**Data Analytics with Cognos - Group 1**

**Development Part 2**

**Au211521243187**

**VIVITHA.R**

**INTRODUCTION:**

The Covid-19 pandemic has shaken the world completely. No one knew what was coming and everyone was running helter-skelter. 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. 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

**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 which is govt website which has provide 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.

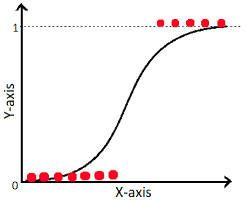
**Step 4: Predicting Covid +ve or -ve**

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.

**ALGORITHM:**

**1. LOGISTIC REGRESSION**

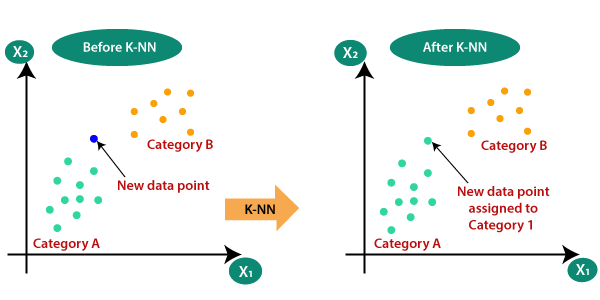
Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class. It is used for classification algorithms its name is logistic regression. it’s referred to as regression because it takes the output of the linear regression function as input and uses a sigmoid function to estimate the probability for the given class.



**2. KNN**

K-Nearest Neighbours is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining, and intrusion detection.

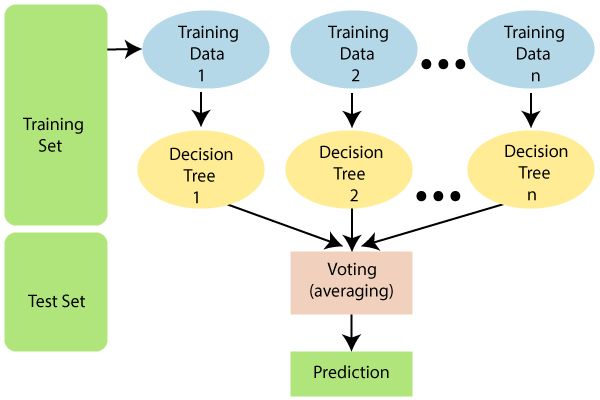
It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data). We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.



**3. RANDOM FOREST CLASSIFIER**

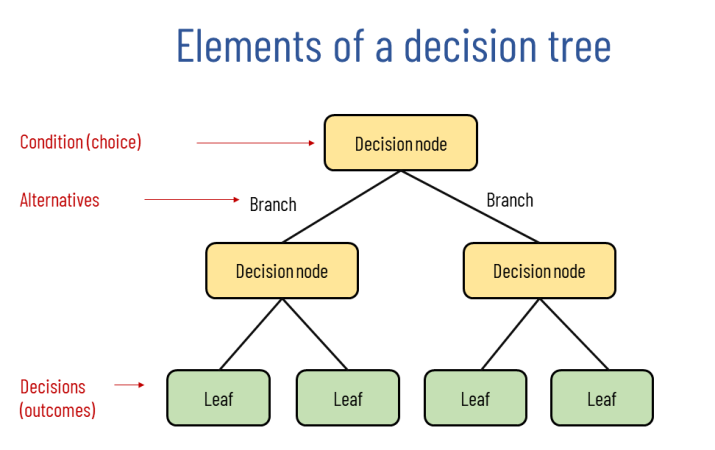
Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.



**4.DECISION TREE**

A decision tree is a flowchart-like tree structure where each internal node denotes the feature, branches denote the rules and the leaf nodes denote the result of the algorithm. It is a versatile supervised machine-learning algorithm, which is used for both classification and regression problems. It is one of the very powerful algorithms. And it is also used in Random Forest to train on different subsets of training data, which makes random forest one of the most powerful algorithms in machine learning.



Non-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. Open Refine for data scrubbing.

