Covid vaccine analysis

ABSTRACT : The Covid-19 pandemic has shaken the world completely. No one knew what was coming and everyone was running helter-skelter. The governments were paralyzed and the infrastructure required to deal with this problem was absent completely. The genome sequence was out. But what the disease entailed and what it will lead out was just anyone's imagination. Till today as we write there are multiple dimensions of it that lay unexplored and need a deep exploration to be found out. Our Project seeks to uncover the mystery using the application of data sciences to solve it. We seek to use data sciences to help authorities and also to give the medical field the insight that data can provide to them to deal with the pandemic better. Data science is the application of data science algorithms and machine learning to train the models to find patterns. Patterns reveal what the common issues are and common symptoms and everything that is common comes out in a visual representation. It's these representations which make complex things easy and digestible to people from non tech backgrounds.

Use of data science in such a pandemic will lead to greater insights in the data we are working on. A huge dataset of people suffering from Corona virus to give us better ways of fighting the pandemic. Data-sciences in our project is being applied to just the Corona virus but its applications are wide ranging and can be applied across sectors of diseases to diagnose better. In-fact data science is the new method of diagnostic and can lead to even better cure for diseases. It's this frontier we seek to find from our project.

1. INTRODUCTION

Covid-19 cases are increasing day by day in and all over the world, millions of people are dying and the economy is experiencing free-fall. It has been spreading like water in the open ocean and it seems like there is no stopping it, But thankfully since the inception of this epidemic many countries have properly managed the databases of each and every patient and their health history. We today have advanced computational infrastructure and data science algorithms through which we can analyse these data sets and gain insight-full information so that we can help the society. People have proposed many interesting models and trend prediction methods.The project will help us in recognizing the insights that will be gained by using data science algorithms on the data, these insights will help us in identifying and giving an idea of how the number of covid cases are impacted as possibility of being diagnosed positive on the basis of the symptoms .

2. LITERATURE SURVEY

According to the research paper [1], the authors, R. Wang, G. Hu, C. Jiang, H. Lu and Y. Zhang, have compared the prediction of patterns by using 3 methods and comparing their graphs with each other. These models are the conventional logical regression model, the Particle Swarm Optimization SIR model and the Lowest Square approach SIR model. The chart ultimately shows some patients with a novel form of X-axis coronary pneumonia, and Y-axis date. By seeing the three patterns we come to know that the data is plotted in the form of a curve.

"The public figures of daily updated confirmed instances of Covid-19 from University John Hopkins were analysed in this study article [2] proposed by V.Z.Marmarelis.[2]. RM as described by Riccati Equation, is the main modelling element for the method. The public figures of daily updated confirmed instances of Covid-19 from University John Hopkins were analysed in this study article [2] proposed by V.Z.Marmarelis etal. [2]. RM, as described by Riccati Equation, is the main modelling element for the method. Further by applying the equation we find 5 different parameters and their dependence on the no. of cases increasing day by day".

Everyone analysed knowledge on coronary disease and sustainable therapy utilising research articles from Gerry Wolfe*, Ashraf elnashar*, Will Schreiber* Izzat Alsmadi*. " Guided by COVID-19 Literary Clustering of the Datasets from Kaggle based on COVID-19. [3] The data were further divided into four: (1) Mobility social distances, (2) Health and COVID; (3) Economic impact; and (4) Vulnerable population, and were utilised in a second dataset from MTI. The document has been analysed and text has been processed in order to produce tokens for clustering and the use of the K-Median method to label data to assist extract and analyse categorised data.

According to Tuli,[4] the epidemic may be tracked extremely efficiently via Shrestha et al Machine Learning (ML) and Cloud Computing, anticipate an outbreak of the illness, and create appropriate policies to regulate its expansion. Then given the array, face extraction and collection is done. They have proposed a Machine Learning model that can be run continuously on Cloud Data Centers (CDCs) for accurate spread prediction and proactive development of strategic response by the government and citizens. The dataset used by them in this case study,World in Data by Hannah Ritchie. They have also used a cloud framework and azure instances for real time analysis of data.

The research paper [5] Francisco Nauber,Bernardo Gois et al. have emphasised the rising popularity of epidemic behaviour prediction research due to their capacity to anticipate the natural course of viruses. This study presents several predictor approaches with machine training, logistic regression, filters, and epidemiological models in order to explain COVID-19's behaviour.

