**PROJECT PHASE-I REPORT**

**ON**

**ANALYSIS AND PREDICTION OF EFFECT OF VIRAL DISEASE IN HUMAN + DEPLOYMENT**

*A report submitted in partial fulfilment of the requirement for the award of*

*The degree of*

**BACHELOR OF TECHNOLOGY**

**In**

**COMPUTER SCIENCE AND ENGINEERING**

A picture containing drawing

Description automatically generated

**Submitted to:**

Manish Sharma

IBM Trainer

**Submitted by:**

Ayan Bhatnagar- 190109900 - BDA

Kartikey Garg-180109045 - BDA

Pradyumn Bahukhandi-180109025 - BDA

Yash Sharma-180109032 - BDA

**SCHOOL OF COMPUTING**

**DIT UNIVERSITY, DEHRADUN**

(State Private University through State Legislature Act No. 10 of 2013 of Uttarakhand and approved by UGC)

**Mussoorie Diversion Road, Dehradun, Uttarakhand - 248009, India.**

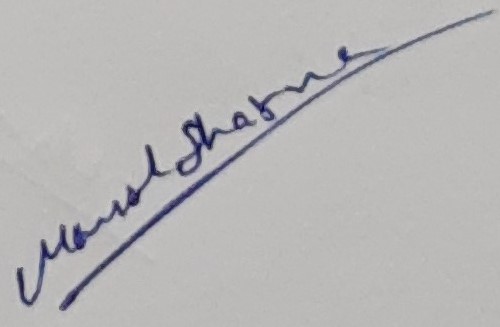
**2020**

**CANDIDATES DECLARATION**

I hereby certify that the work, which is being presented in the Report, entitled **Analysis and Prediction of effect of Viral Diseases in Human + Deployment**, in partial fulfilment of the requirement for the award of the Degree of **Bachelor of Technology** and submitted to the DIT University is an authentic record of my work carried out during the period ***17-August-2020***to ***30-November-2020***under the guidance of **Manish Sharma Sir**.

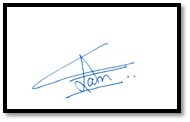
**Date: 22/10/2020**

**Signature of the Instructor**

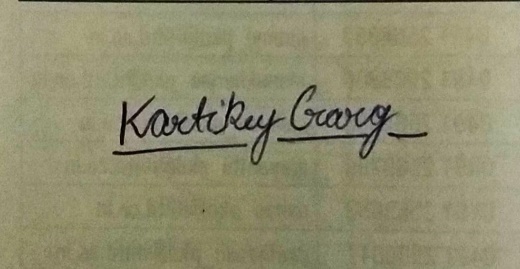
****

**(Manish Sharma IBM Trainer)**

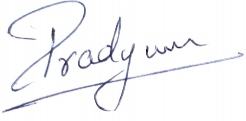
**Signature of the Candidates**



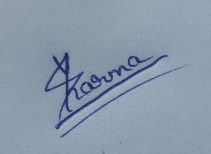
**(Ayan Bhatnagar)**

****

**(Kartikey Garg)**

****

**(Pradyumn Bahukhandi)**

****

**(Yash Sharma)**

**ACKNOWLEDGEMENT**

It is a great opportunity for us to present this report on the partial completion of University Group Project. We express our heartiest thanks to all the faculty members engaged directly or indirectly in guiding the making of this project. We would also like to thank **Manish Sharma Sir** who supervised and assisted us at every moment. We wish to utilize this valuable Project and the professional skills which we have learned in Development and Deployment of this Project in our Professional Future.

**Name: Ayan Bhatnagar**

**Roll No.: 190109900**

**SAP ID: 1000013728**

**ABSTRACT**

This project will describe different features of virus, disease and also presents different visualization for the spread of the infection, and also discusses the potential applications of data analytics components on this viral infection . Firstly, a literature survey and gathering of dataset is done on viruses highlighting a number of factors including its origin, its similarity with previous disease, its transmission capacity, its symptoms, etc. Secondly, data analytics is applied on a dataset of respectiveto data collection and find out the spread of the viral infection, the worldwide increase in the number of confirmed cases over time is modelled. Thirdly the application deployment which will deduce our project on a platform and it will work on the real time data.

**TABLE OF CONTENT**

CHAPTER PAGE No.

Candidate’s Declaration 2

Certificate 3

Acknowledgement 4

Abstract 5

Chapter 1 – Introduction 8

Chapter 2 –Project Description 9

2.1. Purpose 9

2.2. Motivation 9

2.3. Problem statement 9

2.4. Project Perspective 9

2.5 System Requirements 10

Chapter 3 –Literature Review 11

Chapter 4- Methods and Materials 12

Chapter 5 – Conclusion & Road map of Phase-2 23

Bibliography 30

Annexures: 34

Plagiarism report checked and signed by Manish Sir

**List of Figures**

FIGURE NAME PAGE

Figure 1: Showing what all data columns are there 17

Figure 2: Metadata of Datasets 19

Figure 3: Data types in Dataset 19

Figure 4: Head command representing the first 5 value of dataset 20

Figure 5: representation of head plus tail thing 20

Figure 6: Representing no Missing Values 21

Figure 7: Distribution of people based on age and region suffering from HIV 21

Figure 8: IBM Cognos 26

Figure 9: Python NumPy 27

Figure 10: Python Pandas 27

Figure 11: Python matplotlib 28

Figure 12: Python Tkinter 28

Figure 13: wxPython 29

Figure 14: Python Qt 29

Figure 15: GTK 29

**List of Tables**

**NO TABLES**

**Chapter 1 -- Introduction**

The aim of the project is to provide details to each respective user about how infectious a virus on human body with certain pre assumed characteristics like an unhealthy body or under an age group that will help in predicting for future occurrences and consequences.

