PROYECTO FINAL

Minería de datos

SOBRE EL TRABAJO

Este documento muestra la investigación, recolección y análisis de datos sobre COVID-19.

Luis Antonio Vásquez Tiu
Dimmitry Benjamyn Estuardo Recinos Ovalle
Wilson Estuardo Nájera Caxaj

Contenido

Introducción

El siguiente proyecto está enfocado a la predicción, análisis y comprensión del comportamiento de la pandemia Covid19 a nivel mundial y nacional, usando un set de datos que comprenden desde marzo de 2020 hasta diciembre del mismo año. El objetivo del proyecto es lograr identificar puntos de colapso en hospitales con la predicción de muertes en diferentes fechas y países, esto se logra gracias a la gran cantidad de datos obtenidas en una serie de investigaciones que permite hacer predicciones de muertes, modelos de predicción y correlación sobre muertes y casos confirmados y a su vez arboles de decisiones a nivel nacional que muestran de manera detallada y gráfica la relación que puede existir sobre casos y muertes dependiendo de los diferentes países

Descripción del proyecto

El proyecto consta del análisis de una base de datos sobre contagios de COVID-19 a nivel mundial.

La primera parte del análisis está enfocado al incremento de muertes por covid-19 con relación al tiempo transcurrido, esto quiere decir que, con la base de datos y con algoritmos como k-vecinos se podrá comparar el número de muertes con las fechas que se programen en el algoritmo, para luego ser comparadas con una fecha distinta, poder hacer un análisis de incremento o disminución de muertes. Así como la creación de gráficas sobre las muertes sobre diferentes fechas establecidas.

La segunda parte del proyecto consta de la creación e interpretación de un árbol de decisiones que permita visualizar de manera clara y gráfica la información sobre muertes en algún determinado país o región. Para esto es necesario utilizar el algoritmo de árbol de decisiones.

La tercera parte del proyecto consta de un modelo de predicción a través de regresión lineal que permita ver el comportamiento de cantidad de muertes en relación con los casos confirmados de COVID-19. También, la creación de una gráfica de dispersión con la cual pueda ser más fácil entender el modelo.

El archivo necesario para la construcción de este proyecto es el de la base de datos de covid19 a nivel mundial del 2020. Esta base de datos cuenta con los datos necesarios para trabajar la comparación de las muertes correspondientes a fechas y otras variables.

La entrega consta de lo siguiente:

Documentación: Se entregará un documento con las especificaciones del proyecto

Capturas: Capturas de la implementación de los algoritmos de análisis de datos sobre la base de datos anteriormente mencionada.

Archivos: Archivos trabajados para el proyecto como: archivos de RapidMiner y NoteBook de Python.

Objetivos

General

Analizar la estadística de tasa de mortalidad por covid-19 con relación a diferentes variables con el uso de una base de datos de COVID-19 recabados en el año 2020.

Específicos

- Utilizar el algoritmo k-vecinos cercanos para comprobar la relación entre el número de muertes de una fecha establecida por países.
- Crear un árbol de decisión que permita visualizar la cantidad de muertes por regiones.
- Crear una regresión lineal que permita ver el comportamiento de muertes por casos confirmados utilizando Python.

Antecedentes

Página del gobierno de Guatemala

Esta es una página desarrollada por el gobierno de Guatemala para el análisis de los casos y muertes de COVID-19 en el país.

En esta se puede encontrar datos sobre casos y muertes a nivel nacional. Así como gráficas, porcentajes y tasas sobre municipios y departamentos mostrándolos por colores.

Parte 1: Incremento de muertes entorno al tiempo

El siguiente modelo fue desarrollado con el fin de dar una predicción sobre muertes de covid19 en fechas próximas haciendo uso de información como el país, fecha y numero de contagiados de un set de entrenamiento que contiene dicha información de la mayoría de los países de todo el mundo.

Herramientas utilizadas

- RapidMiner
- K-nn

Desarrollo del modelo

Se hizo uso de un set de entrenamiento que contiene datos de contagios, muertes por país y por fecha, con este set de entrenamiento se logra obtener un valor de predicción alto dependiendo de la fecha, cantidad de contagiados y país.

