iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** IX **Month of publication:** September 2020

DOI: https://doi.org/10.22214/ijraset.2020.31624

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 8 Issue IX Sep 2020- Available at www.ijraset.com

Covid-19 Analysis and Prediction Model

Adarsh Sharma[1], Shantanu Pingale[2], Chanchal Mal[3], Sangeeta Malviya[4], Nikita Patil[5]
[1, 2, 3, 4, 5]*Department of Information Technology, Vishwakarma Institute of Technology*

Abstract: From past few months we are facing very dangerous air borne disease COVID-19. Even it is normal cold or low fever or high fever now a days peoples cannot predicate whether it is corona virus or not. As the number of patients are increasing day by day, it is not feasible to test all the patients which are now lakhs on numbers. Government is taking all the measures but it is not possible to reach each and everyone. So, through this model we can detect whether a person is having corona virus or not by entering the body temperature, travel history and other necessary information. This model can predicate the forward and backward transmission process in critical situation.

Keywords: Accuracy, Algorithms, Covid-19, Data processing, Graphs.

I. INTRODUCTION

World is moving through a very stressful situation due to spread of novel Corona Virus. The disease is very dangerous and contagious. The world is trying to cope up with the disease but the spread is so fast that WHO (World Health Organization) has declared it as a medical emergency. The first patient in India was reported on 30 January, 2020. But now the spread is in almost every district of India and also in different countries. It is very important to detect this disease as it is contagious and its seriousness is increasing over the period of time. The administration is facing a lot of stress as they are not able to reach everywhere for detecting the virus infected people. So, there must be some tools to detect the cases of COVID-19. This project will help in detecting the patients suffering from novel Corona Virus. The project analyses the input like body temperature, any other diseases, travel history, etc.

II. LITERATURE REVIEW

COVID-19 is an infectious disease caused by a newly discovered coronavirus. The people who are infected from this will show some symptoms like cough, fever, etc. Older people who are already having some problems like blood pressure, diabetes, asthma, etc are finding it difficult to get recover from the disease. At this time, there is no specific treatment for COVID-19. So, it because very important to detect this disease as it is contagious and also its spread is very fast.

This model will help to detect the infected people. There is also some research done on detecting COVID-19 patients. Machine learning is widely used and accepted method now-a-days. This method has changed the perspective of diagnosis by giving great results to diseases like diabetes and epilepsy. Akib Mohi Ud Din Khanday, Syed Tanzeel Rabani, Qamar Rayees Khan, Nusrat Rouf & Masarat Mohi Ud Din has proposed a paper for detecting corona virus using Text Mining.M. Rubaiyat Hossain Mondal,Subrato Bharati,Prajoy Podder, PriyaPodder also have done research on novel corona virus disease using data analytics.Gitanjali R. Shinde, Asmita B. Kalamkar, Parikshit N. Mahalle, Nilanjan Dey, Jyotismita Chaki & Aboul Ella Hassanien have catagorise different data sets for detection of virus and other diseases by studying different forecasting models. Sina F. Ardabili, Amir Mosavi, Pedram Ghamisi, Filip Ferdinand, Annamaria R.Varkonyi-Koczy, Uwe Reuter, Timon Rabczuk, Peter M. Atkinson from germany also have done some work on the COVID-19 outbreak prediction using machine learning by using some algorithms like GA algorithm.

III.METHODOLOGY

We have used Python Programming language to do machine learning and data science using Jupyter Notebook.

A. Libraries of Python Used

1) *NumPy library:* It stands for Numerical Python, is a library consisting of multidimensional array objects and collection of routines for processing those arrays. By using NumPy mathematical and logical operations in array can be performed.
2) *Pandas:* It is a software library written for python programming language for data manipulation and analysis.
3) *Matplotlib:* It is a graph plotting library for the python programming language and its numerical mathematics extension NumPy.
4) *Seaborn:* It is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attritive and informative statistical graphs.
5) *Warnings:* To ignore all warnings which might be showing up in the notebook due to future depreciation of a feature.
6) *Train_test_split:* To split the dataset into training and testing dataset.
7) *Standard Scaler:* To scale all the features, so that Machine Learning model better adapts to the dataset.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IX Sep 2020- Available at www.ijraset.com