The research paper [6], the authors Yazeed Zoabi, Shira Deri-Rozov and Noam Shomron have acknowledged that accurate SARS-CoV-2 screening allows for fast and efficient COVID-19 diagnosis and reduces the strain on health care systems. Prediction models using many characteristics have been created to assess the likelihood of infection. The model projected 0.90 auROC in the forward-looking test set (area under the receiver operating characteristic curve).

The research paper [7], authors Enis Karaarslan and DoğanAydın mentioned thatThe incident at COVID-19 showed that the world was unwilling to disseminate the virus so rapidly. One crucial factor in mitigating the detrimental impacts of an epidemic or pandemic is the effective use of information technology. They suggested a management epidemic system (EMS), which relies on the unfettered and timely flow of information between states and organisations. They have been using an MPISA paradigm, which allows different platforms to be integrated and gives the solution for issues of scalability and interoperability.

[8] This paper Describes the use of a new epidemiological compartiment-based model for the estimation of the propagation of the coronavirus CO VID-19, that is, SEIAR(Susceptible Exposed Asymptomatic Infectious Recovered). This is accomplished through the heuristic approach of differential

evolution. In this way the day(s) when that number reaches its maximum, the associated value and the future evolution of its spread may be evaluated in approximate order for different situations.

The [9] authors Ayyoubzadeh S et al haveUsed computerised data mining technologies for improved insights on the outbreak of COVID-19 in each country and globally for the management of the health catastrophe. Google Trends website collected data. For estimating the number of positive COVID-19 instances, linear regression and long-term memory (LSTM) models were utilised.

[10] The study document [7] by Amir-Sardar Kwekha Rashid,Heamn N Abduljabbar and Bilal Alhayani shows that in COVID-19 research, hypotheses may be proved to be deterministic, transforming into clear findings and predictions. The outcomes of supervised learning algorithms are better than those of 92.9% of uncontrolled learning algorithms. The assistance for the development of standard diagnostic procedures like IgM, IgG, X-ray chest, CT-scans and RT-PCR can be seen as an artificial intelligence and deep learning. The CNN Algorithms selected to perform this study are MobileNet, DenseNet, Xception, ResNet, InceptionV3, InceptionResNetV2, VGGNet, NASNet.

3. IMPLEMENTATION

3.1 Methodology

We are using Machine Learning to give predictions on the basis of data taken from government website[11], and then we clean the data by using excel cleaning methods and give prediction by using the algorithm with highest accuracy to predict COVID -ve or +ve on basis on 5 major symptoms.

The process can be explain in following points :

1. First, Take the dataset, remove redundant data and organise the data according to our needs.
2. Second, Load the dataset on the Jupyter Notebook and apply data visualization techniques to understand the data better.
3. Third, then we calculate accuracy for various algorithms and plot a graph on the basis of accuracy of various algorithms.
4. Finally, using the accuracy graph we finally use the algorithm with best accuracy in this case (Decision Tree Classifier) to predict the person is either -ve or +ve on the basis of symptoms.

3.2 Description of the Process

We are building our own COVID Prediction System using Jupyter Notebook.

We can describe the process in following steps :

Step 1: Cleaning the dataset

The very first step in our project is to get a reliable and authentic dataset for the prediction and analysis.

Our search for dataset ended on [11] which is govt website which has provided dataset for free use and is absolutely authentic.

Then next thing we did was to clean the dataset and remove unwanted columns from dataset for faster computation.

Step 2: Data Visualization

Here, we use the dataset and check the consistency of the dataset by checking the values out of the dataset randomly.
Then we do data visualization for better understanding of data by the use of various plots, graph and heatmaps.
All this graphs and plots gets us an insight into huge datasets easily.

Step 3: Computing Accuracy
In this step we compute accuracy of all the algorithms by checking the four algorithms mentioned here: Logistic Regression, KNN, Random Forest Classifier, Decision tree Algorithm , we selected these algorithms on the basis of their qualities of regression & classification.

Step 4: Predicting Covid +ve or -ve
In the last step, all we need to do is plot a graph of accuracy of all the algorithms and use the algorithm with best accuracy to predict whether a person has corona or not.
We take input of 5 symptoms in binary values and using our predictor we predict the person is positive or negative on the basis of these 5 symptoms.

