Comparison of Regression Models on Covid-19 Cases

Joseph George[1], Ranjeesh R Chandran[2]

[1]*Asst Prof, Dept. of CSE, ASIET, Kalady, Kerala, India*
[2]*Asst Prof, Dept. of AE & I, ASIET, Kalady, Kerala, India*

---***---

Abstract - *The outbreak of Coronavirus(CoV) disease 2019 (COVID-19) are a large family of viruses that causes illness ranging from common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV) has thus far killed over 1,694 people and infected over 33,514 in India. The basis of this study is to identify and create an optimal Death Prediction System (DPS) which predicts the death rate by processing and applying an effective, optimized and convenient regression methods and measures that allows the government and the health workers for planning the eradication of the threatening disease.*

***Key Words*: COVID-19, Adjusted-R, Root Mean Square Error, Polynomial Regression, Support Vector Regression, Random Forest Regression, Decision Tree Regression.**

1. INTRODUCTION

CoV are an outsized family of viruses that causes illness starting from cold to more severe diseases like Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV). This new type of zoonotic disease is often transmitted between citizenry. Most of the people infected with the COVID-19 shows symptoms within 14 days of exposure to the virus experience mild to moderate respiratory disease and recover without requiring special treatment [1]. People with underlying medical problems like disorder, diabetes, chronic respiratory illness, and cancer are more likely to develop serious illness.

The simplest way to prevent and hamper transmission is by washing hands or using an alcohol based rub frequently and not touching your face. This virus will spread mainly through droplets of saliva, discharge from the infected nose when the person coughs or sneezes, so it's important that you simply also practice respiratory. At this point, there are not any specific vaccines or treatments for COVID-19 the amount of latest cases and therefore the death rate increases rapidly. The idea of this study is to spot and make an optimal Death Prediction System which predicts the death rate by processing and applying effective, optimized and convenient regression methods and measures that allow the officials and therefore the doctors for planning the eradication of the threatening disease.

This study includes collection of COVID-19 data of affected Indian people from the official website. The data set contains the number of confirmed cases, active cases, recovered cases and the death cases from January 30th till May 10th. These data were collected from the website corona.mygov.in/ official website for COVID-19.

2. RELATED WORK

Numerous studies have shown that machine learning and statistical techniques are the powerful tools for predictions. Our study is to predict the number of death from multiple independent variables [2] by finding a regression function that approximates mapping from an input domain to real numbers on the basis of a training sample and to compare them and to identify the best model which fits our data. Multiple regression analysis is used to see if there is a statistically significant relationship between the sets of variables. It is also used to find the trends in those sets of data. Multiple regression analysis differs from a simple linear regression with an increased number of predictors used in the regression.

In Support vector regression (SVR) the best fit line is the hyper plane that has a maximum number of points. To find the hyper plane SVR hyper parameters plays an important role. A kernel helps to find a hyper plane in the higher dimensional space without increasing the computational cost. Cost of computing increases with data dimensions due to difficulty in finding a separating hyper plane in a given dimension and so is required to move in a higher dimension. Hyper plane is the line that will be used to predict the continuous output. A decision boundary is like a demarcation line on either sides of the hyper plane.

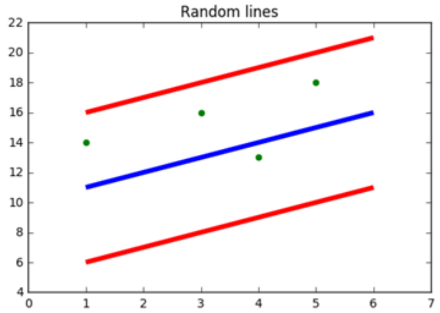

Fig 2.1 SVR Decision boundary and Hyperplane

Consider these decision boundary two red lines being at any distance, say '+a' and '-a' from the hyper plane a blue line.

Our objective, when we are moving on with SVR, is to basically consider the points that are within the decision boundary. Our best fit line is the hyper plane that has a maximum number of points [3].

If the equation of the hyper plane is

$Y = wx+b$

Then the equations of decision boundary become

$wx+b= +a$ and $wx+b= -a$

Thus, any hyper plane that satisfies our SVR should satisfy

$-a < Y- wx+b < +a$

Here, the main aim is to decide a decision boundary at 'a' distance from the original hyper plane such that data points closest to the hyper plane or the support vectors are within that boundary line. Hence, we are going to take only those points that are within the decision boundary and have the least error rate, or are within the Margin of Tolerance. This gives us a better fitting model.

