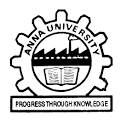
|  |  |  |
| --- | --- | --- |
|  | **COVID-19 CASES ANALYSIS WITH COGNOS** |  |

****

**NAAN MUDHALVAN PROJECT**

Submitted by

REEMASRI M (730321243020)

**SHILOH J S H (730321243023)**

**SUDHESHNA M (730321243026)**

**THENMUGI T (730321243029)**

FIFTH SEMESTER

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

**BUILDERS ENGINEERING COLLEGE, KANGEYAM**

**ANNA UNIVERSITY:: CHENNAI 600 025**

**DECEMBER 2023**



**BONAFIDE CERTIFICATE**

Certified that this is a bonafide record of work done by **REEMASRI M(730321243020) ,SHILOH J S H(730321243023)** ,**SUDHESHNA M(30321243026)** and **THENMUGI T(730321243029)** in AD3521 – Naan mudhalvan during the Academic year 2023- 24 for fifth Semester.

|  |  |
| --- | --- |
|  |  |
| **STAFF –INCHARGE** | **HEAD OF THE DEPARTMENT** |
|  |  |
| Submitted for the Naan mudhalvan vivavoce | held on……………….. |
|  |  |
| **INTERNAL EXAMINER** | **EXTERNAL EXAMINER** |

**ABSTRACT**

The COVID-19 pandemic has brought unprecedented challenges to global healthcare systems, demading the deployment of advanced analytical tools for effective data interpretation and decision-making. This study employs IBM Cognos, a business intelligence tool, to conduct a comprehensive analysis of COVID-19 cases. By integrating diverse data sources and leveraging Cognos' robust analytical capabilities, this research unveils critical insights into the spread, impact, and management of the pandemic. Through the utilization of Cognos' reporting and dashboarding functionalities, this analysis offers a detailed exploration of regional variations in infection rates, mortality patterns, and healthcare resource allocation.

Furthermore, the study explore into the identification of significant correlations between socio-economic factors, vaccination rates, and disease severity. Leveraging Cognos' predictive modeling features, this research also presents forecasts for future infection trends and potential outbreak hotspots, aiding in proactive policy formulation and resource planning. The findings of this study aim to provide valuable guidance to public health authorities, policymakers, and healthcare professionals in their efforts to mitigate the impact of COVID-19 and devise evidence-based strategies for effective pandemic management. By harnessing the power of Cognos for COVID-19 case analysis, this research contributes to the ongoing global discourse on pandemic response and highlights the pivotal role of data-driven insights in shaping public health interventions and preparedness for future health crises.

**TABLE OF CONTENT**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGE NO.** |
| **1** | **PROBLEM STATEMENT** | **1** |
| **2** | **DESIGN THINKING** | **3** |
| **3** | **DATASET DEFINITION** | **4** |
| **4** | **DATA PREPROCESSING AND FEATURE EXTRACTION** | **5** |
| **5** | **PROPOSED ALGORITHM** | **9** |
| **6** | **PROPOSED INNOVATION TECHNIQUE** | **12** |
| **7** | **CONCLUSION AND FUTURE SCOPE** | **16** |

CHAPTER 1

**PROBLEM STATEMENT**

The COVID-19 pandemic has had a profound impact on global health and society, making it essential to analyze and visualize data to better understand the spread, trends, and impacts of the virus. To address this challenge, we propose a project for COVID-19 cases analysis using IBM Cognos, a powerful business intelligence and data visualization tool. The project aims to:

* **Data Integration:**

Collect and integrate COVID-19 data from various sources, including government health departments, international organizations, and other relevant sources, into a centralized database.

* **Data Cleaning and Transformation**:

Clean, preprocess, and transform the raw data to ensure consistency, accuracy, and relevance for analysis.

* **Data Visualization**:

Develop interactive and informative data visualizations using Cognos to represent the following aspects of the COVID-19 pandemic.

a. Geospatial distribution of cases.

b. Trends in cases, recoveries, and fatalities over time.

c. Demographic analysis, such as age groups, gender, and comorbidities.

d. Impact on healthcare systems, including hospitalizations and ICU occupancy.

e. Vaccination rates and coverage.

* **Advanced Analytics**:

Apply statistical and predictive analytics to identify emerging trends, hotspots, and potential future outbreaks.

* **User-Friendly Dashboards**:

Create user-friendly dashboards that allow healthcare professionals, policymakers, and the public to explore and understand the data easily.

* **Scenario Analysis**:

Simulate different scenarios and policy interventions to help policymakers make informed decisions.

* **Data Security:**

Ensure the security and privacy of sensitive COVID-19 data in compliance with regulations.

* **Performance Optimization**:

Optimize the performance of the Cognos system for efficient data retrieval and real-time reporting.

* **Accessibility:**

Make the data and visualizations accessible to a wide range of stakeholders, including individuals with disabilities, to ensure inclusivity.

The primary goal of this project is to provide decision-makers and the public with a powerful tool for monitoring, analyzing, and responding to the COVID-19 pandemic effectively. It will help in understanding the current state of the pandemic, forecasting its trajectory, and developing evidence-based strategies for mitigating its impact.

