**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'*** = ***10,921***.

* **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**-

1. Here we performed ‘***Chi-square’*** test because in our dataset we have more than two ***independent categorical*** variables.
2. After performing Chi-square test we observed that- we have all categorical columns. Here we got ***p\_value*** [Probability value] for columns ***less than*** ***0.05*** so ***all independent columns have relationship with dependent*** ***column*** i.e., 'Test\_result'. We are getting p-value for ***'ID'*** column = ***0.499***. We if p-value is greater than 0.05 then ***we can drop particular column***.

**Machine Learning Approach**

We need to ***predict*** ***whether Covid-19 test will be positive or negative***, basically ***categorical*** ***output***. So we can use algorithms which are best for categorical output here.

Here we are going to use following 4 algorithms:-

1. **Logistic Regression**,
2. **Decision Tree**,
3. **Random Forest**,
4. **K Nearest Neighbors**.

and after comparing all algorithms we will decide which is best fit algorithm for our dataset.

* **Decision Tree Algorithm**- we performed this with help of ***DecisionTreeClassifier***.
* **Random Forest**- we performed this with help of ***RandomForestClassifier***.
* **Logistic Regression** **Algorithm**- we performed this with help of ***LogisticRegression***.
* **K Nearest Neighbors (KNN) Algorithm**- we performed this with help of ***KNeighborsClassifier***.
* After performing ***confusion matrix*** and ***classification report*** on each algorithm we got following results-

### Accuracy of our 4 Algorithms:-

1. **Decision Tree = 98.37%**
2. **Random Forest = 98.37%**
3. **Logistic Regression = 97.81%**
4. **K Nearest Neighbors Algorithms = 97.42%**

From above results we can observe that accuracy of our 4 Algorithms are very close to each other. If we compare all then we can conclude that **'Decision Tree'** or **'Random Forest'** Algorithm are best for our Covid-19 dataset.

* After visualisation of our algorithms, we performed ***Model evaluation*** and ***Optimization*** for our Algorithms like-
* Performing any one of the three methods is usually sufficient for assessing model performance.

1. **KFold cross-validation**,
2. **Cross-validation\_score** and
3. **GridSearch**CV(hyperparameter tuning)

* A ***standard deviation*** of zero (***0.00***) in cross-validation scores usually ***indicates*** that the ***model's performance is*** ***consistent*** across different folds of the cross-validation process. In other words, the model is ***consistently making*** ***predictions*** with ***very similar accuracy*** ***across all subsets*** of the data.
* After our Model evaluation we observed that-

Through cross-validation scoring, we obtained the following mean accuracy scores:

**Logistic Regression = 94.98%**

**Decision Tree = 95.66%**

**Random Forest = 95.68%**

**K Nearest Neighbors Algorithm = 94.59%**

Upon close examination, it's evident that the accuracies of all four algorithms are quite similar. Upon comparison, it is apparent that the ***'Random Forest'*** algorithm ***outperforms the others***, making it the ***most suitable choice for our*** ***Covid-19 dataset***.

* **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**)

**Covid 19 Precautionary Measures and Guidelines**

* **Preventative Measures:**
* **Vaccination:** Getting vaccinated with authorized COVID-19 vaccines is one of the most effective ways to prevent severe illness and reduce the spread of the virus. ***Follow*** your ***country's vaccination guidelines*** and schedules.
* **Mask-Wearing:** ***Wear masks***, particularly in ***indoor settings*** and ***crowded areas***, where physical distancing is challenging. Use masks that meet local guidelines and ***cover both your nose and mouth***.
* **Hand Hygiene:** ***Wash your hands frequently*** with soap and water for at least ***20 seconds***. If soap and water are unavailable, use ***hand sanitizer*** with ***at least 60% alcohol***.
* **Avoid Crowded and Poorly Ventilated Spaces:** ***Limit your time in indoor spaces*** ***with poor ventilation*** and high occupancy. ***Good ventilation*** helps reduce the concentration of virus particles in the air.
* **Cough and Sneezing Etiquette:** ***Cover your mouth and nose*** with a tissue or your elbow ***when coughing*** or ***sneezing***. ***Dispose of used tissues properly*** and ***wash your hands*** immediately.
* **Stay Informed:** ***Stay updated with information*** from reliable sources such as the ***WHO***, ***CDC***, and your ***local health*** ***authorities*** regarding the latest guidelines and recommendations.
* **Treatment and Government Guidelines:**
* **Seek Medical Advice:** If you experience severe symptoms, such as ***difficulty breathing***, ***chest pain***, ***confusion***, or ***bluish lips*** or ***face***, ***seek immediate medical attention***.
* **Medication:** Some treatments like ***antiviral drugs*** and ***monoclonal antibodies*** may be prescribed by healthcare professionals in certain cases. ***Follow your healthcare provider's advice*** regarding treatment options.
* **Quarantine:** If you have been exposed to someone with COVID-19, ***follow local quarantine guidelines*** to prevent potential spread. Quarantine requirements may vary by region.
* **Travel Restrictions:** ***Stay informed about travel restrictions*** and quarantine requirements in your area and at your destination.
* **Testing:** ***Get tested for COVID-19*** as ***recommended by your healthcare provider*** or local health authorities, ***especially if you have symptoms*** or have been exposed to someone with the virus.