Minería De Datos

INFORMACIÓN SOBRE VEHÍCULOS ELÉCTRICOS

Diego Cordero Contreras

Enrique Albalate Prieto

Lucía De Ancos Villa

Pablo Del Hoyo Abad

Mohamed Essalhi Ahamyan

Índice:

1. Introducción

Los vehículos eléctricos han surgido con fuerza en los últimos años, principalmente debido a su colaboración con el medio ambiente. Estos vehículos utilizan una fuente energética que no contamina como la electricidad para cargar sus baterías, al contrario que otros coches de combustión que necesitan diésel o gasolina.

Es necesario expandir su uso a lo largo del mundo y en este trabajo queríamos encontrar información sobre los diferentes motivos que pueden llevar a un comprador a decantarse por el vehículo eléctrico. De esa forma, se desprende nuestro objetivo final: hacer que esa información sintetizada pudiera ser de utilidad para gobiernos o empresas que quieran ayudar a la creación de un futuro más sostenible mediante la mejora de las características que buscan los consumidores.

Dicho lo anterior, nos fue imprescindible fijar un alcance de datos, es decir, una serie de países y regiones seleccionadas a tener en cuenta. No fue viable hacer lo propio con el resto de sociedades mundiales entre otros motivos porque no todas ellas hacen pública este tipo de información.

Para poder considerar como válido un dataset, éste debía contener datos referidos a, al menos:

- Norte América
- Europa
- Asia central

2. Definición del problema

2.1. Proyectos pasados similares

Tras una exhaustiva búsqueda, logramos encontrar ciertos estudios o proyectos centrados a su vez en la identificación de patrones tanto en sociedades como en vehículos que marquen la diferencia en términos de aumento de compras de los últimos. Dos buenos ejemplos son:

- Identifying early adopters in Germany
- Relating consumers' information and willingness to buy electric vehicles

Aún habiendo adquirido gran conocimiento e inspiración desde con ellos, teníamos claro que eran mejorables. Por un lado, el primero de ellos se limita a Alemania y a un intervalo temporal que llega únicamente a 2014, lo que supone su condición de obsoleto ante el rápido crecimiento de los vehículos eléctricos. Por otro, en el segundo estudio (centrado en la India), se tratan únicamente aspectos demográficos como la edad, el sexo o el nivel de estudios de la población.

En nuestro caso nos enfocamos en diversos aspectos relacionados con la economía, pues es bien conocida como el aspecto principal para que un consumidor pueda elegir este tipo de vehículos por encima de otros.

2.2. Descripción de las hipótesis

Una vez establecidos el tema principal, alcance y realizada la documentación de trabajos similares, decidimos los puntos que debían haberse abordado en el nuestro para el momento de su finalización.

Las ideas que a todo interesado en el tema pueden surgirle y que tendrían que tener una confirmación o desmentido por nuestra parte posteriormente fueron:

1. *Se venden más vehículos por el aumento del rendimiento de los coches eléctricos*
2. *Se venden más vehículos por el aumento de los precios del carburante*
3. *Predicción de las ventas de vehículos eléctricos en base al precio de la electricidad y el número de puntos de carga*
4. *Se venden más vehículos en países con más PIB per cápita*

3. Fases del proceso KDD

3.1. Creación de VPS

Es importante citar al comienzo de la explicación del proceso KDD aplicado a nuestro proyecto que se contrató un servidor VPS a modo de almacén compartido por todos los miembros del equipo, donde se creó un contenedor Docker que contiene una imagen de la base de datos MySQL 8.0.

Esta base de datos fue la encargada de albergar las diferentes versiones y combinaciones de los dataset escogidos inicialmente como consecuencia del preprocesamiento de ellos. En síntesis, 3 esquemas principales clasifican dichas modificaciones:

- [RAW]: Contiene los datasets en el mismo estado en el que se descargaron u obtuvieron de forma automática.
- [SILVER]: Por su parte, maneja los datasets resultado de la limpieza y transformaciones estipuladas en el segundo punto del entregable 2. Toma como fuente de material aquello almacenado en el esquema [RAW].
- [GOLD]: Por último, adquiere los datasets limpios pertenecientes al esquema [SILVER] para después guardar en él mismo las tarjetas de datos (cuyas directrices de creación pueden leerse en el tercer punto del entregable 2) requeridas para contrastar las hipótesis antes lanzadas.

3.2. Selección de los datos

Tras un periodo de reflexión sobre cuáles serían aquellos datos que nos llevarían a obtener conclusiones acerca de las hipótesis formuladas con anterioridad, terminamos empleando técnicas de "webscraping" para hacernos de manera automática (ejecutando simplemente un script python) con una serie de datasets relacionados con:

Ventas por modelo de vehículo eléctrico: Este dataset fue extraído del portal especializado en automoción *MarkLines*. Cuenta con referencias desde el 2017 hasta el 2202, y se convirtió en la base sobre la que intentar responder a las preguntas anteriores por la fundamental importancia de su contenido.

Precio de los combustibles fósiles: En este dataset, es posible consultar los precios de la gasolina en una amplia lista de países para un intervalo temporal de entre 2015 y 2023.

Rendimiento de los vehículos eléctricos: Este dataset corresponde a una base de datos pública en la que se encuentran todos los coches eléctricos comerciales actualmente (2023) con ciertas sus características técnicas realmente útiles (batería, PVP de fábrica, etcétera).

Puntos de Carga de Vehículos Eléctricos: El conjunto de datos habla sobre la cantidad de puntos de carga destinados a vehículos eléctricos en los principales países del mundo en función del año. Es importante citar que, en este archivo, principalmente podremos diferenciar entre dos tipos de puntos de carga, los de carga rápida y los de carga lenta.

PIB per cápita: Este archivo nos servirá a la hora de manejar información relacionada con el producto interior bruto y su evolución temporal en diferentes países del mundo (en concreto 196), además de ciertos territorios más amplios como Centroamérica, la zona sur de Europa o las islas del Caribe, entre muchas otras. Los datos están surgen desde 1980 y de previsiones futuras hasta 2028.

Nivel de estudios: En él se refleja el nivel máximo por poblaciones conseguido en términos exclusivamente académicos.

Precio de la luz: Para los años comprendidos entre 2015 y 2023, contiene datos relativos al coste de la electricidad por megavatio/hora. Como curiosidad, muestra los datos de cada país por meses, en lugar de forma anual.

Niveles de CO2: Recoge registros de polución por países y otras zonas mundiales desde los 70 y hasta el 2021.

Para más información sobre los datasets originalmente escogidos, consultar entregable 1.

3.3. Limpieza y transformación

En todos los dataframes hemos tenido que estandarizar ciertos datos que serían de uso común, como es el caso de los países. Hemos elegido como idioma común el inglés, de manera que el nombre de los países deberá traducirse al inglés en caso de ser necesario. Además, cada dataframe en concreto ha tenido su proceso de limpieza y estandarización particular:

***Data_CO2*:** Se han modificado los datos relacionados con la emisión de CO2 con el objetivo de mantener exclusivamente los que están dentro de nuestro alcance, es decir, los comprendidos entre 2015 y 2023. Además, se eliminaron las ultimas líneas, ya que contenían información irrelevante en nuestro caso: datos globales en lugar de clasificados por países.

***Electricity_car_data_clean*:** Este dataframe contiene la información de todos los vehículos eléctricos a la venta en los últimos años, es decir, no contamos con los vehículos eléctricos históricos que lleven sin fabricarse mucho tiempo, sino que sólo los actuales.

Gracias al uso del api de Kaggle, nos ahorramos problemas derivados del manejo de datos nulos o mal recogidos durante la etapa de webscrapping. En adición, debido a que los datos estaban expresados de forma mensual, hubo que transformarlos en anuales valiéndonos de agrupaciones.

Electricity_all_countries: Al contener los datos de cada país por meses y nuestro alcance estar definido en años, decidimos hacer la limpieza calculando la media del precio por año. De esta forma, conseguimos obtener los precios de la electricidad por megavatio/hora clasificados por país y año.

Gasoline_Diesel_prices: Al haber utilizado la técnica de webscrapping, algunos caracteres estaban codificados en utf-16 mientras que nuestra base de datos estaba en utf-8, causando incompatibilidades con el resto. Eliminamos estos caracteres y cambiamos el tipo de datos a los caracteres numéricos.

***Level_Of_Studies*:** Borramos las columnas que no podremos diferenciar, en este caso la distinción entre hombres y mujeres en sus estudios, ya que no sabemos el sexo del comprador de un vehículo. A su vez, corregimos el símbolo de porcentaje que aparecían en él y hemos transformado los valores de las columnas a enteros para después trabajar de manera más fácil.
Por último, se transformó el nombre de las columnas suprimiendo el carácter '__' que aparecía al final del nombre de ellas.

***Model_Per_Year*:** La primera tarea que se tuvo que realizar fue la de establecer un valor común para los nulos. En el conjunto de datos, este tipo de valores aparecen representados tanto por el símbolo '-' como por 'N/A'. A continuación, eliminamos el separador de millares que contenían los números porque obligaba a que el tipo de datos fuera 'category' cuando, en realidad, estábamos interesados en que fuesen de un tipo de dato numérico. En este caso concreto, como la página web donde se han obtenido los datos es estadounidense, utilizaban la coma. Además, descubrimos que, para algunos modelos en un país y año determinados, el número de unidades vendidas era negativo. Como carecía de sentido, decidimos sustituirlos por nulos.

