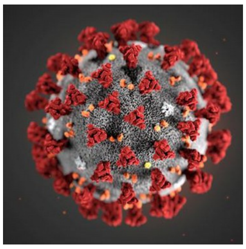

Data Science for COVID-19
South Korea

1. Introduction

Initially, it is convenient to define data science as the combination of multiples that include business strategy, statistics and computer science with the intention of extracting information and scientific experience from the data by conducting controlled experiments whose results allow us to evaluate possible responses to real life problems.

The goal of any data science project is to gain valuable insight based on expectations in order to make better, better decisions.

It is not only the responsibility of data scientists to define the objectives to be achieved for a project. This requires knowledge, mastery and experience on the subject matter of analysis. In such a way that the agents involved are varied from a perspective of problem solving and conflict scenarios.

The main idea is to look for patterns from statistics, which as we well know is a mathematical discipline with different degrees of complexity (advanced). Initially in this project we will consider basic statistics as an elemental analysis technique.
Just as statistics is of great help, it is no less so, the science of informatics as far as programming is concerned, generating powerful algorithms capable of discovering the knowledge that is hidden in the data.

In the present work, we will mainly investigate what we currently know as COVID-19 influenza, a deadly epidemic that covers a large part of the planet.

The research will focus on open data provided by institutions in South Korea

Obviously, it must be recognized that all this has gone very fast, in such a way that the normalization of data together with the awkwardness of some governments has not allowed to normalize the flow of information to carry out studies and machine learning for its real analysis to achieve predictions. Effective.

The current epidemic represents a great challenge for the governments of the world whose fundamental responsibility is to implement adequate health policies and improve the quality of life of the population.

Note:

Please consider the current situation the planet is going through. This work does not have the primary objective of graduating the course, but rather wants to contribute from data science research and add value and knowledge.

It has been tried to introduce in the subject due to the need to integrate knowledge and collective in order to look for a global solution to the problem.

COVID-19 has been a fast-moving disease, now we have the opportunity to help make a better world by understanding and investigating it by shedding light to promptly resolve this pandemic.

Tribute

A tribute to the people who have passed away and their relatives who will keep their memories in their memory.

1.2-Definition of the problem

Currently the main concern of humanity is to stop the advance of the COVID-19 virus, which is a real danger to humanity.

The world health organization has officially declared the danger of the spread of this virus at the PANDEMIC level.

The COVID-19 infection surprised the world with its rapid spread and has had a major impact on the lives of billions of people.

Society as a whole, while becoming aware of the danger, needs both reliable and secure sources of information.

The Internet continues to be the primary source for collecting information globally.

A highly globalized society implies a collective dependence on information.

Specifically in the area of datascience, having reliable and safe sources for creating classification, prediction and description models, which implies having contrasted, updated and easily accessible data.

What is the usefulness of collecting information? It will mainly be useful in that the quality of the information can be evaluated in order to achieve predictive models regarding the outbreak of disease.

Scientific information in this regard is not extensive, resources are scattered. The institutions that are mainly providing elaborated or semi-elaborated datasets are Universities, research centers or health ministries.

The main reason for this research is to explore the existing data from contrasted sources in order to visualize the behavior of the virus in a determined time space (January-present) this year.

The question is: Where is the data to help us understand the current situation? Next, from the available data, what happens with the classification: gender, age, territorial location, nationality, etc. Current statistics tell us of deaths, recovered and infected, but a more detailed description is lacking in terms of the information offered, the state monopolizes the information and does not share the source of the information with society. All this leads us to think about the degree of veracity in terms of data manipulation.

The disaggregation of data adds value to it from the point of view of research.

Today more than ever, this information is essential to start the questions:

Does it affect women more aggressively than men and in what amount?
What is the age that affects with greater virulence?
Do women recover faster than men?
What is the most frequent age in women to contract the infection?

In short, an absence of centralized information available for use is detected quickly and legibly.
Certain public and private universities have sufficient resources that through their research departments manage to publish scientific information.

1.3-Interested

Those interested in this diverse work include students, researchers, health personnel as well as managers of medical centers and professionals involved in decision-making in a health center.

It is also necessary to include politicians and managers of the state bureaucracy, who need to have a different and disinterested vision of events.

2.-Data acquisition

In the problem statement it is explained that one of the main difficulties is obtaining disaggregated data from the information that government agencies report to the population, these are mainly infected, recovered and deceased people.

The main idea is to integrate data as disaggregated as possible in order to understand patterns that to date do not allow a specific vision of the situation.

The data will be acquired taking into account those who present us with the largest number of disaggregation as the main requirement, with the least two variables to be considered. For example, we know that COVID-19 affects more elderly people with respiratory diseases mainly.
Another requirement in the acquisition of data is your temporality, its actuality is privileged, it is not older than five days at most.

We will address the acquisition of data from various sources considering public data and / or open sources.

The following are considered as data acquisition sources:

DS4C: Data Science for COVID-19 in South Korea KCDC (Korea Centers for Disease Control and Prevention) / KOSTAT (Korea Statistics) / KMA (Korea Meteorological Administration) / NAVER DataLab.

Research reference:

a) https://www.kaggle.com/kimjihoo/coronavirusdataset/activity

b) https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

c) https://qap.ecdc.europa.eu/public/extensions/COVID-19/COVID-19.html

d)https://github.com/
e)https://covid19.isciii.es/
f)https://opendatawatch.com/what-is-being-said/data-in-the-time-of-covid-19/

g)https://www.kaggle.com/search?q=covid-19+date%3A1+datasetSize%3Asmall+datasetSize%3Amedium+datasetFileTypes%3Acsv+datasetLicense%3AOther+datasetLicense%3ACommercial+tag%3A%22public+health%22