MEMORIA TRABAJO FINAL
MINERÍA DE DATOS

GRUPO 6

Antonio Sevilla
Javier Nicolás
Álvaro Francisco Toledano

ÍNDICE

DESCRIPCIÓN DE LOS DATOS

El conjunto de datos que emplearemos ha sido extraído de una encuesta sobre consumo en la industria automovilística realizada por parte de la empresa Smartme Analytics. Esta empresa de datos española recoge datos de consumo digital de sus usuarios de forma , recompensándolos con dinero y premios.

La encuesta se realizó a través de la aplicación de la empresa a todos los usuarios interesados en contestar y se completó con información sociodemográfica que los usuarios aportan al registrarse y con información de su consumo digital que se obtiene a través de diferentes procesos que los usuarios permiten.

Dicha encuesta se llevó a cabo con una finalidad comercial. Con ella se puede obtener información acerca de perfiles de consumidores, preferencias y necesidades; y gracias a ello ganar especificidad y efectividad en la oferta de ciertos servicios.

Hemos decidido utilizar estadísticas actuales para poder extraer conclusiones útiles y realistas en un ámbito comercial. Debido a que el dataset original contenía datos privados y confidenciales hemos realizado una limpieza previa para anonimizarlo, la cual no se documenta en la presente memoria.

Nuestro dataset cuenta con 7403 muestras, es decir, usuarios que respondieron la encuesta, con un porcentaje de NAs de 40% repartido en todo el dataset, por tanto se berá realizar un labor de limpieza previa. Cada pregunta de la encuesta se corresponde con una de las 121 variables. Estas se describen a continuación.

- Un ID de usuario para facilitar el estudio de los casos particulares y 2 preguntas binarias que suponen una criba a la hora de seguir realizando la encuesta. Si el participante posee permiso de conducir y si es conductor habitual.

- Una docena de preguntas de carácter personal
  - con respuestas en diferentes grupos (conformación del hogar, formación, CCAA…)
  - numéricas agrupadas por grupos (las edades en generaciones, los ingresos en horquillas)
  - y de tipo binario (sexo).

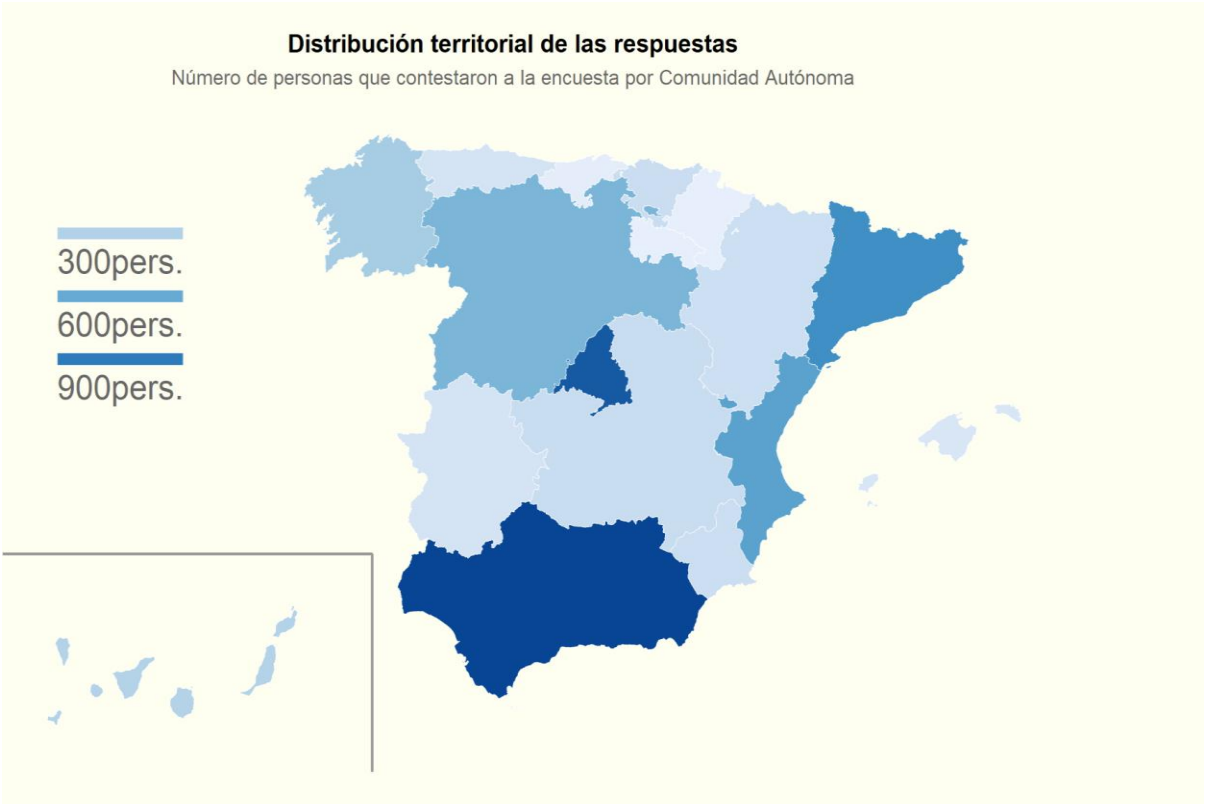
- Una docena de preguntas con respecto al uso de combustible
  - con respuestas en diferentes grupos (tipo de combustible que usa tu coche),
  - numéricas agrupadas por grupos (su conformidad acerca de preguntas relacionadas con el ecologismo)
  - y de tipo binario (si considera poseer un coche que consuma cierto tipo de combustible).

