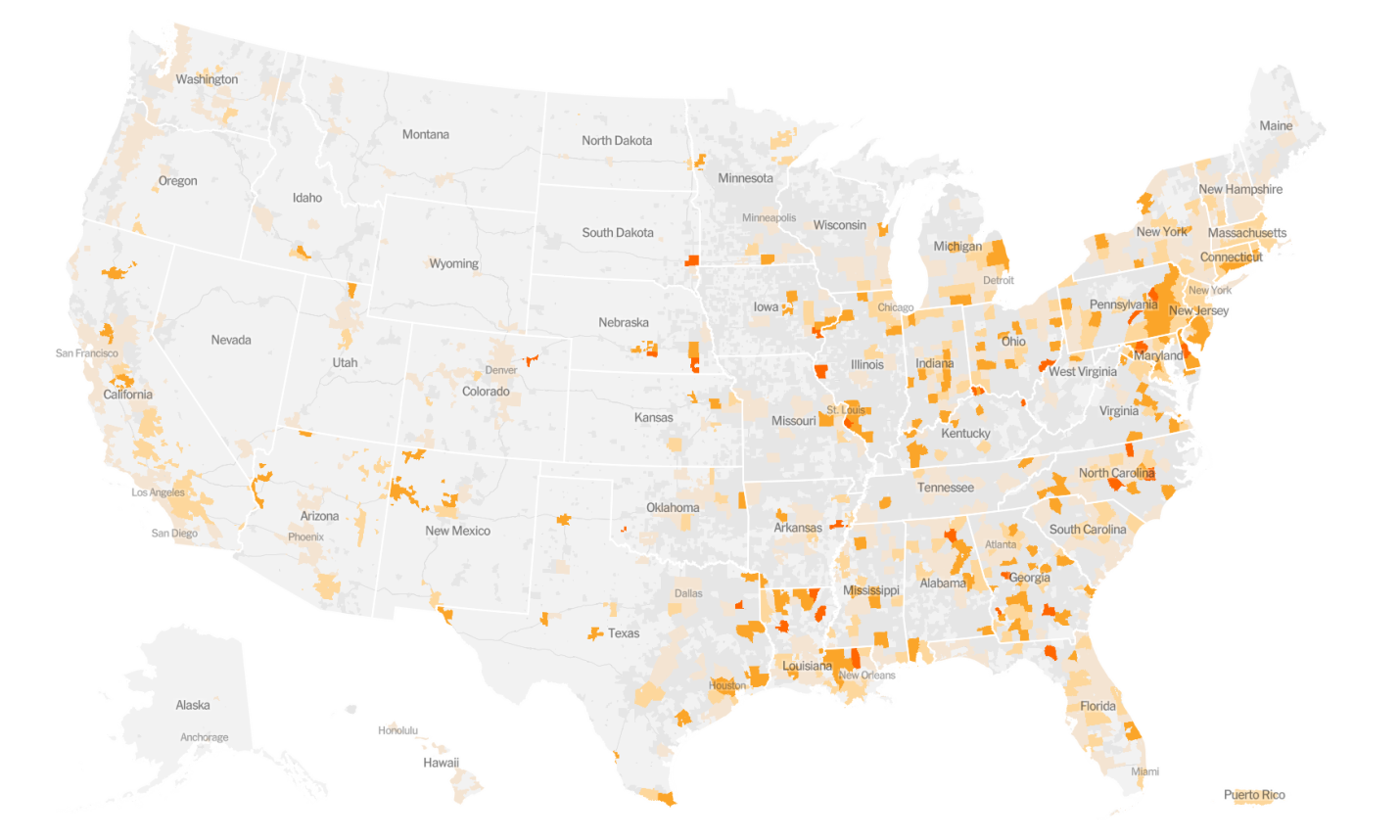
In March, President Trump declared a national emergency over the coronavirus pandemic, effectively freeing up to $50 billion in federal funds to help states and territories fight the spread of the virus, which he said would include expanding access to testing.

Still, there have been many issues with the availability of the coronavirus test. Some people say they’ve been denied tests, and even though Pence announced on March 3 that anyone in the country can be tested for coronavirus, subject to doctor’s orders, the rate of testing still varies widely from state to state.