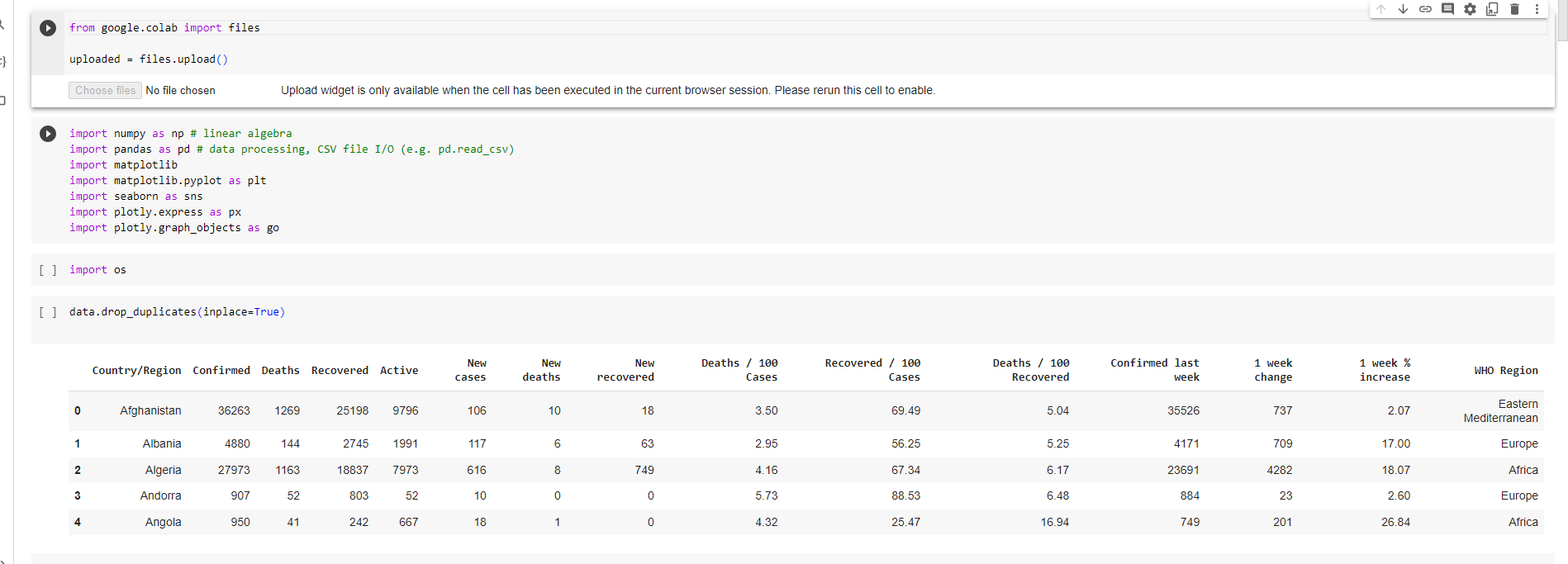
3. Numpy, Pandas, Matplotlib for data exploration, inspection and visualization.

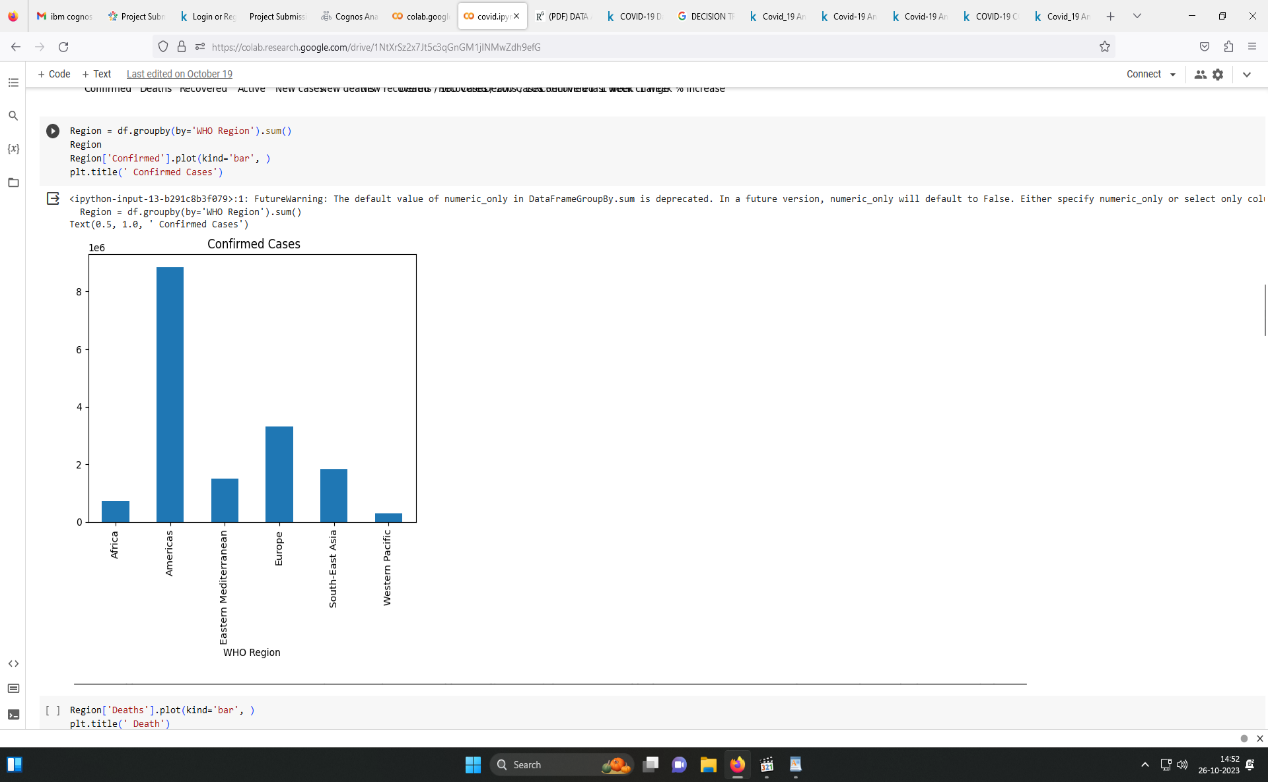
4. For modeling the data we need a decent knowledge of the Scikit library of Python.

