Entrega Final

RETO CAJAMAR AGROANALYSIS

Integrantes del grupo:

-Diego Sepúlveda Millán

-Miguel Ángel Matas Rubio

-Pedro Rodríguez Viñuales

-David Rivera Concepción

Enlace al repositorio

HIPÓTESIS 1

1.Introducción.-

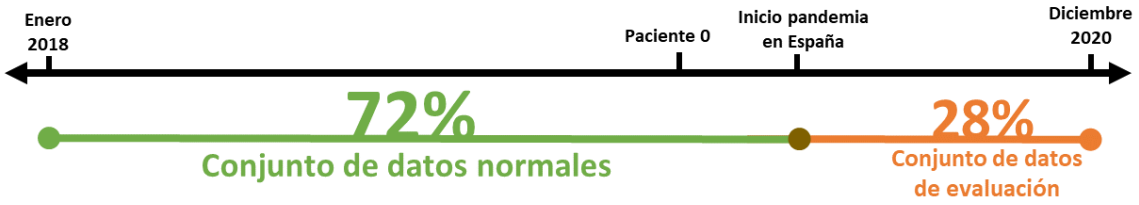
La primera hipótesis planteada es *"La pandemia por la covid-19 provocó escasez de frutas y hortalizas durante dicha pandemia."*

Para responder a esta hipótesis partimos de la información dada por el *Reto Cajamar Agro Analysis*. Hacemos especial hincapié en el *Dataset1* ya que contiene datos tanto del consumo total nacional como por CCAA de frutas y hortalizas en miles de kilogramos por mes, desde 2018 a 2020. El objetivo que se ha perseguido es aplicar técnicas de análisis de datos para afirmar o refutar la hipótesis planteada.

2.Fases del proceso KDD realizadas:

2.1.-Selección de datos:

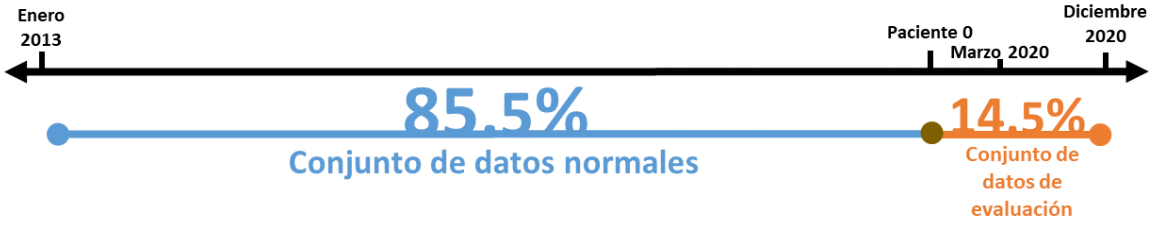
Tras el estudio preliminar de los datos, lo primero que se hizo fue hacer una distinción del tiempo, marcando el punto de inflexión como el inicio de la pandemia en España, el mes de marzo de 2020 para poder tomar los datos anteriores a esta fecha como *datos estables**. El problema, es que al hacer esto, estamos siendo demasiado optimistas al incluir algunos meses anteriores a la pandemia como datos normales. Probablemente estos datos ya podrían estar afectados por los efectos directos o indirectos de la pandemia.



Con el objetivo de poder hacer un estudio más preciso se buscó ampliar la base de datos para poder ensanchar el conjunto de datos de evaluación. Así se sortea el posible problema de que los datos de meses anteriores al inicio de la pandemia en España estuviesen afectados directa o indirectamente por la pandemia global y estos fuesen tomados como parte del conjunto de datos estables.

En la página web del Ministerio de Agricultura, Pesca y Alimentación se encuentran accesibles los datos del *Dataset1* con la particularidad de que el periodo de los datos se amplía pues contiene los datos a partir de 2013. https://www.mapa.gob.es/app/consumo-en-hogares/consulta11.asp

Con esta ampliación podemos tomar un rango mayor como datos normales, situando su inicio en el mes de noviembre de 2019, sin tener una reducción del porcentaje de datos estables inasumible.



Para obtener estos datos se ha creado un scraper programado en Python. Con ayuda de la librería selenium de python, que está enfocada en el testeo de páginas web pero ya que la página funciona

con un formulario que hay que completar para que muestre los datos de cada distinto mes, sin dar opción la descarga automática, es ideal para poder extraer los datos.

*datos estables: se refieren a datos que no presentan patrones o comportamientos anormales que puedan afectar la precisión de las predicciones.

2.2.-Procesamiento de los datos:

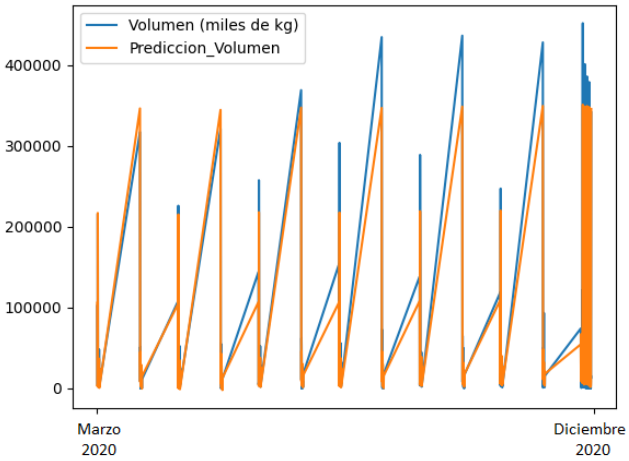
Los datos han sido procesados con pandas. Se limpiaron las columnas innecesarias y se transformó en float los datos cuyos decimales estaban expresados con comas y las unidades de millar acompañadas de un punto. Las columnas mes y producto al ser categóricas se transformaron usando la técnica one-hot. Y algunos valores nulos del primer dataset fueron sustituidos por la media ya que pertenecían a los últimos meses que no entraban en el conjunto de datos de evaluación. En los datos ampliados no hubo necesidad pues no había valores Nan.

Para la comprobación de la calidad de los datos al principio se optó por usar la librería pandera en reemplazo a great expectations pero finalmente se usaron las primitivas de pandas como *"df.dtypes", "df.isnull()", "pd.get_dummies",* para comprobar que la calidad de los datos es apta para el modelo.

2.3.- Minería de datos:

Primero se ha creado un modelo de regresión lineal con el dataset1 sin ampliar. Como variables independientes se ha tomado las columnas que indican la fecha y las columnas producto. Una vez creada la predicción sobre el primer conjunto de datos a evaluar obtenemos una precisión del 81.39%. Este valor no lo tomaremos como una evaluación del modelo en sí, si no como una prueba de que los datos del conjunto de evaluación eran los esperados en un 81.39%, según el modelo limitado que se ha creado con los datos base del dataset1.

