REGRESSION MODEL ANALYSIS ON COVID-19 VACCINATION

Haripriya S
Division of Mathematics
School of Advanced Sciences
Vellore Institue of Technology
Vandalur, Chennai-600127
haripriya.s2021@vitstudent.ac.in

Jayasree M
Division of Mathematics
School of Advanced Sciences
Vellore Institue of Technology
Vandalur, Chennai-600127
jayasree.m2021@vitstudent.ac.in

Keerthana D
Division of Mathematics
School of Advanced Sciences
Vellore Institue of Technology
Vandalur, Chennai-600127
keerthana.d2021@vitstudent.ac.in

Dr. Jaganathan B
Division of Mathematics
School of Advanced Sciences
Vellore Institue of Technology
Vandalur, Chennai-600127
jaganathan.b@vit.ac.in

***Abstract*—The SARS-CoV-2 virus causes COVID-19, a contagious disease. Corona virus Disease is sweeping the globe, and vaccines are in high demand. It has become a triple-mutated virus, making it even more deadly than earlier. Indian Government has launched its Vaccination program on January 16th, 2021 by offering 2 Vaccines-Covid shields and Covaxin. The Government stated that the Ideal Gap between the 1st and 2nd dose should be 4-6 weeks. The COVID-19 database for India is collected from 13th February 2021 – to 13th February 2022 from the government website 'ourworldindata.org'. The dataset deals with the Vaccination rate Vs mortality and was analyzed using the Regression models such as Linear Regression, Robust Regression, Multi linear Regression, Polynomial Regression, Ridge Regression, Lasso Regression and Elastic Net. Throughout the analysis, models were evaluated using the regression metrics R2 square, mean square error (MSE), medium absolute error (MAE) and Root mean square error (RMSE). In this paper, the Linear Regression model is used to predict the number of deaths based on the vaccination ratio.**

Keywords: Covid-19, Linear Regression, Lasso Regression, Polynomial Regression, Ridge Regression, Robust Regression

I.INTRODUCTION

Pandemics have happened throughout history, with new virus strains such as influenza virus causing increased sickness, death, and destruction in the impacted areas.. The coronavirus disease 2019 (COVID-19) pandemic, which has spread to 220 countries and is caused by severe acute respiratory syndrome, Over 63 million laboratory-confirmed illnesses and 1.4 million fatalities have resulted [World Health Organization (WHO)].causing considerable social and economic devastation. According to official estimates, With 524,413 deaths as of May 22, 2022, India has the world largest number of victims of COVID-19 infection (after the United States of America) and the third-country is Brazil..[1]

The World Health Organization (WHO) declared this virus as a pandemic epidemic on February 11, 2020, designating it COVID-19 and declaring that the virus originated in China before spreading to other countries. The top 4 countries that were widely hit by increasing covid 19 cases were the United States, Brazil, India, and France.[2]

To limit the spread of SARS-CoV-2 infection and lessen its effects on health, countries around the world have undertaken enormous control measures, including maintaining social distancing in crowded places and imposing partial and full lockdowns. School and workplace closures, and ordered them to wear face masks in public. Despite the fact that such efforts have helped to flatten the epidemic curve, COVID-19 has affected the society and the e countries economy has fallen down. As a result, long-term preventive actions are critical. Few countries have attempted to develop Herd immunity that avoids epidemics through natural infection. However, such a method has been considered as unethical and impracticable. Many countries closed their borders and didn't allow people from other countries.

Open SAFELY was created to explore parameters linked to COVID-19-related death. It is a health analytics platform in the United Kingdom that covers 40% of all patients in England. Increased death risk from covid 19 is mainly due to the auto immune diseases, chronic diseases, diabetes and the other diseases that are not permanently curable like HIV, AIDS and different types of cancer. As seen by the global outbreak, Individuals with pre-existing cardiovascular illnesses, are more susceptible. People with type 2 diabetes, cancer, and COPD are more likely to have a severe allergic reaction, After being diagnosed with COVID-19, you may develop an infection or die.[4]

