BMS COLLEGE OF ENGINEERING
(Autonomous Institute, Affiliated to VTU)
Bull Temple Road, Basavanagudi, Bengaluru - 560019

A project report on

"COVID-19 PREDICTION USING MACHINE LEARNING"

Submitted in partial fulfilment of the requirements for the award of degree

BACHELOR OF ENGINEERING

IN

INFORMATION SCIENCE AND ENGINEERING

By

MANASI M (1BM18IS052)

PRAMILA DALAVAI(1BM18IS068)

Under the guidance of

SINDHU K

ASSISTANT PROFESSOR

Department of Information Science and Engineering
2021-2022

BMS COLLEGE OF ENGINEERING
(Autonomous Institute, Affiliated to VTU)
Bull Temple Road, Basavanagudi,
Bengaluru – 560019

Department of Information Science and Engineering

CERTIFICATE

This is to certify that the project entitled **“COVID-19 PREDICTION USING COVID-19”** is a bona-fide work carried out by Manasi M(1BM18IS052) and Pramila Dalavai(1BM18IS068) in partial fulfilment for the award of degree of Bachelor of Engineering in **Information Science and Engineering** from **Visvesvaraya Technological University, Belgaum** during the year **2021-2022**. It is certified that all corrections/suggestions indicated for Internal Assessments have been incorporated in the report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the Bachelor of Engineering Degree.

Sindhu K	**Dakshayini M**	**Ravishankar B V**
Assistant Professor	**Professor and HOD**	**Principal**

Examiners

Name of the Examiner	**Signature of the Examiner**
1.	
2.	

ABSTRACT

Covid-19 is a disease that has affected the entire world. It is caused by a new family of viruses called Coronaviridae. This was formerly found in China. Later the disease was carried by infected persons to other parts of the world and soon it was declared a pandemic. A person affected with the disease may experience mild symptoms like cough, cold, breathing problem and body pain. There are no specific medicines discovered but people can get vaccinated to reduce the severity of the disease. People are advised to wear masks, wash their hands regularly and maintain social distancing to reduce the spread of the disease.

The pandemic has impacted life as we know it around the world. The swift spread of the COVID-19 epidemic is remarkable with increasing spread speed. It plays a more and more important role in efficiently and precisely predicting the characteristics of this disease. Hence we focus on building machine learning-based prediction of COVID-19 diagnosis based on symptoms. The dataset contains initial records, on a daily basis, of all the residents who were infected for COVID-19 nationwide. Information like test date and results are added, including clinical symptoms in binary indication. Based on the dataset of tested patients, we developed a model that predicts COVID-19 test results using features like cough, fever, headache, shortness of breath, and sore throat, and also contacts with an infected individual.

ACKNOWLEDGEMENT

We would like to express our deepest appreciation to all those who provided us the support to complete this project. A special gratitude to our project guide, Mrs. K Sindhu, whose contribution in stimulating suggestions and encouragement helped us through every step of building this project. We would also like to extend our sincerest thanks to our HOD, Dr. M Dakshayini who has been a constant source of support in all our endeavors.

Manasi M (1BM18IS052)
Pramila Dalavai(1BM18IS068)

CONTENTS

SL NO	TOPIC	PG NO
1.	Chapter 1: Introduction 1.1 Background Work 1.2 Motivation 1.3 Problem Statement 1.4 Scope 1.5 Existing System 1.6 Proposed System	1 - 5
2.	Chapter 2: Literature Survey	6 – 13
3.	Chapter 3: Requirement Analysis 3.1 Product Functions 3.2 User Specification 3.3 Operating Environment 3.4 Technical Requirements	14 – 15
4.	Chapter 4: Design 4.1 System Architecture 4.2 Data Flow Diagram 4.3 Use Case Diagram	16- 18
5.	References	20-21

Chapter 1

INTRODUCTION

1.1 Background Work

One of the primary concept behind this project is Machine Learning. The main goal of machine learning is to understand the structure of data found in dataset given and fit that data into several models that can be understood and utilized by human. In traditional computing way, algorithms were sets of explicitly programmed instructions used by computers to calculate or solve the given problem statement. Machine learning models instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range.

