PROJECT REPORT

Project : COVID-19 Vaccine Tracker

Guided By:

Dr. Nilamadhab Mishra

Group Members (Group-8):

Aftab Hussain (18BCE10330)

Prajwal Sharma (18BCE10188)

Sarthak Mishra (18BCE10241)

Tushar Nachan (18BCE10282)

Mritunjay Singh (18BCE10163)

ABSTRACT

Predictive Model in R for tracking and predicting release of COVID-19 vaccine and using it with an R app built using shiny package.
In the current situation of pandemic due to COVID-19 our predictive model will let the users to track on all leading pharmaceutical groups, organizations, medical research institutes, biotechnological companies like Novavax, Bharat Biotech, Moderna, Johnson and Johnson, CanSino Biologics,etc. on their progress on developing vaccine for SARS-CoV-2. Our predictive model will predict which organization will be able to produce vaccine sooner. Our model will produce outputs such as ranked list of the organizations, which will also mention vaccine name, mechanism used, sponsor of the vaccine, Trial Phase, Institution name, predicted release date, announced release date, side effects till now (if human trials have started), effectiveness of the vaccine, etc. Also our model will produce plots such as plot on progress versus time of different organizations, vaccine effectiveness versus organization names,etc.
Our model will serve as a relief to the people by giving them an option to use the model to track COVID-19 vaccine and predict its release date with high precision.

Problem Statement

Problem :
In the current situation of COVID-19 pandemic, individuals do think every now and then that when this all will be over and everything gets back to normal with a launch and distribution of a COVID-19 vaccine.

Why it is a problem?
The above mentioned problem is a problem because every individual do want to know about the possible date of launch of COVID-19 vaccine but there does not exist a proper platform to access this particular prediction.

How our project is going to be a solution?
Our project is to design a predictive model in R which will produce outputs such as currents status of organizations on COVID-19 vaccine, predicted dates of the possible launch of the vaccines, histograms and other interactive plots based on our data set and we will display these outputs in R shiny app built using shiny package in R. These will serve as Solution to the above mentioned problem.

The Objective Of The Project

General objective :-

The app with the right data and technology to predict the accurate time of release of COVID -19 vaccine to vanish the spread of disease.

Specific objectives :-

- Providing a predictive model which can be deployed by developers or used by individuals to know the possible release time of the vaccine.
- Using shiny package in R to display the outputs from the model to user via an R web app.
- Comparing quality measures of different vaccines under development through plots like histograms.
- Providing data on the current status of vaccines under development.
- Comparing release time of different vaccines under development via plots like histogram.

Approach/Methodology (Flow Chat)

INPUT DATA → **TOOL/TECHNIQUE**

1.Past Data

2.Present Data

[Ex: Company name , Current Phase , month in Which go to next phase , Institute Name]

1.R Studio

2. Function of ML

3. Prototyping

4.Observation

5. Questions and surveys

6.Testing

↓

OUTPUT

Predict the releasing Time COVID – 19 Vaccine

Approach/Methodology:

Procedure :-

- Building our own data set for training and testing of the predictive model.
- Choosing the best machine learning algorithm on the basis of the analysis of the data set.
- Creating our own machine learning model.
- Training and testing of the machine learning model with the data set we built.
- Creating a web app in R using shiny package in R which will display the outputs of our predictive to the user.

Tools used:-

- Rstudio
- Packages like caTools, ggplot2, ElemStatLearn, class, e1701, rpart, randomForest, cluster and other required packages required according the ml algorithm to be chosen.
- Shiny package to create a web app in R.

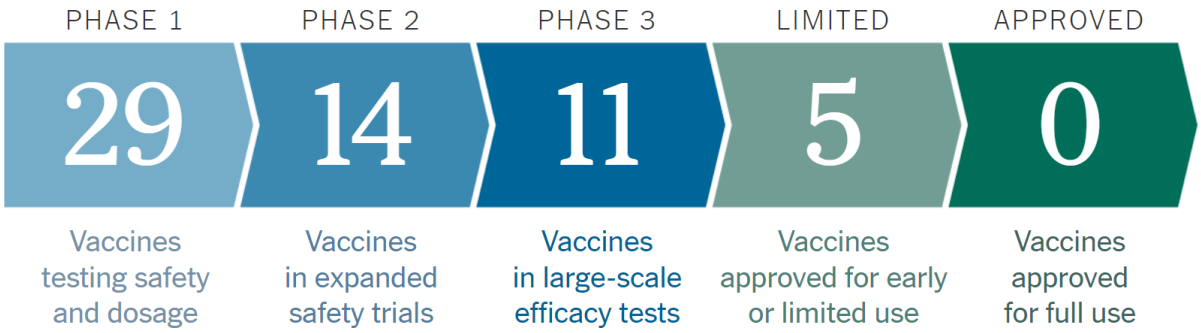
Literature Review/Related Works:

Many firms and professors of IIT's are working on it and many got success
Some of the websites names are :

1. https://github.com/sllloyd/vaccine_predictions
   Vaccine Pipeline Modelling

   The model takes data on existing COVID-19 vaccines in various stages of clinical trials and expert opinions as to their likely success and predicts how many vaccines will get proper regulatory approval and on what timescales. The model uses Monte Carlo techniques to randomly decide an outcome given the input paramaters.

2. https://www.nytimes.com/interactive/2020/science/coronavirus-vaccine-tracker.html
   Coronavirus Vaccine Tracker



This Website gives us an statistical knowledge with details about Vaccines on Phase 1, Phase 2, Phase 3, Emergency Vaccine and approved one's.

