Hybrid Voting Classifier Model for COVID-19 Prediction by Embedding Machine Learning Techniques

GurjotKour[a], Pawanesh Abrol[b], Namrata Kalrupia[c], Jasmeen Kaur[d]

[a,b,c]Department of Computer Science and IT,University of Jammu,Jammu and Kashmir, India
[d]Department of Computer Science & Engineering, Mahant Bachittar Singh College of Engineering & Technology, Jammu,Jammu and Kashmir, India

Abstract: Predictive analytics, including methods of data mining, are typically used to improve predictability levels for results of interest or KPIs (Key Performance Indicators). This research work is based on the COVID-19 prediction which has various phases which include data pre-processing, feature reduction and classification. This paper presents a Hybrid Voting Classifier for prediction of corona infection. In this research work, dataset is collected from authentic data source which is pre-processed to remove missing and redundant values. The collected dataset contains incidences of Mexico COVID-19 cases. The dataset is further processed for the feature reduction using PCA algorithm and k-means algorithm is applied which can cluster similar and dissimilar features. In the last phase voting classifier is applied which is combination of naive Bayes, Random Forest Classifier, Bernoulli naive Bayes, and SVM for the COVID-19 prediction. The proposed model is examined in terms of parameters like accuracy, precision and recall. The performance results show that logistic regression givesan accuracy of 84%, naive Bayes has 82% and voting classification method generates maximum accuracy value of 94%. The recall value of logistic regression is 84%, naive Bayes gives 82% and voting classification gives maximum recall value of 94%. The precision value of logistic regression is 71%, naive Bayes gives 73% and voting classification results in the highest precision value of 94% for COVID-19 prediction. This study also depicts that how these supervised approach options may help to alleviate the enormous strain on the healthcare system's constrained capacity.

Keywords: COVID-19, Prediction, PCA, K-means, Naive Bayes, Random forest, Bernoulli naive Bayes, SVM, Voting Classifier Model

1. Introduction

Severe acute respiratory syndrome corona virus 2 (SARS-CoV-2) is reported as a virus strain due to which the respiratory disease of COVID-19 (corona virus disease 2019) is occurred **(Ardabili et al., 2020)**. This disease is declared as the extremely infectious by WHO (World Health Organization) and the global nations. The corona virus disease 2019 is considered as a public health emergency of international concern. The officials and media make the deployment of epidemiological models for estimating the epidemic, recognizing the peak ahead of time and predicting the mortality rate. Outbreak prediction models are efficient for providing the insights into the damages occurred due to the corona disease. Moreover, new policies are created and the conditions of curfew are computed using the predictive models as a reference. The epidemiological data containing the number of formerly diseased persons and the overall population is employed in the prediction models of corona virus 2019. Many criteria such as pre-infection period and probability of recovering from the virus are considered for predicting the trend of disease spread. But, these models are inefficient to reflect diverse socio-static and economic factors due to which the course of the virus is affected.

Thus, it is important to predict the disease on the basis of economic and social data and to analyze the trends of COVID-19 on the basis of epidemiological data **(Yang et al., 2020)**.The Novel Coronavirus 2019 is confirmed as a medical emergency by WHO. An open access is provided by the researchers and hospitals to the data of this contagion. There are certain confounding factors included in the gathered data because of which prediction of corona disease becomes difficult **(Nadella et al., 2020)**. The confounding factors can be reduced to relevant factors which can help in predicting corona illness easily.

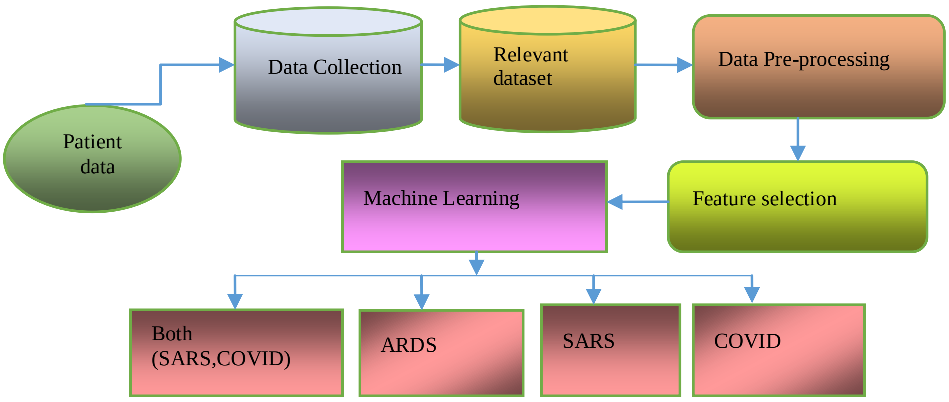


Figure 1: Generalized COVID-19 prediction model

The figure 1 presents generalized COVID-19 prediction model which has various steps which include data collection, data preparation, data pre-processing, feature selection that employ various machine learning algorithms. Generally, the datasets of corona disease are highly imbalanced that lay negative impact on predictive accuracy. In this process, the training data is sampled to be suitable for classification algorithm. The pre-processing method acts effectively in order to tackle an enormous volume of data or in cases at which domain knowledge is deficient. There are 3 stages included in the process of preparing the data in which data is handled, outliers are removed and data is balanced. This is an essential procedure to construct a ML framework.

The selection of suitable attributes is useful to alleviate the data redundancy and to avoid the noisy data. Consequently, the performance of model is enhanced. The attributes based on filter are selected using Relief feature selection. In every iteration n, an instance is chosen from the available instances using this algorithm in the dataset at random. Subsequently, feature relevance is updated on the basis of difference amid selected instances and two close neighbor instances as shown in equation (1).

$$W_n = W_n - (x_n - nearHit_n)^2 + (x_n - nearMiss_n)^2 \qquad (1)$$

Machine learning (ML) has emerged as a promising technology as uncertainties and a high degree of complexity have arisen in the development of outbreak epidemiological models **(Giordano et al., 2020)**. The prolific results and efficient outcomes are acquired from machine learning for developing the SIR (Susceptible-Infected-Recovered) model that has proven to be more reliable and widely applicable. ML is considered as the processing strategy having extraordinary potential for anticipating the corona illness.