- Otra docena de preguntas que definen al tipo de consumidor con respuestas en distintos grupos (preferencia de financiación, tipos de coche que plantea adquirir según antigüedad…)
- Unas 30 preguntas binarias acerca de si el participante considera o no consumir cierta marca.
- Media docena de preguntas binarias sobre el uso o no de ciertas apps por parte del participante.
- Otras 30 con respuesta contínua de tiempo de visualización/escucha de ciertas cadenas de televisión y radio en minutos durante 1 año.

5240 de los participantes tienen carné y de estos solo 4085 son conductores habituales.

Un análisis inicial de los datos nos muestra una importante mayoría de participación de mujeres de cerca del 70%.

Además tenemos una distribución asimétrica de participantes a nivel territorial, debido a la diferencia de densidad poblacional y el alcance de la empresa entre Comunidades Autónomas.



LIMPIEZA

- Primeramente asignamos nombres cortos, descriptivos y manejables a las variables, vectorizando la sustitución de los mismos. Cambiamos las respuestas posibles por una sola palabra también.

- Así mismo, se eliminan palabras clave mediante búsqueda automática y se reformulan las preguntas de forma más sencilla mediante regular-expresions.

- Sustituimos las strings “N/A” por objetos tipo NA que reconoce R.

- Simplificamos las respuestas categóricas posibles para ciertas variables, cuando obtenemos diferencias en la precisión menores al 0.01%. Por ejemplo, los coches en la categoría “Seminuevo” pasarán a “Nuevo” y aquellas respuestas a la pregunta de intención de compra futura que figuraban como “Próximamente” ahora lo harán como “Sí”.

- Por último creamos nuevas variables que nos resultarán de utilidad como predictoras:

    - TV_TOTAL y Radio_TOTAL, variable contínua suma de los tiempos de visualización de todos los canales por parte del usuario.

    - Gama_Coche, variable categórica que determina el grupo al que pertenece un usuario según una agrupación realizada previamente de las marcas.

    - Income2, variable que representa el Income del usuario. Para crearla hemos transformado Income (variable categórica que aludía a un intervalo) en contínua. A fin de conseguir el máximo realismo posible utilizamos el siguiente algoritmo:

        - A los usuarios que pertenecen a la horquilla “hasta 1000€” se les asocia el valor pseudoaleatorio de una exponencial de parámetro 12.

