

الجامعة الإسلامية العالمية ماليزيا
INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA
يُونُزُ بَرِيَّتِيْ اِسْلَامُ اَنْتَا اَبْغُسَا مِلْدِسِيَا

Garden of Knowledge and Virtue

DATA SCIENCE
PROJECT PROPOSAL
SEMESTER 1, 2020/2021
CSC 3305, SECTION: 01

LECTURER: DR. RAINI HASSAN

Prepared by:

Name	Matric number
TASNIM RAFIA	1725826
HASAN TANVEER MAHMOOD	1725413
BILLAH SYED MASHKUR	1723387
K. M. ZUBAIR	1722931

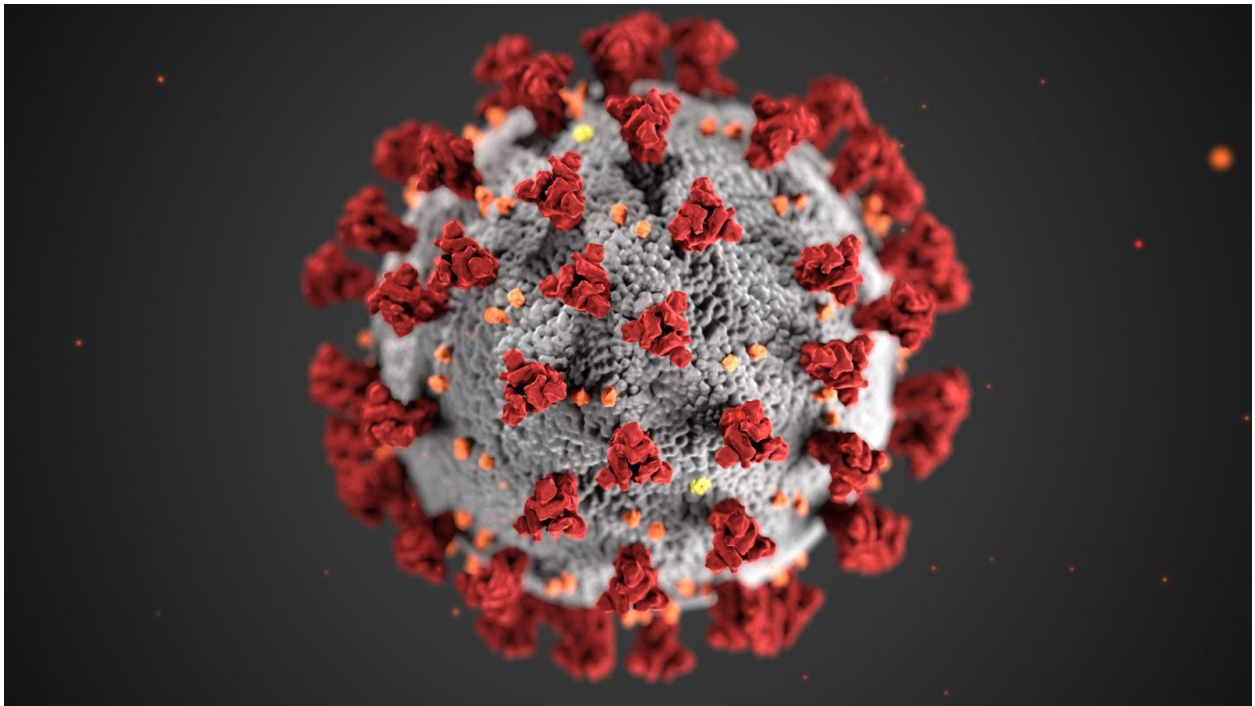
Table Of Content

<u>Task:</u>	<u>Content</u>	<u>Page no:</u>
1	Title	2
2	Introduction	2
3	The Background	2
4	The Problem Statement	3
5	The Research Question and/or Hypothesis	3
6	The Research Objectives	4
7	The Research Significances	4
8	The Relevant Works/Literature Review	4 - 7
9	Methodology	8 - 9
10	References	9 - 10