ML approaches can be commonly partitioned into three principle classes, i.e., supervised, unsupervised and semi-supervised learning. This load of strategies needs information to learn specific classes or forms. This information is generally indicated as training data. The supervised learning is where the number needed to be learned and the task of labelling is recognized. The motivation behind this type of ML approach is to perceive designs to accumulate features utilizing with various classes in the training datasets. This kind of machine learning algorithm seeks to identify patterns in terms of groups of features by which different classes in the training dataset can be differentiated. The developed classification background subsequently can be useful to new data with unlimited class labels **(Jewell et al., 2020)**.

The second category is unsupervised learning, where both classes as well as the job of training data are unfamiliar to those classes. The unsupervised learning in clinical research is generally required for investigation of raw information **(Shoer et al., 2021)**. The reason for this approach is to recognize patterns or clusters in training data that deviate from arbitrary noise and are unlikely to happen unpredictably. The third type is semi-supervised learning which an amalgamation of supervised learning techniques and unsupervised learning techniques. In certain circumstances, it may be tough to completely assign whole training data to their corresponding class labels. There may be no information of the class labels or the efforts expected to discover accurately classified jobs for entire data might be extremely high. This issue can be settled by joining a limited quantity of labeled data with a vast capacity of data which is unlabeled **(Ferguson et al., 2020)**. Incorporating labelled and unlabeled data can generally enhance learning accuracy in cases where restricted amount of data is available for training.

Section 2 presents a comprehensive literature survey which is based on the various models which are designed for the COVID-19 prediction. In section 3, various predictive models are discussed. Research methodology, results and conclusion are presented in section 4, section 5 and section 6 respectively.

2. Literature Review

The literature survey is presented in this section which is based on models which are designed so far for the corona virus disease prediction. The designed models are based on the supervised learning, unsupervised learning and regression. Transfer learning technique is designed with the help of Googlenet in order to predict the infection caused by corona virus through images obtained from chest X-ray **(Haritha et al., 2020)**. GoogleNet, a kind of CNN (Convolutional Neural Network) model was adopted to classify the image. The occurrence of disease was revealed with the positively classified images. The primary health workers utilized this automatic technique in remote places at which there was unavailability of the skilled practitioners. The results demonstrated that the introduced technique had obtained 99% accuracy in training and 98.5% accuracy in testing phase while predicting the corona disease. In the future, the introduced approach would be expanded to develop an enormous dataset containing chest scans of corona affected and healthy patients with the purpose of enhancing the specificity and sensitivity further. The susceptible-exposed-infected-removed (SEIR)-PAD system is designed for evaluating the vulnerable, infected, unprotected, recovered, transmitter, symptomless, and deceased populations **(Sedaghat et al., 2020)**. This approach consists of seven regular mathematical equations that incorporated differentiation with 8 unidentified coefficients. An optimization algorithm was executed to solve these equations in MATLAB in numerical way so that it became suitable for implementation on medical data. The data taken during the epidemic disease was employed for predicting the trends of corona virus disease 2019 in GCC (Gulf Cooperation Council) countries. The efficient outcomes obtained from the suggested system offered insight for managing the pandemic of corona virus disease 2019.

A dynamic hybrid approach is formulated on the basis of SEIRD (Susceptible-exposed-Infected-Recovered-Deceased) and ascertainment rate using parameters chosen in automatic manner **(Ala'Raj et al., 2021)**. The modified SEIRD model and ARIMA (Auto Regressive Integrated Moving Average) model were comprised in this approach. The latter model was assisted in correcting the residuals the initial model people who were infected, recovered and deceased. The formulated approach was capable of analyzing the data for predicting the COVID-19 in confidence intervals. The US COVID statistics dataset was applied to test and authenticate this approach. The output of the formulated approach assisted various sectors and policymakers in mitigating the health risks in efficient manner. The major objective of the author was to develop a system for predicting corona disease and the total deaths caused by exposed individuals based on the reported cases of elderly, diabetic, and smoking patients **(Jarndal et al., 2020)**. This model was constructed using GPR (Gaussian Process Regression) and its comparison was done with ANN (Artificial Neural Network). A reliable data taken from WHO (World Health Organization) was utilized for the implementation of this system. The investigation represented that the promising outcomes were obtained from the developed system using predictive data of three countries. The numbers of deaths because of corona virus disease 2019 were predicted using this system as it had arbitrary number of inputs. The developed system was proved valuable in preparing efficient measures for mitigating the number of deaths. The enhanced epidemiology predictive model-eSEIR is developed for which the well-mixed Susceptible-Exposed-Infected-Removed (SEIR) model was enhanced on the infectious disease dynamics **(Ma et al., 2020)**. An optimization technique was contained in the established model for computing β and γ parameters. Diverse infection rate of latent and infected people and the difference of daily contact number amid these two groups were considered in this model. The epidemic data taken from Italy and China was executed to determine the model with decreased RMSE (root mean square error) in the final curve. The potential epidemic spread was predicted in the US with the deployment of this model. The established model had provided higher accuracy for predicting the Coronavirus Disease 2019 and its transmission in contrast to other models. An unprecedented method was put forward in which logical solution of the transmittable people was combined **(Sedaghat et al., 2020)**. For this, Weibull distribution function was implemented into any SIR like design. The projected system had predicted the patients suffering from the corona, suspected or recovered in a more efficient way in Kuwait and UAE as compared to the original model. The projected technique had offered insights to analyze the genetic dynamic structure having genuine biological data patterns instead of yielding complex numerical procedures. A machine learning scheme is constructed called artificial neural network for predicting the outbreak of Coronavirus disease 2019 in India **(Kumari et al., 2020)**.

