The Covid-19 Prediction in a county using Machine Learning techniques

Merin Joseph  
Graduate student, Computer Science Department  
California State University, Fullerton  
merin.joseph@csu.fullerton.edu  
  
  
Tianhao Shen  
Graduate student, Computer Science Department  
California State University, Fullerton  
tianhaoshen@csu.fullerton.edu  
  
  
Ruiling Wang  
Graduate student, Computer Science Department  
California State University, Fullerton  
rwang18@csu.fullerton.edu  
  
  
Kanika Sood  
Assistant Professor, Computer Science Department  
California State University, Fullerton  
kasood@fullerton.edu

*Abstract*—The Covid-19 Corona Virus, otherwise known as SARS-CoV-2, has caused destruction all around the world since December 2019. It was declared as a health emergency for the entire globe by the World Health Organization (WHO) in January 2022. Millions of people’s lives have been affected. Amid this pandemic, having an early warning system that can predict how much an area will be affected by the disease would be great. This can help the government enforce more safety measures to avoid a tragic situation. This project aims at modelling such a warning system that predicts the positivity rate of Covid-19 in an area. The proposed solution uses supervised Machine Learning techniques such as Random Forest, Linear Regression, Naïve Bayes and Gradient Boosting regression. The prediction can be made by analyzing past data of an area in a particular time with features such as the population of the area, number of tests conducted, number of positive tests, reported cases in that area and so on. Predictable computer applications for predicting the trend of Covid-19 spread can help the public and government to take necessary precautions for suppressing its effect.

Keywords—Covid-19 Corona Virus, Machine Learning, Random Forest, Linear Regression, Naïve Bayes, Gradient Boosting, Cross-Validation

# Introduction

The first case of the Covid-19 Corona Virus, a disease caused by the virus SARS-CoV-2 (Severe Acute Respiratory Syndrome Corona Virus 2), was found when a cluster of patients in Wuhan, China began to experience shortness of breath and fever on December 12, 2019. It is highly contagious in nature and has spread all over the world in a very short period [1]. The number of infected cases began to rise, and all the cases connected to the Huanan Seafood Wholesale Market in Wuhan. This pandemic has caused havoc all around the world and has deleterious effects on the global economy. The disease spreads from one infected person to another by having direct contact between them. Coughing or sneezing may transfer virus-containing tiny droplets from an infected person and hence covering the face using an appropriate mask was highly enforced all around the world to prevent the spread of the virus. The symptoms of the Covid-19 disease include cough, fever, cold, absence of taste/smell, shortness of breath, or even no symptoms at all. The lower and upper tracts of the respiratory organs are mostly affected by this infection. Infected people are at high risk of death due to the severity of this virus’s side effects. With time, many other variants of this virus came into existence, some even more deadly than the original virus-like Omicron, Delta, and many more. The various commonly adopted control measures were keeping the infected individuals in quarantine, covering their faces in public places, keeping hands clean all the time by using a sanitizer or washing our hands with soaps, social distancing, and avoiding contact with other people and so on. The Covid-19 disease can be medically diagnosed by analyzing its symptoms or by running a reverse transcription-polymerase chain reaction (RT-PCR) test. The earlier the infection of this virus is detected, the better it is to treat the patients. Hence, predicting the curve or rate of new positive cases can be particularly useful in implementing preventive control measures.

This project is about predicting the future positive cases for the area based on past data. This is a regression problem where we predict the value of Covid-positive cases of an area based on numerous factors. Given a county, we must be able to predict the Covid-positivity rate for that count. The model is trained using machine learning algorithms used for regression like Random Forest, Linear Regression, Bayesian Ridge regression, and comparing their performance in solving this problem.

# Background

## Cross-Validation

## Linear Regression

Linear regression is a supervised machine learning technique. It is widely used for determining the underlying relationship between independent and dependent variables and applying that knowledge to predicting. It is possible to determine the linear relationship between the independent/input variables and the dependent/target variable. The following equation can be used to express it:

*y = a + b.x* 

In this equation, y is the target variable, x is the independent/input variable, a is the y-intercept, and b is the slope of the line. We are interested in finding the best fit line that can best capture the underlying relationship between the variables.

