Universidad del Desarrollo
Facultad de Ingeniería

CLUSTERIZACIÓN EN BASE A COMPRAS DE VEHÍCULOS MOTORIZADOS

Periodo 2009 al 2021

POR: NICOLÁS PASTOR BUENO Y MARIA ALEJANDRA HERMOSILLA

Capstone project presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESORES GUÍA:

Sr. Alonso Astroza Tagle

Sr. Cristian Candia

Sr. Takeshi Asahi Kodama

Sr. Víctor Landaeta

DICIEMBRE 2022

SANTIAGO

AGRADECIMIENTOS

Agradecemos a los profesores guía de este programa, por el apoyo y dedicación no solo a este proyecto, sino que en todo el periodo que comprendió este magíster. También a nuestras respectivas familias, parejas y amigos que nos animaron a perseguir este objetivo.

TABLA DE CONTENIDO

Resumen

El presente documento, se refiere al estudio de ventas de una gran concesionaria de venta de vehículos, la cual nos proporciona datos crudos para el estudio del comportamiento de compra de vehículos nuevos desde el año 2009 al 2021, generando agrupamientos de estas ventas a través de la generación de clústers, en base a modelos de aprendizaje no supervisado y por otra parte, la predicción de estos clústers en base a modelos de aprendizaje supervisado.

Para realizar este estudio comenzaremos con un análisis de la data disponible y un análisis exploratorio de los datos para poder comprender las tendencias de venta que posee esta concesionaria.

Luego implementaremos algoritmos para la generación de agrupaciones dentro de los registros, para de esta forma, determinar la/s variable/s que contribuye/n mayormente a en la generación de estos grupos. Con el fin de validar el comportamiento de la agrupación realizada a través de modelos de aprendizaje no supervisado para generación de clusters, se implementan modelos de aprendizaje supervisado para así predecir y validar los clusters agrupados en base a modelos de clasificación.

Finalmente presentaremos los resultados de la metodología aplicada y con ello, las principales conclusiones de este estudio.

1. Introducción

El automóvil es un bien muy preciado para los chilenos y ha pasado de ser un lujo a una necesidad. Su uso abarca desde permitir traslados de la población hacia lugares de trabajo y/o lugares de recreación, hasta la generación de puestos de trabajo para diversas industrias.

Todo vehículo en Chile está afecto a un impuesto para circular por las vías del territorio nacional, llamado "permiso de circulación". El valor del permiso de circulación es calculado por el Servicio de impuestos internos de acuerdo con las distintas características del vehículo. Este tributo es pagado de forma anual y puede ser dividido en dos pagos, pudiendo ser el segundo de estos, afectado por el IPC. El ente emisor de autorización de uso de las vías del territorio nacional para un automóvil, son algunas de las 346 municipalidades de Chile, las cuales son beneficiadas por el pago ya que este, va en directo beneficio a sus arcas.

Nuestra motivación es poder analizar la información de compra de vehículos entre el periodo 2009 y 2021 en base a los automóviles que se hayan comercializado a personas naturales en la concesionaria de venta de vehículos nuevos, y verificar el tipo de vehículos con que se clasifican los automóviles en base a la emisión del permiso de circulación.

Es de conocimiento público el cambio de orientación que han tenido los chilenos en lo que a preferencias de compra de automóviles se refiere, por lo que nuestro interés reside en analizar las preferencia de compra entre los años 2009 y 2021 y cómo estas pueden influenciar la decisión de los clientes al momento de seleccionar su siguiente automóvil, teniendo en cuenta los hechos que han marcado en los últimos años a nuestro país y al mundo, como por ejemplo cambio gobierno Michelle Bachelet a Sebastián Piñera (2017-2018), estallido social (2019-2020), pandemia Covid-19 (2020-actualidad), crisis de abastecimiento de chips (2020-actualidad).

Con nuestra propuesta de estudio, proponemos clasificar a los clientes y sus preferencias de compra, para obtener conclusiones respecto a patrones de comportamiento en la compra de los vehículos comercializados por una gran concesionaria de automóviles, con el fin de permitir la toma de decisiones respecto de orientación de campañas de marketing a clientes en específico, respecto de marcas, modelos y especificaciones técnicas que sean correlacionados a su realidad.

