Machine Learning for COVID-19 Data Analysis

Project report for DS-203, Programming for Data Science, Autumn '21

Aziz Shameem
Dept. of Electrical Engineering
IIT Bombay
[redacted]@iitb.ac.in

Kalash Shah
Dept. of Mechanical Engineering
IIT Bombay
[redacted]@iitb.ac.in

Shivam Patel
Dept. of Electrical Engineering
IIT Bombay
[redacted]@iitb.ac.in

***Abstract*—The goal of this study is to investigate and compare the variations in COVID-19 cases and deaths in a select group of countries, in an effort to understand which country was efficient in controlling the rising virus, and to elaborately analyse the performance of India in curbing the spike in cases. We used four regression techniques to predict the number of deaths as a function of other COVID statistics. The second part of the paper focuses on analyzing the ICU admission possibility for a COVID patient, given a certain set of medical parameters. We have employed five classification algorithms here to accurately predict whether a patient needs to be admitted to an ICU or not.**

***Index Terms*—COVID-19, Daily-Cases, Medical Predictions**

I. INTRODUCTION

According to the World Health Organization (WHO, 2020), the coronavirus (COVID-19) outbreak which emerged from central China in late December 2019 has spread to over 218 countries, areas or territories, and has resulted in over 250 million confirmed cases as well as over 5 million deaths across the globe as of October 2021. Given the widespread and ongoing transmission of the novel coronavirus worldwide, the WHO officially declared it a pandemic on March 11, 2020.

The rapid spread of the unprecedented COVID-19 pandemic has put the world in jeopardy and changed the global outlook unexpectedly. Many countries have adopted various ways to deal with the pandemic. Some countries were able to effectively control or restrict the spread of cases, while others, because of a large population, poor infrastructure, or lack of structured policies in general, performed poorly in these testing times. Thus, to limit the spread of coronavirus, it was essential that all the countries came together and adopted the best practices to deal with the pandemic, and minimize the casualties.

Analytics and Data Science techniques played a huge role in predicting the future course of the pandemic and helped governments worldwide, to channelize their resources and impose restrictions, including lockdowns. Data Science algorithms also helped in predicting the rate of mutation of the coronavirus strains and helped assess the severity of disease caused by different variants.

We attempt to solve the problem of analysing the COVID-19 data in a selected group of countries and compare their relative performances through Exploratory Data Analysis. We also analyse the COVID-19 situation in India and various states for a fixed duration of time through different visualisation techniques.

It is important for any country to predict the number of deaths beforehand and optimise the resources to save lives. Thus we attempt to predict the number of deaths in India using different regression techniques. Likewise, a system that can predict the possibility of ICU admission can be very vital as it becomes easier to assign the limited number of ICU beds in a hospitals, based on the confidence with which the program classifies the data-point. Here we use different classification algorithms for our purpose and compare them based on their accuracy levels.

There is a vast amount of data available for the coronavirus, but time-to-time concerns have been raised on the authenticity of the reported data, and possible manipulation of the numbers which may severely distort the accuracy of data science-led predictions. Also, the instantaneous statistics heavily depend upon the variant of the coronavirus active in a particular region, which is very difficult to incorporate in a dataset. Additionally, any computation on a dataset with more than a million points is generally expensive. Thus efforts should be directed to optimise and tune the basic frameworks, rather than using very complex models.

II. PRIOR WORK

Given the urgency and dynamics of the situation, a lot of work has been done on COVID-19 data analysis, and even more, is in pipeline.

Research work done by universities:

1) Carnegie Mellon University: Supporting decision-makers and the public by constructing geographically-detailed, real-time indicators of COVID-19 activity and using them to produce interactive visualizations and short-term forecasts.
2) University College London: Modeling the prevalence of COVID-19 and understanding its impact using publicly-available aggregated, anonymized search trends data.
3) Tel Aviv University: Developing simulation models using synthetic data to investigate the spread of COVID-19 in Israel.
4) Indian Institute of Science, Bengaluru: Mitigating the spread of COVID-19 in India's transit systems with rapid

testing and modified commuter patterns.

5) Indian Institute of Technology, Gandhinagar: Modeling the impact of air pollution on COVID-related secondary health exacerbations.

Projects undertaken by individuals:

1) Statistical analysis of the novel coronavirus (COVID-19) in Italy and Spain by Jeffery Chu: Using data of the daily and cumulative incidence in both countries over approximately the first month after the first cases were confirmed in each respective country, Jeffery analysed the trends, modelled the incidence and estimated the basic reproduction value using two common approaches in epidemiology—the SIR model and a log-linear model- Report

2) The probability of being diagnosed with a COVID-19 infection, using medical symptoms by Yazeed Zoabi, Noam Shomron: The duo established a machine learning approach that trained on records from 51,831 tested individuals (of whom 4,769 were confirmed COVID-19 cases) while the test set contained data from the following week (47,401 tested individuals of whom 3,624 were confirmed COVID-19 cases). Their model predicts COVID-19 test results using only eight features: gender, whether age is above 60, known contact with an infected individual, and five initial clinical symptoms- Report

III. DATASETS

In this report, we have obtained and used datasets on COVID-19 cases in nine countries: India, USA, Brazil, Indonesia, Turkey, Ukraine, Italy, Mexico, and Germany. We have performed a detailed Exploratory Analysis on COVID cases and the vaccination campaign in India. In the second part, we have fitted several ML classification models on a data set, in an effort to try and predict whether a patient was admitted to the ICU, based on several medical parameters.

