iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VIII **Month of publication:** August 2021

DOI: https://doi.org/10.22214/ijraset.2021.37652

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

Analysing the Covid Data Using Various Machine Learning Techniques

Mannat Uppal[1], Mohammed Adhil S[2]
[1, 2]*Department of Biomedical Engineering (Medical Electronics), SSN College of Engineering, Chennai, Tamil Nadu, India*

Abstract*: The outbreak of pandemic covid-19 (corona virus) has shaken this entire world. It has widely spread in all over world. The officials of a country are still in task of controlling it in order to prevent people from severity of causative agent as their variants get mutated. There are need of some approaches and strategies to follow as a precaution. The methods like computational modelling, statistical analysis, quantitative analysis helps in predicting the possibility of impact and make aware of it earlier. In this paper, we developed a regression model that can predict the best fitting model of a large complex data in order to predict the total deaths per million of all world countries. With this experiment we proposed different regression methods and found which is best out of all. With the help of the Regression, we are able to develop the relationship between dependent and independent variables. Different regression techniques we used were: Random Forest, Linear Regression, Extra tree Regression, XgBoost Regression, Ridge Regression, Lasso Regression, Decision Tree Regression and the best result was given by XgBoost Regression.

Keywords*: Covid-19, Machine Learning, Fitting model, Regression, Total death per million.

I. INTRODUCTION

The Corona virus (Covid-19) which belongs to family Corona Viridine and order Noroviruses which was originated in Wuhan, China has made this world in big tragedy, there was no anti-viral drug to completely eradicate it but some precautions like avoiding close contact with the people, isolating themselves for some days when they find symptoms are to be followed [1]. The officials of each country are still in task of controlling it in order to prevent people from severity of causative agents as the variants get mutated. As of now, each and every country got vaccination just for temporary protection but cases seem to be rising and all the country at the midst of next wave. There are need of some approaches and modelling strategies to make people follow a precaution and make them aware instead of heavy loss. Machine Learning techniques is a subset of Artificial Intelligence which allows some computational technique very easier. It allows in inspecting the data, analysis it and learn the data of its own [2]. There are lot of machine learning resources available for easy accessing for engineers in order to use machine learning approaches in their new project and creating new impactful machine learning systems [3]. As told earlier, ML techniques can investigate large amount of complex data. It is essential to display precise results about the profitable chances or hazardous risks and it is important to train the data properly whatever time it needs. The combination of MLT with AI can produce to effective processing of large amount of data. Though several types of ML techniques are available, that can clearly solve the problem of large dataset in quick succession so that it can help in improving the status of health sector and industry side. This machine learning techniques can be used for classification and prediction purposes. Regression analysis, means a machine learning method that allows to predict a continuous variable result from the one (or) more predictor variables. This regression is much more useful in future outcomes, their familiar types of algorithms are lasso, ridge, linear and polynomial [4]. In this work, we collected a covid dataset from an open website, we did some feature engineering, statistical analysis, outliers treating and fitting the model. For model fitting we used regression machine learning techniques like Random Forest, Linear Regression, Extra tree regression, Lasso regression, Decision tree regression to predict the good fitting model for the complex covid data we used. The complete data engineering work performed is explained below. Finally, we tabulated the values of MSE (Mean Square Error) and cross- validation to display which regression model we used will be fitting the large data we used.

II. LITERATURE REVIEW

From the research paper," Machine learning-based prediction of COVID-19 diagnosis based on symptoms, nature partner" by Yazeed Zoabi, Shira Deri-Rozov1 and Noam Shomron, where their model was able to predicted COVID-19 test results with high accuracy using only eight binary features: sex, age ≥ 60 years, known contact with an infected individual, and the appearance of five initial clinical symptoms. They developed a model that detects COVID-19 cases by simple features accessed by asking basic questions. This study may benefit the health system response to future epidemic waves of this disease and of other respiratory viruses in general [5].