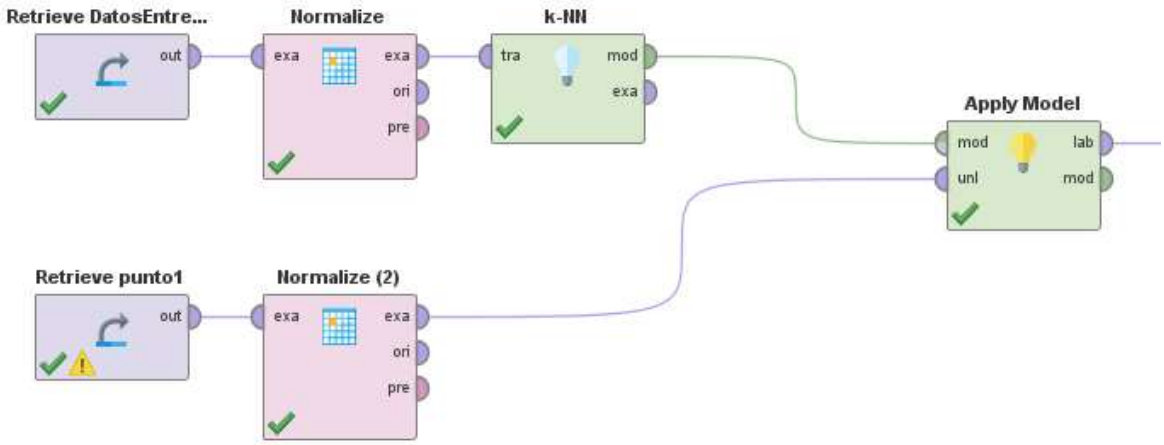
Row No.	deaths	dateRep	day	month	year	cases	countriesAn...	geoId	countryterrit...	popData2019	continentExp
1	6	Dec 14, 2020	14	12	2020	746	Afghanistan	AF	AFG	38041757	Asia
2	8	Dec 13, 2020	13	12	2020	298	Afghanistan	AF	AFG	38041757	Asia
3	11	Dec 12, 2020	12	12	2020	113	Afghanistan	AF	AFG	38041757	Asia
4	10	Dec 11, 2020	11	12	2020	63	Afghanistan	AF	AFG	38041757	Asia
5	18	Dec 10, 2020	10	12	2020	202	Afghanistan	AF	AFG	38041757	Asia
6	13	Dec 9, 2020	9	12	2020	135	Afghanistan	AF	AFG	38041757	Asia
7	8	Dec 8, 2020	8	12	2020	200	Afghanistan	AF	AFG	38041757	Asia
8	26	Dec 7, 2020	7	12	2020	210	Afghanistan	AF	AFG	38041757	Asia
9	10	Dec 6, 2020	6	12	2020	234	Afghanistan	AF	AFG	38041757	Asia
10	18	Dec 5, 2020	5	12	2020	235	Afghanistan	AF	AFG	38041757	Asia
11	5	Dec 4, 2020	4	12	2020	119	Afghanistan	AF	AFG	38041757	Asia
12	19	Dec 3, 2020	3	12	2020	282	Afghanistan	AF	AFG	38041757	Asia
13	48	Dec 2, 2020	2	12	2020	400	Afghanistan	AF	AFG	38041757	Asia
14	11	Dec 1, 2020	1	12	2020	272	Afghanistan	AF	AFG	38041757	Asia
15	0	Nov 30, 2020	30	11	2020	0	Afghanistan	AF	AFG	38041757	Asia
16	11	Nov 29, 2020	29	11	2020	228	Afghanistan	AF	AFG	38041757	Asia
17	15	Nov 28, 2020	28	11	2020	214	Afghanistan	AF	AFG	38041757	Asia
18	0	Nov 27, 2020	27	11	2020	0	Afghanistan	AF	AFG	38041757	Asia
19	12	Nov 26, 2020	26	11	2020	200	Afghanistan	AF	AFG	38041757	Asia
20	13	Nov 25, 2020	25	11	2020	185	Afghanistan	AF	AFG	38041757	Asia
21	17	Nov 24, 2020	24	11	2020	246	Afghanistan	AF	AFG	38041757	Asia
22	8	Nov 23, 2020	23	11	2020	252	Afghanistan	AF	AFG	38041757	Asia