        - A los usuarios en horquillas intermedias se les asocia el valor pseudoaleatorio de una triangular.

        - A los usuarios con un Income de “más de 5000€” se les asocia los valores de una exponencial con mayor acumulación cerca de los 5000€.

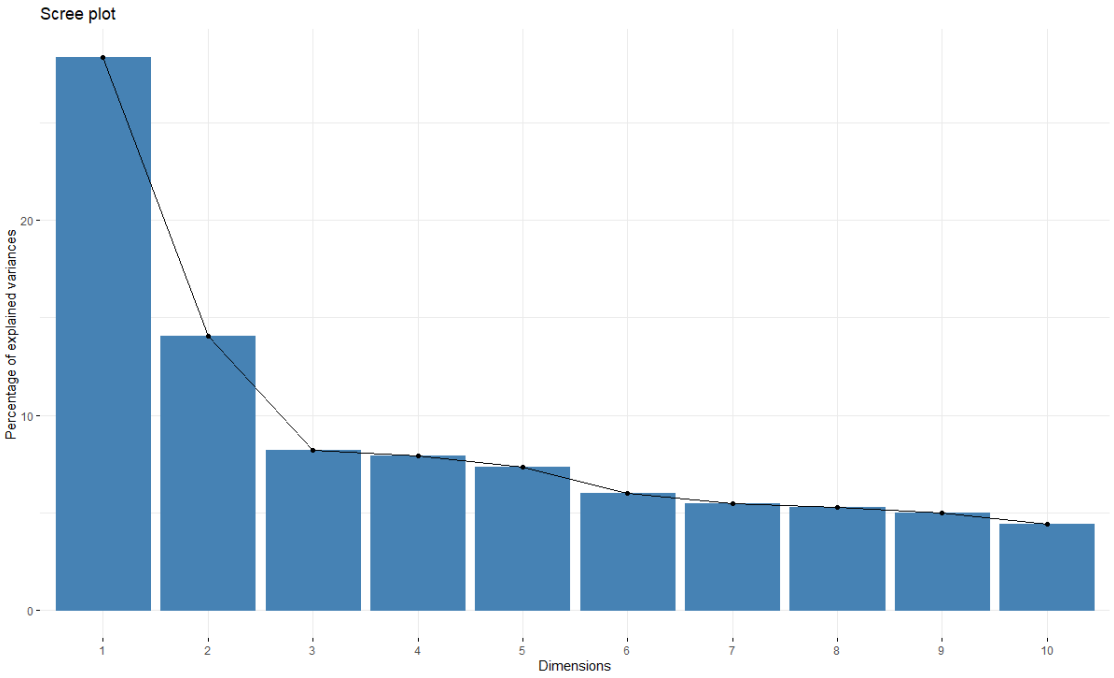
APRENDIZAJE NO SUPERVISADO

Seleccionamos las columnas a utilizar, tratamos los missings.

Para poder realizar una primera reducción de dimensiones comprobamos la independencia de los datos así como el K.M.O. Como existe independencia y el K.M.O. es de entorno al 84%, es recomendable el uso del PCA.

Todas las variables están en una escala numérica de 1 a 10 así que no será necesario un escalado.