CHAPTER-2

**DESIGN THINKING**

**1.Analysis Objectives:**

* Monitor and Visualize COVID-19 Trends
* Identify Hotspots and High-Risk Areas
* Evaluate Vaccine Distribution and Coverage
* Forecast Future COVID-19 Trends
* Ensure Data Quality and Integrity

**2.Data Collection:**

* Identify Data Sources
* Access Data
* Data Extraction
* Data Cleaning and Transformation
* Data Integration

**3.Visualization Strategy:**

Creating an effective visualization strategy for COVID-19 data using IBM Cognos involves designing visualizations that are not only aesthetically appealing but also informative, accessible, and actionable.

**4.Insights Generation:**

Generating insights from data is a crucial step in the data analysis process. It involves extracting meaningful and actionable information from raw data to support decision-making, problem-solving, and understanding of specific phenomena.

CHAPTER-3

**DATASET DEFINITION**

A COVID-19 cases dataset typically refers to a collection of data containing information about the number of confirmed COVID-19 cases, along with various associated details. This dataset may include information such as:

1. Date: The date when the cases were reported.

2. Location: Geographical information, including countries, states, or regions.

3. Total Cases: The total number of confirmed COVID-19 cases.

4. New Cases: The number of new cases reported on a specific date.

5. Deaths: The total number of COVID-19 related deaths.

6. Recoveries: The total number of individuals who have recovered from the virus.

7. Testing: Information about the number of tests conducted.

8. Hospitalizations: The number of COVID-19 patients admitted to hospitals.

These datasets are essential for tracking and analyzing the spread of the COVID-19 pandemic, assessing its impact, and making informed public health decisions. They are often used for research, epidemiological studies, and data visualization to monitor the progress of the pandemic. Data sources can vary, including government health departments, research institutions, and global organizations like the World Health Organization (WHO)

CHAPTER-4

**DATA PREPROCESSING AND FEATURE EXTRACTION**

**Data Collection**:

Obtain COVID-19 data from reliable sources, such as government health agencies, research institutions, or international organizations. Ensure that the data includes relevant attributes, as mentioned in the previous response.

**Data Integration:**

If you are using data from multiple sources, integrate the data into a single dataset, ensuring consistent data formats and structure. Merge data based on common keys, such as date and location.

**Data Cleaning:**

1. Handling Missing Data: Identify and handle missing values in the dataset. This may involve imputing missing values, using methods like mean, median, or interpolation.
2. Removing Duplicates: Check for and remove duplicate records to ensure data integrity.
3. Addressing Outliers: Examine data for outliers that may be due to data entry errors or anomalies and decide how to treat them, whether by removal, transformation, or replacement.

**Data Transformation:**

1. Data Aggregation: Depending on the level of detail in the data, you may need to aggregate it to a higher geographic or temporal level (e.g., summing daily data to create weekly or monthly summaries).
2. Calculating Derived Metrics: Calculate additional metrics like test positivity rates, case fatality rates, and growth rates based on the available data.
3. Normalization: Normalize data when comparing regions with different populations by calculating rates or percentages per capita (e.g., cases per 100,000 people).
4. Time Series Data: Organize data as a time series to enable trend analysis and visualization.

**Data Validation:**

Ensure that the processed data is accurate and consistent with external sources. Cross-verify data to maintain data quality.

**Data Analysis and Visualization:**

Use data analysis tools like IBM Cognos, Excel, Python, or R to create visualizations and perform statistical analysis. This step helps in identifying trends, hotspots, and patterns within the COVID-19 data.

**Geospatial Analysis:**

If your data includes geographic information, consider geospatial analysis to identify regional hotspots, trends, and spatial patterns.

**Predictive Modeling:**

Utilize predictive modeling techniques to forecast COVID-19 trends, such as future case counts or the impact of interventions.