This machine-learning algorithm is simple to implement and understand with comparatively less complexity. This simplicity of the algorithm itself is one of the disadvantages of this technique. It oversimplifies the relationship between independent and dependent variables and assumes the relation to be linear which might not be the case in real-world datasets. Real-world problems or datasets might not always have a linear relationship between input and output variables. In such cases, using Linear Regression would result in a poor understanding of the dataset or inadequate generalization capabilities which in turn means poor accuracy of the model in general.

## Random Forest

The Random Forest Tree algorithm is a collection/ensemble of many decision trees that can be used to solve a problem, with the results of each decision tree being aggregated to get the final target variable value. Both classification and regression problems can be solved with this method. Bagging is an ensemble supervised learning method that combines Bootstrap and Aggregating. Bootstrapping is a technique for obtaining several subsets of the original dataset, each of which is picked at random and has a defined size. The input dataset is subsequently processed through many decision trees that were built during the training period, and the results are pooled or aggregated to generate the final response, mode value for classification issues, and mean value for regression problems. Each decision tree is independent of the others, and they are all executed at the same time.

By mixing several simultaneous weak learners to build a strong learner, Random Forest helps to reduce dataset overfitting. They also aid in the reduction of variance, which improves model accuracy. They can deal with outliers and missing values effectively. Some of the disadvantages of this strategy include the algorithm's complexity and the lengthier training period.

## Bayesian Ridge Regression

Bayesian Ridge Regressor is another machine learning approach employed in this research. It is a Bayesian model based on the assumption that all regression coefficients have the same variance. It uses probability distributions instead of point estimations to formulate linear regression. A probability distribution is used to generate the output/target label. Around x/independent variables, it is considered to be Gaussian distributed. The Gaussian distribution is defined as follows:

*P(y | X, w,* α) = *N(*y | Xw, α) (2)

## Gradient Boosting Regression

It is an ensemble technique of combining several weak learners to form a single strong learner. The weak learners are formed using decision trees in a stage wise manner and it outperforms random forest by generalizing better on the dataset. It allows optimization of an arbitrary differentiable loss function.

## Naïve Bayes

Based on the accuracy of the test kit and the false positive of the coronavirus test, and with the data provided by CDC and John Hopkins University, the figure of people who catch coronavirus will be reduced based on the data we see in the past two years. The algorithm will learn the data pattern from the history and predict the future within. Naïve Bayes takes a lot of data to calculate, but accuracy is guaranteed in most cases.

# Dataset preprocessing

The dataset for the project is collected from California Health and Human Services Open Data Portal. The website provided us with a dataset of size around 40,000 which contained data about Covid-19 cases in US counties from February 2020 to April 2022. This dataset was suitable for our project and could be used for training our model.

## Handling missing values

One way to handle missing values in a dataset is by deleting such rows/columns, but this can lead to a huge loss of data in some cases. Another method is by imputing values wherever values are missing with the mean, median, most frequent value/zero value, whichever suits the situation well. The dataset collected for this project had some missing values in certain features. We used KNN imputer that imputes the average or mode value of ‘n’ neighbors. We set the number of neighbors to be checked as 5. We also imputed the missing values with the mode value of that feature, respectively, in an alternate dataset. The mode value is the most frequently occurring value for that feature.

## Data Transformation

We can transform categorical values to a numerical scale using various encoding methods. In our dataset, we had many features of ‘object’ datatype such as county name and area type. Hence, we used a label encoder to encode the values using a label encoder from the Sklearn preprocessing library.

## Data Cleaning

The irrelevant features were identified using correlation coefficients and were removed. Corr() method from pandas computes the pairwise correlation of columns. The performance after removing irrelevant features and before could be compared and it is observed that the performance or accuracy of the model increased after removing irrelevant features.

# Methodology

A brief introduction for our study is using data to evaluate the object and the model. We collect the related data, mine the data, build our model, train the model, and eventually make predictions based on the model and the data we collected.

## Data Collection

The amount of data we collect to evaluate the future is based on how relative our data is to the prediction we are looking for. Data collection is the process of gathering information of interest, that can solve our problem statement, train or shape our model and give reliable outputs. For this project, we collected Covid-19 related data from a freely available website. The dataset has details about the number of cases found in a US county on a particular date. It contains around 40,000 observations from 2020 to April 2022 in different counties in Unites States of America.