In addition, a mathematical curve fitting system was put forward for determining the performance of constructed scheme. The progression of the epidemic was anticipated at various transmission rates in order to investigate the impact of precautionary administrative measures such as lockdown and social separation on corona virus transmission. The constructed scheme offered superior accuracy to predict the components associated with the corona virus disease 2019 with least MAPE values and cumulative deceased cases. The author focused on performing multi-class organization of the images obtained from chest scans of patients affected with corona disease, patients of pneumonia and unaffected people with the implementation of CNN **(Khan et al., 2020)**. Moreover, Monte-Carlo simulation was applied on the original data distributions to accomplish the task of predicting the diseases by improving the data. The LR (linear regression) of the components of GMM (Gaussian mixture model) was exploited in order to predict COVID-19. The presented approach was implemented and evaluated on chest scans of pneumonia record set collected from Kaggle and the University of Montreal. The presented approach obtained 100% accuracy in training phase and 96.66% in testing phase. Additionally, the LR equations were acquired which assisted in predicting the transmission of corona diseases from GMM. The designed models are the based on the SIR (Susceptible-Infectious-Recovered) and SVM (Support Vector Machine) algorithm **(Mantoro et al., 2020)**. The best or worst scenarios were taken in account to carry out the forecasting of corona virus disease 2019. The live daily reports were used in best case scenario and reports of other country were employed in worst scenario. The SIR model provided similar performance for predicting the disease in both the cases. The recommended approach was useful for the policy makers to tackle and manage the spread of corona virus disease 2019 pandemic. Machine learning system is intended in order to forecast the corona disease on the basis of BLS (Broad Learning System) **(Zhan et al., 2020)**. The key attributes were displayed by leveraging the RF (Random Forest) algorithm. Thereafter, RF-Bagging-BLS (Random-forest-Bagging Broad Learning System) was constructed by integrating the bagging with BLS so that the trend of corona disease was predicted. The constructed system was performed effectively to predict the corona disease with regard to RMSE (relative mean square error), MAE (median absolute error) and MAPE (mean absolute percentage error) in comparison with other algorithms. Thus, the intended approach outperformed the other algorithms. Table 1 provides a brief review of models deployed for analysis of COVID-19 behaviour.

Table1: Different approaches used for analysis of COVID-19 behaviour

Author	Approach deployed	Remarks
Yifan Yang, et.al (2020)	LSTM (Long Short Term Memory) algorithm	SEIR (Susceptible-Exposed-Infected-Removed) was put forward for forecasting the transmission of corona virus disease 2019 in China. **(Yang, et.al., 2020)**
Md Masud Rana, et.al (2020)	Optimal signal processing algorithm	On the basis of the designed gain, the error rate of dynamic system was alleviated so as an absolute COVID-19 prediction technique was built. Outcomes indicated that the formulated system had potential for predicting corona virus disease 2019 in a short term.**(Rana et. al., 2020)**
Andi Sulasikin, et.al (2020)	Holt's exponential smoothing and ARIMA (Auto-Regressive Integrated Moving Average)	The optimal models were determined by performing comparison on time series models to forecast the confirmed cases of corona infection. The result exhibited that the ARIMA offered higher R-Squared value, lower MSE and RMSE and more effective for forecasting the new infections in Jakarta. **(Sulasikin et. al., 2020)**
Huan Zhao, et.al (2021)	BPNN (Back Propagation Neural Network) algorithm	The epidemic scenario in Italy has been forecasted. The comparison of fitted estimate of the framework was done with the actual estimate and the range of curve fit was found to be 0.99. It was seen that the introduced approach was adaptable to predict the COVID-19 trend. **(Zhao et. al., 2021)**
Hua Ye, et.al (2021)	HHO (Harris Hawks Optimization)	The severity of Corona Virus Disease 2019 was predicted. The experimental results confirmed that the designed system had performed well and stable for 4 indexes as well as it displayed the key attributes for determining severe COVID-19 from mild COVID-19. Thus, the designed system was a valuable tool to predict coronavirus disease 2019. **(Ye et. al., 2021)**

The literature survey presented some latest research which is been conducted for the COVID-19 prediction. The researchers used three types of models which are supervised models, unsupervised models and regression models. The COVID-19 prediction is little difficult due to presence of irrelevant features in the dataset. The supervised models use the training dataset for the prediction and training dataset can be prepared from the main dataset due to which it gives high performance as compared to unsupervised learning. The unsupervised learning uses the pre-trained models for the prediction which makes it difficult for the COVID-19 prediction. The algorithms like logistic regression, naive Bayes are popular algorithms used for the COVID-19 prediction. The logistic regression and naive Bayes give accuracy of 84% and 82% respectively for the COVID-19 prediction. The idea is to design voting classification approach based on supervised learning models for the COVID-19 prediction which yields high accuracy as compared to logistic regression and naive Bayes. In section 3, various supervised learning algorithms are presented which can be used to design voting classification model for COVID-19 prediction.

3. Predictive models

Predictive analytics, including methods of data mining, are typically used to improve predictability levels for results of interest or KPIs (Key Performance Indicators).Predictive analysis is particularly useful in a global pandemic situation like corona wave by providing a general approximation of the results, thus easing the load on the already overburdened healthcare system. These specific predictive models or machine learning algorithms may frequently discover any type of relationship between the outcome variable and the predictors, and estimate them carefully to accurately predict events.

The various computing algorithms that have been used in the proposed hybrid voting classifier for prediction of patients' results are naive Bayes, Bernoulli naive Bayes, Random Forest Classifier and SVM.

Naive Bayes

Naive Bayesian network (NB) is an extremely simple Bayesian network. This network consisting of DAGs (Directed Acyclic Graphs) with a single parent (depicting the unmonitored node) and many children (in terms of observed nodes) strongly assumes that the child nodes are independent of their parents. Therefore, the naive Bayes model depends on the estimation given in equation (2).

$$R = \frac{P(a\backslash X)}{P(b\backslash X)} = \frac{P(a)P(X\backslash a)}{P(b)P(X\backslash b)} = \frac{P(a)\prod P(X_r\backslash a)}{P(b)\prod P(X_r\backslash b)} \tag{2}$$

When these two probable outcomes are compared, the larger one specifies that the value of the class label has more probability to be the real label (if R>1: predict a else predict$b$). This algorithm is specially inclined to be excessively affected by probabilities of 0 due to its use of a product computation for deducing the probabilities $P(X,a)$. The Laplace estimator adds one to all numerators.