Procedure :

Part 1

For the EDA part, we obtained several datasets on our topic from various sources, removed NaN values, and devised several types of visualisations to better understand the data. We then drew conclusions based on the graphs/ plots obtained, trying to explain the nature of and variations in the plots.

Part 2

For the ML part, we downloaded the required datasets from Kaggle. We standardised the regression dataset (0 mean and unit variance). The classification dataset was already normalized and feature-scaled by the author. We then performed pre-processing, where we removed columns with negligible variance, replaced NaN values with the column-mean, and removed columns giving undue importance to outliers, or instilling redundancy in the data.

A representation of the final table used :

	AGE_ABOVE65	AGE_PERCENTIL	GENDER	DISEASE GROUPING 1	DISEASE GROUPING 2	ICU
0	1.0	60.0	0.0	0.0	0.0	0.0
1	1.0	60.0	0.0	0.0	0.0	0.0
2	1.0	60.0	0.0	0.0	0.0	0.0
3	1.0	60.0	0.0	0.0	0.0	0.0
4	1.0	60.0	0.0	0.0	0.0	1.0
...	...	...	...	...	...	...
1920	0.0	50.0	1.0	0.0	0.0	0.0
1921	0.0	50.0	1.0	0.0	0.0	0.0
1922	0.0	50.0	1.0	0.0	0.0	0.0
1923	0.0	50.0	1.0	0.0	0.0	0.0
1924	0.0	50.0	1.0	0.0	0.0	0.0

1925 rows × 6 columns

IV. ANALYSIS PIPELINE

A. *Comparison Between Absolute and Relative Parameters*

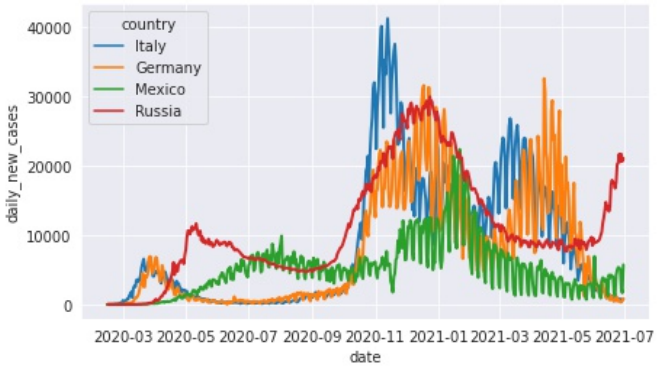

Fig. 1. Absolute number of daily cases in different countries

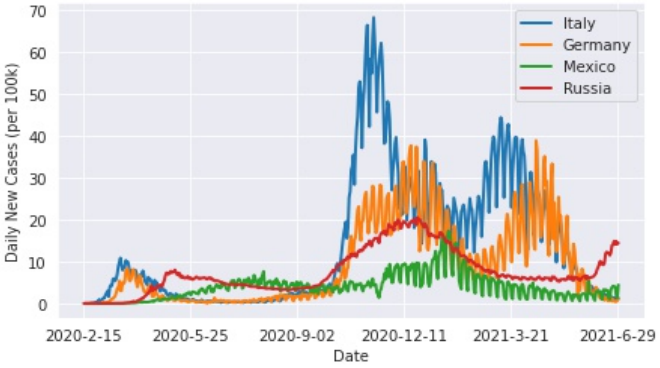

Fig. 2. Relative daily cases in different countries(per 100k pop)

For comparing the spread of COVID-19 virus in different countries, there are two trains of analysis to follow. We observe that the absolute number of cases in different countries depicts the magnitude of spread of the virus, taking into consideration the population. This gives us an idea about the stress on the healthcare system of the country, the estimate about health expenditure of the country as well as the number of patients in the country. On the other hand, relative number of cases gives us the ability to compare the spread of COVID-19 in different countries, irrespective of the population. Dividing the cases by population, and scaling it to per 100k population gives us the ability to compare and contrast the efficiency of any country

in controlling the spread of COVID-19. The relative scaling of parameters gives us an understanding of the quality and relative quantity of healthcare facilities in different countries. As plotted in the above graphs, the absolute number of cases in Russia depicts trend lines almost identical to that of Germany, but the relative number of cases in Russia are always less than that in Germany. This fact depicts that in spite of a large number of cases in Russia, the overall negative effect was not as devastating as it was in Germany.