## Data Preparation

Another term for this step is “data mining”. In this process, we make sure the act of manipulating raw data into a form that can readily and accurately be analyzed. We have the data from the previous step, but not all the data are useful, and some data can be missing. In this process, we are extracting and discovering patterns in large data sets and make the data set ready for the training model. For our dataset, we had certain missing values which is imputed with the mode value of that feature respectively. We have chosen mode value to impute the ‘population’ and ‘cases’ feature since filling them with the maximum possible worst value makes sense.

## Model Choosing

We analyzed the data and its distribution which helped us to pick a proper model for our project that can solve the problem efficiently. Since the prediction of the number of positive Covid-19 cases in an area is a regression problem, where we predict a continuous variable value, we researched many regression techniques available. After some research activity, Random Forest, Linear Regression, Bayesian Ridge regression and Gradient Boosting regression techniques looked apt for our project. We have implemented these three machine learning techniques in our project and their performance are slightly different from one another.

## Model Training

In this step, we adjust the parameters (or weights) of the model so that its performance in solving a certain task increases. There are diverse ways for a program to learn, such as supervised learning and unsupervised learning. We can either explicitly tell the model the output that it is supposed to produce for each input, and, if it does not produce it, then we adjust its parameters so that next time it is more likely to produce the correct output for that input. Or if the data is not labeled, we can use probability distribution for the training. We trained our model first using Random Forest Regressor from sklearn. Since it has multiple independent decision trees, it reduces overfitting and helps in improving accuracy. The input is then passed, and decision trees are built during training time. The nodes are split by making locally optimal choices and end at a leaf node showing the target variable value.

Linear Regression is another technique that we used to train our model. This technique is found very useful in capturing the underlying linear relationship between the input/independent and output/dependent variable. By analyzing the dataset and by plotting a scatter plot of the observations we could find a linear pattern that could be identified using this technique. Scatter plots are used for identifying correlational relationships.

Bayesian Ridge regression technique is also used to train our model and the performance against each technique is compared with the performance of other techniques. This regression technique is based on the Bayes theorem where the conditional probability of features plays a crucial role in determining the value of the target label.

At the Naïve Bayes, we used the data from early 2021 to 2022 present, and other data provided with authoritative sources, Naïve Bayes can learn the pattern of the history and predict the future based on that historical pattern. This algorithm is simple to implement, and have a great accuracy with the amount of data we fed in. Because the data we used are relative to each other, it is simple for correct prediction for these data.

For the Gradient Boosting technique, weak learners are combined to form a strong learner. It is usually an ensemble of decision trees.

## Evaluate the Model

We used root mean square error, R2 value, and Box and Whisker plots of MAE (Mean Absolute Error) to "measure" the objective performance of the model after testing the model against previously unseen data, and this unseen data is meant to be sensitive to model performance in the real world, but still helps adjust the model. This project uses different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses.

## Make Predictions

At this point, we used our trained model to make predictions about the Covid-19 case positivity rate. A simulation of the figures that reflect the real world based on the program. By comparing our own prediction and the results from our training model, we will be able to tell the difference between the two and adjust the area for further and more accurate prediction.

# Results

The model was developed using the pre-processed dataset and trained using three algorithms, Random Forest, Linear Regression and Bayesian Ridge regression. The mean squared value and R2 value of each algorithm against the dataset was observed, and it can be found that Linear regression has the most accuracy in correctly predicting the Covid-19 positivity rate after training. The R2 value of Linear regression was found to be 0.99365 which means an accuracy of 99.365 percent while the R2 value of Bayesian Ridge regression algorithm was 0.99364 (99.364 percent accuracy). The R2 value of Random Forest was found to be 99.089, not too far from the values of other two algorithms.

##### References

|  |  |
| --- | --- |
| [1] | E. Gothai, R. Thamilselvan, R. Rajalakshmi, R. Sadana, T. R., A. Ragavi and R. Sakthivel, "Prediction of COVID-19 growth and trend using machine learning approach," *Elsevier Public Health Collection,* 2021. |