Covid-19 Data Analysis

(Tools and Techniques for Data Science Term Project Report)

Ans Munir
Information Technology University
Lahore
msds20033@itu.edu.pk

1. Introduction

Various approaches have been employed by governments worldwide to address the Covid-19 pandemic. Since it is a novel illness, we do not yet have the tried-and-true methods for managing the current epidemic. We must ascertain the impact of COVID-19 on individuals across various age cohorts. Additionally, there is a theory that suggests individuals with pre-existing medical conditions are more susceptible to severe symptoms than people without a medical background. Thus, it will be advantageous if we can establish a relationship between age, pre-existing conditions, and death. This will help the health authorities save more lives by directing their attention towards those who are most vulnerable to developing severe Covid-19 disease.

In a similar vein, large cities offer better access to healthcare due to their abundance of doctors and other medical services. However, compared to cities, there is a larger risk of death in rural and isolated locations due to inadequate health facilities. As a result, a thorough report on the proportion of deaths among all patients is required. This is particularly significant because true insights cannot be gained from simply looking at patient deaths alone. There are numerous additional variables as well, such as age, pre-existing illnesses, and many other factors.

Our goal is to conduct a thorough examination of the symptoms experienced by Covid-19 patients in order to establish any connections between patient demographics, age, and pre-existing medical conditions, and patient mortality. In order to reduce the number of deaths, this will assist the health authorities in providing medical support to individuals who are more vulnerable.

2. Dataset

The Primary and Secondary Health Care department has gathered data from Punjab's public and private hospitals and clinics. In order to do analysis, the Punjab Information Technology Board (PITB) then gathered this dataset.

There are 35498 Records in the dataset with 60 Attributes.

2.1. Data Pre-processing Steps

We have taken following steps to preprocess the data:

1. Remove duplicate or irrelevant observations.
2. Remove unwanted observations from dataset.
3. Including duplicate observations or irrelevant observations.
4. Fix structural errors.
5. Filter unwanted outliers.
6. Handle missing data.
7. Handle Inconsistent data.

3. Data Analysis

3.1. Gender wise distribution

We have analyzed the data about the gender of patients. Figure 1 shows the distribution of data based on gender. It can seen from the Figure 1 that males are more effected to the Covid-19 as compared to females. One of the reason could be that as females mostly stay at the home therefore, they are less effected with Covid-19.

4. Top age group distribution

We have analyzed the data about the top age groups mostly effected with Covid-19. Figure 2 shows the top age groups effected with Covid-19. It has been shown that people of age 50 are mostly effected.

5. Age distribution with gender

Figure 3 shows the age distribution with respect to age. Both the genders have normal distribution but there is difference between density.

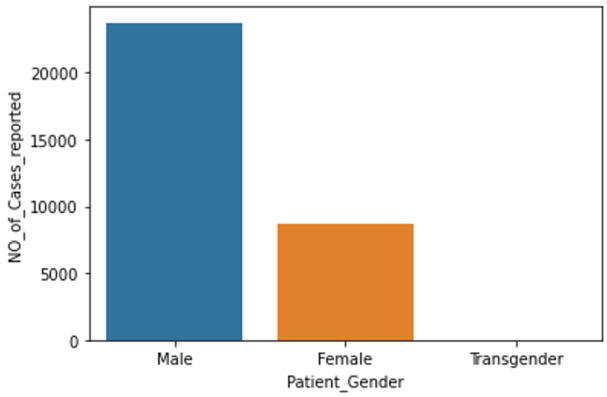

Figure 1. Analysis based on gender distribution.

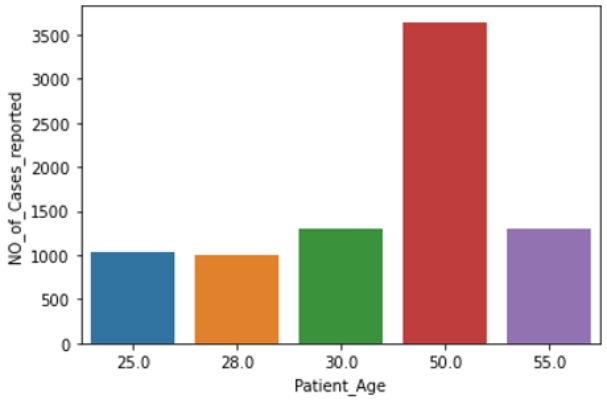

Figure 2. Top age groups effected with Covid-19.

