PROJECT REPORT

Data Science

DEPARTMENT OF COMPUTER SCIENCE

Submitted to:

Dr. Muhammad Atif Tahir

Submitted by:

Syed Owais Tahir Ali (17K-3888)

Syed Muhammad Owais (17K-3858)

Sec: BCS-9A

National University of Computer and Emerging Sciences-FAST Karachi Campus

Table of Contents:

COVID-19 Data Analysis and Predictions

Abstract:

"History repeats but science reverberates", Siddhartha Mukherjee. If history teaches us anything, it's that while pandemics may start small, their impacts can be as disastrous as wars or natural disasters. The difference today is that science gives us the ability to detect pandemics right at the very beginning and to take actions to mitigate their impacts before they spread too widely. Over the past centuries, Humans have endured pandemics like Cholera and Spanish flu. Yet another contagious disease has taken over the world, The Chinese Coronavirus. Since December 2019, millions of patients have been diagnosed with the virus. The future is yet unknown. Predictions for the next couple of days are made by data scientists throughout the world so that the requirements for the necessary resources to deal with the outbreak are fulfilled in time. In this paper, we demonstrate the analysis of the trends in the COVID-19 datasets (global and local, both). We have applied techniques to train the supervised machine learning models effectively and predict the number of cases that are expected shortly. We found that countries whose governments and communities have chosen to implement precautions strictly have halted the spread among their people.

Introduction:

We live in an interconnected, an increasingly globalized world. A big thanks to international jet travel, people and the diseases they carry can be in any city around the planet in a matter of hours. This respiratory virus isn't an outlier, it is the part of our interconnected viral village. It is a highly contagious virus and one sneeze is all it takes to spread throughout the community. By observing the rise in its spread, the World Health Organization (WHO) has declared it a pandemic, meaning that it's spreading worldwide.

Epidemics and Pandemics come in many shapes and forms. For instance, in 2010, a devastating earthquake hit Haiti, forcing thousands of people into refugee camps. Within weeks, the campers were infected with cholera, a bacteria spread by contaminated water, triggering a country-wide epidemic. Pandemics have occurred throughout the human history. However, by far the greatest pandemic killer is influenza. The first recorded pandemic occurred in 1580. The 18th and 19th centuries saw at least 6 pandemics. In terms of mortality, none can compare with the Great Flu pandemic of 1918. A death toll of 50 million people was observed worldwide.

A deadly contagious virus has hit the human race yet again. Coronavirus caused an epidemic of severe acute respiratory syndrome in China. In December 2019, the first case of this severely acute respiratory syndrome was observed in China and it then proliferated faster than public health measures could contain it. It became an international epidemic. Millions of people got affected. Out of the million patients, some couldn't survive, some have recovered and some are still fighting for their lives.

Background:

Coronavirus disease (COVID-19) is an illness caused by a novel coronavirus now called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which was first identified as a respiratory illness cases in Wuhan City, China. It was initially reported to the WHO on the 31st of December, 2019. On January 30, 2020, the WHO declared the COVID-19 outbreak as a global health emergency and then later declared it as a global pandemic, its first such designation since declaring H1N1 influenza a pandemic in 2009.

The illness caused by SARS-CoV-2 was termed COVID-19(the acronym derived from "coronavirus disease 2019"), by the WHO. The name COVID-19 was chosen to avoid stigmatizing the virus's origins in terms of populations, geography, or animal associations

As stated below, machine learning can help accelerate the drug development process, provide insight into which current antivirals might provide benefits, forecast infection rates, and help screen patients faster. Additionally, although not researched on currently, there are several other appropriate application areas. That being said, there are many barriers related to lack of limited training data, the ability to integrate complex structures into DL models, and, perhaps most importantly, access to the available data. I'm not going to detail the techniques below (not that I could as my chemistry/drug-development knowledge is severely lacking), but instead I will aim to summarize the different resources. There actually haven't been too many models (at least publicly documented ones) that have explicitly attempted to model the coronavirus spread. However significant amounts of prior research have studied forecasting the seasonal flu and other outbreaks. Interestingly, a large number of the methods currently being used to forecast disease spread and patient mortality are based on shallow methods.

Problem Statement:

Provided with the abundant data of Covid'19 effected countries, we are determined to estimate the outbreak of this contagious virus for the next 10 days based on the reported data. Keeping in view that the official confirmed cases number has surpassed all the previous prediction (Anastassopoulou C 315), we would be applying two prediction model i.e. Linear Regression and SVM model to compare the results and to provide the mortality rate.