Existe una columna denominada PowerTrain que indica el tipo de motor que utiliza el modelo. Sin embargo, el nivel de granularidad utilizado es mayor al que necesitamos para comprobar la hipótesis. A modo de ejemplificación, la clasificación diferenciaba entre híbrido enchufable e híbrido no enchufable, pero nosotros tratamos ambos casos como eléctricos. Es por ello que decidimos crear una nueva columna denominada "isEv" que indicase si el modelo es eléctrico o no.

Por último, decidimos sustituir los valores nulos por cero ya que, para este estudio, estamos interesados en la cantidad de vehículo vendidos en un país y en un año determinado y no tanto en los modelos concretos. Por lo tanto, al agregar los datos, el valor de cero permite que ignoremos los nulos.

***PIB_per_capita*:** Tras la obtención de los datos relacionados con el producto interior bruto per capita de países y diversas zonas geográficas, se realizó una poda de columnas a modo de filtro de los años manejados. En concreto, únicamente se deseaba tratar datos desde 2015 hasta 2023, pero fue necesario tener en cuenta a su vez registros de los 5 años anteriores y las previsiones para los 5 siguientes, como se explicará a continuación.

Debido a que al estado inicial del dataframe le faltaban ciertas mediciones y estas estaban sustituidas por la cadena de caracteres 'no data', se pensó que una buena solución al problema sería en primer lugar establecer todas las apariciones de dicha cadena por el valor nulo de la librería de Python pandas (nan) y en segundo lugar cambiarlos de forma automática por la media del producto interior bruto del país al que se refiere la fila del dato nulo dentro del rango temporal más arriba mencionado (desde 2010 hasta 2028). El objetivo de todo este proceso fue la generación de valores lo más realistas posibles.

Puntos_De_Carga: En este dataset se borraron algunas columnas de las que se valía la página web desde donde se descargaron los datos para su control pero que en nuestro caso no son de utilidad. Estas columnas son: category, parameter, mode, unit.

A esta eliminación le acompañaron las columnas 'value' con valores decimales, se atribuyen a errores en la recogida de datos.

Por último, se ha elegido para cada país el rango de años entre 2015 y 2022.

3.4. Tarjetas Hipótesis

3.4.1. Primera tarjeta de datos

Para la comprobación de nuestra hipótesis número 1, creamos esta tarjeta de datos que contiene todos los datos extraídos de los datasets anteriores que fueron considerados de más relevancia para este caso concreto. Éstos son los relativos a la venta de coches eléctricos y sus características, exclusivamente.

Para crear esta tarjeta tuvimos que unir información que se encontraba en dos datasets. Ambos hablaban de vehículos: uno, de todas las ventas de vehículos en países; y otro, de las características de coches eléctricos de forma especializada. En el primer datasets agrupamos todos los datos en función del modelo, quedando solo el modelo y las ventas que él mismo tuvo en sus respectivos años. Por otro lado, en el segundo dataset tan solo eliminamos la marca, quedándonos así con el modelo y otros datos que más nos interesaban del mismo.

Para juntarlos hubo problemas, pues en cada dataset las versiones de los propios modelos venían redactadas de formas distintas. Esa es la razón por la que, para cada modelo de coche que tuviera más de una versión, se hizo la media de sus características y se añadió dicho modelo de forma única (ejemplo de ello es el caso del tesla model 3).

De esta manera nos quedamos con el siguiente diccionario de datos:

Nombre del campo	Tipo de dato	Descripción
Model	String	Indica el modelo de coche del que hablan las características.
Year	Int	Es el número de ventas que el modelo tuvo ese año mundialmente.
AccelSec	Float	El tiempo en segundos que el coche tarda en acelerar desde 0kmh hasta 100kmh.
TopSpeed_Kmh	Int	La velocidad máxima alcanzable con el coche en kmh.
Range_Km	Int	Autonomía de la batería en Km bao el ciclo WLTP.
Efficiency_Kwh	Float	Indica la carga del vehículo (con ello el tiempo de recarga) en kwh.
Seats	Int	Número de asientos en el vehículo.
PriceEuro	Int	Precio de venta al público en euros.
FastCharge_KwH	Float	Indica, si tiene, los watios a los que el coche puede cargar con un fast charger.
RapidCharge	Boolean	Indica si tiene carga rápida.
PowerTrain	String	Indica las ruedas del coche que traccionan, delanteras, traseras o 4x4.
PlugType	String	Indica el tipo de enchufe que tiene para cargar.
BodyStyle	String	Indica la carrocería del coche.
Segment	String	Indica el segmento en el que está el coche.

3.4.2. Segunda tarjeta de datos

Por su parte, se creó una tarjeta de datos con la finalidad de agrupar la información relevante para aceptar o refutar la segunda hipótesis. Para ello, seleccionamos los datasets correspondientes a los precios de la gasolina y diésel y a las ventas de coches eléctricos y sus modelos en diferentes países del mundo, ambos del esquema SILVER.

El proceso consistió en unir las dos tablas mencionadas, por medio de la columna que contiene los países. Además, solo hemos conservado el periodo de años común en ambas tablas (2017-2022).

Finalmente, obtenemos el siguiente diccionario de datos:

Nombre del Campo	Tipo de Dato	Descripción
Country	String	Contiene los nombres de los países manejados y se utilizará para identificarlos.
CochesVendidos_[year]	Double	Contiene la cantidad de vehículos oficialmente vendidos para un año concreto ([year]) y un determinado país, es decir, el que aparece en la columna “Country”.
Gasolina_[year]	Double	Almacena el valor del precio de la gasolina para un año concreto ([year]) y un determinado país, es decir, el que aparece en la columna “Country”
Diesel _[year]	Double	Funciona como las columnas Gasolina_[year], pero en este caso almacenando el coste de Diesel.

3.4.3. Tercera tarjeta de datos

Esta hipótesis acabó variando ligeramente. El fin último del cambio fue introducir una hipótesis de distinta naturaleza a las otras existentes, así que quisimos tratar de crear un modelo que predijera el número de puntos de carga que se instalarán en un país dependiendo de las ventas de vehículos eléctricos el año anterior.

Se unieron las tablas que contienen las ventas de coches y de los puntos de carga. Primero, se agruparon las ventas creando una nueva tabla con 2 tipos: Hybrid y Electric en sustitución de a la columna PowerTrain en el dataset de ventas de coches. Además, se tuvieron que pivotar ambas tablas: en la tabla de ventas de coches los años aparecían como columnas; y sin embargo, ahora se modificó creando una nueva columna llamada year donde aparece el año estudiado junto a otra columna llamada Sells, donde aparece el número de ventas.

Finalmente, se fusionaron los dos datasets utilizando como nexo las columnas Country y year (país y año).

Además, se añadieron los datos del precio de la electricidad en cada país por año.

Los datos resultantes pertenecen al intervalo de años entre 2017 y 2022. Sin embargo, queremos realizar un modelo predictivo y teniendo una serie temporal en cada país asumimos que puede existir una evolución en la venta de los vehículos eléctricos. Por tanto, hemos añadido una nueva columna indicando las ventas en el año anterior a costa de sacrificar los valores del año 2017.

El diccionario de la tarjeta de datos es el siguiente:

Nombre del Campo	Tipo de Dato	Descripción
Country	String	Almacenará el nombre de todos los países manejados y que servirá para identificarlos
Type_Vehicle	String	Contiene el tipo de vehículo que se ha vendido, puede ser Hybrid o Electric
year	Int	Contiene el año de estudio en la fila
Fast Charging Point	Int	Es el número de puntos de carga rápidos distribuidos en el país
Slow Charging Point	Int	Es el número de puntos de carga lenta distribuidos en el país
Price Electricity	Int	El precio de la electricidad en el país
Sells_last_year	Int	Es el número de ventas de coches del año anterior
Sells	Int	Es el número de ventas de coches

3.4.4. Cuarta tarjeta de datos

Con el objetivo de contar únicamente con los datos necesarios para contrastar la hipótesis de manera cómoda, clara y rápida se decidió unir la tabla que contiene datos referentes a ventas de coches eléctricos y sus modelos en diferentes países del mundo junto con aquella destinada a reflejar la evolución del producto interior bruto de dichas zonas.

La columna que sirvió de nexo entre ambas fue en la que se detallaban los países a los que los datos de una determinada fila pertenecían. Además, se recortó finalmente el intervalo de años entre los que se puso el foco, desde 2017 hasta 2022. La estructura de la tarjeta de datos desarrollada es la siguiente:

Nombre del Campo	Tipo de Dato	Descripción
Country	String	Almacenará el nombre de todos los países manejados y que servirá para identificarlos
CochesVendidos_[year]	Double	Contiene la cantidad de vehículos oficialmente vendidos para un año concreto ([year]) y un determinado país, es decir, el que aparece en la columna “Country”
PIB_[year]	Double	Contiene el valor del producto interior bruto oficial para un año concreto ([year]) y un determinado país, es decir, el que aparece en la columna “Country”

- [year] representa a cualquier año comprendido entre 2017 y 2022

En adición se requirió un tratamiento adicional que afectó al contenido de la nueva tarjeta de datos, pero no a su estructura. Fue detectada la ausencia en ella de ciertos países que eran comunes a las dos tablas que sirvieron de fuente de información, por lo que acabó comprobándose que tenían ciertos detalles diferentes en su redacción dentro de la columna “Country” de dichas tablas. Esto se resolvió usando un diccionario que aplicara reemplazos una de ellas en la columna referida a países (por ejemplo, desde "China, People's Republic of" a “China”).