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

From the paper," Community detection using unsupervised machine learning techniques on COVID-19 dataset" by Laxmi Chaudhary, Buddha Singh where they used unsupervised learning approach- PCA is used to reduce the dimensionality of country covid-19 dataset with affecting the significant information and using that reduced dataset they performed K-means clustering to detect the community of countries based on cases. It helps the same community of people to follow same aids as preventive measures [6].

Then from the paper," COVID-19 Epidemic Analysis and Prediction Using Machine Learning Algorithms" by Arun Solanki and Tarana Singh, where they considered covid considered as a time series data. The 3 different datasets from 3 different sources were used. The model like ARMA, SERIMAX, Holtwinter's exponential smoothing model, LSTM were used. The virus growth not growth not stabilized and total excepted cases in September middle is 1,40,000. The LSTM virus growth is not stabilized. SARIMA, it made prediction with MAPE value [9].

III.METHODOLOGY

In this section we explained about the feature engineering, pre-processing methods and regression techniques we used in this dataset in order to identify best fit model for this covid dataset.

A. Collecting the dataset

For this work, we collected covid dataset from an open-source website named Our World in Data. It is the website that helps in getting any sort of data against world's problem. It is more useful for researchers of data science field to collect the data periodically and can sort the problems by analysis those data. It is the best website where the data is updated each and every time. The dataset we used has a shape of 101753x60. The Fig 1 tells about the columns that are in dataset we used.

```
Index(['iso_code', 'continent', 'location', 'date', 'total_cases', 'new_cases',
       'new_cases_smoothed', 'total_deaths', 'new_deaths',
       'new_deaths_smoothed', 'total_cases_per_million',
       'new_cases_per_million', 'new_cases_smoothed_per_million',
       'total_deaths_per_million', 'new_deaths_per_million',
       'new_deaths_smoothed_per_million', 'reproduction_rate', 'icu_patients',
       'icu_patients_per_million', 'hosp_patients',
       'hosp_patients_per_million', 'weekly_icu_admissions',
       'weekly_icu_admissions_per_million', 'weekly_hosp_admissions',
       'weekly_hosp_admissions_per_million', 'new_tests', 'total_tests',
       'total_tests_per_thousand', 'new_tests_per_thousand',
       'new_tests_smoothed', 'new_tests_smoothed_per_thousand',
       'positive_rate', 'tests_per_case', 'tests_units', 'total_vaccinations',
       'people_vaccinated', 'people_fully_vaccinated', 'new_vaccinations',
       'new_vaccinations_smoothed', 'total_vaccinations_per_hundred',
       'people_vaccinated_per_hundred', 'people_fully_vaccinated_per_hundred',
       'new_vaccinations_smoothed_per_million', 'stringency_index',
       'population', 'population_density', 'median_age', 'aged_65_older',
       'aged_70_older', 'gdp_per_capita', 'extreme_poverty',
       'cardiovasc_death_rate', 'diabetes_prevalence', 'female_smokers',
       'male_smokers', 'handwashing_facilities', 'hospital_beds_per_thousand',
       'life_expectancy', 'human_development_index', 'excess_mortality'],
      dtype='object')
```

Fig 1. Column name in dataset

B. Software and Language used

For our analysis purpose, we used Google Colaboratory. It is the best platform for data scientist and data analytics to use their knowledge of python in the field of Machine Learning and Deep Learning. We need not download any python library as it is inbuilt software from google cloud. Whatever experiments executed will be stored in our drive automatically, so that we can refer those ever long.

C. Data Analysis and Cleaning

Once we imported the dataset, we imported the needed library to perform our analysis. We did some fearture engineering work like removing inappropriate column, adding values to empty cell of the column. Then we did univariant analysis using plotted box plot from seaborn library for all the columns in the dataset. From the Fig 2, we can understand that some of columns (totally 37 columns) like total cases, new cases, total deaths, new deaths, etc has too many outliers with improper box plot. And the Fig 3, shows the empty plot. Thus, we eliminated those 15 columns.

