A Comparative Study of Machine Learning Approaches for Analyzing COVID-19 Global Data

Mushfiqur Rashid Khan[1], Atanu Das Bapon[2]

[1,2]Dept. of Computer Science and Engineering,
Ahsanullah University of Science and Technology
[1]prottoy.mushfiq@gmail.com, [2]atanudasbapon@gmail.com

Introduction:

"COVID-19", the buzzword of neoteric life and the most threatening disease that has shaken the entire globe has left researchers around the world vigilant and curious like us. Being a completely new and nescient challenge, sentient researchers have undergone relentless labor for tackling it. Our curiosity has driven us to perform a proper analysis of the regional effect of the COVID-19 pandemic around the whole world using several Machine Learning (ML) algorithms to trace down the best one.

Research Area:

Our research confides in the vast scope of Machine Learning. Infliction of Artificial Intelligence (AI) has assisted particular programs to find patterns in voluminous datasets.

Objective:

A plethora of ML algorithms is there to deal with diverse datasets. But not every algorithm can uphold the substantive purpose of dispensation. This bewilderment paves us towards exerting an experiment on a couple of ML approaches to determine which particular algorithm provides the best accuracy score as well as the least processing time for our massive dataset (COVID-19 Global Dataset containing 69,561 data). Hence, our prime objective is to redact a comparative study on some ML approaches to unearth the appropriate one that works best for our collected dataset.

Methodology:

The entire contexture initiated with the acquisition of a large dataset followed several preprocessing steps. The massive dataset-file named "WHO-COVID-19-global-data.csv" had been excerpted from the official website of World Health Organization (WHO). It comprises eight (8) features and 69,561 data from 1 March 2020 to 24 October 2020. Such voluminous data were noisy, incomplete, inconsistent, and duplicate. To ensure data quality (accuracy, completeness, consistency, reliability, interpretability), the necessity of data preprocessing originated.

Data preprocessing steps combined cleaning, integration, reduction, transformation, and discretization of data. We engulfed missing values, eliminated duplicate data, performed integration for merging another dataset, and redacted data-transformation for a better analysis. Data Cleaning enlisted stuffing in missing values, smoothing out noise while identifying outliers, and rectifying inconsistencies in data. To merge data from diverse sources into a coherent data-store, Data Integration was applied. Data Reduction involved the elimination and clustering of redundant features. Normalization, conversion, or construction of attribute or feature type perpetrated Data Transformation. To alleviate the number of values for a conferred continuous attribute, Data Discretization assisted in Binning, Histogram Analysis, Cluster Analysis, and Correlation Analysis. Afterward, 11 ML algorithms were applied while maintaining the 70%-30% training-testing ratio of the dataset. These algorithms included Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), k-Nearest Neighbor (kNN), Decision Tree, Extra Trees, Random Forest, Voting Ensemble, Stochastic Gradient Boosting, AdaBoost, and Bagging. A confusion-matrix was generated to calculate the accuracy, precision, recall, and F1 score. The elapsed time for each algorithm was ascertained later on.

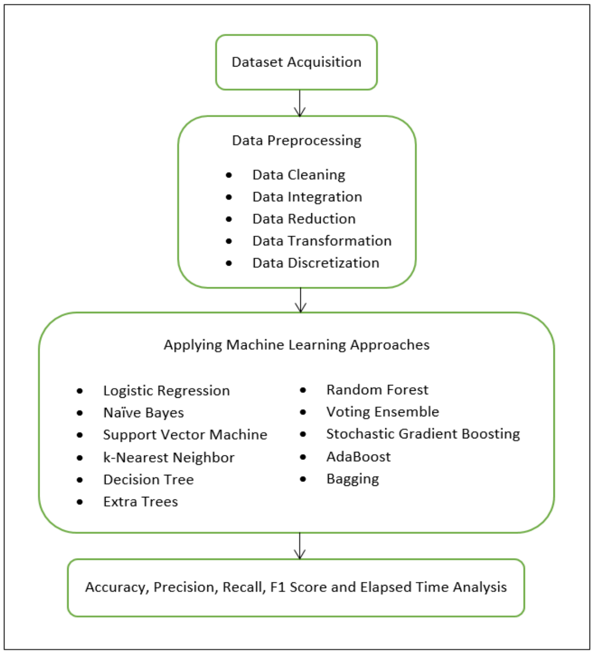


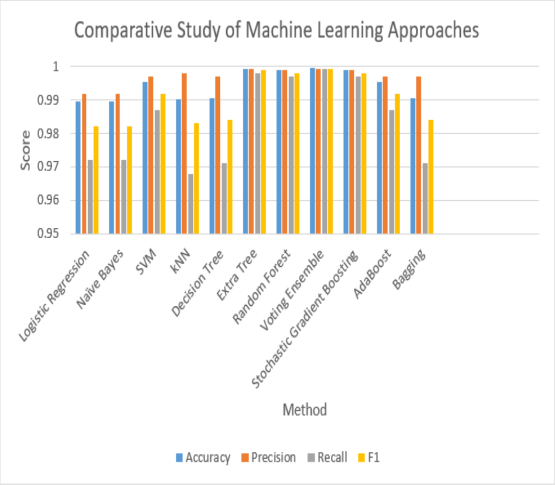
Fig. 1. Flow diagram of our proposed methodology

Results and Analysis:

Concise analysis and comparison among the ML models provided us with a noticeable conclusion. Such recapitulation is illustrated in Table 1.

Table 1. A comparison of various ML algorithms

Method	Accuracy (%)	Precision	Recall	F1 score	Elapsed Time (s)
Logistic Regression	98.97	0.992	0.972	0.982	22.9
Naïve Bayes	98.97	0.992	0.972	0.982	16.3
SVM	99.54	0.997	0.987	0.992	22.8
kNN	99.01	0.998	0.968	0.983	20.9
Decision Tree	99.07	0.997	0.971	0.984	18.0
Extra Trees	99.94	0.999279	0.998	0.999	26.0
Random Forest	99.89	0.999	0.997	0.998	13.4
Voting Ensemble	99.96	0.999282	0.9993	0.9993	21.6
Stochastic Gradient Boosting	99.89	0.999	0.997	0.998	16.7
AdaBoost	99.54	0.997	0.987	0.992	24.2
Bagging	99.07	0.997	0.971	0.984	16.1



It is discernible that Voting Ensemble offers the best performance according to the accuracy (99.96%), precision (0.999282), recall (0.9993), and F1 score (0.9993). In contrast, Random Forest holds the crown for providing the best accuracy according to the elapsed time (13.4 seconds) for our massive dataset. Although data of some regions were unexcavated, preprocessing steps diminished the dominance of this dilemma. The rest of the ML approaches might work better if there had been less scarcity of those data. However, this comparative study clinches to the crying need for investigating the suitable approach considering two distinct aspects.

References:

[1] Khanday, A.M.U.D., Rabani, S.T., Khan, Q.R., Rouf, N. and Din, M.M.U., 2020. Machine learning based approaches for detecting COVID-19 using clinical text data. *International Journal of Information Technology*, *12*(3), pp.731-739.