B. Algorithm Used

1) *K Neighbor's Classifier:* This algorithm is used to solve the classification model problems. K-N-N algorithm basically creates an imaginary boundary to classify the data. When new data points, come in, the algorithm will try to predict that to the nearest to the boundary line.
2) *Super Vector Machine Classifier:* This algorithm is representation of examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. SVM can efficiently perform a non-linear classification, implicitly mapping their inputs into high dimensional feature specs.
3) *Random Forest Classifier:* This algorithm is robust machine learning robust algorithm that can be used for variety of tasks including regression and classification.
4) *Decision Tree Classifier:* This algorithm is a simple representation for classifying examples. It is a Supervised Machine learning where the data is continuously split according to the certain parameters.

C. Parameters Taken in Dataset

The Parameters included in our dataset are age, gender, fever, trestbps, respiration rate, diabetes, hypertension, pulse rate, lung disease, difficulty in breathing, travelled outside India, travel red zone area, senior citizen and target.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 606 entries, 0 to 605
Data columns (total 14 columns):
age              606 non-null int64
gender           606 non-null int64
fever            606 non-null int64
trestbps         606 non-null int64
resp             606 non-null int64
diabetes         606 non-null int64
hypertension     606 non-null int64
PR               606 non-null int64
lungdis          606 non-null int64
dinbr            606 non-null int64
travelledoi      606 non-null int64
travelledrza     606 non-null int64
seniorcitizen    606 non-null int64
target           606 non-null int64
dtypes: int64(14)
memory usage: 66.4 KB
```

Figure 1. Parameters in the dataset

D. Understanding the Dataset

1) *Correlation Matrix*: This matrix will help us to see the correlation between the target and the parameters of the dataset.

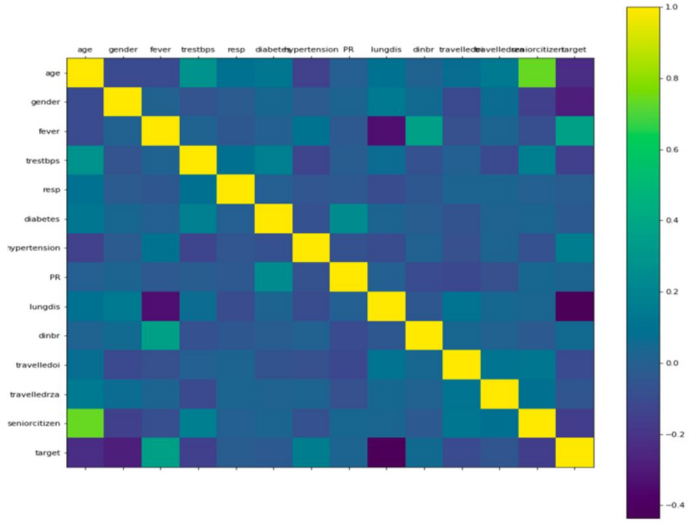


Figure 2. Correlation Matrix between Target and Parameters

2) *Histogram Graph Plotting:* It shows how much each feature and label distributed along different ranges, which further confirms the need for scaling. This all are categorical variable which need to be handled before applying machine learning algorithms.

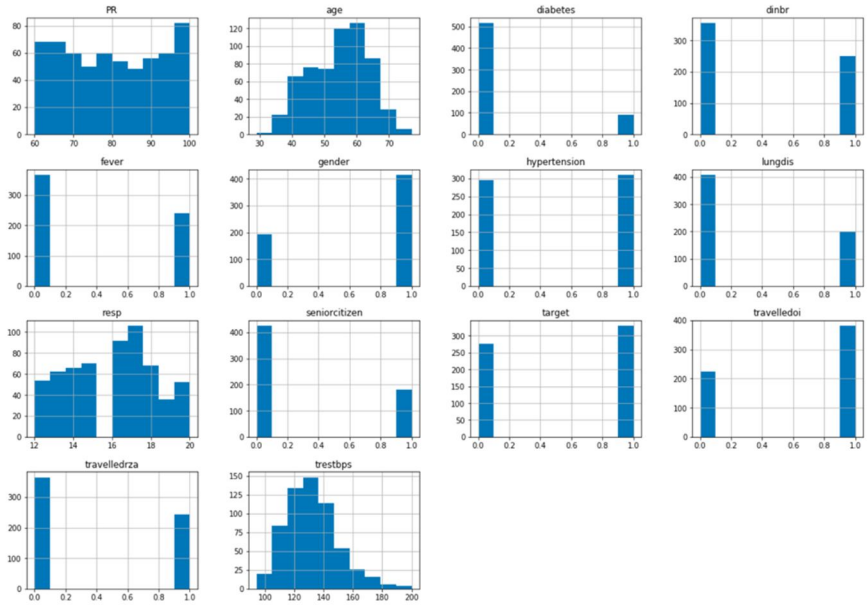


Figure 3. Histogram Graph of each parameter.

3) *Bar Plot for Target Class*