On March 19, Secretary of State Mike Pompeo announced that the U.S. would close its borders with Canada and Mexico, barring entry to all nonessential travelers. The announcement was the latest in an increasing list of travel restrictions. President Trump has also barred entry of all foreign nationals who have been in high-risk countries, including China, Iran, and much of Europe, within the last 14 days. The CDC has advised against all nonessential travel throughout most of Europe, South Korea, China, and Iran, and has advised older and at-risk Americans to avoid travel to any country. Last week, the CDC issued a travel advisory for residents of New York, New Jersey, and Connecticut, asking them to “refrain from non-essential domestic travel” for 14 days.

On March 27, President Trump signed a $2 trillion stimulus plan, the largest in modern American history. The plan will send direct payments of around $1,200 to millions of Americans who earn less than $99,000, along with an additional $500 per child. The plan will also substantially expand unemployment benefits, including extending eligibility to freelance and gig workers, and provide aid to businesses and companies in distress. Treasury Secretary Steven Mnuchin has also extended the tax-filing deadline to July 15.

President Trump has also signed a relief package to establish paid emergency leave for some American workers, expand food assistance, medical aid, and unemployment benefits, and offer free coronavirus testing. The U.S. House and Senate reached a deal last month to provide $8.3 billion in emergency funding to stop the spread of the coronavirus. According to the Washington Post, the money will go toward the development of a vaccine, public-health funding, medical supplies, and research of coronavirus in other countries. Meanwhile, the World Health Organization has said it will take around 18 months to develop a vaccine for COVID-19.



**PURPOSE**

IN THE PAST few days, New York City’s hospitals have become unrecognizable. Thousands of patients sick with the [novel coronavirus](https://www.wired.com/story/chinas-swift-id-of-a-new-virus-is-a-win-for-public-health/) have swarmed into emergency rooms and intensive care units. From 3,000 miles away in Seattle, as Lisa Brandenburg watched the scenes unfold—isolation wards cobbled together in lobbies, nurses caring for [Covid-19 patients](https://www.wired.com/story/coronavirus-has-a-name-the-deadly-disease-is-covid-19/) in makeshift trash bag gowns, refrigerated mobile morgues idling on the street outside—she couldn’t stop herself from thinking: “That could be us.”

It could be, if the models are wrong.

Until this past week, Seattle had been the center of the [Covid-19 pandemic](https://www.wired.com/tag/coronavirus/) in the United States. It’s where US health officials confirmed the nation’s first case, back in January, and its first death a month later. As president of the University of Washington Medicine Hospitals and Clinics, Brandenburg oversees the region’s largest health network, which treats more than half a million patients every year. In early March, she and many public health authorities were shaken by an urgent report produced by computational biologists at the Fred Hutchinson Cancer Research Center. Their [analysis of genetic data](https://bedford.io/blog/ncov-cryptic-transmission/) indicated the virus had been silently circulating in the Seattle area for weeks and had already infected at least 500 to 600 people. The city was a ticking time bomb.

The mayor of Seattle declared a civil emergency. Superintendents started closing schools. King and Snohomish counties banned gatherings of more than 250 people. The Space Needle went dark. Seattleites wondered if they should be doing more, and they [petitioned](https://www.change.org/p/jay-inslee-mandatory-shelter-in-place-in-washington-state-to-fight-coronavirus?recruiter=1057582064&utm_source=share_petition&utm_medium=copylink&utm_campaign=share_petition&utm_term=share_petition) the governor to issue a statewide [shelter-at-home order](https://www.wired.com/story/whats-shelter-place-order-whos-affected/?itm_content=header-recirc). But Brandenburg was left with a much grimmer set of questions: How many people are going to get hospitalized? How many of them will require critical care? When will they start showing up? Will we have [enough ventilators](https://www.wired.com/story/ventilator-makers-race-to-prevent-a-possible-shortage/) when they do?

There’s no way to know those answers for sure. But hospital administrators like Brandenburg have to hazard an educated guess. That’s the only way they can try to buy enough ventilators and hire enough ICU nurses and clear out enough hospital beds to be ready for a wave of hacking, gasping, suffocating Covid-19 patients.

That’s where Chris Murray and his computer simulations come in.