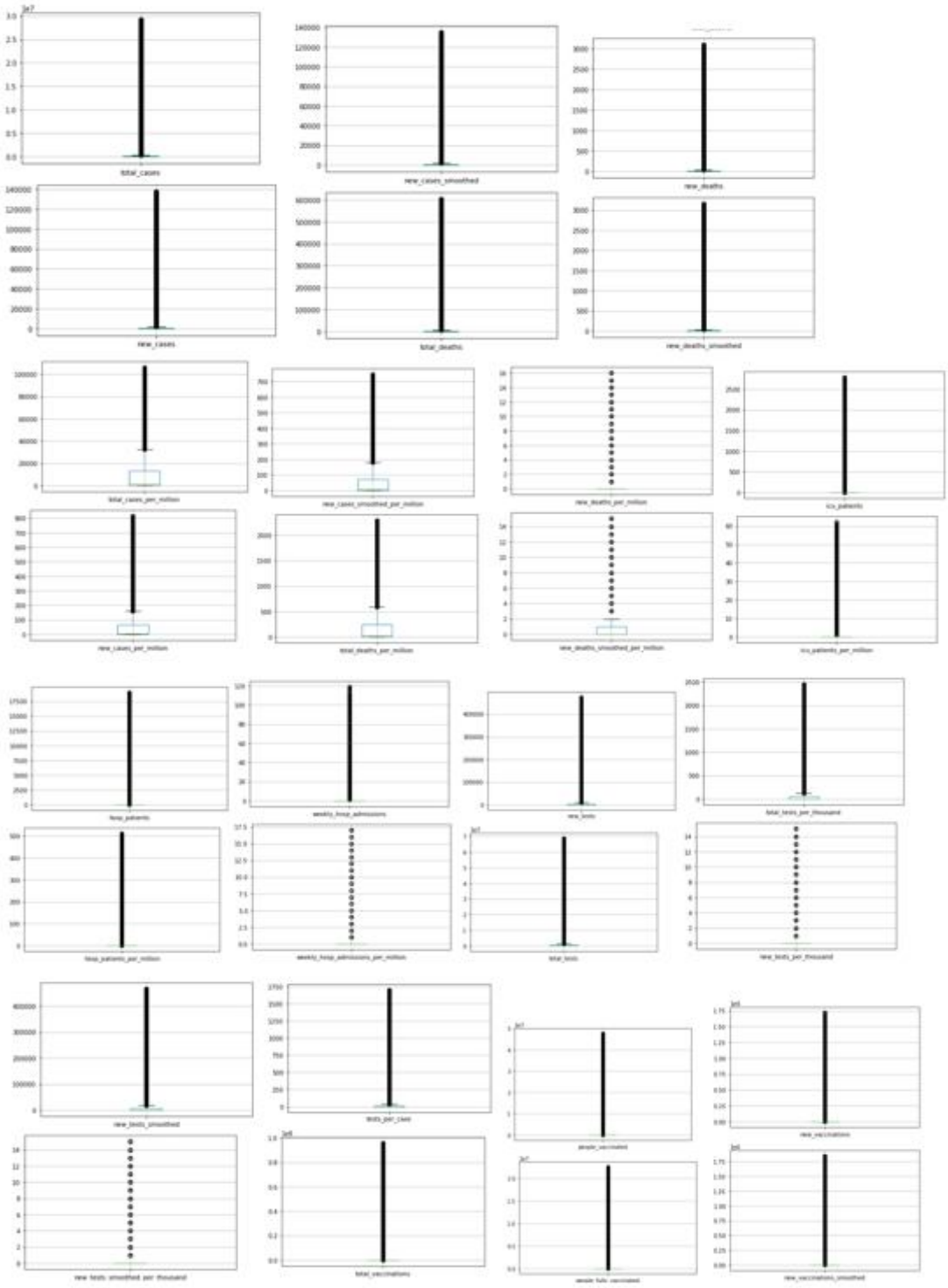


Fig 2 Boxplot of the columns with too many outliers

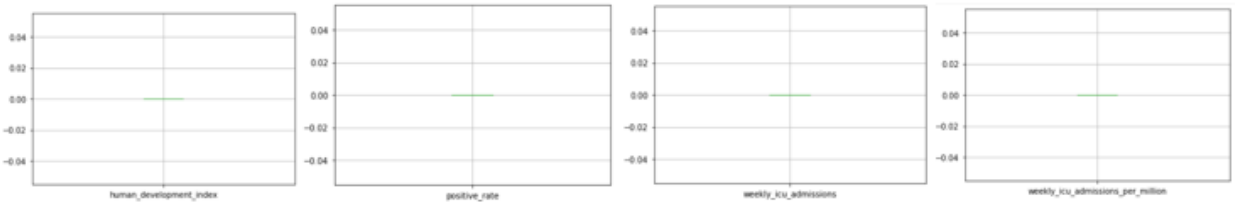


Fig 3 Columns that has box plot

D. Outlier Detection and Treatment

The outlier is detected by Inter-Quartile Range (IQR). IQR is the measure of difference between higher value and lower range. It is explained about the middle half of the data. It is given by the formula, IQR= Q3-Q1, where, Q3- higher value and Q1- lower value. As we told earlier, using IQR totally 37 columns are found to be too more outliers and 4 columns are found to be empty columns. Then we have total of 14 columns in which 13 we considered as independent variable and one dependent variable. From boxplot we detected outliers and we used capping and trimming techniques to remove the outliers. The figure shows how capping and trimming removes outliers.