1.2 Motivation

The motivation for this project comes from considering the drastic growth of pandemic. Now it’s hard to break the covid chain and people still unaware of the characteristics of the virus leading to the wide spread. Because of this, we have used machine learning model that facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs. It also helps doctors to recheck further verification of virus by the proposed method hence training based on symptoms.

1.3 Problem Statement

Training the dataset and the predicting results are compared with the results predicted by doctors and validation is performed between the predicted results and the original data based on some predefined metric and model. The experimental analysis showcases that the proposed machine learning approach is useful in generating suitable results based on the critical disease outbreak. The outbreak of Coronavirus has the nature of potentially exponential growth and so it is difficult to control with limited clinical persons for handling a huge number of confirmed or tested patients with in a reasonable given time. So it is necessary to build an automated model, based on machine learning approach, for corrective measure to self-analyse the rate of symptom before the decision of clinical doctors. It could be a very well promising supplementary confirmation method for frontline clinical doctor’s safety. The proposed methods is comprised of three algorithms with probable accuracy metric in identification of the disease.

1.4 Scope

Developing a machine learning model that could predict whether a patient is suffering from COVID-19 is the main aim behind this analysis. To develop such a model, a literature study and high level design diagrams are used to find a suitable algorithm. The main objective of this model is to predict whether a person has COVID-19 or not, using machine learning techniques. The prediction is performed using a dataset of clinical symptoms of the patient.

The main objectives of our analysis are:

- Identifying the most suitable machine learning technique for predicting based on the clinical reports of patients.

- Preparing an accurate machine learning model that can make predictions of COVID-19 in patients.

- Identifying the features that affects the prediction of COVID-19 in patients.

As such , the following groups may also find use in the project –

1. Data Science Analyst- Trying to study about the extent of spread of disease. Giving reports to help Government and Corporate industries to overcome the upcoming economic crisis.
2. Researchers – For the detail study of the outbreak strategy considering dependent and independent features as input and output features.
3. Users – Self validate their symptoms based on their input and get tested by preventing the spread.

1.5 Existing System

- Automated AI devices like 'Prophet' and Open source can be applied and are viable for determining the spread of Covid-19. It might help nations that need assets to proficiently dispense medical services assets to contain this pandemic.
- Time series forecasting model – This remembers 15 nations for the investigation and considered 13 elements which remember the examination of eight elements for the Containment and Closure classification, two components in Economic classification, and three elements in Health System approaches. This model mostly dissected the effect of the above factors by looking at the estimated number of influenced individuals with the real complete infected cases revealed in those five days. The result of this examination finds the way that out of thirteen approach factors, the nations which focused more on strategies in the monetary classification during the pandemic have helped in controlling the spread of Coronavirus.
- There exists an online AI framework to investigate the unique pattern of the COVID-19 pandemic, work with gauging and prescient displaying, and produce a heatmap representation of strategy measures in 171 nations. The dynamic heatmap diagrams with strategy measures portray changes in COVID-19 measures for every country. 19 measures were inserted inside the three areas introduced on the site, and just 4 of the 19 measures were consistent measures identified with monetary help or venture. Profound learning models were utilized to empower COVID-19 estimating dependent on information gathered utilizing the site.

1.6 Proposed System

Machine learning algorithms have been utilized to develop prediction models in various infectious and non-infectious settings including interpretation of images in predicting the outcome of diseases by remembering understanding of pictures for anticipating the result of illnesses. To control the spread of this infection numerous nations have taken drastic actions yet at the same time couldn't handle the spread. The essential goal of this examination is to arrange the different strategies and clinical factors that deal with the spread of Covid-19. Our proposed model utilizes a basic machine calculation on a dataset acquired about tried cases to more readily comprehend the illness elements. Moreover, we demonstrate the probability of covid-19 cases using various machine algorithms to a dataset obtained about tested cases to better understand the disease dynamics.

The point of this examination is to create a region-level forecast around not-so-distant future illness development for COVID-19 events utilizing openly accessible information, publicly available dataset. The rapid spread of COVID-19 means that government and health services providers have a brief period to plan and plan compelling reaction approaches. and also design effective response policies.

