Data Science Term Project

**COVID-19 Analysis and future forecasting**

**SUBMITTED TO**

Sir Atif Tahir

Farjad Suhail

**Introduction**

Coronavirus virus also known as COVID-19, is a contagious disease caused by respiratory syndrome coronavirus. The first known case was identified in Wuhan, China, in December, 2019. The disease has since spread worldwide, leading to an ongoing pandemic. The symptoms of coronavirus are variable, but often include fever, cough, headache, fatigue, breathing difficulty and loss of smell and taste. Symptoms may begin one to fourteen days after exposure to the virus. At least a third of people who are infected do not develop noticeable symptoms. The older people are generally at higher risk and trends show that casualties have been reported in a very high percentage of people above the age of 60 who have been exposed to the virus. The World Economy has massively suffered because of the ongoing pandemic because month long lockdowns have been imposed in various parts of the world. 20 Million People have become jobless and 3 Million have gone below the poverty line.

**Setting the Research Goal**

Our motive for our Data Science Project is to analyze the impacts and trends of Covid-19 in 5 provinces and capital of Pakistan. Moreover, through data modeling we have also predicted the number of positive cases expected in Pakistan in the upcoming 10 days.

Finally, we have also predicted the positivity ratio in different parts of the country.

**Retrieving Data**

The data was collected after extensive research over the internet. For data analysis and prediction, usually retrieving data is the most difficult task as it is quite difficult to get data similar to what one wants to work upon. We were able to find a data set quite close to what we planned for our project. We got our data from the website <https://opendata.com.pk/>. The data was quite relevant and contained all the useful and had all the prerequisites for performing analyses and predicting future trends of Covid-19 in different parts of Pakistan.

**Data Preparation**

This is where the actual implementation of our project started. Initially, the dataset was loaded. The datasets had many columns that were not of our use so we dropped them. Moreover, some of the columns had empty values so replaced them with 0 so that computation would become possible on that column. The data we have starts from 22nd May, 2020. We have almost a month long data. We cleaned and prepared it for further computation and to make it usable. The dataset had data about the number of active cases, number of deaths, number of tests conducted, number of people who contacted covid-19, became positive in the last 24 hours, number of people that recovered etc.

**Data Exploration**

After preparing our data we then started data exploration. Thus, we worked upon to find the trends of the total number of positive cases in all the provinces of Pakistan and Islamabad. We plotted different graphs to see how covid-19 cases were increasing all over the country. Next up, we plotted a line graph to see the trends of the tests that have come up positive and the number of deaths. The number of deaths and positive tests were drawn on the same graph to visualize the impacts and accepts of covid and see the rapid rate at which positive cases and deaths both were increasing. These explorations gave us a useful insight of the rate at which corona virus is spreading and the impact it is having on us. Around 6% positive cases were observed in the two big provinces of Pakistan. That is how worst the situation was that we observed and the numbers were only increasing.

**Data Modeling**

Our next step in the process was to perform data modeling for our project. Our goal was to predict the positive cases in the upcoming 10 days. So, for this we started off with training and testing data.

Our training data was 70% while the testing data was set to 30%. We have used 2 models to predict the number of positive cases in the next 10 days and the percentage of positive cases in our provinces.

We have used DecisionTree and Naive Bayes classifiers to predict the positive cases. Initially, we imported all the relevant libraries. Next up we trained our model by passing the daysfuture (variable) and punjabcases (variable) list. Then we fitted our model on the X\_train and Y\_train variables we obtained from training the model. Finally, using the predfuture variable we predicted the number of cases to be expected in the upcoming 10 days.

**Presentation and Automation**

The data models of DecisionTree and Naive Bayes models were used and presented in order to predict future trends. We prepared data so that computation could take place on it. After that we did exploration on data. Next up we trained our model using Decision Tree and Naïve Bayes model. We obtained results for our model the number of positive cases to be expected in different provinces of Pakistan and the ratio of positive cases.

**Conclusion**

Our objective for this project was to analyze the various trends related to the coronavirus in different provinces of Pakistan. We successfully demonstrated them and also predicted the future trend of positive cases. By using the Decision Tree and Naïve Bayes model we predicted that positivity rate in Punjab would be just above 6% and in KPK would be around 4.5% in the upcoming 10 days.

**References**

<https://opendata.com.pk/>

<https://towardsdatascience.com/>

<https://towardsdatascience.com/>