3.3 Algorithm

1. Logistic Regression

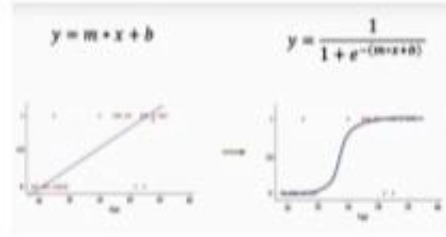
Logistic Regression is a

Classification model, which tries to classify the data based on the probability of it occurring
This algorithm is used in multiple places where classification is required, we have used it to classify if the patient is susceptible to be infected by covid or not
This is one of the classification methods which we have used
It uses Sigmoid function to classify the data

$$sigmoid(z) = \frac{1}{1 + e^{-z}}$$

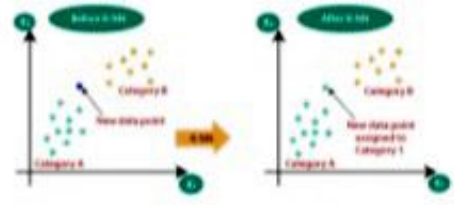
e = Euler's number = 2.71828

Sigmoid function converts input into range 0 to 1



2. KNN

KNN is a supervised machine learning algorithm
KNN forms groups based on the criterias and then decides for the incoming data where to put in in which category
It can be used for regression and for classification too, but mostly for the classification only its used



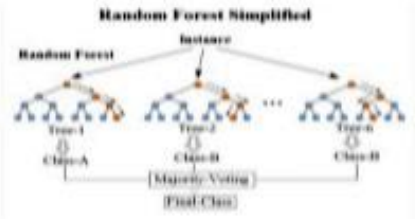
3. Random Forest Classifier

Random forest is a supervised learning algorithm. The "forest" it

builds is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.

Put simply: random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems

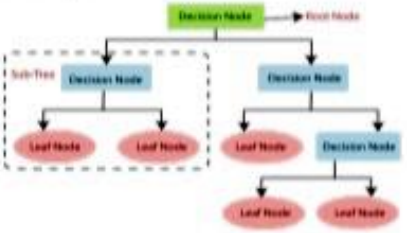


4. Decision tree Algorithm

a.Decision Tree is a supervised learning algorithm

b.Two nodes which are decision node and leaf node are the ones making the decision

c.Repeated if clauses are at work when deciding the classification for the algorithm



4. SYSTEM REQUIREMENTS

4.1 General Description

Data Analytics on Covid-19, as the name suggests is a data analytics on the data such as the people infected,what their age is ,what are the sources that they have been infected from, history of any previous chronic diseases etc. and we wish to obtain almost all the meaningful insights that we can get using various data science and machine learning techniques and by looking at those insights we can arrive at or basically predict the future trends or other crucial information. It requires active internet connection because the project uses various Machine Learning models depending on how we want to train our data.The various tools and library that we intend to use are with the intention that using them we can get the "best of the waste" and provide some services to the society.Hence we look forward to achieve what we have intended and hope the analysis turns out to be a success.

4.2 HARDWARE REQUIREMENTS

1. High Resolution Camera
2. RAM: 4 GB
3. Processor: Intel i5 or Higher
4. 2 GB Graphics Card

4.3 SOFTWARE REQUIREMENTS

1. Windows 7 or Higher
2. Text Editor
3. Python 3.9.0
4. Open CV
5. Jupyter Notebook

4.3.1Non-functional and functional requirements

System functional requirement defines the operations and services to be provided by the system

:-

1. Using **Jupyter Notebook** the csv file is manipulated for getting meaningful insights.

2. **OpenRefine** for data scrubbing.
3. **Numpy,Pandas,Matplotlib** for data exploration,inspection and visualisation.
4. For modeling the data we need a decent knowledge of the **Scikit** library of Python.
5. Training the dataset
6. **Matplotlib,ggplot,Seaborn,Tableau** or d3js for interpreting the data..