Further the reason that led us to choose this topic for Project is not only because its trending but also with the help of this we can provide a detailed analysis of further and ongoing diseases that can ultimately make research and knowledge better for the user as well as the application.

Detection of highly divergent or yet unknown viruses from datasets is a challenging task. When human samples are sequenced, a large proportion of assembled conventional methods find no similarity to known sequences. We wished to explore whether machine learning algorithms could improve the detection of viral sequences in metagenomic sequencing data by performing research on dataset of viruses like Plague, Influenza, Dengue, Ebola, Hantavirus, COVID-19 etc.

**Chapter 2 –Project Description**

**2.1. Purpose**

* Analysis on this dataset can help us to understand human symptoms.
* It helps us figuring out the future scope of viral diseases and their symptoms too.
* Not only this for an accuracy of 80% this limit bounded analysis can be helpful for the further future researches wherein they would have future prospects like AI in health and neurology.

**2.2. Motivation**

Looking at current Scenarios we were eager to work on a project that have several future used. Working on this project after successfully completion this project will help in developing further research and development.

**2.3. Problem statement**

Fighting with the diseases is necessary for human Survival and Development, so a platform must be there that helps human analysing future scope of diseases.

**2.4. Project Perspective**

This project deals with the understanding of diseases the cause harmful effect on human civilization. It aims to visualize the future scope and present scope of viruses on humans.

**2.5 System Requirements**

**Hardware Requirement:**

1. Intel Core i5 or i7
2. RAM 8-16 GB
3. Laptop/Desktop

**Software requirement:**

1. IBM Cognos
2. Anaconda and its components
3. Datasets
4. MS Office
5. GitHub.

**Other Requirement:**

1. Good Internet Connection

**Chapter 3 –Literature Review**

The current status of Technology utilization for Medical analysis is having a vast usage. Researchers form all over the globe are interested in developing, deploying and illustrating the use of technology in HUMAN. Moreover, in this era 3 famous terms “Artificial Intelligence, Deep Learning, Machine Learning” are boosted for Analysis motivation in researchers. We do have Several Regression models, Intelligence based machine models, functions the deals a handsome amount of help in researching on the field of Medical Diagnosis. The relation pf health care with technology begins with the evolution of then first expert system called MYCIN developed in 1976. MYCIN was generally made to use 450 rules which are collected from a medical expert in order to treat bacterial infection by suggesting good antibiotics to the patients ML and AI are also used to augment diagnosis and process of screening of the identified patients with radio imaging technology akin. Studies show the potential of AI and ML tools by suggesting a new model that comes with rapid and valid methods for Diagnosis using Deep Convolutional Networks. Recent studies design an auxiliary tool to increase the accuracy of Diagnosing Algorithms with new automatic disease detection-based algorithm.

**Chapter 4- Methods and Materials**

The methods and materials required to make this project are described in this section.

**3.1 LOGISTIC REGRESSION**

Logistic regression is a technique inherited by Machine Learning from the field of statistics. Classification problems can be solved using this method. Logistic regression and linear regression are very similar to each other as they attempt to predict a response variable Y given a set of X input. It is a form of supervised learning, which tries to predict the responses of unlabeled, unseen data by first training with labelled data, a set of observations of both independent(X) and dependent (Y) variables. It is a technique used to model and analyze the relationship between variables and often times how they contribute and are related to producing a particular outcome together.

The probability that Y, the response variable belongs to a certain category will be modelled by logistic regression. The response variable in many cases will be binary one, so logistic regression will want to model a function y=f(x) that outputs a normalized value that ranges from say 0 to 1 for almost all values of X corresponding to two possible Y values . This is done using the Logistic Function.

Types of Logistic Regression are-

**1. Binary Logistic Regression-** It is the categorical response has only two possible outcomes. Example: Spam or not.

**2. Multinomial Logistic Regression-** It has three or more categories without ordering. Example: Predicting which food is preferred more.

**3. Ordinal Logistic Regression-** It has three or more ordering with ordering. Example: Movie rating from 1 to 5.

**Logistic Function-**

Sigmoid Function is another name used for Logistic Function as shown in equation 2. The Shape is S- curve and it can have any real-valued number and then map it into a value ranging between 0 and 1, never exact between those limits.

1/(1+𝑒^−𝑣𝑎𝑙𝑢𝑒) (2) Where, e = base of natural logarithms and, value = actual numerical value that we want to transform.

**Predictions using Logistic Regressions:**

Input values(x) are combined linearly using weights or coefficients values which further predict an output value(y). A key difference from linear regression is that the output value being modelled is binary values (0 or 1) rather than a numeric value. Below is an example of logistic regression as shown in equation 3:

𝑦 = 𝑒^(𝑏0 + 𝑏1∗𝑥) / (1 + 𝑒^(𝑏0 + 𝑏1∗𝑥)) (3)

Where, y is predicted output, b0 is bias term and b1 is coefficient for single unit value(x). In each column of input data has connected b coefficient that is understandable from training data. **Applications of Logistic Regression:**

1. To classify mail as spam or not.

2. It helps us to determine the presence or absence of certain diseases like cancer based on symptoms and other medical data.