2. Trabajo Relacionado

Para la realización de este estudio nos apoyaremos de entidades oficiales que generen estadísticas y métricas obtenidas desde la Asociación Nacional Automotriz de Chile (ANAC). También nos basaremos en estudios que aborden la temática de segmentación de clientes utilizando modelos de aprendizaje no supervisado y ejemplos que utilicen los mismos algoritmos que abordaremos en este estudio:

- Customer Segmentation and Profiling for Life Insurance using K-Modes Clustering and Decision Tree Classifier.
- Explicabilidad de modelos de Machine Learning con SHAP.
- Clustering Algorithm for data with mixed Categorical and Numerical features.
- Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values.
- Measuring customer similarity and identifying cross-selling products.
- K-prototypes - customer clustering with mixed data types.

3. Hipótesis y Objetivos

Una parte fundamental del análisis del mercado automotriz se basa en la segmentación de clientes. La segmentación facilita la identificación de nuevas oportunidades de abarcar de mejor manera un nuevo mercado y potenciar aquellos segmentos que se abordan actualmente.

Debido al rol que cumplen la segmentación en los estudios de mercado, estos se realizan ampliamente por diferentes entidades. En nuestro país contamos con la entidad ANAC (Asociación Nacional Automotriz de Chile) la cual emite mensualmente un informe del mercado automotor, donde se muestran los resultados de ventas según marca, región, modelos, tipos de vehículos (Vehículo de pasajeros, SUV, Camioneta y Vehículos comerciales). En al figura 1, se puede observar la venta de vehículos por su clasificación en la comparativa Octubre 2021-2022 y acumulado año 2021-2022

Figura 1

Ventas acumuladas al público por segmento del mercado de livianos y medianos.

Segmentos	Octubre			Acumulado Año		
	2021	2022	Var% Mes	2021	2022	Var% Acum
Vehículo de Pasajeros	10.869	5.865	**-46,0%**	98.246	90.923	**-7,5%**
SUV	15.652	12.917	**-17,5%**	142.046	160.127	**12,7%**
Camioneta	7.841	6.966	**-11,2%**	59.159	72.347	**22,3%**
Vehículo Comercial	4.189	2.894	**-30,9%**	36.373	37.740	**3,8%**
Total	38.551	28.642	**-25,7**	335.824	361.137	**7,5%**

Nota. Estudio De Mercado – Anac. (s/f). Anac.cl. Recuperado el 9 de diciembre de 2022, de https://www.anac.cl/category/estudio-de-mercado/

Con este tipo de análisis, podemos tener una visión general del comportamiento del mercado chileno, sin embargo, carece de la prolijidad requerida para enriquecer las estrategias de negocio de una gran comercializadora de vehículos nuevos en nuestro país. Además, existe la posibilidad de que las ventas de esta concesionaria de vehículos, no se ajusten al estándar nacional.

Una forma de caracterizar los clientes que frecuentan esta automotora, es a través de la agrupación de clientes en base a los valores que toman variables tales como: Tipo de vehículo, marca de vehículo, país de origen de la marca, entre otros. Con estas

agrupaciones podemos determinar que variable es la que aporta mayor información para la generación de agrupaciones y que tan efectivos son los clústers creados.

Por este motivo hemos decidido postular la siguiente hipótesis:

> *"Que tan útiles son las clusterizaciones realizadas con modelos de aprendizaje no supervisado de clasificación en base a compras de clientes de una automotora"*

La información de las bases de datos de la automotora contiene una serie de variables que caracterizan cada registro del vehículo que se ha vendido en el periodo de estudio (ej.: marca, año de venta, origen marca, entre otros). Nuestro objetivo consiste en agrupar esta información para construir clústers y luego evaluar qué tanto contribuye cada variable en la generación de estos.

Con la implementación de estas agrupaciones, podremos alimentar modelos predictivos para pronosticar cuál será el comportamiento de los clientes a futuro.

Con la realización de este estudio, nos proponemos responder las siguientes preguntas:

- ¿Es posible implementar modelos de clasificación con los registros disponibles? ¿Cuáles son estos tipos?
- ¿Cual método de clasificación es superior en performance?
- ¿Cuál de las variables aporta la mayor cantidad de información para la generación de los clústers?
- ¿Es posible implementar modelos predictivos con la data disponible? ¿Qué tipo de modelos?
- ¿La información que aportan los clusters, es significativa para los modelos de predicción?
- ¿Cuál modelo predictivo es superior?

