e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:01/January-2023 Impact Factor- 6.752 www.irjmets.com

FIND THE BEST ALGORITHM AMONG DIFFERENT MACHINE LEARNING MODULES USING COVID-19 DATA OF MAHARASHTRA

Pratiksha Jagdhane[*1]

[*1]Student, Vidya Pratishthan's Kamalnayan Bajaj Institute Of Engineering And Technology, Baramati, Maharashtra, India.

ABSTRACT

When COVID-19 hit India of the total reported 10 million cases, approximately 2 million are from Maharashtra, which has many crowded cities. The first COVID-19 pandemic case in the Indian state of Maharashtra was confirmed on 9 March 2020. The drastic increase in cases lead to an unsettling situation for the healthcare system. About half of the cases in the state came up from the Mumbai Metropolitan Region (MMR). This proposed work, Includes data on COVID-19 cases reported in different districts in Maharashtra to carry out this analysis to find appropriate models that can be used to predict future cases. Out of the various classification models, Naive Bayes, Logistic regression, K-Nearest Neighbor, and Support Vector Machine clustering have been employed for the classification of positive and recovered cases.

Method : The work aims to analyze which district in Maharashtra was most affected by COVID-19 cases and what was the recovery and fatality rate in each district. (2020) The project analysis also aims to appropriate models among support vector machine, linear regression, and K-nearest algorithm.

Keywords: COVID-19, Prediction Recovers Cases, Maharashtra, Positive Cases.

I. INTRODUCTION

The coronavirus caused an outburst that was first recognized in Wuhan City, China. Since then, the virus has spread to nearly every country across the world, leading the World Health Organization (WHO) to declare this a pandemic. The name “coronavirus” was derived from the crown-like projections on their surfaces. “Corona” word in Latin means “halo” or “crown”.

This virus spread all across the world and everyone was affected by the current COVID-19 pandemic. However, the impact of the, as well as consequences of the pandemic, it differently depending on the individual's status and as members of society. Many people lost their jobs while some try to acclimate to working online, homeschooling their children, and ordering food via the online platform, others have no choice but to be disclosed to the virus while keeping society functioning.

As of May 8th, 2020, in India, 56,342 positive cases were reported. India a country with a population of more than 1.30 billion and the country with the second largest population in the world faced difficulty in controlling the transmission of extreme acute respiratory syndrome coronavirus 2 among its population. Multiple strategies were implied to handle the outbreak; which included computational modeling, statistical tools, and quantitative analyses to control the spread and the rapid development of a new treatment. The Ministry of Health and Family Welfare of India raised awareness about the recent outbreak and took necessary actions to prevent the spread of COVID-19. The central and state governments were taking various measures and formulating several wartime protocols to accomplish the goal. Moreover, the Indian government executed 55 days lockdown throughout the country that started on March 25th, 2020, to reduce the spread of the virus. (Singhal, 2020) This outbreak is inseparably linked to the economy of the nation, as it has dramatically hindered industrial sectors because people worldwide are currently cautious about employing businesses in the concerned regions.

II. METHODOLOGY

The objective of the work is to analyze how many people were affected by the COVID-19 pandemic and how many people were successfully able to fight against this virus and recovered themselves using machine learning techniques.

Here the aim is to find the best model for the analysis of future predictions among support vector machine, linear regression, Naive Bayes, and K-nearest algorithm by calculating the accuracy of each module.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:01/January-2023 Impact Factor- 6.752 www.irjmets.com

The Data used for this analysis purpose contains statistical data of positive cases, active cases, and recovered cases of all the districts in Maharashtra.