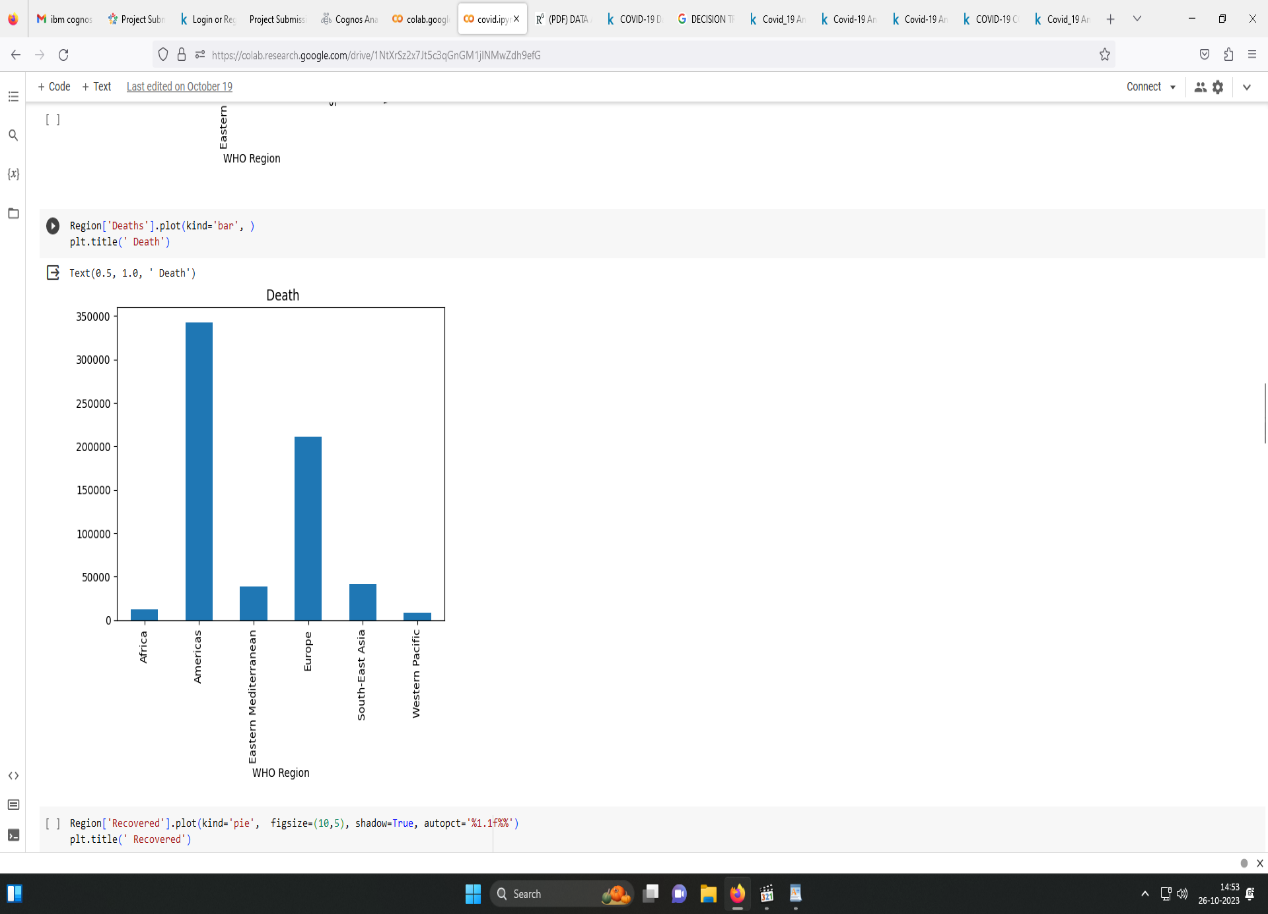
5. Training the dataset

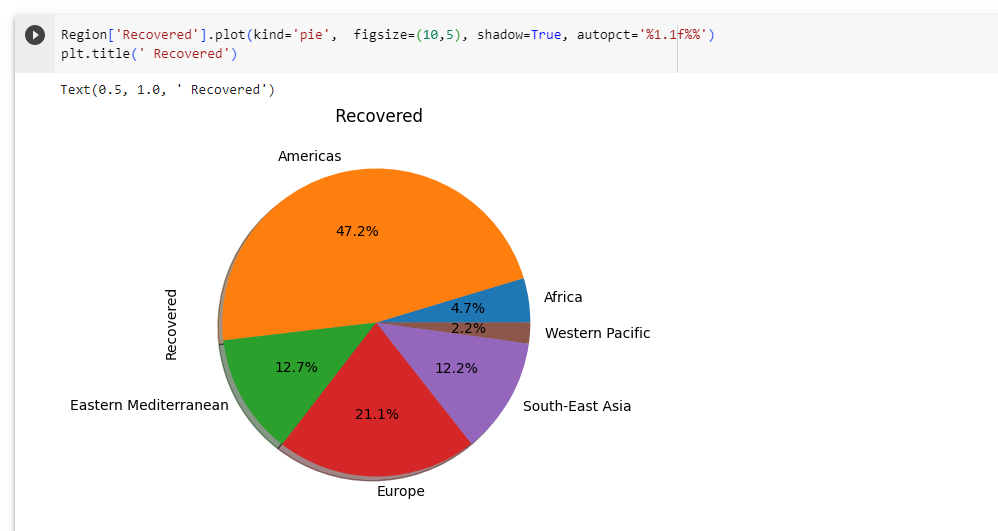
6. Matplotlib, ggplot, Seaborn, Tableau for interpreting the data.