Restricted testing offices and deferred results present a huge variety in announced cases, which delivers an inclination in the model. Hence we created linear, logistic, and decision tree models that can be utilized for the recognizable proof of confirmed cases ratio and potential information disparities and bias in the model.

Additionally, the proposed system will separate its logic and UI into a CLI and GUI respectively. This will allow for it to have customizable GUIs to accompany the predictive analysis and will also allow users to verify or get to know the probability of corona based on 0/1 indicating YES/NO input respectively.

Chapter 2

LITERATURE SURVEY

Amir Ahmad et al [1] proposed how machine learning models used to detect the number of confirmed covid-19 cases can be grouped into 4 categories and also discussed the various challenges in this field. He has identified 4 research themes – traditional machine learning regression, deep learning regression, network analysis and social media and search queries data-based methods. Two approaches of regression have been used to estimate the number of confirmed cases of covid-19 – Time series analysis and relationship between confirmed cases and other factors such as temperature, humidity. Various models of deep learning neural networks have been applied to predict covid-19. LSTM is one such model. This model has been depicted in various research papers for the prediction of confirmed cases in India. It has been found out that the combination of LSTM and gated recurrent unit produce better results than individual methods. Network analysis can be used to study the spread of infectious diseases. Many papers which have used this technique have been discussed. In one paper, network of estimations is used to estimate the spread of covid-19. The data obtained through search queries is used to monitor and predict infectious diseases. Social media search indexes(SMSI) of covid-19 symptoms can be used to predict the number of confirmed cases.

K.C Santosh[2] explained how continuous and unprecedented factors such as hospital capacity, test capacity, population density and demographics can be used to design complex models. The prediction models can use different factors such as number of possible confirmed cases, number of possible hospitalised cases and number of possible death cases, but for this the models require to have important properties such as incubation period , transmissibility and severity. To better understand covid-19 outbreak, data visualisation tools can be used. For example simulations always help better understand the particular events. Comprehensive data exploitation is required for predictive modelling. Missing one/two features can deviate predictive values from actual ones. It was found out that three types of models have been used to describe the characteristics of the disease and forecast accordingly – SEIR/SIR models, Agent-based models and curve-fitting models. Prediction models are expected to tune their hyper parameters over time, so data-driven models which automatically tune hyper parameters are needed.

Furqan Rustam et al [3] demonstrated how ML models can be used to forecast the number of upcoming patients affected by covid-19. Four standard forecasting models, such as linear regression (LR) , least absolute shrinkage and selection operator (LASSO), support vector machine (SVM) and exponential smoothing (ES) have been used to forecast the threatening factors of covid-19 pandemic. Three types of predictions are made by each of these models such as number of newly infected cases, the number of deaths, and the number of recoveries in the next 10 days. The results prove that the ES performs best among all the used models followed by LR and LASSO while SVM performs poorly. The dataset used for the study contains information about the daily reports of the number of newly infected cases, the number of recoveries and the number of deaths due to covid-19 worldwide. This study is an attempt to forecast the number of people that can be affected for the upcoming 10 days. According to the results of LASSO and LR models, the death rates will increase in upcoming days and recovery rate will slow down. SVM produces poor results in all scenarios because of the ups and downs in the dataset values.

Aman Khakharia et al [4] proposed an outbreak prediction system for covid-19 for the top 10 highly and densely populated countries. The proposed model forecast the count of new cases likely to arise for successive 5 days using 9 different machine learning algorithms. By predicting the progression of covid-19, hospitals and healthcare systems can take care of allocating all the required resources beforehand. By using the dataset of 10 highly populated countries, the models were trained and different accuracies for each of the countries were obtained. The highest accuracy obtained was 99.93% which was achieved by ARMA (Auto Regressive Moving Average) model for Ethiopia. The overall best model for prediction was ARIMA (Auto Regressive Integrated Moving Average). Generating high-accuracy prediction that could help in an optimized use of available resources along with pacing up the recovery rate has been the main aim behind this study. This model could help in lowering the cost of dealing with the pandemic and improve the recovery process in regions where it is deployed.