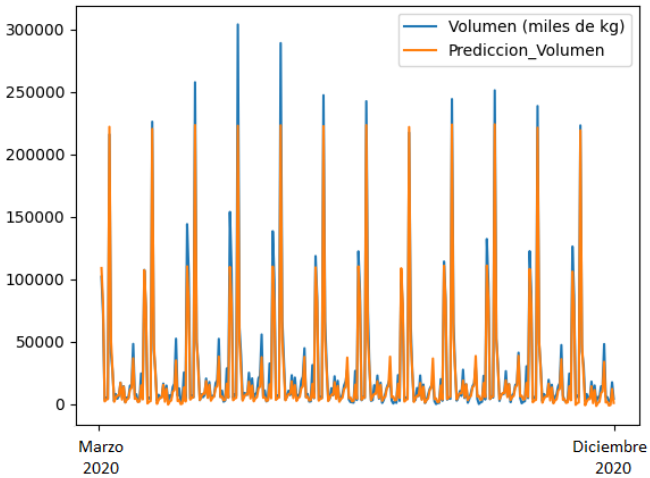
Para ver si las predicciones no llegan o sobrepasan a los datos reales del conjunto de evaluación, se ha creado el gráfico que mostramos a la derecha para poder hacernos una idea aproximada.



Y como podemos ver durante los primeros meses la medida del consumo en unidades de volumen no llega a la realidad por poco pero pronto es superado el volumen real al estimado, y se aprecia más diferenciación.

Para continuar sometemos los datos ampliados al mismo proceso.

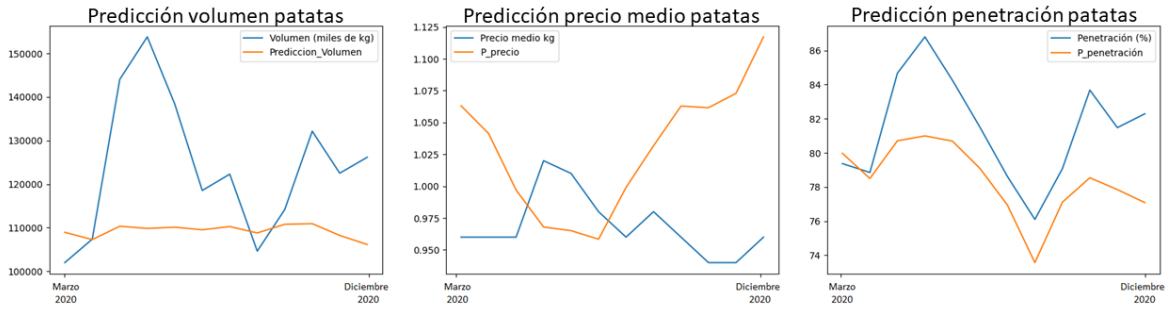
Al crear el modelo con los datos ampliados obtenemos una precisión del modelo de 95.25%. Y al visualizar la distancia entre los datos reales y los de la predicción creamos el gráfico de la derecha. Donde podemos apreciar las pequeñas diferencias aunque los resultados son muy parecidos al del primer modelo con la salvedad de que este contradice aún más la



hipótesis pues la diferencia es menor y sigue siendo el volumen real, superior al esperado.

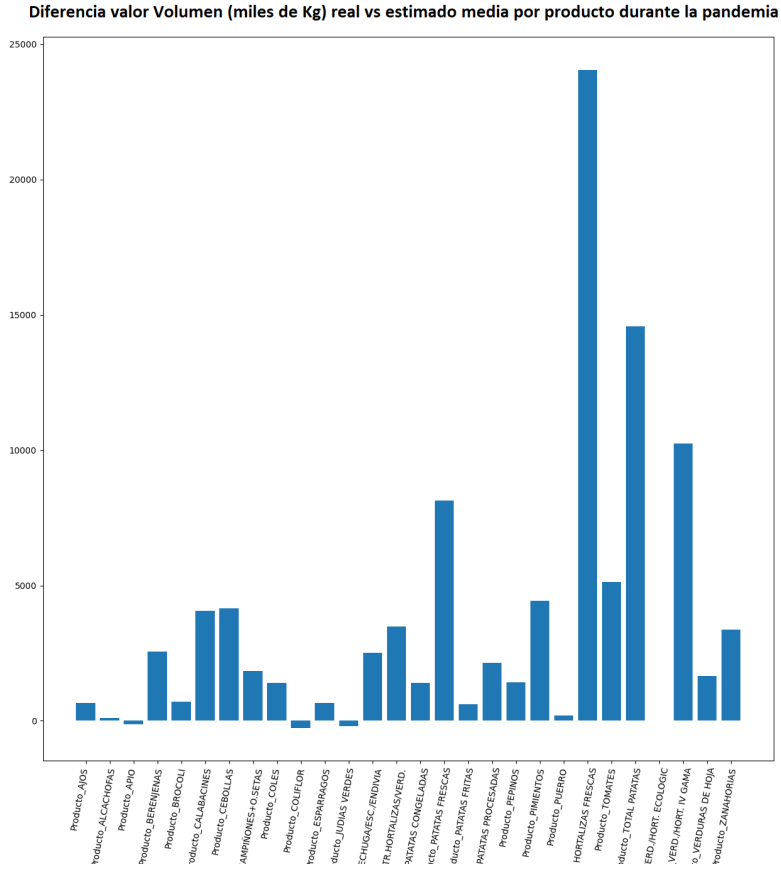
Los datos que difieren más respecto a su predicción son outliers que pertenecen a productos concretos. Lo que indica que se debería analizar producto a producto para profundizar más.

Por ello tomaremos como muestra el total de patatas y así poder ver más claramente los resultados del estudio.

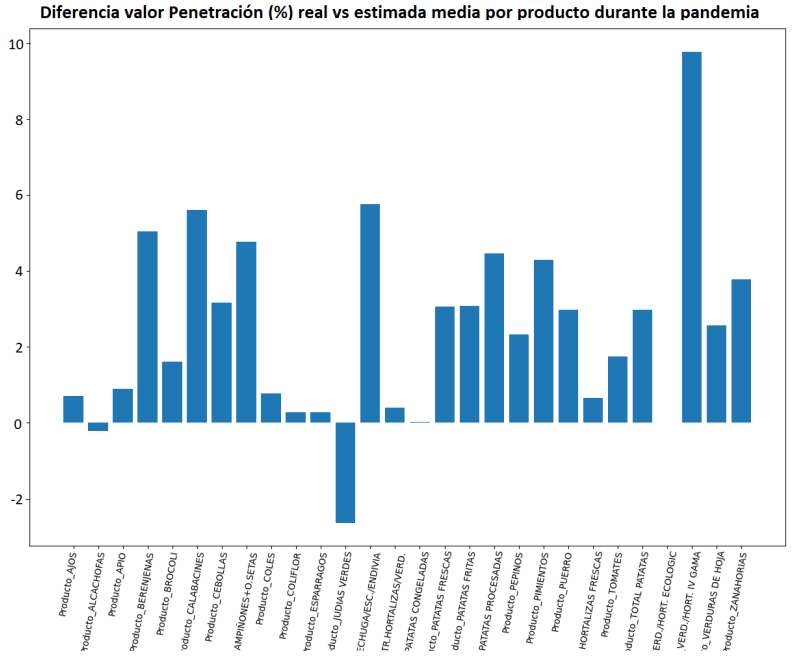


Para comprobar que los datos no han sido corrompidos por anomalías en los precios o un reparto desigual de los productos mostramos también las gráficas de precio y penetración del mismo producto.

Al calcular la diferencia por producto del valor del consumo en volumen (miles de kilogramos) entre los datos reales y los datos de predicción obtenemos el siguiente gráfico.



Mismamente se calcula la diferencia entre la penetración real de los productos en los hogares y su predicción, dando lugar al siguiente gráfico.



3.-Resultados y Conclusiones.