3. We can classify images based on pixel data.

**3.2. SUPPORT VECTOR MACHINE (SVM)**

SVM is a approach which is machine learning , and is used for regression and classification problems. It is capable of taking continuous as well as categorical data. In SVM algorithm support-vector clustering algorithm was developed to categorize unlabeled data because when data is unlabeled, supervised learning is not possible and an unsupervised learning approach is required which attempts to find natural clustering of data to groups. It is the most used clustering algorithms present in industrial applications.

**Maximal-Margin Classifier:**

SVM represents, data points are represented on space of hyper plane, can be categorized into two groups. A hyper plane is a kind line that divide the input variable space. To separate the points in the input variable space by their class, either class 0 or class 1 a hyper plane is selected. Points which are similar come under same group. In SVM with linear kernel the dataset is represented as p-dimensional vector that can be separated by p -1 planes called as hyper-planes. These planes are used to set the boundaries between data groups. According to the boundaries between two classes, suitable hyper-plane is selected. We can assume all our input points can be completely separated by a line.

Example is shown in equation 4:

𝐵0 + (𝐵1∗𝑋1) + (𝐵2∗𝑋2) = 0 (4)

Where, coefficients B1 and B2 determine the slope of the line and the intercept (B0) is found by the learning algorithm, and X1 and X2 are the two input variables.

**Soft Margin Classifier:**

In real world, hyper plane cannot perfectly separate real data that is messy. Soft Margin Classifier is the limitation of making the margin max of the line that separates the classes must be relaxed. It allows some points in the training data to violate the splitting lines.

Additional coefficients set are introduced that give margin room wiggle in each dimension. These variables are called slack variables which increases the complexity of model as there are more parameters for the model to fit to the data to provide complexity.

The amount of violation of the margins allowed is defined by C parameters that determine the magnitude of the twitch allowed across all dimensions. The C=0 is no violation which means that we are back to the inflexible Maximal - Margin Classifier. The more value of C means more violations of the hyper plane which are permitted.

**Kernels of SVM:**

A kernel is used to implement SVM algorithm. The learning of the hyper plane in linear SVM is done by transforming the problem using some linear algebra The inner product between any two vectors is the multiplication sum of each input values pair.

The equation which is used for making prediction involving a new input using dot product between input(x) and each support vector (xi) is calculated as shown in equation 5:

f(x)= B0 + sum(ai∗(x,xi)) (5)

This equation has involved calculating the inner products input vector(x) which is new with all support vectors in training data. The coefficients B0 and ai must be estimated from the training data by learning algorithm.

**Linear Kernel SVM**- The dot product called kernel is written in equation 6 :

𝐾(𝑥,𝑥𝑖)=𝑠𝑢𝑚(𝑥∗𝑥𝑖) (6)

It defines the similarity or distance measure between new data and the support vector. The dot product can be called as similarity measure which is used for linear SVM because the distance has a linear combination of inputs.

**Polynomial Kernel SVM-** Instead of dot product we can use polynomial kernel, it allows for curved lines in the input space, for example shown in equation 7:

K(x,xi)=1+sum(x∗xi)^d (7)

Where, degree of the polynomial must be specified. When d=1 this is the same as linear kernel.

**Radial Kernel SVM-** It is local and can be used to create complex regions inside the feature space, like closed polygons in two dimensional spaces. , for example shown in equation 7:

K(x,xi)=exp(−gamma∗sum((x−xi^2)) (8)

Where, gamma is a parameter that must be specified to learning algorithm. A good value for gamma is 0.1 by default where gamma is often 0<gamma<1.

**Data Preparation for SVM**

Some suggestions for how to prepare for training data when learning an SVM model.

l **Numerical Inputs-** SVM assumes that the inputs are numeric. If we have categorical inputs then we may need to convert them to binary dummy variables.

l **Binary Classification-** Basic SVM is described for binary classification problems. Although, extensions were developed for multi-class classification and regression.

**Properties:**

* Flexibility in choosing a similarity function.
* By Soft Margin Approach over fitting can be controlled.
* Feature selection.
* Large feature spaces are handled i.e. complexity is not dependent on the dimensionality feature space.
* Sparseness of solution is required when dealing with the large data sets

**3.3.1 Data Collection**

Data Collection is the process of measuring information and gathering. This information can then help us to evaluate answers and outcomes. The data can be collected in two ways to create a **Training Data Set.** They are:

1. **Primary Data Collection:** When the data is collected for the first time by person and is therefore original in nature. The training data set is then freshly created. This type of data collection takes a bit more time but generates quite accurate answers.

2. **Secondary Data Collection:** When the data is retrieved from other sources, i.e. already collected data is used to construct a data set. This type of data set is time saving but do not guarantee an accurate answer.

Data Collection is vital step towards any project, research or model as it guarantees an optimal solution or output. When the data is appropriately collected, then only we can hope to achieve the results we want.

**3.3.2 Data Description**

Describing and documenting data is very essential as it confirms that the person who will deal with the collected data is able to understand it. The researcher should be able to understand the data as well as the processes and analysis done to create it. We will evaluate mean, median, mode, standard

deviation etc. to summarize the data. The data we deal with is a large one, which makes it rather mandatory to summarize the respective data. For this “summary of data” we most commonly use measures of central tendency. This includes mean- averaging the values, median finding the middle value from data and mode- most frequently occurring data values. These values of central tendency help in determining the distribution of the data as well as to characterize the large data sets.