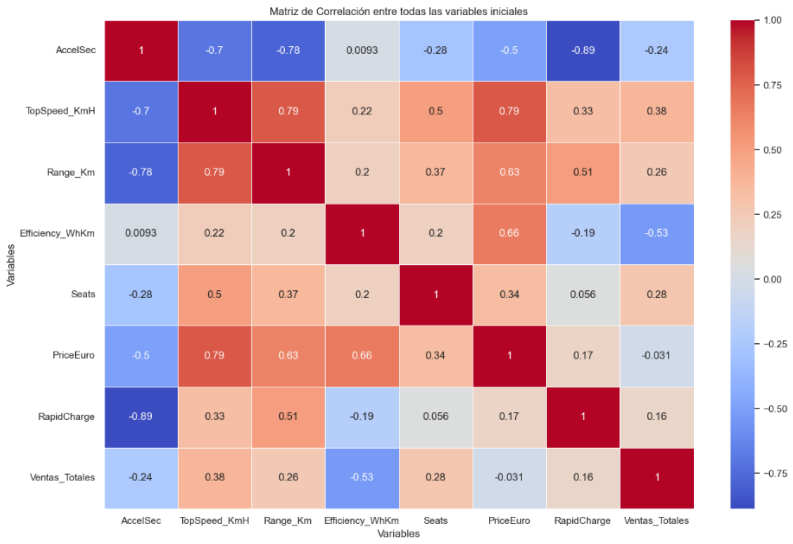
Finalmente, se rellenaron ciertas celdas de filas de la tarjeta de datos final en las que aparecía un 0 representando los coches vendidos de algunos años concretos (únicamente ocurría para Puerto Rico y Uzbekistán). Como valor para completar la cada respectiva sustitución se empleó la media de coches vendidos no nulos (en este caso marcados con 0) de esa misma fila.

3.5. Conocimiento extraído

3.5.1. Hipótesis 1

Como explicábamos anteriormente en otros documentos, la tarjeta de datos de esta hipótesis enlazaba las ventas de modelos entre 2017 y 2022 de vehículos eléctricos con las especificaciones de estos mismos vehículos. Es interesante para este punto ver el notebook de nuestro repositorio donde se explica todo al detalle.

La idea con estos datos es probar nuestra hipótesis que dice que se venden más coches eléctricos con respecto mejoren sus características. . Para ello, probaremos la correlación que exista entre las variables de las especificaciones de cada coche por separado con las ventas del mismo:

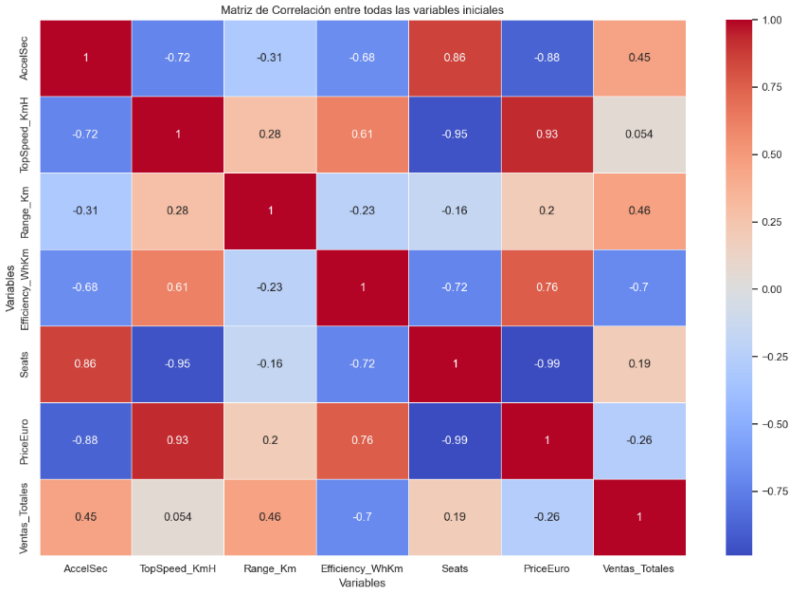


Como podemos ver, esto no nos lleva a ninguna conclusión lógica pues no existe ningún grado de correlación relevante entre venta y ninguna especificación. Una relación que cabe a destacar es la relación entre ventas totales y AccelSec, que pareciera inversamente proporcional, pero esto es debido a que la aceleración es mayor a menor sea este dato, pues se expresa en los segundos que tarda en pasar de 0 a 100 kph. Lo más destacable dentro de esta tabla sería la correlación que hay entre las ventas totales y la velocidad máxima, ¿se compran más coches con respecto a más alta sea su velocidad máxima? En principio, podríamos decir que sí y esto se apoyaría en que también se compran más coches con respecto a mayor aceleración tengan, pero también vemos que se compran más coches con respecto a más asientos tengan, con lo que podríamos estar viendo aquí un conflicto de “públicos”.

Dentro de esta tarjeta se hablan al mismo nivel de coches con un carácter deportivo que de coches con un carácter familiar o incluso empresarial; es por ello que vemos variables tan poco correlacionadas, pues poco o nada tienen que ver las intenciones de la persona que compra un deportivo que de la persona que compra un coche familiar. Es por esto que decidimos dividir esta tarjeta de datos en cinco tarjetas de datos, una por segmento de coches, que, si bien no dividen los coches por deportivos, familiares... Los divide por tamaño de la carrocería, lo que hace que los deportivos estén encasillados únicamente

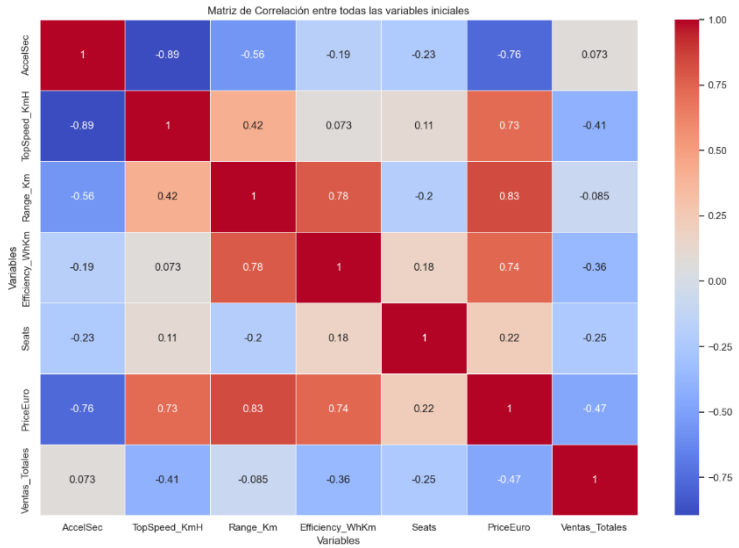
en el segmento D y E (Berlinas de grandes dimensiones como serían en este caso los Teslas con carácter más deportivo).

Correlaciones en el segmento B:



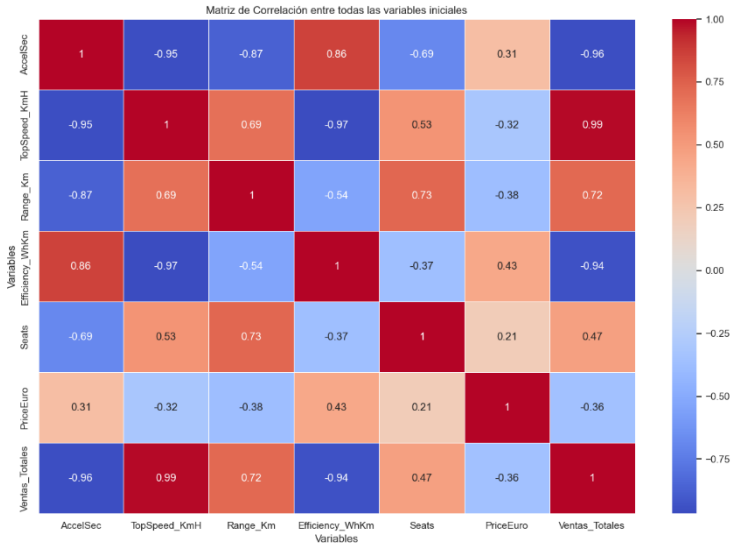
Ahora sí podemos ver algunos grados de relación más interesantes. En el segmento B se engloban los coches compactos, de carrocería más pequeña y por tanto más orientados a la conducción cómoda por ciudad. Podemos ver que las variables con más correlación son siemper las que más debería importarle a un conductor urbano, el precio la autonomía y una carga normal. Más detalles en el notebook

Correlaciones en el segmento C:



El segmento C es un segmento de coches medianos y sobre todo orientados a un público mixto que quiere tanto hacer viajes como poder conducir en ciudad, y esto se ve reflejado en reflejado en las correlaciones, salvo en autonomía. Esto es principalmente debido a que en los coches que hemos estudiado en este segmento se incluyen coches hechos por compañías como Volkswagen o Nissan que ya tenían muchas ventas antes de hacer eléctricos, con lo que el comprador muy probablemente se guíe más por el historial de estas marcas y su fama que por los requisitos técnicos de los mismos.

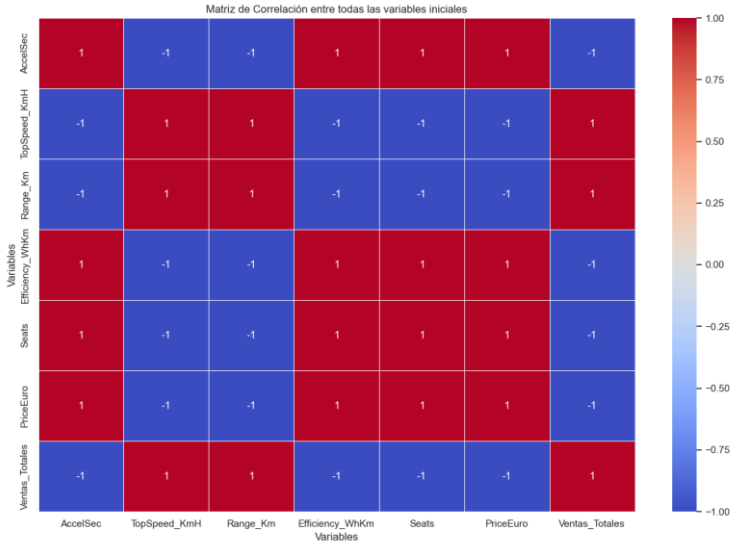
Correlaciones en el segmento D:



Llegamos al Segmento D, el segmento de las berlinas medias y deportivas. En este caso, si bien es cierto que se incluyen vehículos con un carácter más deportivo, hay que decir que son berlinas deportivas, que no buscan el rendimiento en circuito únicamente si no que sobre todo buscan el lujo y la comodidad de hacer largos viajes en carretera más rápidamente. Esto se ve reflejado perfectamente en el desinterés en el precio y en el interés por velocidad máxima, asientos y sobre todo.

Tenemos que hacerm ención de AccelSec, que parece rompe estos esquemas, pero como expresamos en el [notebook], esto autonomía simplemente se debe a que hay un modelo en concreto muy popular que tiene más ventas.

Correlaciones en el segmento F:



Entramos en el segmento F, segmento de berlinas mayores de 5 metros donde solo tenemos dos modelos disponibles, el Tesla model Y y el Tesla model X, con lo que las correlaciones son algo engañosas. No obstante, podemos ver casi lo mismo que en el segmento anterior al ser del mismo carácter, se busca principalmente velocidad y autonomía antes que precio o eficiencia a la hora de cargar el vehículo.

Podríamos añadir el segmento N, pero solo tenemos un vehículo dedicado exclusivamente a uso industrial, y sería precipitado sacar conclusiones con sus datos únicamente

En conclusión ¿Se venden más vehículos eléctricos con respecto mejores sean sus características? A grandes rasgos, podríamos decir que sí, pero para ello primero habría que afinar la hipótesis ¿A qué nos referimos con mejores características? Porque si hablamos de si los compradores buscan cualquier coche que sea lo mejor en todas sus características, la respuesta sería un rotundo no. No obstante, si hacemos una particularización por sector y nos referimos a que los compradores buscan las mejores características que necesiten, entonces podremos decir sin temor a equivocarnos que sí.

En este análisis de datos hemos podido comprobar que, en su mayoría, los compradores eligen su coche en función de las características que más les interesan. Es cierto que hemos visto también que hay casos en los que la fama y la marca predomina sobre cualquier característica, pero en cualquier caso hablamos de casos concretos y no podríamos decir que esto rompa la con la media de compradores que solo se fijan en las características.

3.5.2. Hipótesis 2

Los datos de los que partimos para aceptar o refutar la hipótesis sobre si se venden más vehículos por el aumento de los precios del carburante, son los que encontramos en la tarjeta de datos para dicha hipótesis. Esta tarjeta cuenta con los datos vendidos y los precios tanto de gasolina como Diesel de los años 2017 a 2022.

La idea principal para validar esta hipótesis es estudiar por separado los coches vendidos y el precio de carburante de los países disponibles. Esto se llevará a cabo mediante algoritmos de aprendizaje no supervisado, que darán lugar a clusters o grupos. De tal forma que si el grupo de países en el que el precio de carburante es más elevado coincide con el grupo de países en el que más coches se venden, podremos aceptar la hipótesis y de lo contrario rechazarla.

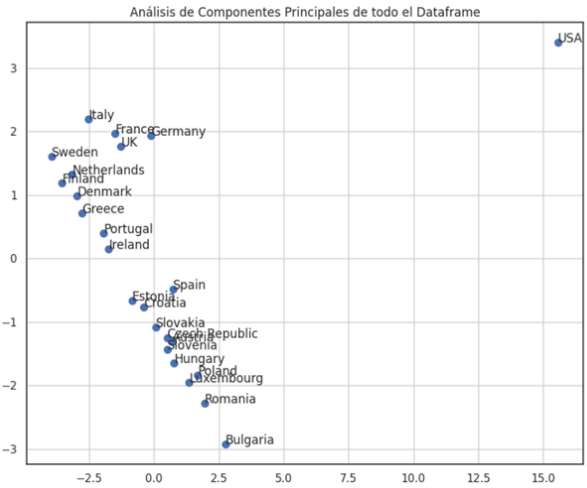
Procesamiento previo de información

Comenzamos con un breve estudio de los datos de partida, que nos ayudará posteriormente al aplicar los algoritmos

Utilizamos la técnica de análisis de componentes principales (PCA) para reducir la dimensionalidad de nuestra tarjeta de datos. Esto nos permitirá retener las características del conjunto de datos que contribuyen más a su varianza (con mayor variabilidad). Consecuentemente, mantendremos un orden de bajo nivel, es decir conservamos la mayor cantidad posible de información.

La visualización de los datos nos será mucho más cómoda, ya que PCA busca la proyección en la que los datos queden mejor representados en términos de mínimos cuadrados. Además, mitigamos el problema de multicolinealidad, es decir el problema de que los datos estén muy correlacionados.

Como nuestro propósito principal en este punto es la visualización, utilizaremos dos componentes principales:



Observamos que el ratio de variabilidad, especialmente por el eje Y, es bueno. Además, podemos ver como muchas de las variables son muy parecidas, exceptuando Estados Unidos o Bulgaria que son la que más se diferencia del resto.

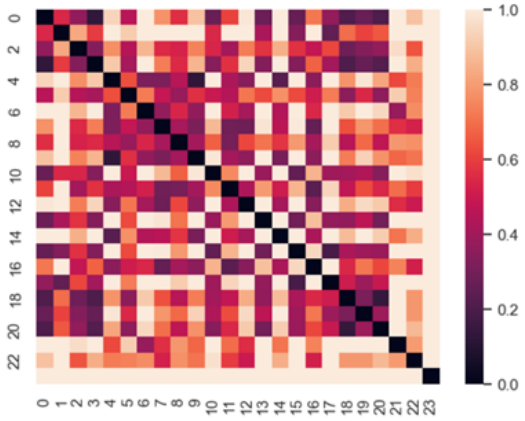
Transformaciones previas

Con nuestro objetivo de estudiar por separado el número de coches vendidos y el precio de carburante, preparamos nuevos dataframes para cada uno. Transformaremos los datos de ambos a una misma escala, en nuestro caso optando por la operación MinMaxScaler, que transforma cada variable para conseguir que los elementos estén un rango determinado [0-1].

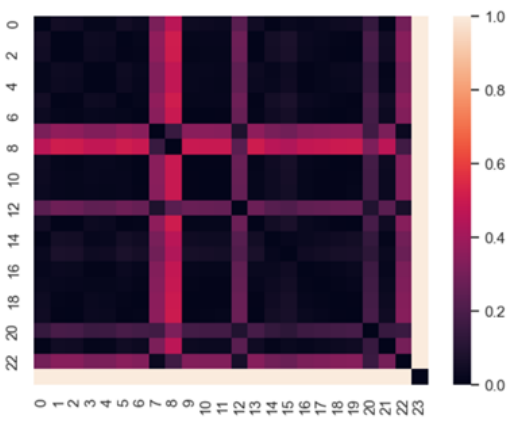
Esto es necesario porque diferentes elementos de nuestras características pueden tener ordenes de dimensión diferentes y hay algoritmos de clustering que son sensibles a esta escalada de variables.

Adicionalmente, volvemos a calcular la PCA de estos nuevos conjuntos de datos y la matriz de similitud de cada uno. Esto se ha llevado a cabo para obtener de nuevo una mejor representación de los datos y para saber cómo de parecidas son las variables entre sí, debido a que estos datos serán necesarios posteriormente para los algoritmos de clustering

Gasolina y Diesel:



Número de vehículos vendidos:



En la matriz de similitud de Gasolina y Diesel podemos observar que los datos tienen mayor parecido entre ellos.

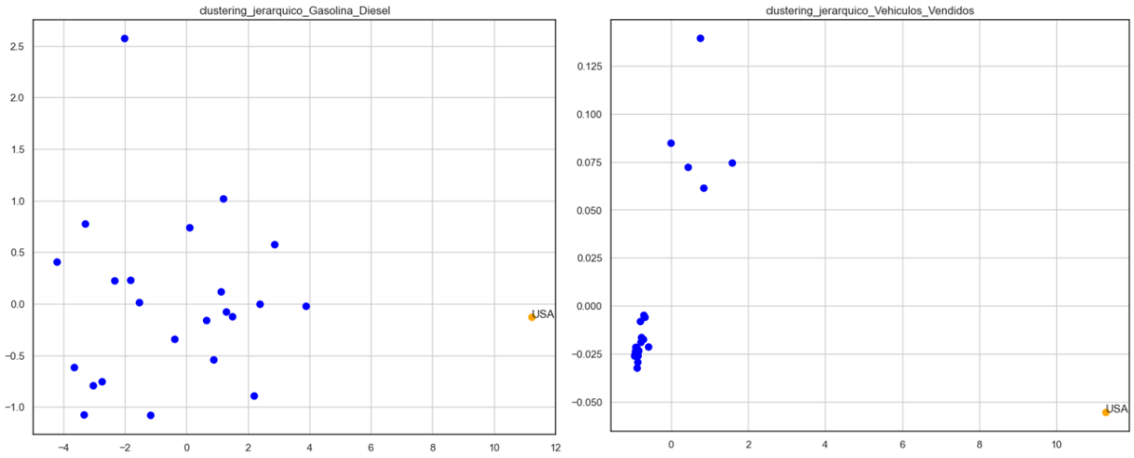
Clustering

Se han implementado diferentes algoritmos de clustering con el objetivo de comparar los resultados y elegir el de mayor coherencia de agrupación según el coeficiente de Silhouette.

