International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

Covid-19 Future Forecasting Using Supervised Machine Learning Models

Mukesh S Rao Dadge[1], Karthik Kumar[2], Jitendra Kumar[3], Deepanshu[4]
Assistant Professor[1], UG Students[2,3,4]
Department of Mechanical, Delhi Technological University, Main Bawana Road Rohini, Delhi, India

ARTICLEINFO

Article History:

Accepted: 01 May 2023
Published: 20 May 2023

Publication Issue
Volume 9, Issue 3
May-June-2023

Page Number
149-156

ABSTRACT

In order to enhance decision-making on the future course of action, machine learning (ML) based forecasting methods have demonstrated its relevance to foresee in unexpected outcomes. Many application fields that required the detection and prioritization of negative aspects for a threat have long employed ML models. To deal with forecasting issues, a variety of prediction techniques are frequently utilized. This work shows how ML models can predict the amount of forthcoming COVID-19 patients who will be afflicted, which is now thought to pose a threat to humanity. Our suggested technique combines a number of approaches in an effort to improve the explore operation's cooperativeness. We create the Covid-19 application in this effort can be able to predict outcomes from Total Confirmed cases, Fatalities and Recoveries.

KEYWORDS: Covid-19, Total Confirmed Cases, Fatalities, Recoveries.

I. INTRODUCTION

By resolving a large number of intricate and challenging real-world situations during the previous ten years. Almost all of the real-world domains were covered by the application fields, which included healthcare, AV (autonomous vehicles), business applications, NLP (natural language processing), intelligent robotics, gaming, climate modelling, speech, and image processing. In contrast to traditional algorithms, which execute programming instructions based on conditional statements like if-else, machine learning (ML) algorithms learn most frequently through the process of trial and error. One of the most important uses of ML is forecasting; several common ML algorithms have been applied in this field to direct the future course of action required in a variety of application domains, such as weather forecasting, illness forecasting, stock market forecasting, and disease prediction. In order to forecast the future health of patients with a certain ailment, a variety of regression and neural network models are widely applicable. Numerous research employing machine learning approaches have been conducted to forecast various illnesses, including coronary artery disease and cardiovascular disease. The project is focused on anticipating verified COVID-19 cases in real time as well as the COVID-19 epidemic and early reaction. These forecasting tools can be a great decision-making tool for handling the current situation and directing early treatments to effectively manage these disorders. In this study, the World Health Organization's COVID-19 coronavirus, better known as SARS-CoV-2, is predicted to spread early using a model. This project targets to develop web application in order to handle the “Forecasting”. This software helps to give the particular response of the SARS-CoV-2. Python-Flask is used as front end which is used to craft the user interface. MySQL is used as back end and used to craft the database and save the particulars. Anybody with a little computer knowledge can approach and deal with the software with ease; hence it can be termed user friendly.

II. RELATED WORK

[1] The definition of reliable and effective techniques that can derive from assessments the stochastic dependency among past and future is necessary due to the growing availability of enormous quantities of historical data and the requirement to perform reliable projections of future behaviour in several academic and applied domains. Since the 1960s, linear statistical techniques have had an impact on the forecasting field. In the forecasting community, machine learning models have gained popularity and have solidified their position as significant rivals to traditional statistical models. In our project we are going to use algorithms to reduce these difficulties of generating timetable. These algorithms incorporate a numeral of strategy, aimed to improve the operativeness of the search operation. The system will take various inputs like number of cases, deaths, recoveries etc.., By relying on these inputs, it will generate possible outcomes for future forecasting. [2] In forecasting the course of patients with a range of illnesses, multiple regression models are widely applicable. However, many scientists are employing these models without verifying the essential suppositions. Far too often, researchers also "overfit" the data by creating models with insufficient sample sizes and too many predictor variables. It is improbable that models created in this manner will pass the validation test on a different patient sample. Without performing such a validation, overfitting is still unknown to the researcher. There are data reduction techniques that can significantly enhance the performance of regression models when the ratio of patients experiencing endpoints to possible predictors is low (let's say fewer than 10). When model assumptions are carefully examined, actions are taken (such as selecting another model or developing the data) when beliefs are violated, and the model formulation process does not lead to overfitting the data, regression models can produce predictions that are more accurate than other techniques such as differentiation and recursive partitioning. [3] The manual system of predicting future forecasting with large number of cases is very time consuming and usually ends up with various classes clashing either… To overcome all these problems, propose to make a web application system. The system will take various inputs like details of cases, recoveries and deaths, depending upon these inputs it will generate a possible response, making optimal predictions of all resources in a way that will best suit any of constraints. [4] Myocardial infarct, cardiovascular coronary disease (CHD), dying due CHD, cerebrovascular accident, heart failure, and death from heart disease prediction equations were created. The calculations suggested that it may be more important to manage several risk variables (including hypertension, overall cholesterol, lipoprotein cholesterol, cigarettes, hyperglycemia, and left heart hypertrophy) than to concentrate only on one risk factor. The employed parametric model was shown to provide a number of benefits over the current conventional regression models. Like logistical regression, it may provide predictions for many time periods, and compared to the proportional hazards Cox model, probabilities can be presented in a simpler way. This application considers a solution to the future forecasting problem. The forecasting problem involves scheduling a trend, each consisting cases of date, the paper introduces the forecasting problem, and then describes the simulated methods. A parallel algorithm