B. *Relative Cumulative Cases in Different Countries*

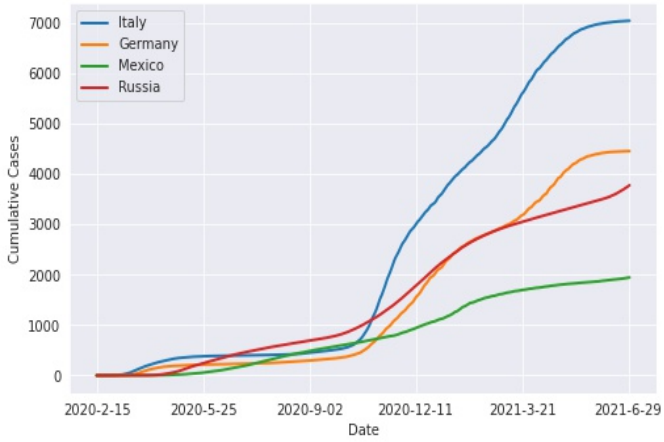

Fig. 3. Relative cumulative cases in different countries(per 100k pop)

The line plots for relative cumulative cases(per 100k population) depict that the relative spread of COVID-19 in Italy has been the greatest out of all the countries so far. By the end of the time duration in consideration, almost 7% of the population in Italy had been infected by the coronavirus at least once. While on the other hand, this number was 4.5% in Germany and 3.8% in Russia. Out of the countries in consideration, Mexico had the least spread of COVID-19 in its population, resulting in less than 2% overall population being infected. It is imperative to note that the total cumulative cases had a slow rate of growth in the initial wave of COVID-19 across the world, but nearing the end of 2020, almost all countries saw a steep increase in cumulative cases, propounding the fact that the second wave had more severity and mortality rates as compared to the first wave.

C. *Distribution of Daily Cases*

The violin plot describes the distribution of the daily cases in different countries during the time duration in consideration. For Mexico, majority of the days had seen very low number of new cases. But Italy and Germany having a high and lean violin plot suggests that Italy and Germany had constantly seen new cases almost everyday, and the maximum number of daily new cases was also higher than that in Mexico. Russia's distribution of daily cases was in between that of Italy, Germany and Mexico, both in terms of maximum daily cases and number of days with zero/very low new cases.

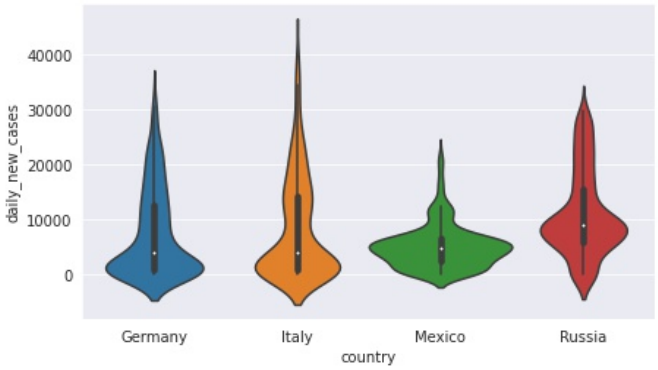

Fig. 4. Violin plot for distribution of daily cases

D. *Mutual Trends between Daily and Cumulative Cases*

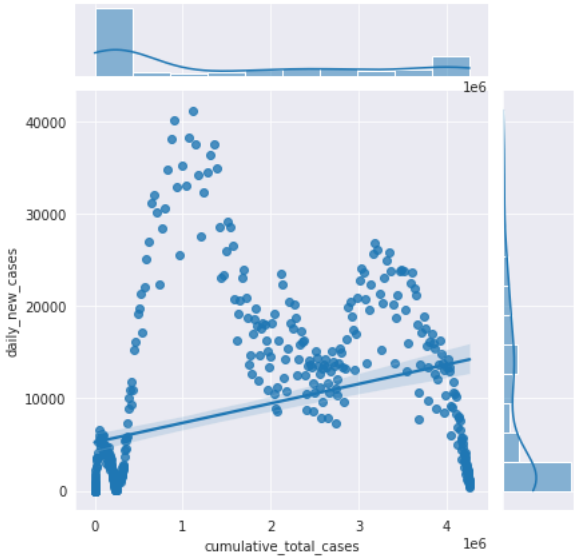

Fig. 5. Relation between Daily New Cases and Cumulative Cases (for Italy)

The joint plot between daily new cases and cumulative cases reveals a striking relation between the cumulative cases and daily new cases. The two peaks/ maximas in the joint plot are reminiscent of the almost well-defined and differentiated 'waves' of the COVID-19 pandemic. Daily cases reached peaks during the intense periods of individual waves of the pandemic, and the cumulative cases also rose proportionally in those periods. The distribution of daily cases and cumulative cases in the bar plots show that cumulative cases were almost constant during the beginning and end of the duration in consideration, outlining zero/low number of cases in that time. Constant cumulative cases define a plateau in the line plots for each country, as can be observed from the line plot on cumulative cases shown previously.

