Analysing features and regression algorithms for predicting excess mortality rate

Hemant Rana[1]
e-mail: hemant.rana@stud.uni-goettingen.de

Abstract – **Since the beginning of Covid several lives have been affected and major changes to the healthcare infrastructure have been made. However estimating changes such as increasing bed counts in hospital, manufacturing of vaccinations, estimating dosages etc. and other aspects related to Covid pandemic are difficult to estimate. One such factor is *excess mortality rate* also known as p-score. It estimates how many extra lives have been lost due to the covid infections. Estimating such a factor can help nations decide to allocate sufficient funds and implement appropriate policies to tackle additional mortality due to covid.**

Keywords **– Regression, Covid-19, Data Analysis**

I. INTRODUCTION

A. The Pandemic

The global pandemic due to corona virus disease started in 2019 has posed to be a great threat to public health globally. It is a result of infection with severe acute respiratory issues that can be rooted back to patients who came into contact with seafood market in Wuhan City, China in the month of December 2019. Some similar findings have been made earlier in cases of SARS-Cov and MER-COV. Although fatality rate of SARS-Cov-2 or Covid 19 is $2\% - 3\%$ which is much less than MER-COV and SARS but the pandemic associated with it has been quite lethal [1]. W.e.f February 11, 2023 the covid-19 has infected $677,378,151$ people around the globe with nearly $6,781,126$ deaths while the remaining $649,914,743$ has recovered.

B. Excess Mortality Rate

It is a term employed in epidemiology and public health surveys and evaluations that indicates the additional number of deaths from all causes during an epidemic than we have expected normally . [2] has provided a study in which the number of deaths during the COVID-19 pandemic is compared to the expected deaths had the pandemic not occurred — It is an important quantity that can be calculated in several ways.

Excess mortality is an extensive measure of an overall impact of the pandemic on deaths than just the confirmed COVID-19 deaths. It signifies not only the confirmed deaths, but also COVID-19 deaths that were not correctly diagnosed and reported as well as deaths from other causes that can be characterized as an impact from the pandemic.

C. Measuring Excess mortality

It can be defined as a difference of reported and expected deaths that may be measured weekly or monthly.

$$ExcessMortality = ReportedDeaths - ExpectedDeaths \quad (1)$$

The parameter *Expected Deaths* can be difficult to estimate and can have various methodologies to be estimated, hence different values for a country for a specific week or month. One of the widely recognized method for calculating expected deaths proposed by Karlinsky et. al. [3] is a regression based method that is fitted on deaths encountered between $2015 - 2019$ and have projected expected deaths for $2020 - 2022$. This method has been adopted for the study as it covers several crucial aspects of mortality i.e. year-on-year trend and seasonal variation.

D. P-Score

P-score can be interpreted as a measure of excess mortality that can be used as a comparable statistic for countries. Since a country can have varied demographics than others, hence just measuring eq.(1) doesn't give a comparable value, hence it can be calculated using following equation [3]

$$P - score = \frac{reported\_deaths - projected\_deaths}{projected\_deaths} * 100 \quad (2)$$

It is given in percentage that indicates how much more the mortality rate has been for current week/month as compared to the previous week/month.

E. Regression

Regression is a statistical method for examining relationship between multiple variables. It is a type of supervised machine learning algorithm. Supervised machine learning algorithms are trained over data that are associated with an output label. Regression is well suited for time series data as it can be mapped for multiple inputs to a single input which can be suitable for analysing relationship between various parameters and p-score for a country.

This report deals with the following classifications of regression

1. **Linear regression**: the dependent variable is continuous in nature. The relationship between input and output follows a linear trend. The output can be estimated as

$$y = \sum \beta_n X_n + \varepsilon \quad (3)$$

Following assumptions are made for linear regression models:

(a) A linear relation between independent and dependent variables should exist.

(b) No outliers should be present.

(c) Error terms should be normally distributed.

(d) Parameters should not be auto-correlated.

