INTERNSHIP REPORT

**COVID 19 CASES PREDICTION**

A Report Submitted to

Jawaharlal Nehru Technological University Kakinada, Kakinada in partial fulfillment for the award of the degree of

## BACHELOR OF TECHNOLOGY

**IN**

**COMPUTER SCIENCE AND ENGINEERING(AI & ML)**

Submitted by

K.Gowtham Chowdary (20KN1A4229)

S.V.Madhava Reddy (20KN1A4255)

G.Guna Sekhar (21KN5A4202)

V.Adithya (20KN1A4262)

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

# NRI INSTITUTE OF TECHNOLOGY

**Autonomous**

##### (Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada) Accredited by NBA (CSE, ECE & EEE), Accredited by NAAC with ‘A’ Grade ISO 9001: 2015 CertifiedInstitution

**Pothavarappadu (V), (Via) Nunna, Agiripalli (M), Krishna Dist., PIN: 521212, A.P, India.**

**2020-2021**

|  |  |
| --- | --- |
| NRI | **NRI INSTITUTE OF TECHNOLOGY**  (An Autonomous Institution, Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada)  Accredited by NBA (CSE, ECE & EEE), Accredited by NAAC with ‘A’ Grade ISO 9001: 2015 Certified Institution  Pothavarappadu (V), (Via) Nunna, Agiripalli (M), Krishna Dist., PIN: 521212, A.P, India. |

**CERTIFICATE**

This is to certify that the “**Internship report”** submitted by Konadapaneni Gowtham Chowdary **(Regd. No.:20KN1A4229),** S.V.MadhavaReddy **(Regd.No.:20KN1A4255**), V.Adithya **(Regd. No.: 20KN1A4262),** G.Gunasekhar **(Regd.No. 21KN5A4202)** submitted during 2020-24 academic year, in partial fulfillment of the requirements for the award of the degree of **BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE AND ENGINEERING,** at **NRI Institute Of Technology (AI & ML), Pothavarappadu.**

**INTERNSHIP COORDIANTOR Head of the Department**

##### (Dr. D. SUNEETHA)

**EXTERNAL EXAMINER**

## CERTIFICATE OF INTERNSHIP

## ACKNOWLEDGEMENT

I take this opportunity to thank all who have rendered their full support to my work. The pleasure, the achievement, the glory, the satisfaction, the reward, the appreciation and the construction of my project cannot be expressed with a few words for their valuable suggestions.

I am expressing my heartfelt thanks to **Head of the Department, Dr. D. SUNEETHA** garu for her continuous guidance for completion of my Project work.

I am extending our sincere thanks to **Dean of the Department, Dr. K. V. SAMBASIVA RAO** for his continuous guidance and support to complete my project successfully.

I am thankful to the **Principal, Dr. C. NAGA BHASKAR** garu for his encouragement to complete the Project work.

I am extending my sincere and honest thanks to the **Chairman, Dr. R. VENKATA RAO & Secretary, Sri K. Sridhar** garu for their continuous support in completing the Project work.

Finally, I thank the Administrative Officer, Staff Members, Faculty of Department of CSE, NRI Institute of Technology and my friends, directly or indirectly helped us in the completion of this project.

**NAME**

**K.GOWTHAM CHOWDARY**

**S.V.MADHAVA REDDY**

**V.ADITHYA**

**G.GUNASEKHAR**

**REGDNO**

**20KN1A4229**

**20KN1A4255**

**20KN1A4262**

**21KN5A4202**

**ABSTRACT**

Prediction of the dynamics of new SARS-CoV-2 infections during the current COVID-19 pandemic is critical for public health planning of efficient health care allocation and monitoring the effects of policy interventions. We describe a new approach that forecasts the number of incident cases in the near future given past occurrences using only a small number of assumptions.Our approach to forecasting future COVID-19 cases involves modeling the observed incidence cases using a Poisson distribution for the daily incidence number, and a gamma distribution for the series interval. We apply our method to predicting the number of new COVID-19 cases in a single state in the India and for a subset of counties within the state to demonstrate the utility of this method at varying scales of prediction. Our method produces reasonably accurate results when the effective reproduction number is distributed similarly in the future as in the past. Large deviations from the predicted results can imply that a change in policy or some other factors have occurred that have dramatically altered the disease transmission over time. Prediction of COVID-19 by using Machine Learning could help increase the speed of disease identification resulting in reduced mortality rate. Analyzing the results obtained from experiments, Random Forest (RF) was identified to perform better compared to other algorithms.