Districts	Positive Cases	Active Cases	Recovered	Deceased	Recovery Rat	Fatality Rate (%)
Ahmednagar	338227	2338	328882	7006	97.2	2.1
Akola	58759	23	57307	1425	97.5	2.4
Amravati	96226	12	94618	1594	98.3	1.7
Aurangabad	155149	433	150452	4250	97	2.7
Beed	103652	123	100721	2801	97.2	2.7
Bhandara	60079	2	58944	1123	98.1	1.9
Buldhana	85501	14	84685	796	99	0.9
Chandrapur	88953	57	87332	1560	98.2	1.8
Dhule	46165	4	45496	654	98.6	1.4
Gadchiroli	30438	7	29729	669	97.7	2.2
Gondia	40517	3	39938	569	98.6	1.4
Hingoli	18474	24	17943	506	97.1	2.7
Jalgaon	139924	15	137163	2714	98	1.9
Jalna	60594	50	59334	1209	97.9	2
Kolhapur	206594	176	200568	5845	97.1	2.8
Latur	92084	61	89584	2433	97.3	2.6
Mumbai	753653	5328	729621	16202	96.8	2.1
Nagpur	493559	72	484288	9128	98.1	1.8
Nanded	90376	16	87696	2658	97	2.9
Nandurbar	40004	0	39053	948	97.6	2.4
Nashik	410142	632	400846	8663	97.7	2.1
Osmanabad	67688	191	65421	1962	96.7	2.9
Palghar	137714	505	133913	3282	97.2	2.4
Parbhani	52347	27	51069	1232	97.6	2.4
Pune	1152592	6955	1125722	19566	97.7	1.7
Raigad	195567	777	190242	4541	97.3	2.3
Ratnagiri	78828	264	76086	2473	96.5	3.1
Sangli	209554	648	203289	5608	97	2.7
Satara	250141	749	242954	6407	97.1	2.6
Sindhudurg	52655	508	50702	1430	96.3	2.7
Solapur	210099	500	203955	5534	97.1	2.6
Thane	608496	3609	593438	11414	97.5	1.9
Wardha	57336	5	55949	1217	97.6	2.1
Washim	41659	4	41015	637	98.5	1.5
Yavatmal	75960	6	74152	1798	97.6	2.4

Figure 1: Data Used for Analysis

Data Collection

The experimental data is collected from Kaggle. The data set is in the Comma Separated Values (CSV) file format. The data set consists of seven features those are Districts, Positive Cases, Active Cases, Recovered, Deceased, Recovery Rate(%), and Fatality Rate(%).

Data Preprocessing

Data are preprocessed to remove Null values, noisy data, and irrelevant data because data are generally incomplete i.e they have lacking attribute values, and also lack certain attributes of interest, or contain only aggregate data.

Data preprocessing is an important data mining technique that involves transforming raw data into an understandable format to get useful information. Real-world data is sometimes incomplete, unpredictable, inconsistent, and/or lacking in certain trends or behaviors, and is likely to contain many errors. Data preprocessing is a proven method of fixing such issues to find useful information.

Data Cleaning

The concept of missing values is crucial to understand to successfully manage the provided data. Due to inappropriate handling, the result obtained by the researcher will differ from the ones where the missing values are present.

Two ways to handle Missing Values

• To handle the null/invalid values. we either delete a particular row if it has a nullified value for a particular feature and a respective column if it has more than 75% of missing values.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:01/January-2023 Impact Factor- 6.752 www.irjmets.com

dataset.dropna(inplace=True) dataset.isnull().sum()

• This strategy can be applied to a feature that has numeric data like the year column or the House team goal column. We can calculate the mean, mode, or median of the features and replace it with the missing values.

Analysis

From the data, it is clear that the most affected districts in Maharashtra are Pune and Mumbai whereas the least impacted district was Hingoli. So people living in high crowed cities were affected more. Also from the data, it can be inferred that the recovery rate is 97.96% and the 2.04% fatality rate. (D) Using this information we are going to find the best machine learning classifier for this analysis which is among the four modules mentioned in the next section.

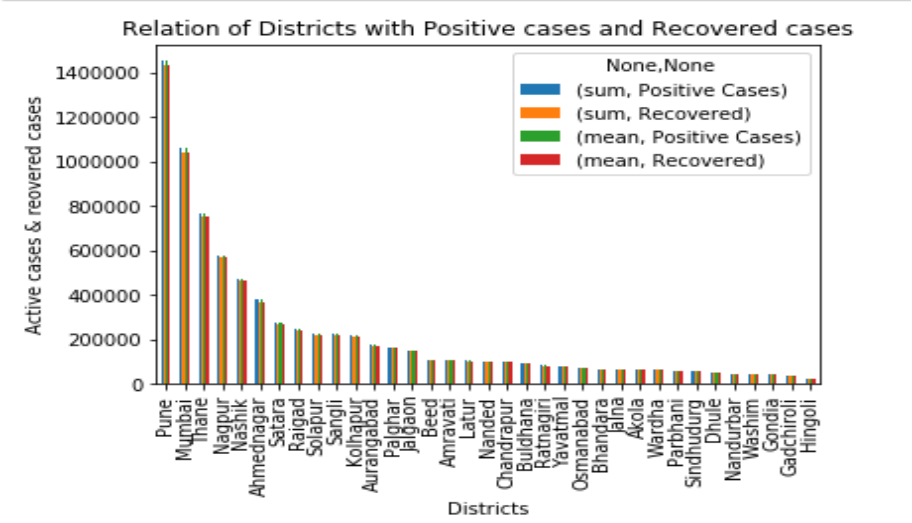

Figure 2: Relation of districts with positive and negative cases.