Infectious diseases such as smallpox, polio, and plague should be managed and avoided. To increase the life expectancy vaccines are mainly used. To control the global

pandemic like covid 19 , developing a effective vaccine is must, given the severe morbidity and death associated with it. Individual predisposing characteristics must be discovered not only to design surveillance processes for vulnerable groups of affected people, but also to develop vaccine tactics. Indeed, because the number of available doses is currently insufficient to reach the target population, health policy measures that take into account individual and community risk differentials are needed to ensure equitable vaccination distribution [4]

Machine learning is critical for future understanding and analysis of the COVID-19 scenario ,because it analyses patterns from input and utilizes them to obtain automatic predictions or actions.Machine Learning can be considered as a benefit in the medical industry because it can handle enormous datasets beyond the human capability, and the resulting knowledge aid clinicians for planning and considering proper therapy. On the basics of the concept, this work has been carried out for analysis of COVID-19 Vaccination vs Mortality. In this paper, results are obtained using Regression models such as Linear, Lasso, Polynomial Regression, Ridge Regression, Robust Regression.[3]

The fact of Covid-19 vaccinations prevents severe sickness and disease-related fatalities has been repeatedly emphasized. After the first dosage, recipients of the Moderna vaccine had 3.7 deaths per 1,000 people per year, and 3.4 after the second; the mortality rate in the unvaccinated comparison group was 11.1 fatalities per 1,000 individuals per year. COVID-19 incidence and deaths by vaccination status were provided weekly from 25 public health departments that routinely link case surveillance to vaccination data from immunization registries (April 4–December 25, 2021).

Between April 4 and December 25, 2021, a total of 6,812,040 unvaccinated COVID-19 cases and 2,866,517 fully vaccinated COVID-19 cases were reported among people aged 18 in 25 U.S. jurisdictions; by December 4th, 94,640 and 22,567 COVID-19–related deaths had been reported among unvaccinated and fully vaccinated people, respectively.
In comparison to the previous week, the number of new COVID-19 cases and deaths declined by 21% and 8% globally for the week of 14 to 20 February 2022. Over 12 million new cases and 67 000 new fatalities were reported across the six WHO regions.

II. LITERATURE REVIEW

[1] In February 2022 Monika Bajaj , VanshikaRustagi, Tanvi, Rajiv Aggarwal, Priya Singh, Mohamed F. AlAjmi, Afzal Hussain, Md. Imtiaz Hassan, Archana Singh, and Indra Kant K. Singh have proposed a study about Effect of vaccination on cases and deaths due to covid in Asian countries using Machine learning techniques. They used the Covid 19 database to predict covid positives and their death rate in this investigation. Models were developed to predict the amount of patients who would succumb to Covid-19. A Quartic polynomial regression was built utilising Karl Pearson's Coefficient for estimate. They employed a Support vector machine technique to examine the efficiency of the models produced in this study to estimate the number of deaths caused by Covid-19. They have also incorporated the Linear Regression model and OLS Regression model and interpreted it with th help of p-value. Finally, with the help of the polynomial regression, they found that Covid cases reduced upon vaccination and the death rate was decreased. They concluded by proving that the Support vector machines algorithm provides more accurate results.

[2] In 2021 Aditi Srivastava, Mohammad Ayoub Khan, Fahad Algarni, Akshika Choudhary, Rijwan Khan, Indrajeet Kumar have evaluated the performance of regression models for Covid-19. To analyse the Covid-19 data, they employed different regression models and the main one among them was support vector regression model and polynomial regression model. Along with them they utilized the regression parameters like R2 square and Mean squared error. Their study proved that Linear Regression , Polynomial Regression and support vector regression models yields the good results for data related to covid19

[3] Varalakshmi Perumal,VasumathiNarayanana, Sakthi Jaya Sundar Rajasekar (2021) predicted that CT scores using images yield better results with AlexNet+Lasso regression when compared with decision tree. Linear regression yields the best result when analyzing the prediction of CC scores using CT scores. With the help of this model , radiologists can save their valuable time and effort and this model mainly aims to detect the Covid-19 with the prediction of CC scores.

[4] Antonio Giampiero Russo, Adriano Decarli , Maria Grazia Valsecchi (2021) proposed a strategy to identify priority groups for Covid-19 vaccination. They used Lasso cross-validated along with conditional Logistic regression to analyse three sub-cohorts of age and comorbidities. They used a method that defines two degrees of classification for vaccination priorities, making it simple for medical authorities to use and achieve better results and the main purpose of this model is to optimize the vaccination program by identifying frail individuals.