Bernoulli Naive Bayes

Bernoulli model uses a binary vector to represent a document. This vector represents a point in the space of words. The Multi-variate Bernoulli model or Bernoulli model depends on binary data. It implies that each token in the feature vector of a document is correlated with the value 1 or 0. If feature vector having m dimensions, m denotes the number of words in the entire vocabulary. The value 1 depicts the occurrence of the word in the specific document, while 0 denotes the absence of the word in this document. The expression in equation (3) is generally used to write Bernoulli trials.

$$P\left(x|\omega_j\right) = \prod_{i=1}^{m} P\left(x_i|\omega_j\right)^b.\left(1 - P\left(x|\omega_j\right)\right)^{(1-b)} (b \in 0,1) \tag{3}$$

Assume$P\left(x_i|\omega_j\right)$is the maximum-likelihood estimate specifying the occurrence of a certain word (or token) $x_i$in class ω_j given by equation (4).

$$\hat{P}\left(x_i|\omega_j\right) = \frac{df_{xi,y}+1}{df_y+2} \tag{4}$$

Where $df_{xi,y}$represents the total record files in the learning set comprising the feature $x_i$and is related to classω_j. Also, df_ydenote the number of documents in the training dataset associating with classω_j. Finally, +1 and +2 represent the parameters of Laplace smoothing.

Random Forest

RF (Random Forest) is an ensemble learning algorithm. An ensemble is a set of diverse classification algorithms which are integrated for generating more powerful model. Generally, this algorithm is planned on the basis of the bagging method that is a statistical technique using which some different training sets are developed starting from a single one. A subset of random attributes is implemented for each classifier to construct a RF on each training set. The DTs (decision trees) are not stable. When these algorithms are developed on diverse training sets with random attributes, various classification algorithms are generated. Hence, the correlation is mitigated among the models due to which the overall performance is maximized. A record is classified by merging all the decision trees with a voting system. Every tree gives a vote to a class for the record and the class having the highest votes is selected as the final outcome using RF. The voting system is deployed to average the DTs (decision trees) predictions that have influence of high variance. Consequently, the accuracy is boosted even with large sets of data. This implies, the random forest is less inclined toward overfitting in comparison with a single decision tree as it averages the tree predictions and provided results with higher robustness. In addition, only few parameters are contained in this algorithm for the construction that is always desirable. This approach has limitation of the computational time that is maximized when the number of trees is increased.

Support Vector Machine (SVM)

The support vector machine is a popular supervised machine learning algorithm. The algorithm is built around the concept of "margin". It indicates periphery of decision boundaries separating two categories of data. By maximizing the margin that establishes the largest feasible distance between the segregating subspace and samples on one of the boundaries of it, an upper constraint on the likely standardization error has been reduced. A pair (w, b) occurs when the training data are linearly divisible as follows:

$$w^T x_i + b \geq 1, for\ all\ x_i \in P \tag{5}$$

$$w^T x_i + b \leq -1, for\ all\ x_i \in N \tag{6}$$

For this purpose, the decision rule provided by $f_{w,b}(x) = sgn(w^T x + b)$ is used. Here, w represents the weight vector and b denotes bias (or -b corresponds to threshold). The squared norm of the separating hyperplane can be minimized to get the best separating hyperplane by linearly separating two classes. The minimization can be represented as a convex quadratic programming (QP) issue:

$$\underset{w,b}{Minimize} \Phi(w) = \frac{1}{2}\|w\|^2 \tag{7}$$

$subjecting\ y_i(w^T x_i + b) \geq 1, i = 1, \ldots . l.$

After finding the best separating hyperplane, the data points lying on the decision surface are termed as support vectors and the final output is obtained by linearly combining merely these points when the data is linearly separable. In the section 4, research methodology is presented which is based on the voting classification for the COVID-19 prediction. The voting classification model deploys supervised machine learning algorithms like naive Bayes, SVM, Bernoulli naive Bayes and random forest. The COVID-19 dataset is very redundant due to which it is preferred to use defined algorithms because they give reliable results in case of labelled information.

4. Research Methodology

This research work is based on the COVID-19 prediction by using labelled dataset. The proposed technique has various steps which include dataset input, pre-processing, feature extraction, classification and performance analysis. The Mexico COVID-19 dataset is used which contains patient records and it is collected from Kaggle[1]. The various supervised machine learning algorithms are used to design voting classification model. The collected dataset contains various irrelevant information due to which it is preferred to use supervised machine learning algorithms as compared to unsupervised machine learning algorithms.

[23] Mexico COVID-19 clinical data. (n.d.). Kaggle: Your Machine Learning and Data Science Community. https://www.Kaggle.com/marianarfranklin/mexico-COVID19-clinical-data/metadata

The proposed research methodology for the COVID-19 prediction attempts to explore the labeled collection of patients' records containing positive and negative test results. The proposed model is implemented on Mexico corona patients' dataset collected from the Kaggle repository. The dataset contains approximately 31 attributes and one target set which is positive and negative.

In the dataset the positive case is labelled with 1 and negative case is labelled with 0. The collected dataset has various missing and redundant values. To remove missing and redundant values mean of the whole dataset is taken. The missing and redundant values will be replaced with the mean value. The pre-processed dataset is taken as input for the feature reduction. PCA has been applied for the feature reduction. The PCA algorithm is utilized to build a low-dimensional representation of the data which defines as much of the variance in the data as possible. For this, a linear basis of reduced dimensionality is discovered for the data, that has maximum amount of variance in the data is maximal.

Figure 2 illustrates the steps in the proposed hybrid voting classification-based model for prediction of corona illness.