Se realizó un set de datos que contenía fechas, cantidad de contagiados y países para obtener una predicción de muertes en dichos datos

Row No.	deaths	dateRep	day	month	year	cases	countriesAn...	geoId	countryterrit...	popData2019	continentExp
1	?	Dec 14, 2020	14	12	2020	100	Afghanistan	AF	AFG	38041757	Asia
2	?	Dec 13, 2020	13	12	2020	200	Afghanistan	AF	AFG	38041757	Asia
3	?	Jun 27, 2020	27	6	2020	12	Ghana	GH	GHA	30417858	Africa
4	?	Jun 26, 2020	26	6	2020	0	Ghana	GH	GHA	30417858	Africa
5	?	Jun 25, 2020	25	6	2020	12	Ghana	GH	GHA	30417858	Africa
6	?	Aug 7, 2020	7	8	2020	1	Liberia	LR	LBR	4937374	Africa
7	?	Aug 6, 2020	6	8	2020	12	Philippines	PH	PHL	108116622	Asia
8	?	Aug 6, 2020	6	8	2020	10	Guatemala	GT	GTM	17581476	America

Estructura del modelo

Se hizo uso de K-nn como algoritmo de predicción, añadido a eso el data set con datos de fechas, contagios y país al que buscamos la predicción de muertes con ayuda de RapidMiner



Resultados del modelo

Resultados de predicción de muertes del modelo implementado

Row No.	deaths	prediction(d...	day	month	year	cases	popData2019	dateRep	countriesAn...	geoId	countryterrit...	continentExp
1	?	62.138	-0.163	1.504	0	0.795	0.026	Dec 14, 2020	Afghanistan	AF	AFG	Asia
2	?	325.441	-0.271	1.504	0	2.200	0.026	Dec 13, 2020	Afghanistan	AF	AFG	Asia
3	?	1.837	1.248	-0.903	0	-0.441	-0.222	Jun 27, 2020	Ghana	GH	GHA	Africa
4	?	0	1.140	-0.903	0	-0.609	-0.222	Jun 26, 2020	Ghana	GH	GHA	Africa
5	?	0.385	1.031	-0.903	0	-0.441	-0.222	Jun 25, 2020	Ghana	GH	GHA	Africa
6	?	0	-0.923	-0.100	0	-0.595	-1.053	Aug 7, 2020	Liberia	LR	LBR	Africa
7	?	28.703	-1.031	-0.100	0	-0.441	2.309	Aug 6, 2020	Philippines	PH	PHL	Asia
8	?	7.898	-1.031	-0.100	0	-0.469	-0.641	Aug 6, 2020	Guatemala	GT	GTM	America