**Keywords:** COVID-19, Machine Learning, Prediction, Supervised Learning, Regression Techniques.

**TABLE OF CONTENTS**

**TOPIC PAGE NO**

### Introduction **8-9**

* 1. [Cancer](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250027) 8
  2. [Breast Cancer 9](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250026)
  3. [Problem Statement](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250025) 9
  4. [Objectives](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250024) 9

1. [System Analysis](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250023) 10-13
   1. [Existing System](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250022) 10
   2. [Proposed System 10-11](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250021)
   3. [Hardware Requirements](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250020) 11
   4. [Software Requirements](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250019) 11
   5. [Feasibility Study](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250018) 12-13
      1. [Technical Feasibility 12](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250017)
      2. [Operational Feasibility 12](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250016)
      3. [Economic Feasibility 13](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250015)
   6. [Functional Requirements 13](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250014)
   7. [Non-Functional Requirements 13](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250013)
2. [Literature Review 14-16](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250012)
3. [System Design](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250011) 17-28
   1. [UML Diagrams 17](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250010) 
      1. [Structural Diagrams 17-21](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250007)
      2. [Deployment Diagrams 21-27](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250006)
   2. [Datasets 27-28](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250003)
4. [Technology Description](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250002) 29-38
   1. [Python 29-30](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250001)
   2. [Machine Learning 30-34](file:///C:\\Users\\Tvspreddy\\Downloads\\Intership%20Report%20(Breast%20Cancer%20Survival%20Prediction).docx" \l "_TOC_250000)
   3. Algorithms Used 35-38
      1. Logistic Regression 35-36
      2. Random Forest 36-37
      3. Decision Tree 37-38

### Software Used **39-43**

* 1. Anaconda 39
  2. Jupyter Notebook 39-40
  3. Python Packages 40-41
  4. Google Colaboratory 41-43

### Coding and Implementation **44-58**

* 1. Coding 44-57
  2. Output 58

### Testing **59-62**

* 1. Testing objectives 59
  2. White Box Testing 59
  3. Black Box Testing 60
  4. Unit Testing 60
  5. Integrating Testing 60
  6. Test Cases 60-62

### Conclusion **63**

1. **Bibliography** 64

**CHAPTER-1**

**INTRODUCTION**

## Introduction

In this era of automation, artificial intelligence and data science have important role in the health care industry. These technologies are so well-connected that medical professionals can easily manage their roles and patient care. All health care organizations work hard to develop an automated system that can be used to accept the challenges faced in health care. Scientists are working on machine learning (ML) to develop smart solutions to diagnose and treat disease. ML is capable of detecting disease and virus infections more accurately so that patients' disease can be diagnosed at an early stage, the dangerous stages of diseases can be avoided, and there can be fewer patients. In the same manner, ML can be used to automate the task of predicting COVID-19 infection and help forecast future infection tallies of COVID-19. In this chapter, we include methods for forecasting future cases based on existing data. ML approaches are used and two solutions, one for predicting the chances of being infected and other for forecasting the number of positive cases, are discussed.

A trial was done for different algorithms, and the algorithm that gave the results with the best accuracy is covered in the chapter. The chapter discusses autoregressive integrated moving average (ARIMA) time series for forecasting confirmed cases for various states in India. Two classifiers, random forest and extra tree classifier (ETC), are selected; both have an accuracy of more than 90%. Of the two, ETC has 93.62% accuracy. These results can be used to take corrective measures by different government bodies. The availability of techniques for forecasting infectious disease can make it easier to fight against infectious disease such as COVID-19.

## 1.1. Introduction to COVID-19

COVID-19 is not just a name now. It has become a deadly widespread virus that has affected tens of thousands of people all over the world. Its origin was Wuhan City, China in Dec. 2019. When people were unaware of the virus, COVID-19 started to spread from one person to another; it has slowly reached almost all countries and has become a pandemic.

COVID-19 is the short form for coronavirus disease 2019, an illness caused by a novel coronavirus (nCoV) now known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2); formerly called 2019-nCoV. COVID-19 was not the formal name of this virus; it was called SARS-CoV-2 by the International Committee on Taxonomy of Viruses because its symptoms were related to the virus that caused the SARS outbreak in 2003. However, this virus had not previously appeared in humans, and this time, they were severely infected by the virus, so to avoid confusion with other viruses, the World Health Organization (WHO) named it COVID-19 to communicate with the public.

