



RETO CAJAMAR AGRO ANALYSIS

UniversityHack 2021 Datathon by Cajamar

INFORME DE RESULTADOS

Equipo: *Datacrop*

Página Web: www.datacrop.es

Plataforma: www.platform.datacrop.es

Usuario: Cajamar

Contraseña: Datacrop21

Universidad: Escuela de Organización Industrial

Formado por: Jimena Areta y Sergio Sillero

Fecha: 14 de Abril de 2021

**Puede ver los resultados obtenidos con una mejor experiencia de usuario a través de nuestra web y/o plataforma.*

1. Introducción

La pandemia que hemos vivido a causa de la COVID-19 ha modificado profundamente los hábitos y la forma de relacionarnos en sociedad.

En España durante los primeros días de la semana del 9 de marzo se comenzaron a escuchar rumores sobre la inmediata norma que obligaría al confinamiento total de la población española. Esta norma entró en vigor el 14 de marzo y afectó a 18.625.000 hogares que sufrieron restricciones de movilidad y cambios bruscos en sus hábitos de vida.

Esta decisión extraordinaria e imprevista ocasionó una primera fase de urgencia para acaparar alimentos y útiles domésticos desembocando en la compra compulsiva de productos como frutas, hortalizas y productos de higiene en la que se vino a denominar “la semana de la histeria”, con las calles vacías, los negocios cerrados y solo largas colas de a uno para acceder a los supermercados, panaderías y farmacias.

El informe que presentamos a continuación describe y analiza cómo se vio afectado el sector agroalimentario tanto por el primer estado de alarma (marzo, abril y mayo de 2020) como por la extensión de sus efectos y de los efectos de la propia pandemia durante los meses posteriores. Para resolver este reto emplearemos diferentes líneas de estudio que abarcan tanto datos de consumo, producción y precios en los diferentes mercados agrícolas españoles, como datos de importaciones y exportaciones de los diferentes productos.

Buscamos también dilucidar si existe una correlación entre las diferentes variables y los casos de la COVID-19 en la Unión Europea, así como reflexionar sobre la aceleración de nuevos patrones de consumo que suponen un desafío para los sistemas alimentarios.

2. Procedimiento y metodología

El procedimiento que se ha seguido en este informe consta de las siguientes partes:

- ❖ **Procesamiento de los datos previo al análisis:**
- ❖ **Documentación con la lectura de diferentes artículos**
- ❖ **Carga inicial de los datasets propuestos para el reto**
 1. Consumo de frutas y hortalizas por comunidades autónomas
 2. Precios de frutas y hortalizas en Andalucía

3. Datos de MercaMadrid y MercaBarna
4. Datos de comercio exterior entre países de la UE y España
5. Datos COVID-19

❖ **Carga de datasets propuestos para enriquecer los datos:**

1. Datos producción (2018/2019/2020)
2. Datos precios medios nacionales (por semanas - 2020)
3. Datos consumo por CCAA y lugar de compra (2013-2020)
4. Datos composición atmosférica (2018/2019/2020)

❖ **Limpieza de los datasets en Python (Jupyter Notebook)**

1. Análisis valores faltantes y columnas vacías
2. Formato de las columnas (fecha tipo datetime y números tipo float)
3. Reescribir productos por falta de encoding (con ñ o tilde)
4. Pasar a csv los datasets limpios
5. Agregaciones necesarias para el análisis y modelos de machine learning

❖ **Análisis de los datos:**

Análisis 1:

1. Creación nueva tabla para comparar los valores de consumo y de precios del 2020 con el 2019 y del 2019 con el 2018 (%).
2. Para ello se ha realizado una unión de las tablas con los productos en común de los 3 años.
3. Estudio de correlación entre las variables mediante un mapa de calor.
4. Agrupar los precios por meses para cada producto y calcular el precio medio en Andalucía, MercaMadrid y MercaBarna.
5. Extracción de los informes excel del MAPA los datos relevantes de producción por producto y año para 2018/2019/2020
6. Estudio de los datos de composición atmosférica de los mismos años en cada estación del país.
7. Correlación entre los compuestos químicos presentes en la atmósfera y la producción.

Análisis 2:

1. Estudio de los precios medios nacionales por producto y en el año 2020.
2. Informes de PowerBI para la representación de la evolución de los precios medios nacionales y por mercado (Madrid, Barcelona, Andalucía)

3. Conclusiones sobre la relación entre el consumo, la producción y la variación de los precios durante el periodo de excepción.

Análisis 3:

1. Separar en dos dataframes diferentes las importaciones de las exportaciones.
2. Cálculo de la variación porcentual en el valor de las importaciones y exportaciones entre el año 2020 y 2019 para la posterior creación de un informe en PowerBI basado en el dataset enriquecido.
3. Exportar los datasets con los nombres: “comercioExterior_Limpio.csv”, “importacionesFinal_UE.csv” y “exportacionesFinal_UE.csv”

Análisis 4:

1. Filtrado el dataset de los datos de la COVID-19 a países de la unión europea para poder compararlo en conjunto con los datos de la pregunta 2
2. Estudio de correlación entre los casos de covid-19 y las importaciones y exportaciones.
3. Exportar los datasets con los nombres: “covid19Datos_EU.csv”

Análisis 5:

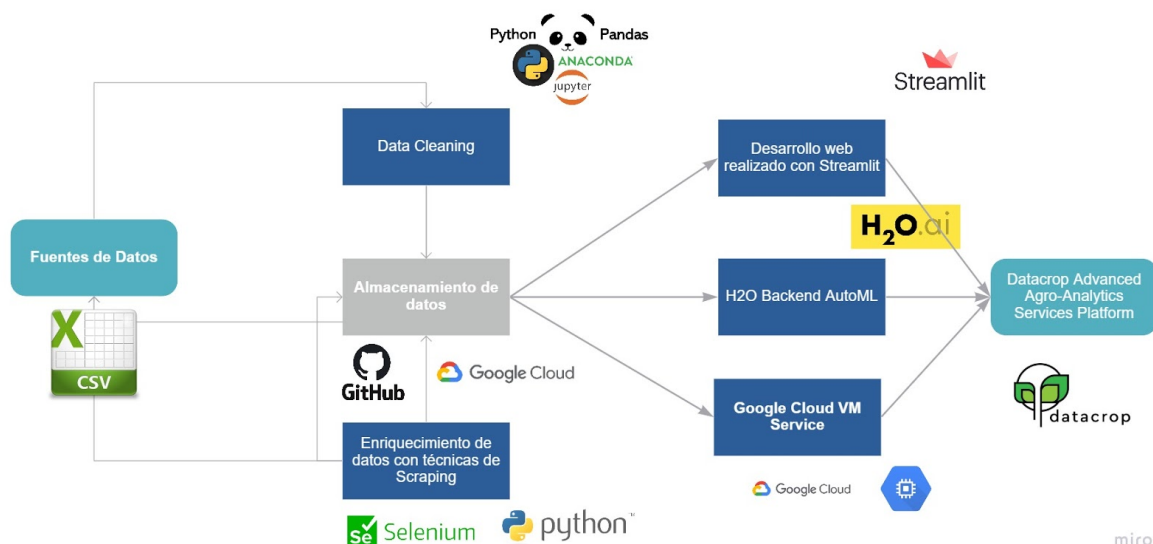
1. Con los datos obtenidos mediante scraping se construyen datasets para los diferentes canales de venta y productos.
2. Se generan informes de PowerBI para representar el consumo en los canales físicos vs online.
3. Construcción de dos modelos de machine learning: uno con datos de 2018/2019/2020 por CCAA y producto y el otro con datos desde el 2013 al 2020 por producto en el canal online(internet)
4. Conclusiones obtenidas del reto y reflexiones del futuro del sector agrícola.

Además de todos los datos que se han obtenido, limpiado y procesado, se han utilizado las siguientes herramientas para enriquecer tanto los datos como su posterior análisis y presentación de los mismos:

- ❖ **Elaboración de informes en PowerBI:** Para las distintas preguntas se ha elaborado un informe presentado tanto en los jupyter notebook como en la web y plataforma.
- ❖ **Se han realizado técnicas de Scraping** para mejorar y enriquecer los datos del reto. Para ello se ha usado librerías como Selenium para automatizar la extracción a través de etiquetas HTML de

los datos que se encontraban alojados en sitios web. Además de la búsqueda de otras fuentes de datos donde se podían bajar los mismos directamente en formato csv.

- ❖ **Desarrollo y despliegue de los resultados obtenidos en la página web:** Creación de una página web para mejorar la experiencia de usuario de cara a presentar los resultados planteados en este informe. Para ello se ha comprado un dominio (www.datacrop.es) donde se ha instalado Wordpress y personalizado con código usando el framework de Bootstrap (HTML5, CSS...)
- ❖ **Desarrollo de la plataforma *Datacrop Advanced Agro-Analytics Service (DAAS)*** con un framework web llamado Streamlit para Python. Dicha plataforma aloja los resultados obtenidos del análisis de las diferentes preguntas del reto de forma mucho más interactiva además de servicios extra de gran valor como: reportes automáticos de análisis previos exploratorios de los datos (EDA) y un servicio propio de *Auto Machine Learning* (AutoML) para generar modelos de inteligencia artificial usando la librería H2O montada en la nube como backend para tal propósito. En ambas herramientas mencionadas anteriormente, podrá subir sus propios datos o elegir los datos del reto para poder probar la utilidad de las mismas. Finalmente hay documentación explicativa en la web de cómo utilizar la plataforma.
- ❖ **La arquitectura usada para montar la plataforma** en la nube ha sido el uso de Google Cloud donde se ha montado un cluster de kubernetes donde alojar todo el código necesario para ejecutar la plataforma y los nodos de H2O para el servicio de AutoML en contenedores, además de usar Cloud DNS para redireccionar nuestro servicio al subdominio www.platform.datacrop.es y aplicar balanceadores de carga y proxy forwarding con NGINX.



- ❖ **Desarrollo de logo para el equipo** con Adobe Photoshop e Illustrator.

3. Resultados obtenidos

En este apartado se presenta una breve descripción de los análisis realizados junto con un enlace para verlos en detalle a través de la [página web](#) y la [plataforma](#) .

Análisis 1: ¿De qué manera se ha visto afectado el consumo y la producción de frutas y hortalizas durante la pandemia con respecto a años anteriores?

Estudio de la variación del consumo por comunidades autónomas y productos como consecuencia del confinamiento y nuevos hábitos de alimentación. Variaciones en la producción por falta de mano de obra y estudio de variables exógenas como la composición atmosférica.

Enlace web: <https://www.datacrop.es/reto-analisis-1/>

Plataforma: <http://www.platform.datacrop.es/> (Iniciar sesión, Análisis de datos, seleccionar análisis 1)
(En caso de duda [ver documentación](#))

Análisis 2: Evolución de los precios en las principales plataformas de distribución de España.

Evolución de los precios medios nacionales del sector hortofrutícola a través de las principales plataformas de distribución (mercaMadrid, mercaBarna y Andalucía). Análisis de las frutas de hueso y los malos resultados de la campaña de la fresa.

Enlace web: <https://www.datacrop.es/reto-analisis-2/>

Plataforma: <http://www.platform.datacrop.es/> (Iniciar sesión, Análisis de datos, seleccionar análisis 2)
(En caso de duda [ver documentación](#))

Análisis 3: ¿Qué efecto ha tenido sobre las importaciones/exportaciones de F&H?¿Ha tenido algún efecto especial el periodo de excepción (Marzo, abril y mayo)?

Estudio de la situación del comercio exterior durante la etapa de la covid-19 entre la Unión Europea y España.

Enlace web: <https://datacrop.es/reto-analisis-3/>

Plataforma: <http://www.platform.datacrop.es/> (Iniciar sesión, Análisis de datos, seleccionar análisis 3)

(En caso de duda [ver documentación](#))

Análisis 4: ¿Existe correlación entre los casos COVID-19 y las importaciones y exportaciones a nivel de la Unión Europea?

Análisis de correlación entre el número de casos de la covid-19 durante los primeros meses de la pandemia y las importaciones/exportaciones entre la UE y España.

Enlace web: <https://datacrop.es/reto-analisis-4/>

Plataforma: <http://www.platform.datacrop.es/> (Iniciar sesión, Análisis de datos, seleccionar análisis 4)

(En caso de duda [ver documentación](#))

Análisis 5: El estado de alarma revoluciona los hábitos de consumo

Localización de nuevos patrones de consumo mediante el estudio de los diferentes canales de venta (Físicos y Online). Acercamiento de la inteligencia artificial al sector agrícola mediante modelos de machine learning que predicen el consumo en base al precio medio de las Frutas y Hortalizas. Conclusión y reflexión sobre el futuro del sector agrícola.

Enlace web: <https://datacrop.es/reto-analisis-5/>

Plataforma: <http://www.platform.datacrop.es/> (Iniciar sesión, Análisis de datos, seleccionar análisis 5)

(En caso de duda [ver documentación](#))

4. Mejoras futuras

Las mejoras que implementaríamos están enfocadas en la plataforma donde desarrollaríamos nuevos servicios como análisis de imágenes satelitales (Sentinel-2) con redes neuronales convolucionales, además de la mejora de los servicios ya existentes como el autoML.

El eje vertebrador de Datacrop es dicha plataforma, donde la intención de la misma es crear una serie de servicios para apoyar la digitalización del sector agrario. Por ello, nuestra intención no solo es presentar los análisis de nuestro reto, sino continuar con el desarrollo de ella como apoyo para el sector en la toma de decisiones en la época post-covid. Esta será totalmente gratuita y open-source.