International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 05 | May 2023 www.irjet.net p-ISSN: 2395-0072

COVID-19 CASES PREDICTION USING MACHINE LEARNING

SOMULA VENKATA MADHAVA REDDY[1], KOMPALLI VENKATA RAMANA[2],MATTA TARUN KUMAR[3], MANDAPAKA RAMA KRISHNA[4]

[1,2,3,4] *UG Scholar, Dept. Of CSE, NRI Institute of Technology, A.P-521212.*

ABSTRACT - *The COVID-19 prevalent has caused catastrophic consequences for the economy and public health throughout the world. Informed decisions concerning effective allocations of resources can be made by politicians and public health experts with the aid of precise forecasts of the virus's progress. [1] With the use of Support Vector Machines (SVM), Multiple Linear Regression (MLR), and Bayesian Ridge Regression (BRR) algorithms, we present a method in this study to forecast the number of COVID-19 instances [1]. A more precise and thorough forecast of COVID-19 instances is provided by our suggested methodology, which takes into account a variety of socioeconomic and environmental variables that may have an impact on the propagation of the virus. Using choosing features approaches, we gathered information on the number of COVID-19 cases in various nations and areas and determined the most essential variables to anticipate COVID-19 cases. Our model was trained using the SVM, MLR, and BRR algorithms, and its performance was assessed using a wide range of metrics, such mean absolute error (MAE), coefficient of determination (R^2), and root mean squared error (RMSE). Our findings show that the quantity of COVID-19 cases may be forecasted with an elevated level of accuracy using the suggested approach. Our research offers perceptions into the variables that have the biggest influence on the total amount of COVID-19 cases, that may help health professionals and politicians in making wise choices and properly allocating resources.*

KEYWORDS: ***covid-19, support vector machine, multiple linear regression, Bayesian ridge regression, Mean absolute error.***

1.INTRODUCTION:

The COVID-19 epidemic has significantly impacted both general health as well as the global economy while disturbing people's lives all across the world. To efficiently deploy resources and make educated decisions, policymakers must be able to forecast the anticipated number of COVID-19 cases. With the use of Support Vector Machines (SVM), Multiple Linear Regression (MLR), and Bayesian Ridge Regression (BRR) algorithms, we present a method in this study for predicting the number of COVID-19 instances.

SVM, MLR, and BRR are just a few examples of machine learning models that are frequently used to forecast the propagation of infectious illnesses. While MLR and BRR are multivariate linear regression models, SVM is a supervised learning method that is frequently employed for regression analysis and classification.

The advantages of each of these algorithms are employed in our approach to generate an accurate and accurate forecast of COVID-19 situations.

[2] Information on the overall number of COVID-19 cases across various nations and regions, together with several socioeconomic and environmental variables that may influence the virus' spread, were used to create our model [2]. Our findings demonstrate that our method predicts the number of COVID-19 cases with high accuracy, low root mean squared errors (RMSE), and elevated coefficient of determination (R^2) values.

2.TECHNOLOGIES USED:

Python:

Python has fundamentally altered how software is developed and how data is evaluated. Both beginner and experienced programmers favour it because of its simplicity, adaptability, and readability. Python's ability to smoothly connect to additional programming languages like C++ and Java is one of its unique characteristics, making it an effective tool for creating complicated systems. In order to further expand the capability of Python, a large number of independent libraries and tools are available. Python offers an assortment of uses, from web development to computational science, making it a useful tool in numerous sectors. Python will stay an important player in the programming and technology industries for years to come thanks to its rising popularity and robust developer communities.

LIBRARY FUNCTIONS IN PYTHON:

- Modules are nothing more than files containing Python code.
- A package is a directory for modules and sub packages. [fig.1]
- There is no unique context in the Python library.