Iman Rahimi et al [5] demonstrated a review and brief analysis of the most important machine learning forecasting models against covid-19. About 920 technical research articles that contain only algorithmic descriptions, review articles, conference papers, case studies and provide managerial insights were selected , which were published as of October 10, 2020. This paper identified the most important subject areas by keywords analysis. The most useful models that researchers have applied for predicting this pandemic were also found out. According to keywords analysis, trends present that studies on Covid-19 will increase in the next few months. Deep learning, SIR and SEIR are the top models that were used by researchers.

Narayana Darapaneni et al [6] proposed a machine learning model to predict covid-19 among a population who have undergone some clinical and blood tests. Here two tasks are performed – a) Predict confirmed cases of covid-19 among suspected cases based on the results of the clinical tests. b) Predict the no of people who need to get admitted to general, semi-ICU and ICU wards among those who predicted positive in the first task. Classification machine learning approach has been used to perform these tasks. These results can be used to build automated systems which predict the likeliness of covid-19. Adding attributes related to flu and liver tests help improve the model by increasing its f1-score, precision and recall. As many countries are facing shortage of testing kits, this model can be used to detect the number of positive cases on a quicker note which can enable the patients to detect it quickly and get admitted accordingly. Also, the healthcare staff can minimize contact with people as the no of persons coming to the hospital for check-up will be reduced.

Vartika Bhadana et al [7] explained how ML models can be used to estimate the no of forthcoming covid-19 confirmed patients. A comparative study of five machine learning algorithms – Linear regression(LR), decision tree, least absolute shrinkage and selection operator (LASSO), random forest and support vector machine (SVM) was performed. Each of these models makes three types of forecasts – the total no of active cases, the total no of deaths and the total no of recoveries in the next 5 days. A six-degree polynomial was used to get better accuracy. The best results were obtained by poly LR and poly LASSO followed by LR, LASSO, random forest and decision tree. Poor result was demonstrated by SVM model. According to the results of poly LASSO and poly LR models, death rate and recovery rate will increase in the coming days. Hence, this research can be a great benefit for people to manage the crisis of Covid-19.

Ashish U Mandayam et al [8] proposed a model to predict the future number of positive cases for better measures and control. Two supervised learning models were used to predict the number of positive cases which may occur in the future. Time-series dataset of covid-19 was used. By the help of predictive analysis and supervised models, prediction of future cases will be helpful for taking much better preventive measures and precautions. The comparison between Linear regression and Support Vector Regression was also done. It was observed that Linear regression performed better with time series data when compared to SVM. Since the dataset used here is linear, SVM cannot handle linear datasets very effectively. The results show that Covid-19 pandemic cases is growing linearly everyday and this poses as a major threat until a year or two. Hence utmost precautions and care needs to be taken to decrease the spread of this pandemic.

Zoabi, Y et al [9] By October 2020, we found that COVID-19 was spread amongst 180 countries exceeding 395000000 patients from the statistical analysis. Around 1,110,000 people died from covid-19, contributing to increased infection rates and delays in critical preventive measures. The machine learning model can be built using clinical signs and symptoms. Initial test screening is done accordingly studying the pattern spectrum of symptoms. This is most important in developing countries that have limited health systems and resources. In contribution to this, the research states that we utilize machine learning models to understand its everyday exponential behavior along with the prediction of the future.

Yan, L., Zhang, H.T et al [10] proposed a machine learning model for predicting survival and death rates based on infection patterns and symptoms. In this study, they identified 3 indicators with thresholds for COVID-19 prognostic prediction. They developed an ml-based model that can predict the survival rates of severe patients with more than 90% accuracy using the last sample and 90% from the blood sample, enabling detection, early intervention, and potential reduction of mortality in high-risk patients with COVID-19. This suggests an advanced symptom-based model as disease severity can be accurately predicted using 3 biomarkers, therefore greatly reducing the space of clinical parameters to be monitored and the associated medical burden.