Tras contrastar los resultados que arroja el modelo de regresión lineal y visualizarlos de distintas maneras, podemos afirmar que durante la pandemia se produjo una subida en las ventas, lo cual se ve reflejado en el volumen consumido, y que esta no está acompañada de una bajada en la penetración en los hogares, lo que sería indicador de que hubo anomalías en la distribución de la comida o no fueron registradas. Hay excepciones como las judías verdes que penetraron con un 2% menos de media en los hogares durante la pandemia pero un 2% es muy poco y es un solo producto como para poder hablar de desabastecimiento, ya que los demás incrementaron su penetración o en casos puntuales la mantuvieron estable positivamente.

Por todo ello, se puede refutar la hipótesis 1, dado que no hubo en ningún momento una cantidad significativa menor en el volumen consumido. Pero no se puede descartar que, durante esos meses, hubiese habido desabastecimiento en días puntuales debido a cuellos de botella generados por el transporte entre comunidades o la propia capacidad de los establecimientos de venta de alimentos, lo que generaría la sensación de desabastecimiento que provocó también la formulación de esta misma hipótesis.

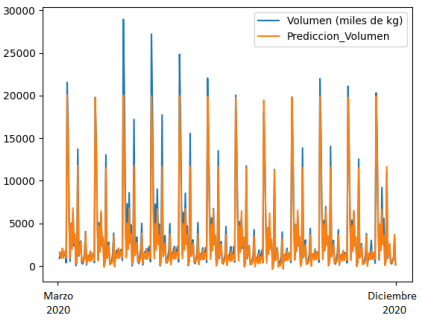
También tenemos en cuenta que España es productora de la mayoría de productos que estamos sometiendo al análisis, por lo tanto para que que se produjese desabastecimiento en los productos como frutas y hortalizas, deberían de haber dado situaciones mucho más anómalas como podría haber sido una aún mayor subida del volumen consumido que provocase desabastecimiento temporal por la compra compulsiva. Por ello para esta hipótesis no se ha tenido en cuenta datasets de importación a España, pues en parte el paro en exportaciones refrendó la posibilidad de desabastecimiento en nuestro país.

3.-1 Alcance.

Para ampliar el estudio y profundizar en el problema, podríamos aplicar el mismo proceso de construcción de modelo de regresión lineal pero por cada comunidad autónoma de manera independiente para asegurarnos de que no hubo desabastecimiento en algunas comunidades en concreto y fue compensado con sobreabastecimiento en otras.

Por otra parte, se podría hacer lo mismo, pero evaluando producto por producto. Aunque sería sencillo y realmente podemos extraer estos datos directamente del modelo que ya tenemos, la complejidad reside en la visualización que debe ser manual de producto por producto, aunque sí podemos automatizar la detección de outliers.

También se debería ampliar aún más la base de datos pues se debería tener en cuenta datos de otros productos que bien pueden estar en otras categorías, pueden ser equivalentes y estar diferenciados simplemente porque están o no procesados de diferentes maneras. Por ejemplo se han extraído los datos de las frutas y hortalizas transformadas y podemos ver productos que simplemente al venderse como troceados o cortados ya pertenecen a otra categoría y si no agrupamos los datasets escapan a nuestro estudio. Por ejemplo vemos que con el dataset de frutas y hortalizas transformadas obtenemos una distribución muy parecida, como muestra el gráfico de la derecha y probablemente encajen con un efecto cremallera.



HIPÓTESIS 2

1.Introducción.-

Durante la pandemia los precios, consumo per cápita y gasto per cápita en las verduras y hortalizas se vieron afectados por las olas. Con esta hipótesis se quiere cuantificar la influencia de las de las olas de contagios en el consumo de este sector continuando lo estudiado sobre los precios en la hipótesis 1.

2.Fases del proceso KDD realizadas:

2.1.-Selección de datos:

En la página web del Ministerio de Agricultura, Pesca y Alimentación se encuentran accesibles los datos del *Dataset1* con la particularidad de que el periodo de los datos se amplía pues contiene los datos a partir de 2013 hasta el 2020 como se ha realizado en la hipótesis 1.

https://www.mapa.gob.es/app/consumo-en-hogares/consulta11.asp

Para esta hipótesis se han seleccionado las columnas '*Gasto per cápita*' y 'Consumo per cápita' referenciando al gasto per cápita o consumo de un producto en un mes en concreto.

2.2.-Procesamiento de los datos:

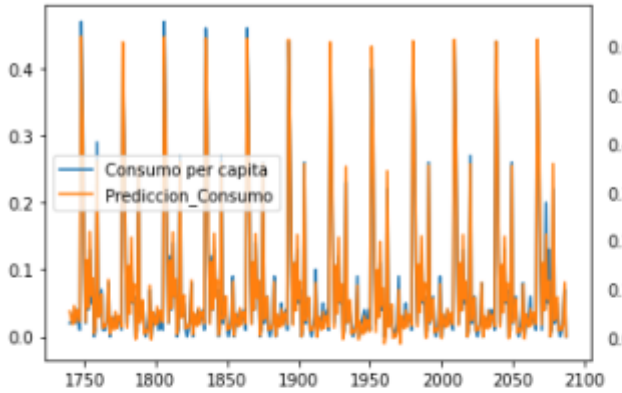
Los datos han sido procesados de forma similar que en la hipótesis 1. Se limpiaron las columnas innecesarias y se transformó en float los datos cuyos decimales estaban expresados con comas y las unidades de millar acompañadas de un punto. Las columnas mes y producto al ser categóricas se transformaron usando la técnica one-hot. Y algunos valores nulos del primer dataset fueron sustituidos por la media ya que pertenecían a los últimos meses que no entraban en el conjunto de datos de evaluación.

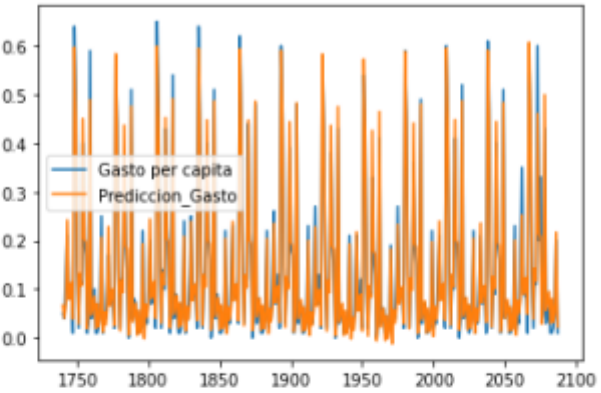
2.3.- Minería de datos:

Se ha realizado un modelo de regresión lineal con el dataset1 ampliado. Como variables independientes se ha tomado las columnas que indican la fecha y las columnas producto y como dependientes las columnas de '*Gasto per cápita*' y 'Consumo per cápita'.

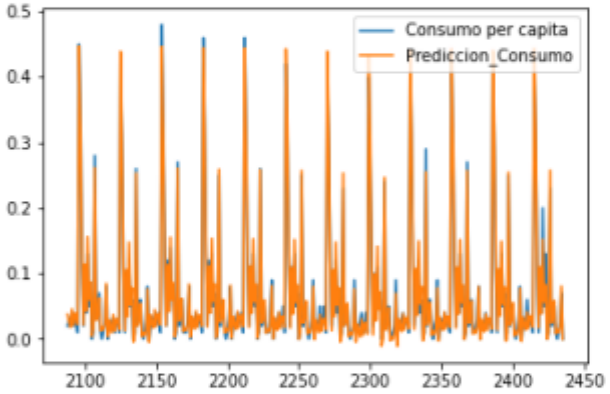
El modelo se entrena con el subconjunto de datos de 2013 a 2017 para así realizar una predicción sobre el valor de las variables dependientes en los mismos meses pero en años posteriores obteniendo una precisión utilizando como test los datos de los años 2018, 2019 y 2020 de 97.71%, 97.57% y 93.56%.