Clustering Jerárquico

En este caso, para el dataset de combustible obtenemos un coeficiente de Silhouette de 0’671 y para el número de coches 0’902. En este enfoque la cantidad óptima de clusters es 2, esto permite dividir los países estudiados en únicamente dos extremos: aquellos con mayor o menor coste de combustible (para clusters generados a partir del dataset sobre el precio de gasolina y Diesel); y aquellos punteros o no en compras de vehículos eléctricos (para clusters generados a partir del dataset de número de vehículos). Esto facilitó las futuras interpretaciones, pues desde ese momento bastaba con observar si en las agrupaciones creadas de países los que estaban en un extremo para un dataset lo estaban también en el otro.

A continuación, se muestran las representaciones graficas de ambos apoyándonos sobre las representaciones de PCA:

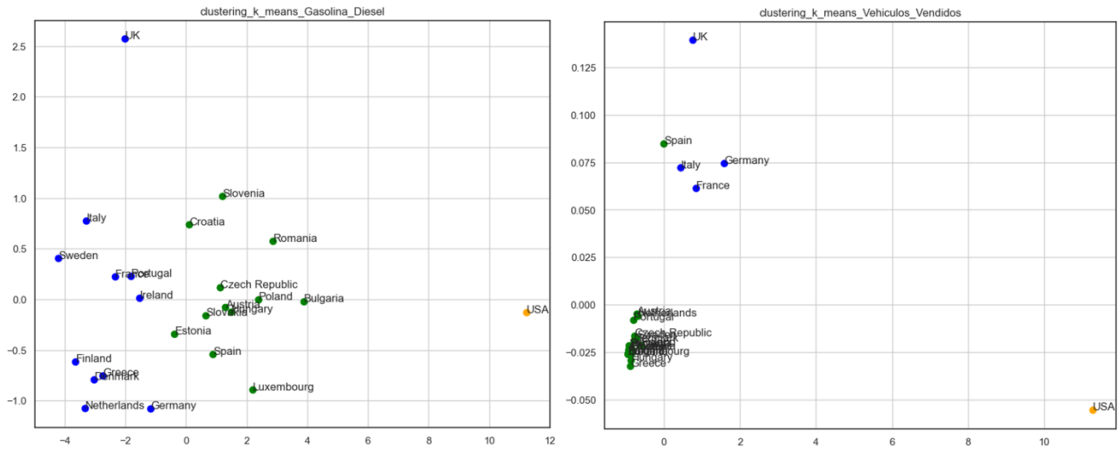


Totalmente al contrario de lo que propone la hipótesis, observamos que el grupo de países en el que más coches se vendes (solo contiene a EEUU), es también el grupo en el que los precios de carburantes son más bajos. Estas conclusiones hacen que la hipótesis sea refutada.

Clustering K-Means

Para este algoritmo de clustering, los resultados obtenidos de coeficiente de Silhouette no son tan buenos como los anteriores. Para el dataset de carburantes obtenemos 0’553 y para el de vehículos tan solo 0’141.

Además, el número óptimo de clusters para ambos datasets en este caso es de tres, asimismo los grupos generados en cada caso son diferentes. Por lo tanto, esto nos complica el análisis de resultados.



De nuevo, el grupo en el que más coches se venden no coincide con el grupo con el mayor precio de combustibles, por lo que también refutamos la hipótesis.

Concluimos también con que el algoritmo de clustering jerárquico obtiene mejores resultados y es capaz de agrupar mejor nuestros datos basándonos en los coeficientes de Silhouette obtenidos.

Clasificación adicional

Con el propósito de hacer un trabajo más completo y demostrar la comprensión de todos los temas vistos en la asignatura a lo largo del cuatrimestre, decidimos aprovechas las etiquetas que incluye clustering a los datos (a todos los miembros de un mismo cluster les pone la misma etiqueta, siendo esta diferente para cada grupo). Dichas etiquetas nos han permitido implementar modelos de aprendizaje supervisado, concretamente de clasificación.

Planteamos seleccionar uno de los países de los que conocemos su número de coches vendidos cada año, pero no el precio de combustible, por ejemplo, Bélgica. Por tanto, este país estará excluido del clustering debido a que la tarjeta de datos de la que partimos para realizarlo contiene a los países de los que conocemos tanto el número de coches vendidos como los precios de carburantes. Con esta metodología evitamos cualquier tipo de data leakage.

Utilizamos la clasificación que nos proporciona el clustering Jerárquico, ya que los resultados obtenidos han resultado ser mejores. Específicamente, haciendo uso del dataframe con los datos del número de coches vendidos en cada país por año, que ahora además incluye una columna “etiqueta”, la cual es 1 si pertenece al grupo de países en los que se venden menos coches y 2 si pertenece a los que venden más coches.

Con este conjunto de datos podemos entrenar y validar un modelo sencillo de clasificación aplicando el algoritmo Gaussian Naive Bayes. Se encargará de clasificar los países de entrada en grupos de países en los que se venden pocos o muchos coches. Comprobamos que el modelo es bueno para predecir, en este caso la precisión es 1, por lo tanto, es un buen modelo.

Posteriormente a partir de los datos de coches vendidos por año en Bélgica, podemos clasificar este país en el grupo que vende menos o más coches. El modelo lo clasifica dentro de los que venden menos.

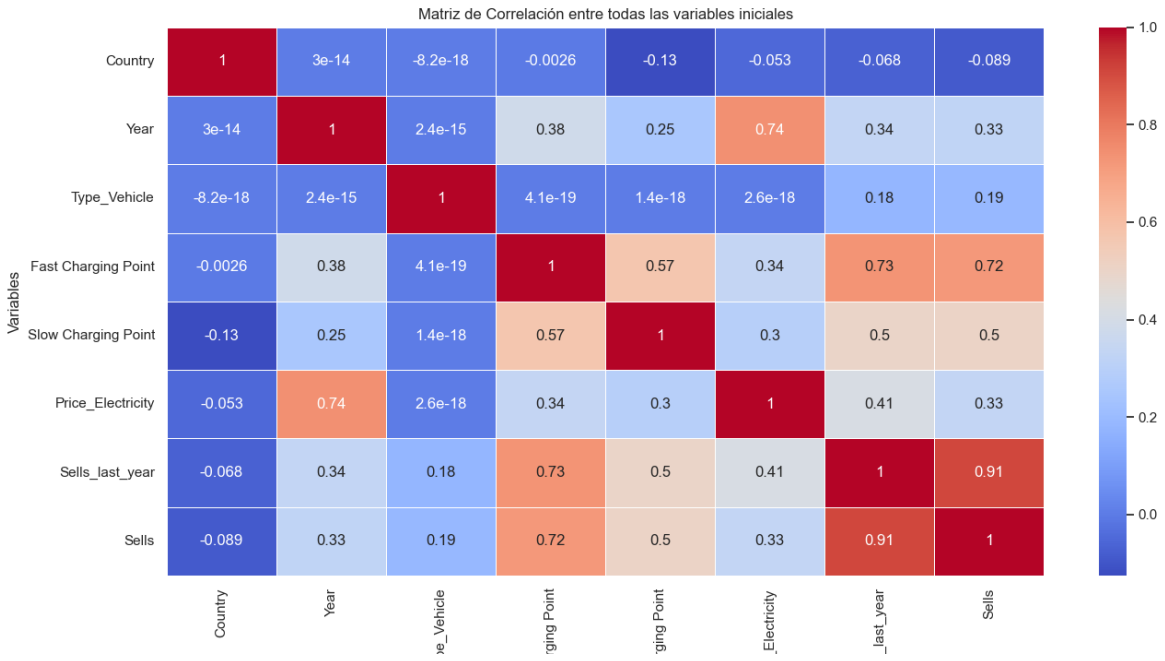
Como conclusión, podemos añadir que este modelo sería útil en el hipotético caso que algún empresario interesado estuviera pensando en que país sería rentable abrir un concesionario de coches eléctricos.

3.5.3. Hipótesis 3

En esta hipótesis explicaremos si se puede predecir las ventas de vehículos eléctricos en base a las condiciones del país para adaptar la tecnología eléctrica. Estas características que mediremos son: puntos de carga de vehículos eléctricos y el precio de la electricidad en el país.

Creemos que puede existir una correlación entre el soporte que da un país con los puntos de carga de cara a que un comprador se decante por un vehículo eléctrico. Es lógico pensar que los compradores decidan en base a las facilidades que les brinde el país para adaptar esta tecnología. Podemos encontrar desde ayudas para la compra de vehículos eléctricos hasta la facilitación en la instalación de los puntos de carga que es lo que estudiaremos.

En esta documentación, trataremos de ser lo más breves, si se necesita más información, acudir al repositorio en busca del notebook que desarrolla de manera más extensa esta hipótesis 3.



Correlaciones entre variables:

- Existe una fuerte correlación entre las ventas del año anterior (Sells_last_year) y Sells con un valor de 0.91.
- Además, también tenemos una buena correlación con los puntos de carga rápida, esto nos indica que este factor es relevante en la decisión de los clientes a la hora de decantarse por la compra de un vehículo eléctrico.
- Aunque con una correlación mucho más débil que los puntos de carga rápida, los puntos de carga lenta tienen cierta correlación con las ventas
- El precio de la luz no tiene tanta correlación con la compra de vehículos. Esto puede deberse a que entre los países que estudiamos apenas hay diferencias en el precio de la luz que lleven a decantarse por este modelo de vehículo.