Data Source:

Four datasets are used in the analyses and visualization of Covid'19 pandemic from the world's largest data science source, Kaggle to demonstrate the outbreak worldwide. The Three global dataset possess the comprehensive information of country/region, province/state, longitude and latitude with their respective numbers of cases, deaths and recoveries from all regions of the world ranging from 1/22/20 to 4/22/20. The covid19 confirmed global, covid19 deaths global, covid19 recoveries global datasets are organized with rows representing the entries from different countries and the columns depict the date in which cases are registered accordingly. Each of the global dataset for a country are a time series. Let I be the Country/City and j the day. A matrix of I by j dimensions would show the total cases against each Country/City on a particular date.

The local dataset demonstrates the date wise entries of confirmed cases, deaths and recoveries from different regions within Pakistan. It also shows the travel history, province and city name to illustrate the reason behind the high peaks of confirmed cases which is mainly due to local social contact and pilgrims from Tehran.

Active cases from the local dataset are calculate by using the formula:

$$\text{Total active} = \text{Cases} - (\text{Deaths} + \text{Recovered})$$

For each of the city/province for active case visualization within Pakistan. To predict the outbreak for future, following models are used.

Support Vector Machine:

An algorithm for two-group classification problems which is categorized under the supervised machine learning approach. It is known as a Support Vector Machine. The model is used for categorizing and predicting new data after training the SVM model with sets of labeled data.

Features:

- This model works well with limited amount of data (thousands).
- Takes in data points and gives out a hyper-plane that separates the classes i.e. it draws a decision boundary between two sets of data.
- It treats linear and non-linear data separately.
- A simple linearly separable data works best in a 2D plane.
- For non-linear data, it uses the kernel trick to differentiate among the data tags.

The kernel trick:

SVM model predictions include the conversion of data into multi-dimensions for fair classification. Every step of transforming the data from lower dimension to a higher one involves multiple steps of complicated calculation. It becomes expensive when every vector of the dataset has to be transformed.

The kernel trick provides a cheaper solution and implements the dot product technique. The kernel function takes input the data and transform in the required form. Kernel functions are of different types:

- Linear
- Non-linear
- Polynomial
- Radial-Bias function
- Sigmoid

Linear Regression Model:

It involves the fitting of linear equations to data and also models the relationship between two variables. It is not necessary for the variables to be dependent on each other but some significant association has to be there.

Linear regression line equation:

$Y = a + bX$

Where, Y = explanatory variable, X = dependent variable. The slope of line is b and a is the intercept.

Features:

- It is commonly used for predictive analysis.
- The predicted output values are continuous and form a slope.

It offers extrapolation technique when the prediction has to be made outside the range of actual data.

Material and Methodology:

For the analysis of the world-wide data, we have coded in Python 3.0 by using Anaconda Navigator 3 in Jupyter Notebook to process the data. The predominant libraries are matplotlib that is responsible for creating a figure with respect to data provided in its function and other legends and labels to make it more readable; seaborn that possesses a high level interface for visualization of Covid’19 outbreak and its statistical information; sklearn library which features multiple predictions models including SVM model by using SVR that is used to predict future peaks in our analyses. mpl_toolkits.mplot3d library is used for 3d animation of the model. Datetime and time libraries are added to rearrange the date format.

In our analysis of the local dataset, the data is grouped with respect to date, city and province to illustrate the number of confirmed cases, recoveries and deaths. Univariate, Bivariate and Multivariate analysis are done to find the correlation among features in local dataset using heatmaps. Additionally, a 3d model is built to demonstrate the number of cases, death and recoveries within Pakistan. A useful of FacetGrid is implemented to exhibit the effect of travel history in the Covid'19 cases, carried out the function of sns.pairplot to get the histogram and scatterplot of individual integer type feature in the set.

In the global dataset, illustrated the top 10 effected countries from this Covid'19 by country-wise summation, grouped by dates using plt.barplots. Also, a pie chart is built to exemplify the percentage of patients in each country.

To predict the outbreak for the next 10 days, we have used SVM and Linear Regression model with test_split ratio of 0.15 and shuffle being False as the trend grows exponentially. Below are the best suitable values for SVM features.

kernel = ['poly','sigmoid','rbf']

c = [0.01, 0.1, 1, 10]

gamma = [0.01, 0.1, 1]

epsilon = [0.01, 0.1, 1]

shrinking= [True, False]

SVM runs the model and chooses the best parameters out of those provided according to the data it gets. Larger C tends to give an optimization result. Both values for shrinking is provided to reduce the kernel dimensions depending upon the model. Svm_estimator plays a major role in this infectious disease forecast as it returns the highest score of accuracy. We trained the model using the original dates reported in the dataset and passed the next 10 days to the model to predict the number of cases for future. Similarly, we fit the same x_train and y_train_confirmed n the linear regression model with the same splitting ratio to get the linear prediction of future for the next 10 days to have a comparative analysis.

Support Vector Machine:

An algorithm for two-group classification problems which is categorized under the supervised machine learning approach. It is known as a Support Vector Machine. The model is used for categorizing and predicting new data after training the SVM model with sets of labeled data.