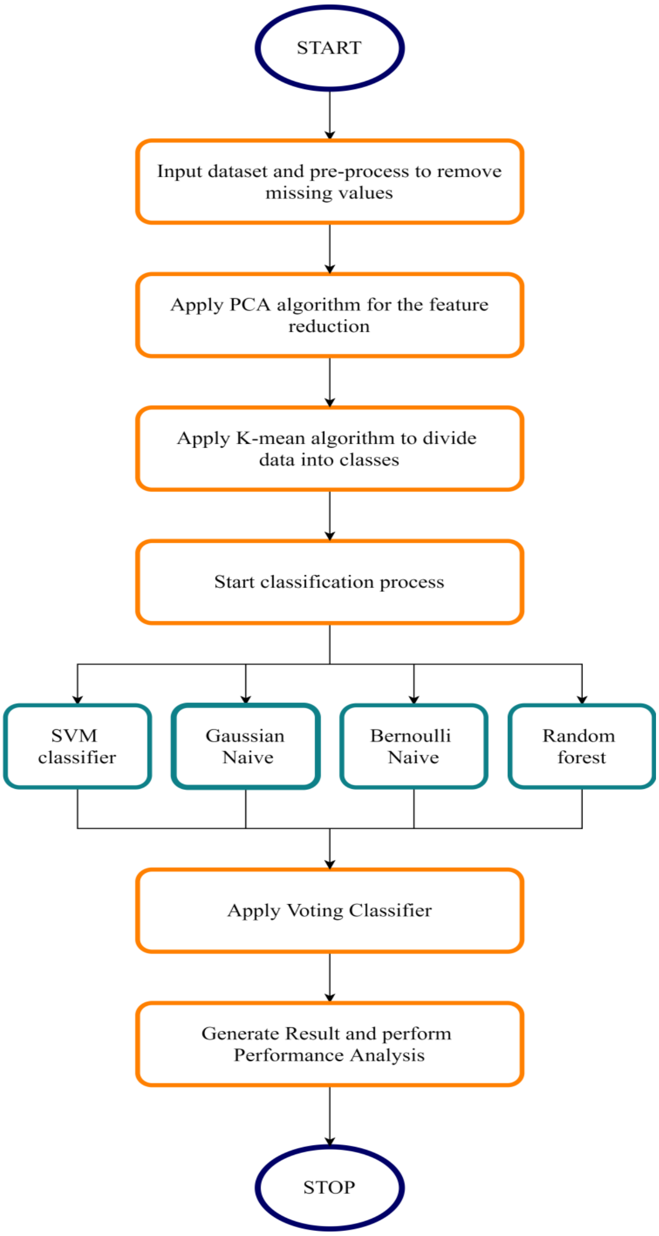
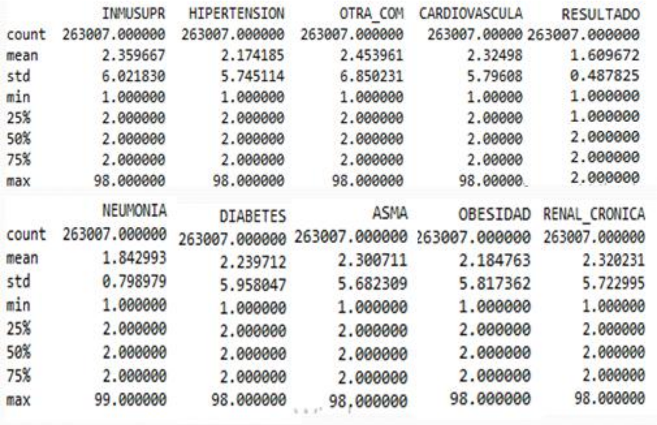

START
Input dataset and pre-process to remove missing values
Apply PCA algorithm for the feature reduction
Apply K-mean algorithm to divide data into classes
Start classification process
SVM classifier
Gaussian Naive
Bernoulli Naive
Random forest
Apply Voting Classifier
Generate Result and perform Performance Analysis
STOP

Figure 2: Proposed epidemiology prediction model

When the features of the dataset get reduced then dataset is processed for the clustering. The k-mean clustering is applied which can cluster similar type of information. It is a simple and extensively utilized clustering algorithm. For the data K-mean algorithm divide whole data into K number of clusters as K value is offered as the input value. The clusters are employed to group the values of the dataset on the basis of their values in the dataset but not on their locations in the dataset, till the location is presented as a specified property. In classification phase, voting classification method is applied for the COVID-19 prediction. The voting classification method is the combination of various classifiers which includes naive Bayes, Random Forest Classifier, Bernoulli naive Bayes and SVM Classifier. A Voting Classifier, a ML (machine learning) model is implemented to train an ensemble of numerous models and predict an output (class) on the basis of their highest probability of chosen class as the output. The findings of each classifier are aggregated and passed into Voting Classifier and the output class is predicted on the basis of the highest majority of voting using this algorithm. Rather than developing the separate dedicated models and investigating the accuracy for each one of them, a single model is constructed whose training is done with the deployment of these models and the output is predicted on the basis of their combined majority of voting for each output class. The proposed model is implemented using python on the Mexico Covid-19 dataset and results of the model are presented in section 5.

5. Result and Discussion

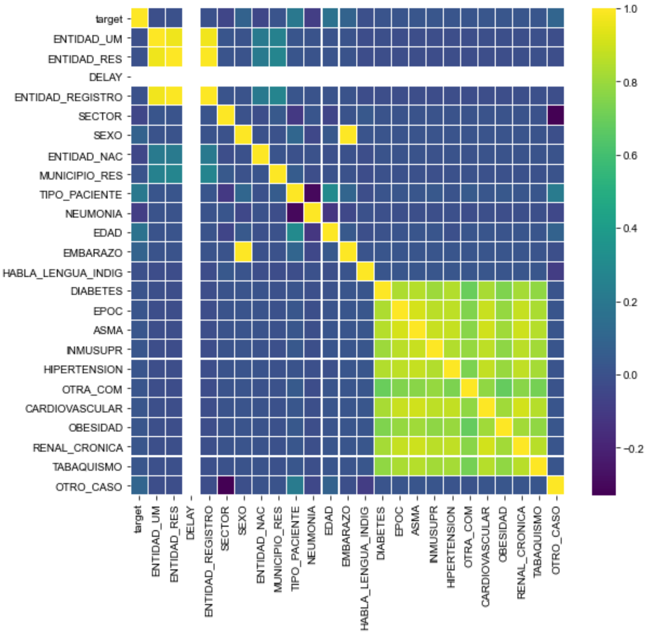
The supervised machine learning methods are applied for the COVID-19 prediction. The supervised machine learning algorithms like naive Bayes, Bernoulli naive Bayes, SVM and voting classification and these algorithms are implemented in python. The dataset is of Mexico COVID-19 patient record which is collected from the Kaggle repository [23].The dataset contains the information of positive and negative cases which is reported by the General Directorate of Epidemiology, Secretariat of Health in Mexico. This dataset is generated from the reported results of the RT-PCR test. The dataset contains approximately 30000 instances and 31 attributes. The 70% of the dataset is used for training and rest 30% for testing. The dataset contains two types of features, first one which are only required for the hospitalization and the second one containing information about the patient's health conditions. The hospitalization features will be removed from the dataset for the prediction analysis.



```
            INMUSUPR    HIPERTENSION        OTRA_COM  CARDIOVASCULA       RESULTADO
count  263007.000000   263007.000000   263007.000000   263007.00000   263007.000000
mean        2.359667        2.174185        2.453961        2.32498        1.609672
std         6.021830        5.745114        6.850231        5.79608        0.487825
min         1.000000        1.000000        1.000000        1.00000        1.000000
25%         2.000000        2.000000        2.000000        2.00000        1.000000
50%         2.000000        2.000000        2.000000        2.00000        2.000000
75%         2.000000        2.000000        2.000000        2.00000        2.000000
max        98.000000       98.000000       98.000000       98.00000        2.000000
            NEUMONIA        DIABETES            ASMA       OBESIDAD    RENAL_CRONICA
count  263007.000000   263007.000000   263007.000000  263007.000000    263007.000000
mean        1.842993        2.239712        2.300711       2.184763         2.320231
std         0.798979        5.958047        5.682309       5.817362         5.722995
min         1.000000        1.000000        1.000000       1.000000         1.000000
25%         2.000000        2.000000        2.000000       2.000000         2.000000
50%         2.000000        2.000000        2.000000       2.000000         2.000000
75%         2.000000        2.000000        2.000000       2.000000         2.000000
max        99.000000       98.000000       98.000000      98.000000        98.000000
```

Figure 3: Mexico Covid-19 Dataset Description

The figure 3 gives a general description of features in the Mexico patient record dataset diagnosed with corona virus infection. The dataset contains the information of positive and negative cases which is reported by the General Directorate of Epidemiology, Secretariat of Health in Mexico. This dataset is generated from the reported results of the RT-PCR test. The dataset contains approximately 30000 instances and around 31 features. The dataset contains two types of features first one which are only required for the hospitalization and the second one that are vital for the prediction analysis containing information about patient's health conditions. In the original dataset, a target set which contains positive and negative classes will be transformed to 0 and 1. The0is used represent the negative case class and 1 is used to represent the positive case class for the corona illness prediction.



.

Figure 4: Correlation Scatterplot of Dataset Attributes

The figure 4 presents matrix of the various attributes of the dataset. The relationship between various attributes in the collected dataset depicting patient’s health conditions at the time of hospital admission is shown in the figure. The target attribute represents the test outcome whether the person is infected or not. The colour in the figure represents relationship value which is range from 0.1 to 1. The 0.1 is the lowest value and 1 is the maximum value in the dataset.

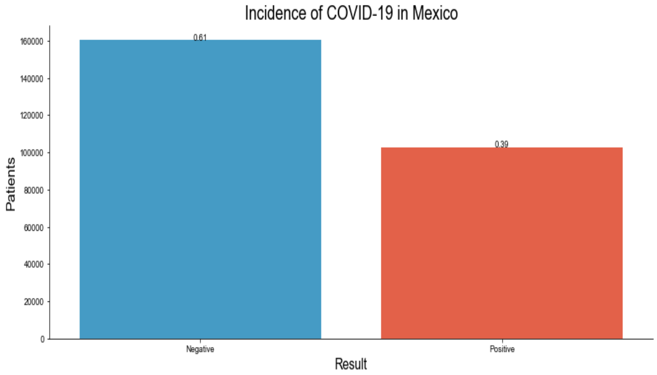
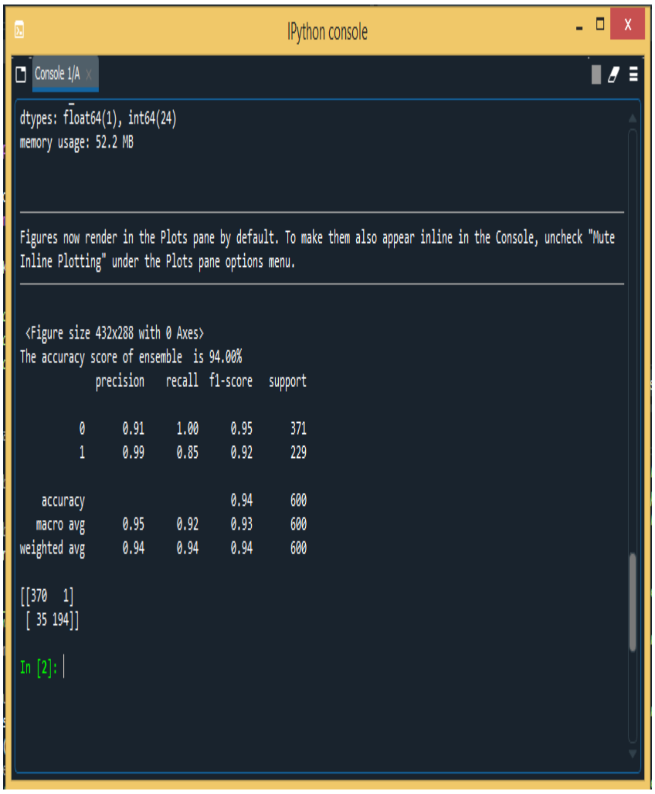


Figure 5: Number of Positive and Negative Cases

As shown in figure 5, the positive and negative cases in the dataset are plotted in the form of bar graph. The y-axis shows the number of patients which are positive and negative. The 61% of the whole dataset are the negative cases and 39% is the positive cases.