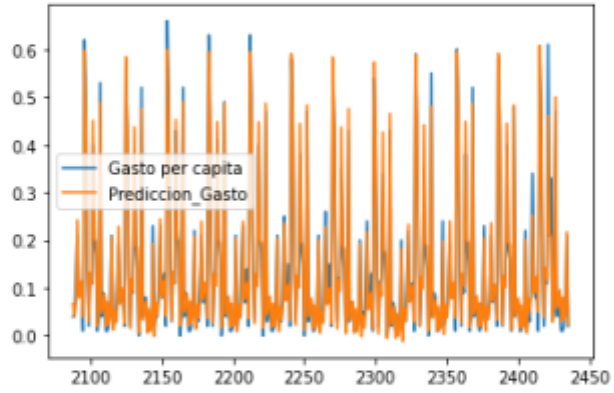
Test Datos 2018



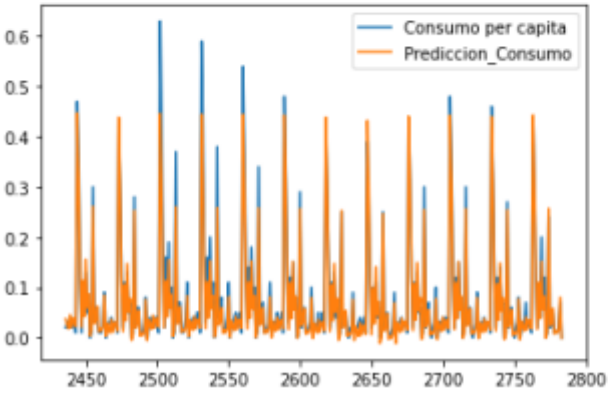


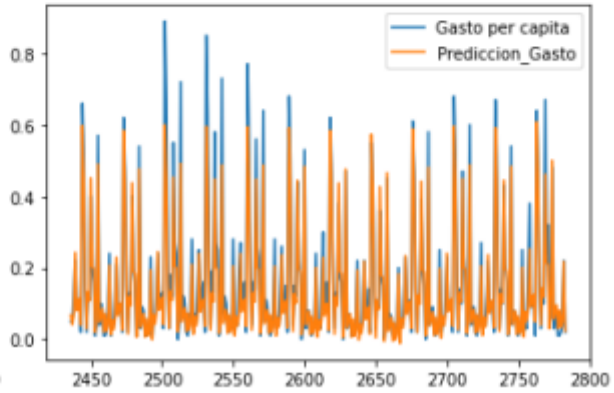
Test datos 2019





Test datos 2020

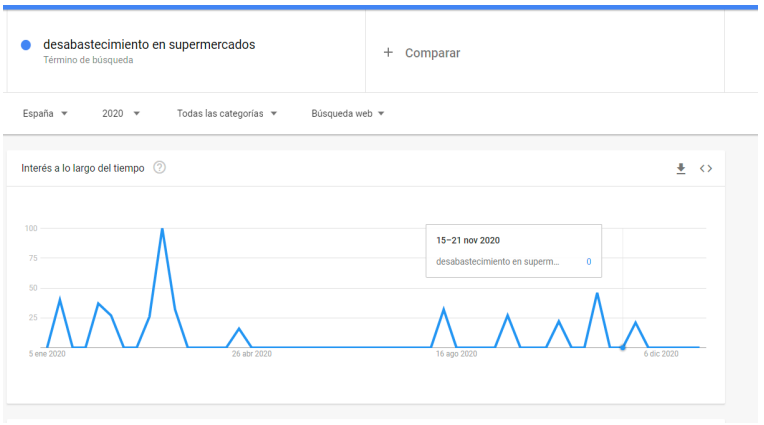




3.-Resultados y Conclusiones.

Se puede apreciar en los resultados que el gasto y consumo estimado para 2020 y lo ocurrido tiene una sustancial diferencia respecto a los años anteriores.

Comparando esta gráfica de google trend sobre el tema “desabastecimiento en supermercados” con la de test datos 2020 se puede apreciar la similitud en los picos, con su máximo valor al inicio en marzo de 2020 disminuyendo en el verano para volver aumentar al finalizar este con la vuelta de la segunda ola.



Por lo que se puede intuir que el incremento del gasto y consumo no tiene porque estar dado por un desabastecimiento de frutas y verduras ya que esto no ha sido demostrado y si por el temor al desabastecimiento.

3.-1 Alcance.

Se podría realizar el mismo estudio producto a producto para ver qué productos son más propensos a ser comprados compulsivamente.

HIPÓTESIS 3

1.Introducción.-

La tercera hipótesis que se planteó fue: “Los super alimentos han influido en la reducción de los casos y la mortalidad del covid-19.”

Con el total nacional de consumo, se comprobará si hay una relación entre la penetración de los alimentos denominados superalimentos de la lista que hemos escrapeado y compararlos con los datos de contagios y fallecimientos, estudiando su correlación.

2.Fases del proceso KDD realizadas:

2.1.-Selección de datos:

Los datos a utilizar son el dataset 1 del consumo Alimentario en España, obtenido del Ministerio de Agricultura, Pesca y Alimentación de España, y el dataset 5 sobre los casos y muertes del coronavirus. Ambos se nos fueron suministrados del campus virtual de la asignatura para la realización del proyecto. Aparte de esto usamos una lista de superalimentos obtenida a través de un scrapper que hemos realizado con la librería selenium de la que ya hemos hablado en la primera hipótesis.

2.2.-Procesamiento de los datos:

Después de examinar los datos del dataset1 del Consumo Alimentario en españa y el dataset 5 sobre los casos y muertes del coronavirus. Decidimos que tras una correcta limpieza de datos construiremos un dataset con los datos de los casos de covid en España mezclado con los datos de los consumos nacionales de los superalimentos. Para esto almacenamos la lista de superalimentos.txt en un array de strings, el cual usaremos para seleccionar las filas de los productos deseados del dataset. Debido a que en el mismo no se encuentran todos los alimentos, si no que había datos de las hortalizas y algunas frutas y verduras, pero por ejemplo alimentos como los huevos no había datos.

Tras reducir los datos del dataset 1 a los superalimentos, comprobamos el estado de los datos, como valores 0 o nulos, o buscar valores redundantes, al hacer esto vimos que los datos de algunos alimentos estaban a 0 en algún mes, lo que se hizo fue calcular la media de el producto concreto y completar los datos. Eliminamos las columnas innecesarias como columnas que indican el país o datos innecesarios.

En el caso del dataset 5 se han seleccionado los datos de España y se ha pasado el formato del mes que era de números, a los nombres (‘Enero’, ‘Febrero’, ‘Marzo’...), a través de un diccionario. También se han eliminado columnas con datos innecesarios para el estudio en cuestión.

Para juntar los dataset se han renombrado las columnas de mes y año para que se llamen igual en ambos data frame, y se ha realizado un merge a los mismos con la librería pandas.