Murray is the director of the Institute for Health Metrics and Evaluation at the University of Washington. With about 500 statisticians, computer scientists, and epidemiologists on staff, IHME is a data-crunching powerhouse. Every year it releases the Global Burden of Disease study—an alarmingly comprehensive report that quantifies [the incidence and impact](https://www.wired.com/2013/11/infoporn-causes-of-death/) of every conceivable illness and injury in each of the world’s 195 countries and territories.

In February, Murray and a few dozen IHME employees turned their attention full-time to forecasting how Covid-19 will hit the US. Specifically, they were trying to help hospitals—starting with the UW Medicine system—prepare for the coming crisis. Brandenburg says the collaboration could turn out to be, quite literally, life-saving. “It’s one thing to know you may be getting a surge of patients,” she says. “If you can make that more tangible—here’s what it’s actually going to look like—then we’re in a much better place in terms of being able to plan for the worst.”

But it’s a big if. During a pandemic, real data is hard to find. Chinese researchers have only published some of their findings on the spread of Covid-19 in Hubei. The ongoing [catastrophe of testing](https://www.wired.com/story/everything-you-need-to-know-about-coronavirus-testing/) for the virus in the United States means no researcher has even a reliable denominator, an overall number of infections that would be a reasonable starting point for untangling how rapidly the disease spreads. Since the [2009 outbreak of H1N1 influenza](http://dx.doi.org/10.1111/1469-0691.12284), researchers worldwide have increasingly relied on mathematical models, computer simulations informed by what little data they can find, and some reasoned inferences. Federal agencies like the Centers for Disease Control and Prevention and the National Institutes of Health have [modeling teams](http://www.centerforhealthsecurity.org/our-work/publications/2020/modernizing-and-expanding-outbreak-science-to-support-better-decision-making-during-public-health-crises), as do many universities.

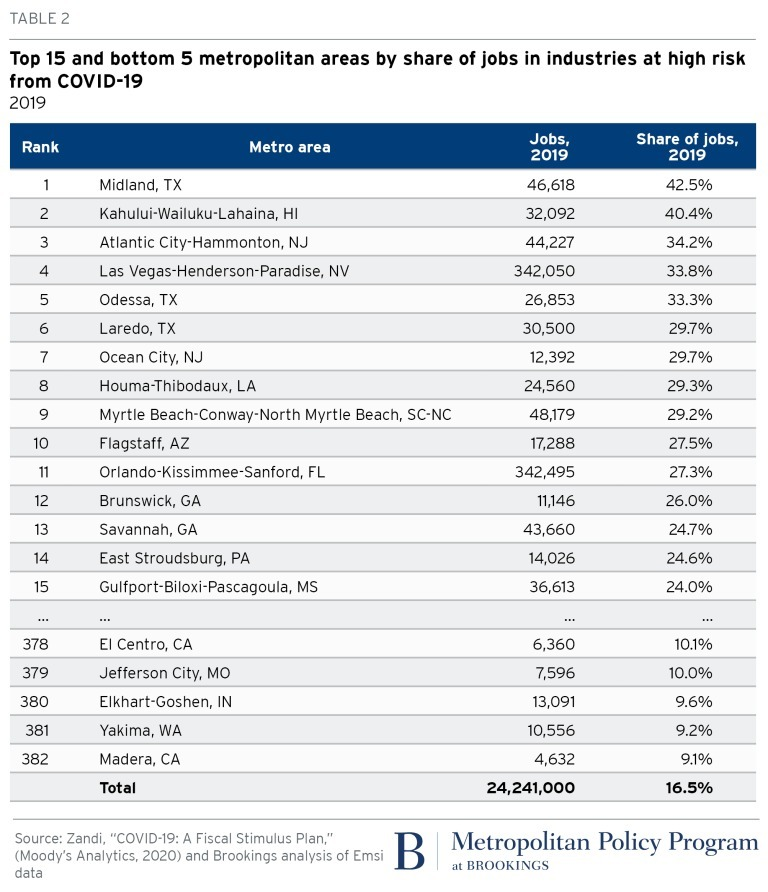
# Why it’s so hard to see into the future of Covid-19

The most difficult thing for an epidemiological model to predict: human behavior.