[5] Ashok Kumar, Pooja Jamdagni, and Poonam Chauhan have investigated on the spread of COVID-19 in India and its states using Regression analysis. The Covid-19 dispersion is evaluated in different states using a linear regression model and a polynomial regression model in this paper. The linear regression model is employed for case fatality rate (CFR) and recovery rate (RR) for the different states, according to the data analysis. The polynomial regression model RMSE and percentage error are used to calculate and the prediction is done for the number of patients in different states in India. With the help of Linear regression they found that West Bengal, Madhya Pradesh, Maharashtra and Gujarat the CFR rate is high and RR for the states of Telangana and Maharashtra is high. In polynomial regression, Gujrat, Delhi, and

Maharashtra, Tamil Nadu states are most affected due to Covid- 19 in India.

III. METHODOLOGY

Regression Models basically provide the relationship between independent variables and a dependent variable and they are generally used to predict the cause-and-effect relationship between the attributes. In India, as the covid cases are increasing, the ratio of people getting fully vaccinated is increasing and as a result of that, the death cases are decreasing. This paper mainly analyses the different Regression models using covid 19 vaccination rate versus death rate data. The Regression models are analyzed based on the regression metrics.

- **Data Collection:**
  The data used in this study is taken from "https://ourworldindata.org/". Data is collected for the whole of India from 13th February 2021 to 13th February 2022, which is exactly one year. Data is collected for a variety of variables, including the cumulative number of persons who have received at least one dose of vaccine, the cumulative total of people who have received all two doses of vaccination, new fatalities per day, and the country's total vaccination ratio. The Dataset mainly deals with the vaccination rate vs mortality rate. The dataset contains 7 attributes with 367 data points.

- **REGRESSION MODELS:**

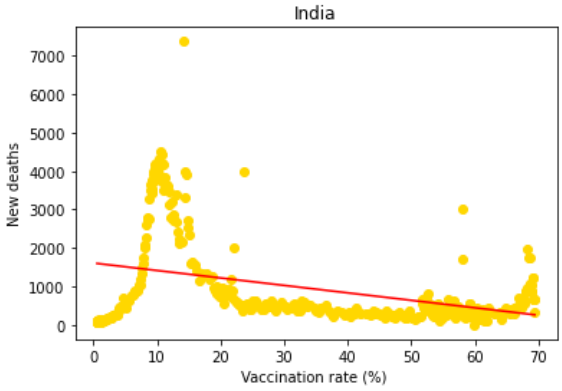
1) Linear Regression:

Linear regression is a statistical method for determining the relationship between two variables known as explanatory and outcome variables. Explanatory variables are known as target variable or label and dependent variables are called as input features. Target variables are plotted on the x-axis and they are used for prediction and input features are plotted on y-axis. n most cases, it produces a straight line called a regression line that shows how the outcome variable changes when the target variable changes.

Linear Regression equation, $y = mx + c$

Where The slope is m,
the intercept is c, t
he target variable is x, and the outcome variable is y.

Simple Linear regression and Ordinary Least squares method are used when we have one or more input and it is generally used to estimate the value of coefficients. Gradient Descent is used when two or more inputs are there and it optimizes the value of coefficients by minimizing the errors.

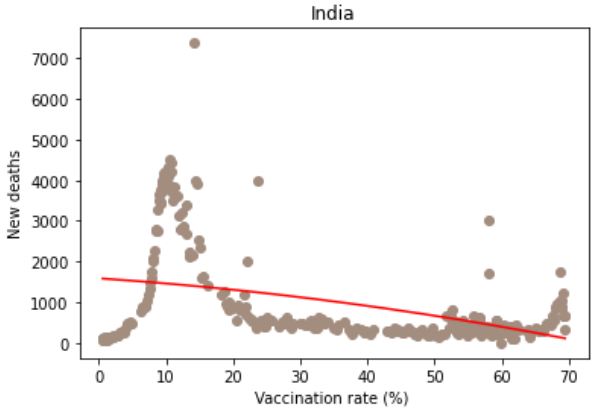


2) Polynomial Regression:

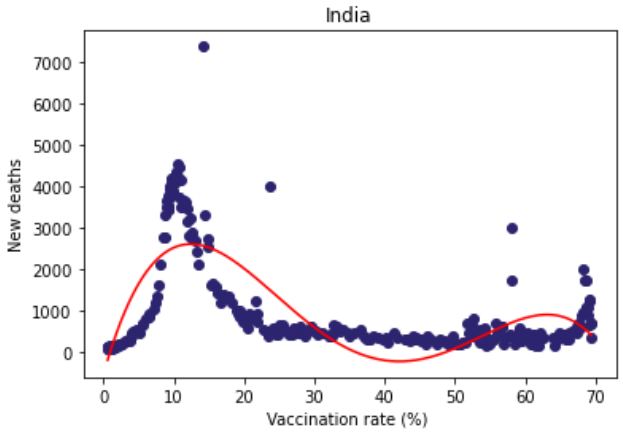
Polynomial Regression is an extension of linear regression and it is used to fit a polynomial equation and represents the curvilinear relationship between the dependent variables and the independent variables. For a good predictor model, polynomial regression is the best technique. The curvilinear relationship indicates that the dependent variable's value changes non-uniformly in relation to the predictor variable.

$$y = x^4 + ax^3 + bx^2 + cx + d$$

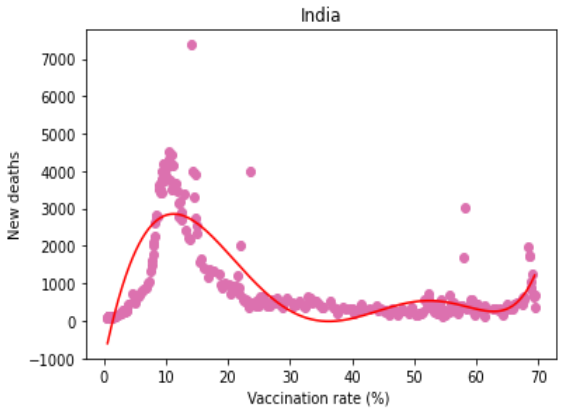
Where a, b, and c are called coefficients of regression and d is called intercept.



The above graph is for Polynomial degree 2.



The above graph is for polynomial degree 4.



The above graph is for polynomial degree 6.

3) Ridge Regression:

Ridge Regression is a model tuning method that provides L2 regularization on multiple regression data that involves multicollinearity. If there is multicollinearity, least squares estimates are unbiased, but because their variances are huge, they may be far from the real value. Ridge regression's main purpose is to reduce standard/inbuilt errors with the help of adding a degree of bias into the regression estimates. To create more reliable estimations, the net effect is employed

$$Y = XB + e$$

Where Y is called as dependent variable,
the independent variable is the X and
regression coefficient is named as B, that has to be estimated and
e represents the residual errors.

$$\min\left(\|Y - X(\theta)\|_2^2 + \lambda\|\theta\|_2^2\right)$$

λ in the ridge function is actually denoted by the value shown here. As a result, changing the values of alpha effectively controls the penalty term. The greater the alpha value, the greater the penalty, and so the size of coefficients is lowered. Ridge regression mainly shrinks the parameters and prevents multicollinearity. The main step in this regression is to standardize the dependent and independent variables. Standardization is done by subtracting the means of their variables and dividing them by their standard deviation.

4) Lasso Regression:

The Least Absolute Shrinkage and Selection Operator is a technique for overcoming overfitting in a regression model, as well as for data model regularization and feature selection. Lasso regression, like Ridge, is primarily used for accurate prediction and employs shrinkage. Shrinkage is the process of reducing the number of data points to the mean. This regression model works effectively with data that is very multicollinear. Lasso regression model uses L1 regularization technique and this technique adds a penalty to the coefficients that result in sparse models with few coefficients that might get eliminated or turn zero from the model.

$$min(\|Y - X\theta\|_2^2 + \lambda\|\theta\|_1)$$

λ is the hypermeter, whose value in the Lasso function is equal to alpha.