3. https://www.raps.org/news-and-articles/news-articles/2020/3/covid-19-vaccine-tracker
   COVID-19 vaccine tracker

   This website basically tells us about the Cadidate, Mechanism, Sponsor, TrialPhase and Institution. This website has almost every COVID-19 Vaccine with their details given in tabular form.

Show 10 entries Search:

	Candidate	Mechanism	Sponsor	Trial Phase	Institution
	Ad5-nCoV	Recombinant vaccine (adenovirus type 5 vector)	CanSino Biologics	Phase 3	Tongji Hospital; Wuhan, China
	AZD1222	Replication-deficient viral vector vaccine (adenovirus from chimpanzees)	The University of Oxford; AstraZeneca; IQVIA; Serum Institute of India	Phase 3	The University of Oxford, the Jenner Institute
	CoronaVac	Inactivated vaccine (formalin with alum adjuvant)	Sinovac	Phase 3	Sinovac Research and Development Co., Ltd.
	JNJ-78436735 (formerly Ad26.COV2-S)	Non-replicating viral vector	Johnson & Johnson	Phase 3	Johnson & Johnson
	mRNA-1273	mRNA-based vaccine	Moderna	Phase 3	Kaiser Permanente Washington Health Research Institute
	No name announced	Inactivated vaccine	Wuhan Institute of Biological Products; China National Pharmaceutical Group (Sinopharm)	Phase 3	Henan Provincial Center for Disease Control and Prevention
	NVX-CoV2373	Nanoparticle vaccine	Novavax	Phase 3	Novavax
	Bacillus Calmette-Guerin (BCG) vaccine	Live-attenuated vaccine	University of Melbourne and Murdoch Children's Research Institute; Radboud University Medical Center; Faustman Lab at Massachusetts General Hospital	Phase 2/3	University of Melbourne and Murdoch Children's Research Institute; Radboud University Medical Center; Faustman Lab at Massachusetts General Hospital
	BNT162	mRNA-based vaccine	Pfizer, BioNTech	Phase 2/3	Multiple study sites in Europe and North America
	Covaxin	Inactivated vaccine	Bharat Biotech; National Institute of Virology	Phase 2	

Showing 1 to 10 of 51 entries Previous 1 2 3 4 5 6 Next

Now the question arises is how is our model different from the models already available on Internet ?

- Our predictive model will take all the past data about that vaccine and will predict the possibility of its approx. manufacturing date and will also predict its failure and delay rate
- There is no app facilitating interactive user interface.

- Our predictive model will let the users to track on all leading pharmaceutical groups, organizations on their progress on developing vaccine for SARS-CoV-2.

Our model will serve as a relief to the people by giving them an option to use the model to track COVID-19 vaccine and predict its release date with high precision.

Scope and Limitations:

Scope :-

There are many approaches to scope a problem. Scoping process is fairly iterative and the scope gets refined both during the scoping process as well as during the project.

Step 1: Goals – Define the goal(s) of the project which is let the users to track on all leading pharmaceutical groups, organizations, medical research institutes, biotechnological companies like Novavax, Bharat Biotech, Moderna, Johnson and Johnson, CanSino Biologics,etc. on their progress on developing vaccine for SARS-CoV-2.
Step 2: Actions – Actions/interventions that this project will inform - Register name, company's name and various details of vaccine. Analyze the data of all trials and make a prediction for release date.
Step 3: Data and its Analysis – Our model will produce outputs such as ranked list of the organizations, which will also mention vaccine name, mechanism used, sponsor of the vaccine, Trial Phase, Institution name, predicted release date, announced release date, side effects till now (if human trials have started), effectiveness of the vaccine, etc. Also our model will produce plots such as plot on progress versus time of different organizations, vaccine effectiveness versus organization names,etc.

Limitations :-

- Our app is a predictive model and will consider the past outliers and fluctuation in pattern of progress of companies and accordingly will

create future predicted possible delays corresponding to different companies. Thus, accuracy will might be compromised a bit due to possible less, none or more delay.

Significance of project:

Our project is significant because in the current situation of COVID-19 pandemic, individuals do think every now and then that when this all will be over and everything gets back to normal with a launch and distribution of a COVID-19 vaccine, our project will serve as a solution which can be used by developers or groups like us designing an app in R using it.
There do not exist any such app out there which provides solution to the mentioned problem with enough precision and with a interactive user interface in R.
There do not exist any such project written in R.

Phases In Project

Phase 1: Research On the Topic (Sep 22, 2020 to Sep 29, 2020)
Phase 2: Gathering information. (Sep 30, 2020)
Phase 3: Building our dataset. (October 1, 2020)
Phase 4: Choosing suitable ML algorithm. (October 6, 2020 to October 6, 2020)
Phase 5: Building our ML model. (October 7, 2020 to October 19, 2020)
Phase 6: Training and testing of the model. (October 20, 2020)
Phase 7: Approval/Disapproval of model followed by modifications in model. (October 21, 2020)
Phase 8: Building a R shiny app for our project (October 22, 2020 to October 24, 2020)

Resource Scheduler Hub Planner used to schedule our human resources and time resources:


September 2020
October 2020
Name
Mo 28 Tu 29 We 30 Th 1 Fr 2 Sa 3 Su 4 Mo 5 Tu 6 We 7 Th 8 Fr 9 Sa 10 Su 11 Mo 12 Tu 13 We 14 Th 15 Fr 16 Sa 17 Su 18 Mo 19 Tu 20 We 21 Th 22
Predective Model For COVID-19 Vaccine Tracking
Gather Inform
Buildi our datase
Learning suitable Machine Learning Algorithm Oct 2nd 2020 - Oct 6th 2020
Building our ML model Oct 7th 2020 - Oct 19th 2020
Traini testing of the
Approval/Dis modification Oct 21st 202
Research on the topic Sep 22nd
Aftab Hussain
Mrityunjay Singh
Prajwal Sharma
Sarthak Mishra
Tushar Nachan
Phase 1 (Predective Model For
Phase 2 (Prede
Phase 3 (Prede
Phase 4 (Predective Model For COVID-19 Vaccine Tracking)
Phase 4 (Predective Model For
Phase 5 (Predective Model For COVID-19 Vaccine Tracking)
Phase 5 (Predective Model For COVID-19 Vaccine Tracking) - General
Phase 6 (Prede
Phase 7 (Pred
Vaccine Tracl
Phase 7 (Pred
Add Resource


October 2020
November 2020
Name
Th 15 Fr 16 Sa 17 Su 18 Mo 19 Tu 20 We 21 Th 22 Fr 23 Sa 24 Su 25 Mo 26 Tu 27 We 28 Th 29 Fr 30 Sa 31 Su 1 Mo 2 Tu 3 We 4 Th 5 Fr 6 Sa 7 Su 8
Predective Model For COVID-19 Vaccine Tracking
Building our ML model Oct 7th 2020 - Oct 19th 2020
Traini testing of the
Approval/Disapproval of ML model and further modification if required Oct 21st 2020 - Oct 27th 2020
Export reports in a
D
Aftab Hussain
Mrityunjay Singh
Prajwal Sharma
Sarthak Mishra
Tushar Nachan
Phase 5 (Predective Model For COVID-19 Vaccine Tracking)
Phase 5 (Predective Model For
Phase 6 (Prede
Phase 7 (Predective Model For COVID-19 Vaccine Tracking)
Phase 7 (Predective Model For COVID-19
Phase 8 (Prede
Add Resource

Phase I : Research On the Topic

I. Understanding Corona Virus:

Coronaviruses are present in many species of animals, such as camels and bats. Mutations of the virus can infect humans.

Coronaviruses typically affect the respiratory system, causing symptoms such as coughing and shortness of breath. Some people, including older adults, are at risk of severe illness from these viruses.

Previous outbreaks of diseases that coronaviruses have caused in humans have been severe. They typically spread rapidly and can cause death in some people.

One example is severe acute respiratory syndrome (SARS), which caused a pandemic in 2002. There were around 8,439 cases and 812 deaths as a result of the virus.

The outbreak of the disease known as COVID-19 is the result of the novel coronavirus, now renamed SARS-CoV-2, that has spread rapidly across many parts of the world.

Effects on the body:

Viruses work by hijacking cells in the body. They enter host cells and reproduce. They can then spread to new cells around the body.

Coronaviruses mostly affect the respiratory system, which is a group of organs and tissues that allow the body to breathe.

Respiratory illnesses affect different parts of this respiratory system, such as the lungs. A coronavirus typically infects the lining of the throat, airways, and lungs.

Early symptoms of coronavirus may include coughing or shortness of breath. In some cases, it can cause severe damage to the lungs.

Usually, the immune system will identify and respond to coronavirus early by sending special proteins, or antibiotics. The immune response to infection has side effects for the body, including fever. During an infection, white blood cells release pyrogens, a substance that causes fever.

A temperature of greater than 100.4°F from an oral thermometer indicates a fever.

Sometimes other symptoms will occur alongside a fever, including:

- breathlessness
- a cough
- muscle pain
- a sore throat
- a headache
- chills
- new loss of taste or smell

These symptoms will usually last until the body fights off the coronavirus.

Symptoms might not show up straightaway. For example, people with COVID-19 may get symptoms 2 to 14 days after infection.

Risks and complications:

Coronavirus can have severe complications, such as pneumonia.

Pneumonia occurs if the virus causes infection of one or both lungs. The tiny air sacs inside the lungs can fill with fluid or pus, making it harder to breathe.

Coronavirus can also damage the heart, liver, or kidneys. In some people, it will affect the blood and immune system. For example, COVID-19 can cause heart, renal, or multiple organ failure, resulting in death.

II. Development of Vaccines:

The general stages of the development cycle of a vaccine are:

- Exploratory stage
- Pre-clinical stage
- Clinical development
- Regulatory review and approval
- Manufacturing
- Quality control

Clinical development is a three-phase process.

During **Phase I**, small groups of people receive the trial vaccine.

In **Phase II**, the clinical study is expanded and vaccine is given to people who have characteristics (such as age and physical health) similar to those for whom the new vaccine is intended.

In **Phase III**, the vaccine is given to thousands of people and tested for efficacy and safety.

Many vaccines undergo **Phase IV** formal, ongoing studies after the vaccine is approved and licensed.

III. Type of vaccines:

Type	**Description**	**Examples of licensed human vaccines**
RNA	Consist of messenger RNA molecules which code for parts of the target pathogen that are recognised by our immune system ('antigens'). Inside our body's cells, the RNA molecules are converted into antigens, which are then detected by our	None

	immune cells.	
DNA	Consist of DNA molecules which are converted into antigens by our body's cells (via RNA as an intermediate step). As with RNA vaccines, the antigens are subsequently detected by our immune cells.	None
Viral vector	Consist of harmless viruses that have been modified to contain antigens from the target pathogen. The modified viruses act as delivery systems that display the antigens to our immune cells. Replicating viral vectors make extra copies of themselves in our body's cells. Non-replicating viral vectors do not.	Ebola
Inactivated	Consist of inactivated versions of the target pathogen. These are detected by our immune cells but cannot cause illness.	Hepatitis A, Influenza, Rabies
Attenuated	Consist of active but non-virulent versions of the target pathogen. These are still capable of infecting our body's cells and inducing an immune response, but have been modified to reduce the risk of severe illness.	Polio, Rotavirus, Measles
Protein subunit	Consist of key antigens from the target pathogen that are recognised by our immune system.	Whooping cough, Hepatitis B
Virus-like particle	Consist of empty viral shells that resemble the target pathogen but contain no genetic material.	HPV