During its early stages, COVID-19 was first identified as only an outbreak of respiratory illness cases in Wuhan City, Hubei Province, China. On Dec. 31, 2019, China reported about this respiratory disease to the WHO. It was declared to be COVID-19, a global health emergency, by the WHO on Jan. 30, 2020. According to records of WHO, in 2009, H1N1 was declared to be a global pandemic after which, on Mar. 11, 2020, COVID-19 was declared a global pandemic by the WHO.

The name COVID-19 was selected because the WHO does not want to associate the origins of the virus in terms of populations, geography, or animals to cause stigma.

According to the WHO and other health agencies, coronaviruses are defined as a collection of viruses whose symptoms ranges from the common cold to more severe diseases. However, nCoV is a new type of virus not been previously seen in humans.

Countries across the globe quickly identified this respiratory disease as the cases of COVID-19 rapidly increased. More and more people were infected with COVID-19 since the day it was identified in China. Since it was declared as the pandemic, the WHO has published guidance regarding this virus for all countries, including how the people may identify whether they are infected by this disease, how to remain unaffected by the virus, what kind of precautions should be taken care, when to go to the hospital, levels of conditions of people who are infected, and symptoms of this virus after a deep examination of infected people.

The WHO continuously shares information with people in different countries about this virus so that the public does not panic. During the early days of COVID-19, the WHO did not suggest avoiding travel. Strict suggestions were to distance from infected persons, wash hands regularly, and, if experiencing coughing or a cold, covering the mouth.

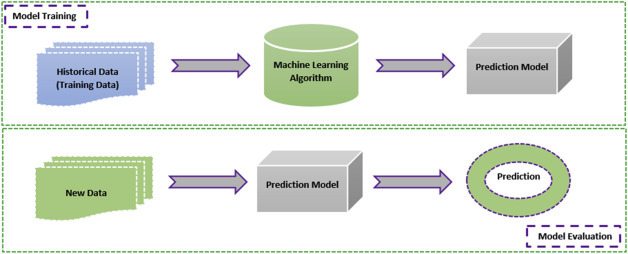
However, later on, travel history became one of the important identifiers of COVID-19, and based on this information, screening of all persons traveling from different countries, especially from infected areas, was done regularly. All persons coming from other countries were recommended to be isolated at home for around 14 days, because that was the symptomatic period of this virus, as mentioned by the WHO. If a person showed any symptoms of illness, he or she was taken to the hospital for treatment.

## 1.2. Introduction to machine learning

According to Arthur Samuel (1959), ML is the field of study that gives computers the ability to learn without being explicitly programmed. Thus, we can define ML as the field of computer science in which machines can be designed that can program themselves.

The process of learning is simply learning from experience or observations from previous work, such as examples, or instruction, to look for patterns in data and with the help of examples, provided the system can make better decisions. The basic aim of ML is to make computers learn automatically with no human intervention and to adjust perform actions accordingly. Shows the process of ML.

Past data are used to train the model, and then this trained model is used to test new data and then for prediction. The trained ML model's performance is evaluated using some portion of available past data (which is not present during training). This is usually referred as the validation process. In this process, the ML model is evaluated for its performance measure, such as accuracy. Accuracy describes the ML model's performance over unseen data in terms of the ratio of the number of correctly predicted features and total available features to be predicted.

[[](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8138040/figure/f0010/)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8138040/figure/f0010/" \t "figure)

### **1.2.1. Some machine learning methods**

ML algorithms can be divided into supervised or unsupervised learning:

Supervised ML algorithms is a type of ML technique that can be applied according to what was previously learned to get new data using labeled data and to predict future events or labels. In this type of learning, supervisor (labels) is present to guide or correct. For this first analysis, the known training set and then the output values are predicted using the learning algorithm. The output defined by the learning system can be compared with the actual output; if errors are identified, they can be rectified and the model can be modified accordingly.

Unsupervised ML algorithms: In this type, there is no supervisor to guide or correct. This type of learning algorithm is used when unlabeled or unclassified information is present to train the system. The system does not define the correct output, but it explores the data in such a way that it can draw inferences (rules) from datasets and can describe hidden structures from unlabeled data.