**3.3.3 Data Pre-processing**

Data Pre-Processing generally means converting data into a much more understandable format. Generally the data which we need to handle is raw data which is then preprocessed by – cleaning, instance selection, normalization, feature extraction and selection. Ultimate result of data preprocessing is the Training Set.

For our model, we preprocessed our data in two steps:

**1. Manual Pre-Processing –** The pre-processing which involved manual efforts are kept under this category. The following steps were done manually, without the aid of any programming language or such technologies. The steps can be listed down as follows:

**Deletion of un-required Columns**

All the text columns like Name, Gender, Age, Program, Year of Study etc. were removed from complete data-set shown in figure 6. These fields are not required for the construction of Training Data Set. Therefore, the unrequired data fields are removed.

df = pd.read\_csv(r'E:\Training Second Year\covid\_19\_india.csv')

df.head()

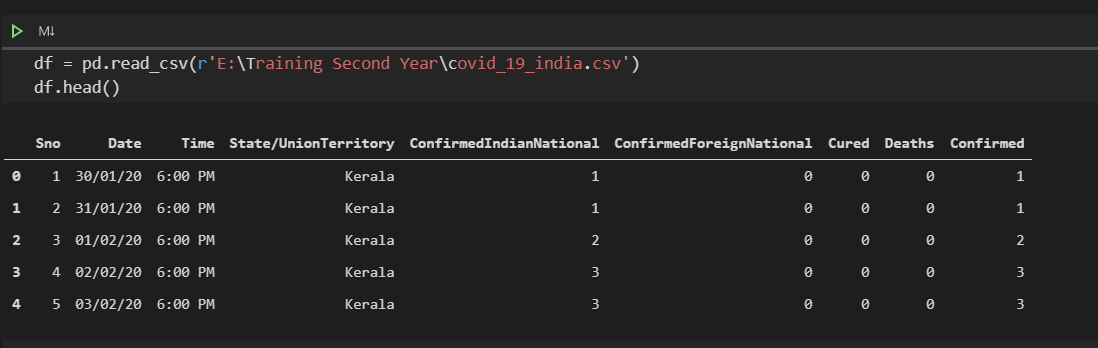


Figure : Showing what all data columns are there

**Pre-Processing through Python Programming –**

In this second step of preprocessing, we used **Python-** a programming language, work on our complete data set and preprocess it. The steps can be listed as follows:

**a) Import libraries**

In Python, various in-built libraries are available using which we can perform various operations.

For our model, we have imported the following libraries shown in figure 10- pandas, matplotlib and sklearn, preprocessing, Imputer.

• **Pandas** is used for various tasks such as :

- Creation of data frames and

- Navigation of data frames.

• **matplotlib** is used for plotting target variable in the form of a bar chart.

• **Imputer** is used for handling missing values in the dataset.

**b) Load Dataset**

• Before starting of our project we had loaded our dataset which was stored in the form of a csv file shown in figure 11.

• The pandas package was then used to convert CSV file to data frame using the *pandas.read\_csv()* function. This function returns a pandas. Data Frame.

**c) Verify Dataset**

Following commands shown in figure 12 and functions are used to verify the performance of the dataset values.

• **describe()** function is used to see some of the core statistics of the dataset, such as -

mean, min, max etc.

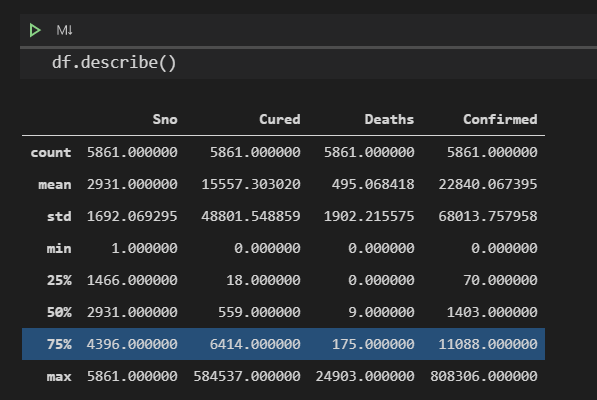


Figure : Metadata of Datasets

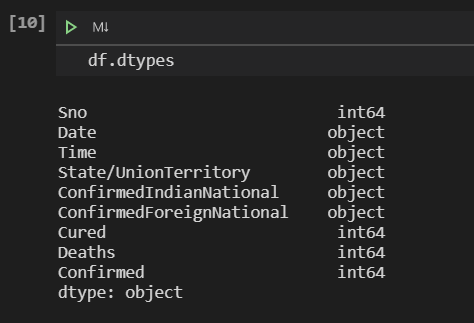


Figure : Data types in Dataset

• **shape** command is used to get dataset size. It returns a tuple with number of rows, and number of columns for the data in the Data Frame.

• **head()** function is used to show the top 5 rows of data in the Data Frame.

