Project Report on Covid-19 Data Analysis & Prediction using Machine Learning.

Submitted By :- Raushan Kumar Chaurasiya

Project Description

1. Introduction

In today's world, where the global community continues to combat the COVID-19 pandemic, the development of accurate and efficient tools for disease prediction has become increasingly crucial. In line with this need, our proposal aims to build a machine learning model to predict whether an individual is COVID-19 positive or not based on a dataset of symptoms. By leveraging the power of machine learning algorithms, we aim to enhance the accuracy and speed of COVID-19 diagnosis, thereby contributing to effective medical treatment and reducing the burden on healthcare systems.

Importance in Today's World:

The COVID-19 pandemic has overwhelmed healthcare systems worldwide, leading to a pressing need for efficient and reliable diagnostic methods. Traditional diagnostic approaches, such as PCR testing, often require time-consuming laboratory procedures and are limited by testing capacity. By developing a machine learning model for COVID-19 prediction, we offer a promising solution that can provide quick and

accurate assessments based on symptoms alone, without the need for extensive laboratory testing. This can greatly improve the early identification of COVID-19 cases, enabling prompt medical intervention and appropriate allocation of healthcare resources.

Improving Medical Treatment:

Accurate disease prediction plays a crucial role in improving medical treatment outcomes. By accurately identifying individuals who are likely to be COVID-19 positive, our machine learning model can enable early intervention, leading to timely medical care and the implementation of necessary preventive measures. This early detection can help prevent disease progression, reduce the severity of symptoms, and improve overall patient outcomes.

Impact on the Medical Field:

The implementation of an effective screening tool based on our machine learning model can have a profound impact on the medical field. It can significantly reduce the burden on healthcare systems by streamlining the diagnostic process, particularly in areas with limited testing capacities. By efficiently identifying COVID-19 positive individuals, healthcare resources can be optimized, ensuring that critical cases receive immediate attention while preventing the unnecessary utilization of resources for low-risk cases. This targeted allocation of resources can help in managing the influx of patients, improving overall healthcare delivery, and potentially saving lives.

Future Applications and Knowledge Gap:

Beyond its immediate application in COVID-19 prediction, our proposed machine learning model can serve as a valuable framework for predicting other diseases based on symptoms. The knowledge gained from developing this model, such as feature selection techniques and algorithmic approaches, can be applied to future disease prediction efforts. This has the potential to fill knowledge gaps in various healthcare domains, allowing for faster and more accurate diagnosis of diseases beyond COVID-19. The ability to predict diseases accurately has far-reaching implications for personalized medicine, public health planning, and the development of targeted treatment strategies, ultimately leading to improved healthcare outcomes for individuals and communities alike.

2. Initial Hypothesis

Based on the dataset provided, which includes variables such as Test_date, Cough_symptoms, Fever, Sore_throat, Shortness_of_breath, Headache, Corona, Age_60_above, Sex, and Known_contact, we can form the following initial hypotheses:

Hypothesis 1:

Patients who were in direct contact with Confirmed (Covid Positive) Patients are more likely to be Corona positive.

This hypothesis suggests that individuals who have had close contact with confirmed COVID-19 positive patients are at a higher risk of being infected themselves. The assumption is that the virus spreads primarily through close proximity and direct contact with infected individuals. By examining the variable "Known_contact" in the dataset, we can investigate whether there is a correlation between known contact with COVID-19 positive patients and the likelihood of testing positive for the virus.

Hypothesis 2:

Shortness_of_breath, Fever, and Cough_symptoms are important factors in determining COVID-19 positive or negative cases.

This hypothesis proposes that symptoms such as shortness of breath, fever, and cough are key indicators in identifying COVID-19 positive cases. These symptoms are commonly associated with respiratory infections and have been identified as prominent symptoms of COVID-19. By analyzing the variables "Shortness_of_breath," "Fever," and "Cough_symptoms" in the dataset, we can explore the relationship between these symptoms and the likelihood of a person being COVID-19 positive or negative.

It is important to note that these are initial hypotheses based on the provided dataset, and further analysis and modeling will be required to validate these assumptions and uncover additional insights related to COVID-19 prediction.

TO VIEW THE REPORT ON EXPLORATORY DATA ANALYSIS , BUILDING AND DEPLOYMENT OF MACHINE LEARNING MODEL ON THE DATASET.(PLEASE REFER TO capstone_covid.ipynb file)

SQL

Problem 1 : Find the number of corona patients who faced shortness of breath.

