**Covid-19 Prediction Analysis**

**Data Source:**

* The data was obtained from the official government website 'ABC' and comprises information about individuals who were subjected to RT-PCR testing for COVID-19. The dataset covers the period from March 11, 2020, to April 30, 2020.

**Data Description:**

* Total records: 2,78,848 individuals
* Columns: 11 columns including 8 features suspected to influence COVID-19
* outcomes
* Target 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:**

1. **Why is your proposal important in today’s world? How can accurately predicting a disease improve medical treatment?**

In the current global scenario dominated by the COVID-19 pandemic, the ability to swiftly and accurately predict the disease status of individuals is crucial. It allows for:

* + - Timely isolation and treatment of positive cases, preventing further spread.
    - Efficient allocation of healthcare resources, ensuring that critical cases receive prompt attention.
    - Reduction in unnecessary testing, thereby conserving resources and minimizing inconvenience to individuals.
    - Early intervention, which can lead to milder symptoms and better overall outcomes.

1. **How is it going to impact the medical field when it comes to effective screening and reducing health care burden.**

Implementing our proposed predictive Modeling approach will revolutionize disease screening in the medical field. By leveraging machine learning, we can:

* Expedite the diagnosis process, especially in regions with limited testing capabilities.
* Alleviate strain on healthcare systems by efficiently identifying COVID-19 positive individuals.
* Ensure that urgent cases receive immediate care, optimizing resource allocation.
* Streamline patient influx, enhancing the quality of healthcare services and potentially saving lives.

1. **If applicable, what is the knowledge gap or how could your proposed method be beneficial for future applications in predicting other diseases?**

* The methodology developed for predicting COVID-19 based on symptoms can serve as a versatile platform for predicting various diseases. The expertise gained in constructing this model, including feature selection strategies and computational methodologies, can be applied to future disease prediction efforts. This has the potential to address knowledge gaps in numerous healthcare sectors, enabling quicker and more accurate identification of disorders beyond COVID-19. The ability to predict diseases accurately holds far-reaching implications for personalized medicine, public health planning, and the development of tailored treatment approaches, ultimately leading to improved healthcare outcomes for individuals and communities alike.

**Section 2: Initial Hypothesis (or hypotheses):**

* **Hypothesis 1:**
* Patients who came in direct contact with Confirmed (Covid Positive) Patients are more likely to be Corona positive.
* This hypothesis suggests that individuals with known contact with confirmed COVID-19 positive patients are at a higher risk of being infected themselves. The belief is that the virus primarily spreads through close proximity and direct contact. This hypothesis can be tested by analysing the variable "Known\_contact" in the dataset to see if 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 essential variables in determining COVID-19 positive or negative cases.
* This hypothesis proposes that symptoms like shortness of breath, fever, and cough are crucial indicators in diagnosing COVID-19 positive cases. These symptoms are commonly associated with respiratory infections and have been recognized as major indicators of COVID-19. This hypothesis can be tested by evaluating the variables "Shortness\_of\_breath," "Fever," and "Cough\_symptoms" in the dataset to study the association between these symptoms and the likelihood of a person being COVID-19 positive or negative.
* It's important to note that these are initial hypotheses and further analysis, including the application of machine learning models, will be necessary to validate and gain deeper insights from the data.
* In **Exploratory Data Analysis (EDA)**, & **Machine Learning (ML)** focus on identifying patterns and important features in the data that may impact a machine learning model.
* Preprocess the data (clean, handle missing values, encode categorical variables).
* Split the data into training and testing sets.
* Apply multiple machine learning models (e.g., logistic regression, decision tree, random forest, etc.).
* Evaluate the models using appropriate metrics (accuracy, precision, recall, etc.).
* Compare the performance of different models and justify your choice of the best model based on relevant cost functions and possibly visualization.

