# FLOW -

# Machine Learning-based Prediction of COVID-19 Diagnosis based on Symptoms.

◇Novel Corona virus- COVID\_19 Introduction-

**▹Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus known to cause illness ranging from the common cold to more severe diseases such as Severe Acute Respiratory Syndrome (SARS).**

**▹A novel coronavirus (COVID-19) was identified in 2019 in Wuhan, China. This is a new coronavirus that has not been previously identified in humans. Anyone can get sick with COVID-19 and become seriously ill or die at any age.**

**▹The pandemic has had widespread social, economic, and health implications, leading to changes in daily life, travel restrictions, and the adoption of remote work and learning.**

**▹The outbreak of COVID-19 affected the lives of all sections of society rapidly leading to physical and mental illness of people. Such intense outbreak put huge burden on HealthCare System and Authorities.**

**▹To prevent the spread of the virus, public health measures include practicing good hand hygiene, wearing masks, maintaining physical distance, and getting vaccinated.**

**▹Given below picture depicts chained strategies effective in controlling disease and its outbreak. In project we will Aim to train model to predict covid report considering certain factors.**

**A diagram of a medical process

Description automatically generated with medium confidence◇Importance of proposal**

**▹Why and how predicting a disease accurately can improve medical treatment?**

**-In today's fast world predicting precise medical condition is very important and crucial because-**

Accurate prediction allows a swift and timely interference, enabling healthcare workers, aiders to implement necessary aid measures.

Conditioned person can be provided with essential medical resources, which can ensure quick support to sensitive high-risk areas.

It will also help in promoting public sense of responsibility, preventive and healthy behaviour.

Accurate prediction significantly reduces the burden of healthcare worker, authorities by managing resources timely and effectively.

**◇Impact on the medical field**

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

**-Predicting accurate disease will effectively reduce stress of medical field-**

Accurate predicting help in effective screening, enabling fast identification of risk areas and effective diagnose.

Accurate prediction can help in optimize resources allocation, infection case numbers, staff arrangements, medical equipment and other necessary facilities.

Healthcare institution can timely forge strategies and can control spread of disease. Providing better resource allocation in sensitive zone.

**◇Significance of proposed method**

**▹what is the gap in the knowledge or what proposed method can be helpful if required in future for any other disease?**

Proposed analysis objective is to fill the gap in comprehending the relation between various parameters influencing spread of COV-19 and forging more informed decision.

A comprehensive study and research can develop adaptable method which may be useful in predicting and understanding infectious disease in future.

**ABSTRACT**

The proposal not only addresses the immediate challenges of COVID-19 prediction but also lays the framework for a more flexible and adaptable method to disease prediction and risk management in the future.

**UNDERSTANDING THE DATASET AND FEATURES**

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 considering 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, suspected to play an important role in the prediction of COVID19 outcome. Outcome variable is covid result test positive or negative.

The following list describes each of the dataset’s features used by the model: -

1.**Ind\_ID** (Individual ID): - A unique identifier for each individual in the dataset.

2.**Test\_date**: - The date when the test was conducted.

3.**Cough\_symptoms**: - Indicates the presence or absence of cough symptoms.

4.**Fever**: - Indicates the presence or absence of fever.

5.**Sore\_throat**: - Indicates the presence or absence of a sore throat.

6.**Shortness\_of\_breath**: - Indicates the presence or absence of shortness of breath.

7.**Headache**: - Indicates the presence or absence of a headache.

8.**Corona:(TARGET\_VARAIBLE): -**Related to the presence or absence of coronavirus. It might indicate test results or other relevant information.

9.**Age\_60\_above**: - Indicates whether the individual is 60 years old or above.

10.**Sex**: - Indicates the gender of the individual.

11.**Known\_contact**: - Indicates whether there is known contact with someone who has coronavirus.