4. Datos y Metodología

En esta sección mostraremos y explicaremos la información que utilizaremos para realizar nuestro estudio, así como los pasos y el procedimiento que seguiremos para determinar si nuestra hipótesis es correcta.

4.1. Datos

La información que utilizaremos se encuentra dispuesta en el servidor centella perteneciente a la concesionaria de vehículos. En esta base de datos se encuentra el dataset **tabla_journey_javier**, el cual alberga 13.641.944 datos de ventas realizadas por la concesionaria desde mayo del año 2009 a diciembre de 2021, tanto a personas naturales como jurídicas; tipos de vehículos, entre otros. Estos datos corresponden a ventas de vehículos nuevos, créditos, ventas de atenciones de servicio técnico como mantenciones, venta de repuestos, etc. Esta tabla contiene las siguientes columnas:

Tabla 1

Listado de columnas de tabla_journey_javier

Nombre Columna	Descripción
Fecha transacción	Fecha de realización de la transacción de venta.
Persona	Indica si el comprador es persona natural o jurídica.
Negocio	Indica si corresponde a vehículo nuevos.
Marca	Marca del vehículo vendido.
Detalle	Procedencia de la venta.
Retail	Indicador interno del negocio de la automotora.
Precio Vehículo	Precio de venta del vehículo.
Margen retail	Margen utilidad sobre precio retail.
Margen distribuidor	Margen utilidad sobre precio distribuidor.
Margen combinado	Margen retail más margen distribuidor.
Año venta	Año en el que se realizó la venta.
Comuna	Comuna de cliente
Dirección	Dirección cliente

Fecha modificación	Fecha modificación registro
Rut	Rut encriptado

En lo que respecta a los datos correspondientes a permisos de circulacion, utilizaremos el dataset **patentes_historico_02**, la cual contiene las siguientes columnas:

Tabla 2

Listado de columnas de patentes_historico_02

Nombre Columna	Descripción
Modelo	Modelo del vehículo
Marca	Marca del vehículo.
Patente	Identificador único de patente
Rut	Rut dueño del vehículo.
Año patente	Año en el que se adquirió la patente.
Tipo	Tipo del vehículo.
Año fabricación	Año de fabricación del vehículo

4.2. Metodología

Con la información dispuesta por concesionaria, hemos logrado determinar el procedimiento general bajo el cual trabajaremos, el cual se compone de las siguientes etapas:

1. Limpieza y filtro de datos
2. Pre-Procesamiento y Enriquecimiento del Dataset
3. Análisis exploratorio de datos y resultados preliminares
4. Implementación modelos de aprendizaje no supervisado de clasificación
5. Generación de modelos de aprendizaje supervisado para predecir y validar clusters generados en base al punto anterior

4.2.1. Limpieza y filtro de datos

Como se mencionó anteriormente, la fuente de datos de ventas proviene de la tabla: **tabla_journey_javier.** Para obtener datos contextualizados que cumplan con las necesidades de nuestra investigación, se aplican los siguientes filtros al dataset:

- Venta de autos nuevos (campo negocio: nuevos)
- Se consideran solo ventas a personas naturales (campo persona: natural)

Figura 2
Filtros aplicados a **tabla_journey_javier**

```
tabla_journey = tabla_journey[
    (tabla_journey["negocio"] == "nuevos") &
    (tabla_journey["year_Venta"] != 2022)&
    (tabla_journey["persona"] != "juridica")
]
```

Nota. Se filtran las columnas "negocio", "year_Venta" (año de venta) y "persona". Para el año de venta se excluyen los registros del año 2022.

De este filtro inicial logramos rescatar las principales marcas comercializadas por la concesionaria en el periodo de estudio:

- Mazda
- Suzuki
- Renault
- Renault-Samsung
- Great Wall
- Changan
- Jac Cars
- Geely
- Haval

Por otro lado, se puede observar que dentro de los datos de las marcas de vehículos, existen 1.529 modelos distintos. La diferencia entre algunos modelos que pertenecen a la misma marca corresponde a una distinción de las características entre autos de una misma línea de modelos, sin embargo, estas diferencias por lo general no son significativas. Por este motivo, se decidió normalizar los modelos de vehículos de cada

marca, con el objetivo de rescatar solo las características relevantes para cada modelo de vehículo. Esto también nos permitirá trabajar con un número más acotado de registros

Figura 3

Normalización de modelos

```
#Suzuki
replace_dict["swift dzire"] = "dzire"
replace_dict["apv2"] = "apv"
replace_dict["ap41v2"] = "apv"
replace_dict["apvx"] = "apv"
replace_dict["ciazx"] = "ciaz"
```

```
#Mazda
replace_dict["cx7-"] = "cx-7"
replace_dict["cx7-r"] = "cx-7"
replace_dict["cx-7r"] = "cx-7"
replace_dict["cx5"] = "cx-5"
replace_dict["rx8"] = "rx-8"
```

```
#Jac cars
replace_dict["a 137 be"] = "a137"
replace_dict["a 137 e"] = "a137"
replace_dict["a 137 se"] = "a137"
replace_dict["137 se"] = "a137"
replace_dict["a 137"] = "a137"
replace_dict["a137 1 3"] = "a137"
```

```
#Great Wall
replace_dict["c30 se"] = "c30"
replace_dict["c30 sr"] = "c30"
replace_dict["c30+"] = "c30"
replace_dict["florid"] = "florid"
replace_dict["florid gross"] = "florid"
replace_dict["florid hcach"] = "florid"
```

```
#Haval
replace_dict["h2 t"] = "h2"
replace_dict["h6"] = "h6"
replace_dict["h6 3gen t"] = "h6"
replace_dict["h6 ca"] = "h6"
replace_dict["h7 t"] = "h7"
replace_dict["jolion t"] = "jolion"
```

```
#Geely
replace_dict["ck3"] = "ck"
replace_dict["ck gb"] = "ck"
replace_dict["lc cross"] = "lc"
replace_dict["lc cross gb"] = "lc"
replace_dict["lc ga"] = "lc"
```

Nota. Limpieza realizada para los modelos de las marcas: Suzuki, Mazda, Jac cars, Great Wall, Haval y Geely.

Esta normalización se realizó para todos los modelos de las 9 marcas comercializadas, tal como muestra la figura 3, con los filtros realizados para modelos de vehículos duplicados y mal normalizados.

Al finalizar la normalización, la cantidad total de modelos disminuyó a 148 modelos, con una cantidad de modelos sin poder limpiar de 7 modelos, lo que nos deja con un porcentaje de limpieza logrado: 90.32%.

4.2.2. Pre-Procesamiento y Enriquecimiento del Dataset

Para enriquecer la información de venta de la concesionaria, agregamos la siguiente información:

- País de origen de las marcas.
- Región de origen de las marcas.
- Tipificación de vehículo (obtenida a partir del dataset de emisión de permisos de circulación)

Para esto, primero obtenemos todos los vehículos con sus permisos de circulación en el periodo de estudio y que pertenezcan a las 9 marcas comercializada por la automotora, para luego agregar la información de origen de cada marca:

Figura 4

Enriquecimiento del dataset.

```
d_country_origin_brand = {
    "jac cars": "china",
    "haval": "china",
    "geely": "china",
    "great wall": "china",
    "changan": "china",
    "mazda": "japon",
    "renault": "francia",
    "samsung": "korea",
    "suzuki": "japon"
}
```

```
d_origin_brand = {
    "jac cars": "asia",
    "haval": "asia",
    "geely": "asia",
    "great wall": "asia",
    "changan": "asia",
    "mazda": "asia",
    "renault": "europa",
    "samsung": "asia",
    "suzuki": "asia"
}
```

Nota.: Clasificación de país de origen y región de origen de marca.

Para agregar la tipificación de vehículos, debemos realizar un cruce entre la base de datos de ventas ya normalizada, con el dataset **patentes_historico_02**. Esto lo realizaremos cruzando la data en base al campo de modelo de vehículo, lo que implica que debemos normalizar los modelos en la tabla de patentes para que coincidan con los de la tabla de ventas, tal como muestra la figura 4, con la clasificación realizada a nivel de código.

Una vez realizada la normalización de todos los modelos, procedemos a incluir los datos en nuestro dataset de ventas.

Figura 5

Tipificación de vehículos

```
#Solo para vehiculos Changan
replace_brand_model(df_patentes,
                    (df_patentes["marca"] == "changan") &
                    (df_patentes["modelo_clean"] == "nd201")
                    ,"md201")

replace_brand_model(df_patentes,
                    (df_patentes["marca"] == "changan") &
                    (df_patentes["modelo_clean"] == "csi")
                    ,"cs1")

replace_brand_model(df_patentes,
                    (df_patentes["marca"] == "changan") &
                    (df_patentes["modelo_clean"] == "cs5")
                    ,"cs55")

replace_brand_model(df_patentes
                    , (
                            (df_patentes["marca"] == "changan") & (df_patentes["modelo_clean"] == "a500") |
                            (df_patentes["marca"] == "changan") & (df_patentes["modelo_clean"] == "m201") |
                            (df_patentes["marca"] == "changan") & (df_patentes["modelo_clean"] == "cm5") |
                            (df_patentes["marca"] == "changan") & (df_patentes["modelo_clean"] == "s300")
                    )
                    , "furgon"
                    , column="tipo")
```

Nota. Ejemplo de normalización y tipificación de vehículos para la marca Changan.

Finalmente, con todas estas modificaciones, se logró culminar en un dataset compuesto por 534.577 registros y 23 columnas, tal como muestra la figura 5, con filtro de modelos de vehículos no normalizados:

Tabla 3

Dataset limpio y enriquecido

Nombre Columna	Descripción	Origen
Fecha transacción	Fecha de realización de la transacción de venta.	**tabla_journey_javier**
Persona	Indica si el comprador es persona natural o jurídica.	**tabla_journey_javier**
Negocio	Indica si corresponde a vehículo nuevos.	**tabla_journey_javier**
Marca	Marca del vehículo vendido.	**tabla_journey_javier**
Detalle	Procedencia de la venta.	**tabla_journey_javier**
Retail	Indicador interno del negocio de la automotora.	**tabla_journey_javier**
Precio Vehículo	Precio de venta del vehículo.	**tabla_journey_javier**
Margen retail	Margen utilidad sobre precio retail.	**tabla_journey_javier**

Margen distribuidor	Margen utilidad sobre precio distribuidor.	**tabla_journey_javier**
Margen combinado	Margen retail más margen distribuidor.	**tabla_journey_javier**
Año venta	Año en el que se realizó la venta.	**tabla_journey_javier**
Comuna	Comuna de residencia cliente.	**tabla_journey_javier**
Dirección	Dirección cliente	**tabla_journey_javier**
Fecha modificación	Fecha modificación registro	**tabla_journey_javier**
Rut	Rut encriptado.	**tabla_journey_javier**
Modelo	Modelo del vehículo	**patentes_historico_02**
Marca	Marca del vehículo.	**patentes_historico_02**
Patente	Identificador único de patente	**patentes_historico_02**
Rut	Rut dueño del vehículo.	**patentes_historico_02**
Año patente	Año en el que se adquirió la patente.	**patentes_historico_02**
Tipo	Tipo del vehículo.	**patentes_historico_02**
Año fabricación	Año de fabricación del vehículo	**patentes_historico_02**

4.2.3. Análisis exploratorio de datos y resultados preliminares

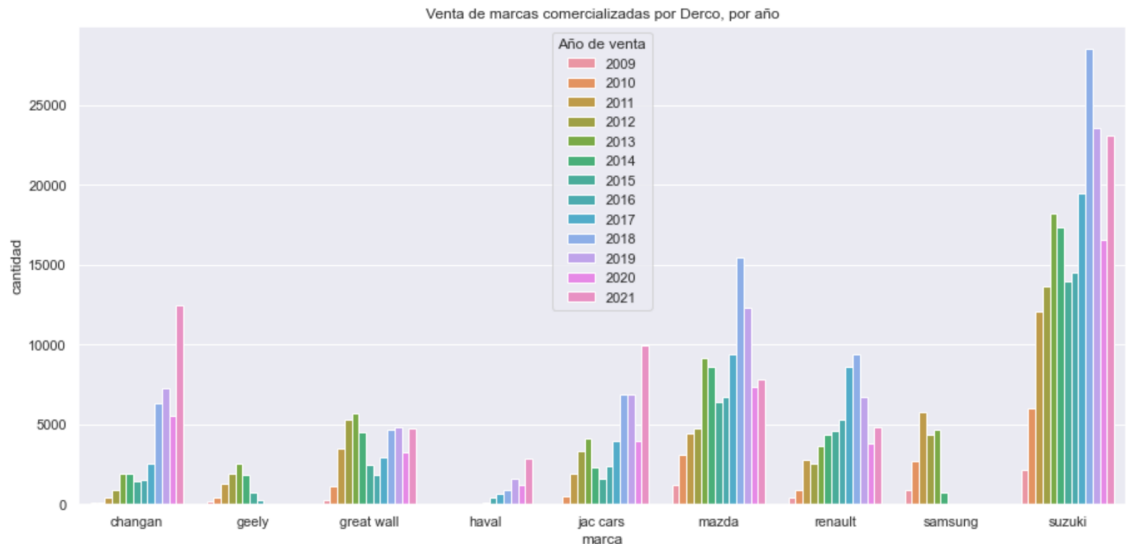
Una vez generado el dataset normalizado y enriquecido con la data mostrada anteriormente, procedemos a generar algunas vistas preliminares que nos otorgarán una visión macro de la distribución y comportamiento de las ventas de la automotora.