In the original dataset target set which contains positive and negative classes will be transformed to 0 and 1. The 0 represents the negative case and positive case will be 1 for the COVID-19 prediction.

In the proposed model, technique of voting classification is applied for the COVID-19 prediction. The execution results of proposed model are shown in Figure 6 illustrates that the data which is taken as input from the authentic source is pre-processed.



```
dtypes: float64(1), int64(24)
memory usage: 52.2 MB

Figures now render in the Plots pane by default. To make them also appear inline in the Console, uncheck "Mute
Inline Plotting" under the Plots pane options menu.

 <Figure size 432x288 with 0 Axes>
The accuracy score of ensemble  is 94.00%
              precision    recall  f1-score   support

           0       0.91      1.00      0.95       371
           1       0.99      0.85      0.92       229

    accuracy                           0.94       600
   macro avg       0.95      0.92      0.93       600
weighted avg       0.94      0.94      0.94       600

[[370   1]
 [ 35 194]]

In [2]:
```

Figure 6: Execution of Voting Classifier

Figure 6 shows the execution results generated by the implementation of Voting Classifier based proposed model. The data is divided into training and testing. The algorithm of voting classifier is applied which can predict the target set for the COVID-19. The precision and recall value is shown 94% and accuracy value is 94 % for the classification.

The results are evaluated on the basis of three parameters i.e. accuracy, recall and precision. The accuracy is the number of samples whose classification is done correctly to the total number of samples available. The mathematical representation of this parameter is expressed as:

$$A_i = \frac{t}{n} \cdot 100 \qquad (8)$$

Where, t depicts the numbers of samples which are correctly classified and n denotes the number of sample cases. Execution time gives the difference between the end time of an algorithm and its initiation time. Precision is the division of positive instances with the total number of instances which are declared positive. The mathematical representation of this parameter is expressed as:

$$\text{Precision=TP/(TP+FP)} \qquad (9)$$

The recall is defined as the true positive instances which are extracted and divided by the total number of positive instances and is expressed as:

$$\text{Recall= TP/(TP+FN)} \qquad (10)$$

Table 2 illustrates the analysis of performance evaluation results of the proposed model in comparison to logistic regression and Naïve Bayes models based on of accuracy metrics like accuracy, precision & recall.

Table 2: Performance Analysis

Parameters	Logistic Regression	Naive Bayes	Voting Classifier
Accuracy	84%	82%	94%
Recall	84 %	82 %	94%
Precision	71 %	73%	94%

The voting classifier is the combination of various classifiers which include Gaussian naive Bayes, Bernoulli naive Bayes, random forest classifier and SVM for the COVID-19 prediction.

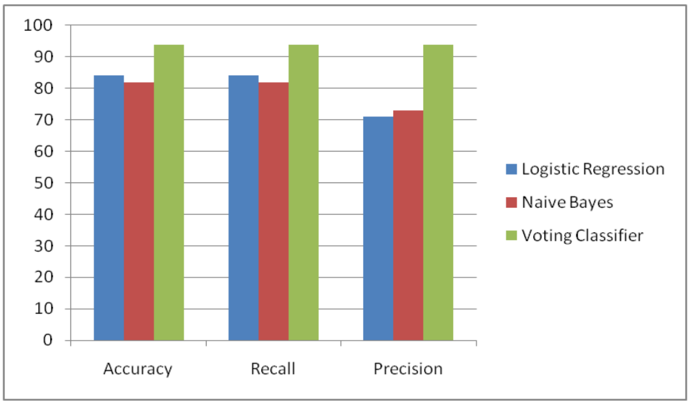


Figure 7: Performance comparisons on the basis of accuracy, precision & recall

As shown in figure 7, the various classifiers like logistic regression, naive Bayes and the proposed voting classifier are compared in terms of accuracy, precision and recall. It is analyzed that in terms of all these three parameters voting classifier gives best performance for COVID-19 prediction.

6. Conclusion

COVID-19 is a deadly infectious disease spreading all across the world. The proposed COVID-19 prediction model has various steps which include pre-processing, feature extraction and classification. The dataset of COVID-19 is collected from the Kaggle and it is of Mexico patients. The Input dataset is transformed in which positive cases are represented with 1 and negative cases are presented with 0. The various attributes which are irrelevant are removed from the dataset and also missing values are removed by taking mean of the whole dataset in the pre-processing phase. The PCA algorithm is used for the feature reduction and reduced features get clustered using K-mean algorithm. The output of the k-mean algorithm is given as input to the voting classification for the COVID-19 prediction. In this research work, hybrid voting classification model is proposed which is a combination of naive Bayes, Bernoulli naive Bayes, Random Forest Classifier and SVM for the COVID-19 prediction. The performance of the proposed model is analyzed in terms of accuracy, precision and recall. It is analyzed that proposed model result upto94% accuracy as compared to existing model for COVID-19 prediction. This study also depicts that how these supervised approach options may help to alleviate the enormous strain on the healthcare system's constrained capacity.

References

Ala'raj, M., Majdalawieh, M., & Nizamuddin, N. (2021). Modeling and forecasting of COVID-19 using a hybrid dynamic model based on SEIRD with ARIMA corrections. *Infectious Disease Modelling*, *6*, 98-111. https://doi.org/10.1016/j.idm.2020.11.007

Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., Rabczuk, T., & Atkinson, P. M. (2020). COVID-19 outbreak prediction with machine learning. *Algorithms*, *13*(10), 249. https://doi.org/10.3390/a13100249

Ferguson, N. M., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., ... & Hinsley, W. (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Imperial College COVID-19 Response Team. Imperial College COVID-19 Response Team, 20

Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., & Colaneri, M. (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine*, *26*(6), 855-860. https://doi.org/10.1038/s41591-020-0883-7

Haritha, D., Swaroop, N., & Mounika, M. (2020, October). Prediction of COVID-19 Cases Using CNN with X-rays. In 2020 5th International Conference on Computing, Communication and Security (ICCCS) (pp. 1-6). IEEE.