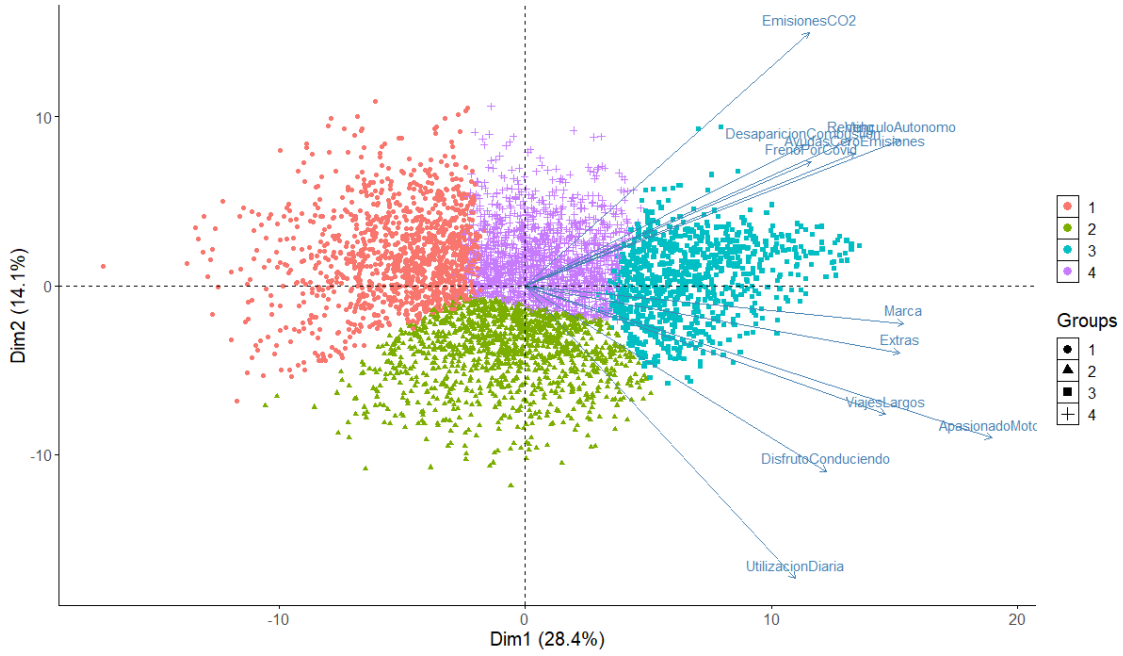
Mediante un gráfico de sedimentación observamos las componentes más relevantes:



Seleccionamos estas 3:

- Dim.1: ApasionadoMotor, Extras, VehiculoAutonomo, Marca
- Dim.2: EmisionesCO2, DisfrutoConduciendo, UtilizacionDiaria
- Dim.3: FrenoCovid (parte de Utilizaci?nDiaria y ApasionadoMotor)

Para encontrar el número óptimo de clusters a emplear usaremos nbclust. Los resultados arrojados sugieren el uso de 3 ó 4. Usaremos 4 sobre las 3 dimensiones creadas anteriormente en el PCA.



La primera dimensión, se puede observar que las variables DesaparicionCombustion, AyudasCeroEmisiones, Renting, VehiculoAutonomo y FrenoCovid se posicionan en la misma dirección de proyección por lo que se consideran con alta correlación entre ellas. En la parte del cuadrante inferior, tenemos Marca y Extras saturadas en la dimensión 1 y ViajesLargos, ApasionadoMotor que comparten algo más con el eje 2.

La segunda dimensión enfrenta las Emisiones de CO2 (parte positiva del eje) con la Utilización diaria y el Disfrute de conducir.

- En este sentido, el cluster 4 se compone de usuarios más concienciados con el medio ambiente.

- En contraposición, el cluster 2 se compone de aquellos individuos que utilizan el coche a diario y disfrutan conduciendo con poca preocupación por las emisiones.

- El cluster 3 mezcla a la gente concienciada con el coche eléctrico y apasionada del motor, que valora la marca y los extras en un vehículo.

- Por el contrario, el cluster 1 contiene a un perfil con un menor acuerdo con las afirmaciones en relación al coche eléctrico y a la importancia de la marca.

Finalmente guardaremos los datos para utilizar las categorías de cluster como variable explicativa.

MÉTODO DE PAGO, INTENCIÓN DE COMPRA Y TIPO DE COMBUSTIBLE

La primera pregunta que intentamos responder tiene una gran importancia a nivel empresarial: ¿Qué método de pago prefieren los consumidores?