* [Share this on Facebook (opens in new window)](https://www.facebook.com/sharer/sharer.php?text=Why+it%E2%80%99s+so+hard+to+see+into+the+future+of+Covid-19&u=https%3A%2F%2Fwww.vox.com%2Fscience-and-health%2F2020%2F4%2F10%2F21209961%2Fcoronavirus-models-covid-19-limitations-imhe)
* [Share this on Twitter (opens in new window)](https://twitter.com/intent/tweet?counturl=https%3A%2F%2Fwww.vox.com%2Fscience-and-health%2F2020%2F4%2F10%2F21209961%2Fcoronavirus-models-covid-19-limitations-imhe&text=Why+it%E2%80%99s+so+hard+to+see+into+the+future+of+Covid-19&url=https%3A%2F%2Fwww.vox.com%2Fscience-and-health%2F2020%2F4%2F10%2F21209961%2Fcoronavirus-models-covid-19-limitations-imhe&via=voxdotcom)
* [SHAREAll sharing options](https://www.vox.com/science-and-health/2020/4/10/21209961/coronavirus-models-covid-19-limitations-imhe)

One of the greatest challenges of the [**coronavirus**](https://www.vox.com/coronavirus-covid19) pandemic is that all levels of policy makers need to make decisions with imperfect information. Scientists still don’t know everything about how this virus is transmitted, and due to the lack of widespread testing, they also don’t know, exactly, how prevalent it is. They don’t know if the virus will show a strong seasonal effect, and decreased during the summer. They don’t know how this will all end.

One way they are trying to answer these questions is through modeling. Specifically, infectious disease models are tools — based on mathematical formulations — that try to game out what’s possible in the future. These models are varied, often confusing to interpret, and are not crystal balls, especially because the ideal data isn’t yet available. But they are a large part of what government leaders use to make decisions, influencing how resources are allocated to health care facilities and how social distancing orders are issued to the public.