Ogundokun, R.O et al [11] Integrated machine learning methods mainly Neural networks and Linear regression to provide real-time predicting COVID-19 disease in India, reckoning the pandemic, tracking COVID-19 disease asperity, predicting the extent of the pandemic together with supporting government and health systems to constitute statistical analysis and strategy towards the eradication of the COVID-19 diseases in India. The aggregation of the prediction from various methods substantially decreases prediction errors and consequently makes available advanced precision.

Ardabili, S.F et al [12] Although ML methods were used in modeling former epidemics like Dengue fever, Zika, Ebola, Swine fever, there is a gap in the literature for reviewed papers dedicated to COVID-19. The study represents ML methods used for outbreak prediction. The ML methods are limited to the basic methods of random forest, naïve Bayes, genetic programming, and classification and regression tree (CART). Although ML has long been established as a standard tool for modeling natural disasters and also weather forecasting, its application in modeling outbreaks is still in the early stages. More sophisticated ML methods like ensemble and hybrids are yet to be explored. Consequently, the contribution of this paper is to study the application of ML for diagnosing the COVID-19 pandemic. This paper aims to investigate the potential ability of the proposed ML models and the accuracy of the proposed models.

Cristina Menni 1,7 [13] Although many people have presented with flu-like, cold, general viral disease symptoms, widespread population testing is not yet available in most countries due to lack of resources. Thus, it is important to identify the combination of clinical symptoms most predictive of COVID-19, to help guide recommendations for isolation and prevent further spread of the corona wave chain. Case reports and mainstream media articles from various states indicate that a number of patients with diagnosed COVID-19 developed a loss of smell and taste.

Zohair Malki et al [14] Compared to the previous work of the other surveys, this paper includes more features that can influence the spread of the COVID-19. The additional features that are included are weather and climatic conditions. This research comprises the best predictive model for daily tested cases in countries with the highest number of positive cases in the world. Proposed a simplified prediction model to predict the number of positive cases to have more readiness in healthcare systems and make forecasts using advanced machine learning models. Each of the models was trained with input features like population density, fertility rate, Age, Intensive Care Unit (ICU) beds along with the additional influence of temperature, humidity, wind, and also hours of sunlight.

Shawni Dutta and Dr. Samir Kumar Bandyopadhyay [15] For assisting health planning systems for Covid-19, an ML framework is proposed in this paper. This will confirm the positive, negative, recovered, and death cases considering the tested individuals present in the dataset. The proposed ml model ensures that it follows the original result regarding this pandemic situation so that economic loss, contact community spread, amount of social distance among people may be detected, and also accurate decisions can be taken accordingly. This method will ensure government authorities yield preventive measures based on our next work for forecasting the occurrence of this disease in the future. In this paper, data mining concepts are exploited for obtaining prediction of confirmed cases and recovered cases are also employed. The actual output and predicted outputs are compared.

Chapter 3

REQUIREMENTS ANALYSIS

The main end result this project aims for is a tool that allows its users to predict the probability of having covid-19 by analyzing their symptoms. And the project seeks to do this by using technologies such as Machine Learning, Python and Streamlit. And to achieve this result, we define the following requirements: -

3.1 Model Functions

- **Import libraries** : The python libraries required for the model are imported.
- **Import dataset** : Extract the dataset of the patient’s symptoms.
- **Create Python Scripts**: The script contains the ML code.
- **Execute the model**: The code is executed and the model is saved in sav format.

3.2 User Specification

- Typical Users of a machine learning based web application like this include people and also clinical doctors who tend to do self analysis. Help people to report and manage their symptom pattern.

- Advanced Users include data analysts and researchers who make use of dataset and include other features to predict survival and death rate. Also helps to know the spread of epidemic and indicate future waves by studying past spectrum.