1. Title: Predicting COVID 19 Cases.

2. Introduction:

Covid-19 was first discovered in the human body on 31 December 2019 in Wuhan city, China. But this is the first time Covid-19 was detected in the human body. In the past, it was discovered in animals' bodies for instance, bats, cats, chicken etc. It is not the same disease as pneumonia, flu, or any other cold fever. Its reproduction is very strong and during the reproduction process in the human body it can change the RNA. So, it is difficult to invent a new vaccine which can destroy the chain of reproduction. But our scientist is working hard to invent the vaccine. For the daily cases and updates visit: [Covid-19](#)



3. Background:

Covid-19, the cause of death of more than one million people, so far, is the most alarmed fact these days. It has become the origin of psychological disruption that is extremely terrifying. Getting rid of the pandemic has become the most widely known desire. To predict the period of discarding these dreadful days, the prediction of the number of patients in next days is genuinely a primary mission. The projection of this project is to forecast the number of a discrete amount of days that will enable to anticipate the pandemic session.

4. Problem Overview:

Coronavirus (COVID-19) is an illness caused by a virus that can spread from person to person. As indicated by the Wuhan Municipal Health Commission, tests from China's Wuhan's Huanan Seafood Wholesale Market were positive for covid-19. Cases demonstrated side effects, for example, fever, dry hack, dyspnoea; radiological discoveries indicated reciprocal lung infiltrates. On 9 January 2020, the China CDC revealed that a novel coronavirus (later named SARS-CoV-2, the infection causing COVID-19) had been distinguished as the causative specialist for 15 of the 59 instances of pneumonia. On 10 January 2020, the main novel coronavirus genome arrangement was made openly accessible. By 20 January 2020, there were reports of affirmed cases from three nations outside China: Thailand, Japan, and South Korea. These cases had all been sent out from China.

On 5 June 2020, WHO published an updated guidance on the use of masks in communities, during home care, and in health care settings in areas that have reported cases of COVID-19. This guidance is aligned with ECDC opinion published on 8 April 2020. On 13 June 2020, officials from the National Health Commission and Beijing Health Commission reported a new cluster of COVID-19 cases in Beijing, People's Republic of China. On 17 June 2020, the European Commission introduced a European methodology to quicken the turn of events, assembling and organizing antibodies against COVID-19. On 2 July 2020, ECDC distributed a fast hazard appraisal on resurgence of instances of coronavirus illness 2020 (COVID-19) in the EU/EEA, the UK and EU competitor and likely applicants, featuring that while diminishing or stable patterns in malady occurrence are being seen in Europe by and large, there is still network transmission detailed, and a few nations are revealing a resurgence of cases or huge limited episodes which is another update that the pandemic is a long way from being done.

So, we decided to utilize the COVID-19 infection as a model to our project. We accept our displaying beginning from around March until July 2020. We have taken four nations in record to do some investigation and computations which are the United States of America. Our information regularly incorporates dates of the first case, at that point how it builds, number of all out dynamic cases, number of new cases, number of recovered and also some data of dead ones among these countries.

5. The Research Question and/or Hypothesis:

The number of Covid-19 patients is evidently detectable to be climbing up most recently. It is assumed based on the number of cases of the current year that next year is going to be bringing about more cases.

6. The Research Objectives:

- Using Data Pre-processing in order to reform the covid-19 dataset to ease the research in the next steps.
- Understanding the data on covid-19 cases for data visualization.
- Extracting out the features useful in order to predict covid-19 cases in the upcoming days.
- Building up a model to point out the risk factors in order to predict covid-19 cases on the upcoming days.
- Finally, testing the model for any errors involved for better prediction and checking out its performance.