which can be implemented on a multiprocessor is presented. This algorithm can provide a faster solution than the equivalent sequential algorithm. Some further experimental results are given.

III.PROPOSED SYSTEM

We propose this application that can be considered a useful system since it helps to reduce the limitations obtained from KNN and Logistic Regression. By providing support through the regression analysis, it can be able to generate best results for attributes without any overlap. The system is developed in a Flask based Python environment. MySQL is used for database management and the models involved in this application are Lasso, Linear Regression and Support Vectors model.

IV.BLOCK DIAGRAM

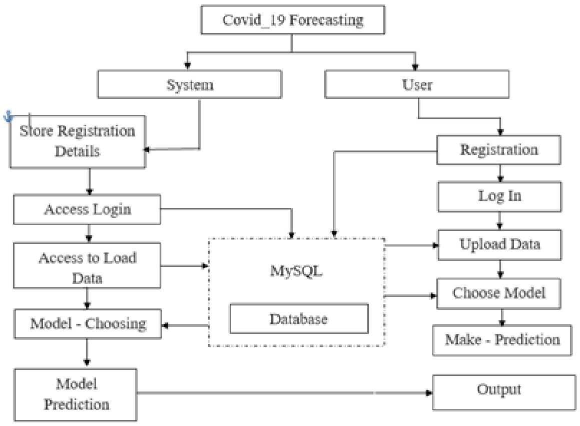


fig. Block diagram of proposed method

Steps:

Registration: The user can Register to the system in which the registered details are stored in Data base.

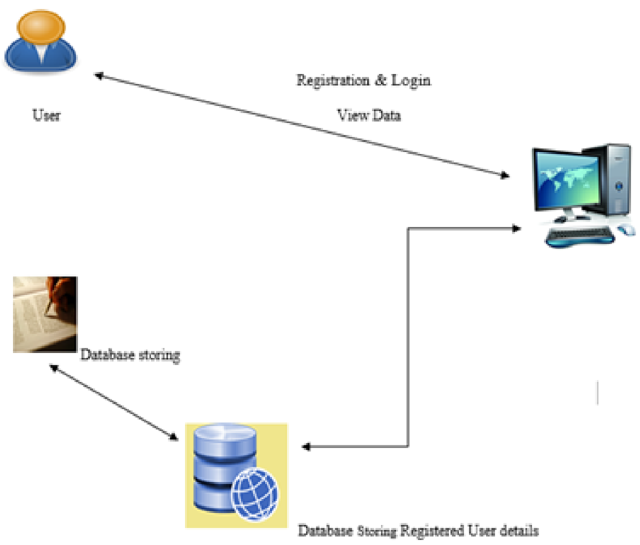
Login: The System allows registered users to get in.

Receive Data: The System receives data given by the user.

Model: The user can select model based on the variables to get response from the data and can predict the outputs from the model.

Predictions: The system can deliver the predicted results and can be displayed to the user.

V. ARCHITECTURE



VI.METHODOLOGY AND ALGORITHM

K-Nearest Neighbor

The algorithms of KNN- K-Nearest Neighbors is one of the Machine Learning (ML) algorithm that is mostly used due to its easier deployment and simplicity. This KNN is much useful for the sake of sorting out any kind of regression and classification challenges for serving any applications [11].