Features:

- This model works well with limited amount of data (thousands).

- Takes in data points and gives out a hyper-plane that separates the classes i.e. it draws a decision boundary between two sets of data.
- It treats linear and non-linear data separately.
- A simple linearly separable data works best in a 2D plane.
- For non-linear data, it uses the kernel trick to differentiate among the data tags.

The kernel trick:

SVM model predictions include the conversion of data into multi-dimensions for fair classification. Every step of transforming the data from lower dimension to a higher one involves multiple steps of complicated calculation. It becomes expensive when every vector of the dataset has to be transformed.

The kernel trick provides a cheaper solution and implements the dot product technique. The kernel function takes input the data and transform in the required form. Kernel functions are of different types:

- Linear
- Non-linear
- Polynomial
- Radial-Bias function
- Sigmoid

Linear Regression Model:

It involves the fitting of linear equations to data and also models the relationship between two variables. It is not necessary for the variables to be dependent on each other but some significant association has to be there.

Linear regression line equation:

$Y = a + bX$

Where, Y = explanatory variable, X = dependent variable. The slope of line is b and a is the intercept.

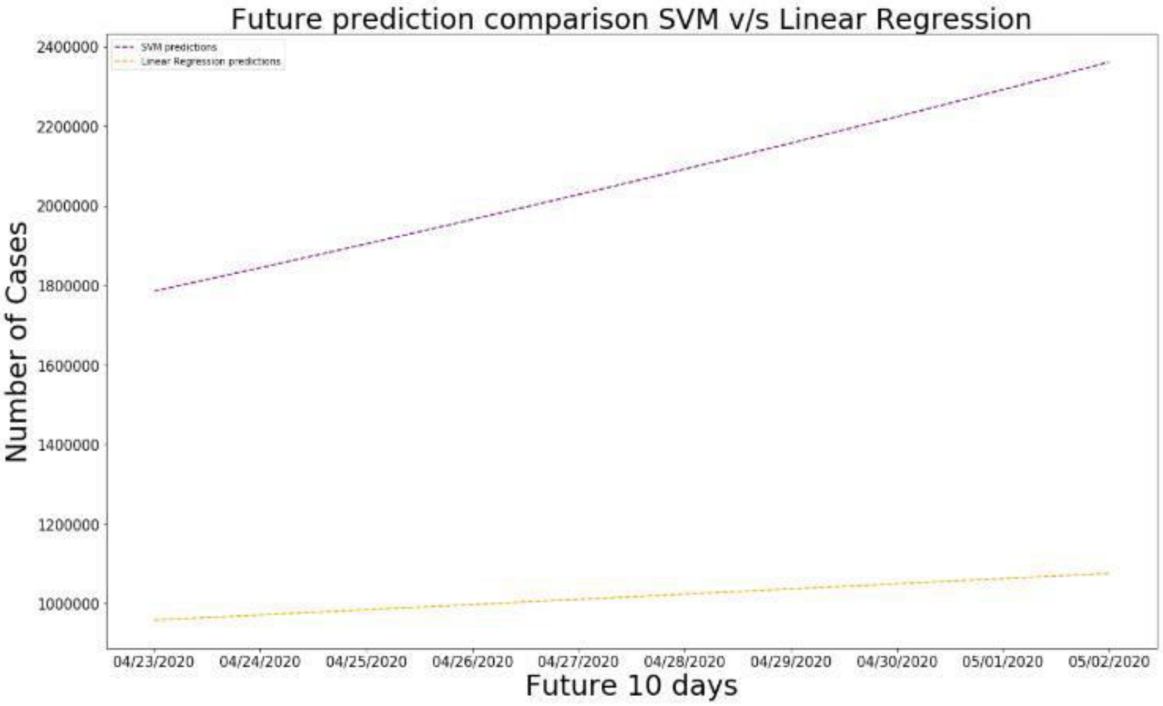
Features:

- It is commonly used for predictive analysis.
- The predicted output values are continuous and form a slope.

It offers extrapolation technique when the prediction has to be made outside the range of actual data.