- Por último, el país, el tipo de vehículo y el año no estan correlacionadas con las ventas. Estos valores son de control para que conozcamos el contexto del resto de los datos, era esperable que no fuera a existir demasiada correlación.

Modelos

Se han elegido los modelos de regresión para plantear este modelo. En concreto kNN, RandomForest y DecissionTree se entrenarán y comprobaremos su efectividad en este caso. Para ello, utilizaremos el error cuadrático medio como medida de rendimiento, al ser un problema de regresión, debemos tratar de acercarnos lo máximo posible a la solución y esta métrica es una de las mejores para esta tarea.

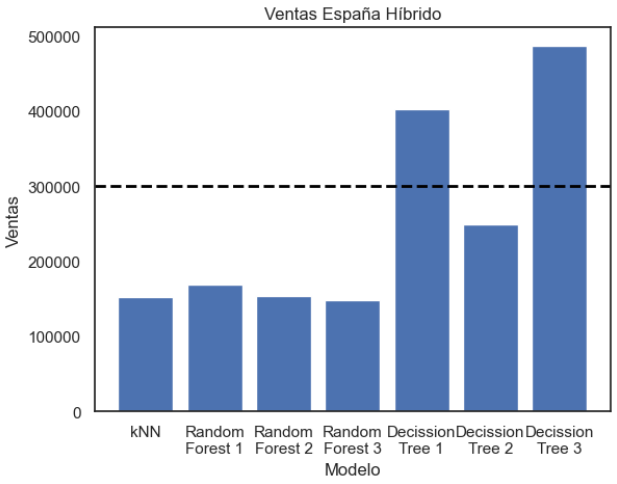
Se ha realizado un entrenamiento, donde se han obtenido los 3 mejores modelos de RandomForest, DecissionTree y el mejor de kNN. Con esos modelos, queremos tratar de aplicarlos en un caso del mundo real, por ello, consultando fuentes online hemos establecido 3 nuevos casos de prueba, que serán:

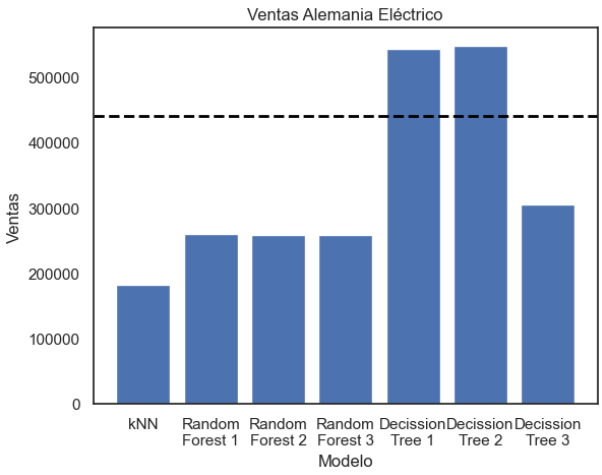
1. **España**: predicción de ventas de vehículo híbrido en 2023
- Ventas de vehículos: **300.000** (entre enero y junio fueron 150.000, suponemos el doble) Fuente: ANFAC
- Puntos de carga: RACE
    - Puntos de carga rápida: 3400
    - Puntos de carga lenta: 13600

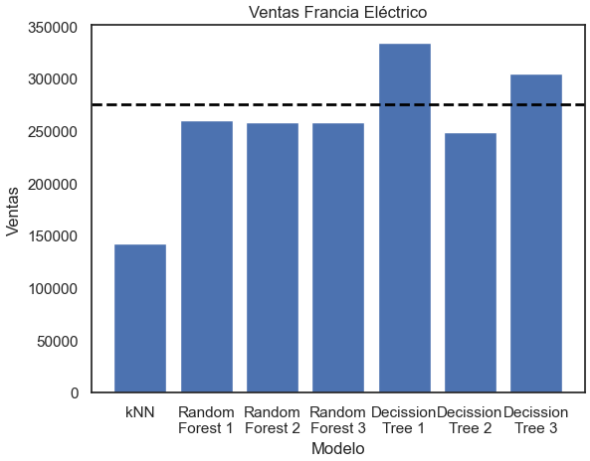
2. **Alemania**: predicción de ventas de vehículo eléctrico en 2023
- Ventas de vehículos: **440.000** (entre enero y junio fueron 220.000, suponemos el doble). Fuente: KBA
- Puntos de carga: Agencia Federal de Redes
    - Puntos de carga rápida: 18500
    - Puntos de carga lenta: 79000

3. **Francia**: predicción de ventas de vehículo eléctrico en 2023
- Ventas de vehículos: **274.000** (entre enero y junio fueron 137.000, suponemos el doble). Fuente: Plateforme automobile
- Puntos de carga: AVERE
    - Puntos de carga rápida: 13700
    - Puntos de carga lenta: 88200

Tras realizar el entrenamiento y obtener los mejores modelos, hemos construido una tabla que resume los resultados obtenidos. La franja discontinua horizontal representa el valor esperado.







Conclusiones

Una vez expuestos los resultados, podemos proceder a un análisis de los resultados.

Si midiéramos los resultados por el valor obtenido por el error cuadrático medio, podríamos pensar que lo que hemos hecho da lugar a un modelo que es incapaz de predecir ningún dato.

Sin embargo, hay que pensar que los valores de las ventas con las que estamos tratando son muy grandes por lo que un valor grande no representa exactamente que el modelo sea poco fiable. Sin embargo, si podemos comparar los modelos entre sí en base a dicho valor.

Primero, podemos ver la diferencia entre usar GridSearch y RandomizedSearch (con optuna no hemos obtenido el error para el mejor modelo), el error cuadrático medio que encontramos en GridSearch es mucho menor que en RandomizedSearch. Es algo que se puede considerar obvio, se puede explicar por el mayor tiempo de cómputo para crear este modelo, que al explorar más soluciones es capaz de encontrar mejores opciones, en GridSearch se han realizado 27000 iteraciones y en RandomizedSearch apenas 5000.

En segunda instancia, podemos comprobar lo fiable que es el modelo en base a los casos de prueba que se han generado para España, Alemania y Francia.

Estos casos han sido generados en base a información encontrada en distintos medios de comunicación y, si bien son estimaciones en base a datos recopilados hasta el momento en el año (por ejemplo las ventas de coches desde enero hasta junio), puede servir como indicativo de lo fiable que es nuestro modelo.

En este caso, predomina DecissionTree, el cual vemos en los gráficos que se queda más cerca por lo general de la línea horizontal divisoria que representa el valor esperado para cada país.

Además, sorprende el hecho de que kNN sea un modelo que genere un error porcentual absoluto medio tan elevado en comparación al resto de modelos. Es posible que se deba a las características irrelevantes para su entrenamiento, el precio de la luz, con apenas correlación puede haber sido un impedimento para obtener mejores resultados. Otro factor puede haber sido el problema de tener muchos países que venden muy pocas unidades de vehículos eléctricos, países como Grecia o Austria venden pocas unidades en comparación con Alemania o Francia, por extensión del terreno y por condiciones económicas.

Sin embargo, debemos cuantificar estos valores, en la siguiente tabla, mostraremos otra métrica que es más comprensible para el usuario y que se puede combinar con el error cuadrático medio, en este caso el error porcentual absoluto medio.

Modelo	Error porcentual absoluto medio
kNN	52%
RandomForest 1	30%
RandomForest 2	32%
RandomForest 3	32%
DecissionTree 1	26%
DecissionTree 2	17%
DecissionTree 3	34%

Finalmente, hemos conseguido obtener un modelo con un 17% de error. Puede parecer poco, pero tomando en consideración el proyecto propuesto, la poca cantidad de datos y la dificultad en su obtención, nos lleva a pensar que una empresa capaz de conseguir un dataset con mayores registros de ventas de coches y agregándole algunos datos más, sería capaz de obtener un buen modelo que le otorgue la información de cara a su planificación empresarial, tendrá una gran ventaja con sus rivales al saber el crecimiento del sector.

3.5.4. Hipótesis 4

Contexto

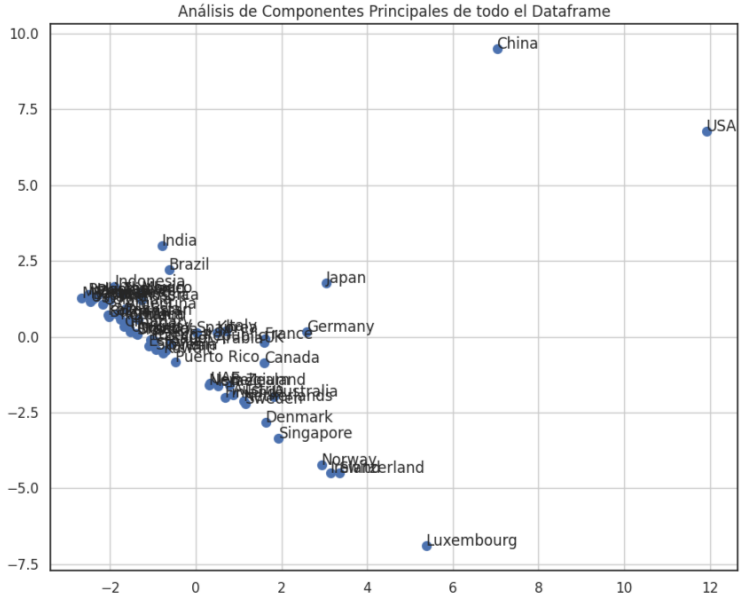
Como fue descrito con anterioridad, la tarjeta de datos empleada en la comprobación de la hipótesis contenía los datos referidos al producto interior bruto (PIB) de multitud de países, representando su evolución económica, y datos que reflejan las cantidades de vehículos eléctricos en dichas zonas. El intervalo final de tiempo que se estableció para llevar a cabo los análisis comienza en 2017 y termina en 2022.