Regression coefficients can be estimated as

$$\hat{\beta} = (X'X)^{-1}X'y \quad (4)$$

2. **Random Forest Regression** [4]: is a supervised learning algorithm based on random forest algorithm that harnesses the capabilities of ensemble learning. Ensemble learning can be defined as a method of combining multiple predictions from different models in order to increase accuracy. Random forest algorithm can be described as follows:

   (a) Pick random k data points from dataset.

   (b) Build a decision tree associated with these k data points.

   (c) Select N number of trees to be built and repeat steps a and b.

   (d) predict for a new sample using N trees and average across all predicted values.

3. **Gradient Boosting Regression**: Gradient boosting is one of the variants of ensemble methods where you create multiple weak models and combine them to get better performance as one holistic model. In Gradient Boosting regression the residuals from first iteration is minimized, based on the following function

$$r_1 = y - mean(y) \tag{5}$$

   For a prediction γ_1 and an initial prediction F_0 to reduce the residuals the following objective function is used, where ν is learning rate

$$F_1 = F_0 + \nu * \gamma_1 \tag{6}$$

4. **K-nearest neighbors regression**: It employs two implementations of K-nearest neighbors, one in which average of numerical target of K nearest neighbors while the other employs inverse distance weighted average of K nearest neighbors. It becomes impractical when number of independent variables increases. The neighbors for a specific cluster can be estimated based on the following distance function

$$d(X,Y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} \tag{7}$$

5. **XgBoost Regression**: Stands for Extreme Gradient Boosting and is based on the concept of ensemble learning. For Xgboost, γ can be defined as minimum reduction of loss required to to split a node, α is L1 regularization on leaf weights and λ is L2 regularization on leaf weights. Consider a residual $r \in \phi$ which is a set of residuals , then XgBoost algorithm follows given steps for building trees:

   (a) Calculating similarity score based on the following equation

$$score = \frac{(\sum_i^n r)^2}{n(\phi)} + \lambda \tag{8}$$

   (b) Calculate the gain factor that determines how the data will be split, let $f(x)$ signifies the similarity score for a tree x

$$gain = f(left\_sub\_tree) + f(right\_sub\_tree) + f(root) \tag{9}$$

   (c) for a given γ (also known as tree-complexity parameter is defined by the user), calculate $gain - \gamma$. If result is positive do not prune the tree and if result is negative then prune and subtract gamma from the gainvalue from the immediate parent.

   (d) Output value can be calculated as

$$output = \frac{\sum_i^n r}{n(\phi)} + \lambda \tag{10}$$

6. **Support Vector Regression**: Objective function in SVR is to minimize the coefficients using L2-norm.

$$Min(\frac{1}{2}\|w\|^2) \tag{11}$$

   with constraints

$$|y_i - w_i x_i| \leq \varepsilon \tag{12}$$

F. Early research

[2] and [3] have previously adopted regression based methodologies to calculate p-scores based on deaths occured between 2015-2019 and deaths occured between 2019-2022. Their data is based on death counts only, whereas p-scores can have strong correlation to some other parameters that can influence mortality due to the pandemic and this report primarily focuses on exploring such parameters and analysing which regression method is more accurate at predicting p-scores.

II. Experiment

A. Tools Used

1. **Language/Framework**: Python 3.9
2. **IDE/Environment**: Jupyter Notebook
3. **Package Manager**: pip
4. **Libraries Used**
   - Visualization: Matplotlib, Plotly
   - Machine Learning: Scikit-Learn, Xgboost
   - Data Processing: Pandas, Numpy

B. Dataset description

For the experiment three datasets have been used from [2]

- Full Covid-19 dataset provided by [2] which contains 62 features covering broad aspects of the pandemic including weekly new cases, total deaths since start of the pandemic, count of people fully vaccinated, positive rate etc. . The features have been recorded for almost all the countries but there are missing values for specific countries for specific time.
- Another dataset from [2] which consists of relative reduction/increase in movement of people to following public places day-by-day
  - retail and recreation

  - grocery and pharmacy
  - residential
  - transit stations
  - parks
  - workplaces
- Excess Mortality dataset from [2] that has been originally used for calculating p-scores for the years since 2019 has been also tested with the regression models. The regression model on this dataset serves as benchmark for our study as the p-scores are originally calculated based on the features present in the dataset. The selected features from this dataset consists of death tolls for all ages from a specific year(2015 - 2022) and additionally the average of deaths that occured between 2015 and 2019 and another feature that specifies the death toll since 2019 when the pandemic started.