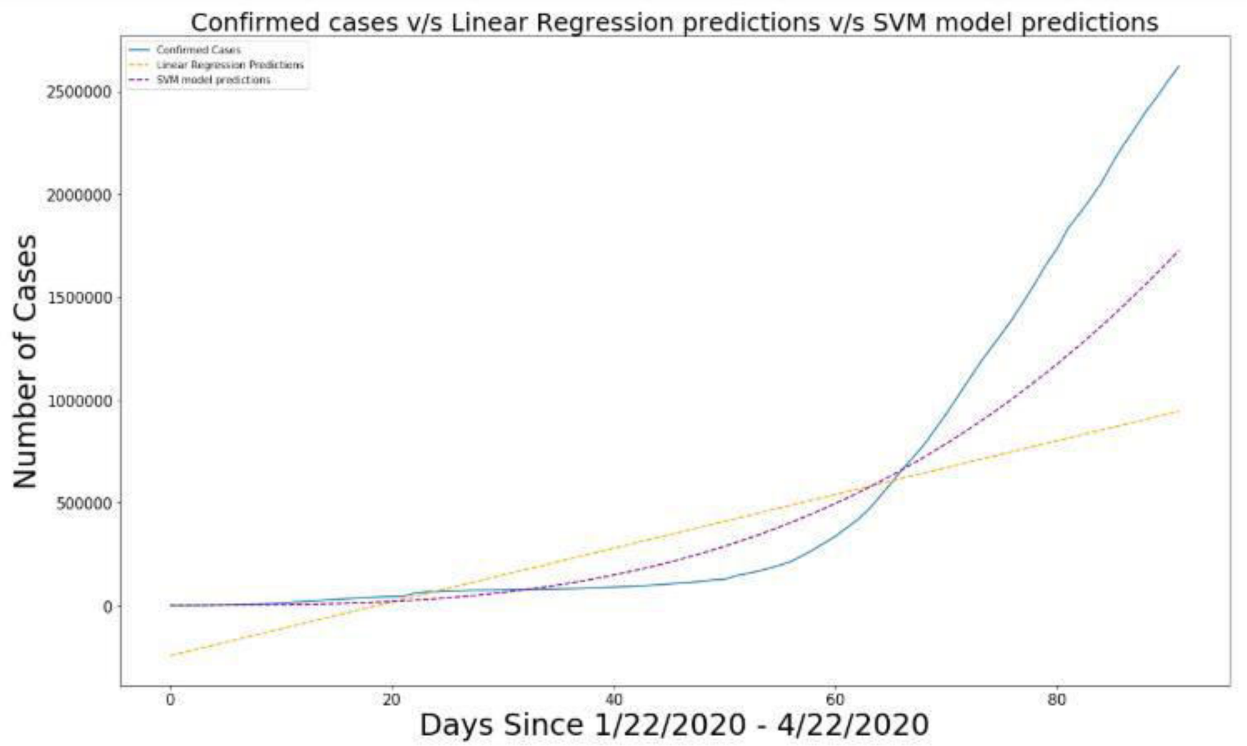
Data and Results:

The fetched datasets had data from all over the globe from January 22nd, 2020 to April 22nd, 2020. A future prediction was made using two supervised machine learning models. Both the models predicted the total number of cases in the next 10 days i.e. from April 23rd, 2020 to May 2nd, 2020.



The above figure depicts the comparison between the two models predictions. A huge difference can be seen among the two. SVM model predictions are way much larger than linear regression model.

Predictions of data from January 22nd, 2020 to April 22nd, 2020 were also made and comparison among the three i.e. Actual data, Linear regression predictions and SVM model predictions was also visualized to see the differences.



The above figure depicts the comparison between the Confirmed cases, linear regression model predictions and SVM predictions. The difference among the three can clearly be seen.

Conclusion:

Covid-19 is a pandemic that has affected the whole world in different ways. Countries are facing serious economical and health crisis. People are dying due to unemployment and malnutrition. Hospitals are out of resources and the contagious virus is not slowing down the spread. But the good news is that every nation is trying to stop the spread by social distancing and obeying the laws imposed by the government. By the Grace of God, many countries have taken control of the situation but some are still struggling.

The information and technology sector is playing its part in making new machines and resources for the help of people. Data scientists are busy in analyzing the trends and forecasting the future. In this paper, a study of global and local data sets has been displayed. Also, a prediction of the future 10 days has also been made using supervised machine learning models.

- Summary of the findings:
  The data of confirmed cases has been trained on two different models i.e. Support Vector machine model and linear regression model. It has been observed that the prediction made by two models vary too much from the actual data. A huge rise was observed in the SVM predictions and the linear regression predictions were much lower than the actual data.

- Limitations of the project:
  The prediction was made only of the future 10 days due to less amount of training data. The results of the linear regression model were flawed and not much could be deduced out of them.

References:

1. Anastassopoulou C, Russo L, Tsakris A, Siettos C. "Data-based analysis, modelling forecasting of the COVID-19 outbreak." *PLoS ONE* (2020): 314-356. Print.

2. He, Benlan, et al. "Prediction of customer attrition of commercial banks based on SVM model." *Procedia Computer Science* 31 (2014): 423-430.

3. Ivanov, Dmitry. "Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case." *Transportation Research Part E: Logistics and Transportation Review* 136 (2020): 101922.
4. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
5. www.kaggle.com
6. https://pandas.pydata.org/
7. https://scikit-learn.org/stable/modules/svm.html
8. https://realpython.com/linear-regression-in-python/