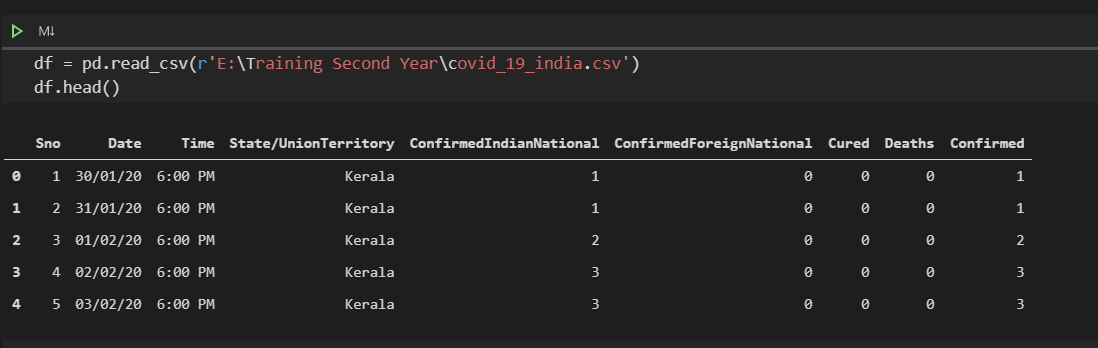


Figure : Head command representing the first 5 value of dataset

• **groupby()** function is used on Output variable of the dataset to find out total number of 0 and 1 category.

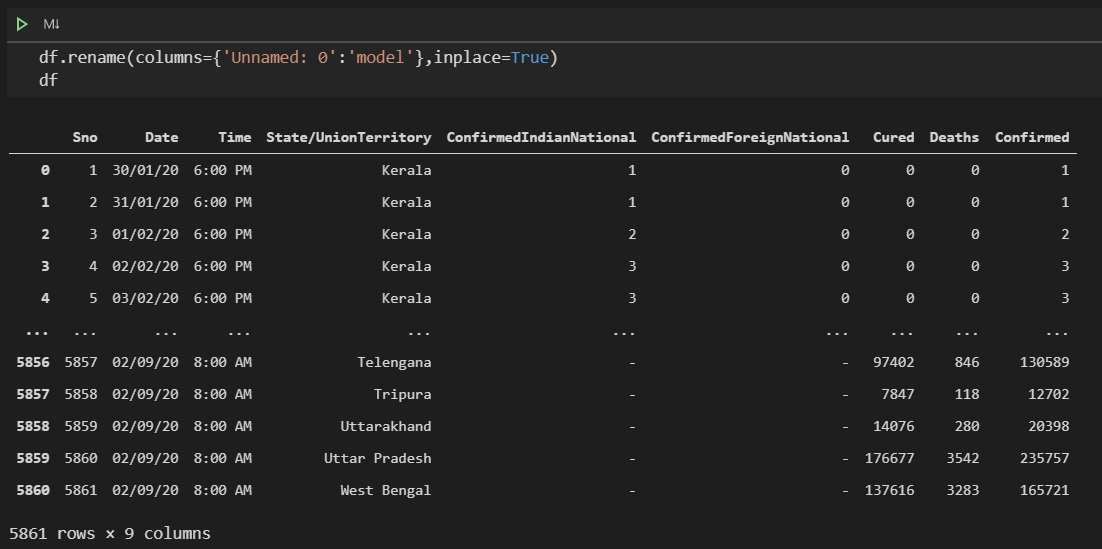


Figure : representation of head plus tail

**Checking of Columns with missing values**

Using isnull() and sum() functions we found the columns with number of missing values. It is required so that we can apply imputer for filling those missing values

