Milestone-III

E19CSE020 E19CSE087 E19CSE135 E19CSE314

October 2021

1 ABSTRACT

Introduction: The main purpose of our model would be to predict the reopening of different institutes. It would take in different Data-sets. We will first make a model for the prediction of the third wave which would be used for a second model for predicting reopening of colleges. These predictions would be really helpful for the students around the country who are waiting for their college to reopen.

Methodology: We make our first model after comparing different algorithms and choosing the one with better results. We use the predictions from the first model to use in our second model.

Results: We get a forecasting of the third Covid wave and predict the reopening of institutes which have not been reopened as of yet.

2 Introduction

Covid-19 has caused a major halt in the lives of college students since 2020. Even after some relief in the covid measures, most of the institutions remain closed. With the recent uncertainty about the covid wave in upcoming months, every student's mind has one question- when will the colleges re-open? Or for some students, how would their college re-open in the near upcoming month? We are making a model that predicts the chances of a college re-opening based on the data related to its area and related covid cases. In addition to this, we are looking forward to adding some extra features like predicting the much anticipated third wave, its impact and how long would it last. Predicting the third wave then acts as a feature for our target prediction. As we all are aware, students are growing increasingly frustrated over the reopening of their respective institutes, but along with this in the back of their heads, they fear the arrival of the third wave. This is where our model would come in handy. The main purpose of our model would be to predict the reopening of different institutes. It would take in different datasets like the vaccination numbers of the institution's faculty members and the students, the Covid count in the area and city of the institution, the Covid status in the region from where the majority of the students come from, etc. The same data would be taken in and compared to the same datasets from the time just before the arrival of the second wave for the prediction of the third wave. These predictions would be really helpful for the students around the country who are waiting for their college to reopen.

2.1 What motivated us

The past 2 years have been really tragic for most of us. Since the arrival of Covid in March 2020, the world has changed dramatically. The lockdown commenced, followed by the tragic second wave. The most affected sector in the world has arguably been the education sector. As the last remorse, all institutions started in the online mode around the globe. Initially, it was very well received by the students, but slowly they started to face the consequences of this- decreased physical activities, increased load on the eyes, decreased practical knowledge, etc. Most of the students, including us, have been eagerly waiting for over a year now for our university to reopen. This has been the main motivation for us to work on this. Through this model, we will be able to help millions of people around the country to get an 'ac-

curately vague' idea of when their institutions would open, hence would help them prepare better for the near future.

2.2 Why are we solving this

There are a lot of sites available which provide us with Covid statistics like the case counts and vaccination counts. Also, there are a lot of websites showing different studies regarding Covid predictions. But there is not a single model present right now which can be used by the people themselves to get their own output corresponding to the place they are concerned about. Our model would solve this problem as it would provide a simple model to the people in hand through which they will be able to get the data according to their needs.

2.3 Background Knowledge

The background research mostly includes all the projects launched by the government during the pandemic, including Cowin, Arogya Setu, and MyGov.in. The listed sites/apps provide us with valuable data of the total number of cases active, total samples tested, total deaths, discharge of a given area, and that of the whole of India. These stats help people be aware of the current Covid-19 conditions in their area. They also include the vaccination stats including the total population vaccinated, age-wise vaccination numbers, and state-wise numbers as well. Similarly, there are a few additional initiatives based on machine learning and artificial intelligence that provide other sorts of updates regarding Covid-19, such as death rates, patient deterioration, current case counts, and future projections based on a statistical analysis of present conditions.

3 Related Work

Effective SARS-CoV-2 screening allows for a speedy and accurate diagnosis of COVID-19, reducing the load on healthcare systems. In order to evaluate the risk of infection, prediction models that integrate many features have been developed. The students are waiting patiently to hear from colleges and universities about the reopening. Several machine learning models are underway that forecast and inform present covid instances. Covid19_projections made by Youyang Gu is one such project. This is a complete library that includes not only the most recent but also all previous updates. In this project, it took into account the infection estimation, r values, reich forecasts, test targets, this machine learning-based project predicts future covid cases in all predicting of another wave hitting a certain area. The underlying stimulator used to generate the predictions is also developed by Youyang Fu, yyg-seir-stimulator. The best-learned values of the various parameters for each region are likewise stored in the simulator repository. Researchers in artificial intelligence (AI) are honing their skills in constructing mathematical models for researching the pandemic utilizing data from around the country. Using real-time data from the Johns Hopkins dashboard, machine learning models were used in tandem with a forecast of predicted COVID-19 reachability over the nations. The dataset was obtained from Johns Hopkins University's official repository3. Daily case reports and daily time series summary tables make up this data. They used time-series summary tables in CSV format for the study, with three tables for confirmed, deceased, and recovered COVID-19 cases, each with six properties. Province/state, country/region, last update, confirmed, death and recovered cases are only a few examples. The CSV files can be found in Github4 repositories. In parallel, state-of-the-art mathematical models based on machine learning are chosen for a computational process to forecast virus spread, for example:

- Support Vector Regression (SVR)
- Polynomial Regression (PR)
- Deep Learning regression models
- Artificial Neural Network (ANN)
- Recurrent Neural Networks (RNN) using Long Short-Term Memory (LSTM) cells.

Also one of the most important and successful similar projects was of our country's Cowin, Arogya Setu

and the 'mygovt.in' website, the site shows the total number of cases active, total samples tested, total deaths, discharged of a given area and that of the whole India. These stats help people be aware of the current Covid-19 conditions in their area. Similarly, there are plenty of other projects based on machine learning and artificial intelligence which give various types of updates about Covid-19, let it be mortality rate, deteriorating conditions of patients, current number of cases, and future predictions based on the statistical analysis of current conditions. And when we ensemble all of the results, a student can get an idea about their college re-opening, judging by all the factors.

4 Methodology

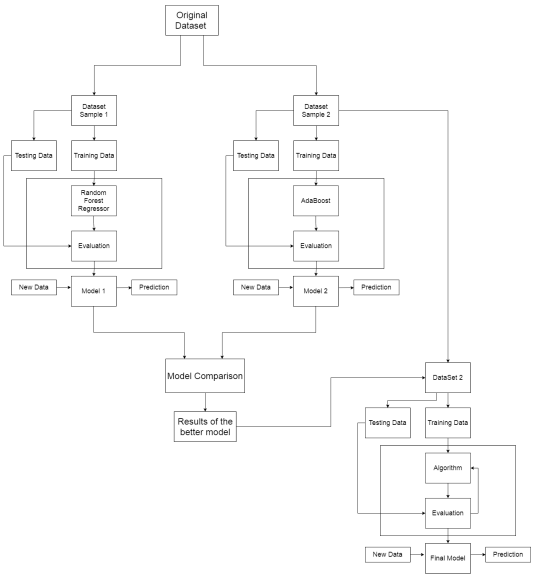

Figure 1: Block Diagram for our approach

We used the Drop function to eliminate the columns that were not relevant, such as time, confirmedindiannational, and confirmedforeginnational, from the dataset we imported from Kaggle. Then, to compare the two datasets, we imported another one and chose the better of the two. The date's type was changed from object to datetime64 as a result of this. We also split the month and year columns from the date column in order to access each individual record and removed the original date column. We then divided all of the states into distinct columns for data segregation. This was done so we could keep track of the unique values. We imported minmax from Sklearn. Finally, for comparability purposes, we normalized the confirmed cases column. State/union territory, number of cured patients, number of confirmed cases, number of fatalities, mortality rate, year, and month are the features utilized.

We divided the entire dataset into two samples, then used the first sample dataset in model 1 and the second sample dataset in model 2. Random Forest is used in the first model, whereas Adaboost is used in the second. Then we compare and contrast the accuracy scores and assessment metrics of both models to determine which is the best fit. Our target will be the total number of confirmed cases.

RANDOM FOREST- Individual decisions trees are combined to make a random forest. Random Forest models are a type of non-parametric model that may be used for both classification and regression. They are one of the most often used ensemble techniques, and they fall within the Bagging method group. Random Forest models combine the flexibility and strength of an ensemble model with the simplicity of Decision Trees. In a forest of trees, we overlook a tree's high variation and are less concerned with each particular aspect, allowing us to build finer, bigger trees with greater predictive power than a manicured tree. Although Random Forest models do not have the same interpretability as single trees, they perform far better, and we don't have to worry as much about fine-tuning the forest's parameters as we do with individual trees.

ADABOOST- Boosting comes in handy when nothing else does. Many individuals nowadays utilize XGBoost, LightGBM, or CatBoost to win Kaggle or Hackathon events. The initial step into the realm of boosting is AdaBoost. One of the earliest boosting algorithms to be used in solving problems was AdaBoost. Adaboost makes it possible to merge several "weak classifiers" into a single "strong classifier." AdaBoost is best used to improve decision tree per-

formance on binary classification issues. AdaBoost is a machine learning method that may be used to improve the performance of any other machine learning technique. It works well with students who are struggling. On a classification task, these are models that reach accuracy slightly above random chance. It's been dubbed discrete AdaBoost recently because it is utilized for classification rather than regression.

ALGORITHM WHICH PROVIDES BETTER RESULTS- Both algorithms' assessment metrics on R2 scores show that Adaboost produces superior results. Because AdaBoost has a higher R2 score, it is a better fit for our project.

5 Results and Analysis

Working with our first model, we used Random Forest regressor as the algorithm. For our first testing, the result was sub-optimal henceforth we modified the hyper parameters of the regressor seeking to attain better results. We tweaked the max depth a few times to observe which depth delivers us the best results. Here you can see the table for the same:

Max Depth	Score
2	0.927
3	0.9278
4	0.94
6	0.918

For our second model, we used AdaBoost as the algorithm. The results were certainly better than the previous model but we still worked with parameters to find improvement in the results. The following table gives an insight on the effect of change in n_estimators for the model:

$n_estimators$	Score
100	0.97
150	0.978
200	0.98
400	0.95

You can visualise the changes in results for both the models in a graphical format.