C. Data Cleaning

Since the dataset is sparse and even for developed countries which have sophisticated data collection infrastructure there still can be null values present. It is a time-series based dataset that stored week-by-week data since Feb 2020 till Dec 2022, hence there can be some fields which may not have any input, in order to process such fields following steps have been carried out

1. **Filling null values**: The null values have been replaced with mean of the numerical values in a column.
2. **Normalizing values**: The values in dataset has been scaled based on min-max normalization, it is necessary as different features of the dataset can have different value ranges, hence in order to translate them to a uniform scale, normalization has been performed.

$$normalized(Col) = \frac{Col[i] - max(Col)}{max(Col) - min(Col)} \tag{13}$$

where $i \rightarrow 1...Length(Col)$

D. Outline

In order to analyse various aspects related to excess mortality rate following experiments have been done:

1. Analysing correlation between various features in the dataset and the excess mortality for a specific country, for this paper focus has been on Germany. Then selecting features having correlation greater than a threshold. Having a greater correlation signifies higher dependence of features and hence poses an interesting analysis case study for predicting p-scores.
2. Analysing prediction results for selected columns from the dataset that may be influential parameters for predicting p-scores as they are effecting population and covid-19 cases associated with it, so they may work well for predicting p-scores.
3. Checking accuracy for predicting p-scores using the changes in movement of people at public places. Movement is an important aspect for analysis as people can contract infections from public places that are too crowded or where people are not following appropriate restrictions.
4. Testing aforementioned linear regression models for checking accuracy of predicting p-scores using stringency index. [5] conducted a study of 79 countries/territories for what measures taken by governments were most effective in containing infections, hence stringency index can have a strong correlation to p-scores. [2] calculates stringency index based on following nine metrics

   - School closures
   - Workplace closures
   - Cancellation of public events
   - Restrictions on public gatherings
   - Closures of public transport
   - Stay-at-home requirements
   - Public information campaigns
   - Restrictions on internal movements
   - International travel controls.
5. Apart from aforementioned regression techniques Logistic regression and K-Nearest neighbors algorithm has also been applied for categorical classification. Here categorical classification refers to encoding p-scores into 0 and 1 based on whether they are negative or positive respectively. Since the available data for a specific country might not be enough to train Regression algorithms on time series data, hence this approach has been adopted to analyse relative change in p-score weekly.
6. A dataset based on p-scores from [2], that has been primarily used by the researchers to calculate excess mortality has been tested with all the regression techniques in order to set a benchmark for negative RMSE (Root mean square error) represented as eq. 14 for the aforementioned experimentations.

$$negativeRMSE = -\sqrt{\frac{\sum_{i=1}^{N}(y_i - t_i)^2}{N}} \tag{14}$$

   where y_i = original labels, t_i = predicted labels and N = number of samples.
7. The accuracy for various regression methods have been tested using cross-validation with $k = 5$, where k is the number of sets the original dataset have been divided.