```
1    330
0    276
Name: target, dtype: int64
```

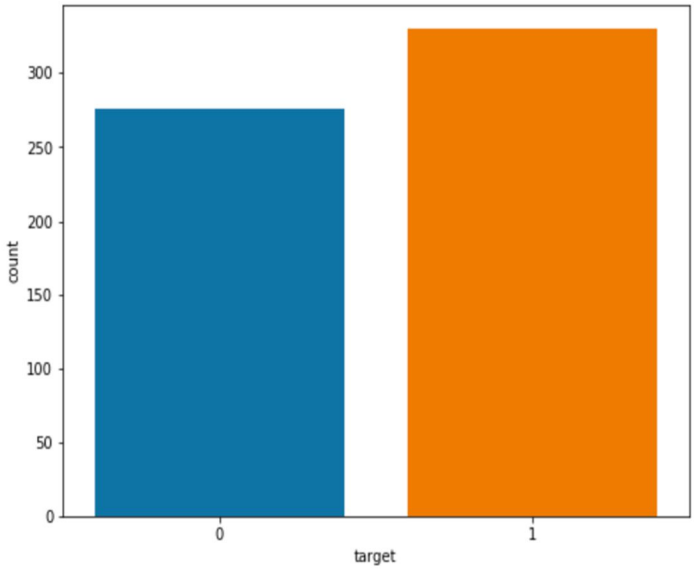


Figure 4. Represents the target with respect to the no. of counts where Target 1 = Infected and Target 2 = Non- Infected.

We can see that the classes are almost balanced, and we can go for the data processing.

4) Graph: Density of disease True or False with respect to ages and Disease Probability with Respect to Ages

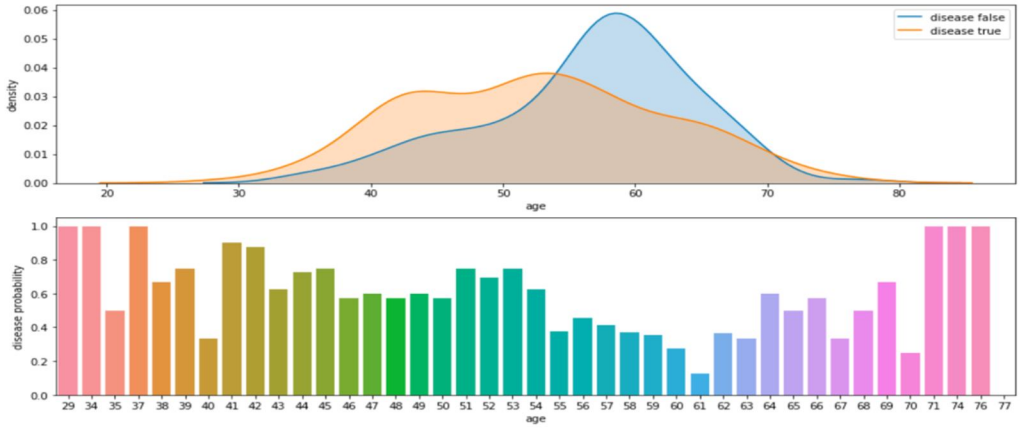


Figure 5. Indicates the density of disease true or false and disease probability with respect to ages.

E. Data Processing

To work with categorical variable, we should break each into dummy columns which can only has value 0s and 1s.
0 for False
1 for True.
To get this done the get_dummies() method from Pandas. Next, we need to scale the dataset for which we file transform method of the scaler scales the data and we can update the columns.

