# Data Mining

# Classification

Team 10: Jeevan Rai & Abhilash Narayanan

**Executive Summary:**

This report is on analyzing various aspects of social structures that were collected during the COVID-19 pandemic. These social structures are referred as variables for the dataset COVID-19\_cases\_plus\_census that go through various data mining processes to discover any possible relationships with the number of cases and/or deaths for various geographic regions. There could be many possible reasons that could depict the observed confirmed cases and/or deaths for a particular region during the pandemic. Discovering such insightful relationships and building an effective predictive model are the two primary problems described in the report. From the general public to leaders of various government and non-government institutions, the information described in the report can be beneficial in many ways.

**Table of Contents**

[Data Mining 1](#_gjdgxs)

[**Classification 1**](#_vmplmdk0au9m)

[1 Business Understanding 3](#_30j0zll)

[2 Data Preparation 3](#_1fob9te)

[3 Modeling 5](#_2s8eyo1)

[3 Evaluation 5](#_1pxezwc)

[**4 Deployment 5**](#_5tqp5cuok6f1)

[5 List of References 6](#_49x2ik5)

[6 Appendix 6](#_2p2csry)

# Business Understanding

A widespread disease COVID-19 (also known as coronavirus disease 2019) started at the end of 2019, and it quickly spread throughout the entire world impacting every aspect of human society. It was first identified in December 2019 in Wuhan district in China. Since then, there have been several kinds of studies conducted on the impact of this pandemic. The data on those studies are available for the general public. Among those datasets, we will be looking at four different datasets. The primary focus of the analysis is based on understanding the impact of this pandemic on various aspects of human society all around the world from 2019. These policies were primarily implemented to enforce “social distancing” amongst people. Social distancing involves measures taken to reduce close contact between individuals to slow the spread of infectious diseases such as COVID-19. By implementing measures like social distancing, mask-wearing, and hygiene practices, the goal is to spread out the number of cases over a longer period, resulting in a flatter curve. Additionally, data mining is performed on these datasets to understand any existing relationships between entities. This can be used to predict recurrence of similar pandemic in the near future and the consequences of such situations. To classify Texas counties as low or high risk in terms of COVID-19 deaths (deaths per 10,000), several classification models are being used. The performance of these models were evaluated, and compared against one another to identify the best performing model.

# Data Preparation

The report uses two datasets “COVID-19\_Global Mobility” and “COVID-19\_cases\_plus\_Census”. The primary objective of the report is to use the datasets to classify the US counties as low or high risk in terms of the COVID-19 confirmed cases and death rates. Since the dataset COVID-19\_cases\_plus\_Census only contains records for the date 01/19/2021, the observations recorded on the same date were extracted from the dataset COVID-19\_Global Mobility. The report focuses only on Texas counties. So, the observations for Texas counties were extracted from both datasets. Both of these datasets are then merged. The final merged dataset has observations on 204 Texas counties.

The merged dataset was checked for missing values. There were feature columns with various numbers of missing values (see below). Since most of the features have too many missing values, they were replaced with the corresponding average values.

| **Features** | **Number of missing values** |
| --- | --- |
| retail\_and\_recreation\_percent\_change\_from\_baseline | 91 |
| grocery\_and\_pharmacy\_percent\_change\_from\_baseline | 104 |
| parks\_percent\_change\_from\_baseline | 155 |
| transit\_stations\_percent\_change\_from\_baseline | 114 |
| workplaces\_percent\_change\_from\_baseline | 9 |
| residential\_percent\_change\_from\_baseline | 93 |

Features with missing values

Next, all the irrelevant or less significant features were dropped from the dataset. See below for basic statistics on the final dataset. These features were used to perform all the classifications.

| **Metric** | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** |
| --- | --- | --- | --- | --- | --- | --- |
| Total Population | 600 | 12,635 | 24,898 | 133,668 | 66,671 | 4,525,519 |
| Hispanic Population | 0.03454 | 0.17705 | 0.25266 | 0.34352 | 0.48529 | 0.99185 |
| Black Population | 0 | 0.01504 | 0.05254 | 0.07161 | 0.10209 | 0.33743 |
| Male (50 and above) | 0.1067 | 0.1564 | 0.1857 | 0.1854 | 0.2116 | 0.3057 |
| Female (50 and above) | 0.1192 | 0.1698 | 0.1949 | 0.2004 | 0.2305 | 0.3554 |
| Income ($50K - $100K) | 0.04964 | 0.09331 | 0.10466 | 0.10275 | 0.11199 | 0.14785 |
| Rent > 50% Income | 0 | 0.01207 | 0.01587 | 0.01745 | 0.0218 | 0.06957 |
| Commute (work outside home) | 0.466 | 0.7021 | 0.7697 | 0.7608 | 0.8285 | 0.9794 |
| Worked at Home | 0 | 0.009951 | 0.014673 | 0.015404 | 0.018924 | 0.045182 |
| Transit Stations % Change from Baseline | -61 | -11.06 | -11.06 | -11.06 | -11.06 | 76 |
| Workplaces % Change from Baseline | -49 | -27 | -21 | -22.69 | -17 | -8 |
| Cases per 10,000 | 231.1 | 590 | 754.7 | 785.7 | 942.4 | 1829 |
| Deaths per 10,000 | 3.309 | 12.016 | 16.478 | 17.705 | 22.083 | 54.608 |
| Death per Case | 0.003789 | 0.015099 | 0.02219 | 0.023541 | 0.029217 | 0.09322 |

Basic statistics on final dataset

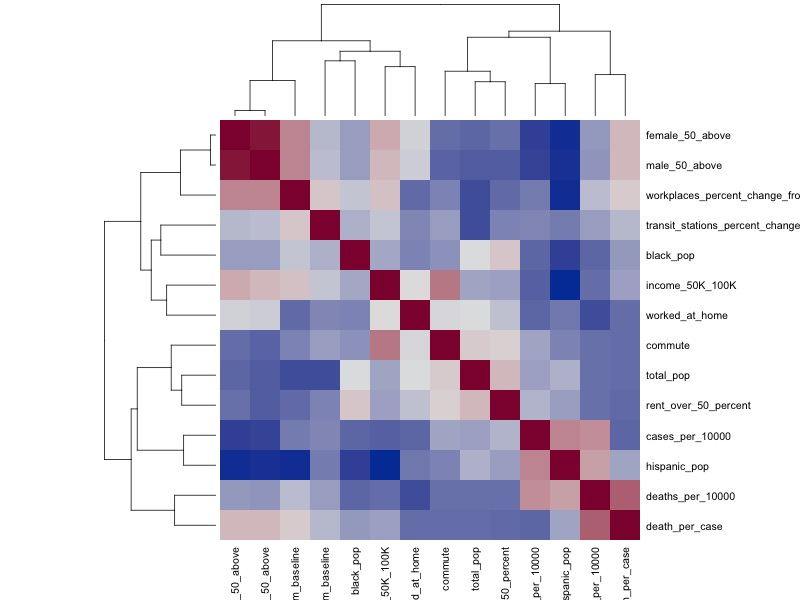
In the table above, the values for the features cases\_per\_10000, deaths\_per\_10000, and death\_per\_case are the number of people per 10,000 out of the total population. The rest of the features (except for “Transit Stations % Change from Baseline” and “Workplaces % Change from Baseline”) are normalized values as their counts were divided by the total population. These feature processing operations were performed to normalize each feature.

All the Texas counties will be evaluated based on deaths per 10,000. Furthermore, the threshold for the counties to be considered as high is 16. That means the counties are classified as high if their values for deaths\_per\_10000 is greater than 16. This threshold value was strategically chosen as it created a very well-balanced training and testing datasets for the counties that were chosen as training and testing datasets. Based on the counties selected for the training dataset, the class labels for TRUE (high) and FALSE were 11 and 12 for the training dataset respectively. Similarly, the class labels were 88 as FALSE and 93 as TRUE for the testing dataset. There were reasons why the classes were defined this way.

Counties for training dataset: (DFW metropolitan: collin, dallas, denton, ellis, johnson, kaufman, parker, rockwall, tarrant, wise), (from project 2: bell, cameron, el paso, hidalgo, nueces), terry, martin, lubbock, wichita, cherokee, hale, maverick, parmer.

# Modeling

Before beginning to perform classification, the chosen features of the final dataset were evaluated in a correlation plot to get a glimpse of their correlations with deaths\_per\_10000, cases\_per\_10000, and death\_per\_case.



Correlation heat map for dataset features

Looking at the above plot, we can see that the feature “worked\_at\_home” seems to have the highest positive correlation to the target variable “deaths\_per\_10000”. At the moment, this is based on the dataset with all the counties included. This assessment will be validated throughout the experimentation of various classification models throughout the report.

#### 

Counties with high deaths per 10,000

The above heat map shows all the counties

# Evaluation

# Deployment

# List of References

# Appendix