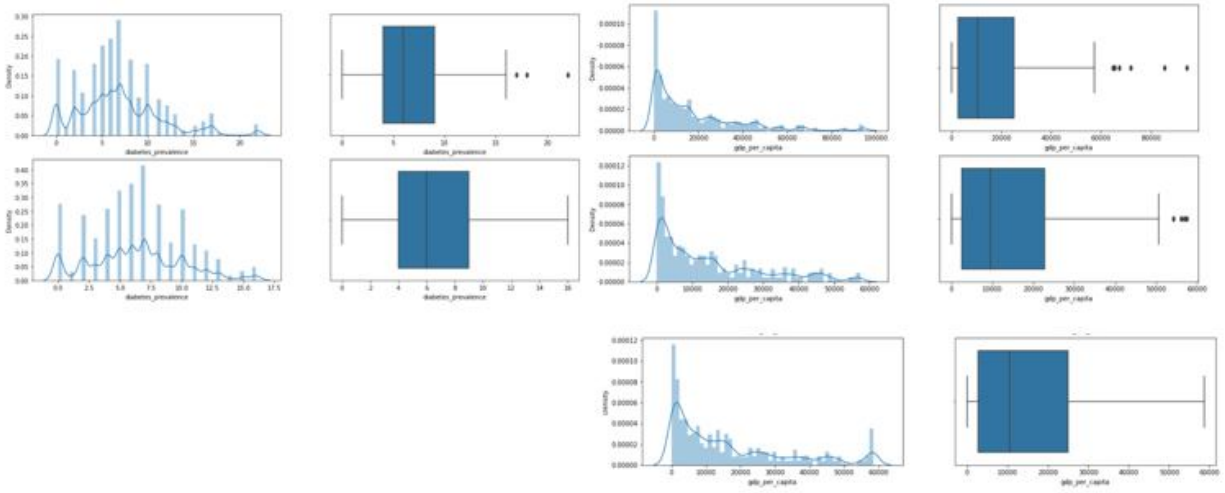


Fig 4 Outliers removal using capping and trimming methods

Reducing the unwanted dimensionality of the data will help in understanding the data easier and helps in easy analysis too.[6]

E. Model Fitting and Evaluation

After removing outliers we did test train split for the available data and we fit and train the data (X,y). After model fitting, we calculated MSE (Mean Square Error) and Cross validation for the regression methods like Random Forest, Linear Regression, Extra tree regression, Lasso regression, Ridge regression Decision tree regression to predict the good fitting model for the complex covid data we used.

Random forest is said to be ensemble-based classification and regression model that uses boot strapping data sampling techniques that divide the data into test and train sets. The selection of its node is based on entropy information, Gini index, gain. Extreme Gradient Boosting (XGB) is another form of gradient algorithm it reduces the risk of overfitting. It is an iterative approach; it can reduce the variance in the predicting model.[8] Linear regression is found to be most popular ML modelling technique which is the relationship between dependent and independent variable. Lasso regression helps to find the predictors subset that reduces the prediction error of a variable whereas ridge regression is used for multi-collinearity and helps in correlating one with other predictors.[3] Extra tree regression uses bagging and Random Forest techniques; it uses combination prediction of many decision trees.

The Mean Squared Error (MSE) explain us how a regression line is close to a set of data points by considering the distances from the data points to the regression line and squaring them to remove any negative signs. It finds the average of set of data points.

$$MSE = 1/n \sum_{i=1}^{n} (Yi - Yp)^2$$

Where, Yi-observed values, Yp - predicted values

Cross Validation means *a* much handy approach for knowing the efficiency of our model, especially in cases where we want to identify and separate overfitting. It is also helpful in knowing the hyper parameters of our model, so that parameters will be resulting in lowest test error. It is otherwise called as K-fold cross validation.

IV. RESULT AND CONCLUSIONS

Using the dynamic data from the open-source website, certain pre - processing steps were applied to clean the data as it is explained in the methods above. After wards certain models were designed on the data to ensure which model fits the best for our data.

The regression models were designed such as Random Forest Regression, Linear Regression, Extra Tree Regressor, Xgboost Regression, Ridge Regression, Lasso Regression, Decision Tree Regression. From all the models we concluded that XgBoost Regression Model fits the best for our dynamic data because it shows the lowest MSE values among all (See Table I)

Table I
Table of MSE and CV score for the models used

S. No.	Regression Model	MSE Score	CV Score
01	Random Forest Regression Model	32569.533	260028.491
02	Linear Regression Model	181391.823	200924.670
03	Extra Trees Regression Model	32387.955	190741.725
04	XgBoost Regression Model	105697.7410	195936.598
05	Ridge Regression Model	187293.645	194524.257
06	Lasso Regression Model	181397.908	200669.246
07	Decision Tree Model	33344.073	427931.561

After performing various 7 types of different models, we concluded that XgBoost Regression model has the lowest MSE values which means it fits the best for our dynamic data.