Semisupervised ML algorithms are algorithms that are between the category of supervised and unsupervised learning. Thus, this type of learning algorithm uses both unlabeled and labeled data for training purposes, generally a small amount of labeled data and a large amount of unlabeled data. This type of method is used to improve the accuracy of learning.

Reinforcement ML algorithms is a type of learning method that gives rewards or punishment on the basis of the work performed by the system. If we train the system to perform a certain task and it fails to do that, the system might be punished; if it performs perfectly, it will be rewarded. It typically works on 0 and 1, in which 0 indicates a punishment and 1 indicates a reward.

It works on the principle in which, if we train a bird or a dog to do some task and it does exactly as we want, we give it a treat or the food it likes, or we might praise it. This is a reward. If it did not perform the task properly, it might be scolded as a punishment by us.

**1.3. Use of machine learning in COVID-19**

ML is used in various fields, including medicine to predict disease and forecast its outcome. In medicine, the right diagnosis and the right time are the keys to successful treatment. If the treatment has a high error rate, it may cause several deaths. Therefore, researchers have started using artificial intelligence applications for medical treatment. The task is complicated because the researchers have to choose the right tool: it is a matter of life or death.

For this task, ML achieved a milestone in the field of health care. ML techniques are used to interpret and analyze large datasets and predict their output. These ML tools were used to identify the symptoms of disease and classify samples into treatment groups. ML helps hospitals to maintain administrative processes and treat infectious disease.

ML techniques were previously used to treat cancer, pneumonia, diabetes, Parkinson disease, arthritis, neuromuscular disorders, and many more diseases; they give more than 90% accurate results in prediction and forecasting.

The pandemic disease known as COVID-19 is a deadly virus that has cost the lives of many people all over the world. There is no treatment for this virus. ML techniques have been used to predict whether patients are infected by the virus based on symptoms defined by WHO and CDC.

ML is also used to diagnose the disease based on x-ray images. For instance, chest images of patients can be used to detect whether a patient is infected with COVID-19.

Moreover, social distancing can be monitored by ML; with the help of this approach, we can keep ourselves safe from COVID-19

**1.3.1. Different techniques for prediction and forecasting**

Various ML techniques are used to predict and forecast future events. Some ML techniques used for prediction are support vector machine, linear regression, logistic regression, naive Bayes, decision trees (random forest and ETC), K-nearest neighbor, and neural networks (multilayer perceptron).

Similarly, some ML techniques used to forecast future events are naive approach, moving average, simple exponential smoothing, Holt's linear trend model, Holt-Winters model, Seasonal Autoregressive Integrated Moving Average Exogenous Model (SARIMAX) and Autoregressive Integrated Moving Average Model (ARIMA).

Each technique has unique features and is used differently based on the accuracy results. The model with the best accuracy during the model evaluation process is chosen for prediction or forecasting. In the same way, we identified and used the ETC for the symptom-based prediction of COVID-19 and the ARIMA forecasting model to forecast the number of confirmed cases of COVID-19 in India, because they had the best accuracy results among all classifier and forecasting methods we used when we evaluated model performance. Shows a flowchart of the ML process. It defines how data are collected and pre-processed, and then are divided into a training dataset and test dataset for training and performance evaluation.

**PROBLEM STATEMENT:**

My goal was to develop public models that would reasonably predict the growth of the coronavirus for the next day. This goal can be broken down and formalized into two parts:

* To build a model that will predict the next day’s number of coronavirus cases based on the last n day’s number of coronavirus cases and deaths.
* To build a model that will predict the next day’s number of coronavirus deaths based on the last k day’s number of coronavirus cases and deaths.

While I learned that n and k should be the same value after rigorous testing of my models, when I first defined the problem, I assumed that n and k may be different numbers as to not limit the possibilities of improving my individual models.

**OBJECTIVES:**

The main goal of this thesis is to develop a machine learning model that could predict whether a patient is suffering from COVID-19. To develop such a model, a literature study alongside an experiment is set to identify a suitable algorithm. To assess the features that impact the prediction model.

**SYSTEM ANALYSIS**

**EXISTING SYSTEM:**

With the outbreak of COVID-19 in 2019, which lead to many deaths and global crisis. Many service-based organisations focussed on prediction of new cases, deaths casued by COVID-19 virus for next dew days. Those organisations may include:

* WHO
* CDC(Centers for Disease Control and Prevention)
* IHME(Institute for Health Metrics and Evaluation)

The existing system for prediction of covid-19 cases may also include with many government developed application in order to alert people**.**

**PROPOSED SYSTEM:**

ML achieved a milestone in the field of health care. ML techniques are used to interpret and analyze large datasets and predict their output. These ML tools were used to identify the symptoms of disease and classify samples into treatment groups. ML helps hospitals to maintain administrative processes and treat infectious disease.