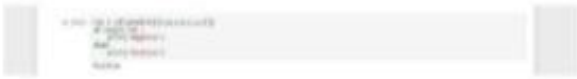
Non-functional Any features or qualities of the system capable of evaluating its operation are the requirements. They are clarified by the following points:

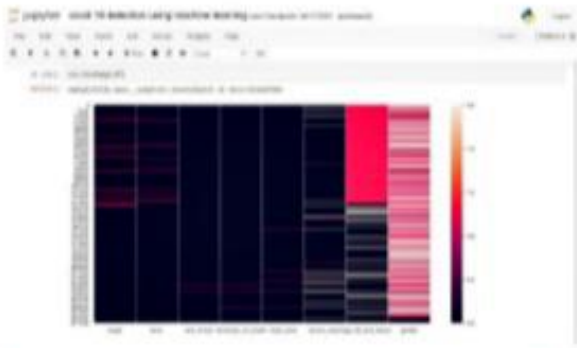
1. **RELIABILITY**:- The insights that we are aiming to obtain should be highly reliable with minimum faults or miscalculations.Every parameter of the dataset is mentioned and observed properly and the insights that we arrive at, are cross checked from practical/previous observations.
2. **SCALABILITY**:- Since new records are added to our dataset on a daily basis our model should be scalable to adopt the dynamic nature of our dataset.
3. **SECURITY**:- Our project is mainly dependent on the covid19 database from an open source data repository ,there is a high chance of data loss due to hackers or attackers.So our system should be secured by using anti-malware software,regular backup etc.
4. **MAINTAINABILITY**:- The system requires good maintainability from our side due to the dynamic nature of the dataset.Since there might be days when there is a sudden surge in the number of daily cases abruptly and we need to be ready for such data too.

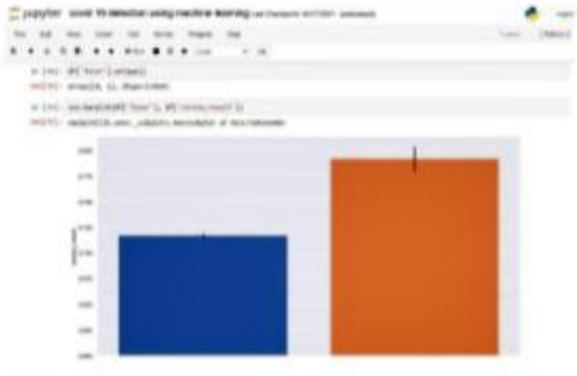
4.3.2 USER REQUIREMENTS

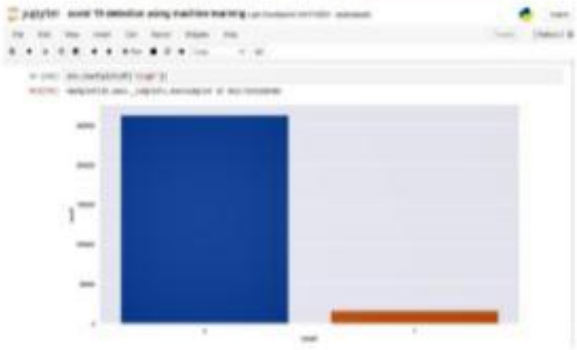
1. The data analysis system shall input and accurately compare the given parameters with the previously stored data.
2. Upon comparing the new input parameters the probability of having covid or not is displayed as a percentage.
3. A front-end interface for taking the symptoms parameters from the patients is present.
4. The user's parameters are compared against the test cases on which the model has been trained.
5. The user shall keep his/her devices connected to our database.

5. RESULTS









The screenshots above show the code and results of the various phases of the Data Analysis done by us on our Covid-19 dataset. The implementation of data analysis has been carried out by various algorithms based on their accuracy. When analysis was done by using various algorithms the most accurate results were yielded by the random forest classifier algorithm. We,while carrying out the analysis, took into consideration the major characteristic features like cough,fever,etc. which largely affect the result of whether the person is positive or negative based on these symptoms.In the later phases we were also able to determine whether the person was covid negative or positive based on his input data which is being taken by a small tkinter interface.

6. CONCLUSION

The Covid - 19 Pandemic is a huge struggle for all of us. The project we are making will seek to find the answers to the most pertinent questions as to what is it that makes the covid 19 such a tragedy and what all people are the ones who are most affected by it. It will seek to find the appropriate response which can be mounted by the authorities concerned and we can reach to a place of proper discussion about the problem and solve it in the best possible manner out there. It will also lead to a solution to any medical condition we might encounter later on in our lives where we can apply data sciences for medical diagnostics. This project saves on the already limited resources that India have and prevents the spread as people can use it to get an idea that they should go and get tested .It also helps unhealthy and infected people to isolate themselves. Using this system we can effectively and efficiently mitigate the burden on our healthcare system which is completely stressed out.

7. REFERENCES