**TABLE: -** Report of characteristics estimate of important features in table.

|  |  |  |
| --- | --- | --- |
| FEATURES | Each count  TOTAL: - 278848 | PERCENTAGE (%) |
| Sex (male/female). | female 130158  male 129127  None 19563 | Female: - 46.68%  Male: - 46.31%  Unknown gender: - 7.02% |
| Age≥60\_above\_years(true/false) | None 127320  No 125703  Yes 25825 | Yes: - 9.62%  No: - 45.10% |
| Cough (true/false) | False 236620  True 42228 | True: - 15.14%  False: - 84.86% |
| Fever (true/false) | False 257096  True 21752 | True: - 7.81%  False: - 92.2% |
| Sore throat (true/false) | False 276922  True 1926 | True: - 0.71%  False: - 99.31% |
| Shortness of breath (true/false) | False 277271  True 1577 | True: - 0.57%  False: - 99.43% |
| Headache (true/false) | False 276434  True 2414 | True: - 0.87%  False: - 99.13% |
| Corona (positive / negative) | negative 260227  positive 14729  other 3892 | Negative: - 93.3%  Positive: - 5.2%  Other: -0.14 |

*~STEPS: -*

*1.Anomalies: -*

*-In gender, Age, Corona column we can spot unknown or unclassified values presence. Which are as follows: -*

*>> In Gender columns ~7.2% unclassified value.*

*>> In Corona ~0.14% diagnosis is reported as “other”. In our set we aim to classify report either positive or negative.*

*>> Age classification (above 60 or not) is filled most unknown values (~45.66%)*

*>>These errors and other occurred errors Which we will rectify as we proceed further in feature engineering.*

*2.Distribution: -*

*-From above table it is clearly seen data in majority of features are highly imbalance.*

*-Nearly all variables are binary classified.*

*-It is important to note that almost all dimensions are object data type including Target variable.*

*-Accordingly, we will proceed in selection and training of model.*

*3.Contact Information: -*

*-Individual contact information with infected person or region is classified in “Known contact” variable by three categories which is shown in snippet below: -*

*A screenshot of a computer

Description automatically generated*

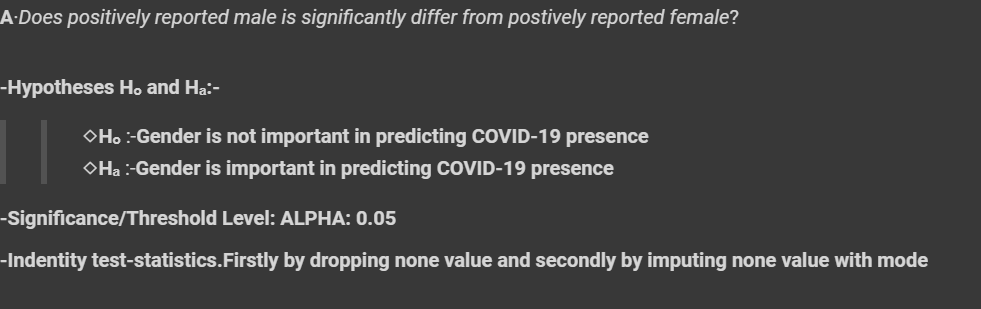
*In above attachment we can observe known contact in three categories (Abroad, Other, confirmed with contact).*

4. Analysis: -

*-We will visualize data by plotting Count plot, Bar plot, Cross tab for further gains.*