IV. Advantages and disadvantages of vaccine production platforms:

Platform	Target	Existing, licensed human vaccines using the same platform	Advantages	Disadvantages
RNA vaccines	Spike protein	No	No infectious virus needs to be handled; vaccines typically very immunogenic, rapid production possible	Safety issues with reactogenicity have been reported
DNA vaccines	Spike Protein	No	No infectious virus needs to be handled; easy scale up; low production costs; high heat stability; has been tested in humans for SARSCoV-1; rapid production possible	Needs specific delivery devices to reach good immunogenicity.

Recombinant protein vaccines	Spike protein	Yes, for baculovirus (influenza, HPV) and yeast expression (HBV, HPV).	No infectious virus needs to be handled; adjuvants can be used to increase immunogenicity.	Global production capacity might be limited; antigen/epitope integrity needs to be confirmed; yields need to be high enough.
Viral vectorbased vaccines	Spike protein	Yes, for VSV (Ervebo) but not for other viral vectored vaccines	No infectious virus needs to be handled; excellent pre-clinical and clinical data for many emerging viruses including MERS-CoV	Vector immunity might negatively impact on vaccine effectiveness (depending on the vector chosen)
Live attenuated vaccines	Whole virion	Yes	Straight forward process used for several licensed human vaccines; existing infrastructure can be used	Creating infectious clones for attenuated coronavirus vaccine seeds takes time due to large genome size; safety testing will need to be extensive
Inactivated vaccines	Whole virion	Yes	Straight forward process used for several licensed human	Large amounts of infectious virus need to be handled (could be mitigated by using an

			vaccines; existing infrastructure can be used; has been tested in humans for SARS-CoV-1; adjuvants can be used to increase immunogenicity	attenuated seed virus); antigen/epitope integrity needs to be confirmed

Related Works:

- https://github.com/sllloyd/vaccine_predictions
  Vaccine Pipeline Modelling

- https://www.nytimes.com/interactive/2020/science/coronavirus-vaccine-tracker.html

  Coronavirus Vaccine Tracker

- https://www.raps.org/news-and-articles/news-articles/2020/3/covid-19-vaccine-tracker

  COVID-19 vaccine tracker

Reference:

- **https://www.who.int/immunization/programmes_systems/policiesstrategies/vaccine_intro_resources/nvi_guidelines/en/**

- **https://youtu.be/BtN-goy9VOY**

- **https://vac-lshtm.shinyapps.io/ncov_vaccine_landscape/#**

- **https://marlin-prod.literatumonline.com/pb-assets/journals/research/immunity/SARS-CoV-2%20vaccines%20status%20report.pdf**

Phase II: Gathering Information For Our Data Set

Our Model basically takes a dataset having all the past data which helps it to predict the future COVID vaccine release date, There are many factors affecting the release date of covid vaccine, our dataset contains all that factors and after keeping that factors in its memory it helps to give an approx. date for the release of vaccine. Such factors are as follows:

1. Funder
2. Developer
3. Current Phase
4. Technology used
5. Phase 1 start date
6. Phase 1 end date
7. Phase 2 start date
8. Phase 2 end date
9. Phase 3 start date
10. Phase 3 end date
11. Phase 4 start date
12. Phase 4 end date
13. Phase ½ overlap
14. Phase 2/3 overlap
15. And many more…

References for data set are as follows :

1. Factors affecting development rate of vaccine: https://www.bcg.com/publications/2020/covid-vaccines-timelines-implications

2. WHO's useful links to build our dataset : https://www.who.int/publications/m/item/who-target-product-profiles-for-covid-19-vaccines https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines

3. Funding of vaccines :

https://www.devex.com/news/funding-covid-19-vaccines-a-timeline-97950

4. Work Flow of a number of COVID Vaccines : https://racap.com/media/Covid-19/COVID-19_VX_10162020_F.pdf?v=6tLEaP76XtgzTBXPaPA3rDPUOKfSUjaQAWO9bhcTpWg

5. Feasibility Count : http://vaccinedevelopment.org.uk/decision-guide.html

6. Used to Build dataset : https://en.wikipedia.org/wiki/COVID-19_vaccine

Phase III: Building Our Data set

One of the things that we learned in our journey to build this predictive model is that the efficiency and accuracy of a predictive model depends a lot on the data set we're using to in our predictive model. We observed that the data set also drives the choice of machine learning algorithm or other mathematical model to be chosen for the project.

The general objective of our project is to give a predicted date for the possible completion of all the four phases with manufacturing the vaccine at global scale of a covid19 vaccine. So, we have a dependent variable plus_time whose values are in unix epoch time which equals to the more time that probably will be taken by the vaccines developers to complete all the phases and manufacture the vaccines at global scale. So, to predict this dependent variable as precise as possible, on the basis of our gathered knowledge and expert views we tried to add as much as possible column fields which represents the factors affecting rate of development and possible approval of the vaccines.
The column fields used to store the information about the vaccines:

1. **Candidate_name:** Stores the name given to the vaccine.
2. **Developer:** Stores the name of the developers or researchers working on the corresponding vaccine.
3. **Current_Phase:** Stores the current clinical trial phase the vaccine is.
4. **Technology:** Stores the technology being used to develop the vaccine.
5. **Countries:** Stores the name of the countries where the clinical trials of the corresponding vaccine are taking place.
6. **Status:** Stores the recruitment status of the clinical trials.