5) Robust Regression:

Robust regression approaches differ from least squares regression in that they require fewer assumptions. In order to produce a better fit to the majority of the data, these strategies seek to lessen the influence of outlying occurrences. Outliers have a tendency to skew the least-squares fit in their favor by being given significantly more "weight" than they need. You would expect the weight assigned to each observation in a data set of n observations to be on average 1/n. Outliers, on the other hand, may be given significantly more weight, resulting in distorted regression coefficient estimations. Outliers are difficult to spot as a result of the distortion since their residuals are considerably lower than they would be otherwise. Outliers' influence is down-weighted in robust regression, making their residuals larger and simpler to spot.
It uses the Random sample consensus (RANSAC) method to estimate the parameters.

6) Elastic Net:

A linear regression model was trained using regularizes L1 and L2. This combination allows for the learning of a sparse model with few non-zero weights, comparable to Lasso, but yet retaining Ridge's regularization skills. Elastic-net is useful when there are numerous features that are related to one another. Elastic-net is more likely to choose both, but Lasso is more likely to choose one at random. Elastic-Net can benefit from the trade-off between Lasso and Ridge because it inherits some of Ridge's rotational stability. Elastic Net uses both Lasso and Ridge Regression regularization to remove all irrelevant coefficients but not the informative ones.

$$\boldsymbol{Elastic\ Net\ R = Lasso\ R + Ridge\ R}$$

➢ Metrics for evaluating Regression Models:

1) Mean Squared Error (MSE):

Mean Squared Error is otherwise called MSE, It is an important loss function for algorithms to fit or optimize the least square of regression models. In MSE "Least squares" help to minimize the mean squared error between predictions and expected values. MSE helps users to find the mean or average of the squared differences of the dataset. While squaring the differences negative terms are avoided and MSE is beneficial.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2$$

- i^{th} observed value is called as y.
- corresponding predicted value is called ŷ.
- n is called the no. of observations.

. While squaring the difference between expected and predicted values has the effect of inflating large errors. The larger the resulting squared positive error. The advantage of MSE is differentiable using graphs, and we can use it as a loss function. The disadvantage of MSE is while calculating is squared unit of output. And MSE is not robust to outliers.

2) Mean Absolute Error (MAE):

Mean Absolute Error otherwise called "MAE", is one among the famous parameter. The prediction is done by the matching unit of error score and unit of target value. MAE is used to calculate the average of the absolute error values. MAE takes only positive values.

$$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|$$

- n = the number of errors,
- $|y_i - x_i|$ = the absolute errors.
- $\sum$ is absolute value of residual

The use of MAE is the same unit as the output variable, robust to outliers. The drawback of the MAE is graph is not differentiable so we need to implement various optimizers like Gradient descent so it can be differentiable.

3) Root Mean Squared Error (RMSE):

Root mean squared error is otherwise called "RMSE", it is one of the famous metrics. It is also known as the square root of mean square error. RMSE units are the same as the original units and target value which is predicted.

$$RMSE = \sqrt{MSE}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2$$

Where the number of observations is called n,
the expected value is called y and
y^ (y cap) is the predicted value. The mean of the square root of mean squared error values is not calculated in RMSE. The square root is the inverse of the square operations, where the RMSE square root reverses the operation to the result remains positive. The use of RMSE is the interpretation of loss is easy. The drawback of the RMSE is robust to outliers and is not efficient when compared to mean absolute error (MAE).

4) R Squared (R2):

R2 score is one of the metrics in regression models which tells the performance of the model. Where the context of MAE and MSE depends and R2 score is independent. The calculations of R2 square of predicted line is better than mean/average line. R2 is also called as Coefficient of determination otherwise called as goodness of fit.