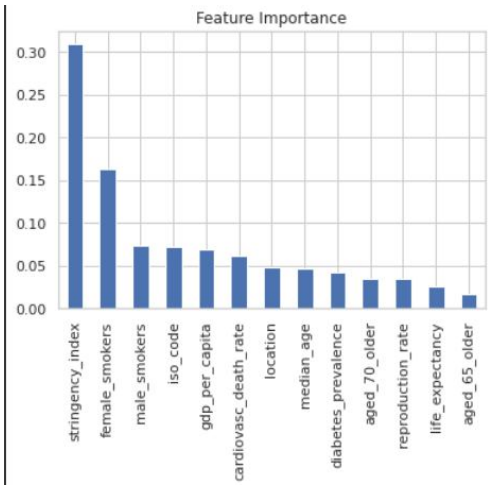


Fig 4 Random Forest Regression Graph

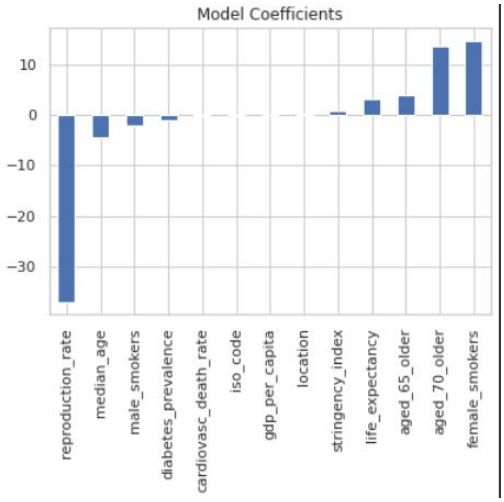


Fig 5 Linear Regression Graph

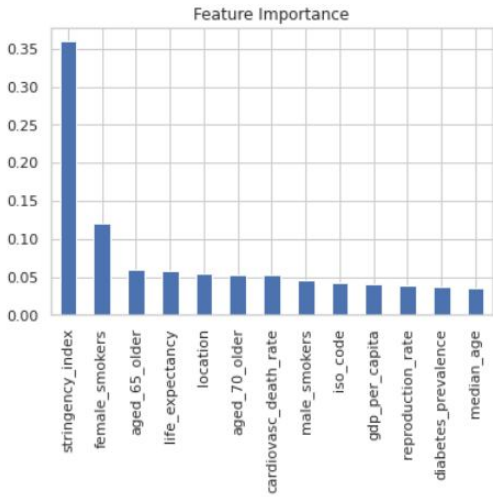


Fig 6 Extra Tree Regressor Graph

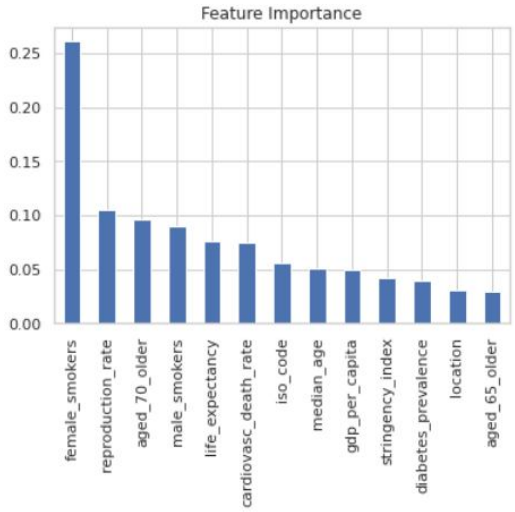


Fig 7 Xgboost Regression Graph

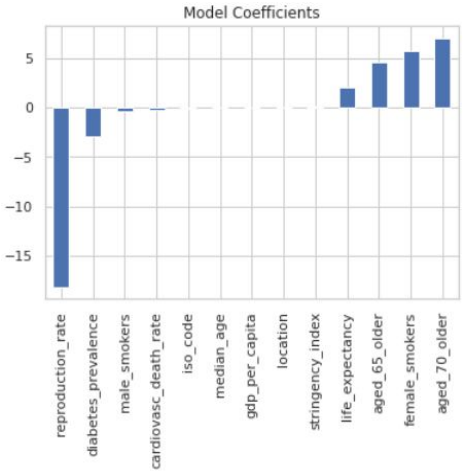


Fig 8 Ridge Regression Graph