7. The Research Significances:

- Predicting the possibility of new covid-19 cases in the upcoming days would allow us to be prepared in order to take possible safety needed.
- Pointing out the risk factors would highly assist to predict covid-19 cases on the next day that could allow the governments of the countries to take necessary steps.
- Predicting new covid-19 cases, could also help doctors at the hospitals to assume approximately how many patients could be admitted and take preparations accordingly.
- Prediction of covid-19 cases would also assist to point out the safe zones or locations to travel to.

8. The Relevant Works/Literature Review:

As the Covid-19 is the most popular concern of this year, we could find some literature that is somewhat relevant to our topic.

No	Year	Authors	Research problem/ Application	Main Techniques applied	Results	Future work
1	2020	Gitanjali R., Asmita B., Parikshit N., Nilanjan Dey, Jyotismita Chaki & Aboul Ella Hassanien	Forecasting models for Covid -19 to control the global outreach of pandemic.	Researchers have used the best Machine Learning models for prediction and used natural progression of disease	The Mathematical models inferred 44% of the transmission happened without showing any prior symptoms.	These models and predictions will help to prepare against probable tragedies and consequences in other

				for better accuracy. Big Data Analysis and Deep Learning is used to analyze big data from WHO or other national resources. Used Population Statistics to view data.	The Data Science models forecasted suspected members of Covid-19, death counts, Impact on Public Health, Identify individuals who are at paramount risk and predicted the infection rate in China. Population Statistics helped to make reliable predictions having less computational overhead.	similar pandemics in the future.
2	2020	Akib Mohi, Syed Tanzeel Rabani, Qamar Rayees Khan, Nusrat Rouf, Masarat Mohi	Machine learning based approaches to detect COVID-19 using clinical text data	The authors used machine learning algorithms like support vector machine (SVM), multinomial Naïve bayes (MNB), Logistic	Logistic regression and Multinomial Naïve Bayes has 94% precision, 96% recall, 95% f1 score and a testing accuracy of 96.2% and algorithms like	Recurrent neural network can be used for better accuracy and the authors believe that the more data they get the more accurate their models will

				regression, Decision tree, Random Forest, Bagging, Adaboost and Stochastic gradient boosting for classification to complete this research.	random forest and gradient boosting have 94.3% accuracy.	become. They also tend to use Deep learning approach in the future.
3	2020	Dasari Naga Vinod, S.R.S. Prabaharan	Using Data Science to achieve fast diagnosis of Covid-19	Using AI with Chest X-Ray and CT scan to identify the Covid-19. Using Decision tree classifier for faster prediction. Utilizing X-ray pictures and CT scan images with deep learning strategies.	The recommended approach achieved an accuracy of 93% in CT scan images and produced 88% precision score for chest x-ray images and also provides faster results during training and inference.	This research can be used to detect other diseases or infections that are related to lungs.
4	2020	Zohair Malki, El-Sayed Atlam, Aboul Ella Hassanien,	Using weather data with Machine learning to predict	The authors used machine learning models like	It is experimentally proved that the weather variables,	The authors are planning to consider additional features

		Guesh Dagne, Mostafa A. Elhosseini, Ibrahim Gad	mortality rate.	Linear Regression, Lasso Regression, Ridge Regression, Elastic Net, Least Angle Regression, Lasso Least Angle Regression, Orthogonal Matching Pursuit, Bayesian Ridge, Automatic Relevance Determination, Passive Aggressive Regressor, Random Sample Consensus, TheilSen Regressor, Huber Regressor etc to understand the effect of temperature and humidity on the spread of Covid-19.	temperature and humidity have an inverse relationship to the number for confirmed cases. The ML model estimates the relationship based on the number of confirmed cases in Italy as of May 17, 2020. For every one-unit increase in the independent variable temperature, the number of COVID-19 cases increases by about 143%.	like wind speed and rainfall to improve their model's accuracy. Essential measures can be taken by predicting mortality rate on similar pandemics in the future.
--	--	---	-----------------	--	---	--

The research mentioned on the table uses different techniques for predicting Covid-19 outbreaks. Most commonly used technique is machine learning, and the model is regression. As Logistic regression and Multinomial Naive Bayes have a higher accuracy, we're planning to use these algorithms for our project.