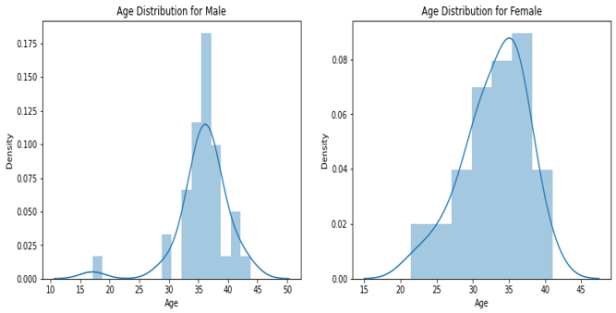

Figure 3. Gender Based Age distribution.

6. Top affected districts

Figure 4 shows the top 10 districts with covid-19 patients. Lahore has the highest number of patients.

7. Symptoms based analysis

Figure 5 shows the top districts with fever. And Figure 6 shows the top districts with the patients who have fever, shortness of breath and cough. Figure 4 shows Lahore as the top district with covid-19 patients. However, Figure 5 and 6 has the top patients with symptoms but Lahore is not the top district here.This implies that even though Lahore has a larger patient population, its residents likely have milder symptoms because of the city's superior medical facilities.

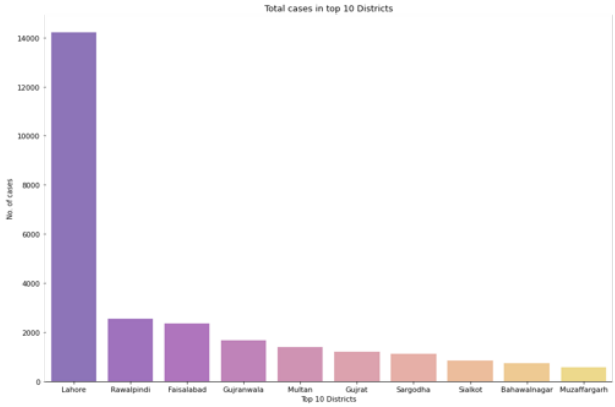

Figure 4. Top districts effected with Covid-19.

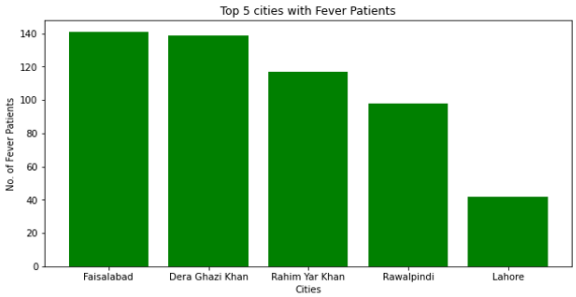

Figure 5. Top 5 districts with fever patients.

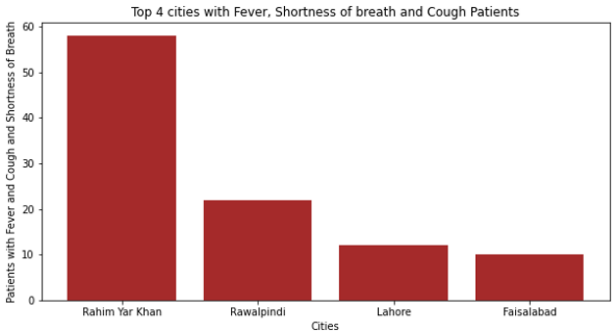

Figure 6. Top 4 districts with fever, shortness of breath and cough patients.

7.1.

Predictive Modeling

We have trained different models to predict the chances of death give the symptoms and bio data of patients like age. We have used different methods for prediction.

- Logistic Regression.
- K Neighbor Classifier.

Methods	Accuracy
Logistic Regression	94%
K Neighbor Classifier	97%
Decision Tree	94.9%
SVM	95%
Ensemble (Bagging)	95%
Naive Bayes	83.7%

Table 1. Table with accuracy of different machine learning methods on covid-19 data.

- Decision Tree.
- SVM.
- Ensemble (Bagging).
- Naive Bayes.

8. Conclusion

We have analysed Covid-19 data and attempted to establish a relationship between patient data and disease severity. This dataset indicates that males were more likely than females to have COVID-19. Additionally, we discovered that older individuals were more affected by COVID-19 than younger individuals. Large cities can have a higher patient population with milder symptoms. This demonstrates how patients' COVID-19 severity was decreased by larger cities' superior healthcare facilities.