

# CS685A Data Mining : Assignment-1

Harshvardhan Pratap Singh (20111410), hrshsengar20@iitk.ac.in

September 25, 2020

## Abstract

In this assignment we studied COVID-19 pandemic data. This is a brief report for the analysis of results and conclusions.

## 1 Introduction

Purpose of this report is analysis and mining of covid data. Finding pattern in it, how it is increasing countrywide and spreading. What are some common pattern between the districts and what is the growth of it with respect to each district and state over the Time by analysing it weekly, monthly, and overall. what are the district where situation is in control and cases are not increasing very fast and where government has to give special attention in order to control its spread. What are some hotspots that can be declared as containment zone and in which districts borders should be seized.

## 2 Analysis

**neighbor-districts-modified.json** Looking at the neighbor-districts-modified.json, people can get to know what are the districts in covid portal that are sharing boundaries and can directly affect each other, because they are neighbours.

**cases-time.csv** looking at cases-week.csv, cases-month.csv, cases-overall.csv people can get to know about number of covid cases for each district in a time series manner weekly, month and overall over the period of time 15 March 2020 to 5 September 2020. This can be used to monitor where cases are increasing weekly, monthly very fast and require special attention from government and in deciding the rules for lockdown. And policies can be decided such as if in some district cases are increasing drastically.

**edge-graph.csv** edge-graph.csv tells us that whether districts are connected with each other or not.

**neighbor-time.csv** neighbor-week.csv , neighbor-month.csv, neighbor-overall.csv are giving information about the neighbour districts of a particular districts. Entry neighbormean is giving information about what are the average number of cases of all the neighbouring districts weekly, monthly and overall and entry neighborstdev is giving information about how much every neighbouring districts cases are close to average number of cases.

That means whether cases are distributed equally among the neighbours or some neighbour is red zone containment zone area.

**state-time.csv** state-week.csv, state-month.csv, state-overall.csv are giving information about the state and other districts in the state of a particular districts. With this information we can get to know what is state mean and based upon that can decide that whether this district is coldspot in its state or not and also by looking at statestddev, if statestddev value is high that means cases are not equally distributed and some districts in this state have much more cases than other districts and they can affect their neighbours in future so these should be lockdown until situation improves there.

**zscore-time.csv** zscore-week.csv, zscore-month.csv, zscore-overall.csv gives information with respect to number of cases in a district against its neighbours and other districts in state. It is a standard score and that's why it allows comparison of scores on different kinds of cases by standardising the distribution. Comparison between cases of two or more districts can be easily done through this score.

**method-spot-time.csv** method-spot-week.csv, method-spot-month.csv, method-spot-overall.csv are giving information about hotspot and coldspot districts per week, per month, and overall in neighbourhood and state. This data can be used to know what cities to give special attention neighbourhood and state. And where lockdown can be less restrictive.

**top-time.csv** top-week.csv, top-month.csv, top-overall.csv this is giving information about hotspot and coldspot districts per week, per month and overall countrywide for top five hotspot districts and coldspot districts.

### 3 Observations

Statewise top 5 hotspots [Lucknow, Srinagar, Indore, Ahmedabad, Thiruvananthapuram] and top 5 coldspots [Kargil, Diu, Krishna, Kinnaur, Wayanad]

Neighbourhood wise top 5 hotspots [Davanagere, Prayagraj, Raipur, Jhansi, Gwalior] and top 5 coldspots [Sambhal, West Jaintia Hills, Siang, Osmanabad, LongLeng]

top-overall.csv gives information about top 5 hotspots and coldspots.

### 4 Conclusions

We can pick top districts where situation is not under control and then work on them in finding what can be possible problems whether health infrastructure is an issue or awareness is an issue or whether essentials community act is following properly there or not or districts are economically weaker and government should provide supply of sanitizer and mask. Through this one can monitor that rules and policies are implemented properly or not. What are the districts because of which their neighbours are getting affected and cases are increasing rapidly and should be controlled and also what state borders should be seized and require more doctors. By analysing output files of each csv file one can get to know many information like in which city travel should be avoided completely and much more.