3.3 Operating Environment

- Windows 7/8/10 (All Python compatible versions of Windows)

3.4 Technical Requirements

- Python (Application)

– Numpy

– Pickle

– Matplotlib

– Pandas

– Sklearn

– Seaborn

_ Collab Notebook

_ Streamlit (Virtualization and Code Implementation)

_ Git and GitHub/ Gitlab / Bitbucket (Management of Project and Version Control)

Chapter 4

DESIGN

4.1 System Architecture

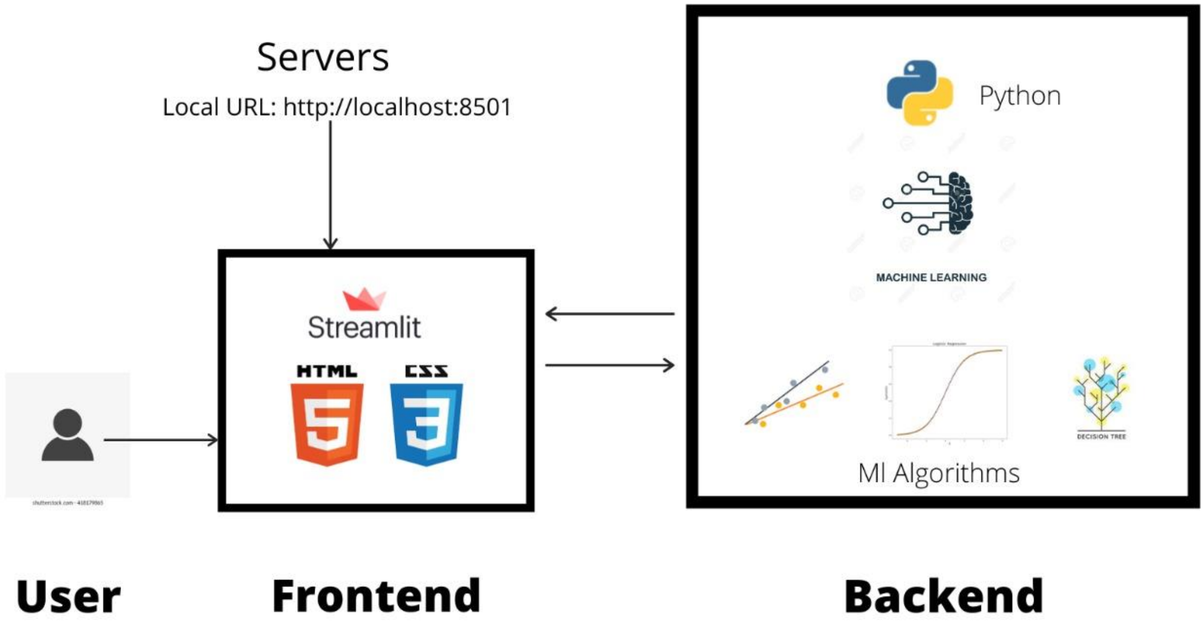

Figure 4.1: This diagram demonstrates a high-level view of the underlying architecture used in the application.

The above diagram presents the differentiation between the client application and the servers it interacts with while also presenting the underlying technologies involved in each.