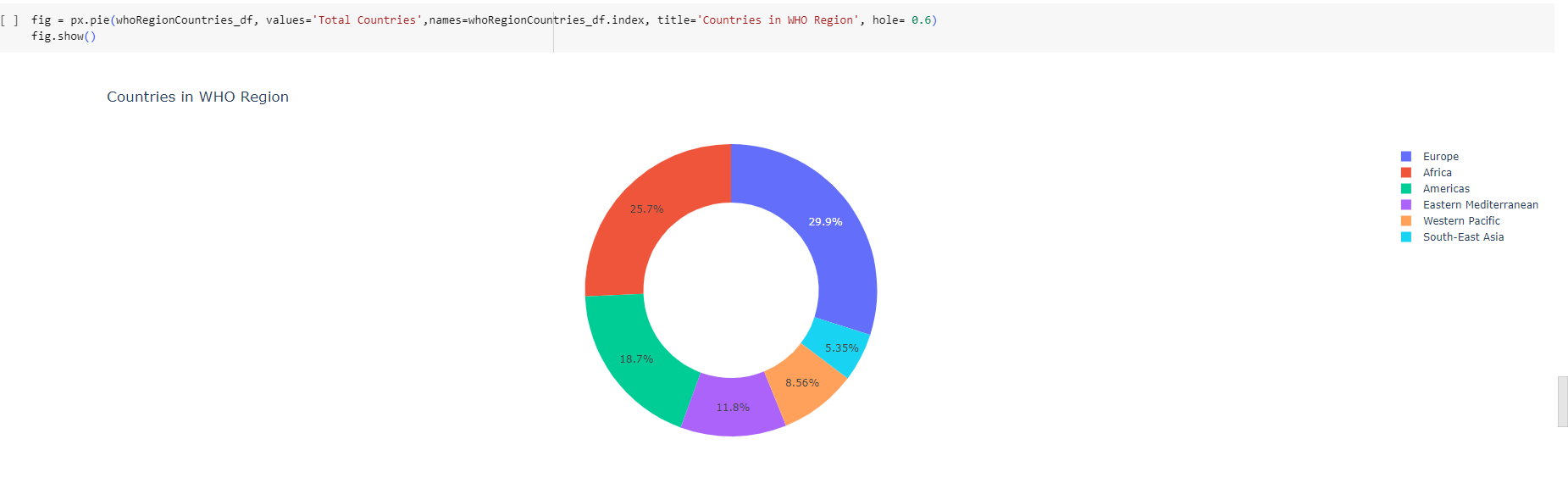
**RESULTS:**

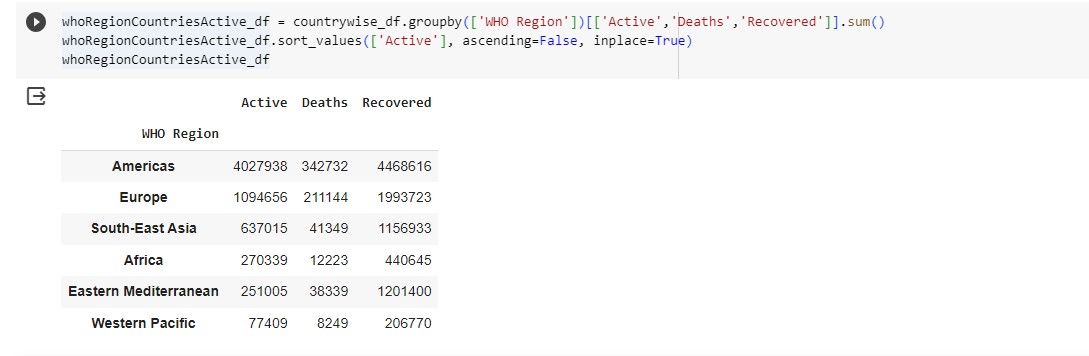
****

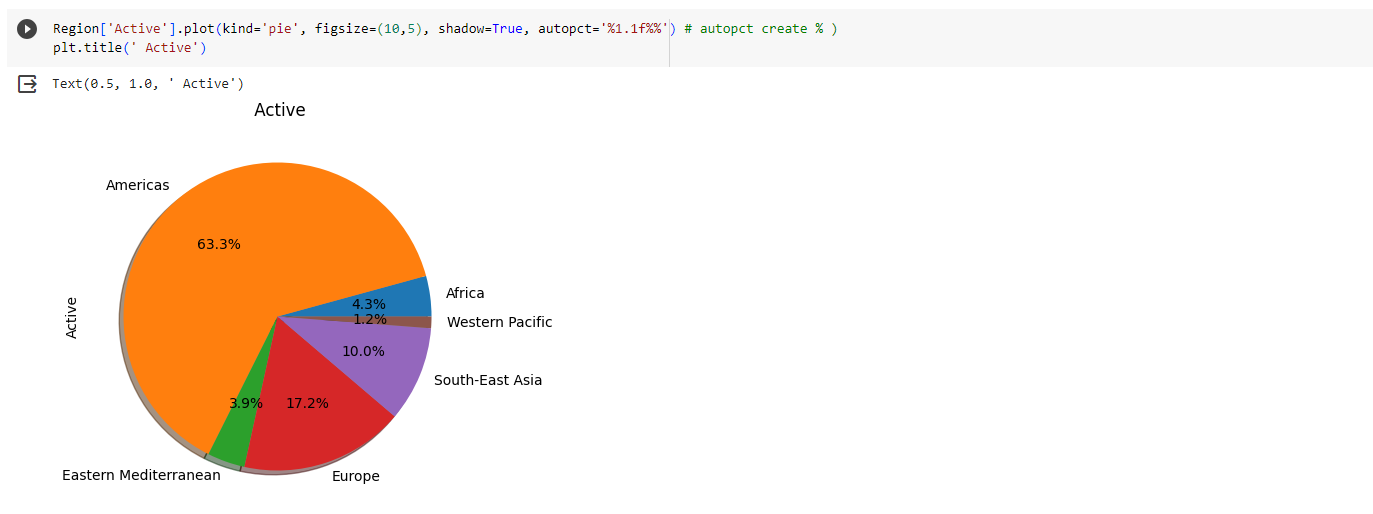
****

****

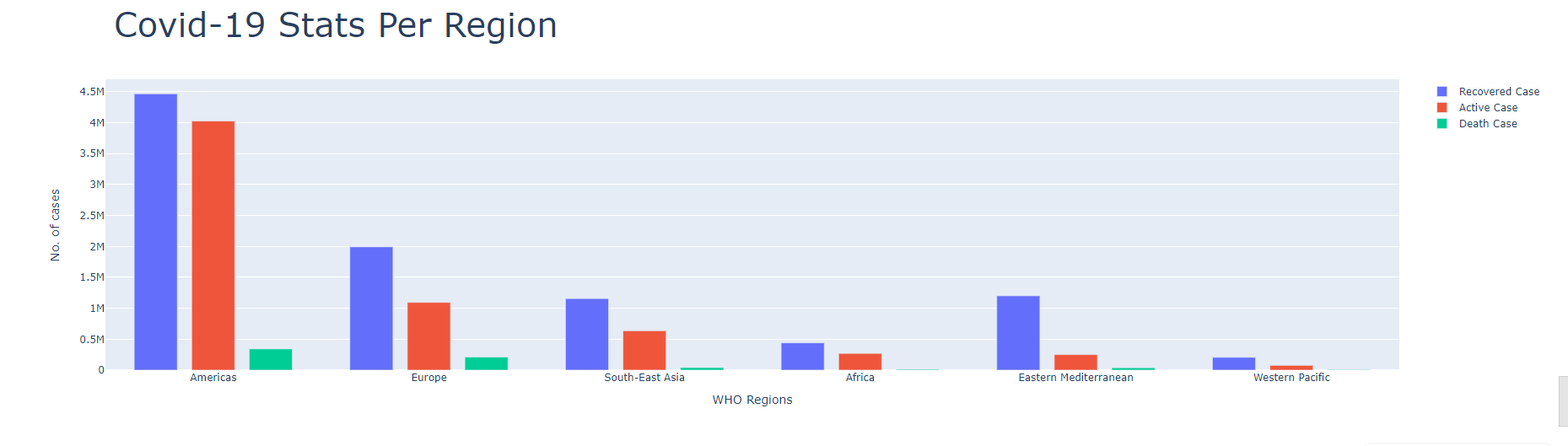
****

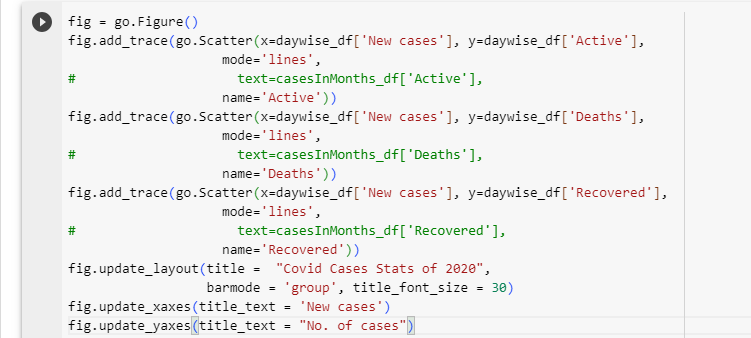