TABLE I
Negative RMSE scores for Germany

Labels	Linear	Random Forest	Gradient Boosting	KNN	XgBoost	SVR
Custom columns	-3.7209381	-2.6314331	-2.6540210	-3.8817687	-2.7393178	-3.8300641
Correlation based columns	-6.4383633	-8.4675361	-7.4753865	-149.62925	-8.8904862	-211.83578
Public place footfalls	-0.0721714	-0.0797623	-0.0814559	-0.0772989	-0.0900402	-0.0866271
Stringency Index	-0.0721805	-0.0752048	-0.0727818	-0.0777807	-0.0770083	-0.0859858
Benchmark	-0.1845428	-0.1231846	-0.1262362	-0.1171516	-0.1436238	-0.1164801

TABLE II
Negative RMSE scores for Norway

Labels	Linear	Random Forest	Gradient Boosting	KNN	XgBoost	SVR
Custom columns	-3.1752809	-2.7346037	-2.5200055	-3.1061231	-2.6787432	-3.4190999
Correlation based columns	-1.3452337	-6.7098250	-4.9696657	-121.49028	-5.9407268	-141.45277
Public place footfalls	-0.0712364	-0.0750005	-0.0771901	-0.0776774	-0.0788875	-0.0888292
Stringency Index	-0.0713697	-0.0733457	-0.0728607	-0.0774373	-0.0733458	-0.0987043
Benchmark	-0.2838882	-0.1826682	-0.1772600	-0.1606830	-0.2031037	-0.1446229

TABLE III
Negative RMSE scores for United States

Labels	Linear	Random Forest	Gradient Boosting	KNN	XgBoost	SVR
Custom columns	-3.9423090	-3.3402384	-3.4045701	-3.6752702	-3.3764367	-4.1374055
Correlation based columns	-12.190794	-3.3674934	-3.3333009	-5.5145587	-3.0693027	-14.855634
Public place footfalls	-0.0840952	-0.0869504	-0.0881588	-0.0878562	-0.0928645	-0.0927208
Stringency Index	-0.0873868	-0.0971000	-0.0921098	-0.0933310	-0.1045065	-0.1038840
Benchmark	-0.0859579	-0.0892501	-0.0886089	-0.0898545	-0.1036534	-0.0977049

TABLE IV
Negative RMSE scores for Denmark

Labels	Linear	Random Forest	Gradient Boosting	KNN	XgBoost	SVR
Custom columns	-2.9568002	-3.1987805	-3.2350285	-3.2344993	-3.4784595	-3.0030515
Correlation based columns	-1.0590322	-7.9468074	-6.7897446	-113.10409	-6.0121810	-104.29708
Public place footfalls	-0.0759651	-0.0849735	-0.0845248	-0.0837376	-0.0934693	-0.0878726
Stringency Index	-0.0598937	-0.0637728	-0.0632187	-0.0654565	-0.0660613	-0.0792702
Benchmark	-0.2222599	-0.1141613	-0.1132692	-0.1105022	-0.1357136	-0.1065477

TABLE V
Negative RMSE scores for France

Labels	Linear	Random Forest	Gradient Boosting	KNN	XgBoost	SVR
Custom columns	-13.339408	-3.8296905	-5.0724072	-10.212418	-4.1612004	-19.679310
Correlation based columns	-12.972170	-4.0960176	-4.8243750	-9.0856251	-4.1226950	-19.305988
Public place footfalls	-0.0566983	-0.0636437	-0.0625117	-0.0636841	-0.0684979	-0.0743738
Stringency Index	-0.0524178	-0.0543273	-0.0534826	-0.0565980	-0.0555010	-0.0661581
Benchmark	-0.0794880	-0.0785659	-0.0725337	-0.0818447	-0.0792876	-0.0842854