**PROJECT CODE**

**import pandas as pd**

**import numpy as np**

**import matplotlib.pyplot as plt**

**import seaborn as sns**

**from sklearn import preprocessing**

**import time**

**from datetime import datetime**

**import scipy.stats as stats**

**import statsmodels.api as sm**

**US= pd.read\_csv('US.csv')**

**X = US.iloc[:, :-1] #Take all the columns except last one**

**y = US.iloc[:, -1] #Take the last column as the result**

**US.head()**

**US.info()**

**size\_mapping = {'Snohomish':1,'Cook':2}**

**US['county'] = US['county'].map(size\_mapping)**

**US**

**size\_mapping = {'Washington':1,'Illinois':2 ,'California':3}**

**US['state'] = US['state'].map(size\_mapping)**

**US**

**X = US.iloc[:, :-1]**

**y = US.iloc[:, -1]**

**print(X)**

**print(y)**

**from sklearn.impute import SimpleImputer**

**import numpy as np**

**imputer = SimpleImputer(missing\_values=np.nan, strategy='constant')**

**imputer = imputer.fit(X.iloc[:, 1:])**

**X.iloc[:, 1:] = imputer.transform(X.iloc[:, 1:])**

**from sklearn.preprocessing import LabelEncoder, OneHotEncoder**

**labelencoder\_X = LabelEncoder()**

**X.iloc[:, 0] = labelencoder\_X.fit\_transform(X.iloc[:, 0])**

**onehotencoder = OneHotEncoder(categorical\_features = [0])**

**X = onehotencoder.fit\_transform(X).toarray()**

**# Encoding the Dependent Variable**

**labelencoder\_y = LabelEncoder()**

**y = labelencoder\_y.fit\_transform(y)**

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

**X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size = 0.2)**

**from sklearn.preprocessing import StandardScaler**

**sc\_X = StandardScaler()**

**X\_train[:, 3:] = sc\_X.fit\_transform(X\_train[:, 3:])**

**X\_test[:, 3:] = sc\_X.transform(X\_test[:, 3:])**

**from sklearn.linear\_model import LinearRegression**

**linreg = LinearRegression()**

**linreg.fit(X\_train,y\_train)**

**y\_pred = linreg.predict(X\_test)**

**y\_pred**

**from sklearn import metrics**

**print(metrics.r2\_score(y\_test, y\_pred))**

**sns.heatmap(US.isnull(),yticklabels=False,cbar=False,cmap='viridis')**

**US[['date', 'fips', 'cases', 'deaths']] = US[['date', 'fips', 'cases', 'deaths']].fillna(0)**

**sns.heatmap(US.isnull(),yticklabels=False,cbar=False,cmap='viridis')**

**from sklearn.preprocessing import LabelEncoder**

**labelencoder\_X = LabelEncoder()**

**US.iloc[:, 1] = labelencoder\_X.fit\_transform(US.iloc[:, 1])**

**US['day']=pd.DatetimeIndex(US['date']).day**

**US['year'] = pd.DatetimeIndex(US['date']).year**

**US['month'] = pd.DatetimeIndex(US['date']).month**

**US.head()**

**US.columns**

**US.head()**

**sns.heatmap(US.isnull(),yticklabels=False,cbar=False,cmap='viridis')**

**US.groupby('date')["day","year","month"].sum()**

**sns.heatmap(US.isnull(),yticklabels=False,cbar=False,cmap='viridis')**

**US.groupby('date')["day","year","month"].max()**

**US\_per\_day = US.groupby('date')['day','year','month'].max()**

**US\_per\_day.head()**

**US\_per\_day.describe()**

**US\_per\_day['day'].max()**

**US\_per\_day['month'].max()**

**US\_per\_day['year'].max()**

**US['county'].value\_counts()**

**US['date'].value\_counts().plot(kind='bar',figsize=(20,10))**

**group\_data = US.groupby(["county"])["deaths", "cases"].sum().reset\_index()**

**group\_data = group\_data.sort\_values(by='deaths', ascending=False)**

**group\_data = group\_data[group\_data['deaths']>100]**

**plt.figure(figsize=(15, 5))**

**plt.plot(group\_data['county'], group\_data['cases'],color='red')**

**plt.plot(group\_data['county'], group\_data['deaths'],color='green')**

**plt.title('Total deaths(>100), Cases by County')**

**plt.show()**

**import numpy as np**

**import pandas as pd**

**import matplotlib.pyplot as plt**

**import plotly.express as px**

**from scipy.optimize import curve\_fit**

**from datetime import datetime, timedelta**

**data=pd.read\_csv("covid\_19\_data.csv")**

**data=data.drop('Last Update', axis=1)**

**data=data.drop("SNo",axis=1)**

**data=data.rename(columns={"ObservationDate": "date", "Country/Region": "country", "Province/State": "state","Confirmed":"confirm","Deaths": "death","Recovered":"recover"})**

**def plot\_predict(country, stat, future\_days):**

**def avg\_err(pcov):**

**return np.round(np.sqrt(np.diag(pcov)).mean(), 2)**

**# function to be minimized**

**def f\_sigmoid(x, a, b, c):**

**# a = sigmoid midpoint**

**# b = curve steepness (logistic growth)**

**# c = max value**

**return (c / (1 + np.exp(-b\*(x-a))))**

**inception = 0**

**# hardcoding day 0 for several countries based on observations**

**if country=="South Korea": inception = 8**

**if country=="US": inception = 28**

**if country=="Italy": inception = 20**

**country\_data = data[data["country"]==country].iloc[: , [0, 2, 3 ,4, 5]].copy()**

**country\_graph = country\_data.groupby("date")[['confirm', 'death', 'recover']].sum().reset\_index()[inception:]**

**y = country\_graph[stat]**

**x = np.arange(len(y))**

**# fitting the data on the logistic function**

**popt\_sig, pcov\_sig = curve\_fit(f\_sigmoid, x, y, method='dogbox', bounds=([12., 0.001, y.mean()],[60., 2.5, 10\*y.max()]))**

**print(popt\_sig)**

**peakday = datetime.strftime(datetime.strptime(country\_graph["date"][inception], "%m/%d/%Y")+timedelta(days=int(popt\_sig[0])), "%m/%d/%Y")**

**plt.figure(figsize=(16,8))**

**x\_m = np.arange(len(y)+future\_days)**

**y\_m = f\_sigmoid(x\_m, \*popt\_sig)**

**print("Predictions:")**

**for i in range(1,5):**

**pday = datetime.strftime(datetime.strptime(country\_graph["date"][inception], "%m/%d/%Y")+timedelta(days=len(y)+i-1), "%m/%d/%Y")**