**Report Generation:**

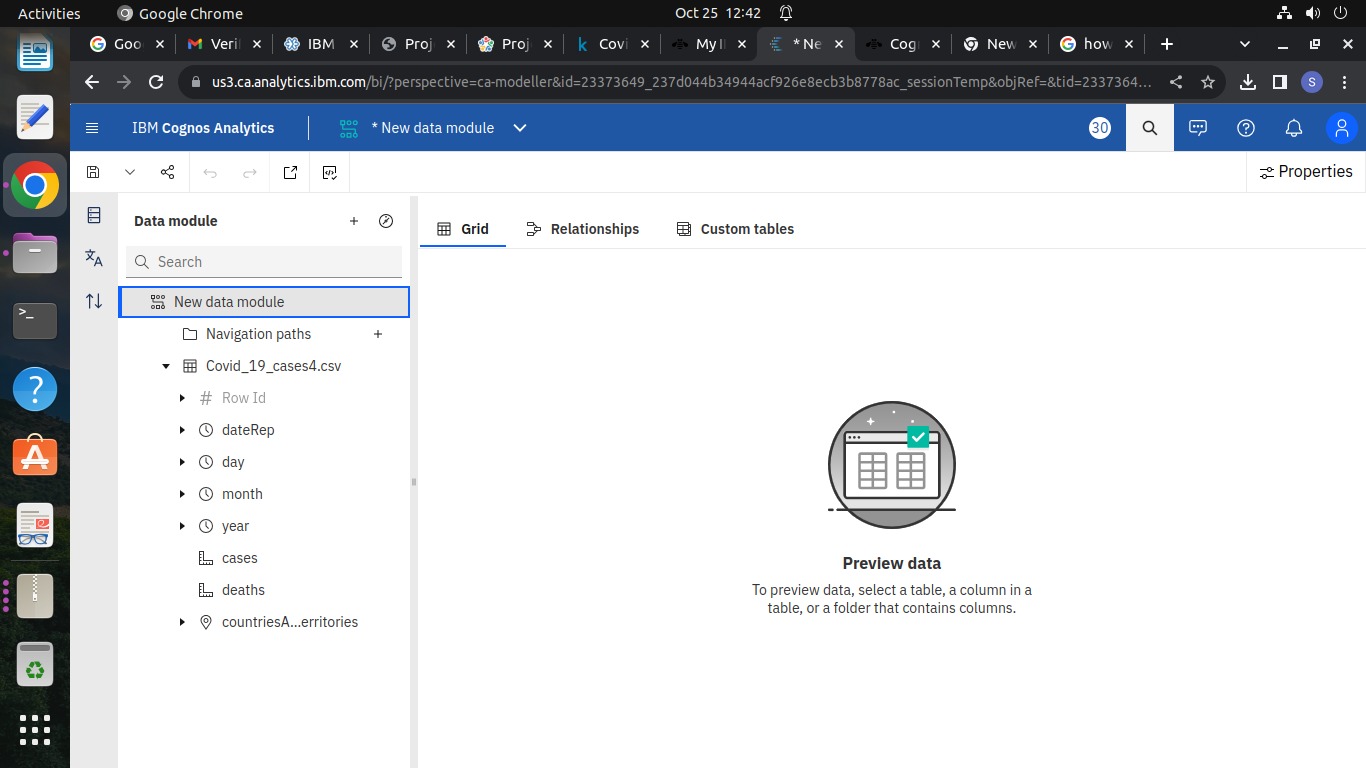
Create reports and dashboards that present the findings of your analysis in a clear and understandable format. This could involve using visualization tools and BI platforms like IBM Cognos.