**Section 3: Data analysis approach:**

1. **What approach are you going to take in order to prove or disprove your hypothesis?**

To validate or refute our hypotheses, we will employ a two-pronged approach:

* **Exploratory Data Analysis (EDA):** This initial step involves exploring the dataset to understand its characteristics. We will perform summary statistics, visualizations, and correlation analyses to identify any prominent trends or patterns that may support or challenge our hypotheses.
* **Machine Learning Modeling:** Following EDA, we will apply machine learning classification algorithms. These models will predict COVID-19 status based on the provided features. By quantitatively evaluating the impact of various variables on the outcome, we can determine if our hypotheses hold true.
* By combining both EDA and machine learning Modeling, we aim to conduct a comprehensive analysis that leverages the strengths of both approaches.

1. **What feature engineering techniques will be relevant to your project?**

* **Handling Missing Values:**
* Across all columns, 'True' and 'False' entries are recorded in two distinct alphabetical formats. Rectifying these typographical errors is imperative. Additionally, there are instances of null ('None') values in certain columns which require removal.
* Upon thorough evaluation, it was determined that there are no 'Missing Values' in our dataset. Instead, we have instances of "None" as values in numerous columns. It is essential to substitute these entries with the respective column's 'Mode'.
* Once the 'None' values and any other irregularities have been addressed, it is crucial to save this file at this juncture for MySQL analysis.
* **Feature Encoding:**
* In this phase, we transformed categorical attributes into a format suitable for numerical analysis.
* The 'astype' method was applied to rectify incorrect data types within the columns, ensuring they are in the appropriate format.
* The 'map' method was utilized to translate categorical variables into numerical representations within our dataset columns.
* As there are no categorical columns present.so ,here no need for feature scaling and feature transformation.
* We can also use other encoding techniques like get\_dummies or label encoder but here we are using the above methods.e scaling or transformation in this co

1. **Please justify your data analysis approach.**

* **Data Understanding & Exploration steps we performed-**
* We imported all necessary libraries require like pandas, numpy, matplotlib, seaborn, plotly and sklearn.
* 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.
* In our observation- Except 'Ind\_ID' column, every column is 'object' datatype.
* For our understanding we renamed to columns as 'Corona' to 'Test\_result' and 'Ind\_ID' to'ID'.
* In our observation- We have more number of female patients records (1,30,158) in this dataset.
* 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.
* After evaluating Test\_result column we found that Most number of covid-19 tests are examined on '20-04-2020' = 10921.
* **Comprehensive Understanding:**
* EDA allows us to gain a thorough understanding of the dataset. By exploring summary statistics, visualizations, and correlation analyses, we can uncover initial patterns and relationships within the data. This provides crucial context for building effective machine learning models.
* **Validation through Modeling:**
* Machine learning models provide a quantitative assessment of the relationships between variables and the target variable. By applying classification algorithms, we can objectively test our hypotheses and assess the predictive power of various features. This approach enables us to validate our initial assumptions in a rigorous and data-driven manner.
* **Objective Evaluation:**
* Through Machine Learning Modeling, we can objectively assess the impact of various features on the target variable. This allows us to quantify the significance of our findings and make informed decisions based on empirical evidence.

1. **Identify important patterns in your data using the EDA approach to justify your findings**.

* **Data Visualisation:** 
  1. We looked at our columns using a heatmap and noticed:
* Cough and Fever are closely related.
* After that, Sore throat and Headache are also somewhat connected.
* Values closer to 0 mean weaker connection, while closer to 1 means stronger connection.
* Correlation tells us how strongly two things are linked.
  1. When we checked symptoms against test results, we found:
* Among positive test results, Cough is the most common symptom.
* Shortness of breath is the rarest symptom in positive test results.
  1. When we examined the "Known Contact" column, we found that the 'Other' category had the highest count.
* **Feature Selection:**
  1. In this context, we applied chi-square test is employed to discern whether there exists a notable correlation or interdependence between two categorical variables.
  2. It accomplishes this by contrasting the actual occurrence frequencies of the categories with the anticipated frequencies that would occur if the variables were unrelated.
  3. The resulting p-value from the test offers insights into the likelihood of obtaining the observed outcome, or one even more extreme, assuming the variables are unrelated.
  4. A low p-value signifies a substantial correlation. This test proves particularly advantageous in scenarios involving hypothesis testing, as it aids in the evaluation of whether a noteworthy relationship exists between two categorical variables.
  5. Here we did chi square test because we have all categorical columns. Here we got P\_value for columns less than 0.05 so all independent columns have relationships 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.