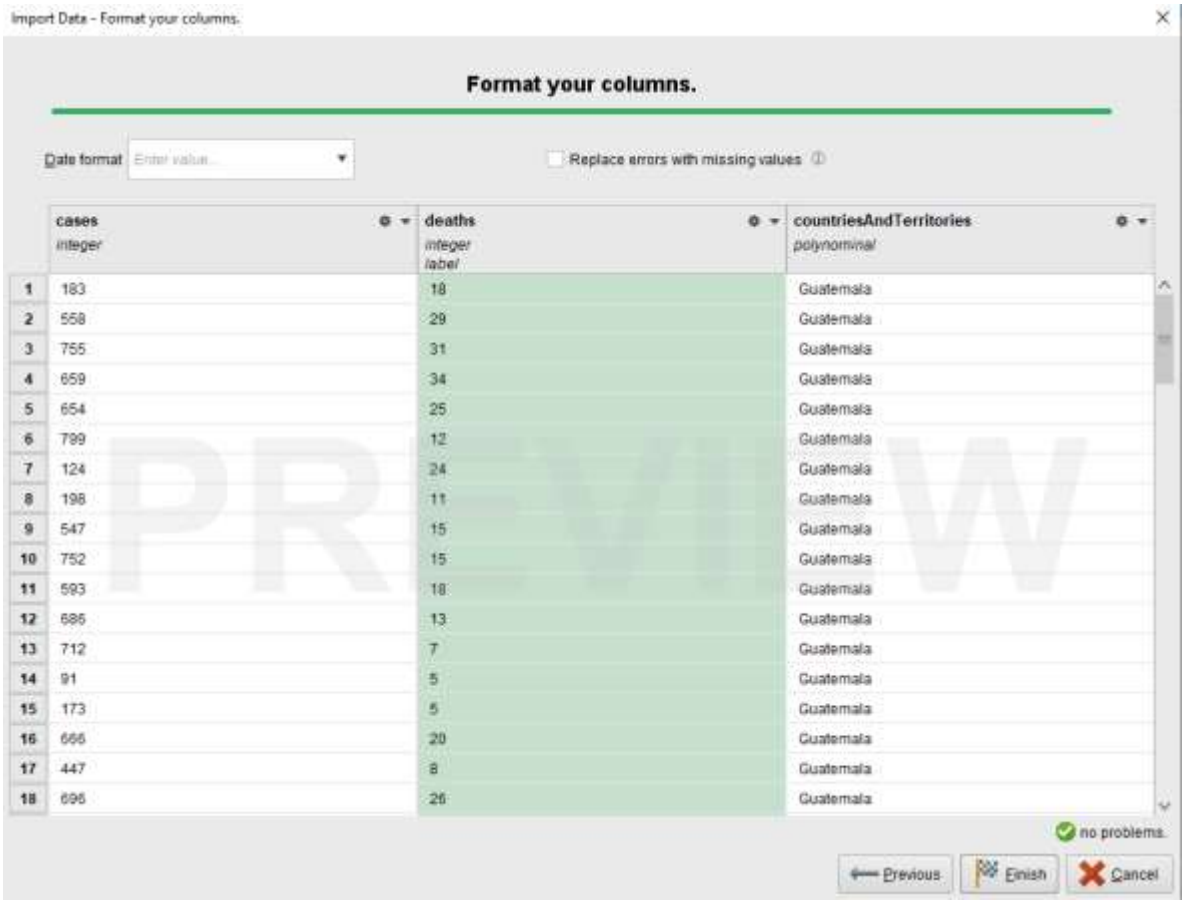
Conclusiones

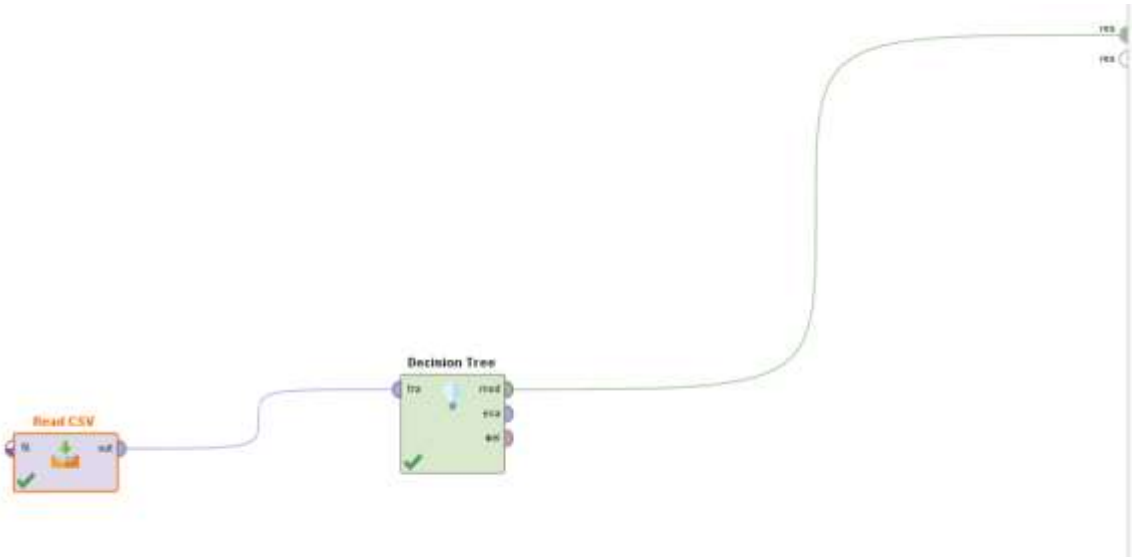
Gracias al set de entrenamiento podemos obtener cambios significativos en la cantidad de muertes al modificar la fecha, el país o el número de contagios, el modelo muestra mayor o menor cantidad de muertes dependiendo del país o de la fecha en la que se busca la predicción.