4.2 Data Flow Diagram

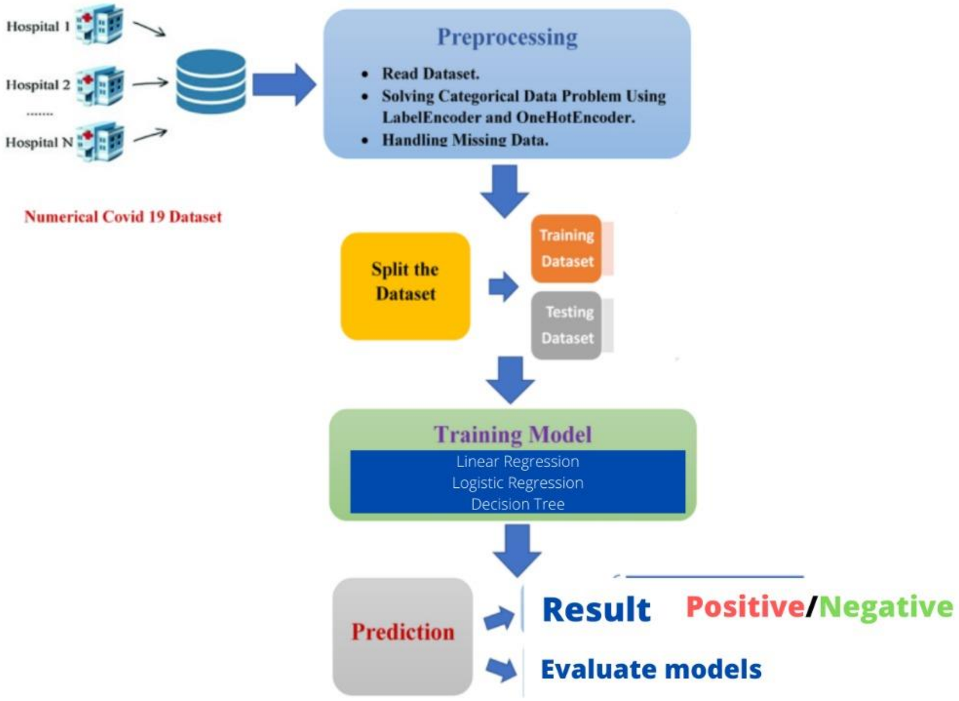

Figure 4.2: This diagram represents the flow of data within the application

The above diagram demonstrates how data within the application moves from the internal components of the application to the external interfaces with which the application interacts with and where the user stands in this flow of operations.