Procesamiento previo de información

Lo primero que se decidió fue revisar la correlación existente entre las variables (o columnas de la tarjeta de datos) con las que se trabajaría. Los resultados obtenidos denotaban una total relación entre las cifras económicas de PIB de los países, pues casi en todas las combinaciones de pares de ellas el coeficiente calculado era 1 o una cantidad muy cercana. Esto tiene sentido al pensar que rara vez el crecimiento o decrecimiento de una zona geográfica en términos monetarios sufre modificaciones aleatorias o repentinas de forma representativa. Por otro lado, en el caso de las variables enfocadas en ventas de vehículos eléctricos ocurría todo lo contrario (valores casi nulos), por lo que queda clara la ausencia de ningún tipo de tendencia social que afecte a este sector del mercado.

Luego, como aproximación a los pasos posteriores y para seguir recolectando información potencialmente útil sobre el conjunto de datos, se extrajo los componentes principales de ellos. Esta técnica busca la proyección según la cual los datos queden mejor representados en términos de mínimos cuadrados, de tal forma que es viable la reducción de dimensionalidad de los dataset mientras se ordenan las restantes según su importancia. En concreto fueron solamente dos el número de características con las

que se trabajó, pues ellas podrían después ser graficadas de forma cómoda y sencilla. De esta manera se probó que ciertos países sobresalían del conjunto restante como es el caso de China, EEUU o Luxemburgo.

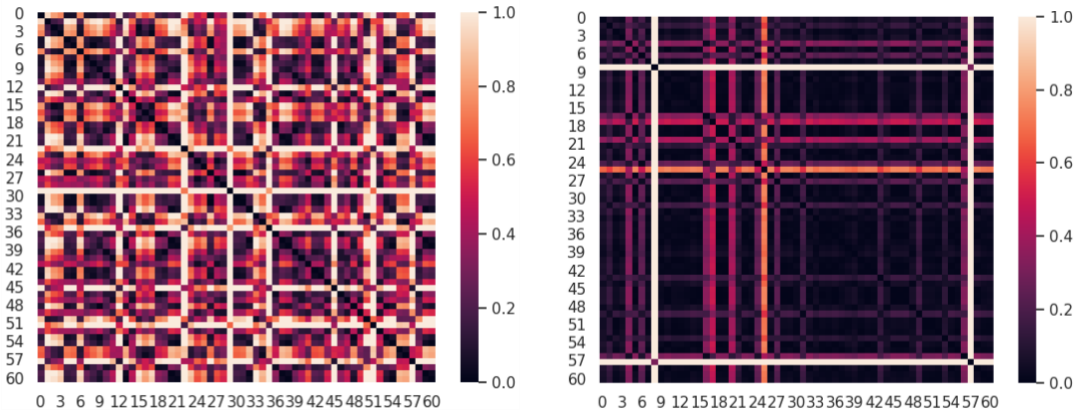


Tras estos dos acercamientos, se tomó la decisión de que el uso de técnicas de clustering podría suponer un buen medio para resolver el problema y arrojar pruebas visuales que confirmaran o desmintieran que “Se venden más vehículos en países con más PIB per cápita”: se pensó en representar gráficamente separados por grupos (o clusters) los países estudiados tanto en el ámbito económico como de venta de coches eléctricos. Así resultaría evidente si el grupo de aquellos con mejor estadísticas monetarias contarían a su vez con las más altas tasas de adquisiciones de vehículos.

Con ese objetivo en mente, se separó la tarjeta de datos en dos diferentes dataframes nuevamente, uno para el producto interior bruto y el otro para ventas de coches. Además, tuvo que llevar a cabo un proceso de normalización sobre ambos para evitar problemas derivados de órdenes de dimensión diferentes: fue elegido un MinMaxScaler, capaz de transformar a una escala con mínimo = 0 y máximo = 1.

Posteriormente a este procesado y tras la documentación sobre modalidades de clustering vistas durante el curso, se pensó que sería de gran ayuda en el futuro contar tanto con los componentes principales de los dos dataframes generados como con la matriz de similitud. Al igual que antes, en el primer caso se extrajo sólo dos componentes y volvió a emplearse un StandardScaler debido a que al transformar los datos de manera que tengan una media de 0 y una desviación estándar de 1 se evita que ciertas características dominen sobre otras simplemente debido a sus unidades o magnitudes. En el segundo caso la construcción de dichas matrices surgió por la necesidad de comprobar cómo de parecidos eran entre sí los valores de las variables usadas (se

utiliza distancia euclídea para ello) para más adelante confeccionar los cluster de manera exacta.



Como información adicional que apoya la teoría propuesta al analizar las correlaciones de las variables, la matriz de similitud de los datos sobre PIB contiene tonos mucho más claros que los oscuros expuestos en la matriz referida a vehículos, lo que denota mayor relación o parecido.

Clustering

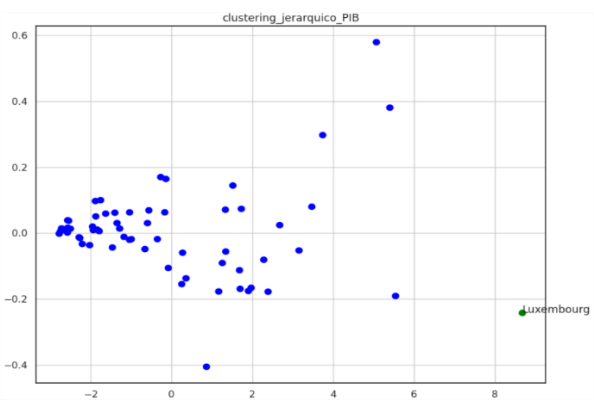
Con el fin de contrastar diferentes enfoques proporcionados por las técnicas de clustering estudiadas durante el curso decidió implementarse tres de ellas. De esa forma se vio la posibilidad de obtener el mismo resultado (mejor opción posible pues quedaría la hipótesis podría aceptarse o rechazarse directamente) o al menos seleccionar el más lógico y cercano a la realidad.

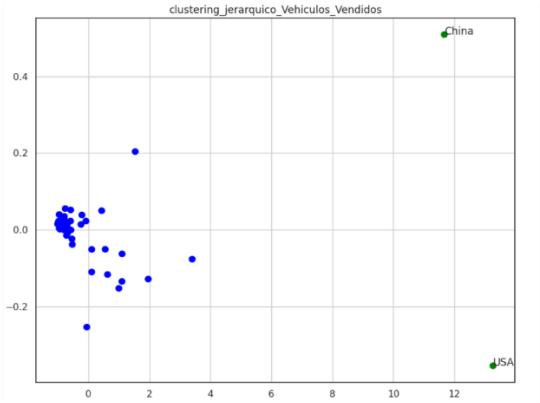
Todos esos enfoques coincidieron en declarar como cantidad óptima de clusters o clases el 2. Esto abrió la puerta a dividir los países estudiados en únicamente dos extremos: aquellos con más o menos poder adquisitivo (para clusters generados a partir del dataset sobre el PIB); y aquellos punteros o no en compras de vehículos eléctricos (para clusters generados a partir del dataset restante). Esto facilitó las futuras interpretaciones, pues desde ese momento bastaba con observar si en las agrupaciones creadas de países los que estaban en un extremo para un dataset lo estaban también en el otro (más ricos junto con más compras de coches y viceversa).

Es digno de mención que se empleó como métrica de evaluación de clusters el coeficiente de Silhouette, que posee un rango de -1 a 1, representando valores cercanos a este último que los objetos (países en este caso) están bien ajustados a su propio cluster y, al mismo tiempo, están separados de otros clusters.

Clustering Jerárquico

Se comenzó por este tipo de clustering. Para el dataset sobre PIB arrojó dos agrupaciones o clusters con un coeficiente de Silhouette de 0’656, mientras que para el de Vehículos Vendidos un valor de 0’934. Ambos resultados fueron bastante prometedores. Las representaciones gráficas generadas (empleando los puntos que nacen de los componentes principales de los datasets) son las siguientes:

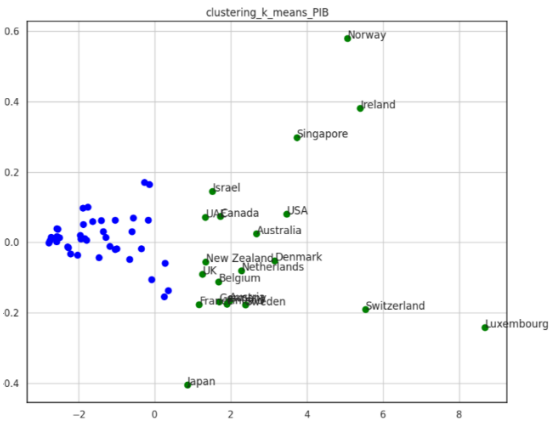


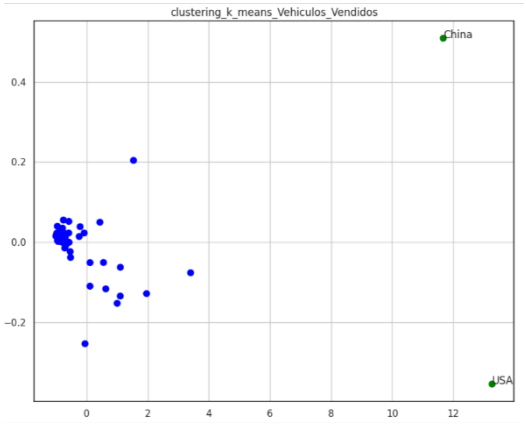


En términos económicos, es curioso cómo se clasifican los países: únicamente se introduce dentro del grupo de los que mejores datos de producto interior bruto a Luxemburgo. Mientras tanto, en términos de adquisiciones de vehículos eléctricos destacan por encima del resto China y EEUU, algo más esperable.