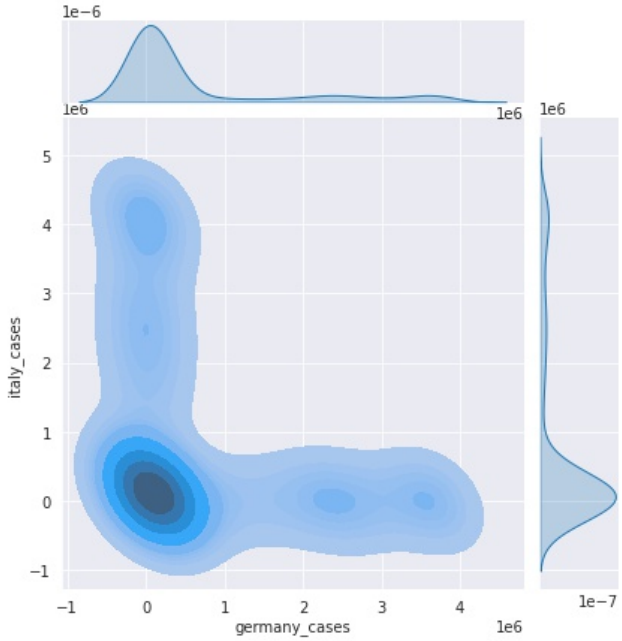

Fig. 6. Relation between Cumulative Cases in Italy and Germany

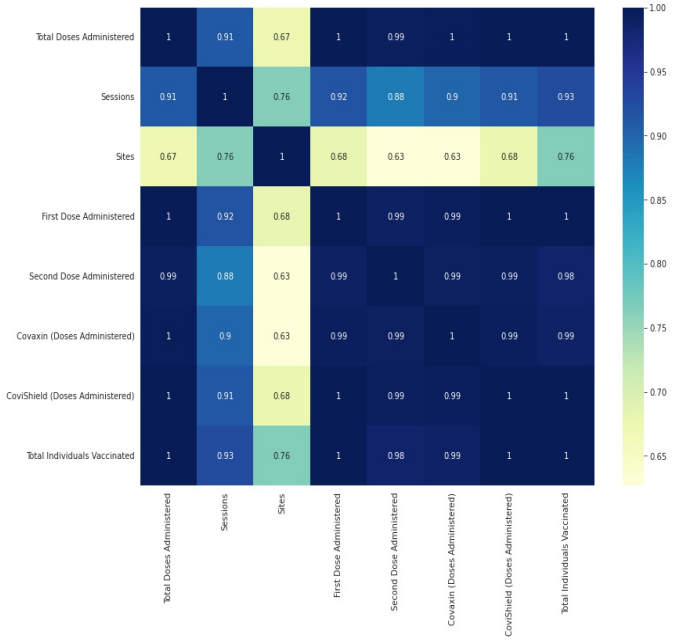

Fig. 7. Correlation between vaccination parameters in India

E. Correlation between Cumulative Cases in different countries

The Kernel Density Estimation for the distribution of the daily cases in Italy and Germany portray the intuitive result that the number of daily cases for most of the countries has been on the lower side of the range of daily new cases. The maximum density of the contour plot is towards the origin of the graphs, showing that the majority of days had very few new cases in each country. The distribution of daily cases in each of the countries shows this fact individually.

F. Correlation between Vaccination Parameters in India

The correlation matrix shows that there is a high correlation between the number of first and second doses administered, which highlights the efficient and effective vaccine drive which the government has undertaken. Also, as expected, there is a complete correlation between the number of Covishield and Covaxin shots administered, depicting that there wasn't any bias, shortage or excess of any type of vaccine, which would have led to unequal trends in administering the doses of different vaccines. Also, the highly correlated bottom right 5x5 matrix supports this observation.

G. Cumulative Cases in Indian States

Plotting the confirmed cases in different states of India reveals startling observations. Firstly, we notice that the cases in different states are of different magnitude because of factors such as population, healthcare facilities and other socio-economic parameters. But more importantly, there is a plateau region forming in the total confirmed cases in different states throughout the time period under consideration. This

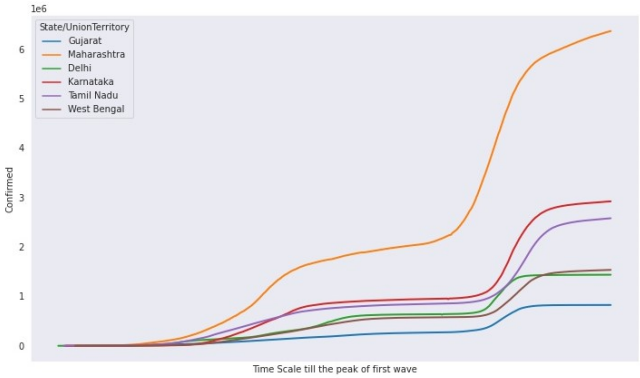

Fig. 8. Cumulative cases of COVID-19 in Indian states

corresponds to the time between the two 'waves' of COVID-19 spread in India. Even more strikingly, we observe that the rise in the second wave was far greater than that in the first wave, which testifies the virulence of the newer Delta strain in the second wave.