Eliminamos las observaciones incompletas de la variable a predecir. Agrupamos los métodos de pago posibles en “Financiación” y “Al Contado”, que tenían una presencia similar, obviando otras opciones como “Renting” con una muestra mucho menor. Convertimos a clase factor todas las variables de tipo character. Creamos las particiones de los datos para el entrenamiento, el test y el cross-validation.

Como variables explicativas utilizamos Metodo_Pago, Sex, Age, Income, Composicion_Hogar, Propiedad_Coche, Marca_Coche, Tipo_Coche, Canal_Compra, Tiempo_Coche, Estado_Coche, Combustible_Coche, Intencion_Compra, Habitat, TV_TOTAL, Radio_TOTAL, Province, además de las variables de uso de aplicaciones y las categorías del Clustering del principio.

Utilizaremos un árbol de decisión, estudiando primero iterativamente a qué nivel de precisión ofrece cada C.P. Podemos observar que cuando nos aproximamos al 0 obtenemos mejor precisión a la vez que mayor riesgo de sobreajuste. Así mismo a partir de cierto valor de C.P. la precisión converge. Como nos interesa un árbol lo más sencillo posible, aceptamos dicho valor. El primer esquema obtenido es demasiado grande y poco visual, por lo que reduciremos el C.P. a costa de perder precisión.

Estudiamos iterativamente la repercusión en la precisión sobre el conjunto de test de cada nivel de C.P. Justo al contrario que en el set de entrenamiento, cuanto mayor es el C.P. se obtiene mayor precisión. Eso se debe a que el modelo es más general, se está evitando el sobreajuste; ergo aceptamos C.P. = 0.008 como valor para el parámetro. Metemos las predicciones en el conjunto de testeo.

El árbol obtenido al final solo tiene en cuenta 4 variables: Canal_Compra, Estado_Coche, Marca_Coche e Income.

Del resto de variables, Income destaca por encima del resto de aquellas que no aparecen en el árbol, seguida de Age, Tipo_Coche, Tiempo_Coche, Intencion_Compra. Obtenemos una precisión del 71.5%.

Posteriormente aplicamos el método KNN. Aplicamos el mismo proceso de tratamiento de las variables que para el árbol, obviando la C.V. para conseguir un modelo más específico. Estudiamos iterativamente el valor de k que arroja una mejor precisión. Observamos que el porcentaje de error converge desde los 50 vecinos. Minimizamos el error para k desde 50 hasta 120 en el conjunto de entrenamiento. Finalmente probamos en el data test de testeo con el k que daba error mínimo, es decir k = 60; obteniendo una precisión de entorno al 70%, lo cual es bastante aceptable.

Además aplicamos un random forest, probando iterativamente distintas combinaciones de variables y árboles para maximizar la precisión. En este caso decidimos utilizar 5 variables y 80 árboles.

RESULTADOS MÉTODO DE PAGO

	ÁRBOL DE DECISIÓN	KNN	RANDOM FOREST
PRECISIÓN	**0.7146**	**0.7047**	**0.6961**
SENSIBILIDAD	**0.5765**	**0.6261**	**0.5450**
ESPECIFICIDAD	**0.8371**	**0.7769**	**0.8347**

Obtenemos buenas precisiones en los tres casos, por tanto debemos fijarnos en cómo clasifican. En este caso, en Random Forest se detecta en mayor proporción la FINANCIACIÓN por tanto no estará tan balanceado. Tenemos el mismo caso para el árbol que obtenemos una mayor especificidad. Por tanto, decidimos quedarnos con el modelo KNN ya que obtenemos una precisión de alrededor del 70% y con una especificidad y sensibilidad muy balanceada. Además, en este caso la precisión es la mejor medida que podemos utilizar por el contexto de los datos y las preguntas formuladas.