**Section 4: Machine learning approach:**

1. **What method will you use for machine learning based predictions of COVID19?**

* Logistic Regression: This is a simple and interpretable algorithm that can be used for binary classification tasks, such as predicting COVID-19 status (positive/negative).
* Decision Trees and Random Forests: These ensemble methods are versatile and can handle both classification and regression tasks. They are capable of capturing complex relationships in the data.
* K-Nearest Neighbors (KNN): KNN is a simple and intuitive algorithm for classification tasks. It can be effective in scenarios where similar cases tend to have similar outcomes.

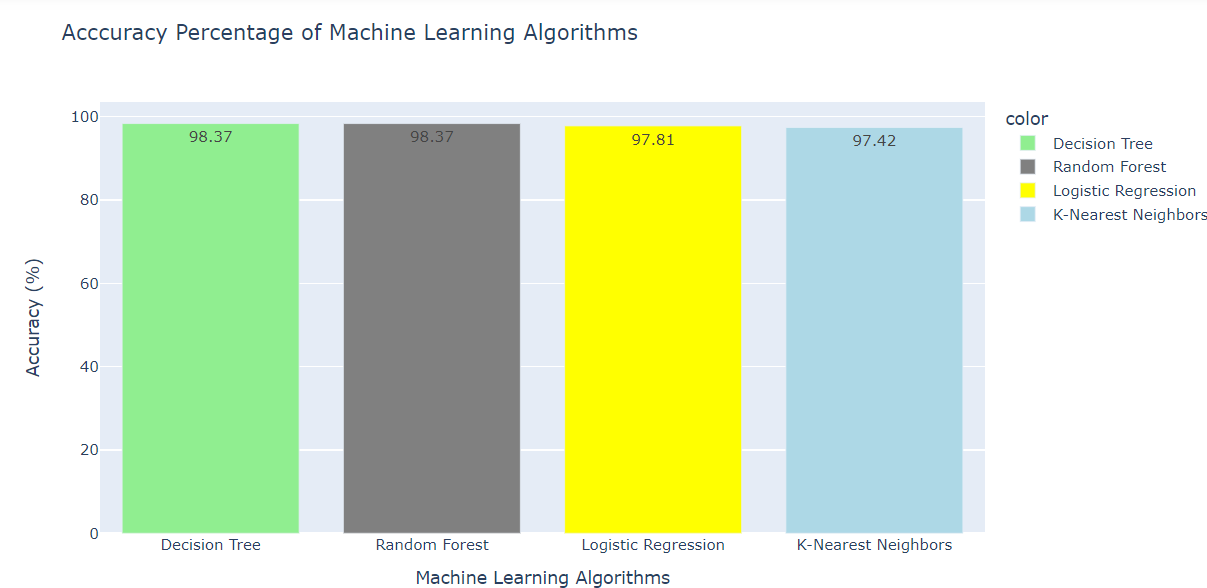
1. **Please justify the most appropriate model.**

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.

1. **Please perform necessary steps required to improve the accuracy of your model.**