It can take following values:

- **Not yet recruiting:** The study has not started recruiting participants.
- **Recruiting:** The study is currently recruiting participants.
- **Enrolling by invitation:** The study is selecting its participants from a population, or group of people, decided on by the researchers in

advance. These studies are not open to everyone who meets the eligibility criteria but only to people in that particular population, who are specifically invited to participate.

- **Active, not recruiting:** The study is ongoing, and participants are receiving an intervention or being examined, but potential participants are not currently being recruited or enrolled.
- **Suspended:** The study has stopped early but may start again.
- **Terminated:** The study has stopped early and will not start again. Participants are no longer being examined or treated.
- **Completed:** The study has ended normally, and participants are no longer being examined or treated (that is, the last participant's last visit has occurred).
- **Withdrawn:** The study stopped early, before enrolling its first participant.
- **Unknown:** A study on ClinicalTrials.gov whose last known status was recruiting; not yet recruiting; or active, not recruiting but that has passed its completion date, and the status has not been last verified within the past 2 years.

7. Completion_Date: Stores the unix epoch time when the study is estimated to be completed on.

8. phase0_date: Stores the unix epoch time for date of clinical registration of a vaccine.

9. phase1_start_date: Stores the unix epoch time for the date phase 1 of a vaccine started.

10. phase1_end_date: Stores the unix epoch time for the date phase 1 of a vaccine ended.

11. phase2_start_date: Stores the unix epoch time for the date phase 2 of a vaccine started.

12. phase2_end_date: Stores the unix epoch time for the date phase 2 of a vaccine ended.

13. phase3_start_date: Stores the unix epoch time for the date phase 3 of a vaccine started.

14. phase3_end_date: Stores the unix epoch time for the date phase 1 of a vaccine ended.

15. phase4_start_date: Stores the unix epoch time for the date phase 4 of a vaccine started.

16. phase4_end_date: Stores the unix epoch time for the date phase 4 of a vaccine ended.

17. phase1/2_overlap: Stores time for which phase 1 and phase 2 were overlapping in milliseconds.

18. phase2/3_overlap: Stores time for which phase 1 and phase 2 were overlapping milliseconds.

19. time_taken_till_now: Stores time taken till now in development process in milliseconds.

20. fund: Stores amount of spent till now.

21. disadvantage_count: Stores number of disadvantages of technology used.

22. advantage_count: Stores number of disadvantages of technology used.

23. feasibility_count: Stores number of feasibility criteria satisfied by the vaccine.

24. prv_success: Stores number of successful vaccines produced by the developer or funder.

25. phase1_trials: Stores the number of recruits in phase 1 of the clinical trial.

26. phase2_trials: Stores the number of recruits in phase 2 of the clinical trial.

27. phase3_trials: Stores the number of recruits in phase 3 of the clinical trial.

28. num_cases: Stores current number of cases in the country where trials are going on.

29. **above_60:** Stores whether trials includes recruits older than 60 years.

30. **current_scale:** Stores the numbers of doses produced by the manufacturers currently.

31. **likely_scale:** Stores the numbers of doses will be able to produce till the end of the year.

32. phase1_age: Stores the age group of the recruits in phase1.

33. phase2_age: Stores the age group of the recruits in phase2.

34. phase3_age: Stores the age group of the recruits in phase3.

35. **phase0_time:** time taken in phase 0.

36. **phase1_time:** time taken in phase 01.

37. **phase2_time:** time taken in phase 02.

38. phase3_time: time taken in phase 03.

We used the following references to fill our data set:

- WHO official site providing a pdf file : "Draft landscape of COVID-19 candidate vaccines" : https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines

Factors affecting development rate of vaccines

https://www.bcg.com/publications/2020/covid-vaccines-timelines-implications

Funding for vaccines

https://www.devex.com/news/funding-covid-19-vaccines-a-timeline-97950

Reference for our column : Feasibility count

http://vaccinedevelopment.org.uk/decision-guide.html

To know about the advantages and disadvantages of vaccine technologies

https://marlin-prod.literatumonline.com/pb-assets/journals/research/immunity/SARS-CoV-2%20vaccines%20status%20report.pdf

Related Works:

https://vac-lshtm.shinyapps.io/ncov_vaccine_landscape/#
https://www.raps.org/news-and-articles/news-articles/2020/3/covid-19-vaccine-tracker

https://www.nytimes.com/interactive/2020/science/coronavirus-vaccine-tracker.html
https://biorender.com/covid-vaccine-tracker

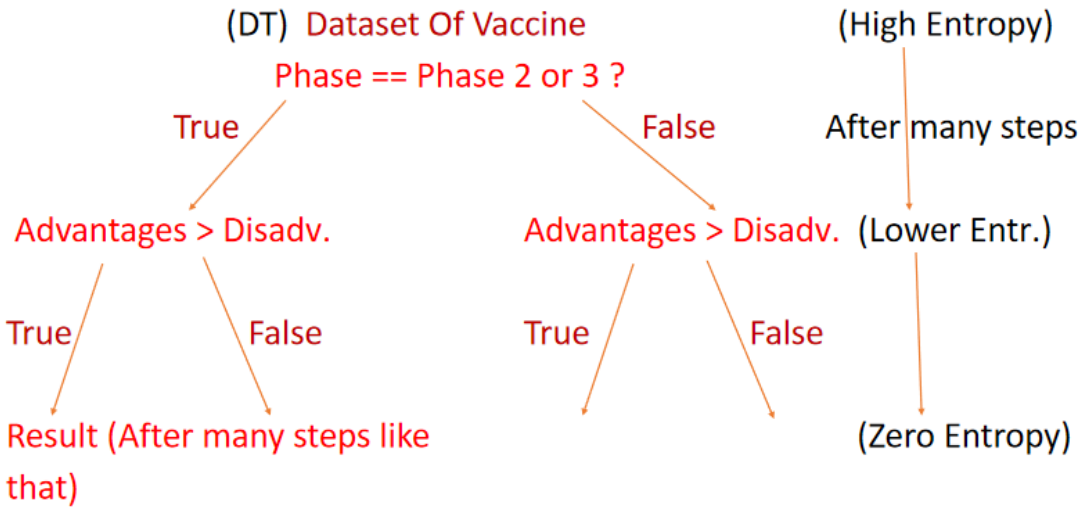
Phase IV: Choosing Suitable Machine

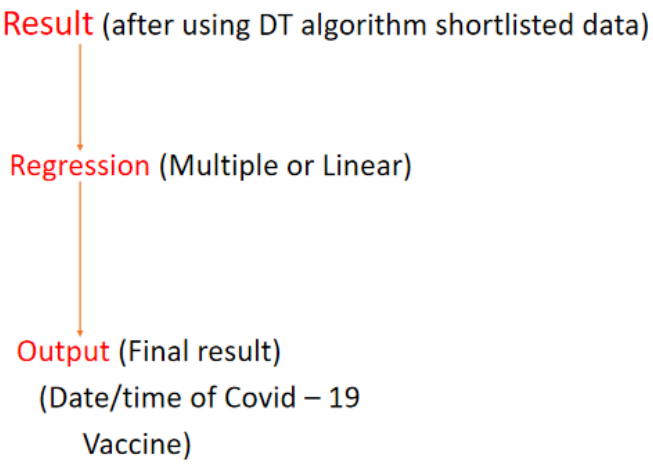
Learning Algorithm

- Well you can see that our dataset look quite messy and the entropy (Entropy is the measure of randomness and unpredictability in the dataset) is high in this case (phases, Candidate Name, Current Phase, Technology , Trial Number , advantage count , etc).

- First we classifying data , short listing it through Decision Tree (DT) Algorithm then use Multiple Regression (ML) Algorithm to Predict the date/time of releasing COVID-19 Vaccine.

- Classification (Give categorical solution yes or no , 0 or 1) >> Naïve Bayes , Logistic Regression (these two use when we have less calculation) , Decision Tress (when you have more of data attributes and you what to short) , Random Forest (DT is a form of RF when you have a lot of data then we use Random Forest) >> Regression (Use when continuous value need to predict like ‘date’ , ’profit’ , etc).



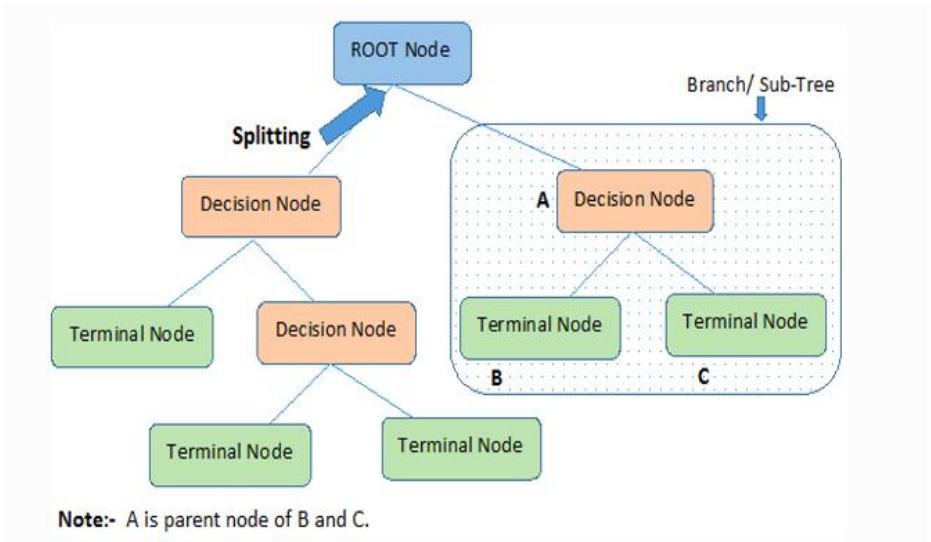


Short Information about Decision Tree (DT)

-> Belongs to the family of supervised learning algorithm.
-> Goal is to create a training model that can used to predict the class or value of the target.

Important Terminology related to Decision Trees

Root Node: It represents the entire population or sample and this further gets divided into two or more homogeneous sets.
Splitting: It is a process of dividing a node into two or more sub-nodes.
Decision Node: When a sub-node splits into further sub-nodes, then it is called the decision node.
Leaf / Terminal Node: Nodes do not split is called Leaf or Terminal node.
Pruning: When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.
Branch / Sub-Tree: A subsection of the entire tree is called branch or sub-tree.
Parent and Child Node: A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.