The pandemic disease known as COVID-19 is a deadly virus that has cost the lives of many people all over the world. There is no treatment for this virus. ML techniques have been used to predict whether patients are infected by the virus based on symptoms defined by WHO and CDC.

MODELS: Support Vector Machine (SVM), Multiple Linear Regression, Bayesian Ridge.

**SOFTWARE REQUIREMENTS:**

* **PYTHON:** Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured, object-oriented and functional programming.
* **JUPYTER:** Jupyter notebook is an open-source IDE that is used to create Jupyter documents that can be created and shared with live codes. Also, it is a web-based interactive computational environment.
* **MATPLOTLIB:** Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.
* **Scikit-learn:** Scikit-learn is a free software machine learning library for the Python programming language.[3] It features various classification, regression and clustering algorithms including support-vector machines, random forests, etc.
* **PANDAS:** Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.
* **NUMPY:** NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

**HARDWARE REQUIREMENTS:**

* **Processor:** 64 bit, quad-core, 2.5 GHz minimum per core
* **Ram:** 4 GB or more
* **Hard disk:** 20 GB of available space or more.
* **Display:** Dual XGA (1024 x 768) or higher resolution monitors

## 

## FEASIBILITY STUDY

Preliminary investigation examines project feasibility; the likelihood the system will be useful to the organization. The main objective of the feasibility study is to test the Technical, Operational and Economical feasibility for adding new modules and debugging old running system. All systems are feasible if they are given unlimited resources and infinite time. There are aspects in the feasibility study portion of the preliminary investigation.

* Technical Feasibility
* Operation Feasibility
* Economic Feasibility

### TECHNICAL FEASIBILITY

To determine whether the proposed system is technically feasible, we should take into consideration the technical issues involved behind the situation. Technical feasibility center on the existing computer system and to what extent it can support the proposed addition. Python and its libraries are technology software which are used to develop Data Analytics. So, there is no need for additional purchase of any software and these are open source softwares which are freely available in Internet.

### OPERATIONAL FEASIBILITY

Proposed projects are beneficial only if they can be turned out into information systems that will meet the user’s operating requirements. Operational feasibility aspects of the project are to be taken as an important part of the application implementation. This system is operational feasible since the users are familiar with the technologies and hence there is no need to gear up the personnel to use the system. Also, the system is very friendly and easy to use.

### ECONOMIC FEASIBILITY

To decide whether a project is economically feasible, we have to take into consideration various factors as:

* + - * Cost benefit analysis
      * Long-term returns
      * Maintenance costs

The proposed system is computer based. It requires average computing capabilities which is very basic requirement and can be afforded by an organization; it doesn’t incur additional economic overheads, which renders the system economically feasible.

## FUNCTIONAL REQUIREMENTS

The functional requirements of the given system are as follows

* Dataset collection
* Create model
* Train the model
* Test the model
* Take video input from user
* Process and classify the video
* Print the output gesture in text format

## NON-FUNCTIONAL REQUIREMENTS

### USABILITY:

The system is designed with completely automated process. Hence there is no or less user intervention.

### RELIABILITY:

The system is more reliable because of the qualities that are inherited from the chosen platform python. The code built by using python is more reliable.

### PERFORMANCE:

The system is developed in the high-level language and using the advanced technologies it will give response to the end user efficiently.

### SUPPORTABILITY:

The system is designed to be the cross platform supportable. The system is supported in wide range of hardware and any software platform

**Literature Review**

A systematic literature through the guidelines of Claes Wohlin and Barbara Kitchenham, has been conducted to analyze and answer RQ1. This literature review focuses on the understanding of several machine learning algorithms and also identifying appropriate machine learning algorithms that can be used for prediction. There are several steps that we performed in our research, which are:

**1. Identifying the key words:** We have identified the following keywords which are Supervised Machine Learning algorithms, COVID19, classification, prediction.

**2. Formulating the search strings:** From the above identified keywords, primary keywords were selected to formulate the search string.

**3. Locating the literature:** Using search string, the search was performed on various digital database platforms such as Google scholar, IEEE and Science Direct.