H. Correlation between Daily Cases Worldwide

Various studies are conducted and research has been undertaken all around the world to predict the spread of COVID-19 in one country given its spread in another country. Various theories are put forth, based on the biological strain of the infection in any country, climatic factors like temperature and humidity, geographical proximity of a country, inflow of tourists into one country, foreign laws and travel restrictions, etc. The above correlation matrix depicts the relation in daily new cases in nine different countries. We notice that the daily cases in Italy, Germany and Ukraine are closely

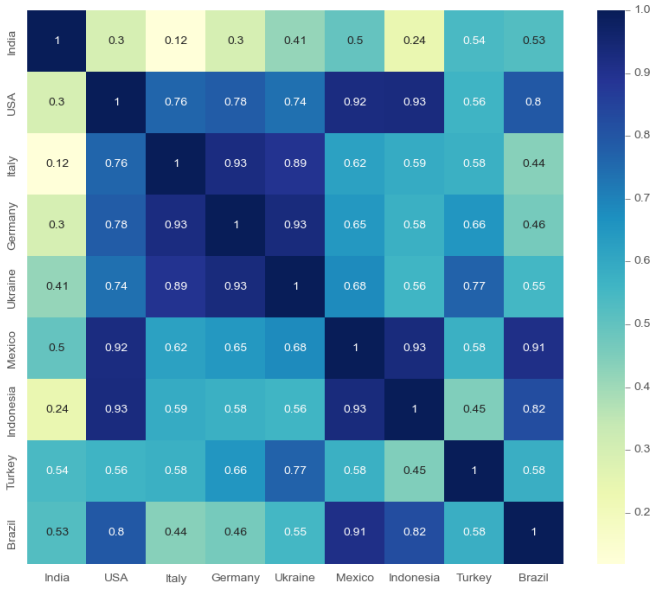

Fig. 9. Correlation Matrix for international daily new cases

correlated to each other, i.e. trends in daily new cases were very similar in those countries. This can be accredited to geographical proximity, almost similar healthcare facilities, similar populations, etc. On the other hand, we also notice that cases in India are almost uncorrelated with any other country to a significant extent. This could be resulting from different strains of the coronavirus in these countries, different climatic conditions, different healthcare facilities, etc.

I. Comparison of Monthly Cases in Different Countries across the Time Scale

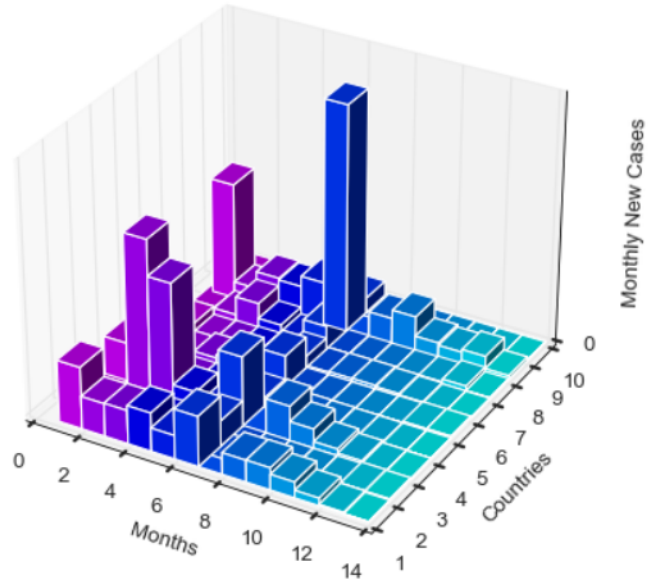

Fig. 10. 3D plot of monthly cases in different countries

List of countries - Brazil(1) Turkey(2) Indonesia(3) Mexico(4) Ukraine(5) Germany(6) Italy(7) USA(8) India(9) | Months - February' 20 to June' 21 (ordered from 14 to 0 in the graph)

We observe that the number of monthly cases of all countries grew steeply in October 2020. The most affected country was USA. Then for some months, the cases dropped sharply and remained zero or very low. But, after the first wave, the secondary waves came independently and staggered in time in different countries, with different severity and durations.