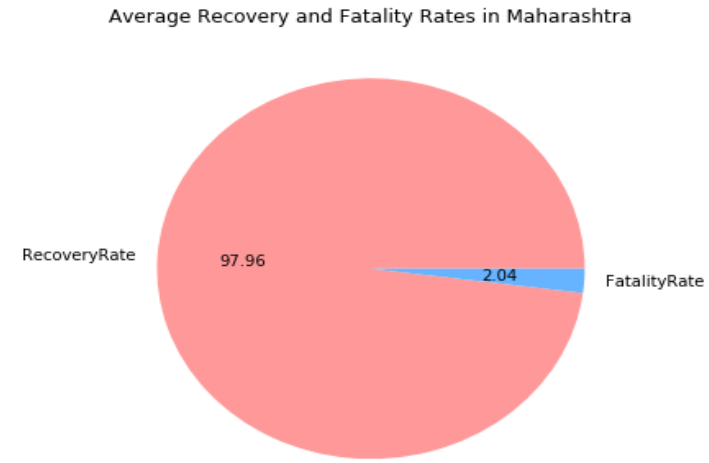

Figure 3: Average Recovery and Fatality Rate

III. MODELING AND ANALYSIS

Classification is a supervised learning method in machine learning where the model tries to predict the correct label of a given input data. In classification, the model is completely trained using the training data, and then the model is evaluated on test data before being used to perform prediction on new unseen data. (al K. A., 2020) Classes in given data are sometimes called labels or targets or categories. Classification in predictive modeling is the task of the corresponding mapping of function (f) from input variables that is (X) to discrete output variables that are (y).

Following are the types of Classification Algorithms in Machine Learning.

• Logistic Regression

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:01/January-2023 Impact Factor- 6.752 www.irjmets.com

- Naive Bayes
- Support Vector Machine
- K-Nearest Neighbour

Support Vector Machine

Support Vector Machine or SVM is the most famous Supervised Learning algorithm in machine learning, which is used for Regression and Classification problems. Nevertheless, the primary role of SVM is the Classification of problems in Machine Learning.

The purpose of the Support Vector Machine algorithm is to assemble the decision boundary or best line that can separate n-dimensional space into classes so that we can simply put the latest data point in the accurate category in the future. This best line or the decision boundary is called a hyperplane. Support Vector Machine selects the outermost points/vectors that support creating the hyperplane. These severe cases are called support vectors, and therefore the algorithm is known as a Support Vector Machine.

Logistic Regression

Linear regression is a most popular and straightforward Machine Learning algorithm. Linear regression is a statistical approach that is used in machine learning for predictive analysis. It is used to predict real /continuous numeric variables such as age, salary, product price, etc.

Linear regression algorithm demonstrates a linear connection between a dependent variable (x) and one or more independent variables (y), therefore called linear regression. As the linear regression algorithm shows the linear relationship, this means it encounters how the value of the dependent variable (x) is changing in accordance with the value of the independent variable (y).

Gaussian Naive Bayes Classifier

Naive Bayes classification technique based on Bayes Theorem with an assumption of independence among predictors. This classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. In Naive bayes, even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability. Naive Bayes is simple to implement and particularly useful for very large data sets. Naive Bayes, Along with simplicity is known to outperform even highly sophisticated classification methods.

K-Nearest Neighbour

K-Nearest Neighbour (K-NN) in Machine Learning algorithms is one of the easiest algorithms based on the Supervised Learning technique.

The k-NN algorithm supplies all the known data and organizes a new data point based on the likeness. This implies that when new data is found then it can be smoothly classified into a good suite category by using K- NN algorithm.

The k-NN algorithm considers the similarity between the new data and available cases and put the new case into the category that is equivalent to the available categories.

K-NN does not make any assumption on underlying data and which means it is a non-parametric algorithm.

The algorithm at the training phase just holds the dataset and when it obtains new data, then it organizes that data into a category that is much identical to the new data

K-NN algorithm is also known as the lazy learner algorithm as it does not learn from the training set directly rather it stores the dataset and at the moment of classification, it conducts an action on the dataset.