**4. Following the Inclusion and Exclusion criteria for selection:** From the collected literature such as articles and conference papers, the inclusion and exclusion criteria is implemented to confine our research.

**Inclusion Criteria**

• Papers related to prediction of COVID-19 using Machine Learning algorithms.

• All articles should be in English language.

**Exclusion Criteria**

• Incomplete articles

| **No.** | **Study** | **Objective** | **Type of model** | **Result** | **Quality assessment** |
| --- | --- | --- | --- | --- | --- |
| 1 | Yang et al., 2020 | To forecast COVID-19 patterns in China using a SEIR and AI model | SEIR model and AI model | · The model was effective in forecasting COVID-19 cases. | 95% CI |
| 2 | Liang et al., 2020 | To forecast the risk of critical illness at hospital admission and identify survival of COVID-19 patients | Statistical software: LASSO, logistic regression model | · The score gives an estimation of the probability of critical disease progression for a hospitalized patient with COVID-19. | AUC (accuracy) was 0.88, 95% CI. |
| 3 | Yan et al., 2020 | Relieving clinical burden and potentially reducing the mortality rate of COVID-19 | Machine learning tool: XGBoost | To predict patients with higher risk and potentially reduce mortality rate | Overall  accuracy was 0.90 |
| · Survival prediction accuracy was 100%. |
| · Mortality forecast accuracy was 81%. |
| 4 | Gong et al., 2020 | To predict the early detection of cases at high risk for progression to serious COVID-19 | Statistical analysis | · Results helped in COVID-19 patient identification for effective management. | Training cohort: |
| · AUC was 0.912, 95% CI. |
| Validation cohort: |
| · AUC was 0.853, 95% CI. |

**Keywords and Definitions**

**Implementation and architecture**

Machine Learning is a subset of Artificial Intelligence (AI) and was evolved from pattern recognition where the data can be structured for the understanding of the users. Recently, many applications have been developed using Machine Learning in various fields such as healthcare, banking, military equipment, space etc. Currently, Machine Learning is a rapidly evolving and continuously developing field. It programs computers using data to optimize their performance. It learns the parameters to optimize the computer programs using the training data or its past experiences. Using the data, it can also predict the future. Machine Learning also helps us in building a mathematical model using the statistics of the data. The main objective of Machine Learning is that it learns from the feed data without any interference of humans that is, it automatically learns from given data(experience) and gives us the desired output where it searches the trends/patterns in the data. It is broadly classified into four types:

* Supervised Machine Learning.
* Unsupervised Machine Learning.
* Semi-Supervised Machine Learning.
* Reinforcement Machine Learning.

**Supervised Machine Learning**

Supervised Learning is a Machine Learning model that is built to give out predictions. This algorithm is performed by taking a labelled set of data as input and also known responses as output to learn the regression/classification model. It develops predictive models from classification algorithms and regression techniques.

**Classification** predicts discrete responses. Here, the algorithm labels by choosing two or more classes for each example. If it is done between two classes then it is called binary classification and if it is done between two or more classes then it is called multi- class classification. Applications of classification includes hand writing recognition, medical imaging etc.

**Regression** predicts continuous responses. Here, the algorithms returns a statistical value. For example, a set of data is collected such that the people are happy when 5 6 Chapter 2. Background considered the amount of sleep. Here, sleep and happy are both variables. Now, the analysis is done by making predictions. The types of popular regression techniques are:

* Linear regression.
* Logical regression.

**Unsupervised Machine Learning**

Unlike the supervised learning, there is no supervisor here and we only have input data. Here, the basic aim is to find certain patterns in the data that occur more than others. According to the statistics, it is called density estimation. One of the methods for the density estimation is called clustering. Here, the input data is formed into clusters or groupings. Here, the assumptions are made such that the clusters are discovered which will match reasonably well with a classification. This is a datadriven approach that works better when provided with sufficient data. For example, the movies in Netflix.com are suggested based on the principal of clustering of movies where several similar movies are grouped based on customer’s recently watched movie list. It mostly discovers the unknown patterns in the data but most of the time these approximations are weak when compared with the supervised learning.

**Semi-supervised Machine Learning**

The name “semi-supervised learning” comes from the fact that the data used is between supervised and unsupervised learning. Semi-supervised algorithm has the tendency to learn both from labelled and unlabeled data. Semi-supervised machine learning gives high accuracy with a minimum annotation work. Semi-supervised machine learning uses mostly unlabeled data together combined with labelled data to give better classifiers. As less annotation work is enough to give good accuracy, humans have less work to do here.