J. Partition of Worldwide Cases Country-wise

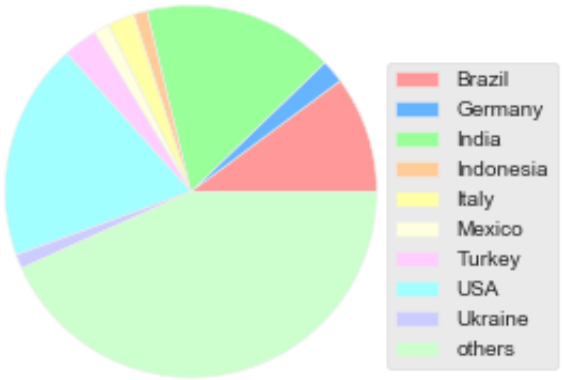

Fig. 11. Pie chart representing breakup of international cases

For the nine countries in consideration, we observe that India and USA are responsible for the majority of International COVID-19 cases. Given the fact that the population of USA is almost a quarter of that in India, an equal number of COVID-19 cases in both countries depict a very grave situation in USA compared to that in India. India, Brazil and USA form up a larger proportion of total cases worldwide. This can be attributed to larger populations of those countries, and more severity and penetration of coronavirus in the society.

V. MACHINE LEARNING MODELS

Part 1 - Regression

In the first part, we employed a total of four regression models for our purpose, namely linear, polynomial, lasso and ridge regression techniques including implementing Batch Gradient Descent from scratch.

General expression for target variable prediction using regression:

$$\hat{y_i} = w_0 + w_1 x_1 + w_2 x_2 + .. + w_n x_n \tag{1}$$

where w_i denote the coefficients and x_i the data points.

Optimizer - Batch Gradient Descent : Based on variable learning rates for α, we employed the following equation :

$$w_i' = w_i - \alpha \frac{\partial L}{\partial W_i} \tag{2}$$

where w_i' and w_i denote the updated and current coefficients respectively, and L is the cost function, for different number of iterations. Here we notice that after $\sim$ 80 iterations the loss function approximately remains constant.

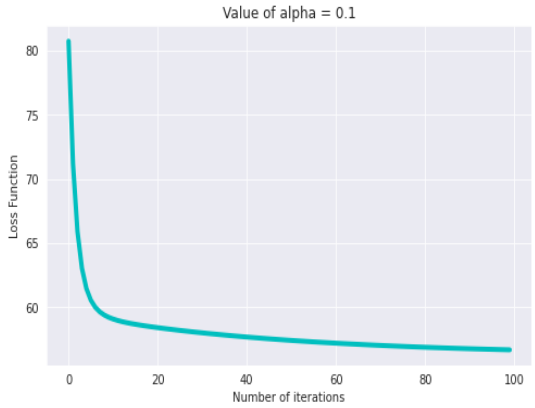

Fig. 12. Plot of Loss function (OLS) vs number of iterations for α= 0.1

Ordinary Least Squares : Ordinary Least Squares (OLS), perhaps one of the simplest regression models, is a type of linear least-squares method for estimating the unknown parameters. OLS minimizes the sum of the squares of the differences between the observed dependent variable (values of the variable being observed) in the given dataset and those predicted by the linear function of the independent variable. The above given figure shows a strong linear relation between the number of cured cases and the number of deaths.
OLS has the following loss function

$$L = \frac{\sum_i^n (y_i - X^T w)^2}{n} \tag{3}$$

Polynomial Regression : Polynomial Regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted by $E(y|x)$.The general form of predicted target variable in polynomial regression:

$$\hat{y_i} = w_0 + w_1 x_i + w_2 x_i^2 + .. + w_n x_i^n \tag{4}$$

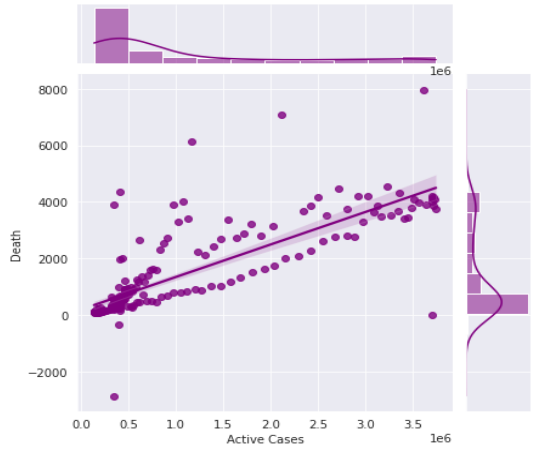

Fig. 13. Joint plot of Deaths vs Active Cases

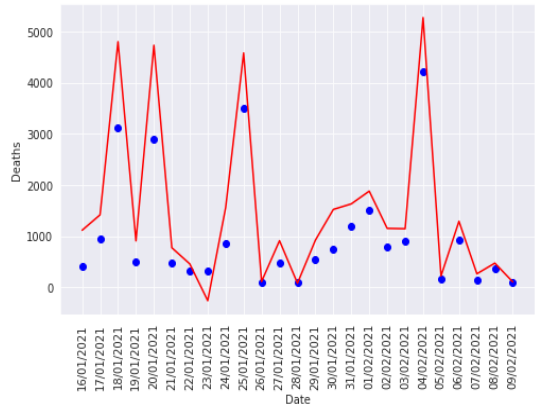

Fig. 14. Polynomial regression of degree 2

For degree = 2, we used the instance of the class `PolynomialFeatures()`. Further, we employed the `.fit_transform()` method to appropriately transform the train and test data according to the input of `.fit()` method of polynomial regression.