4.3 Use Case Diagram

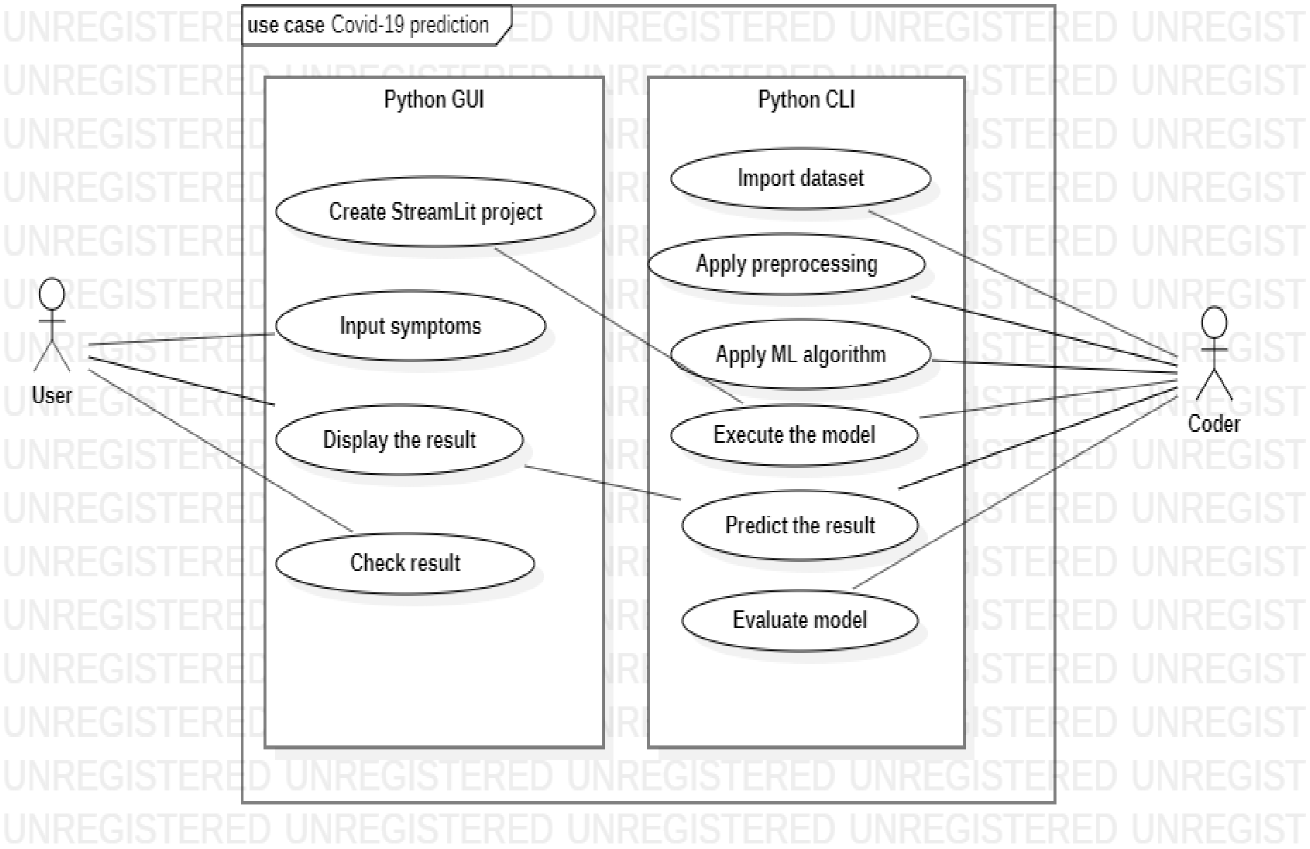

Figure 4.3: This diagram demonstrates all of the functionalities that are available to a user of the application

REFERENCES

[1] Ahmad, A., Garhwal, S., Ray, S.K., Kumar, G., Malebary, S.J. and Barukab, O.M., 2020. The number of confirmed cases of covid-19 by using machine learning: Methods and challenges. *Archives of Computational Methods in Engineering*, pp.1-9.

[2] Santosh, K.C., 2020. COVID-19 prediction models and unexploited data. *Journal of medical systems*, *44*(9), pp.1-4.

[3] Rustam, F., Reshi, A.A., Mehmood, A., Ullah, S., On, B.W., Aslam, W. and Choi, G.S., 2020. COVID-19 future forecasting using supervised machine learning models. *IEEE access*, *8*, pp.101489-101499.

[4] Khakharia, A., Shah, V., Jain, S., Shah, J., Tiwari, A., Daphal, P., Warang, M. and Mehendale, N., 2021. Outbreak prediction of COVID-19 for dense and populated countries using machine learning. *Annals of Data Science*, *8*(1), pp.1-19.

[5] Rahimi, I., Chen, F. and Gandomi, A.H., 2021. A review on COVID-19 forecasting models. *Neural Computing and Applications*, pp.1-11.

[6] Darapaneni, N., Singh, A., Paduri, A., Ranjith, A., Kumar, A., Dixit, D. and Khan, S., 2020, November. A Machine Learning Approach to Predicting Covid-19 Cases Amongst Suspected Cases and Their Category of Admission. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)* (pp. 375-380). IEEE.

[7] Bhadana, V., Jalal, A.S. and Pathak, P., 2020, December. A Comparative Study of Machine Learning Models for COVID-19 prediction in India. In *2020 IEEE 4th Conference on Information & Communication Technology (CICT)* (pp. 1-7). IEEE.

[8] Mandayam, A.U., Rakshith, A.C., Siddesha, S. and Niranjan, S.K., 2020, November. Prediction of Covid-19 pandemic based on Regression. In *2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)* (pp. 1-5). IEEE.

[9] Zoabi, Y., Deri-Rozov, S. and Shomron, N., 2021. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj digital medicine, 4(1), pp.1-5.

[10] Yan, L., Zhang, H.T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jin, L., Zhang, M. and Huang, X., 2020. A machine learning-based model for survival prediction in patients with severe COVID-19 infection. MedRxiv.

[11] Ogundokun, R.O. and Awotunde, J.B., 2020. Machine learning prediction for covid 19 pandemic in india. medRxiv.

[12] Ardabili, S.F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A.R., Reuter, U., Rabczuk, T. and Atkinson, P.M., 2020. Covid-19 outbreak prediction with machine learning. Algorithms, 13(10), p.249.

[13] Real-time tracking of self-reported symptoms to predict potential COVID-19 Cristina Menni 1,7 ✉, Ana M. Valdes.

[14] Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches Zohair Malki,a El-Sayed

Atlam,∗,a,b Aboul Ella Hassanien,c Guesh Dagnew,d Mostafa A. Elhosseini,a,e and Ibrahim Gadb

[15] Machine Learning Approach for Confirmation of COVID-19 Cases: Positive, Negative, Death and Release Shawni Dutta1 and Dr. Samir Kumar Bandyopadhyay2