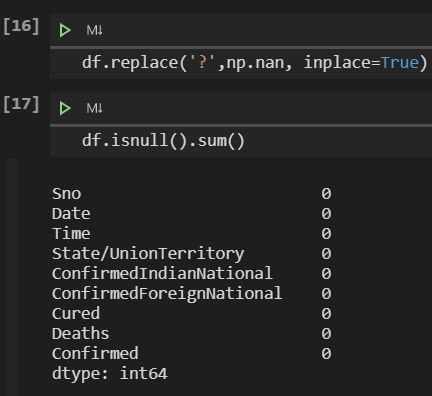


Figure : Representing no Missing Values

**Understanding the content**

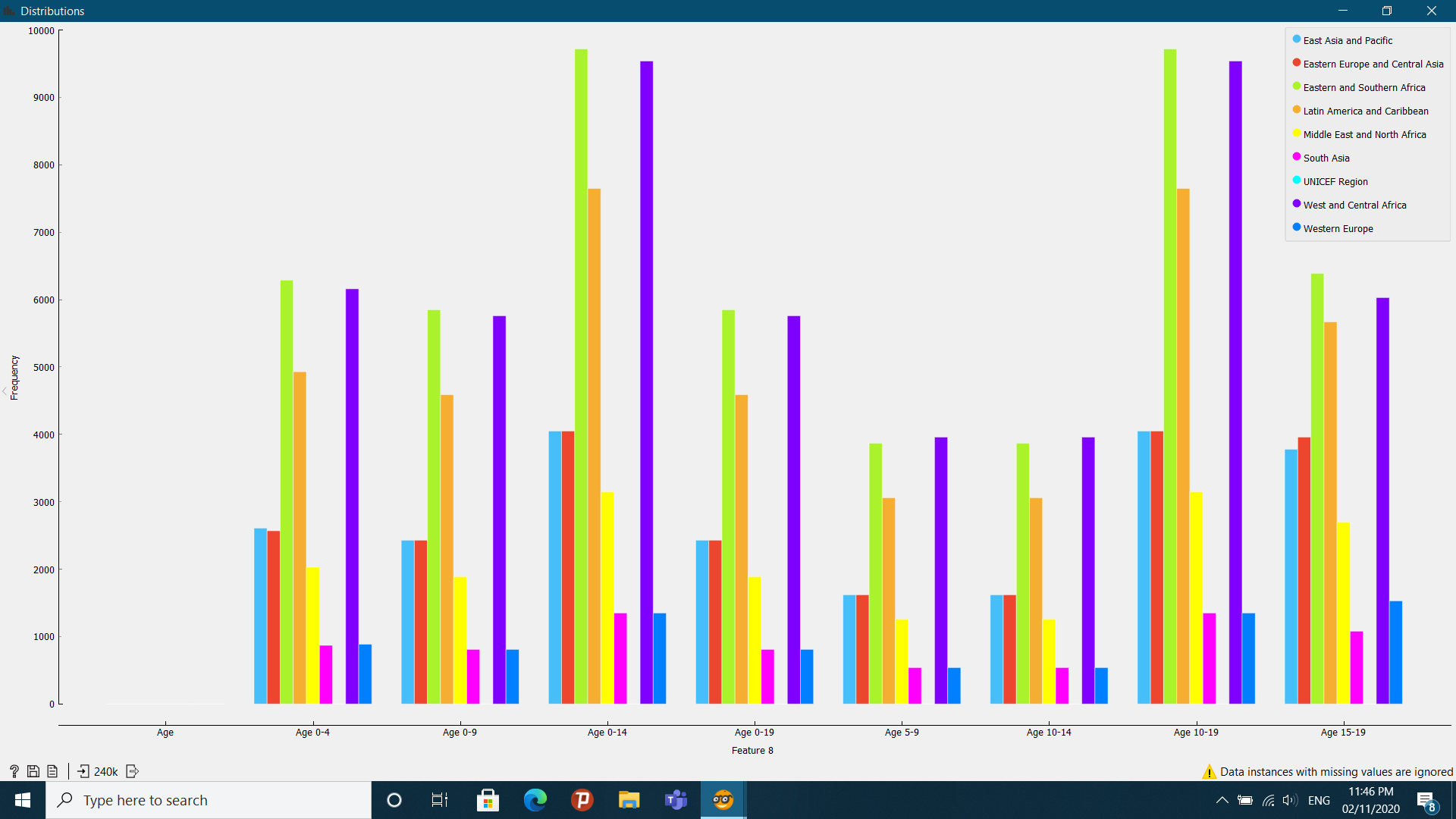


Figure : Distribution of people based on age and region suffering from HIV

After successfully completion of the data preparation and cleaning part, we headed towards Data Understanding Part. As shown in the figure above the figure represents the distribution of people age based on frequency suffering from HIV and split by their region. As a warning is shown that data instances missing are ignored is not actually missing values but actually not applicable so this is the representation of 240K values of dataset done in the ORANGE software. IN the upper right corner the regions are mentioned. From this distribution we understand that the age froup from 0 to 14 and 10m to 19 have a very large number of HIV Infections specially in East Asia and Pacific and West and Central Africa. We also made a conclusion that Africa Suffers a large number of HIV Infections. Whereas South Asia and western Europe suffers the minimum number of HIV Infections.

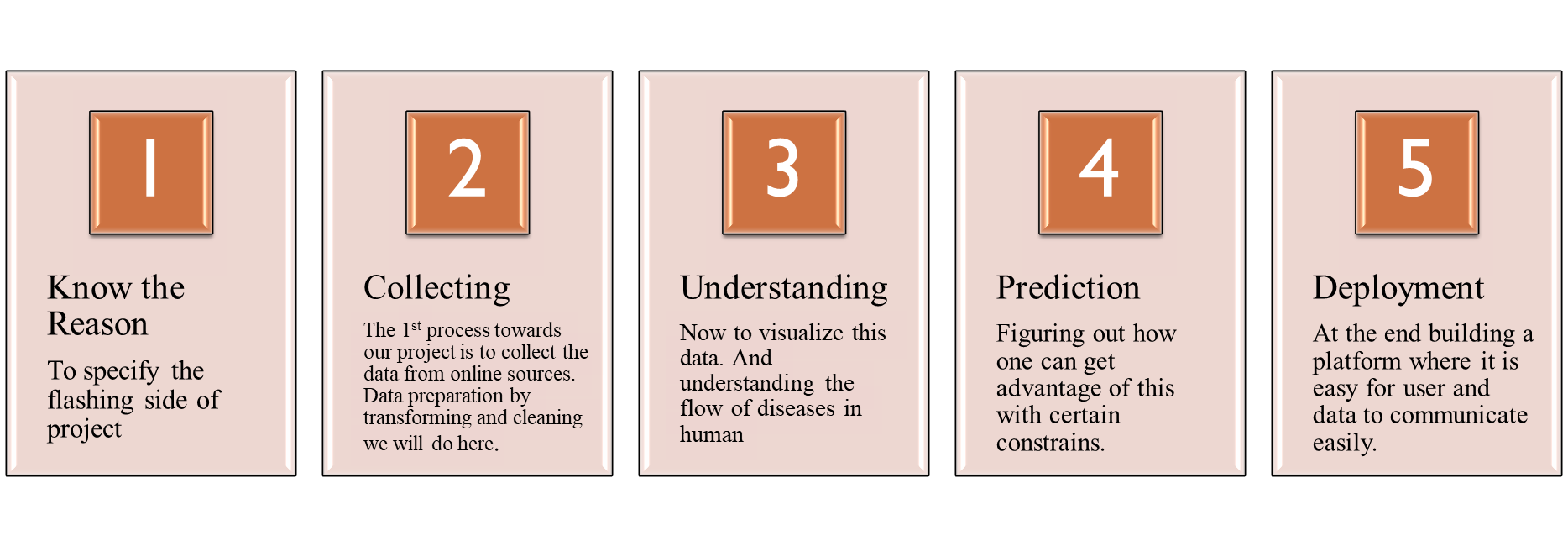
**Chapter 5 – Conclusion & Road map of Phase-2**