Fig. 1: Accuracy of various models over threshold

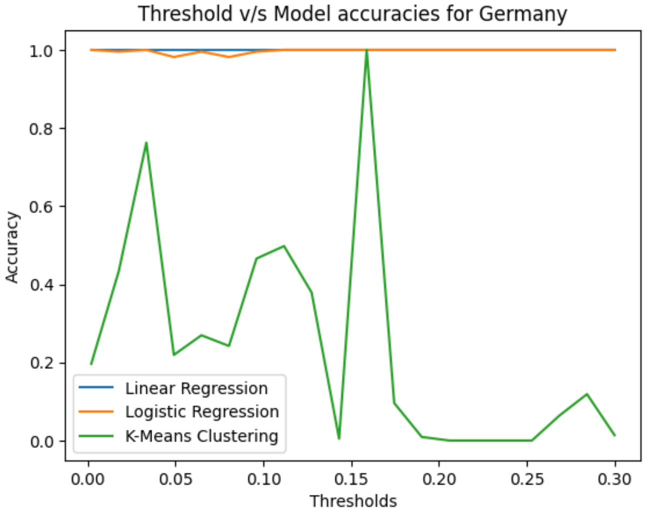


Fig. 2: Model accuracy for multiple countries with selected columns based on threshold

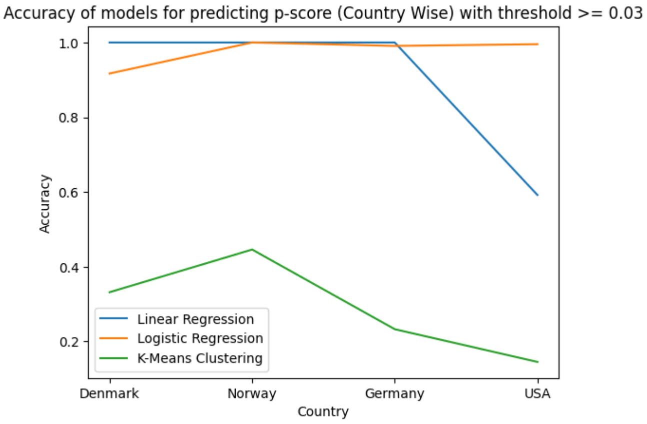


Fig. 3: Regression model accuracies for Germany based on deviation of negative RMSE from benchmark

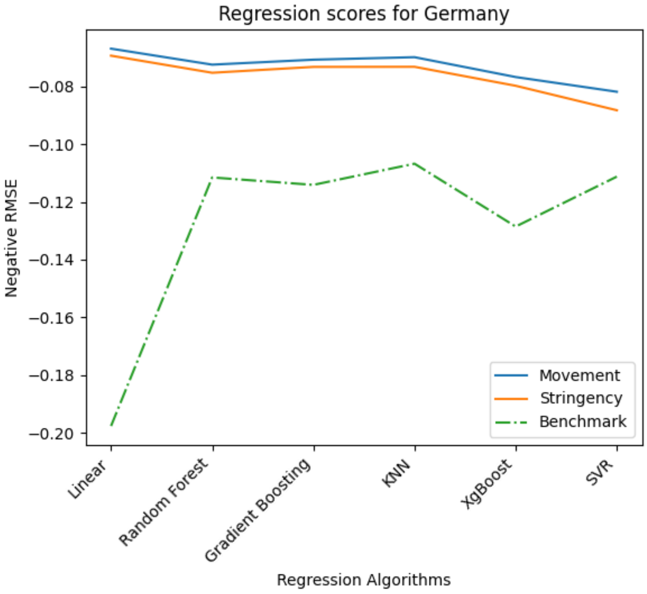


Fig. 4: Regression model accuracies for USA based on deviation of negative RMSE from benchmark

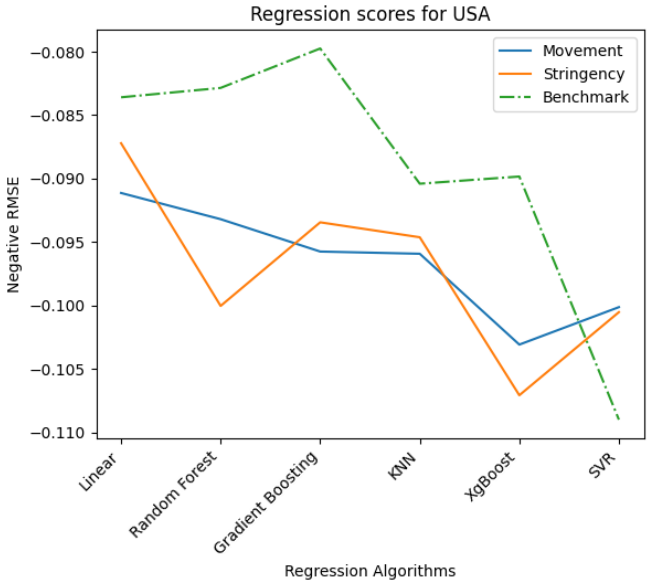


Fig. 5: Regression model accuracies for Denmark based on deviation of negative RMSE from benchmark