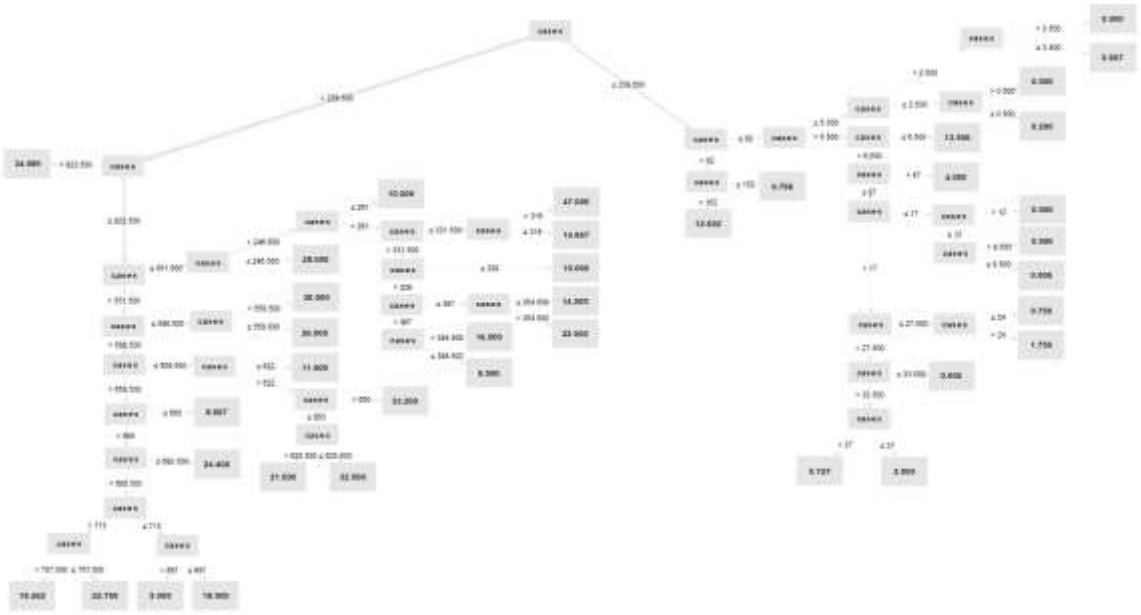
Gracias al set de entrenamiento se logra obtener una predicción útil al poder buscar por fecha o por país.

Parte 2: Árbol de decisión de muerte en relación a región

Se realizo la carga del documento csv para posteriormente colocar su conexión con el árbol de decisiones.







Conclusiones

Gracias a los datos obtuvimos podemos observar que mientras mayor sea el número de casos podremos encontrar un incremento en el número de muertes aplicado a la región de Guatemala**.**

Parte 3: Regresión lineal casos-muertes

La tercera parte del proyecto consta de un modelo de predicción a través de regresión lineal que permita ver el comportamiento de cantidad de muertes en relación a los casos confirmados de COVID-19. También, la creación de una gráfica de dispersión con la cual pueda ser más fácil entender el modelo.

El modelo de predicción es desarrollado utilizando una regresión lineal entre las variables "cases" y "deaths" de la base de datos, los cuales son casos y muertes, respectivamente.