Jarndal, A., Husain, S., Zaatar, O., Al Gumaei, T., & Hamadeh, A. (2020, November). GPR and ANN based Prediction Models for COVID-19 Death Cases. In 2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI) (pp. 1-5). IEEE.

Jewell, N. P., Lewnard, J. A., & Jewell, B. L. (2020). Predictive mathematical models of the COVID-19 pandemic. *JAMA*, *323*(19), 1893. https://doi.org/10.1001/jama.2020.6585

Khan, Y., Khan, P., Kumar, S., Singh, J., & Hegde, R. M. (2020, December). Detection and Spread Prediction of COVID-19 from Chest X-ray Images using Convolutional Neural Network-Gaussian Mixture Model. In 2020 IEEE 17th India Council International Conference (INDICON) (pp. 1-6). IEEE.

Kumari, P., & Toshniwal, D. (2020, November). Real-time estimation of COVID-19 cases using machine learning and mathematical models-The case of India. In 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS) (pp. 369-374). IEEE.

Ma, Y., Xu, Z., Wu, Z., & Bai, Y. (2020, October). COVID-19 Spreading Prediction with Enhanced SEIR Model. In 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE) (pp. 383-386). IEEE.

Mantoro, T., Handayanto, R. T., Ayu, M. A., & Asian, J. (2020, October). Prediction of COVID-19 Spreading Using Support Vector Regression and Susceptible Infectious Recovered Model. In 2020 6th International Conference on Computing Engineering and Design (ICCED) (pp. 1-5). IEEE.

Nadella, P., Swaminathan, A., & Subramanian, S. V. (2020). Forecasting efforts from prior epidemics and COVID-19 predictions. *European Journal of Epidemiology*, *35*(8), 727-729. https://doi.org/10.1007/s10654-020-00661-0

Rana, M. M., Abdelhadi, A., Ahmed, M. R. U., & Ali, A. (2020, August). Secure IoT communication systems for prediction of COVID-19 outbreak: an optimal signal processing algorithm. In 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 135-139). IEEE.

Sedaghat, A., Band, S., Mosavi, A., & Nadai, L. (2020, November). COVID-19 (Coronavirus Disease) Outbreak Prediction Using a Susceptible-Exposed-Symptomatic Infected-Recovered-Super Spreaders-Asymptomatic Infected-Deceased-Critical (SEIR-PADC) Dynamic Model. In 2020 IEEE 3rd International Conference and Workshop in Óbuda on Electrical and Power Engineering (CANDO-EPE) (pp. 000275-000282). IEEE.

Sedaghat, A., Band, S., Mosavi, A., & Nadai, L. (2020, November). Predicting COVID-19 (Coronavirus Disease) Outbreak Dynamics Using SIR-based Models: Comparative Analysis of SIRD and Weibull-SIRD. In 2020 IEEE 3rd International Conference and Workshop in Óbuda on Electrical and Power Engineering (CANDO-EPE) (pp. 000283-000288). IEEE.

Shoer, S., Karady, T., Keshet, A., Shilo, S., Rossman, H., Gavrieli, A., Meir, T., Lavon, A., Kolobkov, D., Kalka, I., Godneva, A., Cohen, O., Kariv, A., Hoch, O., Zer-Aviv, M., Castel, N., Sudre, C., Zohar, A. E., Irony, A., … Segal, E. (2021). A prediction model to prioritize individuals for a SARS-Cov-2 test built from national symptom surveys. *Med*, *2*(2), 196-208.e4. https://doi.org/10.1016/j.medj.2020.10.002

Sulasikin, A., Nugraha, Y., Kanggrawan, J., & Suherman, A. L. (2020, July). Forecasting for a data-driven policy using time series methods in handling COVID-19 pandemic in Jakarta. In 2020 IEEE International Smart Cities Conference (ISC2) (pp. 1-6). IEEE.

Yang, Y., Yu, W., & Chen, D. (2020, July). Prediction of COVID-19 spread via LSTM and the deterministic SEIR model. In 2020 39th Chinese Control Conference (CCC) (pp. 782-785). IEEE.

Yang, Z., Zeng, Z., Wang, K., Wong, S., Liang, W., Zanin, M., Liu, P., Cao, X., Gao, Z., Mai, Z., Liang, J., Liu, X., Li, S., Li, Y., Ye, F., Guan, W., Yang, Y., Li, F., Luo, S., … He, J. (2020). Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *Journal of Thoracic Disease*, *12*(3), 165-174. https://doi.org/10.21037/jtd.2020.02.64

Ye, H., Wu, P., Zhu, T., Xiao, Z., Zhang, X., Zheng, L., Zheng, R., Sun, Y., Zhou, W., Fu, Q., Ye, X., Chen, A., Zheng, S., Heidari, A. A., Wang, M., Zhu, J., Chen, H., & Li, J. (2021). Diagnosing coronavirus disease 2019 (COVID-19): Efficient Harris hawks-inspired fuzzy k-nearest neighbor prediction methods. *IEEE Access*, *9*, 17787-17802. https://doi.org/10.1109/access.2021.3052835

Zhao, H., Li, Y., Chu, S., Zhao, S., & Liu, C. (2021, April). A COVID-19 Prediction Optimization Algorithm Based on Real-time Neural Network Training—Taking Italy as an Example. In 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC) (pp. 345-348). IEEE.

Zhan, C., Zheng, Y., Zhang, H., & Wen, Q. (2021). Random-forest-Bagging broad learning system with applications for COVID-19 pandemic. *IEEE Internet of Things Journal*, *8*(21), 15906-15918. https://doi.org/10.1109/jiot.2021.3066575