****

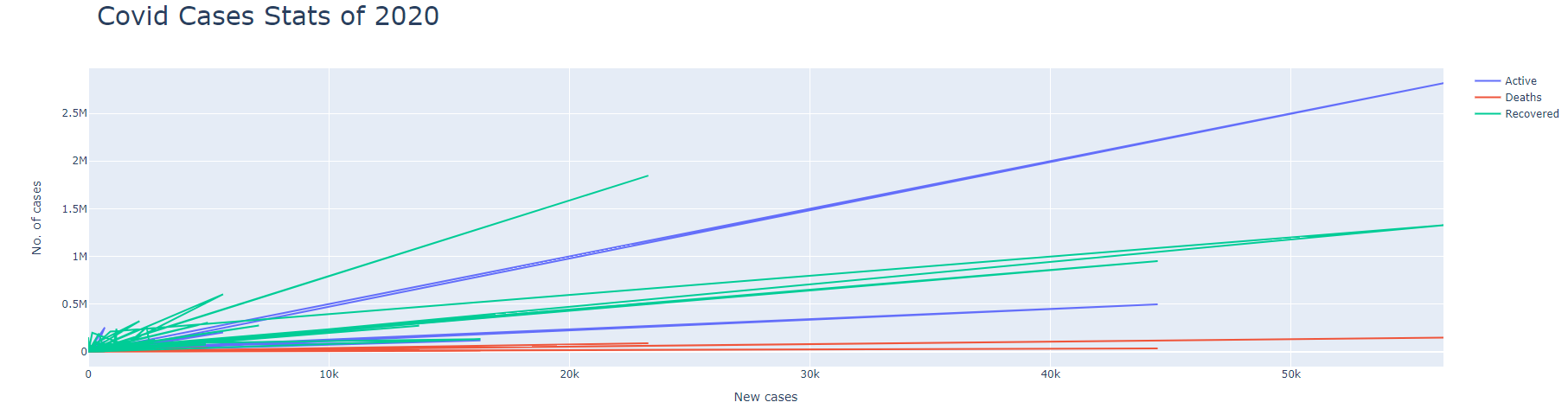








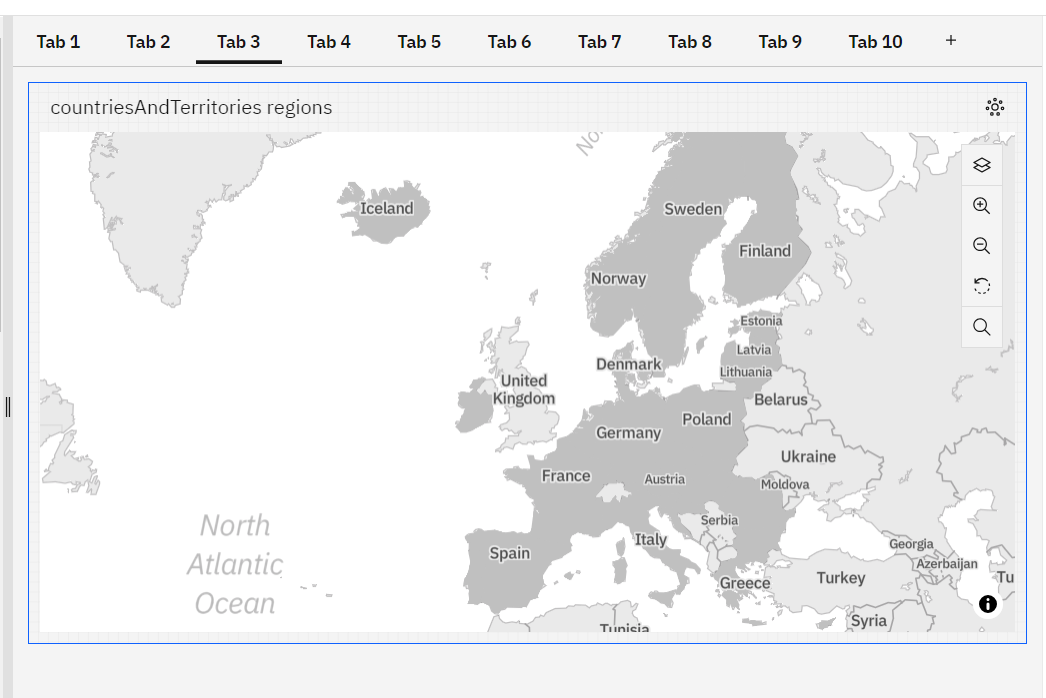




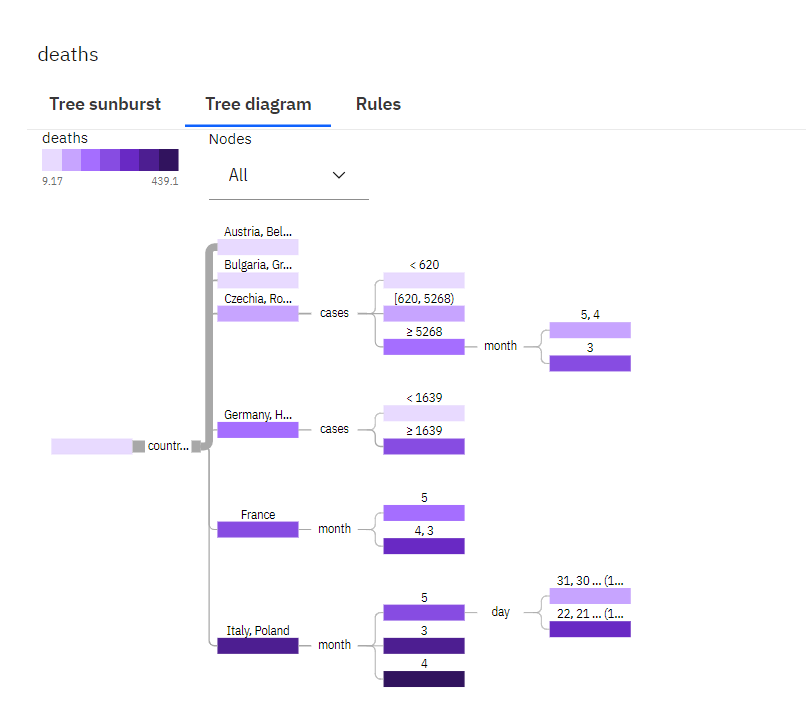
## The screenshots above show the code and results of the various phases of the Data Analysis done by us on our Covid-19 dataset. The implementation of data analysis has been carried out by various algorithms based on their accuracy. When analysis was done by using various algorithms the most accurate 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.