```
In [71]:
dataset = pd.get_dummies(dataset, columns = ['gender', 'fever', 'diabetes', 'hypertensi
on', 'lungdis', 'dinbr', 'travelledoi', 'travelledrza', 'seniorcitizen'])

In [73]:
standardScaler = StandardScaler()
columns_to_scale = ['age', 'trestbps', 'PR', 'resp']
dataset[columns_to_scale] = standardScaler.fit_transform(dataset[columns_to_scale])
```

Figure 5. We have separated the Dummy Columns and Scaled Columns.

F. Machine Learning

We split the dataset into two parts as 67% training data and 33% testing data.
For Analysis and Prediction used Four Major Algorithm of Machine Learning.

1) K Neighbor's Classifier:

This classifier looks for the classes of K nearest neighbors of a given data point and based on the majority class, it assigns a class to this data point.
Then, I plot a line graph of the number of neighbors and the test score achieved in each case.

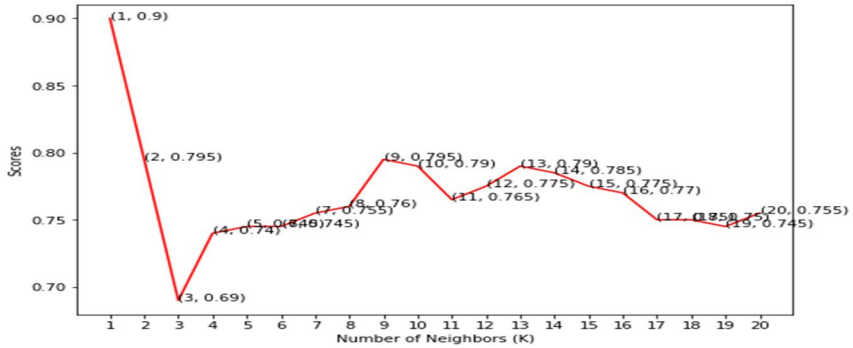


Figure 6. K Neighbors Classifier scores for different K values.

The score for K Neighbors Classifier is 90.0% with [1] nieghbors.

2) *Support Vector Classifier:* This classifier aims at forming a hyperplane that can separate the classes as much as possible by adjus ting the distance between the data points and the hyperplane. There are several kernels based on which the hyperplane is decided. We tried four kernels namely, linear, poly, rbf, *and* sigmoid.

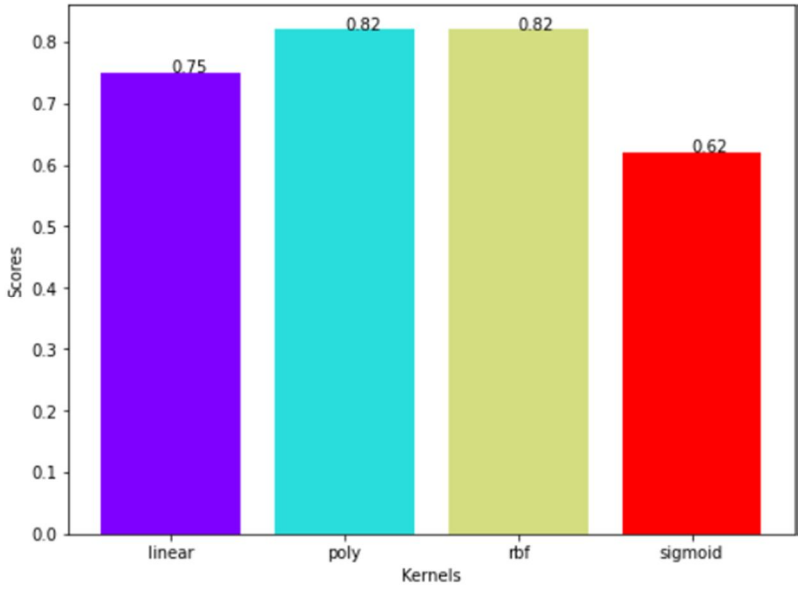


Figure 7. Support Vector Classifier scores for different kernels.