Para el desarrollo de la regresión lineal fue necesario la utilización del lenguaje de programación Python con las siguientes librerías:

- Pandas
- Numpy
- Matplotlib
- Stastsmodel

Desarrollo:

1. Importar las librerías y el archivo necesario
- Primero se importan las librerías necesarias y el archivo como base de datos

```
In [2]: # librerias necesarias
        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
```

```
In [3]: # importar el archivo
        path = "/home/luii/Documents/Mineria de datos"
        file = "proyecto Final/DatosCovid2020.csv"
        data = pd.read_csv(path + "/" + file)
        data.head(10)
```

Creación de relación de variables

- Con el uso de la librería "statsmodels" se crea la relación entre las dos variables casos y muertes y con el uso de "rsquared_adj" se crea la validación del modelo

```
: ml = smf.ols(formula = "cases~deaths", data=data).fit()
```

```
: ml.params
```

```
: Intercept    154.312871
  deaths        38.411694
  dtype: float64
```

```
: # validacion del modelo
  # coeficiente de determinacion ajustado (R^2 ajustado)
  # para cero no hay relacion 1 si hay relacion
  validacion = ml.rsquared_adj
  print("El modelo tiene una presicion de: " + str(round(validacion*100,2)) + "%")
```

```
El modelo tiene una presicion de: 55.29%
```

Validación matemática

- El uso de la opción “predict” se crea la validación matemática del incremento de muertes de la base de datos.

```
In [8]: # validacion matematica = prediccion de variable en la simulacion
        prediccion_ventas = ml.predict(pd.DataFrame(data["deaths"]))
        prediccion_ventas
```

```
Out[8]: 0        384.783034
        1        500.018115
        2        576.841503
        3        538.429809
        4        768.899971
                    ...
        61895    154.312871
        61896    192.724565
        61897    154.312871
        61898    154.312871
        61899    154.312871
        Length: 61900, dtype: float64
```

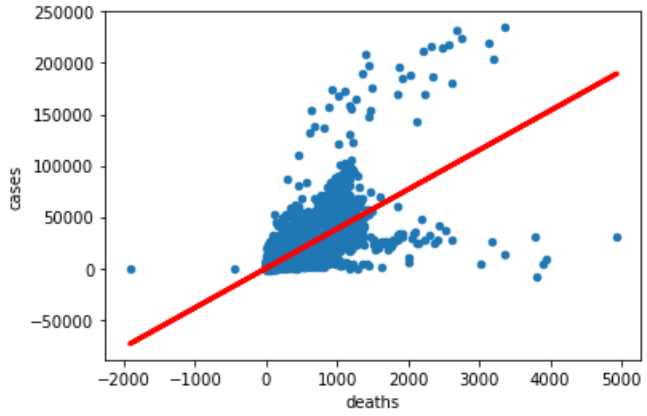
Creación de gráfica

- Con el uso de la librería “matplotlib” se crea una gráfica de dispersión para tener un mejor análisis de los datos.

In [9]:
```
# Impresión del modelo

%matplotlib inline
data.plot(kind="scatter", x="deaths", y="cases")
plt.plot(pd.DataFrame(data["deaths"]), prediccion_ventas, c="red", linewidth=3)
```

Out[9]: [<matplotlib.lines.Line2D at 0x7f50d2dc5d00>]



Conclusión del modelo

La relación entre los casos confirmados y las muertes por COVID-19, según la regresión lineal, es de 0.55, esto puede ocurrir por la fecha en la que se confirmaron los casos, puesto que al principio de la pandemia es sabido que murió más gente de la que muere hoy en día.

Podemos incluir otra variable para ver si es la que determina la cantidad de muertes, o crear una regresión lineal múltiple para ver el comportamiento del modelo con diferentes variables en conjunto.

Conclusión

- Según el set de entrenamiento y los datos obtenidos podemos apreciar que la cantidad de casos es una variable la cual tiene influencia sobre el número de muertes, ya que al momento de apreciar un aumento de casos también se aprecia un aumento de muertes en todos los países.

Recomendaciones

- Para la implementación de KNN es necesario reducir la cantidad de variables del set de entrenamiento y no sobrecargar el algoritmo.

- Para la implementación del árbol de decisiones es necesario reducirla base de datos a variables necesarias para el análisis de este.

- Para la implementación de la regresión lineal es necesario hacer uso de la validación con r^2 ajustado, para obtener todos los posibles resultados y no solo los de probabilidad más alta.

Anexos

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
datos = pd.read_csv("C:/Users/wilso/Desktop/GuatemalaDatosCovid2020.csv")
datos.head(10)
```

	Unnamed: 0	dateRep	day	month	year	cases	deaths	countriesAndTerritories	geoId	countryterritoryCode	popData2019	continentExp
0	23932	14/12/2020	14	12	2020	183	18	Guatemala	GT	GTM	17581476.0	America
1	23933	13/12/2020	13	12	2020	558	29	Guatemala	GT	GTM	17581476.0	America
2	23934	12/12/2020	12	12	2020	755	31	Guatemala	GT	GTM	17581476.0	America
3	23935	11/12/2020	11	12	2020	659	34	Guatemala	GT	GTM	17581476.0	America
4	23936	10/12/2020	10	12	2020	654	25	Guatemala	GT	GTM	17581476.0	America
5	23937	09/12/2020	9	12	2020	799	12	Guatemala	GT	GTM	17581476.0	America
6	23938	08/12/2020	8	12	2020	124	24	Guatemala	GT	GTM	17581476.0	America
7	23939	07/12/2020	7	12	2020	198	11	Guatemala	GT	GTM	17581476.0	America
8	23940	06/12/2020	6	12	2020	547	15	Guatemala	GT	GTM	17581476.0	America
9	23941	05/12/2020	5	12	2020	752	15	Guatemala	GT	GTM	17581476.0	America