**Visualization using IBM® Cognos Analytics:**

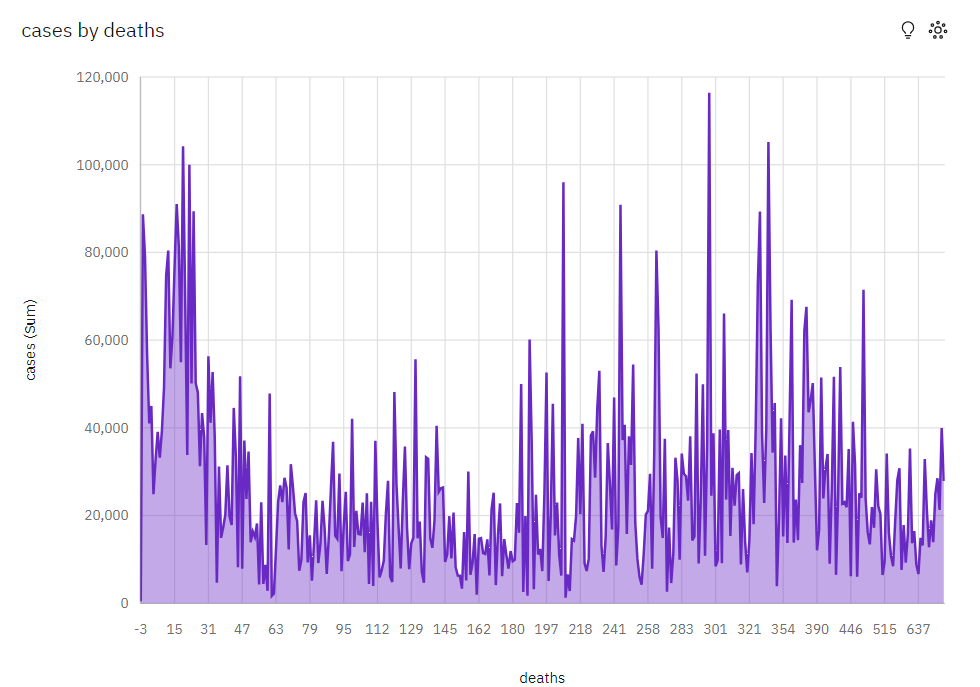
1.map visualization



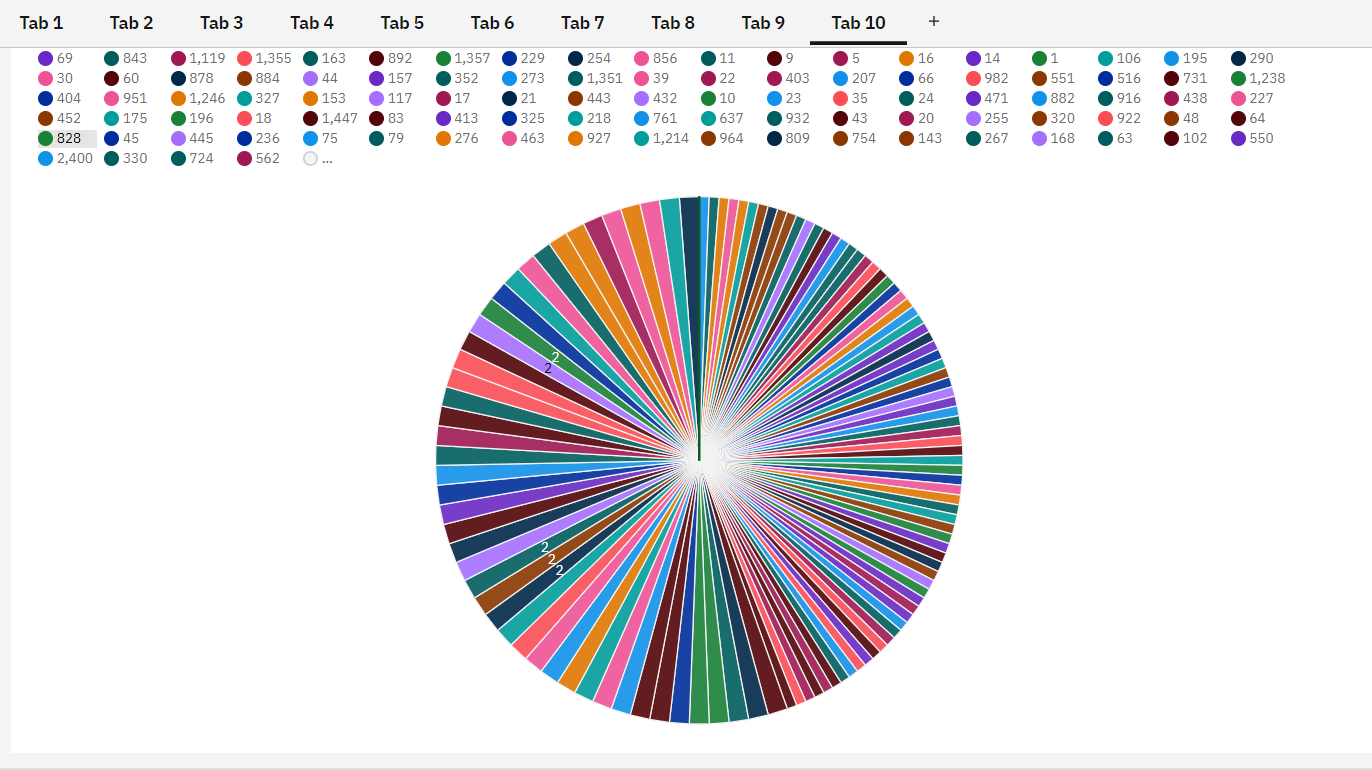
2.DECISION TREE VISUALIZATION



3.AREA VISUALIZATION