As can be seen from the plot above, the poly kernel performed the best for this dataset and achieved a score of 82%.

3) *Decision Tree Classifier:* This classifier creates a decision tree based on which, it assigns the class values to each data point. Here, we can vary the maximum number of features to be considered while creating the model. I range features from 1 to 22 (the total features in the dataset after dummy columns were added).

Once we have the scores, we can then plot a line graph and see the effect of the number of features on the model scores.

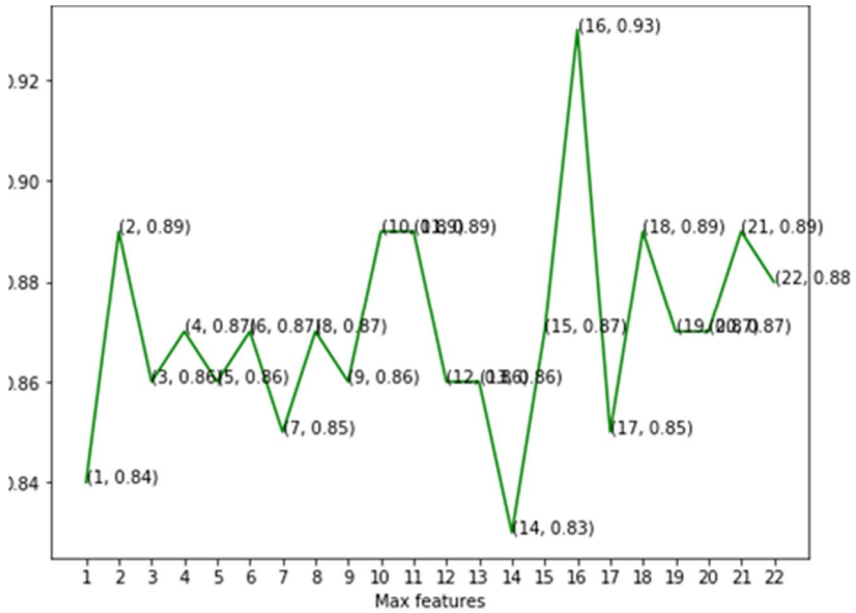


Figure 8. Decision Tree Classifier for different number of maximum features.

From the line graph above, we can clearly see that the maximum score is 93% and is achieved for maximum features being selected to be [16].

4) *Random Forest Classifier:* This classifier takes the concept of decision trees to the next level. It creates a forest of trees where each tree is formed by a random selection of features from the total features. Here, we can vary the number of trees that will be used to predict the class.

We calculate test scores over 10, 100, 200, 500 and 1000 trees.

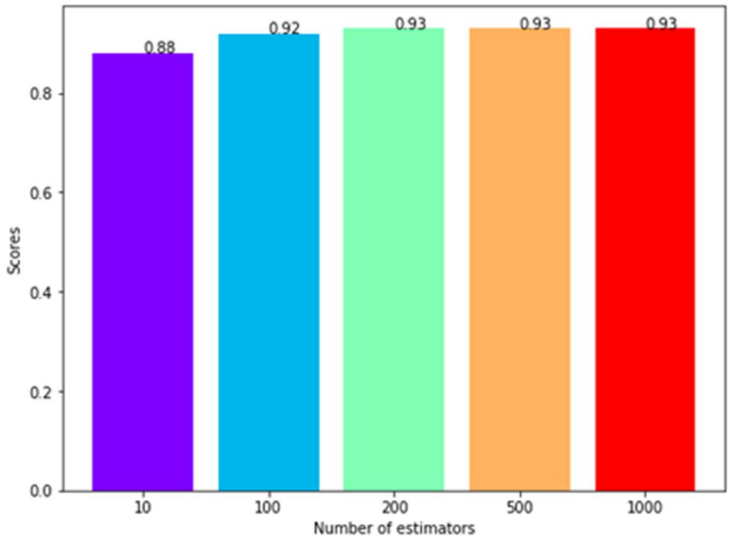
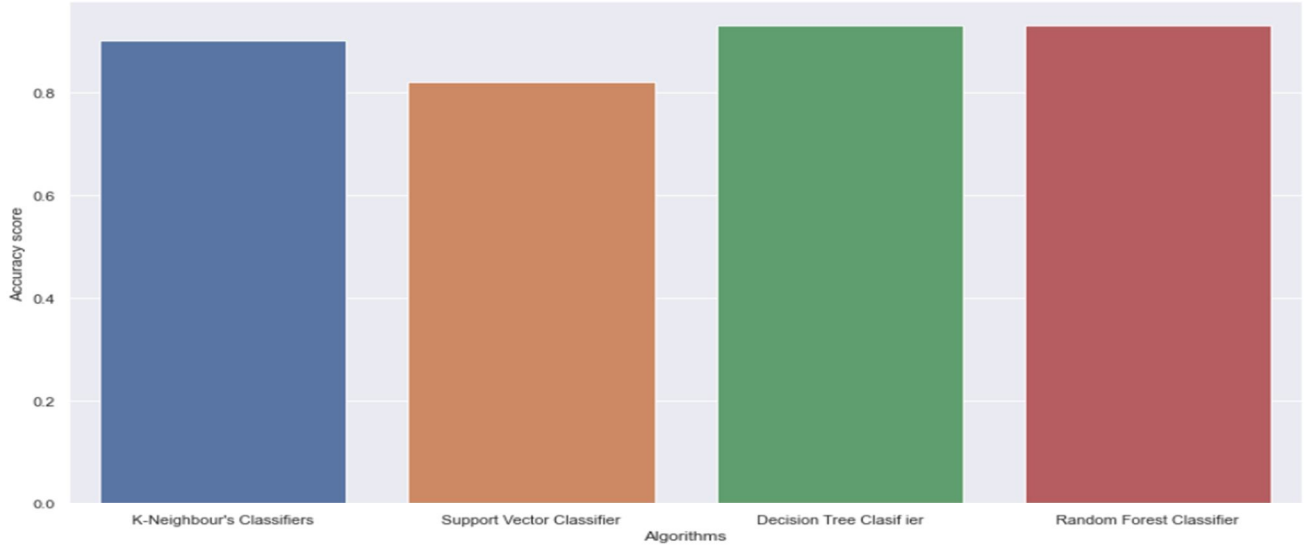


Figure 9. Random Forest Classifier for different number of estimators

Taking a look at the bar graph, we can see that the maximum score of 93% was achieved for 200, 500 and 1000 trees.

IV. RESULT

The graph given below tells about the best algorithm for our model by comparing their accuracy scores.