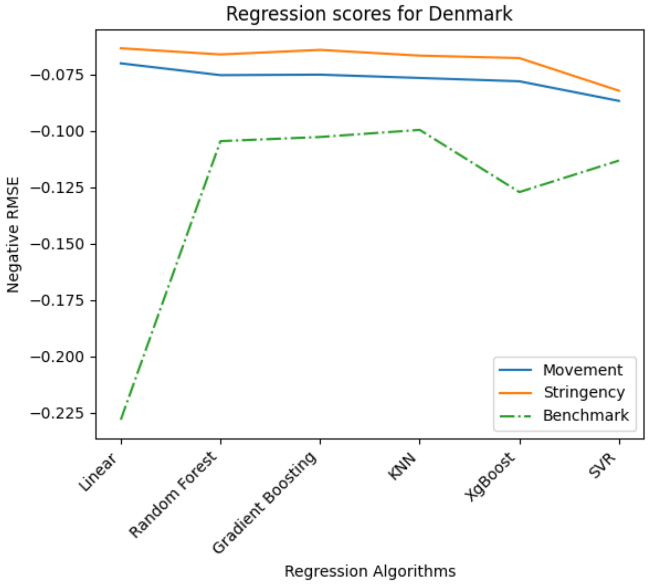


Fig. 6: Regression model accuracies for France based on deviation of negative RMSE from benchmark

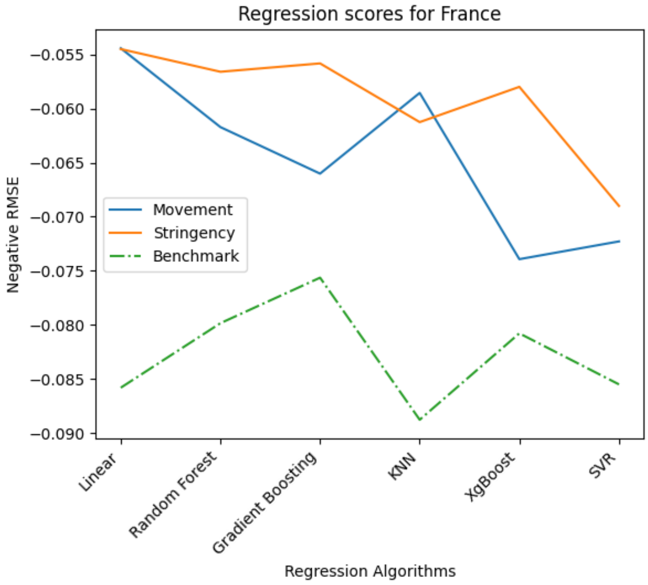
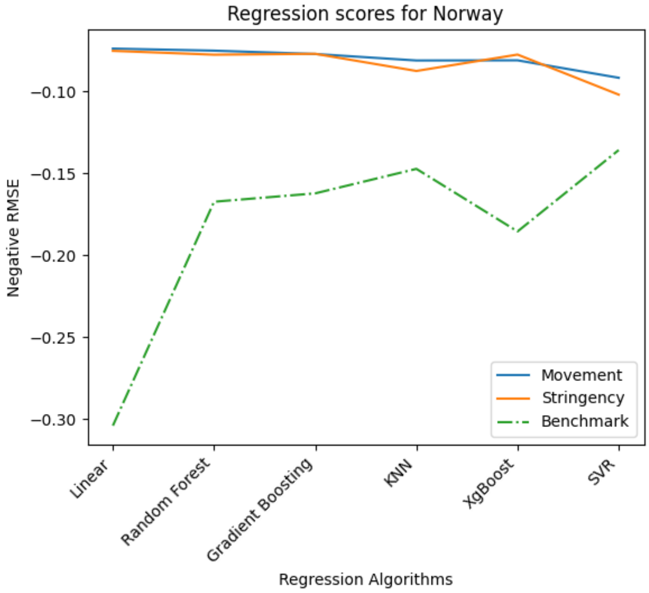