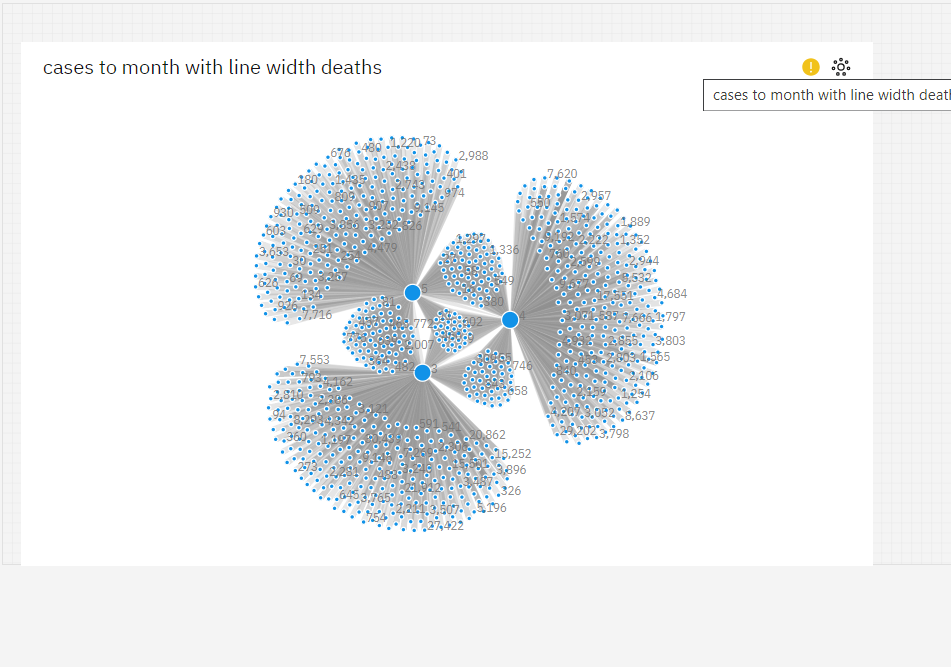
4.PIE CHART VISUALIZATION



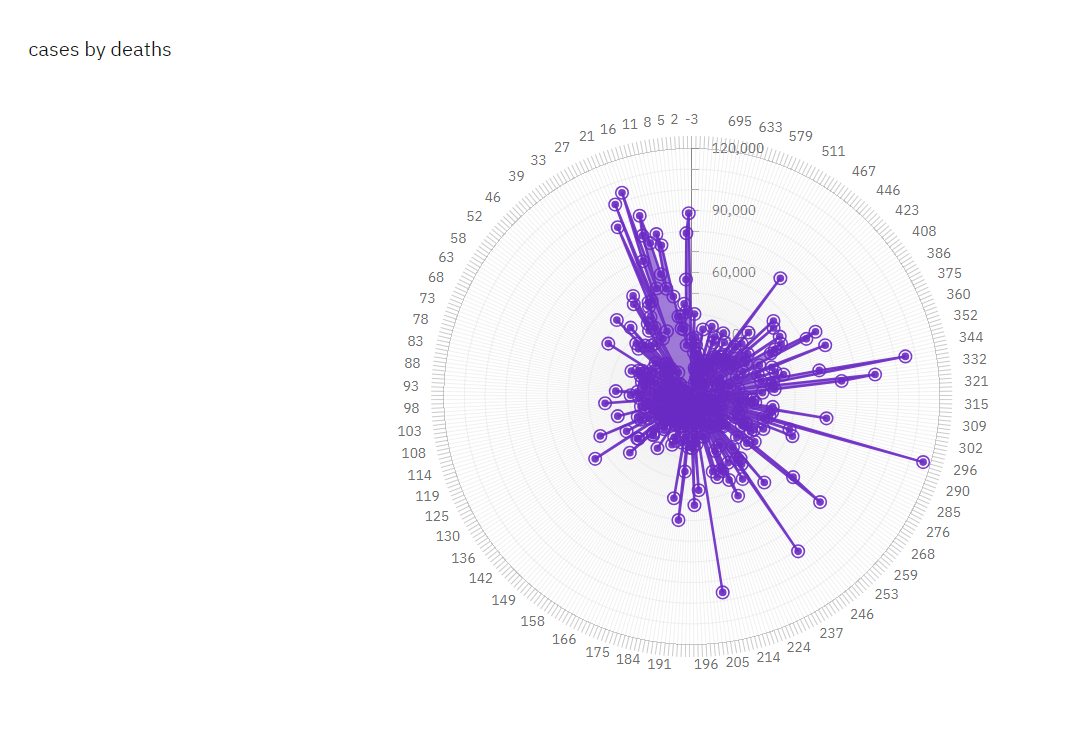
5.BAR VISUALIZATION



6. Network visualization



7.Radar visualization



**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.