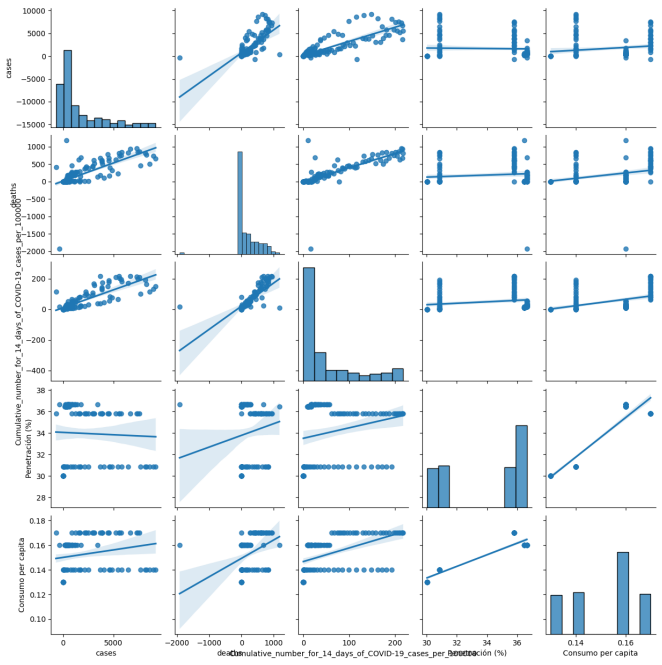
2.3.- Minería de datos:

Con el nuevo data frame creado con los datos deseados, se decidió realizar un estudio manual de las gráficas a través de la representación de las correlaciones de las gráficas entre los distintos valores. Para la mayor comprensión de las gráficas decidimos realizar unos nuevos data frames específicos para cada uno de los superalimentos con los que habíamos acabado obteniendo tras el procesamiento de los datos en el data frame resultante. Para ello gracias al uso de máscaras seleccionamos las columnas deseadas para la realización de las gráficas que relacionan todos los datos que deseamos estudiar, en este caso fueron:

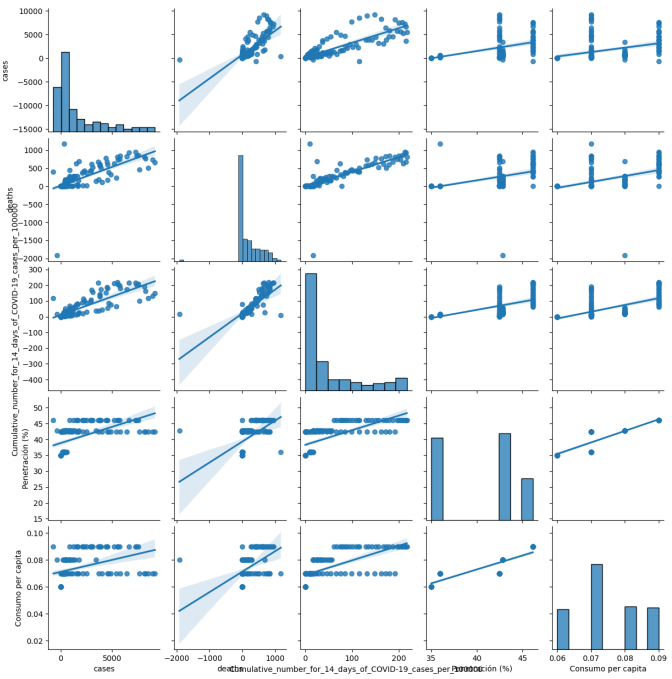
- ‘cases’: Son los casos de infectados que había ese día registrados en España.
- 'deaths': Son las muertes que hubo ese día en España registradas como causa de muerte por covid.
- 'Cumulative_number_for_14_days_of_COVID-19_cases_per_100000': Como indica el nombre de la columna en inglés, es el número acumulativo de casos Covid en los últimos 14 días por cada 100000 personas.
- 'Penetración (%)': Es la penetración del superalimento en concreto en ese mes. En sí es el porcentaje de hogares que consumieron el producto.
- 'Consumo per capita': Es el consumo del superalimento específico por habitante durante el mes.

Tras esto para la generación de estas gráficas hemos utilizando las librerías de python seaborn y matplotlib.