```
print("The Accuracy for K-Neighbours's Classifier is {}%".format(knn_scores[0]*100))
print("The Accuracy for Super Vector Classifier is {}%".format(svc_scores[1]*100))
print("The Accuracy for Decision Tree Clasifier is {}%" .format(dt_scores[15]*100))
print("The Accuracy for Random Forest Clasifier is {}%" .format(rf_scores[2]*100))
```

```
The Accuracy for K-Neighbours's Classifier is 90.0%
The Accuracy for Super Vector Classifier is 82.0%
The Accuracy for Decision Tree Clasifier is 93.0%
The Accuracy for Random Forest Clasifier is 93.0%
```

So, after comparing the four most popular algorithm of Data Science Machine Learning. We can say that Decision Tree Classifier and Random Forest Classifier is Best for our Model with an Accuracy of 93%.

V. CONCLUSION AND DISCUSSION

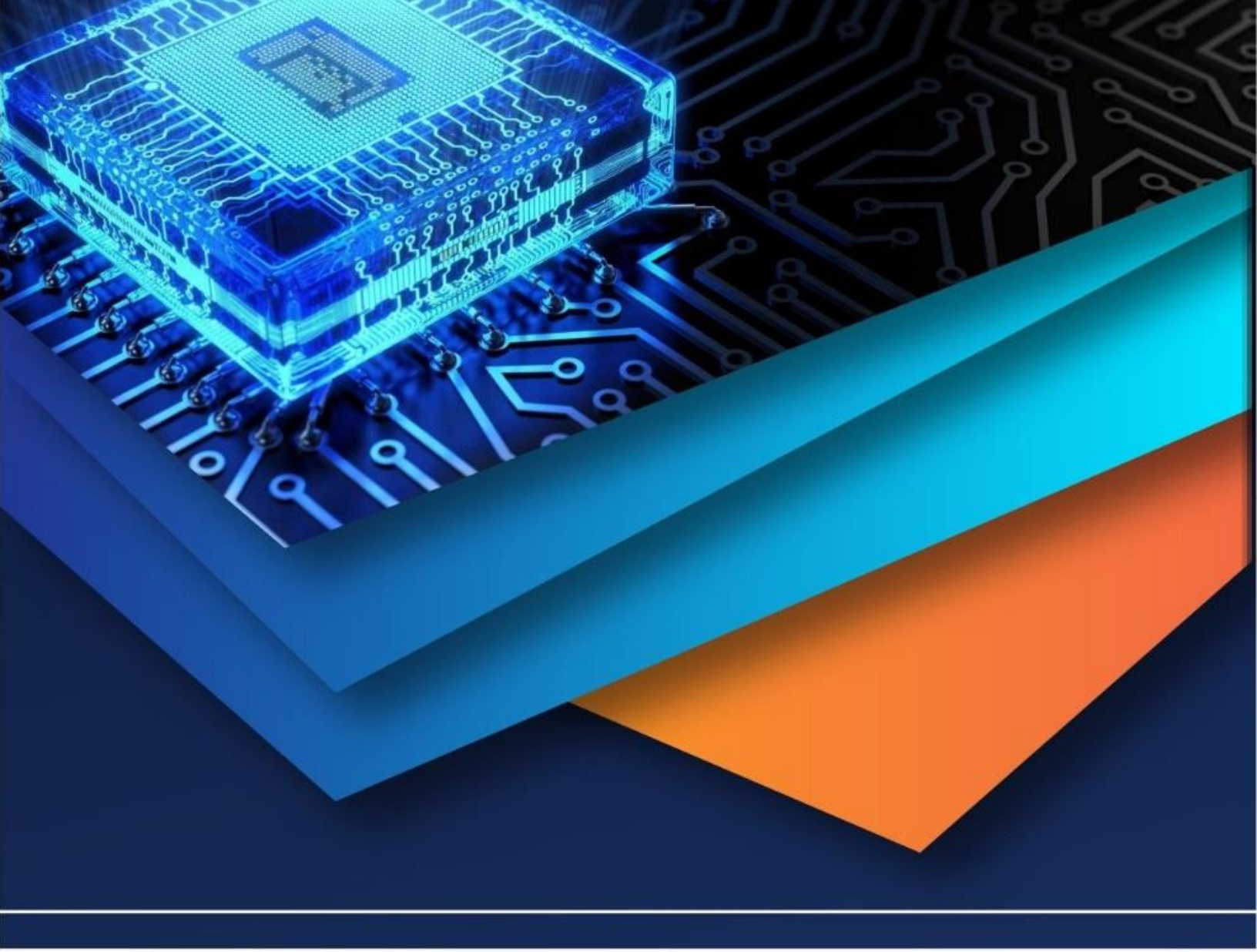
The day-by-day increase in the death rate due to the pandemic has put the normal life on hold. By using this tool, it will be easier to detect the infected people from the Corona Virus. The four algorithms K-Nearest Classifier, Super Vector Classifier, Decision Tree Classifier and Random Forest Classifier are used in this project. By going through the accuracy values for these different algorithms, a conclusion can be made that Decision Tree Classifier and Random Forest Classifier are better than the other two as these algorithms return 93% accuracy rate. This project can be a very useful contribution to the society.

VI. FUTURE SCOPE