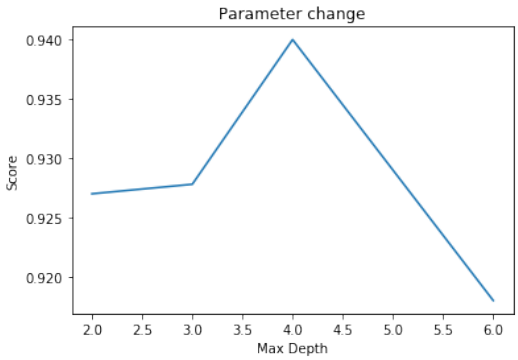

Figure 2: Model 1 scores

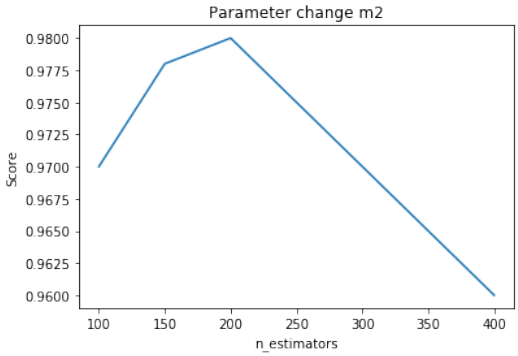

Figure 3: Model 2 scores

We changed a few other parameters for both the models to arrive at the best results for each to then compare the models to each other.

As read in one of the related works, that is, Analysis and Predictions for Covid-19(India), we observed the data analyst used the features such as states/regions, total confirmed cases, total cured, total deaths, date, etc to forecast the upcoming Covid wave. For their predictions, they have used algorithms such as Decision Tree Regressor, Random Forest, Linear Regression Model. Their model shows an accuracy of around 95%. For our model, we have used similar features as those mentioned above. We too have used one of the algorithms which they have, that is, random forest. For reference, you can see the results of their model compared with our below.

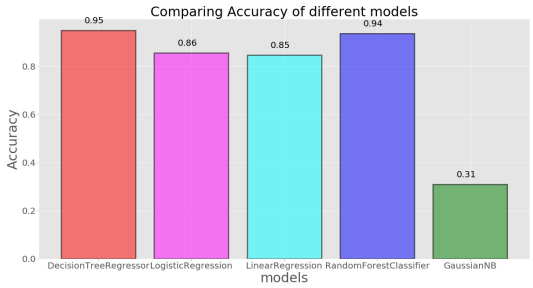

Figure 4: Related Model scores

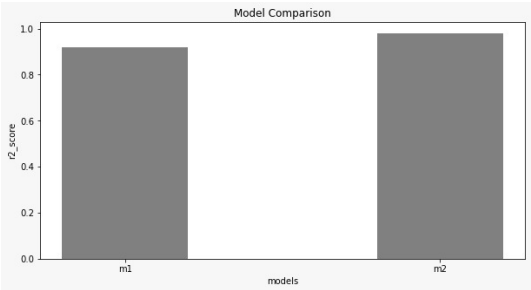

Figure 5: Comparison with our scores. (M2 represents their model here)

As per the two algorithms in our own project, the first as mentioned earlier is random forest regressor which resulted with an R2 score of around 0.92 and as for our other algorithm we used AdaBoost. Upon comparing, we observed that our second algorithm namely AdaBoost resulted in a better accuracy. It produced an R2 score of around 0.98. You can find the comparison of the two algorithms below.

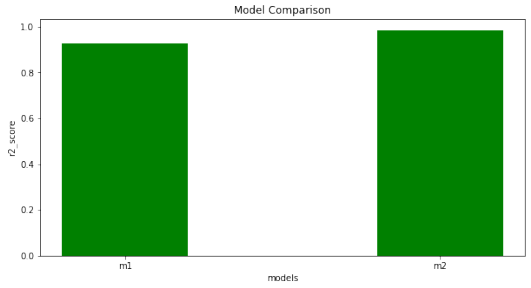

Figure 6: Scores of our two models

It can be seen that our second model provides a better accuracy in comparison to the previously mentioned model (Analysis and Predictions for Covid-19).

6 Conclusion

We are working on a model that forecasts the likelihood of a college reopening based on data from the area and similar situations. We are also excited to include some new features, such as projecting the much-anticipated third wave, its impact, and how long it will persist. The ability to forecast the third wave thus becomes a characteristic of our target prediction. These forecasts would be extremely beneficial to students around the country who are awaiting the reopening of their colleges. The full dataset is divided into two samples, with the first sample dataset used in model 1 and the second sample dataset used in model 2. The accuracy scores and assessment metrics of both models are then compared and contrasted to decide which model is the best match. There are several websites that give Covid information such as case counts and immunization counts. However, there is currently no model available that individuals may use to generate their own output relating to the location they are worried about. Our solution would tackle this problem by providing a simple model to the individuals in charge, allowing them to obtain data tailored to their needs. Finally, we determine which model is better for predicting the third wave.

7 References

- covid19projections
- Ai System for covid predictions
- Covid prediction ML algorithm
- covid19projections
- Analysis and Predictions for Covid-19(India)