Query: -

```
select count(Ind_Id)
from covid_data
where Corona = "positive"
and Shortness_of_breath = "TRUE";
```

OUTPUT :-

There are 1162 corona patients who faced shortness of breath.

Problem 2 : Find the number of negative corona patients who have fever and sore_throat.

Query: -

```
select count(Ind_Id) as NoOfPatients
from covid_data
where Corona = "negative"
and Fever = "TRUE"
and Sore_throat = "TRUE";
```

OUTPUT :-

There are 121 corona negative patients who have fever and Sore throat.

Problem 3 : Group the data by month and rank the number of positive cases.

Query: -

```
select monthname(Test_date) as Month,
count(Ind_Id) as positivecases,
RANK() OVER(order by count(Ind_Id) DESC) AS 'rank'
from covid_data
where Corona = "positive"
group by monthname(Test_date);
```

OUTPUT :-

	Month	positivecases	rank
▶	April	8863	1
	March	5833	2

In April we found most number of corona positive cases(8863), And in march 5863 corona positive cases.

Problem 4 : Find the female negative corona patients who faced cough and headache.

Query: -

```
select * from covid_data
where Sex = "female"
and Corona = "negative"
and Cough_symptoms = "TRUE"
and Headache = "TRUE";
```

OUTPUT :-

Ind_Id	Test_date	Cough_symptoms	Fever	Sore_throat	Shortness_of_breath	Headache	Corona	Age_60_above	Sex	Known_contact
13756	2020-03-22	TRUE	TRUE	TRUE	FALSE	TRUE	negative	No	female	Abroad
17289	2020-03-22	TRUE	TRUE	TRUE	FALSE	TRUE	negative	No	female	Abroad
17657	2020-03-23	TRUE	FALSE	TRUE	FALSE	TRUE	negative	No	female	Abroad
19554	2020-03-23	TRUE	TRUE	FALSE	FALSE	TRUE	negative	No	female	Other

19615	2020-03-23	TRUE	FALSE	TRUE	TRUE	TRUE	negative	No	female	Contact with confirmed
20248	2020-03-23	TRUE	TRUE	FALSE	FALSE	TRUE	negative	Yes	female	Abroad
20253	2020-03-23	TRUE	TRUE	FALSE	FALSE	TRUE	negative	No	female	Contact with confirmed
37904	2020-03-27	TRUE	TRUE	TRUE	TRUE	TRUE	negative	No	female	Contact with confirmed
40616	2020-03-27	TRUE	FALSE	FALSE	TRUE	TRUE	negative	No	female	Contact with confirmed
40752	2020-03-27	TRUE	TRUE	FALSE	FALSE	TRUE	negative	No	female	Contact with confirmed
43650	2020-03-28	TRUE	FALSE	TRUE	TRUE	TRUE	negative	No	female	Contact with confirmed
49678	2020-03-29	TRUE	TRUE	FALSE	FALSE	TRUE	negative	No	female	Contact with confirmed
51034	2020-03-29	TRUE	FALSE	FALSE	FALSE	TRUE	negative	Yes	female	Contact with confirmed
52740	2020-03-29	TRUE	TRUE	TRUE	FALSE	TRUE	negative	No	female	Contact with confirmed
57155	2020-03-30	TRUE	FALSE	TRUE	FALSE	TRUE	negative	No	female	Abroad
58101	2020-03-30	TRUE	FALSE	TRUE	FALSE	TRUE	negative	No	female	Contact with confirmed
70026	2020-04-01	TRUE	TRUE	FALSE	FALSE	TRUE	negative	No	female	Contact with confirmed
76125	2020-04-01	TRUE	TRUE	TRUE	TRUE	TRUE	negative	No	female	Contact with confirmed
84586	2020-04-03	TRUE	FALSE	TRUE	FALSE	TRUE	negative	No	female	Abroad
86104	2020-04-03	TRUE	FALSE	TRUE	TRUE	TRUE	negative	No	female	Contact with confirmed
86109	2020-04-03	TRUE	FALSE	FALSE	TRUE	TRUE	negative	No	female	Contact with confirmed
86145	2020-04-03	TRUE	FALSE	TRUE	TRUE	TRUE	negative	No	female	Contact with confirmed
100555	2020-04-05	TRUE	FALSE	FALSE	FALSE	TRUE	negative	No	female	Contact with confirmed
122243	2020-04-09	TRUE	TRUE	FALSE	TRUE	TRUE	negative	No	female	Contact with confirmed
133573	2020-04-11	TRUE	FALSE	TRUE	TRUE	TRUE	negative	No	female	Contact with confirmed
135059	2020-04-11	TRUE	TRUE	TRUE	FALSE	TRUE	negative	Yes	female	Abroad
137748	2020-04-12	TRUE	TRUE	TRUE	FALSE	TRUE	negative	No	female	Contact with confirmed
145561	2020-04-13	TRUE	FALSE	FALSE	FALSE	TRUE	negative	No	female	Other
149178	2020-04-13	TRUE	FALSE	FALSE	FALSE	TRUE	negative	No	female	Abroad
163207	2020-04-16	TRUE	FALSE	FALSE	FALSE	TRUE	negative	Unknown	female	Contact with confirmed
168703	2020-04-16	TRUE	FALSE	TRUE	TRUE	TRUE	negative	Unknown	female	Contact with confirmed
244441	2020-04-25	TRUE	FALSE	FALSE	FALSE	TRUE	negative	Unknown	female	Contact with confirmed