Analyze Confusion Matrix

This matrix is used to make the in-depth analysis of statistical data faster and the results easier to read through clear data visualization. The tables can help analyze faults in data mining and statistics. The analysis helps users decide what results indicate how errors are made rather than merely assessing performance. Confusion matrices use a simple format to log predictions. In rows of a confusion matrix for a machine learning model, the possible predictions are aligned on the right-hand side, and the actualities are along the top. Results of the matrix can include the correct indication of a positive as a true positive or a negative as a true negative, as well as an incorrect positive as a false positive or an incorrect negative as a false negative.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:01/January-2023 Impact Factor- 6.752 www.irjmets.com

```
[1 0 0 0 3 0 1 0 0]
[[3 1 0 0 0 0]
 [0 1 0 0 0 0]
 [1 0 0 0 0 0]
 [0 0 0 1 0 0]
 [1 0 0 0 0 0]
 [1 0 0 0 0 0]]
              precision    recall  f1-score   support

           0       0.50      0.75      0.60         4
           1       0.50      1.00      0.67         1
           2       0.00      0.00      0.00         1
           3       1.00      1.00      1.00         1
          13       0.00      0.00      0.00         1
         587       0.00      0.00      0.00         1

    accuracy                           0.56         9
   macro avg       0.33      0.46      0.38         9
weighted avg       0.39      0.56      0.45         9
```

Figure 4: Confusion matrix for Naïve Bayes classification

IV. RESULTS AND DISCUSSION

After testing and training the dataset, their predictive accuracy is determined to find out which model is the best classifier for classifying the predictions accurately. As shown below, the Naive Bayes classifier model has the best predictive accuracy among the four models and the Logistic regression classifier is the least predictive accuracy.

The accuracy of the four classifiers is as shown below :

```
Accuracy of Logistic regression classifier on training set: 0.38
Accuracy of Logistic regression classifier on test set: 0.22

    Accuracy of GNB classifier on training set: 0.58
    Accuracy of GNB classifier on test set: 0.56

   Accuracy of K-NN classifier on training set: 0.46
   Accuracy of K-NN classifier on test set: 0.44

Accuracy of Support Vector Machine classifier on training set: 0.77
Accuracy of Support Vector Machine classifier on test set: 0.33
```

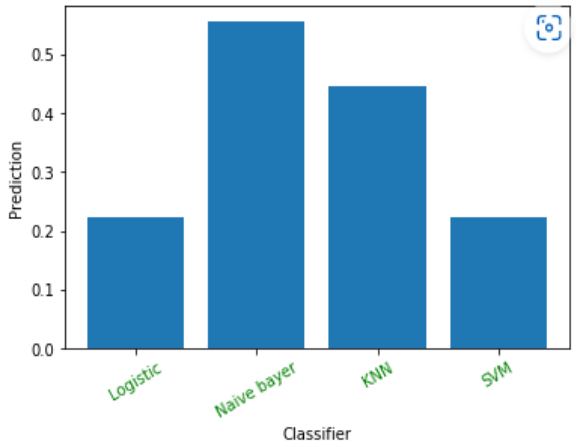

Figure 5: Classifier results graph

V. CONCLUSION

From the analysis, it is clear that the most affected districts in Maharashtra are Pune and Mumbai whereas the least impacted district was Hingoli. So people living in highly crowded cities were affected more and have to be careful as the thread of the pandemic isn't over yet. Moreover, it can be inferred that the recovery rate is 97.96% and the 2.04% fatality rate. After testing and training the dataset, their predictive accuracy is determined to find out which model is the best classifier for classifying the predictions accurately. As shown

above, the Naive Bayes classifier model has the best predictive accuracy among the four models and the Logistic regression classifier is the least predictive accuracy. The Naïve Bayers algorithm can be used for future predictions.

VI. REFERENCES

[1] Al, K. A. (2020). Machilearning-basedsed approaches for detecting COVID-19 using clinical text data. Int J Inf Technol 12(3):731–739.

[2] D, A. (2020). Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Intern Emerg Med 15(8):1435–1443.

[3] L, B. (2020). Machine learning for coronavirus COVID-19 detection from chest x-rays. Proced Comput Sci 176:2212–2221.

[4] Singhal, T. (2020). A review of coronavirus disease-2019 (COVID-19). The Indian Journal of Pediatrics pages 1–6.