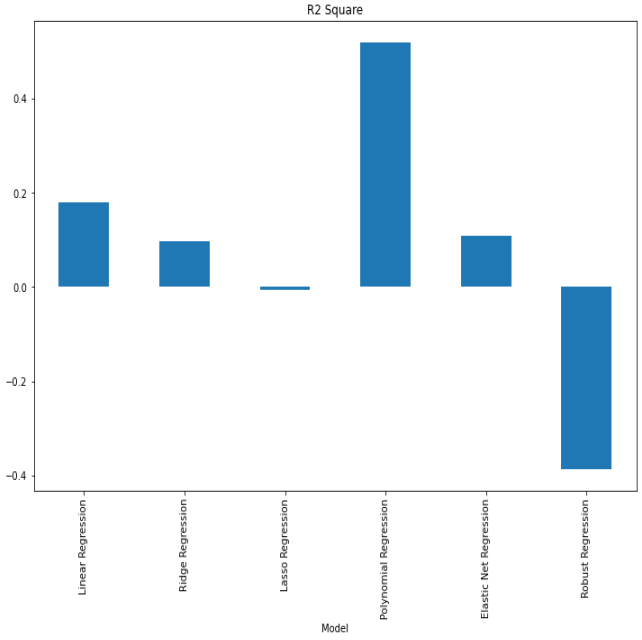
$$R^2\ squared = 1 - \frac{SSr}{SSm}$$

Where SSR is called the of Regression line's squared Sum of Error and the SSM is called as Mean line's Squared sum of error.
When R2 score is zero the regression line and mean line is overlapping then the model performance is bad, ad not capable of output column. In other case R2 score is 1 and the division term is zero the regression line cannot make any mistake and it is perfect but in the real world it is not possible. When R2 score is 0 to 0.8 the performance model is capable to explain the variance of data in normal case.

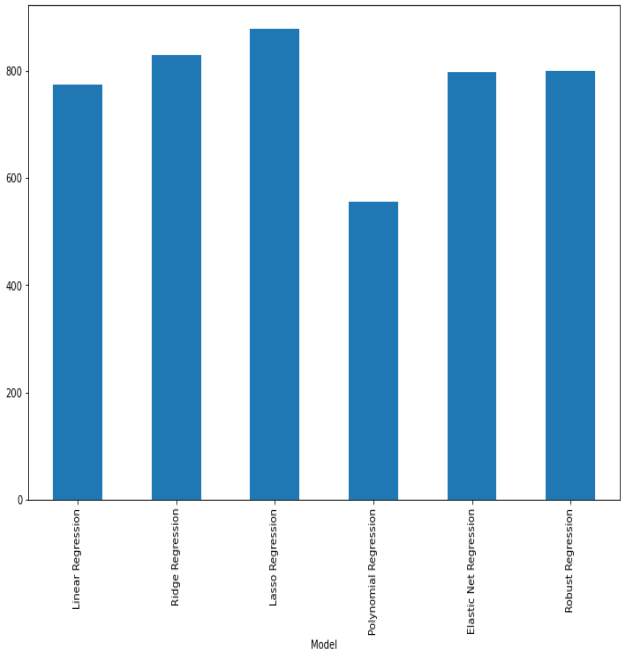
IV. RESULT:

MODEL	MSE	MAE	RMSE	R2 SQUARE	CROSS VALIDATION
Linear Regression	774.108828	1.15E+06	1072.98921	0.180993	-535.457944
Polynomial Regression	555.31634	6.75E+05	821.293772	0.520163	0
Robust Regression	731.479269	1.46E+06	1207.95549	-0.038004	-70.21936
Lasso Regression	879.352371	1.41E+06	1189.1311	-0.005904	-553.233656
Ridge Regression	829.439	1.27E+06	1126.67496	0.096986	-536.740314
Elastic Net	798.199665	1.25E+06	1118.756581	0.109635	-553.434276