LASSO Regression : Lasso regression is a regularization technique that uses the L1-norm of the weights. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). Loss function in LASSO regression:

$$L = \frac{\sum_{i=1}^n (y_i - \hat{y_i})^2 + \lambda \sum_{j=0}^p |w_i|}{n} \tag{5}$$

We used hyperparameter tuning for lasso regression using `GridSearchCV()` method via the `Pipeline()` method which sequentially applies a list of transforms and a final estimator, as we used polynomial features for prediction.

Ridge Regression : Ridge regression is a regularization technique that uses the L2-norm of the weights. When multi-collinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. Loss function in Ridge regression:

$$L = \frac{\sum_{i=1}^n (y_i - \hat{y_i})^2 + \lambda \sum_{j=0}^p w_i^2}{n} \tag{6}$$

Again, we used hyper parameter tuning for ridge regression using `GridSearchCV()` method via the `Pipeline()` method.

Part 2 - Classification

In the second part, we discuss five of the many ways to classify data, and then use these algorithms on a data set to try and predict whether a patient needed to be admitted to the ICU, given a set of medical parameters.

Hypothesis Testing : We performed a χ^2 - contingency test, based on the α value as 0.05, to check the influence of certain medical factors on the ICU admission.

Logistic Regression : In the Machine Learning world, Logistic Regression is a kind of parametric classification model. This means that logistic regression models are models that have a certain fixed number of parameters that depend on the number of input features, and they output categorical prediction, like for example if a plant belongs to a certain species or not. In Logistic Regression, we don't directly fit a straight line to our data, instead, we fit an S-shaped curve, called Sigmoid, to our observations

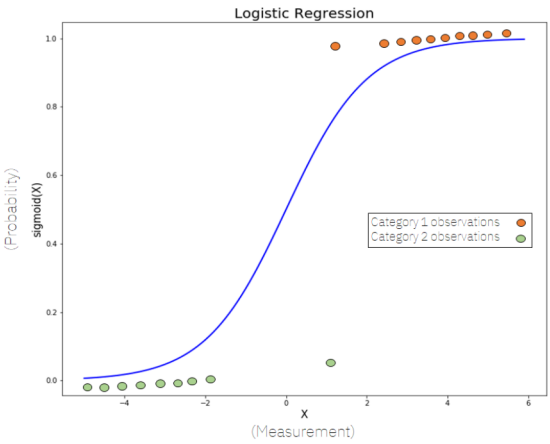

Fig. 15. Logistic Regression

In simple terms, logistic regression models, based on some input features, classify inputs into two or more baskets.

Support Vector Machine : SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.
At first approximation what SVMs do is to find a separating line(or hyperplane) between data of two classes. SVM is an algorithm that takes the data as an input and outputs a line/ plane/ hyperplane that separates those classes if possible. It is often described as a large margin classifier, since it chooses the partition that maximizes the distance from it to the closest data point.

Random Forest Classifier : Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a reliable and accurate result most of the time. It is also one of the most widely used algorithms, because of its simplicity and diversity.
Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.
Put simply, random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.
Random forest adds additional randomness to the model while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Gradient Boosting Classifier : Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness in classifying complex datasets.
Gradient boosting systems have two other necessary parts: a weak learner and an additive component. Gradient boosting systems use decision trees as their weak learners. Regression trees are used for the weak learners, and these regression trees output real values. Because the outputs are real values, as new learners are added to the model the output of the regression trees can be added together to correct for errors in the predictions. The additive component of a gradient boosting model comes from the fact that trees are added to the model over time, and when this occurs the existing trees aren't manipulated, their values remain fixed.

Multilayer Perceptron : A multilayer perceptron (MLP) is a class of feedforward Artificial Neural Network (ANN). An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

VI. INFERENCES AND EVALUATION

A. Regression Models

As we used the `train_test_split()` method of the `sklearn.preprocessing()` module, each runtime would produce a different result due to a variable splitting. These values have been averaged over five runtimes.
On training the dataset with the four regression models, here is the result obtained on the test set, for the number of deaths :

Regression	R^2 score	RMSE (Train)	RMSE (Test)
Linear	0.854	993.531	783.723
Polynomial	0.744	4258.634	3801.276
LASSO	0.799	216587.986	191365.640
Ridge	0.803	386380.629	356320.621

1) Generally, R^2 score is a measure of goodness of fit. Thus, Linear Regression, the best performing algorithm could account for $\sim$ 85 % of the variability in data.
2) RMSE accounts for the separation between the observed data and the predicted data. Thus, Linear Regression with the least RMSE proves to be the best fit amongst the four for this situation.
3) As the RMSE of the train and test data are quite similar for each of the four regression techniques, it indicates a balanced fit and avoids excessive overfitting or underfitting of the train data.
4) This might have happened because we naturally expect the number of deaths to be directly proportional to the new cases, active cases and negatively correlated to the vaccinated persons and discharged cases. Thus for the given dataset, a simple algorithm like Linear Regression performs the best.