**5.1 CONCLUSION**

In our current project phase, we have collected datasets from various sources like CERN, Kaggle, UNICEF, Data World etc. The datasets that we have collected are in the form of xlsx, csv so the datasets we are having till today are of COVID, Dengue, Influenza, Pneumonia, Ebola, Hanta virus, Zika virus, HIV Aids. After collecting the datasets, we have performed data cleaning on these datasets. Data cleaning help us to remove improper data type which will make sure that our data is within the correct range not only that we are also able to do prediction with 80% accuracy. By performing data cleaning we have removed the inconsistencies which are caused due to user entry errors, or by corruption in transmission or storage so our datasets can now be used to generate graphs, charts, lists and other types of visualization which will help us to understand various aspects like which virus is more effective during which years, what are the chances of its returning in the upcoming seasons, what are the symptoms of a particular virus in humans and many more. In this phase the data collection and data cleaning of the data has been completed now in phase-2 we have a plan to implement various machine learning models on the collected data and also, decided to use various tools like IBM Cognos, R, Orange etc in order to generate different types of visualizations.

**5.2 FUTURE SCOPE**

**Workflow:**

****

In the future the study may be improved by including following points:

* At each visualization we will focus on the relationship between human and virus. More specifically we will study the effect on human based on gender, age, time, intensity, place. We will also look for symptoms reflected and their commonness. We will neglect the analysis above and below a specified age. We will relate the age with disease categorized by place.
* We will be studying more virus datasets in order to extend the usage of our model and to generate visualizations with more details.
* After data cleaning we have decided to do data processing in which we will filter out the missing values for this we will use some packages and libraries from python like:Numpy,Scipy,Scikit-learn,Theano,TensorFolw,Keras,PyTorch,Pandas,Seaborn,Matplotlib etc.We will be adding more libraries (if required) in the future for further advancement of our algorithms.
* When we are done with our data processing we will understand the data flow with the help of visualizations and then we will predict the data according to certain constraints.

After performing all these operations on the datasets deployment phase will come in which we will be creating either the web or an android application(whatever suits best for our model) and then deploy our model on that platform which will help the user to communicate with our data easily.

In Phase -1 we have gathered the data or API data sets from various organization and government official web-sites. We have filter out the many null/void data set or in other word we can say “Data Cleaning”.

Following is the planning of our Phase-2 part of our project:

* Phase-2

The flow-diagram showing the process phase 2. The following data will be analysed on various platform like IBM Cognos, Python, R studio, Orange.

Mainly the implementation will be analysing the data.

**Filter Data Sets/API from Phase-1**

**Implementation of NumPy, Pandas & Matplotlib**

**Phase-3**

**Phase-3**

**Using Of IBM Cognos For Analysis**

**Phase-3**

**Phase-3**

1. **IBM Cognos**

We will be using the IBM congos for the analysing the data and give the meaningful outcome. In IBM congnos analysis studio we can use for multidimensional analysis and exploration of large data sources and it is user friendly interactive environment to answers the data query.

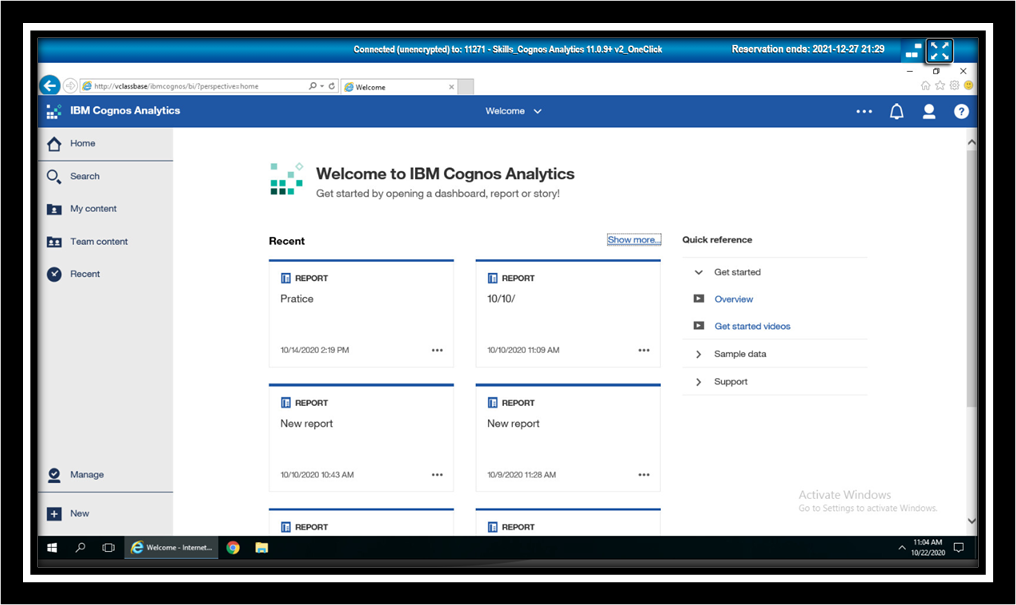


Figure : IBM Cognos

1. **Exploration**

During our Phase 2 process we will be using OLAP (Online Analytical Processing) exploration refer to the term dicing and slicing of the data. For example, we may look the death ratio for year 2006 to 2008 by gender. You may notice the rise or drop in the death ratio by clicking on the 2006 we can drill down to show the death result by quarter to 2006. It will help us to focus on the data that can answer our queries.