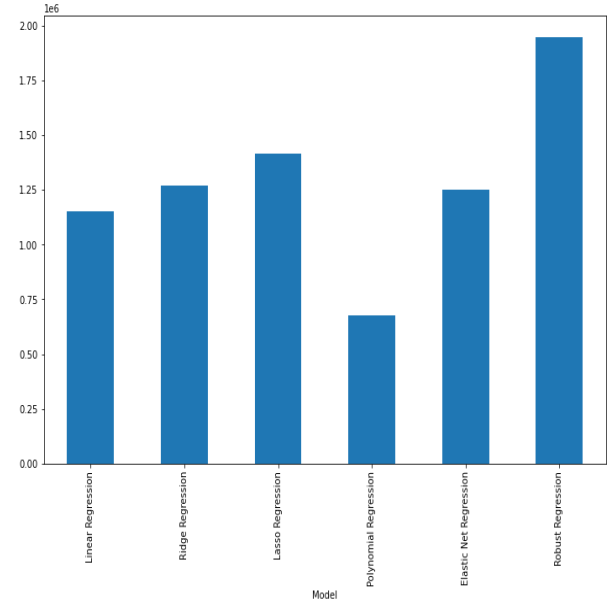
➢ **R2 SQUARE:**



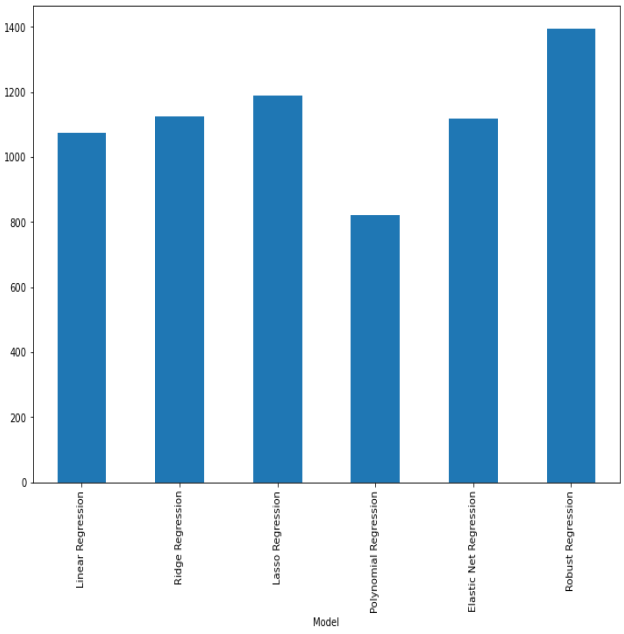
➢ **Mean Absolute Error (MAE):**



➢ **Mean Squared Error (MSE):**



➢ **Root Mean Squared Error (RMSE):**



V. CONCLUSION

The Regression Model analysis was conducted on the covid 19 vaccination dataset of India from 13th February 2021 to 13th February 2022. The dataset contains details about the vaccination rate and new deaths per day. Different regression models were applied to the dataset and the models were analyzed based on the regression parameters like R2 square for finding the best fit and Mean absolute error (MAE) for finding the minimum error, along with

Root mean squared error (RMSE) and Mean Squared Error (MSE). Based on the metrics, we obtain that R2 square gives the
maximum value for polynomial regression and minimum value for Robust Regression. Root mean squared error (RMSE) gives the minimum value for polynomial regression., and maximum value for Robust regression. Mean Squared error (MSE) gives the minimum value for Polynomial Regression and maximum value for Lasso Regression. Mean absolute Error (MAE) gives the minimum values for polynomial regression and maximum value for Robust Regression.

Therefore, from the analysis we conclude that Polynomial regression is the best model and Robust regression model is the poor model for theCovid-19 vaccination vs mortality data used in this paper.

Reference:

1) Rustagi,V, BajajM, Singh P, Aggarwal R., AlAjmi M. F, Hussain A & Singh I. K. "*Analyzing the Effect of Vaccination Over COVID Cases and Deaths in Asian Countries Using Machine Learning Models*". *Frontiers in Cellular and Infection Microbiology, 1380,* (2022).

2) KhanMA, Khan R, Algarni, F, Kumar I, Choudhary A, & Srivastav, A. "*Performance evaluation of regression models for COVID-19: A statistical and predictive perspective"*. *Ain Shams Engineering Journal,* 13(2), 101574, (2022).

3) PerumalV, Narayanan V, & Rajasekar S. J. S. "*Prediction of COVID Criticality Score with Laboratory, Clinical and CT Images using Hybrid Regression Models*". *Computer Methods and Programs in Biomedicine*, 209, 106336, (2021).

4) RussoA. G, Decarli, A, & Valsecchi M. G. *"Strategy to identify priority groups for COVID-19 vaccination: A population-based cohort study"*. *Vaccine*, 39(18), 2517-2525, (2021).

5) Chauhan, Poonam, Ashok Kumar, and Pooja Jamdagni. *"Regression analysis of covid-19 spread in India and its different states." medRxiv* (2020).