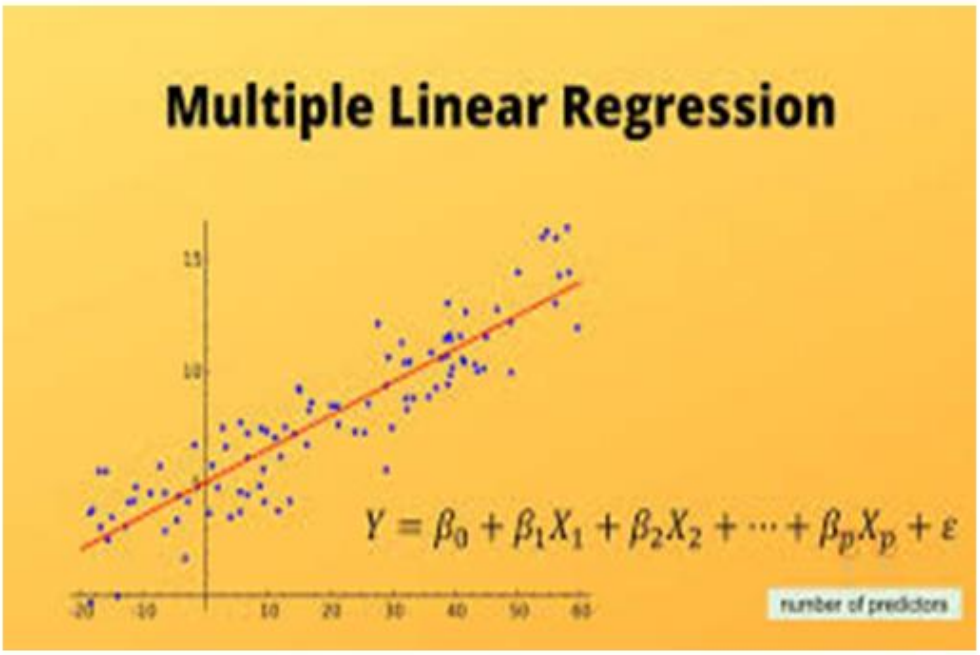
Assumptions while creating Decision Tree

(1) In the beginning, the whole training set is considered as the root.
(2) Feature values are preferred to be categorical. If the values are continuous the they are discretized prior to building the model.
(3) Records are distributed recursively on the basis of attribute values.
(4) Order to placing attributes as root or internal node of the tree is done by using some statistical approach.
Decision Trees follow Sum of Product (SOP) representation. For a class, every branch from the root of the tree to a leaf node having the same class is conjunction (product) of values, different branches ending in that class form a disjunction (sum).

Short Information about Multiple Regression (MR)

Multiple linear regression is a statistical method of predicting or explaining a continuous variable (sometimes called the dependent variable) as a linear combination of one or more variables (sometimes called independent variables).
The model is of the form
$Y = \sum bi\ xi+e$

It makes several assumptions:
Independent errors (e)
Linearity - that is, that the model is linear in its parameters (the B)
Additivity of terms (although you can add interaction terms)
Errors have mean 0 and constant variance across all levels of x
normality of errors (necessary only for some aspects of how the model is used)



Assumptions in Multi Regression

- Regression residuals must be normally distributed.
- A linear relationship is assumed between the dependent variable and the independent variables.
- Absence of multicollinearity is assumed in the model, meaning that the independent variables are not too highly correlated.
-
- Why Multiple Regression –
- This model is directly interpretable. It's easy to get ideas from a regression model about how to change strategies.
- Regression model is easily explainable.
- Regression models are computationally cheap and extremely fast. Especially when you're running data through a model millions (or more) times a day, these differences in speed add up to real dollars and cents.

Phase V: Building Our ML Model

Building our ML model

1. Deciding a best algorithm for your dataset is one of the tough work to predicting a value.
2. For that we have different type in ML to solve that kind of problems.
3. Firstly we see our dataset attributes and matches to the properties of the ML algorithm's according to our goal and how to achieve that using these algorithm's.
4. Problem's in ML - (1). Classification (categorial solution 1 or 0 , True or False)
   (2). Regression (continuous Value need to predict)
   (3). Clustering (data need to organized)

5. In our case firstly we have to short the data because we can't handle to much to predict the data/time .Then from that shortlisted data we have to predict the date/time that's why we select that following steps.

Classification >> Decision Tree (DT) >> Shortlisted data >> Regression >> Multiple Regression (MR) >> Result (In the for of Date/time)

Decision Tree (DT) algorithm :-

```
# Importing the dataset
dataset = read.csv('dataset.csv')
dataset = dataset[2:3] # Splitting the dataset into the Training set and Test set
# install.packages('caTools')
# library(caTools)
# set.seed(123)
# split = sample.split(dataset$DependentVariable, SplitRatio = 2/3)
# training_set = subset(dataset, split == TRUE)
# test_set = subset(dataset, split == FALSE)
```

```
# Feature Scaling
# training_set = scale(training_set)
# test_set = scale(test_set)
# Fitting Decision Tree Regression to the dataset
# install.packages('rpart')
library(rpart)
regressor = rpart(formula = ourFormula, data = dataset , control =
rpart.control(minsplit = 1))
# Predicting a new result with Decision Tree Regression
y_pred = predict(regressor, data.frame(Level = level))
```

Multiple Regression (MR) algorithm :-

```
# Splitting the dataset into the Training set and Test set
 # install.packages('caTools')
# library(caTools)
# set.seed(123)
# split = sample.split(dataset$DependentVariable, SplitRatio = 2/3)
# training_set = subset(dataset, split == TRUE)
# test_set = subset(dataset, split == FALSE)
# Feature Scaling
# training_set = scale(training_set)
# test_set = scale(test_set)
# Fitting Multiple Linear Regression to the Training setregressor = lm(formula =
ourFormula, data = training_set)
# Predicting the Test set results
y_pred = predict(regressor, newdata = test_set)
```

Phase VI : Testing and Training of our ML

model

Training testing of the model

To develop a models on machine learning principles a training data is used that can help machines to read or recognize a certain kind of data available in various formats like texts, numbers and images or videos to predict as per the learned patterns.