This model predicates day to day infected cases .The occurrence of disease is clear. Short-term projection of new cases is done. Analysis of model can be done easily. To assess correlation between predicted and actual values of cases can be done easily. Accuracy can be increased using dataset. As dataset used of coronavirus this model can be also trained in other disease(dataset).

REFERENCES

[1] Machine Learning methods to aid in Coronavirus Response at https://towardsdatascience.com/machine-learning-methods-to-aid-in-coronavirus-response-70df8bfc7861
[2] Machine learning : answer to coronavirus at https://www.analyticsinsight.net/machine-learning-answer-coronavirus/
[3] Novel Corona Virus 2019 Dataset at https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
[4] Covid-19 pandemic in India at https://en.m.wikipedia.org/wiki/COVID-19_pandemic_in_India#:~:text=The%20COVID%2D19%20pandemic%20in,reported%20on%2030%20January%202020.
[5] Forecasting Models for Coronavirus Disease (COVID-19) at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7289234/
[6] Machine learning based approaches for detecting COVID-19 using clinical text data at https://link.springer.com/article/10.1007/s41870-020-00495-9
[7] Data analytics for novel coronavirus disease at https://www.researchgate.net/publication/342195015_Data_analytics_for_novel_coronavirus_disease
[8] COVID-19 Outbreak Prediction with Machine learning at https://www.medrxiv.org/content/10.1101/2020.04.17.20070094v1.full.pdf



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY