Minería De Datos

Entregable 1

Integrantes del grupo:
Diego Sepúlveda Millán
Miguel Ángel Matas Rubio
Pedro Rodríguez Viñuales
David Rivera Concepción

0.- Descripción breve del problema planteado.

Desde UNIVERSITYHACK 2021 DATATHON proponen el reto de responder algunas cuestiones interesantes sobre cómo ha afectado al pandemia COVID-19 a la venta y el consumo de verduras y hortalizas a través del análisis de 5 datasets que son proporcionados, que describimos en el siguiente apartado.

Las preguntas a resolver, no siendo estas limitantes a la hora de analizar y procesar los datos, son:

- ¿De qué manera se ha visto afectado el consumo y la demanda de frutas y hortalizas durante la pandemia con respecto a años anteriores?
- ¿Qué efecto ha tenido sobre las importaciones/exportaciones de frutas y hortalizas? ¿Ha tenido algún efecto especial el periodo de excepción (Marzo, ººAbril y Mayo)?
- ¿Existe correlación entre los casos COVID-19 y las importaciones/exportaciones a nivel de la Unión Europea?

1.-Descripción breve de los datos originales:

Inicialmente vamos a contar con dos datasets principales dentro de las 5 opciones de datos a elegir. Utilizaremos los siguientes datasets: *Dataset1.-DatosConsumoAlimentarioMAPAporCCAA.txt* y *Dataset5_Coronavirus_cases.txt.*

El primer conjunto de datos, contiene toda la información de consumo alimentario mensual de todas las comunidades autónomas de España, desde Enero del 2018 hasta Noviembre del 2020. Tiene un total de 26635 líneas. Tenemos las siguientes variables:

- Año, Mes, CCAA, Producto, Volumen (miles de kg), Valor (miles de €), Precio medio kg, Penetración (%), Consumo per cápita, Gasto per cápita.

El quinto conjunto de datos, contiene estadísticas internacionales de COVID-19 por país, desde enero de 2020 hasta noviembre de 2020. Las variables son las siguientes:

- Fecha, día, mes, año, casos detectados de covid, muertes confirmadas por covid, País o territorio, id del país o territorio, código país o territorio, popdata2019, acumulado de casos COVID-19 por cada 100000 en un rango de 14 días.

  (Cuenta con 58690 entradas, pero no todas ellas cuentan con todos los datos por lo que tendremos que someter al dataset a una limpieza de datos)

Como podemos observar, tenemos que hacer una limpieza de datos en el segundo dataset, debido a que solo necesitamos datos de nuestro país. Bastaría con eliminar los países innecesarios, con lo cuál es una tarea muy simple.

2.-Antecedentes o trabajos similares:

Efecto de la COVID-19 en el gasto monetario de los hogares (INE)

En este estudio se ha comparado el gasto monetario en los hogares en el año 2019 y el gasto en 2020, previo al confinamiento, durante al confinamiento y después del confinamiento.

Los resultados de este estudio nos puede guiar en la evolución de los efectos del covid en el gasto.

2.1.-Planteamiento de la hipótesis y de los objetivos a perseguir:

-H1 La pandemia por la covid-19 ha provocado escasez de verduras y hortalizas que se vio reflejado en una subida de precios.

-H2

Además de las preguntas planteadas por el propio reto, pretendemos resolver si los alimentos con vitamina C ayudaron durante la pandemia, ya sea a una menor tasa de contagio, de virulencia, personas hospitalizadas o casos de muertes. Para ello agruparemos los alimentos por grupos de vitaminas que aportan al ser humano, con el fin de encontrar patrones de mejora. En cuanto a la diferencia respecto a un modelo estadístico, crearemos un modelo predictivo con el cual podremos hacer predicciones y crear diferentes contextos para explorar las diferentes posibilidades dando la posibilidad de crear diferentes escenarios.

2.2.-Posibilidades de enriquecimiento de los datos:

En MAPA (Ministerio de Agricultura, Pesca y Alimentación), se pueden consultar datos de carácter similar de años anteriores a los datos ofrecidos o datos del 2021. A través de este enlace *[Base de datos de consumo en hogares] - Alimentación - mapa.gob.es*