Clustering K-Means

Para el dataset sobre PIB arrojó dos agrupaciones o clusters con un coeficiente de Silhouette de 0’638 (ligeramente menor que antes), mientras que para el de Vehículos Vendidos un valor de 0’934 (exactamente igual que antes). Ambos resultados volvieron a ser bastante prometedores. Las representaciones gráficas generadas (empleando los puntos que nacen de los componentes principales de los datasets) son las siguientes:





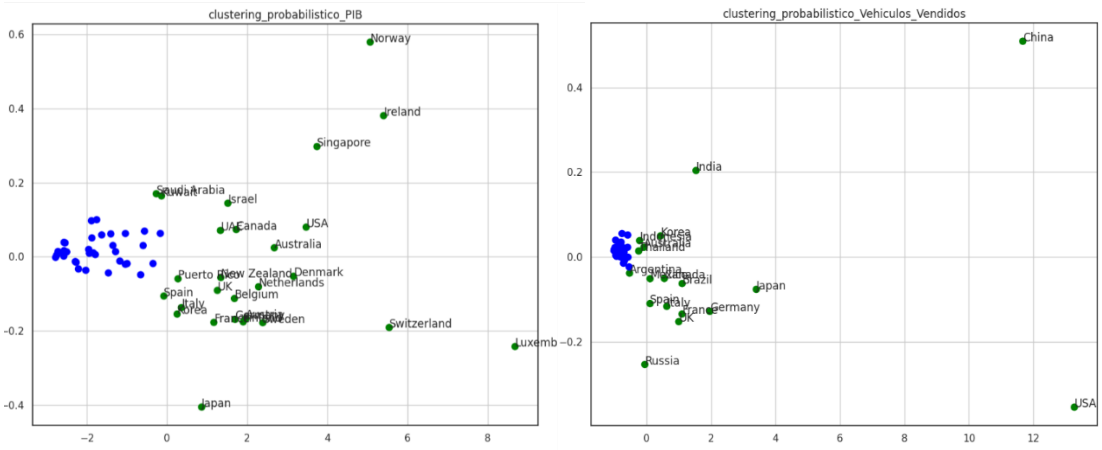
En términos económicos, esta vez se sitúa a un número más equilibrado de países en cada uno de los dos grandes grupos, algo que es más realista que el caso anterior. Mientras tanto, en términos de adquisiciones de vehículos eléctricos destacan por encima del resto China y EEUU, es decir, lo mismo que ocurría antes y lo que significa que esta clasificación es bastante confiable.

Clustering Probabilístico

Para acabar con el análisis se tomó la decisión de emplear un tercer tipo de clustering, a pesar de ya contar con resultados ciertamente determinantes. A la hora de utilizar esta técnica apareció un problema, y es que al basarse en probabilidad (la aleatoriedad se establece desde el momento mismo de creación de la clase que genera los clusters), los resultados que ofrece pueden variar de una ejecución a otra del código. Esto es lo que ocurre de manera frecuente con las clasificaciones llevadas a cabo sobre el dataset de vehículos vendidos (no en el del PIB, que se mantuvo estable, al menos).

En adición, en este caso no puede presentarse como argumento de fiabilidad el coeficiente de Silhouette debido a que ni es utilizado: simplemente se maneja el “bic” (Bayesian Information Criterion). Este criterio compara el ajuste de diferentes modelos, pero penalizando la complejidad del modelo (número de parámetros), de modo que el modelo con menor valor de BIC debe elegirse.

Sin embargo, teniendo todo eso en cuenta estos son las gráficas conseguidas:



Conclusión del Clustering

Posteriormente al estudio de las tres posibilidades presentadas más arriba, se consideró que la más acorde con lo que realmente sucedió es la ofrecida por el Clustering K-Means. No obstante, ni en esa solución ni en las demás hubo una clara relación entre los países clasificados como económicamente sobresalientes y aquellos destacados en ventas de vehículos no contaminantes, por lo que se llegó a la conclusión de que la

hipótesis que venía intentándose contrastar debía rechazarse (a pesar de que inicialmente era lógico pensar que sería cierta).

El conocimiento oculto que se desprende de todo lo comentado sobre la hipótesis es que aquellas potencias mundiales con mejores indicadores monetarios son partidarios de realizar inversiones en otros campos diferentes de la automoción eléctrica y sostenible.

Contraste de hipótesis adicional

Al poco tiempo de acabar el estudio realizado sobre la tarjeta de datos, se pensó en añadir una última comprobación. Echando un vistazo rápido sobre los datos referidos a la cantidad de coches que no usan la combustión a lo largo de los años manejados, surgieron algunas dudas sobre si realmente con el paso del tiempo había aparecido un incremento en las ventas de ellos o más bien todo lo contrario.

En concreto, se lanzó la hipótesis de que las ventas de vehículos eléctricos del año 2019 superaron a las de 2022. A partir de dicho momento, se trabajó para verificar si dicha afirmación era fruto de la mera casualidad (H0) o detrás existían razones de peso que tuvieran como consecuencia ese resultado (H1).

La primera idea que se tuvo fue evitar el uso del muestreo, dado que éste es recomendable para casos en los que se tenga una cantidad tan ingente de datos que suponga un verdadero reto su manejo, algo que para nada ocurría en nuestro dataset (no obstante, se probó de igual forma esa técnica sobre él, confirmando las sospechas de su mejorable funcionamiento).

```
Media de CochesVendidos_2019: 1137015.08
Desviación típica de CochesVendidos_2019: 2687430.56

Media de CochesVendidos_2022: 949283.93
Desviación típica de CochesVendidos_2022: 2372786.12

Intervalo de confianza para CochesVendidos_2019:  [462598.2426135732, 1811431.9213208533]
Intervalo de confianza para CochesVendidos_2022:  [353827.8357554013, 1544740.0330970578]

Ttest_indResult(statistic=0.4089867008441667, pvalue=0.6832896051346736)
```

Finalmente, y con un grado de garantía del 95%, ciertas conclusiones fueron extraídas. A pesar de que la media de coches vendidos en 2019 sea mayor que la de 2022, la de ambos se sitúa en una zona central del intervalo de confianza del otro. Además, dichos intervalos son bastante similares, de manera que casi podría realizarse una perfecta superposición.

El estudio terminó tras la ejecución de la prueba “T de Student” (manejamos dos años diferentes y sus varianzas son distintas) a través del método ttest_ind() proporcionado

por Scipy. El p-value obtenido es muy alto como para considerar lo ocurrido estadísticamente significativo, así que al ser improbable que no haya sucedido por casualidad, se rechaza la hipótesis alternativa formulada inicialmente como H1 y se acepta la nula (H0).

4. Conclusión del proyecto

Finalmente tenemos un proyecto que comenzamos en septiembre para tratar de dar respuesta a una serie de hipótesis sobre una serie de datos que cogimos en ese momento, y es sin duda impresionante ver cómo todo eso ha evolucionado. No solo las hipótesis si no también los datos han evolucionado y cambiado según veíamos preguntas más interesantes o déficits de información en algunos lados. Nos ha resultado también bastante interesante como, a pesar de tener todos los datos que queríamos, cualquier pregunta o hipótesis que te plantees es imposible de responder simplemente con un sí o un no, todas llevan mucho trabajo detrás sobre los datos que se quieren investigar y no es tan fácil como buscar datos que afirmen lo que en un principio dices.

Como resumen a todo el tema en general de los vehículos eléctricos lo podemos resumir rápidamente, es un mercado demasiado actual con pocos datos. Y es que no podemos hablar de la venta masiva de vehículos eléctricos hasta por lo menos el 2016, con lo cual en tan solo 6/7 años que tiene de vida este mercado es tanto difícil como precipitado intentar sacar conclusiones de los datos que se tienen. No obstante, hemos sido capaces de sacar unas conclusiones a las hipótesis iniciales contrastadas de tal manera que podrán servir en el mercado actual como, esperemos, en el mercado a futuro.

Este ha sido uno de los proyectos más grandes que hemos tenido a lo largo de toda la carrera que nos ha llevado todo el cuatrimestre. Sin duda podemos sacar muchas cosas positivas de él

5. Bibliografía

https://github.com/Mohamed11302/MineriaDeDatosYSistemasMultiagentes

Dataset	Enlace
Rendimiento de los vehículos	Kaggle
Precio gasolina	Trading economics
Precio diésel	The global economy
Precio luz	Ember Climbate
Puntos de Carga	IEA
PIB per cápita	IMF
Ventas por modelo de coche	Marklines
Nivel CO2	Our world in data
Nivel de estudios	World population reviews

6. Participación

Alumno	Participación (%)
Diego Cordero Contreras	20
Enrique Albalate Prieto	20
Lucía De Ancos Villa	20
Pablo Del Hoyo Abad	20
Mohamed Essalhi Ahamyan	20