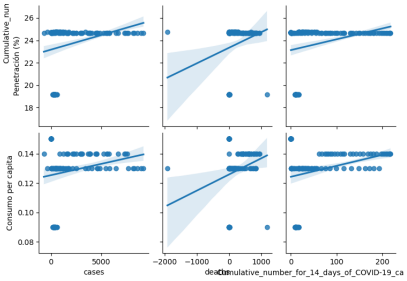
Gráficas correlación Aguacates



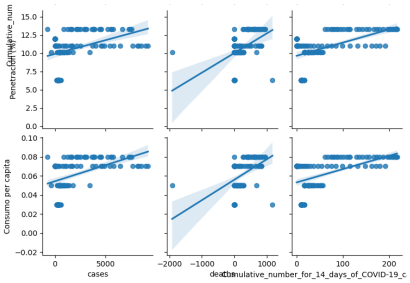
Gráficas correlación Ajos



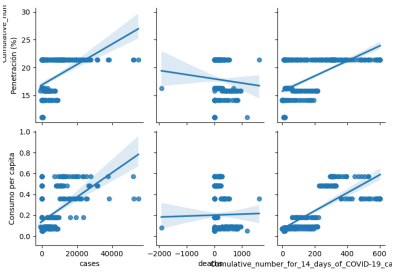
Gráficas correlación Brócoli



Gráficas correlación Coliflor

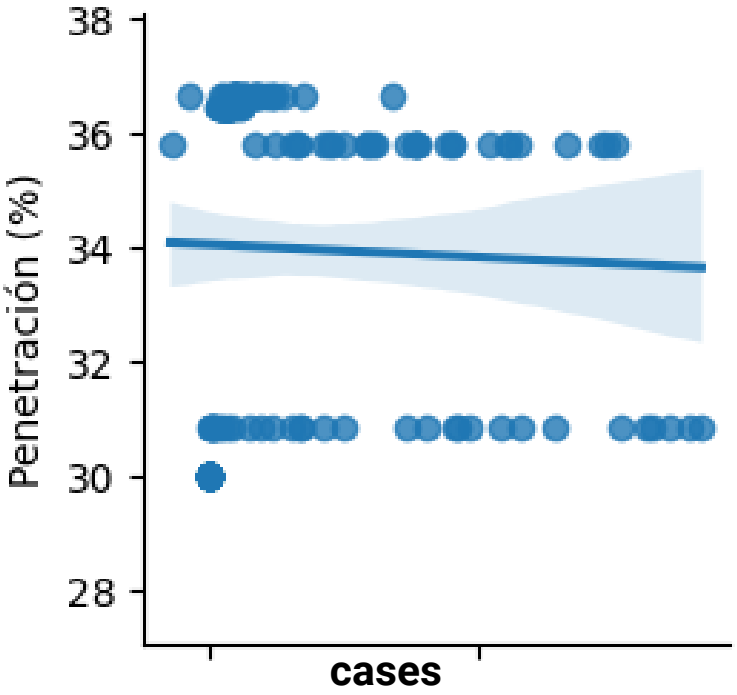


Gráficas correlación Uvas



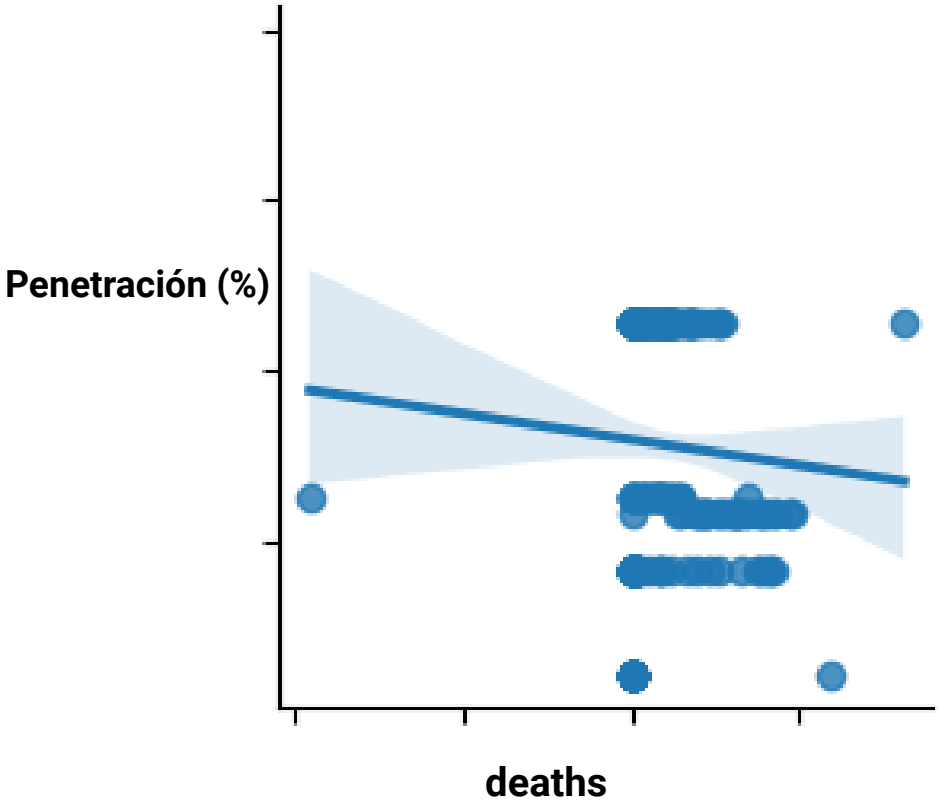
Tras el estudio de las gráficas, hemos observado que el porcentaje de penetración del aguacate es ligeramente inversamente proporcional al número de casos de COVID-19. Estos datos tras estudiar la temporalidad del aguacate podrían indicar que realmente hay algún tipo de relación entre su ingesta y el contagio del covid, el problema es que la relación no es suficientemente concluyente y nos faltan otros datos con los que podríamos apoyar la hipótesis. Pero en cuanto a un problema de que el resultado fuera debido a la temporalidad, no es sustentable debido a que el aguacate es más consumido precisamente en los meses que se acabó el confinamiento y subieron los contagios, lo cual nos hace ver que aquí podría haber algo, que con otros bases de datos se podría profundizar más.

Gráfica Relación Penetración del Aguacate con el número de Casos Covid



También hemos observado que la relación entre el porcentaje de penetración de las uvas y la tasa de fallecimientos debidos al covid es inversamente proporcional. Tras estudiar la temporalidad lo más posible es que fuera debido a que la temporada de uvas es a partir de octubre, cuando la pandemia estaba en un punto en el que ya habían descendido el número de muertes semanales y aumentado la cantidad de casos debido a la retirada de restricciones.

Gráfica Relación Penetración de las Uvas con las muertes



3.-Resultados y Conclusiones.

La hipótesis no creemos que podamos aceptarla ni rechazarla, aunque en parte podríamos pensar aceptarla debido a que hemos encontrado una relación inversa con el consumo de algunos de los mencionados superalimentos y los fallecimientos o contagios. Si tuviéramos a nuestra disposición información de más superalimentos, a lo mejor podríamos haber aceptado en mayor medida la hipótesis.

Los resultados en general dan una relación directa debida a la temporalidad de los acontecimientos de la pandemia, por lo que hemos obviado analizar esas gráficas en mayor profundidad, y nos hemos centrado en las que no cumplían esto.

En primer lugar pensamos que el aguacate aunque su curva en la relación inversa no es muy pronunciada, pensamos que esta puede haber sido suavizada por la temporalidad de la venta de aguacate la cual al relacionarse con la temporalidad en la que hubo mayor número de casos, debería de haber sido una relación más directa que inversa.

En segundo lugar la gráfica obtenida de la relación entre la penetración de las uvas y la tasa de muertes es debido simplemente a que los momentos con mayor penetración que son a partir de octubre fueron cuando menos muertes hubo, por lo que eso explica que la gráfica sea así, harían falta otro tipo de datos e información para poder estudiar este tipo de cosas con alimentos que dependen

tanto de la temporada del año, al depender el número de muertes y casos también de la temporalidad.

Conclusión, no tenemos los suficientes datos sobre los superalimentos para tener una conclusión general sobre los mismos, y debido a la dependencia en cuanto a la temporalidad de los datos tanto del covid como de algunos alimentos esto lleva a una mayor dificultad para dar veracidad a las relaciones encontradas entre los mismos.

HIPÓTESIS 4

1.Introducción.-

Como última hipótesis tenemos la siguiente: *"En las comunidades autónomas donde es rica la ingesta de verduras y frutas, hubo una menor tasa de mortalidad por covid*".

Averiguar si el consumo de verduras y de frutas por comunidad autónoma, tiene relación con la tasa de mortalidad por coronavirus. Para ello se siguen los procesos KDD nombrados anteriormente para poder hacer un estudio de los datos.

2.Fases del proceso KDD realizadas:

2.1.-Selección de datos.

En la parte de selección de datos para esta hipótesis hemos decidido utilizar la base de datos de consumo de productos agrícolas por comunidad autónoma que se nos dio en la asignatura. Para complementar estos datos, hemos dado uso a un database de casos COVID por comunidad autónoma para así poder realizar el estudio requerido. La base de datos es la siguiente: https://portalestadistico.com/?pn=portalestadistico&pc=AAA00&idp=57&idpl=1348&idioma=

2.3.- Procesamiento de los datos.

El procesamiento de datos se ha hecho con la librería de python pandas. Primeramente se han limpiado los dos datasets que van a ser usados, para ello se eliminaron las variables que no van a ser utilizadas o que no son importantes para esta hipótesis. Eliminando todos los valores igualados a cero, NaN y normalizando todas las variables de los datasets, por ejemplo, escribir el mes en número o el formato a la hora de escribir las comunidades autónomas, hemos hecho un *merge* de ambos conjuntos de datos con pandas para tener el data frame resultado para nuestro estudio. Cabe destacar que los datos sólo comprenden los meses de Marzo a Junio del año 2020, cuando la pandemia comenzó y en su momento más álgido, el confinamiento. Todo el código se encuentra subido en nuestro repositorio de GitHub.

2.2.- Minería de datos.