B. Classification Models

On training the dataset with the five classification models, here is the result obtained on the test set :

Classification Algorithm	F1 Score	Accuracy
Gradient Boosting Classifier	0.904860	90.6494%
Random Forest Classifier	0.876072	87.7922%
SVM Classifier	0.864541	87.0130%
MLP Classifier	0.838285	84.9351%
Logistic Regression	0.650888	84.6753%

1) F1 score is defined as the harmonic mean between precision and recall. It is used as a statistical measure to rate performance.
2) Accuracy indicates the fraction of test cases that were classified correctly.
3) All the models have accuracies in the range of 84-90 %, which indicates that the number and type of features are adequate, and are not prone to high bias or variance.
4) Gradient Boosting Classifier incorporates a number of learning algorithms, thus increasing the randomness and excluding the possibilities of any bias due to lack of randomness in initialization. It, therefore, outperforms the other models.
5) In the Logistic Regression model, we have fixed the threshold at 0.5, with no room for margin(as opposed to other models like SVM or GBC). This may explain its inferior performance.

C. Evaluation Metrics

We have used the following evaluation metrics:

1] RMSE

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2} \tag{7}$$

2] R^2 Score

$$1 - \frac{RSS}{TSS} \tag{8}$$

3] Accuracy

$$\frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

4] Precision

$$\frac{TP}{TP + FP} \tag{10}$$

5] Recall

$$\frac{TP}{TP + FN} \tag{11}$$

6] F1 Score

$$\frac{2 * precision * recall}{precision + recall} \tag{12}$$

VII. FUTURE WORK

Preventive treatment of patients with predictive analysis and diagnosis of diseases from patients' body parameters is the goal of modern medicine. Continuation of this project would be directed in the direction of predictive diagnosis, which will not only enable doctors to treat even the mildest of symptoms, but also provide an estimate of the healthcare resources that would be required in the upcoming time. This can prove to be of vital importance in times such as pandemics. Our machine learning models can be extended easily to cover newer objectives, with far deeper and more complex algorithms and methods.
We would be able to achieve this when access to the vast amount of healthcare data is made public, and adequate steps are taken to ensure the authenticity of data. Additionally, deep neural networks can be utilised for this process to optimise the predictive accuracy.

CONTRIBUTION OF THE TEAM MEMBERS

This project was undertaken by Aziz Shameem, Kalash Shah and Shivam Patel, three sophomores pursuing a minor in Artificial Intelligence and Data Science. The project consists of three parts : General Covid-19 EDA including comparison among countries, an in-depth EDA of coronavirus cases and vaccinations in India, and machine learning. The ML part included training various regression and classification models on datasets, followed by comparison and documentation of their performances.
The first part was executed by Aziz and Shivam, the second entirely by Kalash, and the third part required combined efforts from all three.
This report, along with all of its figures and graphs, and the video submission, w made together by all three of us.

ACKNOWLEDGEMENT

We would like to thank our professors, Prof. Amit Sethi, Prof. Manjesh K. Hanawal, Prof. Sunita Sarawagi and Prof. S. Sudarshan for teaching us this course, without which we would not have been able to complete this project. We would also like to thank all the TA's involved with this course, who helped us in doing the various assignments.

REFERENCES

[1] "Hyperparameter Tuning" [Online]. Available : https://alfurka.github.io/2018-11-18-grid-search/

[2] "Polynomial Regression : Python Implementation" [Online]. Available : https://www.geeksforgeeks.org/python-implementation-of-polynomial-regression/

[3] "Model Evaluation Metrics" [Online]. Available : https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/

[4] "ICU Prediction Dataset" [Online]. Available : https://www.kaggle.com/afamos/covid-19-icu-admission-prediction-92-accuracy

[5] "Covid Analysis Dataset : Global" [Online]. Available : https://www.kaggle.com/josephassaker/covid-19-global-data-analysis-visualization/data

[6] "Covid Analysis Dataset : India" [Online]. Available :https://www.kaggle.com/n1sarg/covid-19-india-analysis/data

[7] "Covid Vaccinations, India Dataset" [Online]. Available :https://www.kaggle.com/sudalairajkumar/covid19-in-india?select=covid_vaccine_statewise.csv

APPENDIX

The Python notebook(s) which contain the entire code for the above analysis, along with all the datasets, and this report is present in the below given GitHub links.

- https://github.com/Kalash1106/DS-203-Project
- https://github.com/Aziz-Shameem/DS203-Covid-Analysis
- https://github.com/patel-shivam/Covid-19-Analysis