```
totalGastoCompleta = datos.copy().drop(columns = ['dateRep', 'day', 'Unnamed: 0','month','year','geoId','countryterritoryCode','popData2019','continentExp'])
totalGastoCompleta.to_csv('C:/Users/wilso/Desktop/datos_Nuevos_Guate.csv')
```

Ilustración 1 Eliminación de última columna para knn

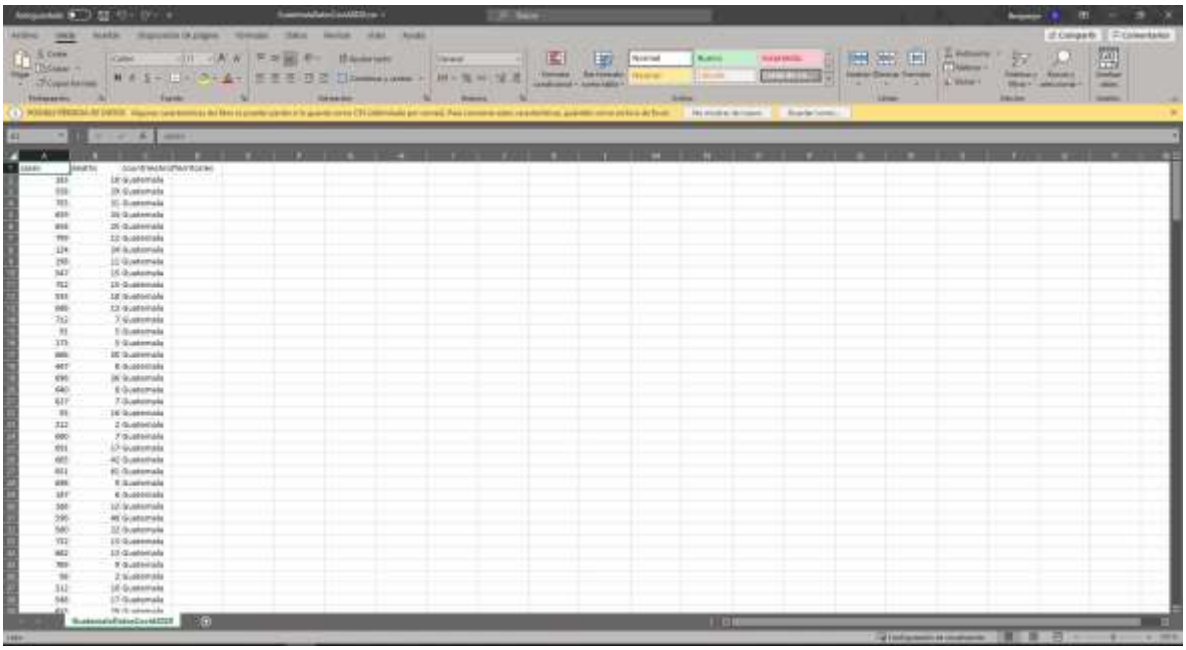


Ilustración 2 Reducción de variables para el árbol de decisiones


ExampleSet (/Local Repository/covid19data)

Open in Turbo Prep Auto Model

Row No.	deaths	dateRep	day	month	year
1	6	Dec 14, 2020	14	12	2020
2	9	Dec 13, 2020	13	12	2020
3	11	Dec 12, 2020	12	12	2020
4	10	Dec 11, 2020	11	12	2020
5	16	Dec 10, 2020	10	12	2020
6	13	Dec 9, 2020	9	12	2020
7	6	Dec 8, 2020	8	12	2020
8	26	Dec 7, 2020	7	12	2020
9	10	Dec 6, 2020	6	12	2020
10	18	Dec 5, 2020	5	12	2020
11	5	Dec 4, 2020	4	12	2020
12	19	Dec 3, 2020	3	12	2020
13	48	Dec 2, 2020	2	12	2020
14	11	Dec 1, 2020	1	12	2020
15	0	Nov 30, 2020	30	11	2020
16	11	Nov 29, 2020	29	11	2020
17	15	Nov 28, 2020	28	11	2020
18	0	Nov 27, 2020	27	11	2020

Ilustración 3 Base de datos de COVID-19 2020

Ilustración 4 Página donde se obtuvo la base de datos

Link para base de datos COVID-19 2020

https://data.europa.eu/data/datasets/covid-19-coronavirus-data?locale=es

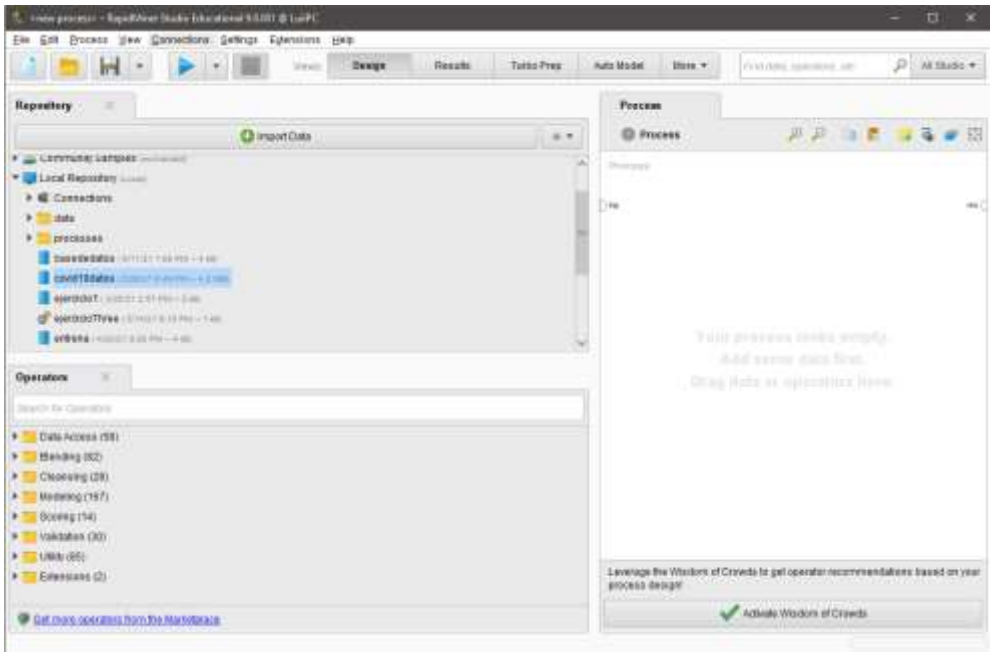


Ilustración 5 Herramienta Rapid Miner

Ilustración 6 Herramienta Jupyter Notebook

Enlace al proyecto en GitHub

https://github.com/LuiiVasquez/proyectoMdD