Algorithmic Procedure

The algorithmic procedure [11] of KNN has been given in the following pointers:

- The data needed to be first loaded
- Neighbors count is selected by the initialization of K
- Estimating the displacement intermediary to the current and query examples from the information and include that estimated with the example index to a well sequenced assemblage.
- Arrange the indices along with the sequenced assemblage of displacements till the greatest starting from the lowest.

- Select the initial entities of K with the help of arranged assemblage.
- Acquire the labels of every chosen entity of K
- If the application context is classification, the mode of the labels corresponding to K is returned.
- The mean of the labels pertaining to K is returned if the application context is regression.

Choosing the right value for K

For easing the selection of K which is very apt for the considered information, one needs to execute the algorithm of K-Nearest Neighbors for many iterations by varying with unique K values and one needs to select the K value which is able to lessen the flaw counts during the operation of prediction using that algorithm [11].

Benefits of using these KNN algorithms

According to [11], by using KNN algorithms, we are able to draw the following benefits:

- It's easier to deploy and has much simplicity.
- There is nil necessity for constructing a prototype, tuning of numerous of variables or making supplementary presumptions.
- Since KNN is much useful for sorting out any kind of regression and classification challenges for serving any applications, it is much flexible.

VII. LOGISTIC REGRESSION

An established technique for calculating adjusted odds ratios is logistic regression. ML (i.e., Maximum Likelihood) software is usually always used to fit logistic models [24]. If the model is roughly accurate and the sample size is sufficient (e.g., a minimum of 4-5 variables are present in single parameter available in every single level of the result) [25-27], it offers meaningful statistical judgments. However, ML estimation can fail in the presence of small or sparse data sets, an unusual exposure or outcome, or significant underlying effects, particularly when these issues coexist [28-30]. Because of this, ML estimates of finite odds ratios may be infinite and ML estimators are not even roughly unbiased. The covariates are responsible for separating the outcomes and this might result in the emerging of infinite estimates [31, 32].

The relation among many independent variables that are either continuous or categorical and a dichotomous dependent variable is modelled using binary logistic regression [33]. Binary logistic regression has various guesses that must be addressed in order to get a reliable outcome [34].

- Linearity: The connection between the explanatory factors and the response variable's logit should be linear.
- Independent mistakes: Correlation between the mistakes is inappropriate.
- Explanatory variables shouldn't have a lot of correlation with one another to avoid multicollinearity.
- There shouldn't be any anomalies, values with significant leverage, or points with a lot of influence.

Explanatory variables shouldn't have a strong correlation with one another, according to one of the premises of logistic regression. For the findings to be considered legitimate, the assumptions of the logistic regression model must be fulfilled. Invalid statistical conclusions may result from issues with the model, like excessively inflated standard errors, inaccurately low or high t-statistics, and parameter estimates with nonsensical signs [35]. While it may be feasible to construct experiments where the explanatory factors are orthogonal to one another, observational data does not allow for this. Then, according to another study paper [36] , "collinearity is a basic rule in the data set originating from the uncontrolled processes of the data spawning system and is simply a harsh and inevitable life event" the fields of non-experimental sciences. Reviews commonly collect correlated variables for analysis. [37] Firth's penalized likelihood equation and the concept of a double penalty maximum likelihood estimator were coupled in order to provide stability the estimates under circumstances of multicollinearity. [38] A new method was proposed for computing the diminution coefficients of the Liu-type logistic estimator. A ridge type approximation was also reported in [39], which in some cases has lower cumulative average squared errors than the highest likelihood estimator.

VIII. SUPPORT VECTOR MACHINES

Support Vector Machine is widely regarded to SVM, which is nothing but acronym of it. SVM is one among the mostly known supervised kind of learning methodologies which is much utilized for the sake of sorting out the issues related to regression/ classification. Nonetheless, when it comes to ML, SVM is widely found deployed for sorting out any classification issues from perspectives of different applications [13]. The main aim of these approaches is to discover a hyper place in the space of N-dimension that is able to categorize the particulars of data in an unique way [14].

From general sense, the SVMs are of two types. The two types are as follows: Non-Linear SVMs and Linear SVMs.