* Visualization

We will be using the visualization to communicate comparisons between relationships and our query. It will emphasize and clarify our data. Forecasting our data modelling which will be corresponding to visualization.

1. **Analysing Large Data**

In Analysis studio it will helps us to find the meaningful details while keeping summaries in view to maintain a clear overview of our data.

* NumPy

NumPy full form is Numerical Python is a perfect tool for performing basic and scientific computing and advanced array operations.



Figure : Python NumPy

It offers many useful features for performing many operations on n-arrays and matrices in Python. It helps to process arrays that store same data type values, and makes math operations on arrays much easier. The vectorization of mathematical operations on the NumPy array type can increases performance and accelerates the execution time of any data.

* Pandas

Pandas is a library will helped us to work with labelled and relational data and query counterintuitive. It's based on two data structures: "Series" which is one-dimensional, like items list and "Data Frames" – which is two-dimensional, like a table with multiple columns. It enables us to convert the data structures to Data Frame objects, and also allows the handling of missing data and adding/deleting columns from the Data Frame, imputing the missing files, and plotting the data with histogram or plot box. It’s a must-have for data manipulation, visualization, and wrangling.



Figure : Python Pandas

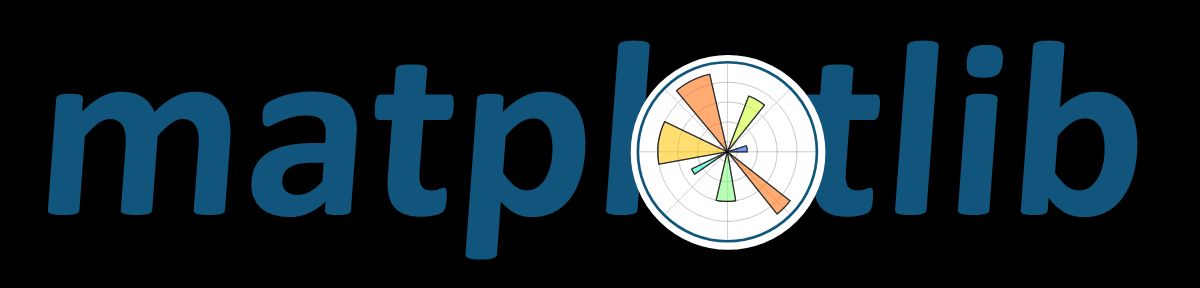
* Matplotlib

Figure : Python matplotlib

Matplotlib is a library used in python programming language and it is also used for the numerical, mathematical extension of the NumPy. It also provides an object-oriented API for embedding plots into data sets and the application for using general purpose GUI toolkit like:

* Tkinter



Figure : Python Tkinter

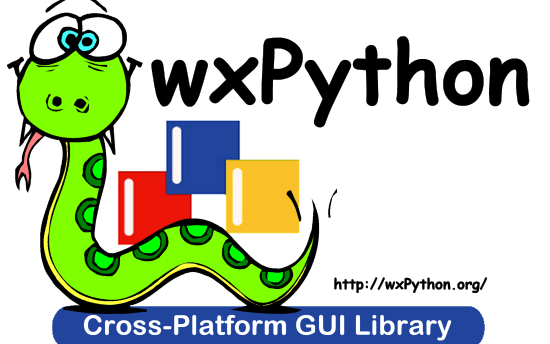
* wxPython 

Figure : wxPython

* Qt



Figure : Python Qt

* GTK+.

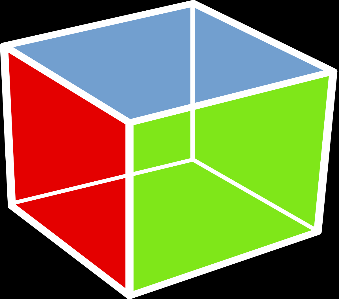


Figure : GTK

**Bibliography**

* 1. <https://home.cern/>
  2. <https://www.kaggle.com/>
  3. <https://www.unicef.org/>
  4. <https://data.world/>
  5. <https://catalog.data.gov/dataset>
  6. <https://www.kdnuggets.com/datasets/index.html>
  7. <https://www.ibm.com/in-en/products/cognos-analytics>
  8. Wikipedia
  9. Europa
  10. NCBI-NLM(National Library of Medicine)
  11. <https://api.covid19india.org/>

**Annexures :**

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5666085/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3130384/>

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0239960>

<https://www.sciencedirect.com/science/article/pii/S2468042720300087>

<https://www.researchgate.net/publication/328746395_Reconstruction_and_prediction_of_viral_disease_epidemics>

<https://www.who.int/globalchange/summary/en/index5.html>

<https://www.frontiersin.org/articles/348571>

<https://jvi.asm.org/content/83/11/5296>

<https://www.nature.com/articles/s41467-020-16153-4>