**print("%s: %d" % (pday, y\_m[len(y)+i-1]))**

**#print(country\_graph)**

**# creating the matplotlib visualization**

**plt.plot(x\_m, y\_m, c='k', marker="x", label="sigmoid | error: "+str(avg\_err(pcov\_sig)))**

**plt.text(x\_m[-1]+.5, y\_m[-1], str(int(y\_m[-1])), size = 10)**

**plt.plot(x, y, c='r', marker="o", label = stat)**

**plt.xlabel("Days")**

**plt.ylabel("Total Infected")**

**plt.legend(prop={'size': 15})**

**plt.title(country+"'s Data", size=15)**

**plt.axvline(x[-1])**

**plt.text(x[-1]-.5, y\_m[-1], str(country\_graph["date"][len(y)+inception-1]), size = 10)**

**plt.axvline(int(popt\_sig[0]))**

**plt.text(int(popt\_sig[0]), 1, "peak: day " + str(int(popt\_sig[0])) + " (" + peakday + ")", size = 10)**

**plt.show()**

**# See the results for different countries**

**#plot\_predict("Mainland China", "confirm", 10)**

**#plot\_predict("South Korea", "confirm", 10)**

**plot\_predict("US", "confirm", 30)**

**data=data.rename(columns={"ObservationDate": "date", "Country/Region": "country", "Province/State": "state","Confirmed":"confirm","Deaths": "death","Recovered":"recover"})**

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

**X\_train,X\_test,y\_train,y\_test =train\_test\_split(X,y,test\_size=0.2,random\_state=1)**

**from sklearn import metrics**

**from sklearn.linear\_model import LogisticRegression**

**logmode1 =LogisticRegression()**

**print(logmode1.fit(X\_train ,y\_train))**

**y\_pred = logmode1.predict(X\_test)**

**from sklearn.metrics import confusion\_matrix**

**confusion\_matrix = confusion\_matrix(y\_test,y\_pred,[1,0])**

**print(confusion\_matrix)**

**from sklearn.metrics import confusion\_matrix**

**from sklearn.metrics import accuracy\_score**

**print('Logistic regression accuracy: {:.3f}'.format(accuracy\_score(y\_test, logmode1.predict(X\_test))))**

**import seaborn as sns**

**forest\_cm = metrics.confusion\_matrix(y\_pred, y\_test, [1,0])**

**sns.heatmap(forest\_cm, annot=True, fmt='.2f',xticklabels = ["Left", "Stayed"] , yticklabels = ["Left", "Stayed"] )**

**plt.ylabel('True class')**

**plt.xlabel('Predicted class')**

**plt.title('Logistic regression')**

**plt.savefig('US')**

**BIBLIOGRAPHY**

* [**https://medium.com/analytics-vidhya/predicting-the-spread-of-covid-19-coronavirus-in-us-daily-updates-4de238ad8c26**](https://medium.com/analytics-vidhya/predicting-the-spread-of-covid-19-coronavirus-in-us-daily-updates-4de238ad8c26)
* [**https://www.kaggle.com**](https://www.kaggle.com)
* [**https://www.thecut.com/2020/04/which-states-have-coronavirus-and-how-do-you-minimize-risk.html**](https://www.thecut.com/2020/04/which-states-have-coronavirus-and-how-do-you-minimize-risk.html)
* [**https://www.wired.com/story/the-mathematics-of-predicting-the-course-of-the-coronavirus/**](https://www.wired.com/story/the-mathematics-of-predicting-the-course-of-the-coronavirus/)
* [**https://www.vox.com/science-and-health/2020/4/10/21209961/coronavirus-models-covid-19-limitations-imhe**](https://www.vox.com/science-and-health/2020/4/10/21209961/coronavirus-models-covid-19-limitations-imhe)

**CONCLUSION**

This project depicts the situation of the world due to covid-19 the pandemic which has destroyed the whole world into a great loss be it population , economic damage etc.My project includes importing data, cleaning data,splitting into train /test or cross validation sets continued by pre processing and transformation. after all this the model is built up through which I can evaluate through various comparison models such as heat maps, bar graphs etc. This is how analysis is done and compared with other models leading to some prediction whether what could be the situation of the world after some time. AI is not yet playing a significant role in the fight against covid-19 at least from the epidemiological, diagnostic and pharmaeceutical point of view. Its use is constrained by a lack of data and by too much noisy and outlier data. The significant efforts of all effected countries have been to shut down their economies through lockdowns, enforcing social distancing and cancelling events. After prediction from the whole project it can be concluded that the condition of us will be stabIlize but would need time.