*-Followed by 2 hypothesis testing –*

*>i) Relation between Gender and Covid \_outcome.*

**

*>ii) Whether Age>60 above and Corona test are independent or not.*

*A screenshot of a computer

Description automatically generated*

*5. Encoding: -*

*-Identify the type of variables we have. Given data comprises of*

*>one integer column- ‘Ind ID’*

*>One date columns- ‘Test date’*

*>rest are categorical entities which are further classified as: -*

*A screenshot of a computer

Description automatically generated*

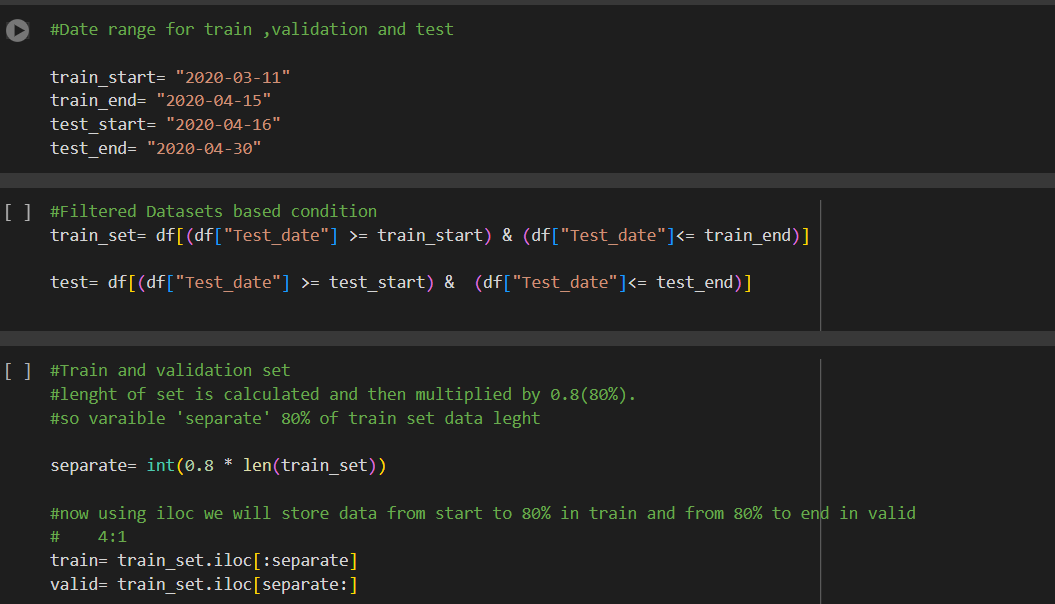
*- To make machine understandable encode value using ‘Nominal encoding’ and ‘Label encoding’ for Target value.*

◇Split of Dataset: -

*Provided set is divided into three parts Train set, Validation set, Test set.*

*Objective: -*

*>>Initially we split data in train and test. Then train further trained set divided into 4:1 into train and validation.*

**

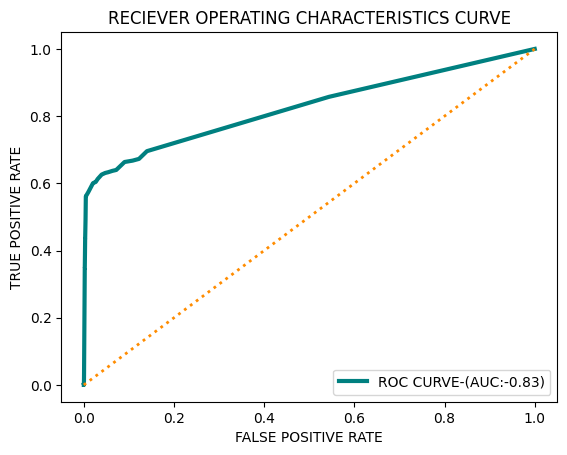
*>> we train 6 model and validate it’s performance by validation set on trained data. We aim to which model is performing is better. We have selected models for classification as shown below-*

*A screenshot of a computer

Description automatically generated*

*>>Further, we proceed with our chosen “Random Forest Classifier and XGBoost Classifier” model for ‘K-fold cross validation’ and Hyperparameter tuning using “Random search cv” to enhance model performance.*

*>>With final “XGBoost Classifier” model we will plot ROC-AUC (Recursive Operating Characteristics) curve to examine model’s overall capability of predicting true instances.*

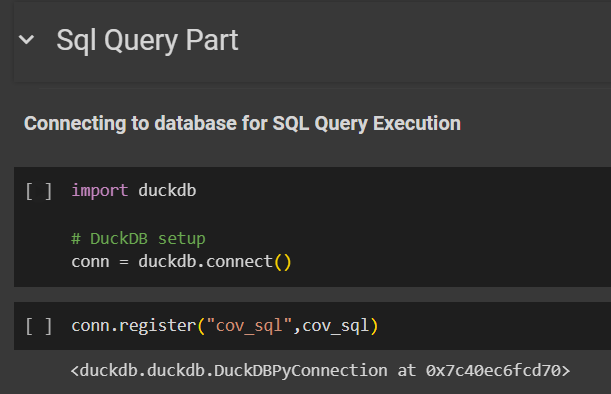
**

*6.SQL-QUERIES: -*

*-To perform SQL queries, we will use Duck DB*

*-Duck DB is OLAP database used to analyze data in fast and efficient way. It leverages a SQL query execution engine capable of running queries on dataset.*

*-First step to using database system is to insert into that system. Duck DB provides several data ingestion methods that allow you to easy fill up the Database. In below section we can see calling of Duck DB and establishing connection-*

**

END!!