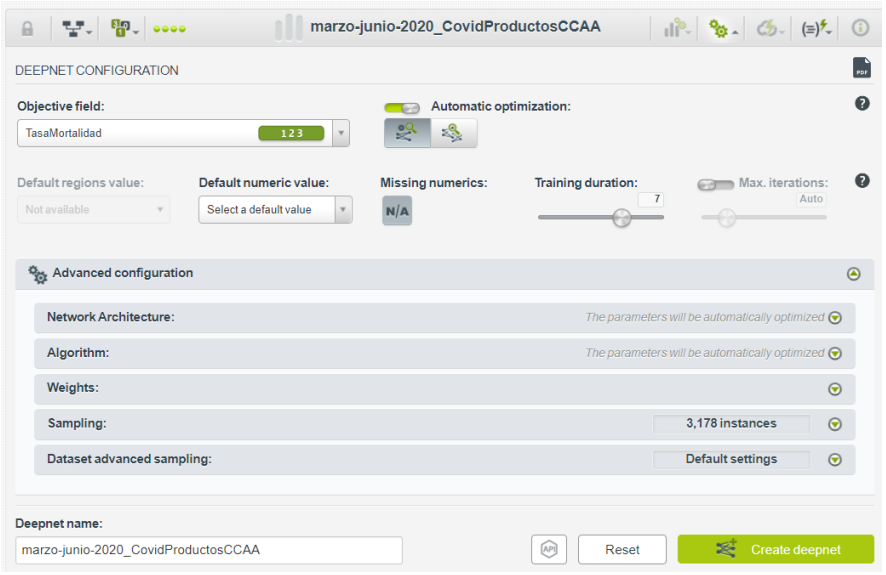
En esta parte hemos visto que era más conveniente utilizar BigML para todo el proceso que conlleva la minería, esto se debe a que en esta hipótesis se requiere un análisis más exhaustivo de los datos para poder llegar a una conclusión correcta y con sentido.

Inicialmente, comprobando las posibles relaciones entre distintas variables, podemos comprobar que si establecemos una relación

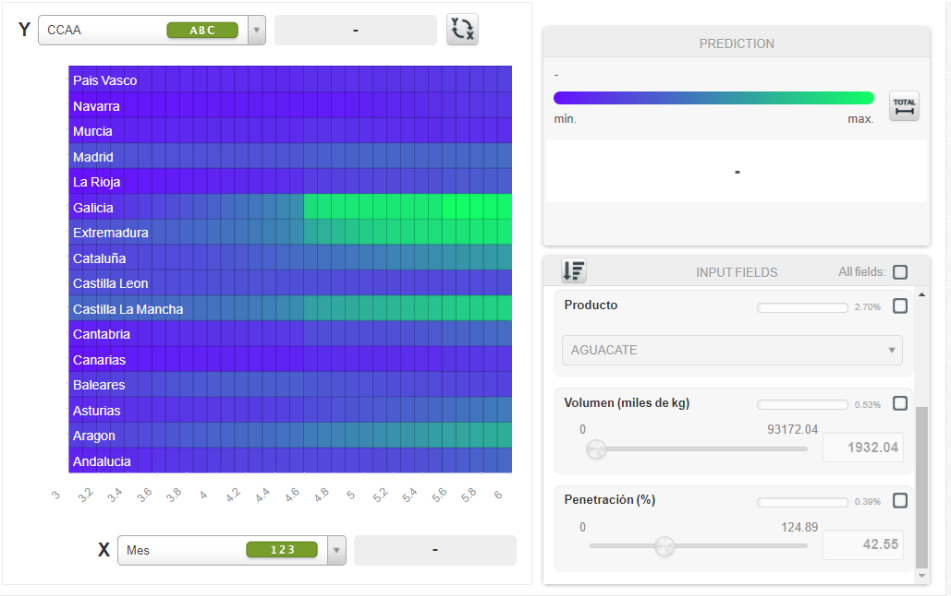
entre la tasa de mortalidad y CCAA, siendo el color de los conjuntos el mes, la tasa de mortalidad va subiendo en la mayoría de las Comunidades Autónomas cuanto más nos acercamos al mes 6.

Habiendo investigado sobre todas las relaciones entre las variables y habiendo discutido sobre las posibles soluciones a nuestro problema, hemos hecho uso de las redes neuronales que nos aporta BigML en su sistema de aprendizaje supervisado.

Como campo objetivo queremos la tasa de mortalidad para observar los demás campos de la gráfica y ver cómo se comporta el valor predicho por la *deepnet*. Además se ha añadido más duración al entrenamiento para sacar unas predicciones más ajustadas a la realidad. Los demás valores han sido establecidos como default dentro de los parámetros que la herramienta considera que son óptimos.

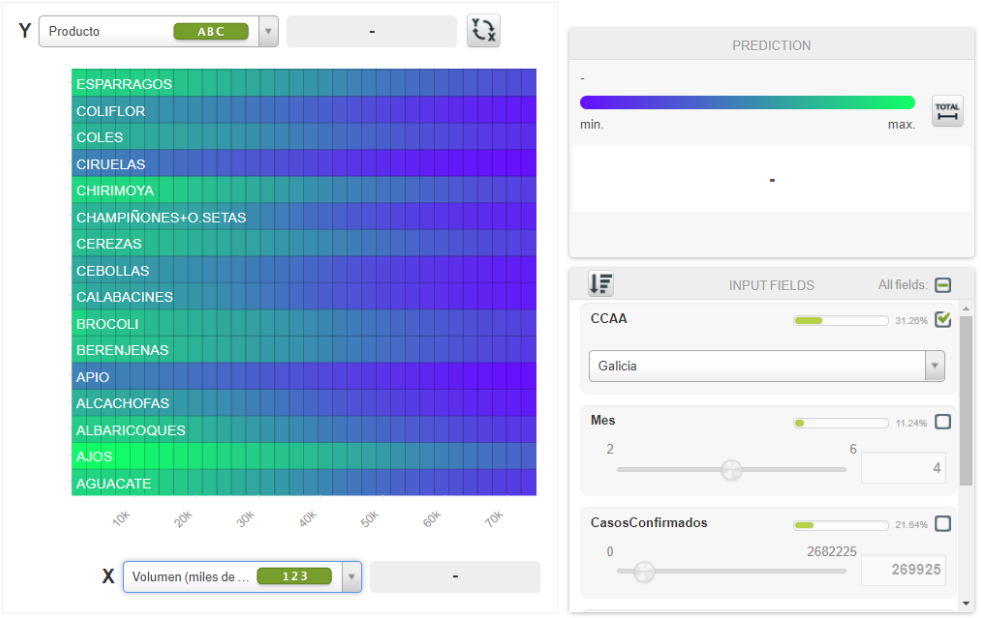


Habiendo entrenado la red neuronal, teniendo los valores CCAA en el eje Y y el mes en el eje X, podemos observar fácilmente que en los últimos meses, Galicia, Extremadura, Castilla La-Mancha y Aragón tienen una mayor tasa de mortalidad, siendo la máxima aproximadamente un 18%. Teniendo en cuenta esto, buscamos alguna anomalía que explicase este suceso.