Decision tree regression models ID3 which employs a top-down, greedy search through the space of possible branches with no backtracking. Here, the tree is incrementally developed by breaking down the training data set into smaller and smaller subsets. The final tree model is with decision nodes and leaf nodes. Each tested attributes in a decision node forms its branches. Leaf node represents a decision on the numerical target. Here we use Standard Deviation Reduction to construct a decision tree for regression. Tree construction begins from the root node in top down manner by involving partitioning the data into subsets that contain instances with similar values. Homogeneity of a numerical sample is calculated using standard deviation which will be zero for complete homogeneous numerical samples. The standard deviation reduction is based on the decrease in standard deviation after a data set is split on an attribute. Finding attribute with highest standard deviation reduction is the principle behind constructing a decision tree.
This additive model random forest makes predictions by considering the output value of all decisions trees constructed from a sequence of base models.

$g(x) = f_0(x) + f_1(x) + f_2(x) + ...$

The technique of combining multiple decision tree models to obtain better predictive performance is called ensemble [4]. Here, each base model is a simple decision tree, constructed independently using a different sub sample of the data.

Polynomial regression model [5] is nth degree polynomial, due to the curvilinear relationship between independent and dependent variables.

$y = a + b_1 x^1 + b_2 x^2 + e$

Here y is dependent variable on x, y-intercept is a, e is the error rate.

In general the nth degree model is

$y = a + b_1 x^1 + b_2 x^2 +....+ b_n x^n$

3. MATERIAL AND METHODS

This study includes collection of COVID-19 data of affected Indian people from the official website. The data set contains the number of confirmed cases, active cases, recovered cases and the death cases from January 30th till May 4th. Totally there were 71 records with 3 independent variables. This 71 dataset is divided into 70% for test set and 30% for train set as an input for various regression model creation and evaluation. Preprocessing is required inorder to build a better model with good prediction.

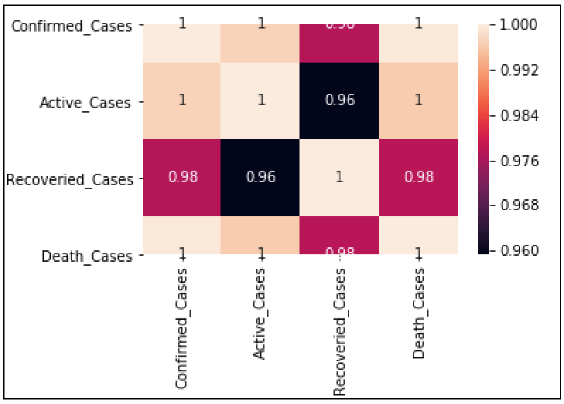

Fig 3.1: Correlation Matrix

Here, missing values, duplicate records, encoding categorical variables are taken care using preprocessing tools. The data is analyzed using principal component analysis an exploratory data analysis to find the insights to refine the important features variable selection that will be used in our model. Once EDA is complete and insights are drawn, its feature is used for our modeling. Once the model is fine tuned and is finalized by validated with k-fold validations to generalize the model. Finally the model performance is evaluated using the r-Squared value. The process is continued for other regression algorithms like support vector machine, ID3, Random Forest and polynomial regression and their performances is compared.

Process:
Step 1: Collect the dataset.
Step 2: Load and merge the dataset.
Step 3: Preprocess the dataset.
Step 3: Divide the dataset into test set and train set.
Step 4: Train the model using training set.
Step 5: Reduce the error and fine tune the model
Step 6: Test the model performance.

Step7: Repeat the process using various regression algorithms.
Step 8: Compare model performances.

The various Regression models are tested with these 71 records with 3 independent variables. This 71 dataset is divided into 70% for test set and 30% for train set as an input for various regression model creation and evaluation.

Over-fitting is where your model is too complex for your data it happens when your sample size is too small. If you put enough predictor variables in your regression model, you will nearly always get a model that looks significant. While an over-fitted model may fit the idiosyncrasies of your data extremely well, it won't fit additional test samples or the overall population. The model's p-values, R-Squared and regression coefficients can all be misleading. Basically, you're asking too much from a small set of data. Here, using cross validation to detect over-fitting this partitions your data, generalizes your model, and chooses the model which works best. One form of cross-validation is predicted R-squared. Most good statistical software will include this statistic, which is calculated by removing one observation at a time from your data, estimating the regression equation for each round and by using the regression equation to predict the removed observation.

