PREDICTIVE MODELLING FOR COVID-19 DIAGNOSIS

**Data Source:**

* The dataset was sourced from the 'ABC' government website, containing records of individuals who underwent RT-PCR testing for COVID-19. The dataset spans from 11th March *2020* to 30th April 2020.

**Dataset Description:**

* Total Records: 2,78,848 individuals
* Columns: 11 columns, including 8 features suspected to influence COVID-19 outcomes.
* Outcome Variable: COVID-19 test result (positive or negative).

**Problem Statement:**

* A speedy and accurate diagnosis of COVID-19 is made possible by effective SARS-CoV-2 screening, which can also lessen the burden on healthcare systems. There have been built prediction models that assess the likelihood of infection by combining a number of parameters. These are meant to help medical professionals all over the world treat patients, especially in light of the scarcity of healthcare resources. The current dataset has been downloaded from ‘ABC’ government website and contains around 2,78,848 individuals who have gone through the RT-PCR test. Data set contains 11 columns, including 8 features suspected to play an important role in the prediction of COVID19 outcome. Outcome variable is covid result test positive or negative. We have data from 11th March 2020 till 30th April 2020. Please consider 11th March till 15th April as a training and validation set. From 16th April till 30th April as a test set. Please further divide training and validation set at a ratio of 4:1.

**Section 1: Questions to Answer**

**Covid-19 Dataset observations with SQL Queries**

* **Our dataset contains records of** **2,78,848** **participants**.
* All Covid-19 tests are performed in month of **March** and **April** of year **2020**.
* Dataset also contains information of persons who are **above** **60 years old**.
* **Knows contact** of those who participated in this test are

1. **Abroad**

2. **Contact with confirm**

3. **Other**

* Both **Males** and **Females** are included in this Covid-19 tests.
* **Symptoms** which are associated with Covid-19 as per our dataset are-

1. **Cough**
2. **Fever**
3. **Sore Throat**
4. **Shortness of Breath**
5. **Headache**

**Data Analysis Approach**

* **Data Understanding & Exploration steps we performed-**

1. We imported all necessary libraries require like pandas, numpy, matplotlib, seaborn, plotly and sklearn.
2. We imported our dataset that is in .csv file format and we made copy of it so that any errors will not damage original dataset.
3. In our observation- Except 'Ind\_ID' column, every column is 'object' datatype.
4. For our understanding we renamed to columns as 'Corona' to 'Test\_result' and 'Ind\_ID' to'ID'.
5. In our observation- We have more number of female patients records (1,30,158) in this dataset.
6. After checking unique values in each column we found alphabetical discrepancies in many columns. So, with the help of ‘Replace’ function we removed those alphabetical discrepancies.
7. After evaluating Test\_result column we found that Most number of covid-19 tests are examined on '20-04-2020' = 10921.

* **Handling Missing/Incorrect values-**

1. **In all columns True and False are written in two different alphabetical types. We need to update this typing errors. We also have null(None) values in few columns, we need to remove them.**
2. **After evaluating we observed that we don’t have any ‘Missing Values’ in our dataset. Instead we have "None" as values in many columns. Here we need to replace them with 'Mode' of particular column.**
3. **After replacing ‘None’ values and all anomalies we need to save this file at this stage for MySQL analysis.**

* **Data Visualisation-**

1. **We visualised our columns with help of heatmap and we observed that-**

* Cough symptoms and Fever are highly correlated.
* After that Soar throat and Headache have next higher values.
* Values which are close to 0 are having less correlation and values which are more towards 1 are highly correlated.
* Correlation shows strength of relationship between two variables.

1. **After comparing all Symptoms with Test result we observed that –**

* Among all 'positive test results' most common symptom is 'Cough'.
* Shortness of breath' is least common symptom in 'positive test results'.

1. **After Analysing “Known Contact” column we found that ‘Other’ category in that columns is giving highest count.**

* **Feature Encoding-**

1. **In this step we converted categorical values to continuous values.**
2. **We used ‘astype’ method to change wrong datatype of our column in appropriate type.**
3. **We used ‘map’ method to encode categorical values to continuous values in our columns.**
4. **Since there are no categorical columns, so feature scaling and feature transformation is not required here.**

* **Feature Selection-**

**Machine Learning Approach**

* **Observations based on following SQL questions**-
* Q1. **Find the number of corona patients who faced shortness of breath**.
* We have **1,164** **'Covid Positive' patients** who faced **'Shortness of breath'**.
* Q2. **Find the number of negative corona patients who have fever and sore throat**.
* There are **142** **'Covid Negative'** participants who experienced **'Fever'** and **'Sore Throat'**.
* Q3. **Group the data by month and rank the number of positive cases**.
* There are a greater number of **'Covid Positive'** cases found in month of **'April'** i.e., **8,881** compared to month of 'March' where there are **5,848** positive cases.
* Q4. **Find the female negative corona patients who faced cough and headache**.
* There are **69** **'Female Covid Negative patients'** who experienced **'Cough'** and **'Headache'**.
* Q5. **How many elderly corona patients have faced breathing problems**?
* There are **263** **'Elderly Covid Patients** (60 years and older) who faced **'Breathing Problem'**.
* Q6. **Which three symptoms were more common among COVID positive patients**?
* After comparing values, **Top 3 symptoms** which are common in **'Covid Positive Patients'** are-

1. **Cough** (**6,584**)

2. **Fever** (**5,559**)

3. **Headache** (**2,235**)

* Q7. **Which symptom was less common among COVID negative people**?
* ‘**Least common'** **symptom** among **'Covid Negative'** participants is = Headache (**179 participants**)
* Q8. **What are the most common symptoms among COVID positive males whose known contact was abroad**?
* **'Most common'** **symptoms** among **'Covid Negative Males'** whose known contact was **'Abroad'** are =

1. **Cough** (**532**)

2. **Fever** (**407**)

3. **Headache** (**129**)