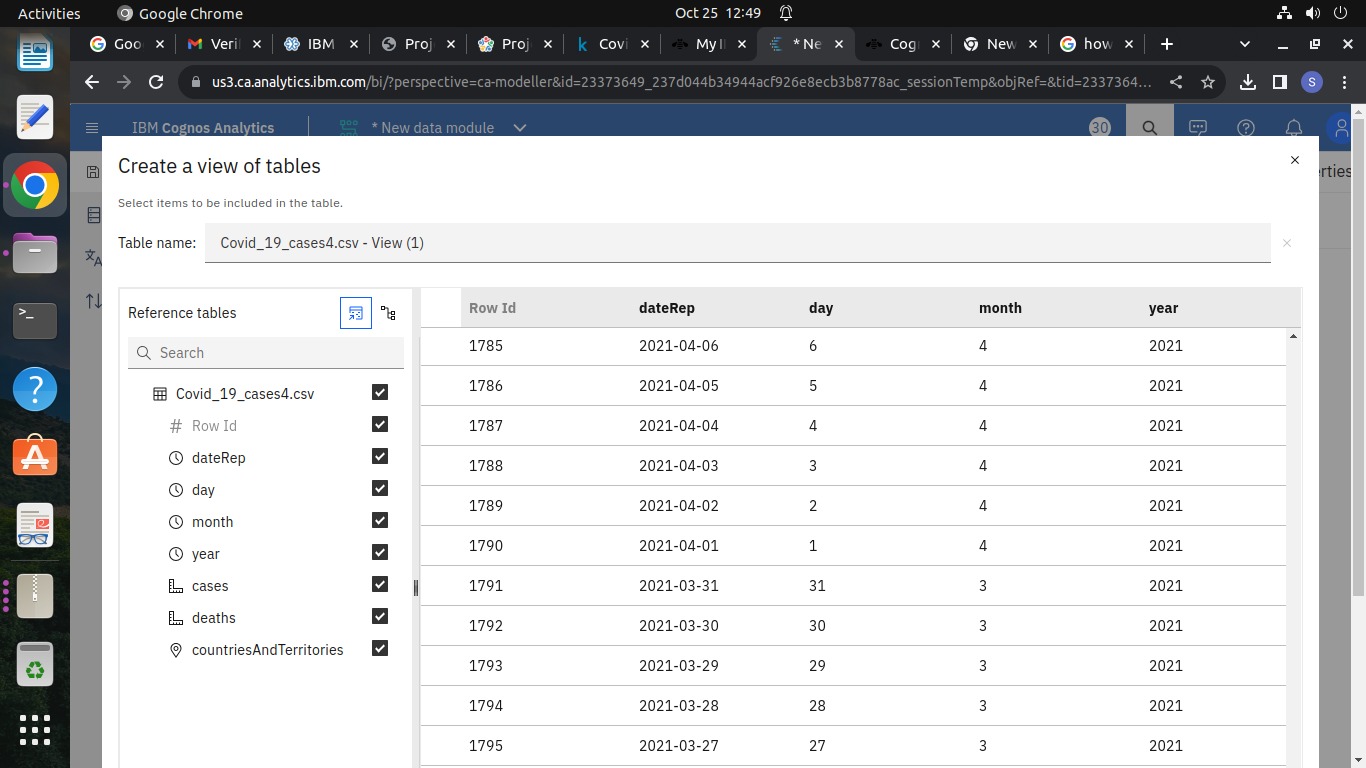
**Regular Updates:**

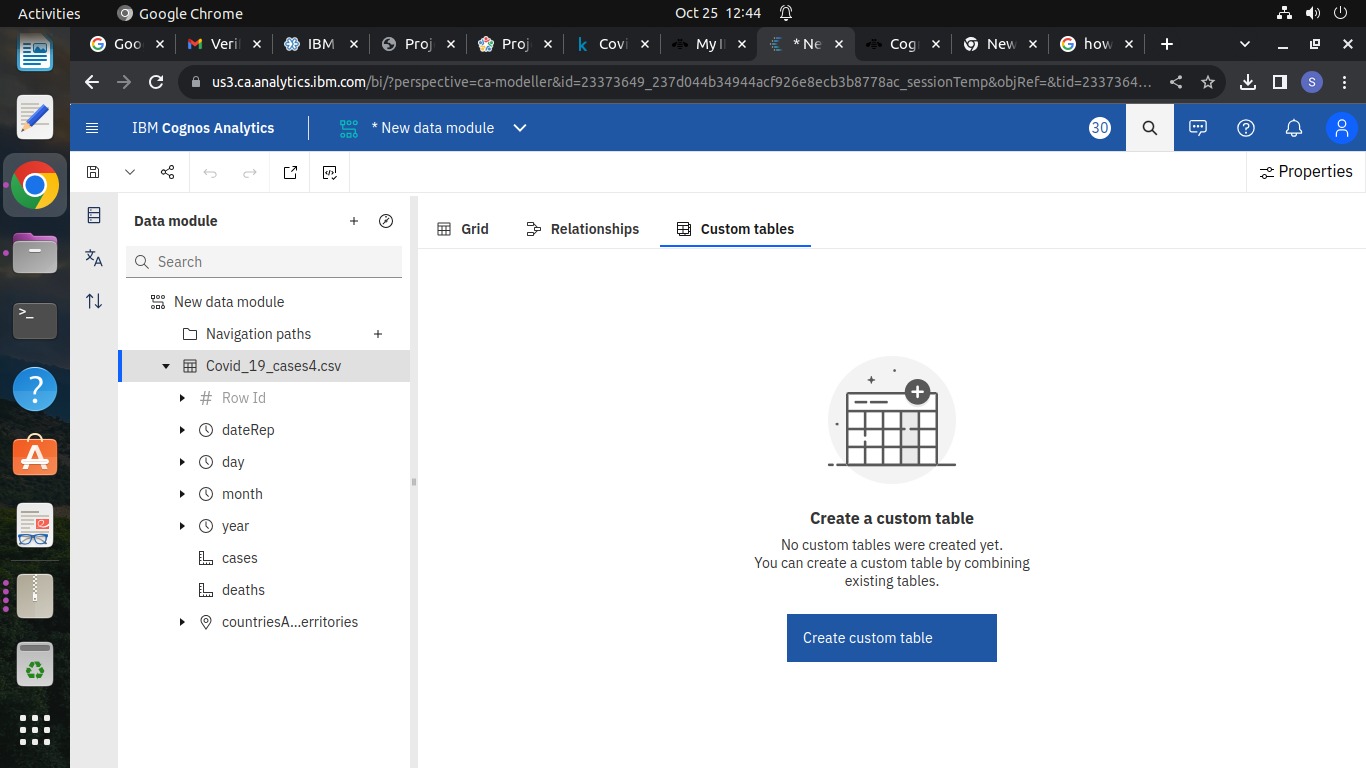
Ensure that your data processing pipeline is designed for regular updates to keep the analysis current as new data becomes available.

**Feature Extraction:**

Extract relevant features from the data that are essential for COVID-19 analysis. Some potential features may include  
 a. Date: Extract the date of the reported COVID-19 cases.  
 b. Region/Country: Extract the region or country where the COVID-19 cases were reported.  
 c. Number of Cases: Extract the total number of reported COVID-19 cases.  
 d. Number of Deaths: Extract the total number of deaths due to COVID-19.  
 e. Number of Recoveries: Extract the total number of recovered COVID-19 cases.  
 f. Demographic Information: Extract relevant demographic information such as age, gender, and pre-existing health conditions, if available.







CHAPTER-5

**PROPOSED ALGORITHM**

**SUPPORT VECTOR MACHINE**

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

**Hyperplane:** There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, the best decision boundary helps to classify the data points and that best boundary is known as the hyperplane of SVM

**Vector Machines:**

**Linear Separation:**

SVMs are originally designed for binary classification problems, aiming to find a hyperplane that best separates two classes of data points.

**Margin:**

The margin is the distance between the hyperplane and the nearest data point from either class. SVM aims to maximize this margin.

**Support Vectors:**

Support vectors are the data points that are closest to the hyperplane. These data points play a critical role in defining the margin and the decision boundary.

**Kernel Trick:**

SVMs can handle nonlinear data by transforming the input data into a higher-dimensional space. This transformation is often done using a kernel function (e.g., polynomial, radial basis function) that allows SVM to find nonlinear decision boundaries.

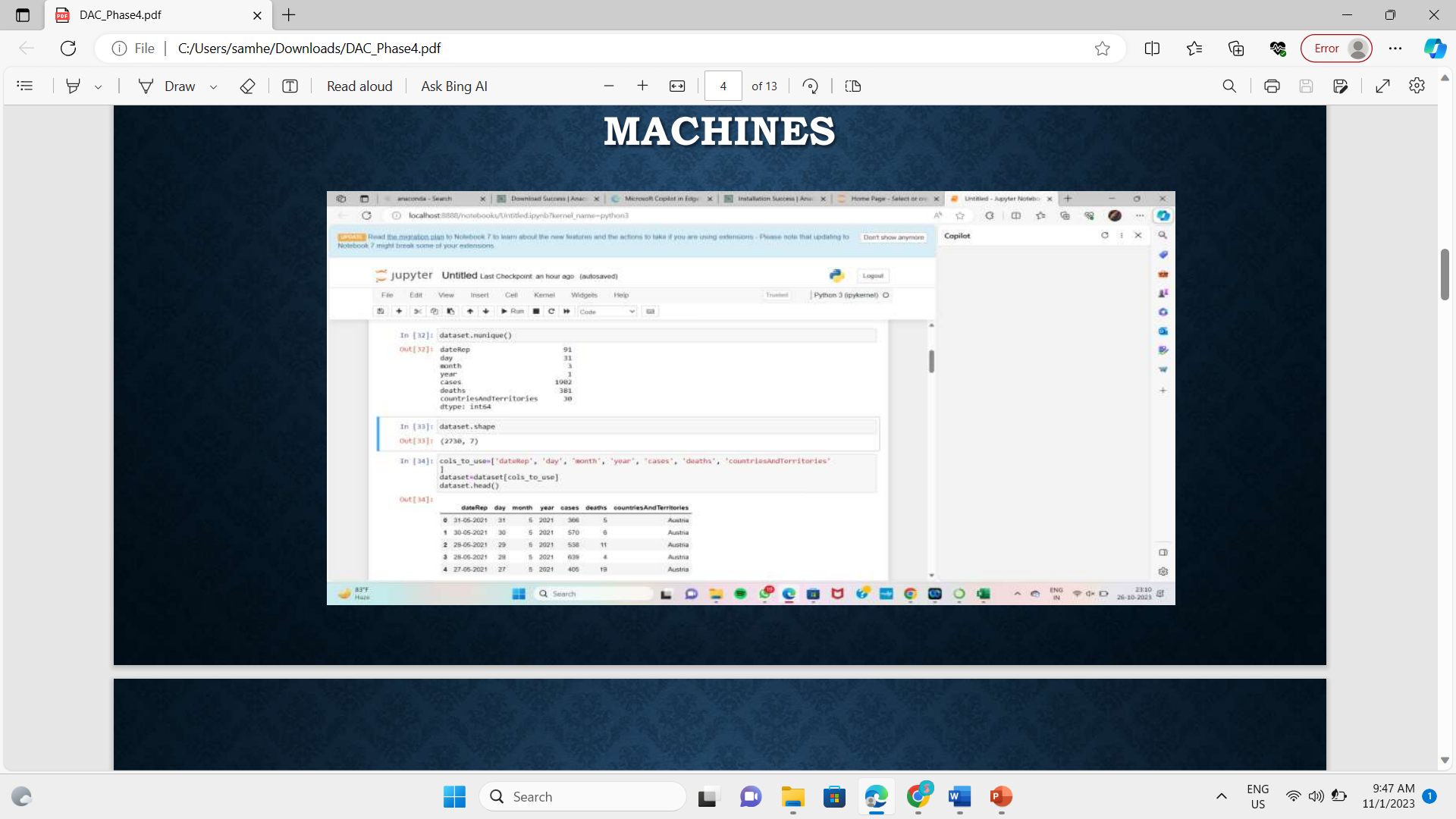
**C Parameter:**

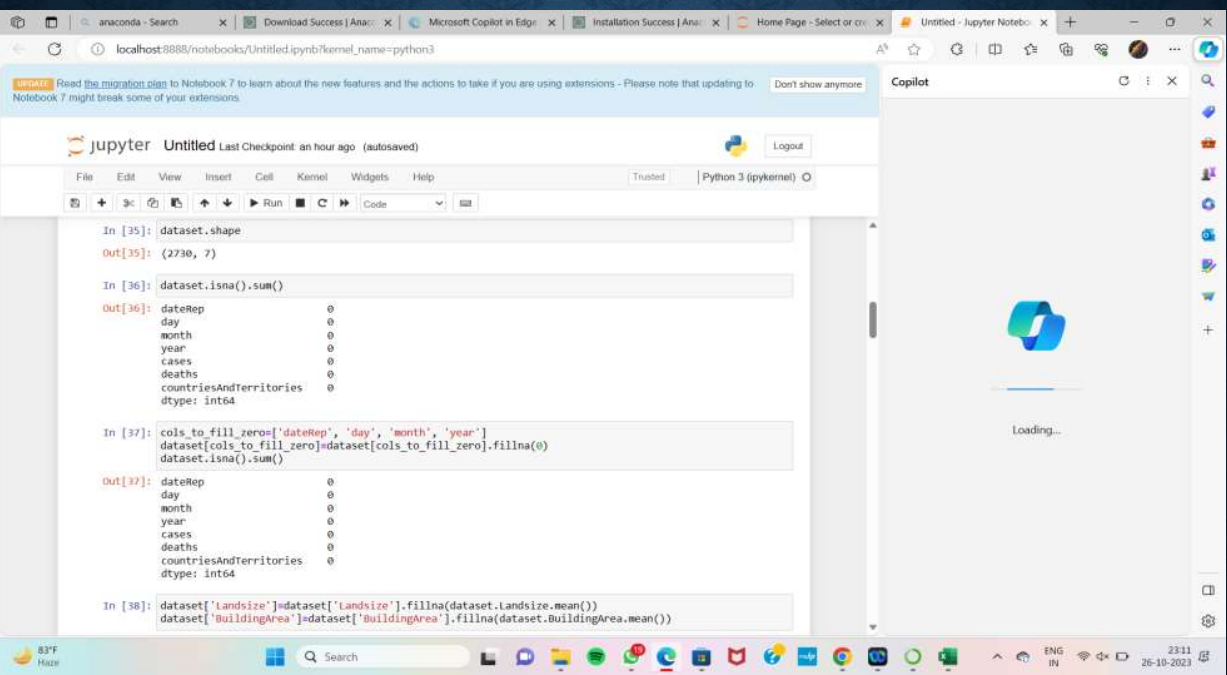
The C parameter in SVM controls the trade-off between maximizing the margin and minimizing the classification error. A small C value prioritizes a wider margin (potentially allowing some misclassification), while a large C value prioritizes correctly classifying as many points as possible (potentially leading to a smaller margin).

**Soft Margin SVM:**

In practical applications, data may not be linearly separable. In such cases, a soft margin SVM allows for a certain degree of misclassification to find a balance between maximizing the margin and minimizing errors.

**Multi-Class Classification:** SVMs can be extended to handle multi-class classification problems. Common techniques include one-vs-one and one-vs-all strategies**.**

****

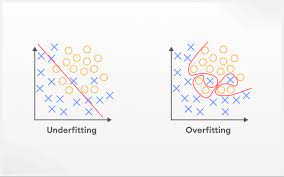
****

CHAPTER-6

**PROPOSED INNOVATION TECHNIQUE**

While developing machine learning models you must have encountered a situation in which the training accuracy of the model is low but the validation accuracy or the testing accuracy is too low. This is the case which is popularly known as overfitting in the domain of machine learning also this is the last thing a machine learning practitioner would like to have in his model. In this article, we will learn about a method known as regularization which helps us to solve the problem of overfitting. But before that let’s understand what is underfitting and overfitting. While developing machine learning models you must have encountered a situation in which the training accuracy of the model is low but the validation accuracy or the testing accuracy is too low. This is the case which is popularly known as overfitting in the domain of machine learning also this is the last thing a machine learning practitioner would like to have in his model. In this article, we will learn about a method known as regularization which helps us to solve the problem of overfitting. But before that let’s understand what is underfitting and overfitting.

* Overfitting is a phenomenon that occurs when a machine learning model is constrained to the training set and not able to perform well on unseen data. That is when our model learns the noise in the training data as well. This is the case when our model memorizes the training data instead of learning the patterns in it.
* Underfitting on the other hand is the case when our model is not able to learn even the basic patterns available in the dataset. In the case of the underfitting model is unable to perform well even on the training data hence we cannot expect it to perform well on the validation data. This is the case when we are supposed to increase the complexity of the model or add more features to the feature set.



**CODING:**

import numpy as np  
import matplotlib.pyplot as plt  
import pandas as pd  
import seaborn as sns

import warnings  
warnings.filterwarnings('ignore')

dataset=pd.read\_csv('./Covid\_19\_cases4.csv')  
dataset.head()

dataset.nunique()

dataset.shape

cols\_to\_use=['dateRep', 'day', 'month', 'year', 'cases', 'deaths', 'countriesAndTerritories'  
]  
dataset=dataset[cols\_to\_use]  
dataset.head()

dataset.shape

dataset.isna().sum()

cols\_to\_fill\_zero=['dateRep', 'day', 'month', 'year']  
dataset[cols\_to\_fill\_zero]=dataset[cols\_to\_fill\_zero].fillna(0)  
dataset.isna().sum()

dataset['Landsize']=dataset['Landsize'].fillna(dataset.Landsize.mean())  
dataset['BuildingArea']=dataset['BuildingArea'].fillna(dataset.BuildingArea.mean())

dataset.dropna(inplace=True)  
dataset.isna().sum()

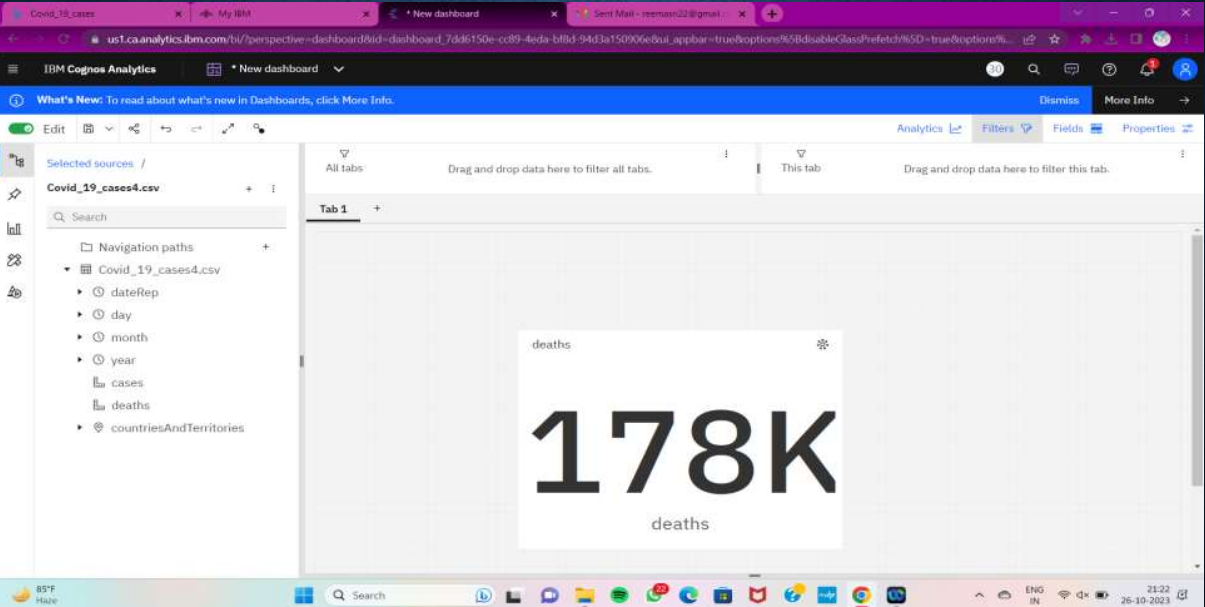
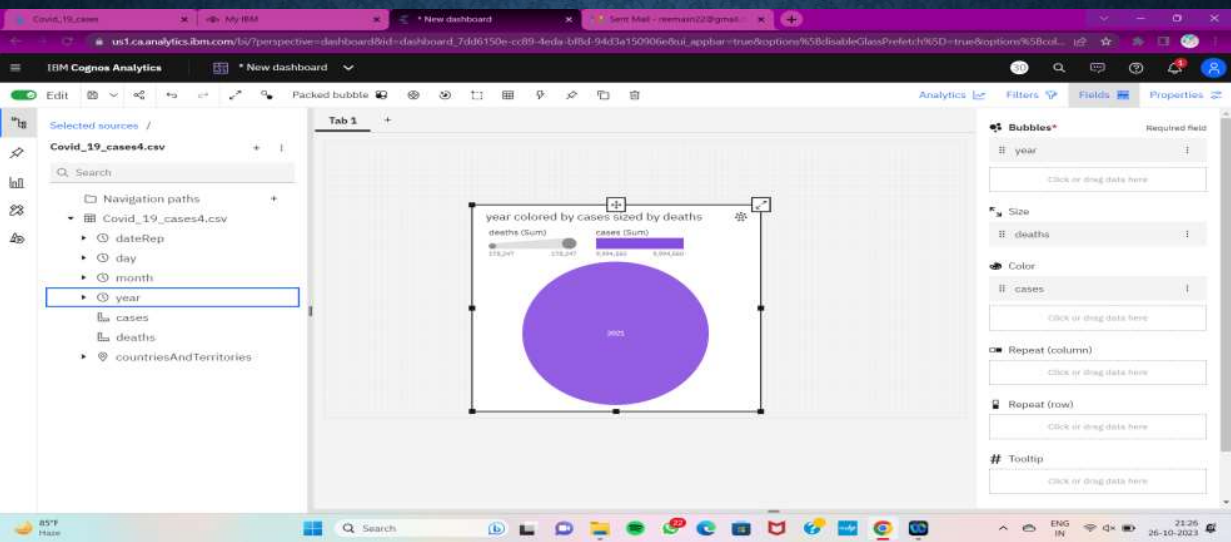
from sklearn.model\_selection import train\_test\_split  
train\_X,test\_X,train\_Y,test\_Y=train\_test\_split(X,Y ,test\_size=0.3, random\_state=2)from sklearn.linear\_model import LinearRegression