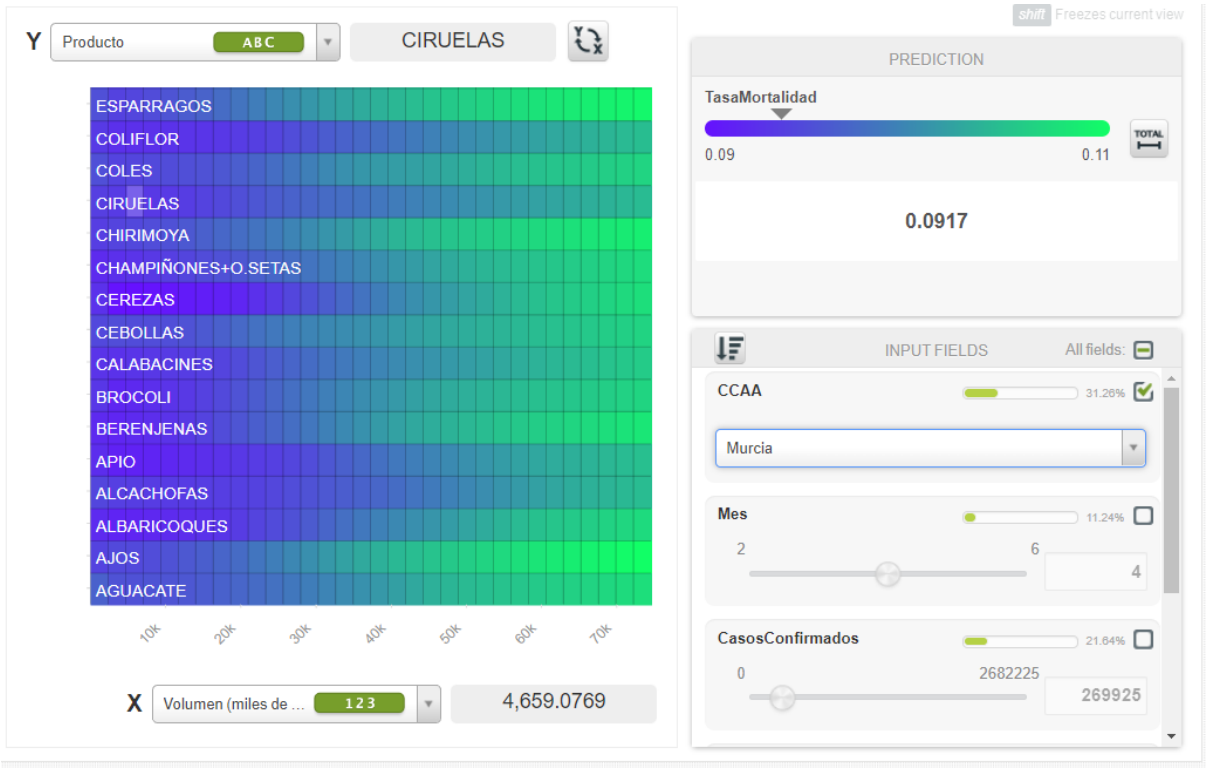


Para ello decidimos estudiar la tasa de mortalidad de Galicia con el volumen de productos consumidos en dicha comunidad autónoma para intentar validar nuestra hipótesis. Se obtiene el siguiente resultado.

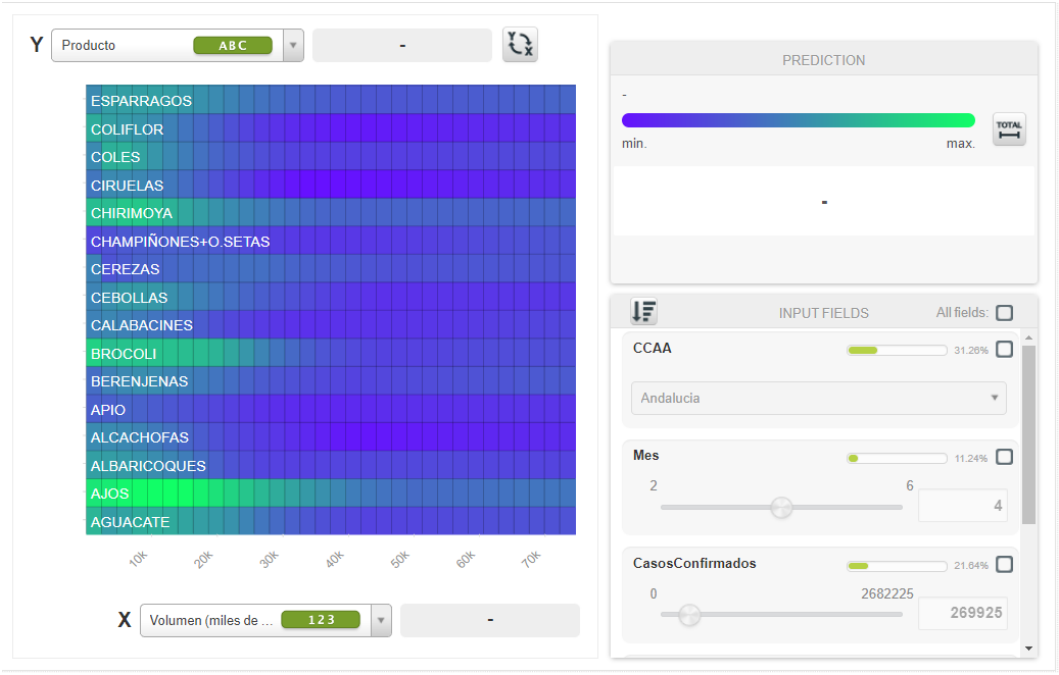
Podemos observar una ligera bajada de la tasa de mortalidad cuanto más volumen de productos se consumen. Algunos productos hacen que la tasa de mortalidad baje de manera más pronunciada que otros, pero en un ámbito general esta se ve reducida.



De la manera contraria, graficando los resultados de la comunidad autónoma de Murcia, obtenemos justamente los resultados inversos.



Sin haber sacado ninguna relación de los pasos realizados anteriormente, a continuación visualizamos la gráfica predictiva del volumen de productos consumidos pero esta vez a nivel nacional, sin seleccionar ninguna comunidad autónoma en específico.



A continuación, en el siguiente apartado se citan los resultados y conclusiones sacados del análisis y de las predicciones realizadas con BigML.

3.-Resultados y Conclusiones.

Tras haber intentado establecer relaciones entre posibles variables o intentar encontrar productos los cuales pudiesen haber ayudado en la recuperación del COVID-19 mediante su ingesta, hemos llegado a la conclusión de que con los datos proporcionados y tratados, no podemos establecer que haya una relación fija entre los productos consumidos y la tasa de mortalidad en las distintas Comunidades Autónomas, es decir, **nuestra hipótesis se rechaza**. El consumo de alimentos agrícolas en todo el país es muy equitativo, lo cual dificulta el análisis de los datos. En cuanto a las predicciones del apartado anterior y por qué cuando la tasa de mortalidad se reduce en algunas comunidades con la ingesta de alimentos y en otras aumenta, tampoco podemos establecer ninguna relación real, con lo cual lo damos como una simple coincidencia. Además, hay que tener en cuenta la posición geográfica de todas las comunidades autónomas de España, la demografía de estas y las posibles restricciones COVID que tuviesen a nivel autonómico en el confinamiento.

3.-1 Alcance.

En cuanto a la visión de futuro en esta hipótesis, ayudaría una base de datos de covid más actualizada a lo que era el dia a dia en el confinamiento, pero sabiendo que hubieron fraudes en los datos y cantidades de positivos y muertes falsas o que no se ajustaban a la fecha, esta hipótesis no creemos que se pueda probar como cierta nunca. La ingesta de frutas y verduras es crucial para el ser humano, pero no por consumir X producto en mayor cantidad se reducirá la tasa de mortalidad por coronavirus.