Estudiamos mediante el mismo sistema de trabajo la intención de compra. Para el árbol se utilizan las variables PCA_CLUST_Opino, Tiempo_Coche, Age, Province, Marca_Coche y se tiene un éxito de entorno al 0.6. Empleando un KNN con las mismas variables del árbol con k = 80, se obtiene una precisión de 0.6 también. Mediante una regresión logística también con las mismas variables se obtiene una precisión similar. Siendo este el modelo que peores resultados nos arroja, obtenemos en ambos casos resultados muy similares. Por tanto, decidimos elegir el modelo de árbol de clasificación ya que es el modelo más simple y visual y podemos entender bien cómo se clasifican los datos.

RESULTADOS INTENCIÓN DE COMPRA

	ÁRBOL DE DECISIÓN	KNN	REGRESIÓN LOGÍSTICA
PRECISIÓN	**0.6036**	**0.598**	**0.6255**
SENSIBILIDAD	**0.5921**	**0.6745**	**0.6857**
ESPECIFICIDAD	**0.6157**	**0.5167**	**0.5616**

Repetimos nuevamente el procedimiento para predecir el tipo de combustible. Para el árbol se utilizan las variables Marca_Coche, Tipo_Coche, Estado_Coche, Tiempo_Coche y Province. Se tiene un éxito de entorno al 0.67. Mediante un random-forest estudiamos el mejor número de variables a emplear, resulta ser 7. Además con 130 árboles obtenemos una

precisión. Esta será finalmente del 0.66. Empleando un KNN con las mismas variables del árbol con k = 150, se obtiene una precisión de 0.64.Tras analizar los resultados, rechazamos el modelo KNN ya que, a parte de tener peor precisión que el resto su balance sensibilidad-especificidad es peor que la del resto. Por este mismo motivo tanto el árbol como el random forest nos arrojan resultados muy similares, pudiendo elegir los dos, vemos un poco más balanceado el RandomForest.

RESULTADOS TIPO DE COMBUSTIBLE

	ÁRBOL DE DECISIÓN	KNN	RANDOM FOREST
PRECISIÓN	**0.6667**	**0.6441**	**0.6643**
SENSIBILIDAD	**0.7842**	**0.8901**	**0.7386**
ESPECIFICIDAD	**0.5084**	**0.3065**	**0.5642**

REGRESIÓN INCOME

Como última pregunta, vamos a intentar utilizar árboles de regresión para predecir los ingresos (como variable contínua) a partir de los datos de los clientes, tanto personales como de uso del vehículo. Esta información puede resultar de suma utilidad a la hora de conceder créditos y en muchos ámbitos financieros y empresariales.

Una vez entrenado el modelo, vemos que quedan bastantes nodos terminales y que se utilizan muchas de las variables. La primera ramificación se produce según la actividad laboral de los encuestados. Después, el nivel de estudios, y a continuación la marca de coche.

A la hora de podar el árbol decidimos usar 6 nodos (en vez de 4 como usaríamos normalmente) a fin de reducir el error todo lo posible. Para determinar el M.S.E. en la predicción utilizamos el conjunto de test. El error es del orden de 330000.

Intentamos mejorar el modelo usando Boosting. Probamos diferentes niveles de profundidad en la interacción y diversas cantidades de árboles para encontrar la combinación que minimiza el error. Observamos que los mejores resultados se producen cuando interaction.depth = 1. En ese caso, el número de árboles no parece tener mucho impacto. En el mejor caso que hemos obtenido el error cuadrático medio obtenido es del orden de 290000, aproximadamente un 12% mejor que con el árbol inicial.

Con una diferencia abismal, la variable con mayor influencia es la marca de coche. La segunda, lejos de la primera pero con mucho mas impacto que el resto, es el sexo. Así pues,

tiene sentido inferir el sueldo de una persona según la marca de su coche. Por otro lado, se puede apreciar que desgraciadamente sigue habiendo una brecha salarial por sexos.

Reescalando los datos obtenemos un resultado consistente; además el error 3 se ubica en torno a 0.008, es decir, menos de un 1% en términos de percentil.

La gran mayoría de variables en nuestro dataset son categóricas, lo que dificulta la aplicación de otros modelos de regresión, por ello aunque las conclusiones sean razonables el error es grande.