**Reinforcement Machine Learning**

Reinforcement learning learns its behavior from a trial and error method in a dynamic environment. Here, the problem is solved by taking an appropriate action in a certain situation to maximize the output and to obtain the acquired results. In Reinforcement Learning, there is presentation of the input or output data. Instead, when the desired action is chosen, the agent is immediately told the reward and the next state are not considering the long terms actions. For the agent to act optimally it should have the knowledge about states, rewards, transitions and actions actively. Formally, the model consists of:

* a discrete set of environment states, S;
* a discrete set of agent actions, A;
* a set of scalar reinforcement signals; typically {0;1} or the real numbers.

**ALGORITHMS:**

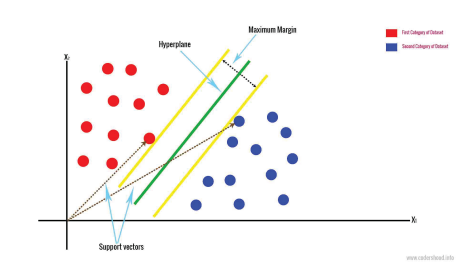
During our research, we have investigated three algorithms through which we have performed supervised classification and regression techniques

* Support Vector Machine
* Linear Regression
* Bayesian

**Support Vector Machines (SVM)**

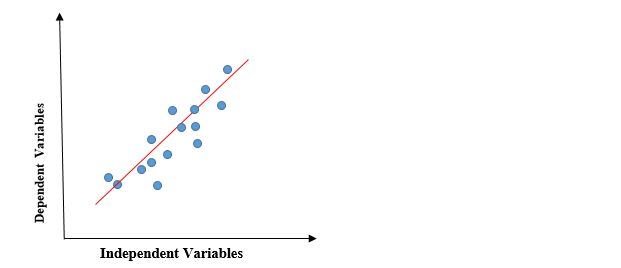
Support Vector Machines performs classification by constructing N-dimensional hyper plane that separates the data into two categories [12]. In SVM, the predictor variable is called an attribute and the transformed attribute is called a feature. Selecting the most suitable representative data is called feature selection. A set of features describing one case is called a vector.

The ultimate goal of SVM modelling is to find the optimal hyper plane that separates the clusters where on one side of the plane there is target variable and on the other side of the plane other category. The vectors which are near the hyper plane are the support vectors.



## **Linear Regression**

Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called ****simple linear regression****. And if there is more than one input variable, such linear regression is called ****multiple linear regression**.** The linear regression model gives a sloped straight line describing the relationship within the variables.

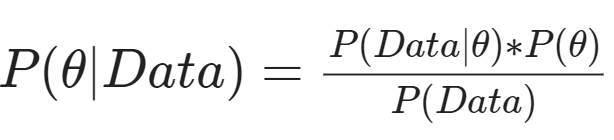


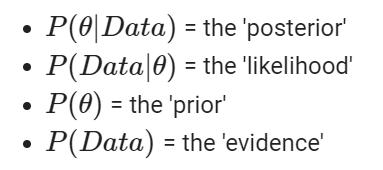
The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (**independent variable**) increases, the value of y (**dependent variable**) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.

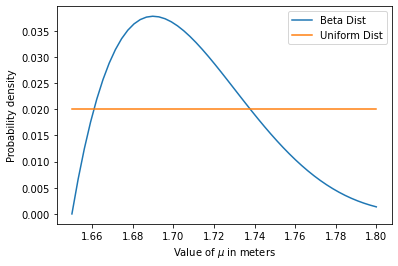
**Bayesian**

Bayesian inference is a method to figure out what the distribution of variables is (like the distribution of the heights h). The interesting feature of Bayesian inference is that it is up to the statistician (or data scientist) to use their prior knowledge as a means to improve our guess of how the distribution looks like.

Bayesian inference depends on the principal formula of Bayesian statistics: Bayes’ theorem. Bayes’ theorem takes in our assumptions about how the distribution looks like, a new piece of data, and outputs an updated distribution. For data science, Bayes’ theorem is usually presented as such:

Statisticians also gave each component of this theorem names:





**Architecture**

Getting Data

Data Cleaning and preprocessing

Visulaization and EDA

Training Model and Testing Model

Prediction of COVID-19 cases for upcoming days