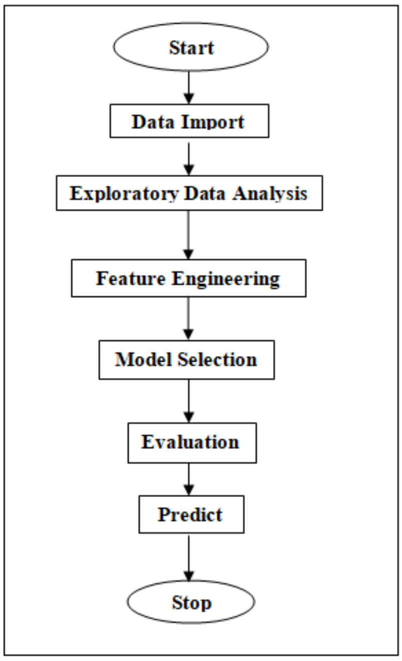

Fig 3.2: Prediction model

4. EXPERIMENTAL RESULT

COVID-19 data of affected Indian people collected from the official website contains 71 records with 3 independent variables. This 71 dataset is divided into 70% for test set and 30% for train set as an input for various regression model creation and evaluation. It is found that the polynomial regression is having a low error rate and highest R^2 value for both training and set predictions. Also, the graph shows how good the model predicting the number of deaths for the test set input value such as active cases, recovered cases is given. The table below shows the performances of various regression models.

Models	Split	RMSE	R^2	Adj-R^2
Random Forest Regression	Train (70%)	14.1	0.9995	0.9994
	Test (30%)	30.99	0.9956	0.9952
Decision Tree Regression	Train (70%)	32.08	0.9981	0.9974
	Test (30%)	45.06	0.9907	0.9898
Polynomial Regression	Train (70%)	7.64	0.9997	0.9998
	Test (30%)	9.89	0.9998	0.9995
SVR linear kernel	Train (70%)	44.42	0.9946	0.9944
	Test (30%)	41.4	0.9922	0.9914
SVR poly kernel	Train (70%)	184.8	0.9071	0.9032
	Test (30%)	237.61	0.7427	0.717
SVR rbf kernel	Train (70%)	53.15	0.9923	0.992
	Test (30%)	39.1	0.9931	0.9923
Multiple Linear Regression	Train (70%)	15.44	0.9994	0.9993
	Test (30%)	21.76	0.9978	0.9976

Table -1: Performance Comparison

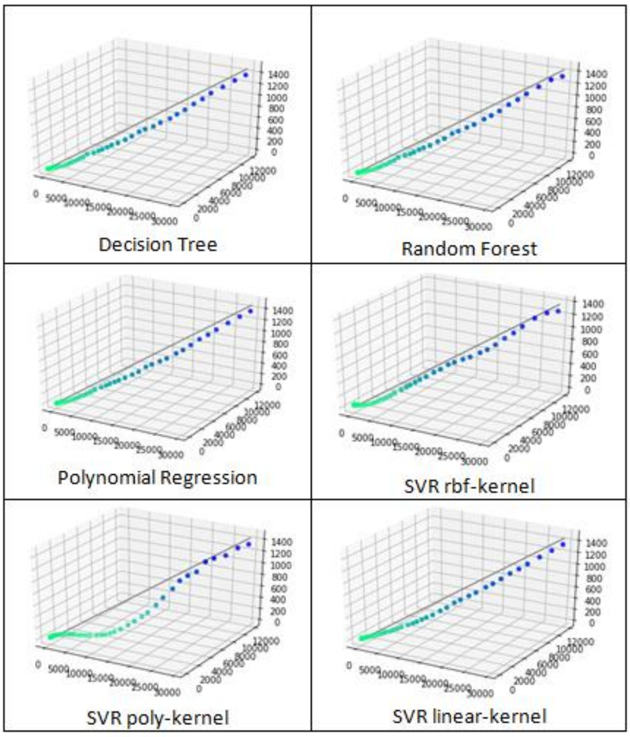

Fig 4.1: Predicted output

5. CONCLUSION

This model predicts the death rate accurately with an error rate of 7.64% during training and 9.89% during the test time. It is found that the polynomial regression is having a low error rate and highest R^2 and adj-R^2 during training and test phase, which helps to accurately predict the death rate which helps to take necessary measures for health workers for planning the eradication of the threatening disease.

REFERENCES

[1] Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation and Diagnosis for COVID-19, Feng Shi, et al, IEEE Reviews in Biomedical Engineering, 10.1109/RBME.2020.2987975 April 2020.

[2] Human age estimation with regression on discriminative aging manifold, Y Fu, TS Huang - IEEE Transactions on Multimedia, 2008.

[3] Travel-time prediction with support vector regression, by Chun- Hsin Wu, Jan-Ming Ho and D.T. Lee, IEEE Transactions on Intelligent Transportation Systems (Volume: 5 , Issue: 4 , Dec. 2004)

[4] An Ensemble Random Forest Algorithm for Insurance Big Data Analysis, by Weiwei Lin, Ziming Wu, Longxin Lin, Angzhan Wen, Jin Li, IEEE Access (Volume: 5) Page(s): 16568 - 16575, August 2017.

[5] Least squares orthogonal polynomial regression estimation for irregular design, Waldemar Popiński, Pages 631-647, 31 Dec 2018.