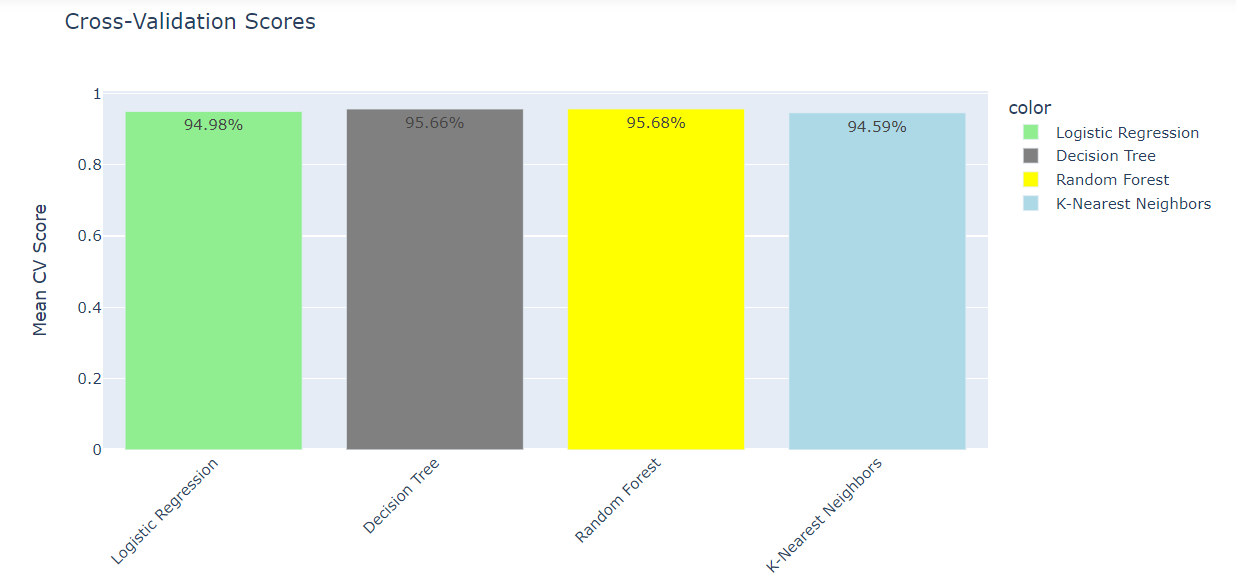


We obtained the following scores for our four algorithms using precision, recall, F1-score, and confusion matrix:

* Logistic Regression = 97.81%
* Decision Tree = 98.37%
* Random Forest = 98.37%
* 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.

**4.Please compare all models (at least 4  models).**



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.
* KFold cross-validation,
* Cross-validation\_score and
* GridSearchCV(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.

A.) 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.

A.) 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.

A.) 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.

A.) There are 69 'Female Covid Negative patients' who experienced 'Cough' and 'Headache'.

Q5. How many elderly corona patients have faced breathing problems?

A.) here are 263 'Elderly Covid Patients (60 years and older) who faced 'Breathing Problem'.

Q6. Which three symptoms were more common among COVID positive patients?

A.) After comparing values, Top 3 symptoms which are common in 'Covid Positive Patients' are-

Cough (6,584)

Fever (5,559)

Headache (2,235)

Q7. Which symptom was less common among COVID negative people?

A.) ‘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?

A.) 'Most common' symptoms among 'Covid Negative Males' whose known contact was

'Abroad' are =

Cough (532)

Fever (407)

Headache (129)

**COVID-19 Safety Protocols 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.

**Physical Distancing:** Maintain physical distance (e.g., at least 6 feet) from individuals who do not live in your household, especially in crowded places.

**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.

**Regular Cleaning:** Clean and disinfect frequently-touched surfaces in your home, workplace, and public areas.

**Treatment and Government Guidelines:**

**Isolation:** In case of a positive COVID-19 test result or the presence of symptoms, adhere to the local directives for self-isolation to prevent transmission to others.

**Seeking Medical Attention:** If you encounter severe symptoms like breathing difficulties, chest pain, confusion, or bluish lips or face, promptly seek professional medical assistance.

**Medication:** Certain treatments such as antiviral medications and monoclonal antibodies may be prescribed by healthcare experts in specific situations. It is important to follow your healthcare provider's recommendations regarding treatment choices.

**Quarantine:** If you have come into contact with an individual who has tested positive for COVID-19, follow the applicable local guidelines for quarantine to reduce the potential for transmission. Quarantine requirements may vary depending on your location.

**Testing:** Under the guidance of your healthcare provider or local health authorities, undergo COVID-19 testing, especially if you exhibit symptoms or have been in close proximity to someone with the virus.