A PROJECT REPORT ON

**COVID19 ANALYSIS, NEW CASE PREDICTION AND**

**RECOVERY RATE IN WORLD AND INDIA**

USING MACHINE LEARNING

Submitted in partial fulfillment for the requirement of the award of

TRAINING

In

**DATA ANALYTICS, MACHINE LEARNING AND AI USING PYTHON**



Submitted by

**Jayanth Vunnam** (CHAITANYA BHARATHI INSTITUTE OF

TECHNOLOGY (CBIT), HYDERABAD)

Under the guidance of

**Mr. Bipul Shahi**

**ACKNOWLEDGEMENT**

My sincere gratitude and thanks towards my project paper guide

Mr. Bipul Shahi, Corporate trainer, Developer, Traveler, IoT, Artificial

Intelligence, Robotics, Cloud Computing, and Android Apps.

It was only with his backing and support that I could complete the report. He provided me with all sorts of help and corrected me if ever seemed to make mistakes. I have no such words to express my gratitude. I acknowledge my sincere gratitude to the lecturers, research scholars, and lab technicians for their valuable guidance and helping attitude even in their very busy schedule. And at last but not the least, I acknowledge my dearest parents for being such a nice source of encouragement and moral support that helped me tremendously in this aspect.

I also declare to the best of my knowledge and belief that the Project Work has not been submitted anywhere else.

**INTRODUCTION**

COVID-19 is a communicable, infectious disease caused by a newly discovered form of coronavirus. People infected with COVID-19 would experience mild to moderate respiratory problems along with fever, dry cough, and tiredness. So far, the virus has infected 48.9 lakh people worldwide and 3.24 lakh people have succumbed to it; of which 1.1 lakh of the infected people and 3,303 deaths are from India (as of 19th May 2020). The pandemic has caused massive disruption of the healthcare system. The knowledge gained from the previous similar viruses like the Middle East respiratory syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS) appears to not be of much use in the treatment of COVID-19. The recent pandemic has caused a major shift in the lifestyle of people and demands significant improvement in health system immunity to counter the effects of the pandemic. Many countries around the world have taken extreme preventive steps for the containment of the virus. In this context, there is a need for the people of India to take up healthy practices in their lifestyle to control the spread of the virus.

The exponential increase in the number of cases every passing day has given rise to mass panic and fear among the people. Keeping tabs on the ongoing cases would not only provide insights into the trends and patterns occurring in the data, but that data would also act as the source data for predicting the new, forthcoming trends about the spread of the pandemic. These trends would act as valuable sources of information for the government and industries alike to understand the need of the hour and create laws and policies accordingly to meet the needs of the common public. Thus, predictive analysis will aid in the betterment of people’s life during these tough times.

**CONCEPTS**

**MACHINE LEARNING:**

It is the field of study that gives computers the capability to learn without being explicitly programmed. As it is evident from the name, it gives the computer that makes it more similar to humans: Theability to learn.

It is a type of artificial intelligence that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

Machine Learning is defined as **“Field of study that gives computers the capability to learn without being explicitly programmed”**.

Steps In Machine Learning:

1. Training Data
2. Validation Data
3. Testing Data
4. Predicting

Types of Machine Learning:

1. **Supervised learning:** In this type of machine learning, [data scientists](https://searchenterpriseai.techtarget.com/definition/data-scientist) supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified.

Applications:

* Binary classification
* Multi-class classification
* Regression modeling
* Ensembling.

1. **Unsupervised learning.** This type of machine learning involves algorithms that train on unlabeled data. The algorithm scans through datasets looking for any meaningful connection. Both the data algorithms train on and the predictions or recommendations they output are predetermined.

Applications:

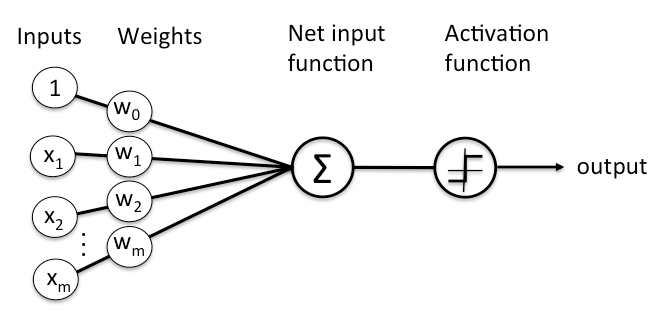
* Clustering
* Anomaly detection
* Association mining
* Dimensionality Reduction

1. **Semi-supervised learning.** This approach to machine learning involves a mix of supervised and unsupervised learnings. Data scientists may feed an algorithm mostly labeled [training data](https://searchenterpriseai.techtarget.com/feature/Using-small-data-sets-for-machine-learning-models-sees-growth), but the model is free to explore the data on its own and develop its own understanding of the data set.
2. **Reinforcement learning.** [Reinforcement learning](https://searchenterpriseai.techtarget.com/definition/reinforcement-learning) is typically used to teach a machine to complete a multi-step process for which there are clearly defined rules. Data scientists program an algorithm to complete a task and give it positive or negative cues as it works out how to complete a task. But for the most part, the algorithm decides on its own what steps to take along the way.

**Neural Networks:**

Neural networks are a set of algorithms, modeled loosely after the human brain that is designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling, or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text, or time series, must be translated.

Neural networks help us cluster and classify. You can think of them as a clustering and classification layer on top of the data you store and manage. They help to group unlabeled data according to similarities among the example inputs, and they classify data when they have a [labeled dataset to train on](https://pathmind.com/wiki/supervised-learning).

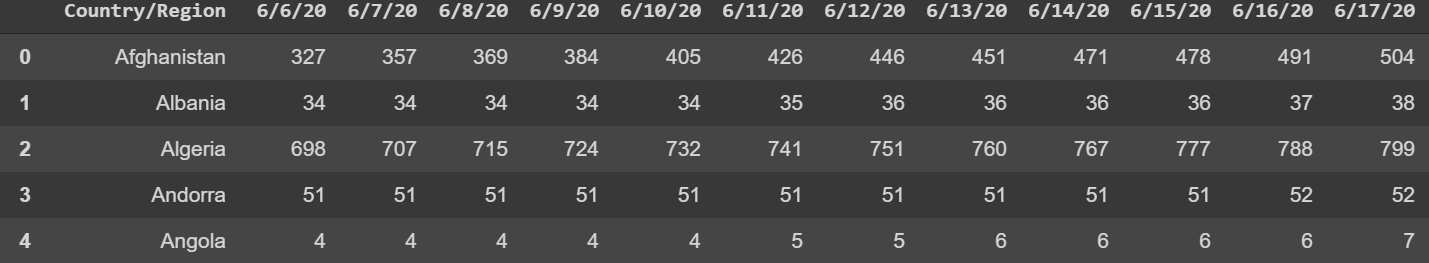


**METHODOLOGY:**

To start off with the project, the datasets were downloaded from the HDX (Humanitarian Data Exchange) website which contains data that is compiled by the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) from various sources [1]. All of these data sets contained information regarding the statistics of “Confirmed cases”, “Recovered Cases” and “Number of deaths” of the pandemic, COVID-19, from 22nd of January,2020 to 15th July, 2020. Three notebooks were created to make working with the data simpler, namely the “Confirmed\_cases”, “Recovered\_cases” and the “Deceased\_numbers” where the respective datasets will be worked upon. All the steps that are carried out hereafter have been carried out in all the notebooks, in a similar way.

Now, the first task was to make the data useful for analysis i.e data cleaning. Since these data sets contained values regarding the number of cases, there weren’t any missing or duplicate data, and hence, there was no requirement for cleaning the data. Feature engineering was then carried out to eliminate a few columns that had minimum/no impact on the analysis and prediction of cases. Each of the data sets was then split into two parts, the world and India data set respectively.



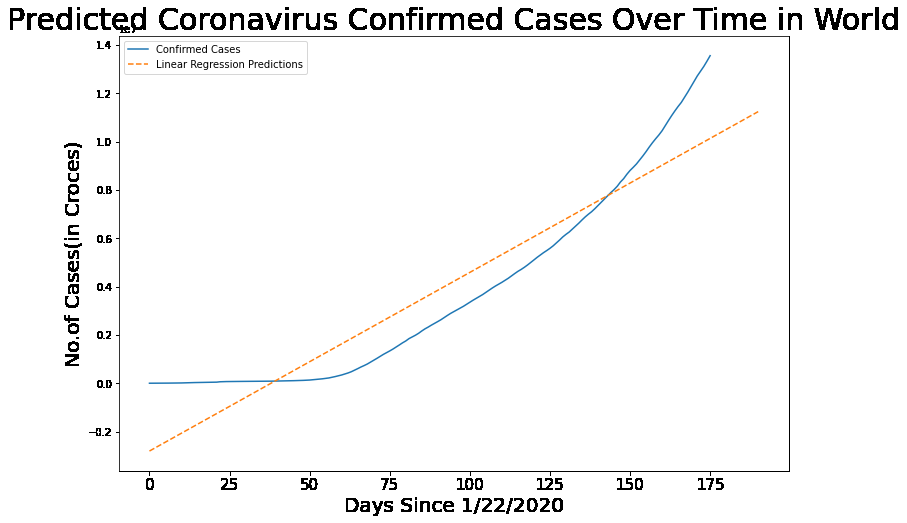


Then, plots related to the different categories were plotted to understand the nature of the data. Moving on to the next part, that is making predictions. The world and the Indian Dataset were then split into a test set and training set each and various models were used to make predictions, namely:

* Linear Regression
* Polynomial Regression
* Bayesian Ridge
* Support Vector Regression (SVR)

**Linear Regression:**

It is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between the dependent and independent variables, they are considering and the number of independent variables being used.

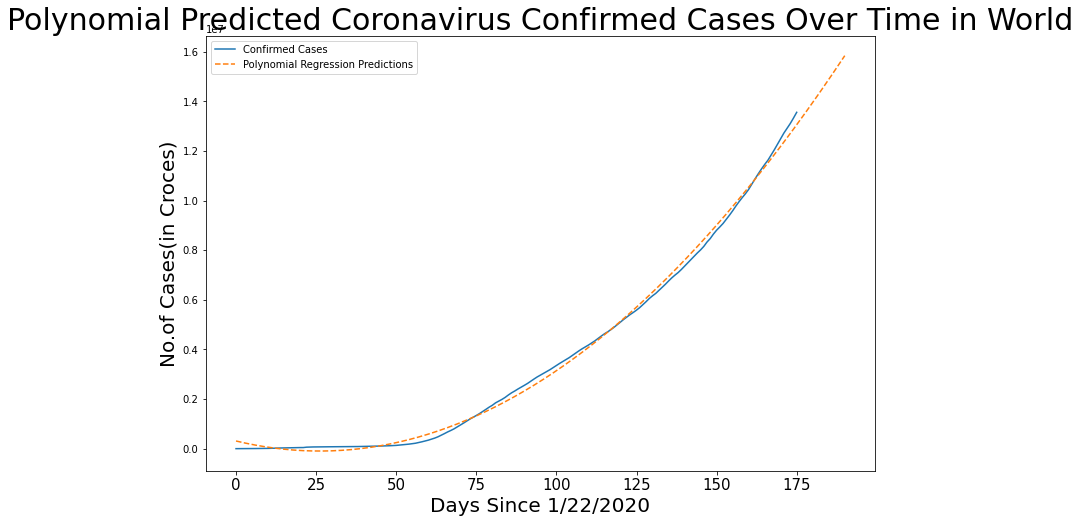


**Polynomial Regression:**

Polynomial regression is a form of [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) in which the relationship between the [independent variable](https://en.wikipedia.org/wiki/Independent_variable) x and the [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) y is modeled as an nth degree [polynomial](https://en.wikipedia.org/wiki/Polynomial) in x.

Polynomial regression fits a nonlinear relationship between the value of x and y.

In Polynomial regression, the original features are converted into Polynomial features of the required degree (2,3,..,n) and then modeled using a linear model.



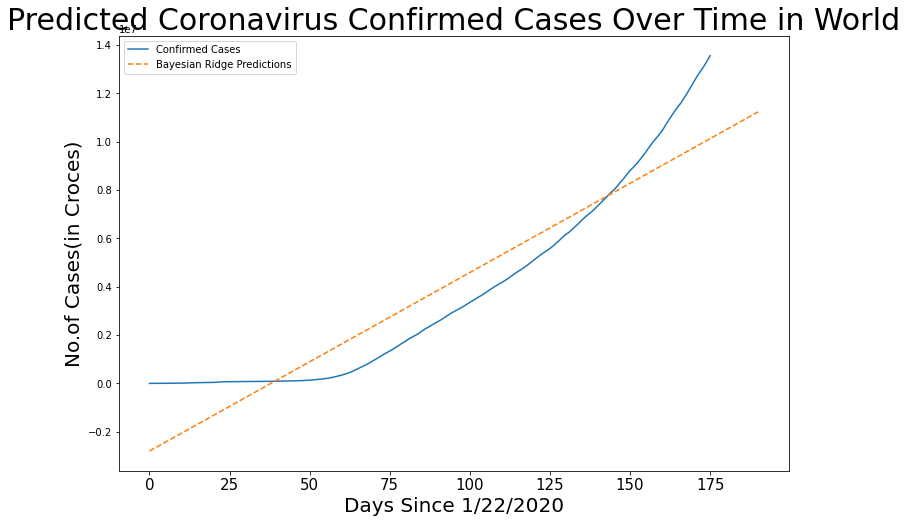
**Bayesian Ridge Regression:**

It is an approach to [linear regression](https://en.wikipedia.org/wiki/Linear_regression) in which the statistical analysis is undertaken within the context of [Bayesian inference](https://en.wikipedia.org/wiki/Bayesian_inference). When the regression model has [errors](https://en.wikipedia.org/wiki/Errors_and_residuals_in_statistics) that have a [normal distribution](https://en.wikipedia.org/wiki/Normal_distribution), and if a particular form of [the prior distribution](https://en.wikipedia.org/wiki/Prior_distribution) is assumed, explicit results are available for the [posterior probability distributions](https://en.wikipedia.org/wiki/Posterior_probability_distribution) of the model's parameters.

Bayesian regression can be implemented by using regularization parameters in estimation. The [BayesianRidge](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.BayesianRidge.html#sklearn.linear_model.BayesianRidge) estimator applies Ridge regression and its coefficients to find out a posteriori estimation under the Gaussian distribution.

The formula or equation to calculate the posterior probability is:

P(c/x) = (P(x/c) \* P(c)) / P(x)

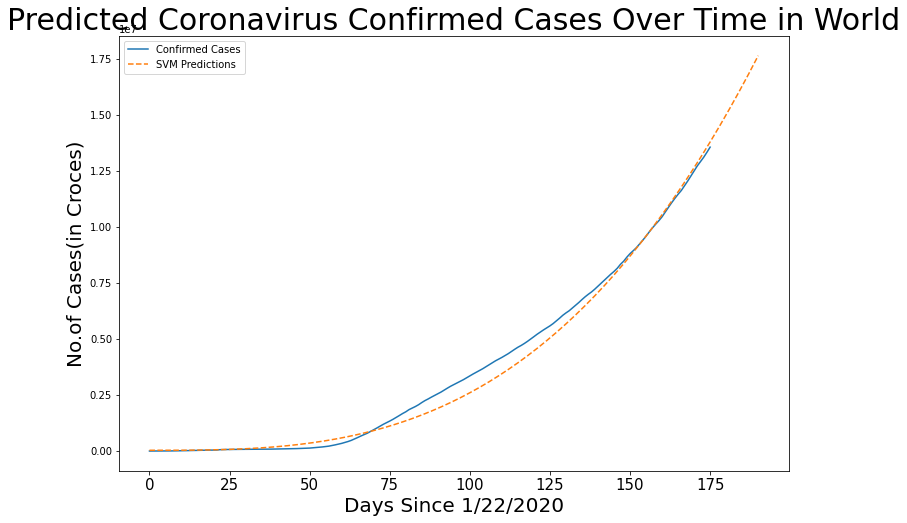


**Support Vector Regression:**

Support Vector Regression (SVR) is quite different from other Regression models. It uses the Support Vector Machine (SVM, a classification algorithm) algorithm to predict a continuous variable.

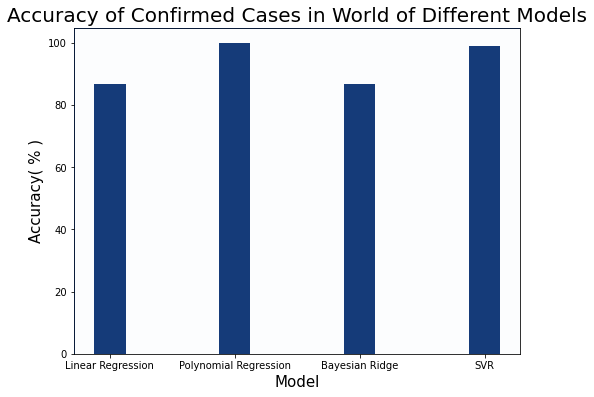
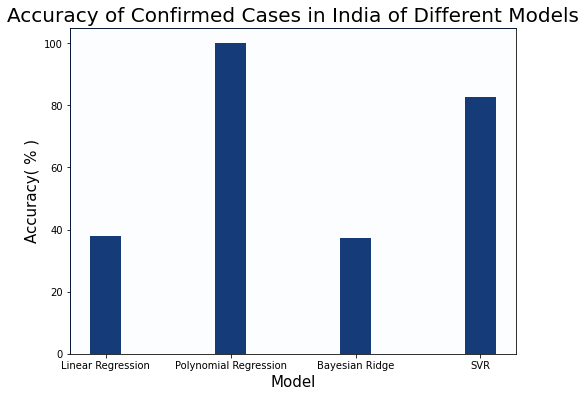
While other linear regression models try to minimize the error between the predicted and the actual value.

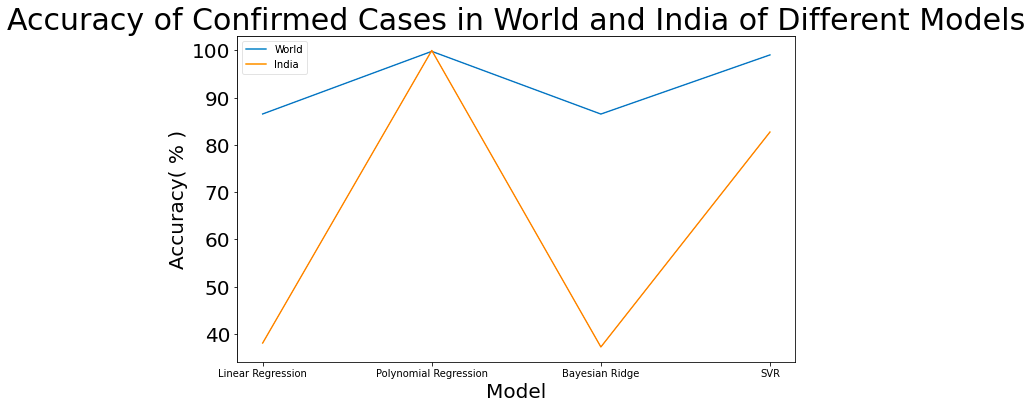
Support Vector Regression tries to fit the best line within a predefined or threshold error value. What SVR does in this sense, it tries to classify all the prediction lines in two types, ones that pass through the error boundary( space separated by two parallel lines) and ones that don’t. Those lines which do not pass the error boundary are not considered as the difference between the predicted value and the actual value has exceeded the error threshold, (epsilon). The lines that pass are considered for a potential support vector to predict the value of an unknown.

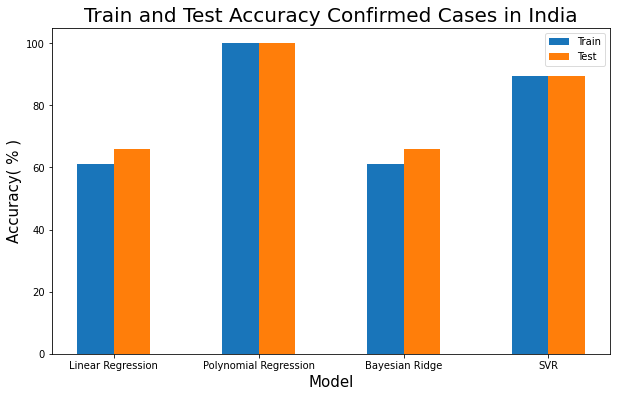


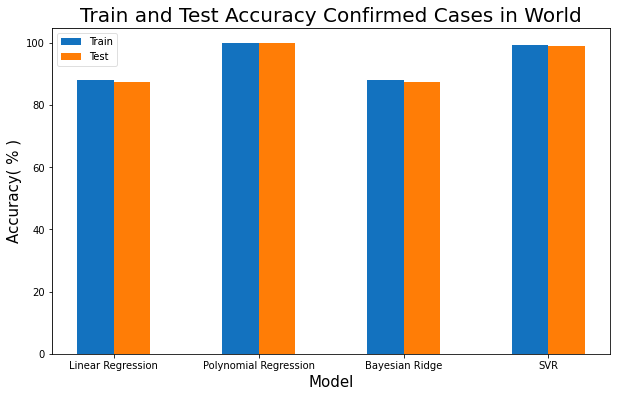
These models were used to predict the “Confirmed”, “Recovered” and “Deaths” in both the world and Indian test and training dataset respectively, till the 30th of July, 2020. The obtained results were then analyzed using different graphs and the models’ accuracies were compared based on the different error metrics, like Mean Absolute Error (MAE), Mean Squared Error (MSE), R2 values.

**Comparison of Different Models:**



****



****

**CONCLUSION**

**Results:**

On comparing the above graphs and results, it can be concluded that the polynomial regression model gave the highest accuracy for both the world and Indian data set, with amazing accuracy of 99% for both the data sets. Also, the next best model was the SVR model which showed an accuracy of 99% for the world dataset and 82% for the Indian dataset. Furthermore, the recovery rate of the COVID-19 infected patients was calculated by dividing the number of recovered patients each day with the no of confirmed cases on that day.

**Advantages and Disadvantages:**

These models have very high accuracy and can be used to predict the trends in the data with almost 100% accuracy. But, the catch is that the data input might change exponentially, like in the Indian Data set, where the number of cases has increased exponentially in the last one month, causing a decrease in the prediction accuracy of the model.

**Scope:**

The scope of the project lies in building models that make more accurate predictions based on the live input of the data so as to build a more accurate and dependable source of information for the government and other organizations that are dependent on such data for creating policies and decisions to meet the needs of the people.

**GitHub Profile:** <https://github.com/Jayanth710/COVID-19-Case-Prediction-and-Analysis>

**BIBLIOGRAPHY**

1. <https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases>
2. <https://www.kaggle.com/>
3. <https://towardsdatascience.com/>