9. Methodology:

In this project we are trying to do our analysis by using real data sets. So, we collect our data from different online dataset sources. Moreover, this data set can give more or less accuracy to analyze the real outcomes. To avoid less accuracy, we endeavour to collect more data.

9.1 Dataset

we have collected our data from: [U.S.A \(Covid- 19\)](#)

9.2 Tools

R, Rstudio and Python.

9.3 Machine Learning Algorithm

We are going to use Naive Bayes. Naive Bayes algorithm is based on the Bayes Theorem, which is used to calculate conditional probability.

$$P(A|B) = (P(B|A) P(A))/P(B)$$

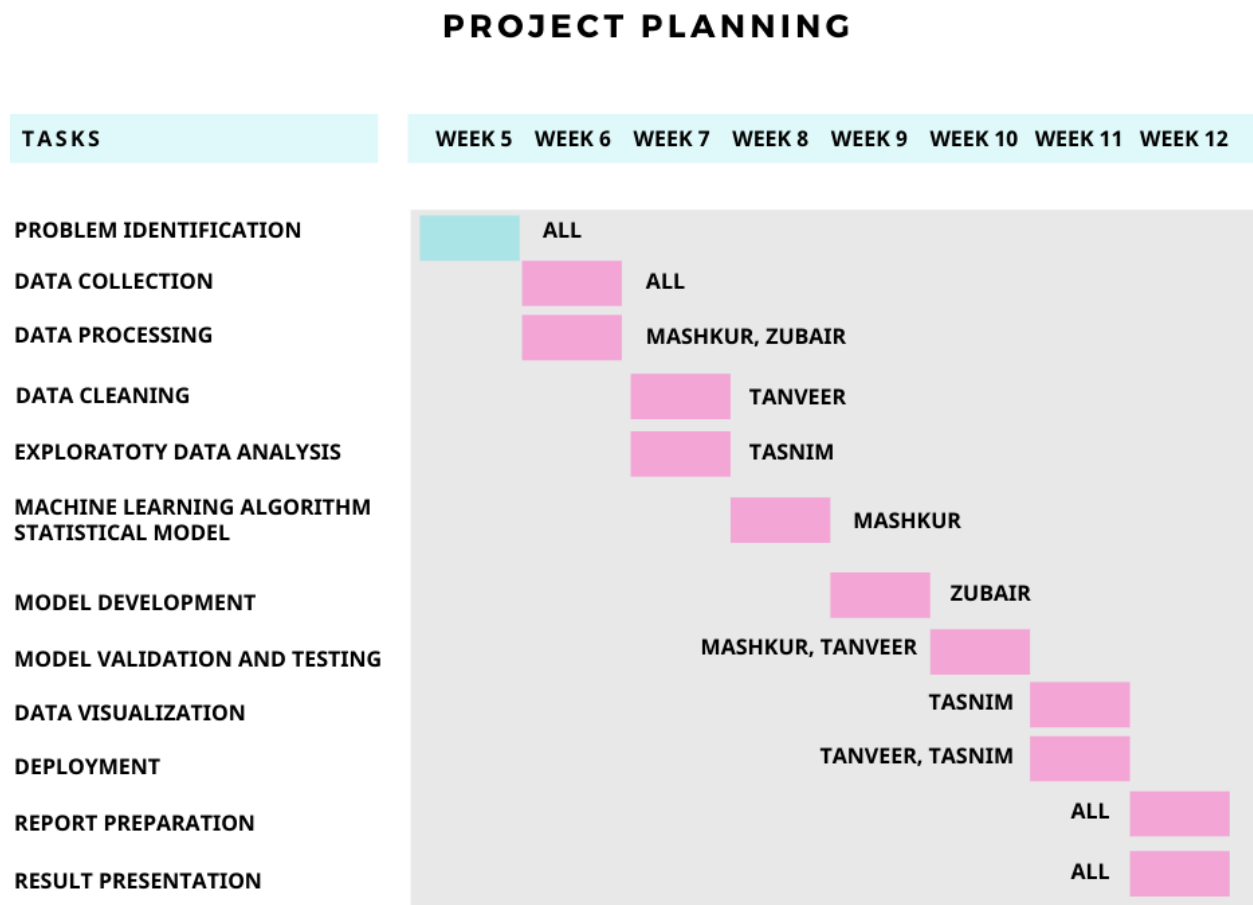
Formally, the terminologies of the Bayesian Theorem are as follows:

1. A is known as the proposition and B is the evidence
2. $P(A)$ represents the prior probability of the proposition
3. $P(B)$ represents the prior probability of evidence
4. $P(A|B)$ is called the posterior
5. $P(B|A)$ is the likelihood

Training and testing Datasets

80% of the [U.S.A \(Covid- 19\)](#) Dataset will be used for training and 20% for validation.

9.4 Project Plan:



10. References

1. Malki, Z., Atlam, E., Hassanien, A., Dagnew, G., Elhosseini, M., & Gad, I. (2020, July 17). Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. Retrieved November 09, 2020, from <https://www.sciencedirect.com/science/article/pii/S0960077920305336>
2. R., G., B., A., N, P., & Dey, N. (2020). Forecasting Models for Coronavirus Disease. Retrieved November 09, 2020, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7289234/>
3. F. Wu, S., N. Chen, M., A. Kumar, V., P. Verma, A., S. Chakraborti, A., A. Sarwar, M., . . . Friedman, J. (1970, January 01). Machine learning based approaches for detecting COVID-19 using clinical text data. Retrieved November 09, 2020, from <https://link.springer.com/article/10.1007/s41870-020-00495-9>
4. Vinod, D., & Prabakaran, S. (2020, July 30). Data science and the role of Artificial Intelligence in achieving the fast diagnosis of Covid-19. Retrieved November 09, 2020, from <https://www.sciencedirect.com/science/article/pii/S0960077920305786>

5. Boudrioua MS, Boudrioua A. Predicting the COVID-19 epidemic in Algeria using the SIR model. medRxiv. 2020;p. 2020.04.25.20079467. Available from: <http://medrxiv.org/lookup/doi/10.1101/2020.04.25.20079467>.
6. You C, Deng Y, Hu W, Sun J, Lin Q, Zhou F, et al. Estimation of the time-varying reproduction number of COVID-19 outbreak in China. International Journal of Hygiene and Environmental Health. 2020 feb;228:113555. Available from: <https://www.sciencedirect.com/science/article/pii/S1438463920302133>
7. Hogan A. How much of the coronavirus does it take to make you sick? Stat News; 2020. [Access date: 15/07/2020]. Available from: <https://www.statnews.com/2020/04/14/how-much-of-the-coronavirus-does-it-take-to-make-you-sick/>
8. https://www.google.com/search?q=covid+19+case+number+malaysia&rlz=1C1CHZL_enMY836MY836&oq=covid+19+case+number+&aqs=chrome.1.69i57j0i20i263i457j0l6.13544j0j7&sourceid=chrome&ie=UTF-8#wptab=s:H4sIAAAAAAAAAAONgVuLVT9c3NMwySk6OL8zJecTYxMgt8PLHPWGpyklrTl5jLOYS901NyUzOzEt1ySxOTSxO9clPTizJzM8T0uNic80rySypFFLhEpRCNUeDQYqfC1VISIOLA65XhotXilM_V9_AsiA3uwyompsLweXZxcSRX5ZaVJaZW6r6IVcQ5vyg_L7Ess6i0WME3MSexsjgzEQCmPmiMuQAAAA
9. <https://www.cebm.net/covid-19/coronaviruses-a-general-introduction/>