Difference between testing and training

Training Data is kind of labeled data set or you can say annotated images used to train the artificial intelligence models or machine learning algorithms to make it learn from such data sets and increase the accuracy while predating the results. (Training)

While on the other hand, after using the training data sets each machine learning model needs to be tested to check the accuracy and validate the model prediction. Testing data is quite different from training data, as it is a kind of sample of data used for an unbiased evaluation of a final model fit on the training dataset to check model functioning. (Testing)

Why Training data is Important

Training data is important because without such data a machine cannot learn anything and if you want to train model you have to feed the curated data sets allowing machines learn from the repetitive or differentiated patterns and predict accordingly.

As much as quality training data is feed into the AI model or ML algorithms with the right algorithm you will get the more accurate results. The accuracy of model prediction mainly depends on the quality and quantity of training data sets used to train such models.

How we use this In practical whey

```
set.seed(123)
split = sample.split(dataset$DependentVariable, SplitRatio = 0.8)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)
```

Randomly selecting the data

Divide the data into training and testing

we have to work on training set and it splited

Phase VII: Testing and approval of ML model

Approval/Disapproval of ML model and further modification if required

Well till now we can't read or analysis our dataset so we can't say that the path we have selected is perfect.

Our path what we have decided to go throw it is like :-

Classification (Give categorical solution yes or no , 0 or 1) >> Decision Tress (when you have more of data attributes and you what to short) >> Shortlisted data >> Regression (Use when continuous value need to predict like 'date' , 'profit' , etc) >> Multiple Regression (ML) >> Output (Date/time)

May be the path can be go like that :-

Regression >> Multiple Regression (ML) >> Output (predicted date/time)

Approval/Disapproval of ML model and further modification if required

Well till now we can't read or analysis our dataset so we can't say that the path we have selected is perfect.

Our path what we have decided to go throw it is like :-

Classification (Give categorical solution yes or no , 0 or 1) >> Decision Tress (when you have more of data attributes and you what to short) >> Shortlisted data >> Regression (Use when continuous value need to predict like 'date' , 'profit' , etc) >> Multiple Regression (ML) >> Output (Date/time)

May be the path can be go like that :-

Regression >> Multiple Regression (ML) >> Output (predicted date/time)

Approval/Disapproval of ML model and further modification if required

Well till now we can't read or analysis our dataset so we can't say that the path we have selected is perfect.

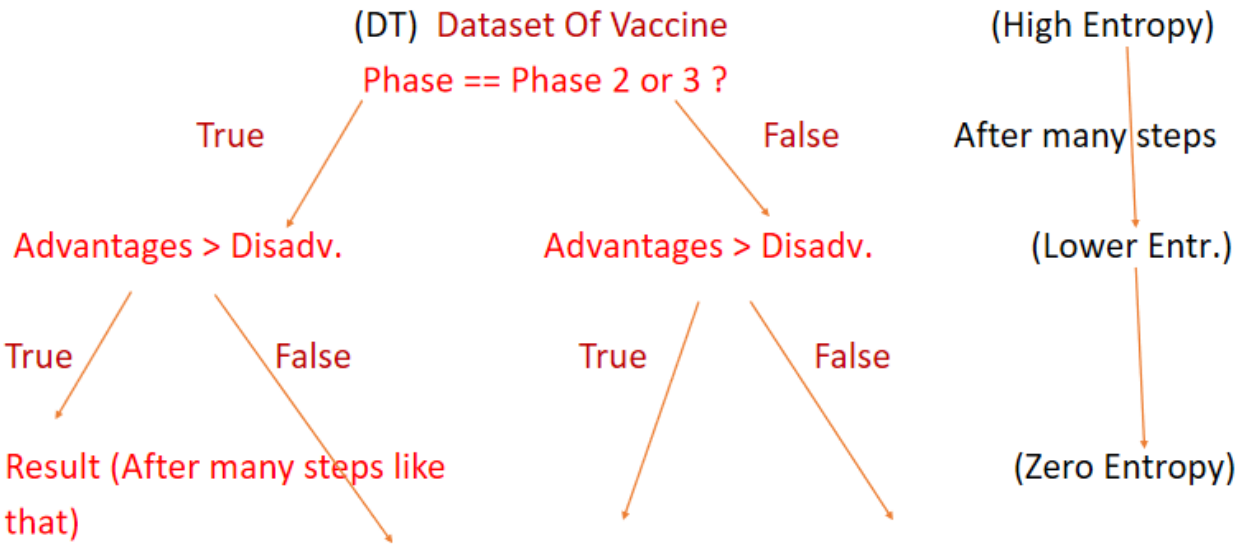
Our path what we have decided to go throw it is like :-

Classification (Give categorical solution yes or no , 0 or 1) >> Decision Tress (when you have more of data attributes and you what to short) >> Shortlisted data >> Regression (Use when continuous value need to predict like 'date' , 'profit' , etc) >> Multiple Regression (ML) >> Output (Date/time)

May be the path can be go like that :-

Regression >> Multiple Regression (ML) >> Output (predicted date/time)

Flow Chart of Our Plane



Result (after using DT algorithm shortlisted data)

Regression (Multiple or Linear)

Output (Final result)
(Date/time of Covid – 19 Vaccine)

Flow Chart Of 2nd Priority

Regression

Multiple Regression

Output (Final result)
(Date/time of Covid – 19 Vaccine)

Phase VIII : Building R shiny web app to

display our analysis

Shiny is an R package that makes it easy to build interactive web apps straight from R. You can host standalone apps on a webpage or embed them in R Markdown documents or build dashboards. You can also extend your Shiny apps with CSS themes, htmlwidgets, and JavaScript actions.

Shiny combines the computational power of R with the interactivity of the modern web.

We will be using it to built an interactive web app to display our analysis.