SVM Kernel

For easing the data manipulation operations in SVMs, SVM kernels are utilized. It is nothing but a function which considers the space of the lower dimensional input and converts it into a maximum space of dimensional input. In other words, it could be said that it transforms the non-segregable issue into segregable issue [15]. This SVM kernel is found much beneficial in the non-linear segregation issues. The main purpose of using this SVM kernel is to carry out a few sophisticated/ difficult data conversions and subsequently discover the operation by which the data segregation is done depending upon the defined outcomes or labels [15].

Applications of support vector machine

According to [16], the application of SVMs are discussed in the below pointers.

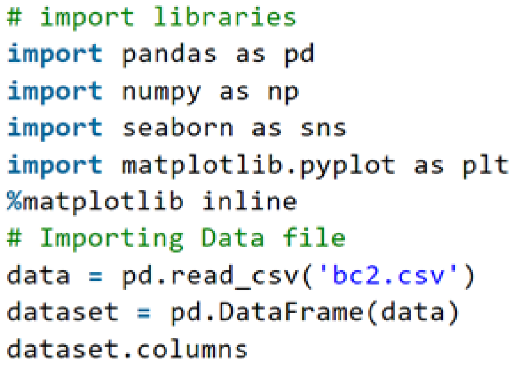
- **Hypertext as well as Text organization** - As the classification methodology, it is deployed to either discover significant data or one could tell the needed data to organize the text contents.
- **Face examination** - It is deployed to identify the face of the human beings in accordance with the model as well as the devised classifier.
- **Bio-informatic context –** SVMs are utilized for the context of medical field and also in laboratories and heath care facilities. For instance, SVMs are found utilized for the locating and classifying purposes when the sequences of amino acids are known.
- **Grouping** - SVMs are found deployed for several grouping needs. For instance, a decision can be arrived just by comparing any data to facilitate grouping to certain or needed class.
- **Recall of Hand writing** - SVMs are utilized for identifying the hand writing of any human beings by comparing the input and samples.

Example of Python-based implementation for Support Vector Machines

Primary Aim for the implementation: By making use of historical information pertaining to a diseased person who have been known to have cancer, the health care professionals are easily empowered to distinguish between the benign as well as the malignant cancer formations provided that the individualistic features are known [15].

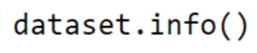
Considered Dataset: [17]

Python Implementation Screenshots: Few stages in a typical Python Implementation have been shown in the below screenshots found in below fig. 1 (a), (b), and (c).



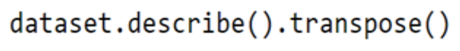
```
# import libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
# Importing Data file
data = pd.read_csv('bc2.csv')
dataset = pd.DataFrame(data)
dataset.columns
```

(a)



```
dataset.info()
```

(b)

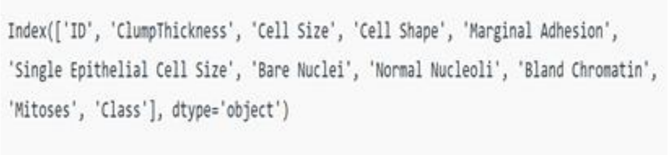


```
dataset.describe().transpose()
```

(c)

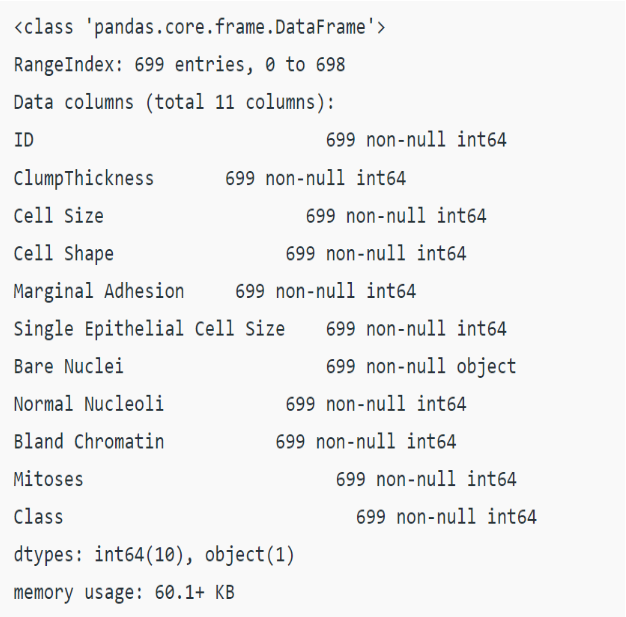
Fig. 1 Python Implementation Screenshot

Few outcomes of a typical Python Implementation have been shown in the below screenshots found in below fig. 2 (a), (b), and (c).



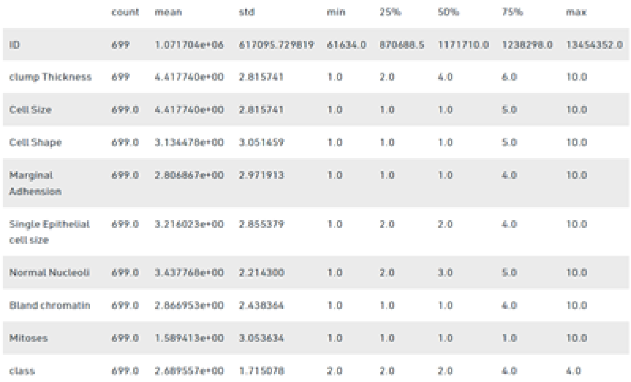
```
Index(['ID', 'ClumpThickness', 'Cell Size', 'Cell Shape', 'Marginal Adhesion',
'Single Epithelial Cell Size', 'Bare Nuclei', 'Normal Nucleoli', 'Bland Chromatin',
'Mitoses', 'Class'], dtype='object')
```

(a)



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 699 entries, 0 to 698
Data columns (total 11 columns):
ID                              699 non-null int64
ClumpThickness        699 non-null int64
Cell Size                      699 non-null int64
Cell Shape                   699 non-null int64
Marginal Adhesion     699 non-null int64
Single Epithelial Cell Size    699 non-null int64
Bare Nuclei                    699 non-null object
Normal Nucleoli             699 non-null int64
Bland Chromatin            699 non-null int64
Mitoses                         699 non-null int64
Class                             699 non-null int64
dtypes: int64(10), object(1)
memory usage: 60.1+ KB
```

(b)



	count	mean	std	min	25%	50%	75%	max
ID	699	1.071704e+06	617095.729819	61634.0	870688.5	1171710.0	1238298.0	13454352.0
clump Thickness	699	4.417740e+00	2.815741	1.0	2.0	4.0	6.0	10.0
Cell Size	699.0	4.417740e+00	2.815741	1.0	1.0	1.0	5.0	10.0
Cell Shape	699.0	3.134478e+00	3.051459	1.0	1.0	1.0	5.0	10.0
Marginal Adhension	699.0	2.806867e+00	2.971913	1.0	1.0	1.0	4.0	10.0
Single Epithelial cell size	699.0	3.216023e+00	2.855379	1.0	2.0	2.0	4.0	10.0
Normal Nucleoli	699.0	3.437768e+00	2.214300	1.0	2.0	3.0	5.0	10.0
Bland chromatin	699.0	2.866953e+00	2.438364	1.0	1.0	1.0	4.0	10.0
Mitoses	699.0	1.589413e+00	3.053634	1.0	1.0	1.0	1.0	10.0
class	699.0	2.689557e+00	1.715078	2.0	2.0	2.0	4.0	4.0

(c)

Fig. 2 Python Implementation Output Screenshot

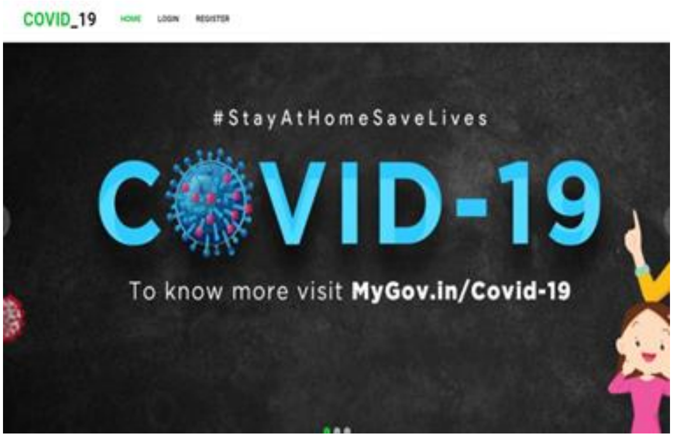
Benefits of SVM

The benefits [18] of SVMs have been discussed below:

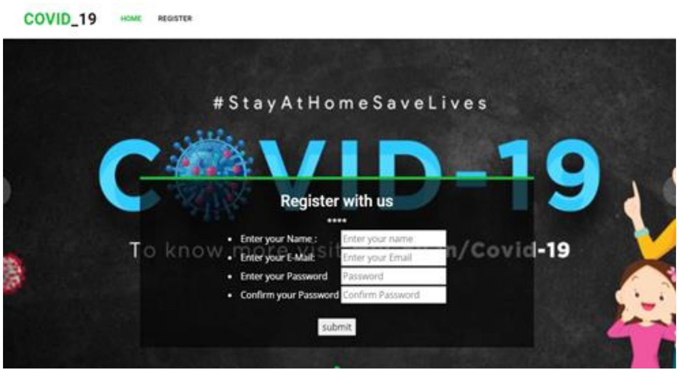
- SVMs are always efficient when it comes to spaces of maximum dimensions.
- SVMs are also effective in the aspect of its memory as it utilizes a subset of training entities in the function deployed for decision.
- SVMs are able to perform well even when the magnitude of dimensions is larger when compared to the samples count.
- SVMs provides the users with the flexibility to make use of several general as well as the unique kernel functions depending upon the needs and application context.

IX. RESULTS AND DISCUSSION

HOME PAGE:



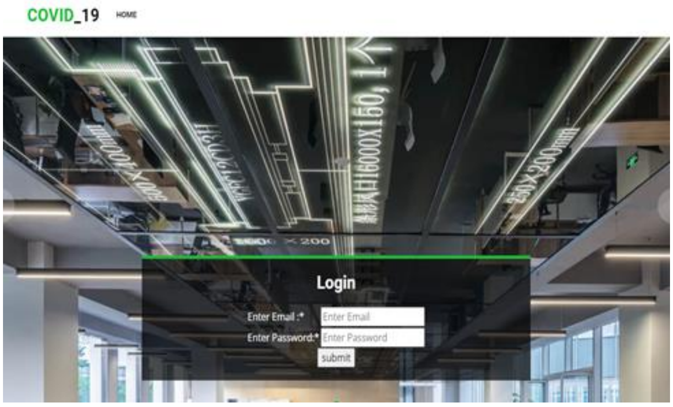
REGISTRATION PAGE:



Register with us

- Enter your Name : Enter your name
- Enter your E-Mail: Enter your Email
- Enter your Password Password
- Confirm your Password Confirm Password

LOGIN:

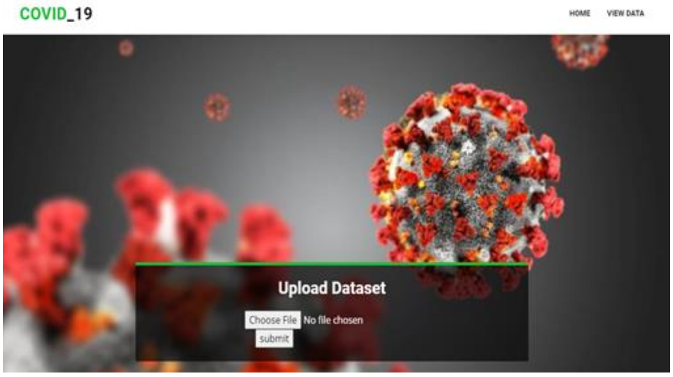


Login

Enter Email :* Enter Email

Enter Password:* Enter Password

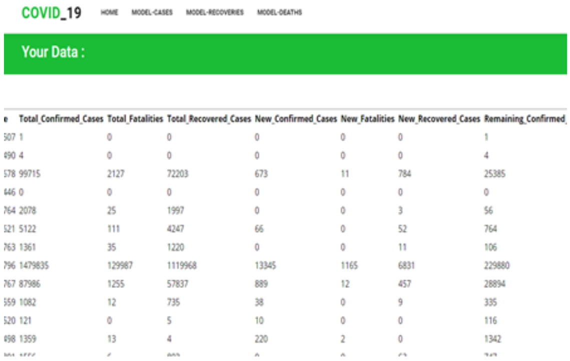
UPLOAD PAGE:



Upload Dataset

No file chosen

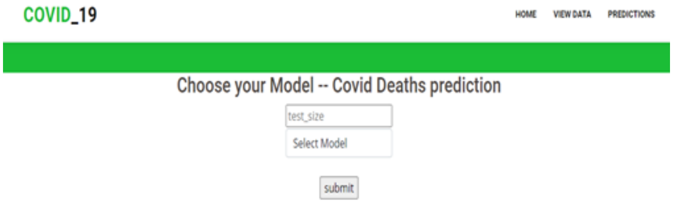
VIEW DATA:



Your Data :

e	Total_Confirmed_Cases	Total_Fatalities	Total_Recovered_Cases	New_Confirmed_Cases	New_Fatalities	New_Recovered_Cases	Remaining_Confirmed_
507	1	0	0	0	0	0	1
490	4	0	0	0	0	0	4
578	99715	2127	72203	673	11	784	25385
446	0	0	0	0	0	0	0
764	2078	25	1997	0	0	3	56
521	5122	111	4247	66	0	52	764
763	1361	35	1220	0	0	11	106
796	1479835	129987	1119968	13345	1165	6831	229880
767	87986	1255	57837	889	12	457	28894
559	1082	12	735	38	0	9	335
520	121	0	5	10	0	0	116
498	1359	13	4	220	2	0	1342

MODEL SELECTION PAGE:

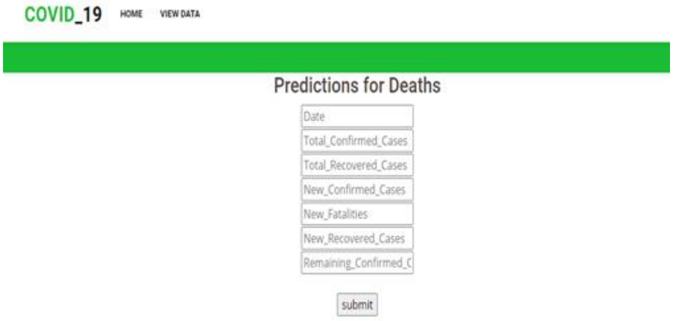


Choose your Model -- Covid Deaths prediction

test_size

Select Model

MODEL PREDICTION PAGE:



Predictions for Deaths

Date

Total_Confirmed_Cases

Total_Recovered_Cases

New_Confirmed_Cases

New_Fatalities

New_Recovered_Cases

Remaining_Confirmed_C

X. CONCLUSION

Our method to forecast the (SARS COV-2) Severely Acute Respiratory Syndrome Coronavirus 2 has been successfully developed for this purpose. Utilizing a setting that is user-friendly Flask via Python programming. The system is likely to collect information from the user to predict the requirements. The user, on other hand who is registered can allowed to view his/her uploaded dataset and can choose model to get response predictions of Covid-19 from the data. Multiple illnesses may be predicted with this program, which can be expanded. We intend to apply the most precise and appropriate ML algorithms for forecasting as we investigate the prediction process utilizing the new dataset. We'll be concentrating a lot of our future effort on real-time live forecasting.

XI. REFERENCES

[1]. S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “Statistical and machine learning forecasting methods: Concerns and ways forward,” PloS one, vol. 13, no. 3, 2018.

[2]. F. E. Harrell Jr, K. L. Lee, D. B. Matchar, and T. A. Reichert, “Regression models for prognostic prediction: advantages, problems, and suggested solutions.” treatment reports, vol. 69, no. 10, pp. 1071–1077, 1985.

[3]. P. Lapuerta, S. P. Azen, and L. LaBree, "Use of neural networks in predicting the risk of coronary artery disease," 1995, pp. 38–52, Machines and Health Analysis, vol. 28, no. 1.

[4]. K. M. Anderson, P. M. Odell, P. W. Wilson, and W. B. Kannel, "Cardiovascular disease risk profiles," Vol. 121, No. 1, pp. 293-298, American heart journal, 1991.

[5]. T. Noel, H. Al Moatassime, H. Asri, H. Mousannif, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," Procedia Computer Science, vol. 83, pp. 1064–1069, 2016.

[6]. F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus covid-19," Plos one, vol. 15, no. 3, p. e0231236, 2020.

[7]. WHO. Naming the coronavirus disease (covid-19) and the virus that causesit. [Online]. Available: https://www.who.int/emergencies/diseases/novel coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it

Cite This Article :

Mukesh S Rao Dadge, Karthik Kumar, Jitendra Kumar, Deepanshu, "Covid-19 Future Forecasting Using Supervised Machine Learning Models", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 3, pp.149-156, May-June-2023.