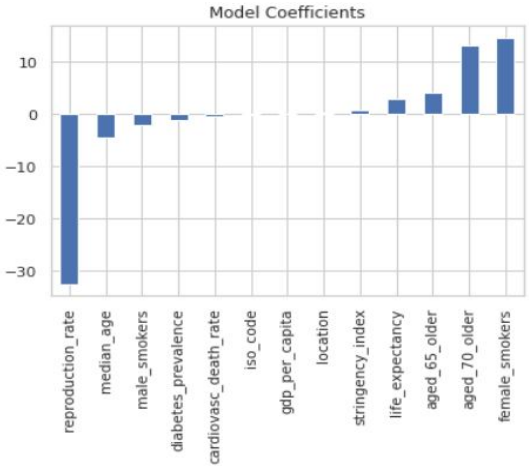


Fig 9 Lasso Regression Graph

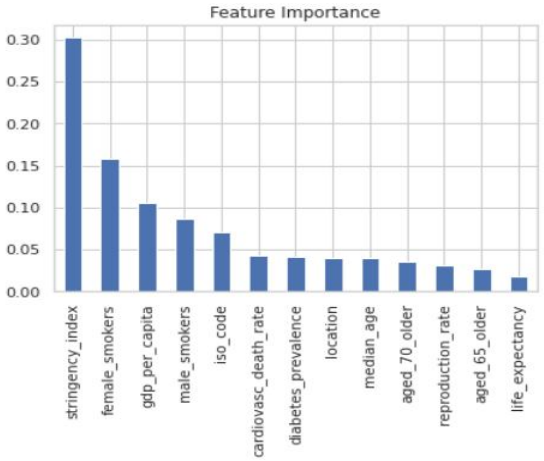


Fig 10 Decision Tree Regression Graph.

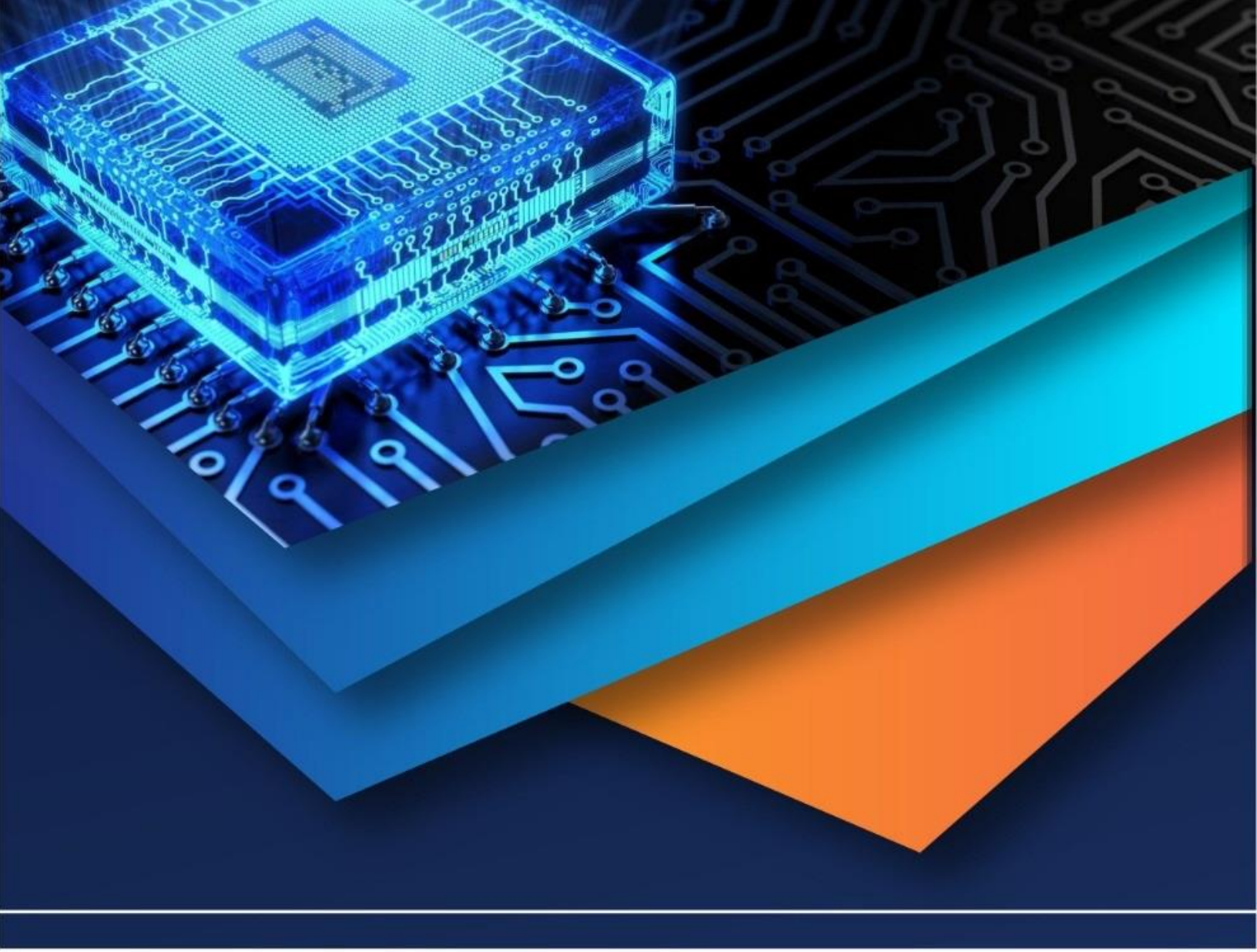
The graph shown above (Fig 4 to Fig 10) tells about the feature importance (otherwise called as Model Coefficients) of the featured columns in descending order manner.