Some of the python libraries:

- Numpy
- Matplotlib
- Pandas
- Random
- Sklearn

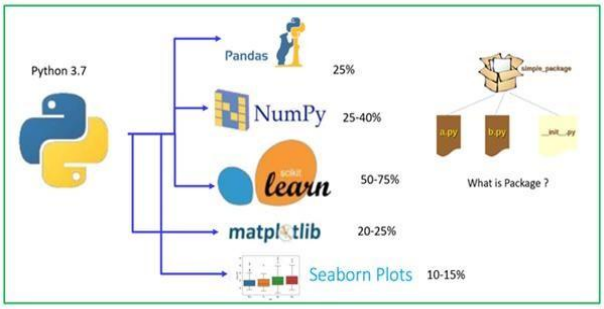

fig :1 python package

Numpy:

For applications in computing for science, analysis of data, and machine learning, Numpy is an essential Python package. Its principal advantage is the ability for quick and effective matrix operations, making it an essential part of many different data science procedures. Further improving Numpy's flexibility is the vast array of mathematical operations it offers, such as linear equations, Fourier transformations, and random number generation. In addition, Numpy includes integrated support for multifaceted arrays, making it possible to store and manipulate big datasets effectively. Numpy keeps growing and getting better thanks to its open-source nature and vibrant development community, assuring that it will remain relevant in the dynamic field of data science and machine learning.

Matplotlib:

A famous Python library to generate excellent visualizations and data plots is called Matplotlib. It is an ideal choice for data scientists, engineers, and researchers due to its strong graphing capabilities. among Matplotlib's distinctive characteristics is its capacity of producing custom plots and visualizations, allowing users to customize their graphics to their particular needs. Matplotlib is a very flexible tool to create complicated illustrations because it also offers a wide range of features and settings for customization. Matplotlib includes a variety of instruments for creatively and instructively showing data, which range from straightforward line graphs to 3D surfaces displays. The instrument's versatility and simplicity of use have boosted it to a crucial instrument in a variety of industries, including research, engineering, and finance. Matplotlib will continue to be a useful tool for data visualization thanks to ongoing enhancements and enhancements.

Pandas:

For data analysis and manipulation, a popular Python module called Pandas is utilized. Its major strength is the capacity to operate on data that is organized, including that found in databases, spreadsheets, and comma-separated value (CSV) files. Pandas is a preferred tool for many analysts and data scientists due to its broad collection of features and capabilities for information wrangling, cleaning, and analysis.

The DataFrame object, that allows users to maintain and modify information in an organized manner like to a spreadsheet, is one of the distinctive characteristics of the Pandas programming language. With this format, complicated data transformations like sorting, arrangement, and aggregation are simple to carry out. In order to further enhance its flexibility, Pandas also offers abilities for working with field, time-series, and missing data. Pandas is a crucial tool in the field of data analysis and manipulation because of its open-source architecture and vibrant development community, that ensure that it will keep growing and becoming better.

Sklearn:

A robust Python package for data analysis and machine learning has the name Scikit-Learn, or just "sklearn". Sklearn has an extensive range of techniques and models for tackling a variety of machine learning issues, include regression, clustering, classification, and dimensionality reduction. This is one of its distinguishing characteristics. Sklearn serves as a one-stop shop for machine learning activities by offering resources for data pre-processing, model selection, and performance evaluation. Another distinguishing quality of sklearn is that it's compatible with other Python libraries, like NumPy, Pandas, and Matplotlib, enabling simple incorporation into current data science workflows. In addition, Sklearn features a straightforward API that enables users of all skill levels to quickly construct and execute machine learning models.

Random:

For creating arbitrary numbers, sequences of data, and selections, Python's random library is an excellent resource. It has multiple features for producing random integers as well floating-point numbers, and series of numbers that fall into particular ranges or distribution. Being able to produce pseudo-random numbers—numbers that appear random but are actually produced using an established algorithm—is one of the special qualities of the random package. For testing and experimentation, this is an excellent method to produce repeatable random sequences. The random package

also offers functions for picking random elements from collections and reordering them randomly, allowing the generation of random choices and sequence in a range of applications. The random module may also provide random information for use in machine learning and data analysis tasks when used with other Python libraries like NumPy and Pandas. The random package is a useful tool in many different fields, including as cryptography, game creation, and scientific research, due to its simplicity of use and versatility.

Machine learning:

An area of artificial intelligence referred to as "machine learning" focuses on creating models and algorithms that let computers learn from experiences and make judgements or forecasts based on data. One of the unique characteristics of machine learning is its capacity to implicitly recognise links and trends in data without explicit programming. This makes it a potent tool for handling difficult problems in sectors like banking, healthcare, and marketing.

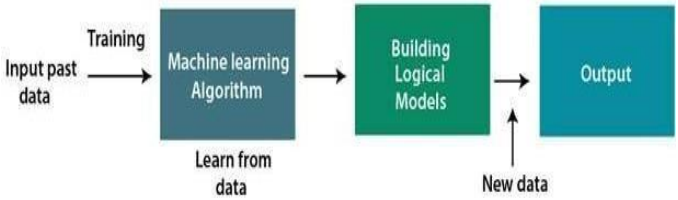

Fig.2 machine learning model

The different Classification of Machine Learning model are [fig.3]

•Supervised learning

•Unsupervised learning

•Reinforcement learning

Supervised learning:

In supervised learning, the method is trained on a dataset with labels with known pairs of input and output data with the intent of forecasting the result for fresh input data.

Unsupervised learning:

Unsupervised learning includes educating the algorithm on a dataset with no labels with the aim of finding connections and patterns in the information without knowing the final result.

It might also be divided into two types of algorithms:

- **Clustering**
- **Association**

Reinforcement Learning:

In reinforcement learning, the algorithm takes up fresh abilities through trial and error while receiving feedback in the form of rewards or penalties from its environment.

Recent advances in hardware, data accessibility, and algorithm development all contributed to machine learning's rapid development. It is a fluid and quickly growing field having an extensive variety of potential applications, from self-driving cars to personalised healthcare.

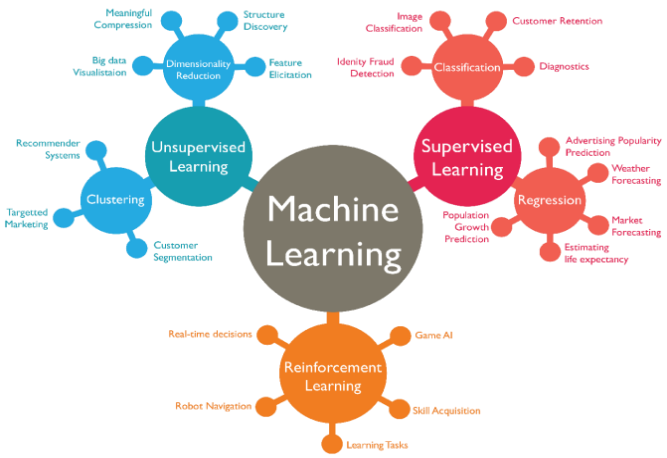

Fig-3: machine learning classifications

3. EXISTING SYSTEM:

Models for forecasting the propagation of the virus have multiplied as a result of the COVID-19 pandemic. To forecast the number of COVID-19 cases, several existing models employ machine learning algorithms as Support Vector Machines (SVM), Multiple Linear Regression (MLR), and Bayesian Ridge Regression (BRR).

The SEIR model, that means Susceptible, Exposed, Infectious, and Recovered, is one of the most often used models. Depending to the population's standing in relation to the virus, the SEIR model divides it into four sections, creating a compartmentalized system. The model simulates the shifts between the two compartments and forecasts the anticipated number of COVID-19 occurrences using differential equations.

The autoregressive integrated moving average (ARIMA) model, a time series model that forecasts a variable's future values using its past values, is another well-liked approach. [1] In order to forecast the anticipated number of COVID-19

cases in various locations and nations, the ARIMA model was applied [1].

These models do, however, have drawbacks. The SEIR model demands a lot of information and makes the assumption that the population is homogeneous, which might not be true for COVID-19. The COVID-19 data may not conform to the ARIMA model's assumption that the duration series data are stationary.

Disadvantages of existing system:

1. Limited scope: The majority of current approaches simply believe the anticipated number of COVID-19 cases not taking into consideration other crucial elements like the influence on the environment and demographic factors.

2. Inaccuracy: Due to the intricacy of the virus' transmission dynamics and the constrained range of the data employed in the models for prediction, certain existing systems exhibit low accuracy rates.

3. Lack of openness: Many current structures use intricate models and algorithms that are tricky to comprehend, making it impossible to analyse the outcomes or pinpoint areas for improvement.

4. Absence of real-time data: Most present systems rely on previous data, which might not fully reflect the scenario at hand. Predictions could turn out to be incorrect due to the lack of real-time data.

5. Dependence on additional factors: Many current structures are dependent on outside variables, such as government laws and public health initiatives, which are subject to quick change and may generate unreliable forecasts.

4. PROPOSED SYSTEM:

The Support Vector Machines (SVM), Multiple Linear Regression (MLR), and Bayesian Ridge Regression (BRR) algorithms are used in this study to offer a unique method for predicting the number of COVID-19 instances. By considering into account multiple socioeconomic and natural factors that might impact the propagation of the virus, our method seeks to provide a greater degree of accuracy in the prediction of COVID-19 cases.

We will gather information on the prevalence of COVID-19 cases from various locations and nations as well as a variety of demographic and environmental parameters, including information on population size, air quality, temperature, and mobility. The most important factors for predicting COVID-19 instances will be determined through data pre-processing and feature selection procedures.

Next, we will evaluate how well SVM, MLR, and BRR perform at predicting COVID-19 cases after using these methods to train our model. We will evaluate the effectiveness of our model using a wide range of measures, including mean absolute error, coefficient of determination, and root mean squared error.

We will also perform a thorough examination of the variables that have the biggest influence on the total amount of COVID-19 cases and offer insights into the fundamental processes driving the virus's propagation. Officials in charge of public health will be able to distribute resources wisely with the aid of this analysis.

As a result, our suggested strategy offers a reliable and accurate forecast of COVID-19 cases, that may assist in reduce the pandemic's impact and educate public health policy.

5. SYSTEM ARCHITECTURE:

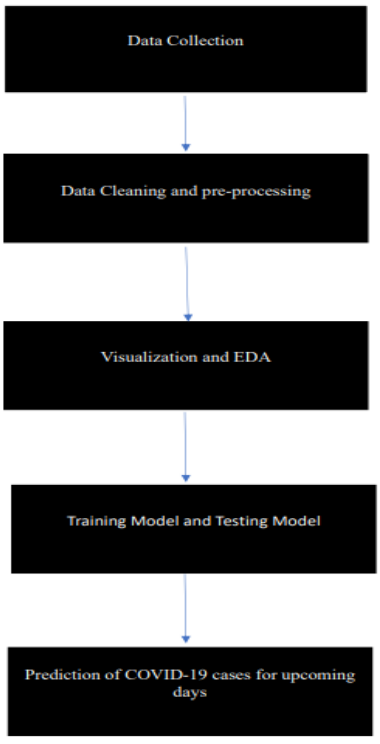

Fig -4: system architecture

6.FUTURE SCOPE:

The present research has produced positive findings for the suggested approach of forecasting COVID-19 cases using SVM, MLR, and BRR algorithms. Though there is always space for enhancement, additional study in this field may result in forecasts of the virus's spread that are both more precise and thorough.

Integrating more sophisticated machine learning techniques, such as deep learning and ensemble learning, is one of the research's next goals in order to increase the precision of COVID-19 case predictions. These methods can handle intricate non-linear interactions among data and identify patterns that traditional linear models might overlook.

To offer more personalized predictions of COVID-19 instances, another future objective is to add greater detail data, such as data at each person's level.

This may help in identifying individuals who are at high risk for acquiring the virus and in the establishment of focused prevention strategies.

Furthermore, the suggested approach can be used to forecast additional crucial COVID-19 parameters, especially mortality and hospitalization rates. This can aid in identifying areas and groups that are more vulnerable to severe COVID-19 results and helping to allocate resources appropriately. Lastly, the suggested method can be used to other fields like finance and economics to forecast the impact of epidemics on the world economy. It can also be used to forecast the development of additional infectious illnesses.

7. RESULTS:

	Date	Polynomial Predicted # of Confirmed Cases Worldwide
0	11/15/2022	638007284.000000
1	11/16/2022	638535995.000000
2	11/17/2022	639072115.000000
3	11/18/2022	639615863.000000
4	11/19/2022	640167461.000000
5	11/20/2022	640727127.000000
6	11/21/2022	641295083.000000
7	11/22/2022	641871547.000000
8	11/23/2022	642456742.000000
9	11/24/2022	643050885.000000

	Date	Bayesian Ridge Predicted # of Confirmed Cases Worldwide
0	11/15/2022	719793169.000000
1	11/16/2022	721895745.000000
2	11/17/2022	724002411.000000
3	11/18/2022	726113172.000000
4	11/19/2022	728228031.000000
5	11/20/2022	730346993.000000
6	11/21/2022	732470061.000000
7	11/22/2022	734597239.000000
8	11/23/2022	736728532.000000
9	11/24/2022	738863944.000000

	Date	SVM Predicted # of Confirmed Cases Worldwide
0	11/15/2022	640947307.000000
1	11/16/2022	641494622.000000
2	11/17/2022	642043002.000000
3	11/18/2022	642592448.000000
4	11/19/2022	643142960.000000
5	11/20/2022	643694541.000000
6	11/21/2022	644247190.000000
7	11/22/2022	644800910.000000
8	11/23/2022	645355700.000000
9	11/24/2022	645911563.000000

8. CONCLUSION:

Since the COVID-19 pandemic has presented serious threats to both the world economy and public health, it is imperative that predictions of the virus's progress be precise and thorough. Using Support Vector Machines (SVM), Multiple Linear Regression (MLR), and Bayesian Ridge Regression (BRR) methods, we presented a method to forecast the number of COVID-19 instances in this study.

[3] In order to provide a more precise and thorough prediction of COVID-19 cases, our suggested strategy included numerous demographic and environmental elements that may have an impact on the propagation of the virus [3]. The most important factors that predict COVID-19 cases were found using feature selection approaches after we collected information about the number of COVID-19 cases from various nations and regions.

Our model was trained using the SVM, MLR, and BRR algorithms, and its performance was assessed using a variety of assessment measures, including mean absolute error (MAE), coefficients of determination, and root mean square error (RMSE). Our findings showed that the proposed method was highly accurate in estimating the number of COVID-19 cases.

Insight into the variables that most influence the prevalence of COVID-19 cases was also revealed by our study, which can help healthcare providers and policymakers allocate resources wisely and make well-informed decisions.

In conclusion, the strong and precise COVID-19 case prediction provided by our suggested approach can help to lessen the pandemic's effects and guide public health policy. To increase the precision of COVID-19 case predictions, more research is required to hone our approach and incorporate these new traits.

8. REFERENCES:

1. Farhan Mohammad Khan, Rajiv Gupta. "ARIMA and NAR based Prediction Model for Time Series Analysis of COVID-19 cases in India", Journal of Safety Science and Resilience, 2020

2. "Intelligent Data Communication Technologies and Internet of Things", Springer Science and Business Media LLC, 2022

3. Kamila Tomczak. "Transmission of the 2007– 2008 financial crisis in advanced countries of the European Union", Bulletin of Economic Research, 2022

9. BIOGRAPHY:

SOMULA VENKATA MADHAVA REDDY
is currently studying B.Tech with specification of **ARTIFICAL INTELLIGENCE AND MACHINE LEARNING** in NRI Institute of Technology. He done internship on AI and ML.

KOMPALLI VENKATA RAMANA
is currently studying B.Tech with specification of COMPUTER SCIENCE in NRI Institute of Technology. He done internship on AI and ML.

MATTA TARUN KUMAR
is currently studying
B.Tech with specification of COMPUTER SCIENCE in NRI Institute of Technology. He done internship on AI and ML.

MANDAPAKA RAMA KRISHNA
is currently studying
B.Tech with specification of COMPUTER SCIENCE in NRI Institute of Technology. He done internship on AI and ML.