***COVID19 Death rate Analysis using Data Mining Techniques***

***SUBMITTED BY***

***AKSHAY V(300169348)***

***JOSEPH CHARLES (300166618)***

***PROFESSOR***

***LONGLONG HUANG***

***COMP 431***

***17/Dec/2020***

**Abstract**

This project is about understanding and predicting the pattern of the spread of Covid-19 in different parts of the United States of America. We study how population and urbanization of region studied by The National Center for Health Statistics can be cause of the quick spread of the disease. The focus of this project is the effect on deathrate in these regions due to the urbanization and the density of population. We also study effect of the confirmed Covid-19 cases in all different region and how this interacts with urbanization of the area to affect the deathrate.

**Introduction**

December 8th, 2019 a patient in the city of Wuhan was admitted in a hospital for Pneumonia like symptoms (Kantis, Kiernan, & Bardi, 2020). Later in a months’ time neighboring countries started reporting similar cases. And the same occurred in US, January 1st, 2020 first case of COVID-19 was reported (Schumaker, 2020). Following this most countries including U.S declare the COVID-19 as National emergency on March 13th, 2020 (Schumaker, 2020) imposed strict lockdown which forced many businesses, educational institution, and other organization to shut down its operations temporarily. Lockdown and use of PPE’s in public places might have helped authorities control the spread of the virus but still it was spreading fast. Since then, this Virus has been under microscope and various studies has been conducted on it. Various theories have been proposed about the Virus and lot of the details about the Virus still seem to be a mystery. Thousands of deaths were reported in few months and most global organization confirmed Vaccine could be the only way the break this chain. In this report we will analyses the spread of the disease in the U.S based on the Urbanization, population the area and the confirmed cases and its effect on the COVID-19 using various Data mining techniques.

**Hypothesis**

Urbanization and Population density influences the number of COVID-19 confirmed cases in a region and thus impacts the Deathrate.

**About the Dataset**

The dataset was collected from Kaggle.com which is has up to date information of the COVID-19 cases in the U.S. The Dataset contains many information’s such as State, County, FIPS-code etc., which we chose to ignore for this analysis. Our predictor variables are **Total Population, NCHS Urbanization** and **confirmed cases**. **Deathrate** is the response variable. Dataset after removing the empty rows have a total of 3196 rows.

**Tools Used**

The Data collected is in csv format for which Microsoft Excel was used to filter few empty rows that was present in the dataset. R program was used for rest of the preprocessing and analysis. In R program we used few data mining and machine learning packages like Tidyverse, GGplot2, Caret, Rweka and C50.

**Data Preprocessing**

The Dataset had many empty values which we felt should be taken care or it can affect the outcome of the project. We used median values of each columns for as the value of empty cells for the same. Variable like NCHS Urbanization was transformed as factors. And some variables like State, county etc., had to be ignored for this project. Deathrate column was numeric which was transformed to three-layer class “Low”, “Medium” and “High”. We used K-nearest neighbor clustering technique to categorize deathrate value into three cluster. Based on centroid value of each three clusters we chose an approximate value of 70, any value that falls below this value will be categorized as “Low”, all value that falls between 71 to 150 will be categorized as “Medium”, and all value above 150 will be considered as “High”.

**Methodology**

**Rule Based Classifier**

Rule based classifier algorithm classifies record based on “if, then” rule. The classifier has two pats, one is antecedent or condition and the rule consequent (TAN, STEINBACH, & KUMAR, 2006). The IF part is the antecedent. Then part is the rule consequent.

**Decision Tree**

This algorithm uses recursive binary partitioning technique starting from the root node by splitting based on purity measure of the class using Impurity calculation techniques like Gini Index, Entropy and Error classification. It further expands to branches called as Internal Node using same impurity measuring technique resulting to leaf Node which is the goal or prediction.

**Random Forest**

Random forest is similar approach like decision tree, but the difference is it generates recursive random vectors of the training data and build decision tree for every vector and combines all the decision tree to create the final model. It does have some disadvantage as it generates many data frame from the training dataset and process each of them it consumes lot of processing time and space. It is effective with large datasets since it uses random sampling technique.

# Bibliography

Kantis, C., Kiernan, S., & Bardi, J. S. (2020, October 09). *UPDATED: Timeline of the Coronavirus*. Retrieved from Think global health: https://www.thinkglobalhealth.org/article/updated-timeline-coronavirus

Schumaker, E. (2020, October 22). *Timeline: How coronavirus got started*. Retrieved from ABC news: https://abcnews.go.com/Health/timeline-coronavirus-started/story?id=69435165

TAN, P.-N., STEINBACH, M., & KUMAR, V. (2006). *Introduction to Data Mining.* Boston: Pearson Education, Inc.

**Dataset Head**



