V. DISCUSSION AND FUTURE WORK

The field of Machine learning and its technique is a trending and most needful tool know a day in all sectors. Especially, in health care sector it is too much needed one in-order to analyse the disease growth, its causality and also helpful in knowing the need of antibodies for certain disease, etc. Currently, covid situation seems to be endless crisis and so knowledge of machine learning techniques is too much needed to analyse and work on it to eradicate them. Still some better and fast working algorithms can be developed so that it can analyse large group of data in short computational Time and help researchers, data scientist, data analytics to get desired result soon. Some newly developed algorithm in future may provide better MSE and CV values in our data.

REFERENCES

[1] Nihala Naseefa Chathappady House, Sheeba Palissery," Corona Viruses: A Review on SARS, MERS and covid-19, Microbiology Insights, Volume 14:18, doi:10.1177/11786361211002481.
[2] Pramila P. Shinde, Dr. Seema Shah, "A Review of Machine Learning and Deep Learning applications", 4th International Conference on Computing Communication Control and Automation (ICCUBEA), 2018.
[3] Iqbal H. Saker, "Machine Learning: Algorithms, Real world Applications and Research Directions", SN Computer Science, 2021.
[4] R.Saravanan, Pothula Sujatha, "A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification", Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018.
[5] Yazeed Zoabi, Shira Deri-Rozov, Noam Shomron,"Machine Learning based prediction of Covid-19 diagnosis based on symptoms", npj Digital Medicine, Springer, 2021, 4:3 ; https://doi.org/10.1038/s41746-020-00372-6.
[6] Laxmi Chaudhary, Buddha Singh, "Community detection using unsupervised machine learning techniques on Covid-19 dataset", Social Network Analysis and Mining, Springer,2021.
[7] Heba Elgazzar, Kyle Spurlock, Tanner Bogart," Evolutionary clustering and community detection algorithms for social media health surveillance", Machine Learning with applications, vol.6, 2021.
[8] Sumayh S. Aljameel, Irfan Ullah Khan, Nida Aslam, Malak Alijabri, Eman S. Alsulmi, "Machine Learning Based Model to predict the Disease Severity and Outcome in Covid-19 Patients", Hindawi Scientific Programming, Vol.2021, 2021.
[9] Arun Solanki and Tarana Singh, "COVID-19 Epidemic Analysis and Prediction Using Machine Learning Algorithms", Emerging Technologies for Battling Covid-19, Springer, 2021, pp 57-78.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY