NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY,

BELGAUM, APPROVED BY AICTE & GOVT.OF KARNATAKA)

LA2 PROPOSAL

On

COVID-19 Predictions

Submitted in partial fulfilment of the requirement for the award of

Degree of Bachelor of Engineering

In

Computer Science and Engineering

Subrina Pradhan	1NT19CS189
Hrishita Rauniyar	1NT19CS083
Soumya Dn	1NT19CS184

TABLE OF CONTENTS

INTRODUCTION

WHAT IS DATA MINING?

The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data Mining.

Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Using a broad range of techniques, you can use this information to increase revenues, cut costs, improve customer relationships, and reduce risks and more.

Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data. Data mining is also called Knowledge Discovery in Database (KDD). The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.

Data Mining is the process of investigating hidden patterns of information to various perspectives for categorization into useful data, which is collected and assembled in particular areas such as data warehouses, efficient analysis, data mining algorithm, helping decision making and other data requirement to eventually cost-cutting and generating revenue.

Data mining is the act of automatically searching for large stores of information to find trends and patterns that go beyond simple analysis procedures.

DESCRIPTION OF PROJECT

Corona viruses are a large family of viruses which may cause illness in humans. In humans, several corona viruses are known to cause respiratory infections ranging from the common cold to more severe diseases. The most recently discovered corona virus causes corona virus

Disease COVID-19.

COVID-19 is the infectious disease caused by the most recently discovered corona virus. This new virus and disease were unknown before the outbreak began in Wuhan, China, in December 2019.

The global reaction to the threat has generally increased the perception of the severity of the illness and the threat it poses to the peoples. However, there is a widespread lack of clarity on specifics related to the illness. The general view of the virus is that it is spreading rapidly. The most common symptoms of COVID-19 are fever, tiredness, and dry cough. These symptoms are usually mild and begin gradually.

TASKS WE WILL IMPLEMENT IN THIS PROJECT

Yes we will be implementing data mining tasks on our required project such as classification, clustering, association rule mining, and prediction.

COVID-19 IN WEKA

1. J48 DECISION TREE

A decision tree is an algorithm that produces a graphical tree-like structure, wherein instances are classified using a root node having a test condition (e.g., the person has sore throat or not) and branches that determine answers. A leaf node represents a class to which all the instances belong; if the instances belong to different classes, it is called a test node, which consists of a condition added to the attribute's value, and a test node can be further represented in two or more sub trees. One kind of decision tree algorithm is J48 DT, which is a common and simple decision tree algorithm used for classification purposes; it uses a divide and conquer approach, which divides the instances into sub ranges based on the values of the attributes.

2. KNN ALGORITHM

This algorithm relies on the distance metric used to determine the nearest neighbors of the given instance, and the most commonly used metric is the Euclidean distance, which is expressed in the following formula: d xi , xj = s n ∑ r=1 wr ar(xi) − ar xj where an example is defined as a vector x = (a1, a2, a3, . . . , an), n is the number of the example's attributes, ar is rth attribute of the example and its weight is referred to as wr , and d xi , xj are the two examples. To compute the class label of an example, the following formula is used: y (di) = arg max k ∑ xj∈ kNN y xj, ck (7) where di is the example by which the

algorithm will determine the class in which it belongs, the term xj is one of the k-NNs present in dataset, and y xj, ck indicates whether the xj belongs to the class ck. The result of Equation is the class that has the most members of the k-NN, and is also the class wherein the example belongs. Euclidean distance is mostly used as a default distances.

DATA MINING TASKS

Logistic Regression (LR)

Logistic regression is used to determine the association between categorical dependent variables against the independent variables. LR is used when the dependent variable has two values such as 0 and 1, yes and no or true and false and thus it is called binary logistic regression. However, when the dependent variable has more than two values, multinomial logistic regression is used. A mathematical model of a set of explanatory variables for LR is used to predict a transformation of the dependent variables. LR transformation is written mathematically as:

$$i=\text{Logisticregression}\ (p)\ =\ln(p1-p).$$

Decision Tree (DT)

Decision tree (DT) is used for classification tasks in data mining and successful technique due to its ability to handle both categorical and continuous data, simplicity and comprehensibility, DT builds tree into phases which include growth and pruning phases respectively.

Naive Bayes (NB)

Naive Bayes is one kind of data mining classification algorithm and used to discriminate dataset instances based on specified features or attributes.

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

Above equation determines the theorem used.

K-Nearest Neighbor (K-NN)

K-nearest neighbor is a non-parametric and supervised data mining classifier used for regression and classification tasks. In both tasks, the input variables consist of the K closes training dataset in the feature space. K-NN relies on labeled input data to learn a function so that to produce appropriate output when inputted unlabeled data.

GOALS OF DATA MINING TASKS:

Given in the data set required it involved techniques such as knn, decision tree which is used to analyze data for expected relationships between given data set.

Pattern functions in this data set is to analyze the records and delete frequently occurring patterns. Patterns are being extracted from the dataset that does not exist.

We extracted the information given in a data set from kaggle and transform the information into a comprehensible structure for further using of the given algorithm.

DATA SET

Dataset Collection and Description

The dataset was obtained from Kaggle website.We used the dataset of covid-19 patients. The dataset has instances and attributes.

Dataset Preparation

The dataset was prepared, and cleaned where only relevant attributes were extracted from the original dataset. The extracted dataset has data instances with attributes we considered only two states of the patient which include released and deceased while isolation state was excluded. The missing value in the dataset reduces the prediction power and produces biased estimates leading to an invalid conclusion.

METHODS

Python programming language was used for data mining predictive tasks. Python is a well-known general purpose and dynamic programming language that is being used for different fields such as data mining, machine learning and internet of things .Data mining algorithms are being implemented using python with the help of the special purpose libraries.

ASSESSMENT

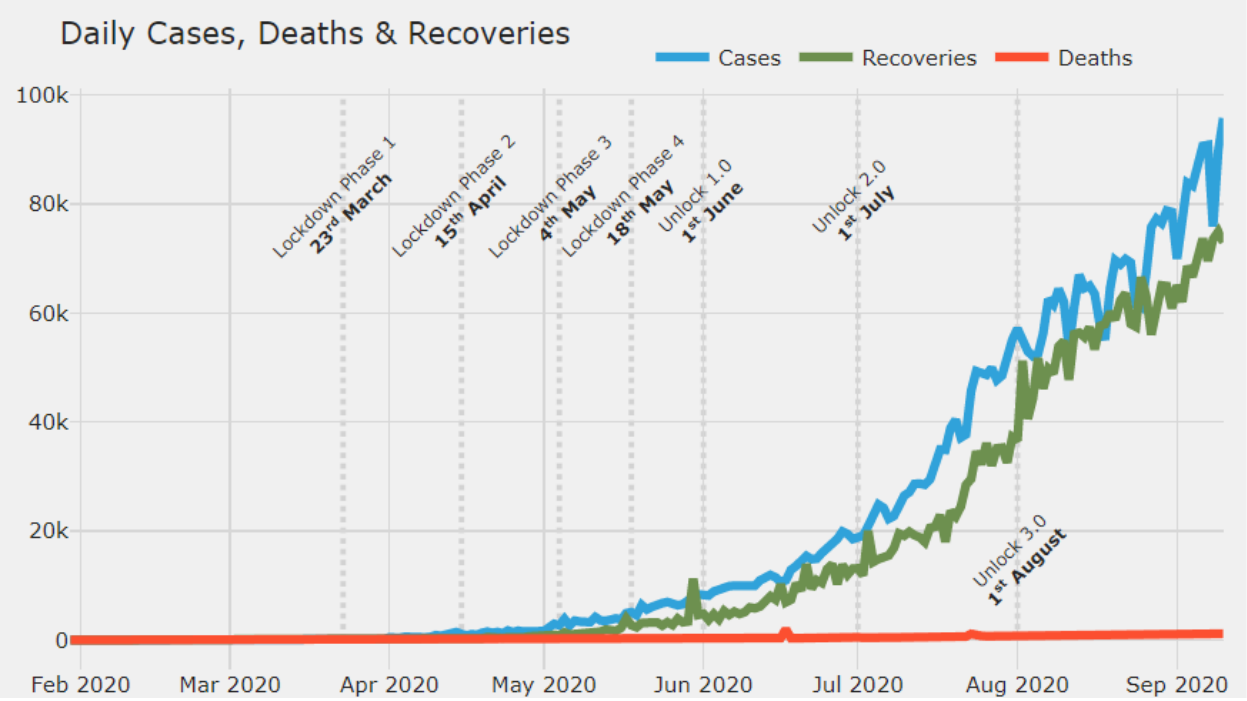
1. Isorisk maps

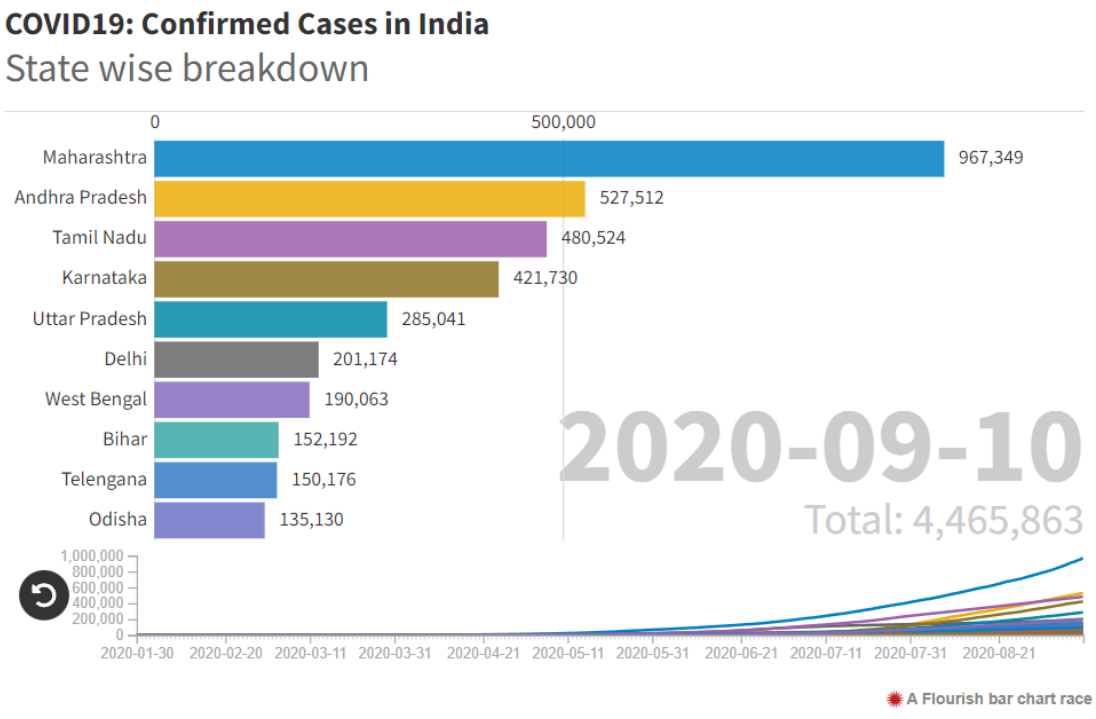
The Individual Risk (IR) is calculated to quantify the is-risk curves which estimate risk at specific points by taking into account both the effect zone and the probability. The total individual risk at each point is evaluated as the sum of the individual risks, at that point, of all incident outcome cases associated with the hazard source. The effect zone will include all the droplets coming from the release source. Probabilities of failure have to be set according to the situation being analyzed thanks to the help of further professionals as experts of infection transmission, occupational medicine, and biological risk in workplaces, risk assessment and management of indoor environment.

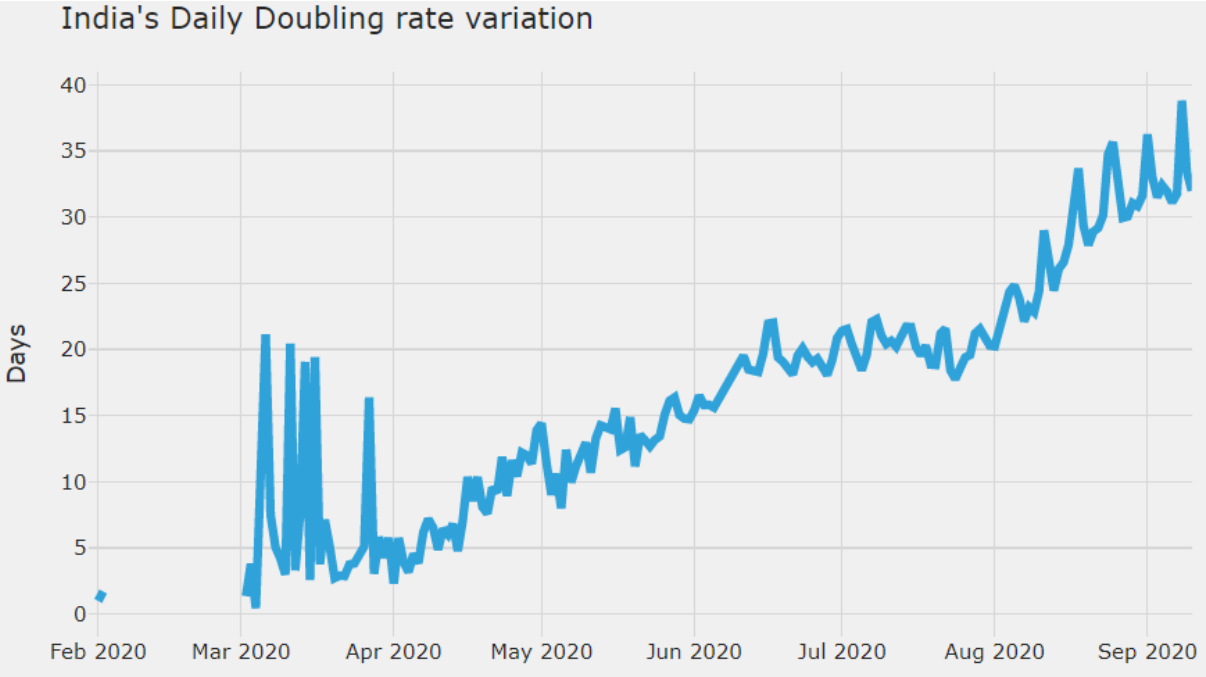
2. CFD model

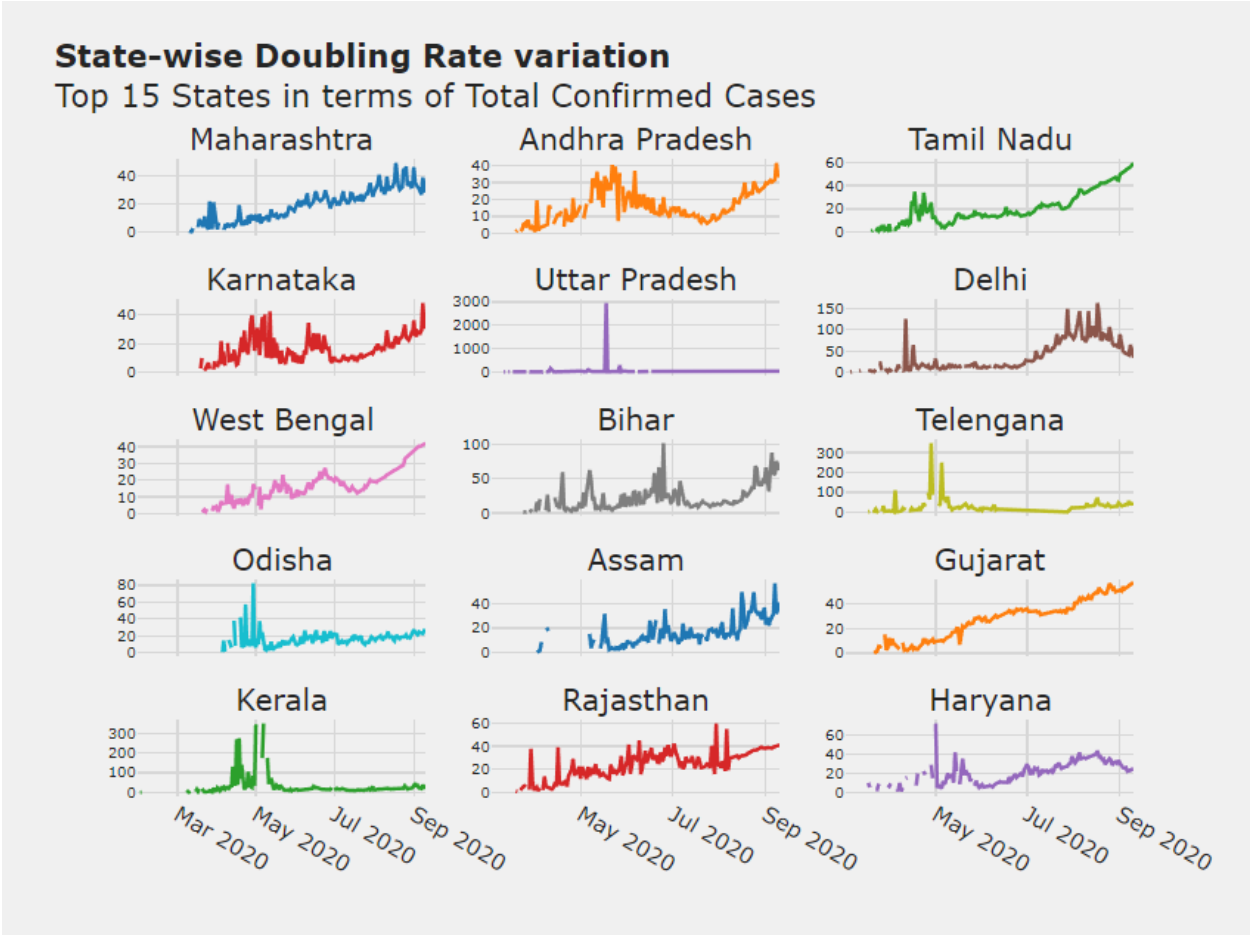
To quantify the is-risk curves dimension, CFD simulation of a sneeze was carried out. Human exhalation such as coughing, sneezing and breathing can be considered as instantaneous airflows produced from a single source with a quite symmetrical and conical geometry. Although coughing and sneezing have gained much attention as potential, explosive sources of infectious aerosols, these are relatively rare events during daily life

PRESENTATIONS AND VISUALIZATION






India's Daily Doubling rate variation
Days
40
35
30
25
20
15
10
5
0
Feb 2020
Mar 2020
Apr 2020
May 2020
Jun 2020
Jul 2020
Aug 2020
Sep 2020


State-wise Doubling Rate variation
Top 15 States in terms of Total Confirmed Cases
Maharashtra
Andhra Pradesh
Tamil Nadu
Karnataka
Uttar Pradesh
Delhi
West Bengal
Bihar
Telengana
Odisha
Assam
Gujarat
Kerala
Rajasthan
Haryana
Mar 2020
May 2020
Jul 2020
Sep 2020

ROLES

Here in this project our group mainly focused on data analysis and visualization of the data set for exploring purposes. We have implemented various techniques such as logistic regression, decision tree, multiple regression. Here, in our project data mining algorithms were implemented by programming language python with the help of special purpose libraries.

SCHEDULE

DATE	TASKS TO BE COMPLETED
03/01/2022	Tasks completed by chosen date
17/01/2022	Tasks completed by final report
17/01/2022	Tasks completed by class presentation

BIBLIOGRAPHY

1. ncbi.nlm.nih.gov/pmc/articles/PMC7306186/
2. ncbi.nlm.nih.gov/pmc/articles/PMC7102847/
3. https://my.clevelandclinic.org/health/diseases/21214-coronavirus-covid-19
4. link.springer.com/article/10.1007/s42979-020-00216-w
5. https://www.kaggle.com/sudalairajkumar/covid19-in-india
6. https://www.kaggle.com/imdevskp/covid19-corona-virus-india-dataset