Fig. 7: Regression model accuracies for Norway based on deviation of negative RMSE from benchmark



III. Observations

- From fig. 1 it is evident that Linear regression for time series prediction and logistic regression for binary classification have resulted in similar accuracies over various thresholds whereas for K-Nearest neighbors(for 2 clusters/ binary classification), the accuracy has been highest for correlation value = 0.15 whereas it has been lowest for correlation value = 0.25.
- From fig. 2 it can be concluded that Linear and Logistic regression have high accuracies for Denmark, Norway and Germany whereas for USA Linear Regression's accuracy is quite low. K-Nearest neighbors proved to be ineffective for all the countries.
- For Germany the negative-RMSE values (for all regression techniques) for Public place footfall and stringency index is more similar to benchmark than custom columns and correlation based columns. Similar is the case for Norway, United States, Denmark and France. Whereas in correlation based columns the regression results deviates a lot from benchmarks. Custom columns although proves to provide better results than correlation based columns but still are significantly deviant from benchmarks.
- From Fig. 3 it can be observed that SVR and KNN provides better accuracy as compared to any other regression algorithm whereas Linear regression performs the worse whereas from fig. 4 it can be concluded that for USA SVR, KNN and Linear regression works well whereas Random forest and Gradient boosting are worse.
- From Fig. 5 it can be deduced that in case of Denmark SVR, KNN and Random Forest works well whereas Linear Regression performs worse. For France (Fig. 6 SVR is well suited whereas like others linear regression delivers worse results. For Norway as can be seen from fig. 7 KNN and SVR has lower deviation from benchmark whereas linear regression delivers the worst performance.

IV. Results

It can be seen from observations that binary classification using K-Nearest neighbors is ineffective in predicting the trend of excess mortality rate. Using Logistic regression for binary classification has high accuracy in some cases but binary cases doesn't completely informs about the p-scores hence other regression methods needs to be employed to predict more discrete values. From various regression results it can be observed that for many countries stringency index and change in people's movements have been quite influential in predicting p-scores whereas values like highly correlated features including new covid cases proved to be ineffective, hence it can be concluded that policies implemented by governments and awareness among people regarding the hazards of the pandemic proved to be effective. From regression algorithm results it can be clearly seen that SVR and KNN regressions performed better than any other regression techniques whereas Linear Regression proved to be the worst performer with USA being an exception.

REFERENCES

[1] Y. Shi, G. Wang, X.-p. Cai, J.-w. Deng, L. Zheng, H.-h. Zhu, M. Zheng, B. Yang, and Z. Chen, "An overview of covid-19," *Journal of Zhejiang University. Science. B*, vol. 21, no. 5, p. 343, 2020.

[2] E. Mathieu, H. Ritchie, L. Rodés-Guirao, C. Appel, C. Giattino, J. Hasell, B. Macdonald, S. Dattani, D. Beltekian, E. Ortiz-Ospina, and M. Roser, "Coronavirus pandemic (covid-19)," *Our World in Data*, 2020. https://ourworldindata.org/coronavirus.

[3] A. Karlinsky and D. Kobak, "Tracking excess mortality across countries during the covid-19 pandemic with the world mortality dataset," *Elife*, vol. 10, p. e69336, 2021.

[4] Chaya, "Random forest regression," 2020.

[5] D. Lewis, "What scientists have learnt from covid lockdowns.," *Nature*, vol. 609, no. 7926, pp. 236–239, 2022.