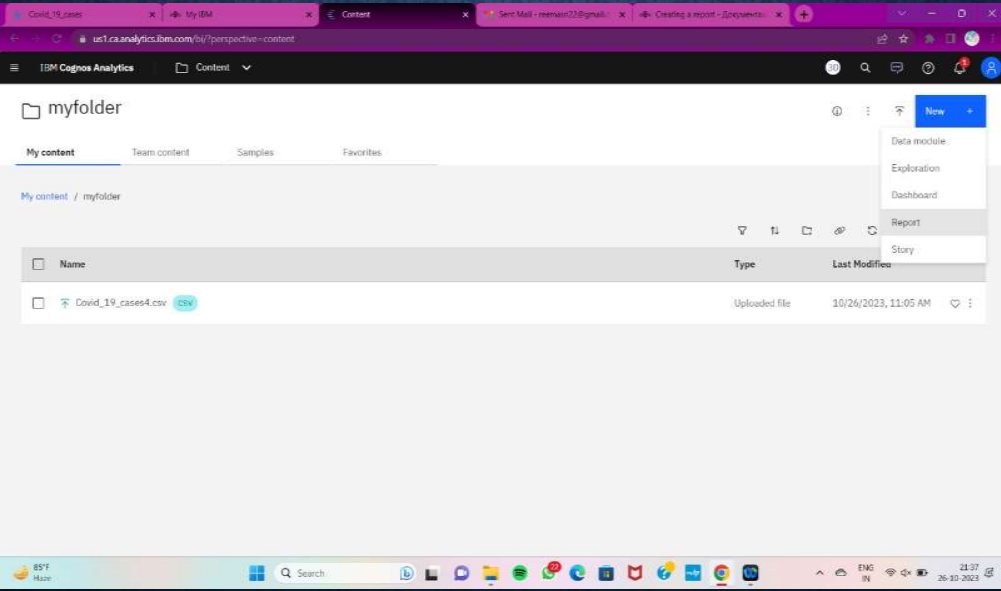
reg = LinearRegression().fit(train\_X, train\_Y)

from sklearn import linear\_model  
lasso\_reg=linear-model.Lasso(alpha=50, max\_iter=100,tol=0.1)  
lasso\_reg.fit(train\_x, train\_y)

ridge\_reg.score(train\_x, train\_y)

dataset=pd.get\_dummies(dataset,drop\_first=True)  
dataset.head()

****

****

CHAPTER-7

**CONCLUSION AND FUTURE SCOPE**

**CONCLUSION:**

"In this COVID-19 cases analysis conducted using Cognos, we have gained valuable insights into the impact and progression of the pandemic.

Key findings include:

1. Geographic Distribution: We observed variations in the number of COVID-19 cases across different regions, with some areas experiencing higher infection rates than others.
2. Trends Over Time: Over the course of the analysis, we noted fluctuations in the number of daily new cases, demonstrating the dynamic nature of the pandemic.
3. Demographic Patterns: Our analysis revealed that the virus affects various age groups differently, with older individuals being more susceptible to severe outcomes.
4. Testing and Outcomes: By examining testing rates and outcomes, we identified areas where more testing may be needed, as well as the recovery rates among COVID-19 patients.
5. Impact of Public Health Measures: We assessed the impact of public health measures, such as lockdowns and mask mandates, on the spread of the virus and observed their effectiveness in curbing the infection rate.
6. Recommendations: Based on our analysis, we recommend targeted testing and vaccination campaigns in high-risk areas, continued monitoring of the situation, and data-driven decision-making to manage the pandemic effectively.

This analysis, powered by Cognos, provides a comprehensive view of the COVID-19 situation, enabling us to make informed decisions and take appropriate actions to mitigate the impact of the pandemic. It highlights the importance of data-driven insights in managing public health crises."

**FUTURE SCOPE:**

The future scope of COVID-19 cases analysis with Cognos, , remains significant as the pandemic continues to evolve and as more data becomes available. Here are some aspects of future scope:

1. Advanced Predictive Analytics: Utilizing Cognos for predictive modeling to forecast COVID-19 trends, helping healthcare authorities and policymakers anticipate surges, resource needs, and plan interventions.
2. Vaccination and Immunization Analysis: Continued analysis of vaccination coverage, effectiveness, and its impact on reducing COVID-19 cases and deaths, especially concerning booster shots.
3. Public Health Decision Support: Supporting decision-makers with real-time data, enabling them to make informed choices about restrictions, mask mandates, and other preventive measures.
4. Variant Tracking: Analyzing the prevalence and impact of new COVID-19 variants, assessing their resistance to existing vaccines, and informing strategies for variant management.
5. Healthcare Capacity Management: Monitoring healthcare system capacity, such as ICU beds and ventilators, to ensure preparedness for potential surges in cases.
6. Epidemiological Studies: Conducting in-depth epidemiological research to better understand transmission dynamics, risk factors, and the long-term health effects of COVID-19.
7. Economic Impact Assessment: Analyzing the economic consequences of the pandemic and evaluating the effectiveness of stimulus packages and relief measures.
8. Data Integration: Integrating COVID-19 data with other healthcare and demographic datasets for a more comprehensive understanding of the virus's impact on public health.
9. Global Collaboration: Collaborating with international health organizations to analyze and share data to combat the global spread of the virus.