Conclusion :- 32 female negative corona patients who faced cough and headache.

Problem 5 : How many elderly corona patients have faced breathing problems?

Query: -

```
select count(*)

from covid_data

where Age_60_above = "Yes"

and Shortness_of_breath = "TRUE";
```

OUTPUT :- 286 elderly corona patients have faced breathing problems.

Problem 6 : Which three symptoms were more common among COVID positive patients?

Query: -

```
select

    (select count(*) from covid_data where Cough_symptoms = "TRUE" and corona =
"positive") as Cough_count,

    (select count(*) from covid_data where Fever = "TRUE" and corona = "positive")
as Fever_count,

    (select count(*) from covid_data where Sore_throat = "TRUE" and corona =
"positive") as Sore_count,

    (select count(*) from covid_data where Shortness_of_breath = "TRUE" and corona
= "positive") as shortBreath_count,

    (select count(*) from covid_data where Headache = "TRUE" and corona =
"positive") as headache_count,

    (select count(*) from covid_data where corona = "positive") as positive_count
```

```
from covid_data

limit 1;
```

OUTPUT :-

	Cough_count	Fever_count	Sore_count	shortBreath_count	headache_count	positive_count
▶	6584	5558	1523	1162	2232	14696

Conclusion :- Cough_Symptoms, Fever and Headache are the most common symptoms among Corona positive patients.

Problem 7 : Which symptom was less common among COVID negative people?

Query: -

```
select

    (select count(*) from covid_data where Cough_symptoms = "TRUE" and corona =
"negative") as Cough_count,

    (select count(*) from covid_data where Fever = "TRUE" and corona = "negative")
as Fever_count,

    (select count(*) from covid_data where Sore_throat = "TRUE" and corona =
"negative") as Sore_count,

    (select count(*) from covid_data where Shortness_of_breath = "TRUE" and corona
= "negative") as shortBreath_count,

    (select count(*) from covid_data where Headache = "TRUE" and corona =
"negative") as headache_count,

    (select count(*) from covid_data where corona = "negative") as negative_count

from covid_data

limit 1;
```

OUTPUT :-

	Cough_count	Fever_count	Sore_count	shortBreath_count	headache_count	negative_count
▶	34987	15816	365	384	147	260008

Conclusion :- Headache, Shortness of Breath, Sore throat symptom was less common among COVID negative people.

Problem 8 : What are the most common symptoms among COVID positive males whose known contact was abroad?

Query: -

```
select

    (select count(*) from covid_data where Cough_symptoms = "TRUE" and corona =
"positive" and Known_contact like '%Abroad%') as Cough_count,

    (select count(*) from covid_data where Fever = "TRUE" and corona = "positive"
and Known_contact like '%Abroad%') as Fever_count,

    (select count(*) from covid_data where Sore_throat = "TRUE" and corona =
"positive" and Known_contact like '%Abroad%') as Sore_count,

    (select count(*) from covid_data where Shortness_of_breath = "TRUE" and corona
= "positive" and Known_contact like '%Abroad%') as shortBreath_count,

    (select count(*) from covid_data where Headache = "TRUE" and corona =
"positive" and Known_contact like '%Abroad%') as headache_count,

    (select count(*) from covid_data where corona = "positive"  and Known_contact
like '%Abroad%') as positive_Abroad

from covid_data

limit 1;
```

OUTPUT :-

	Cough_count	Fever_count	Sore_count	shortBreath_count	headache_count	positive_Abroad
▶	1063	820	228	202	319	1866

Conclusion :- Cough and Fever are the most common symptoms among COVID positive males whose known contact was abroad.