Ventas por marca en el periodo de estudio

Lo primero que nos gustaría estudiar con la data obtenida de la automotora, es como se distribuyen las ventas a través del tiempo, con el fin de hacernos una idea general del comportamiento de las ventas según las marcas que comercializa la automotora.

Figura 6

Ventas anuales por marca



Nota. Distribución de las ventas por marca de vehículos del año 2009 a 2021.

Tal como muestra la figura 6, las ventas por marca van aumentando progresivamente cada año alcanzando un peak en los años 2018 y 2021.

Podemos observar que la mayor cantidad de las ventas se concentran en la marca Suzuki, seguido por Mazda con una diferencia significativa en las ventas entre estas dos marcas.

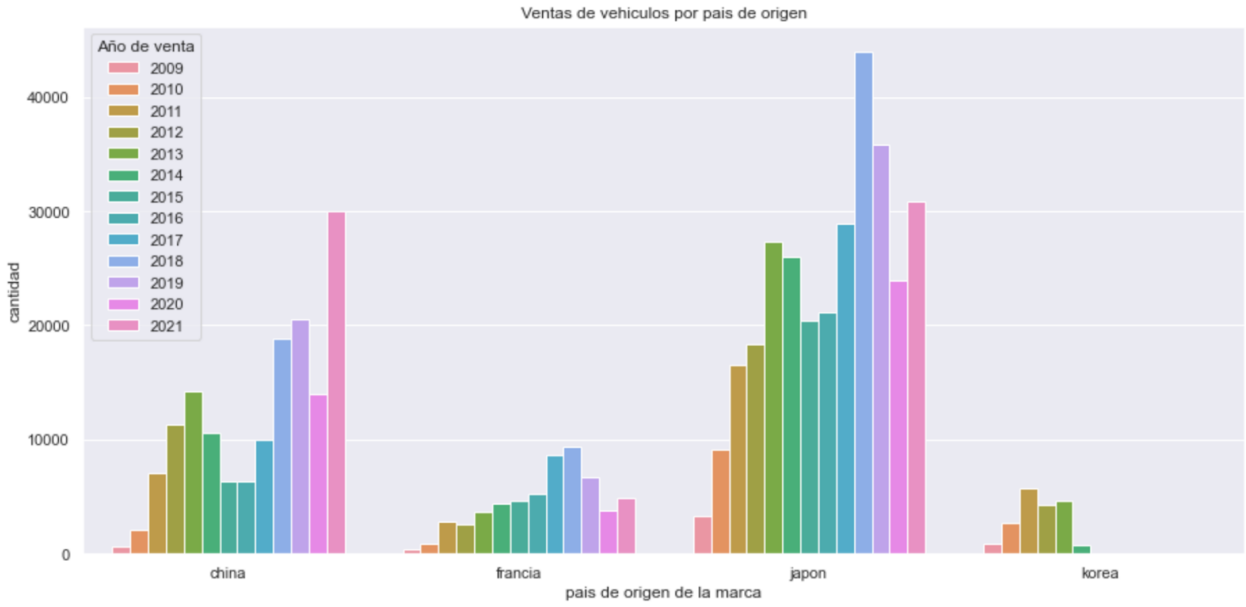
Por otro lado, podemos observar que las marcas Geely y Samsung ya no son comercializadas por la automotora hasta el fin del periodo de estudio.

Ventas por país de origen de la marca de vehículos

El siguiente gráfico, muestra las ventas de cada región de origen de las marcas de vehículos, donde cada barra corresponde a la venta de un año en particular del periodo en análisis:

Figura 7

Ventas anuales por país



Nota. Distribución de las ventas por país de origen de la marca de vehículos desde el año 2009 a 2021.

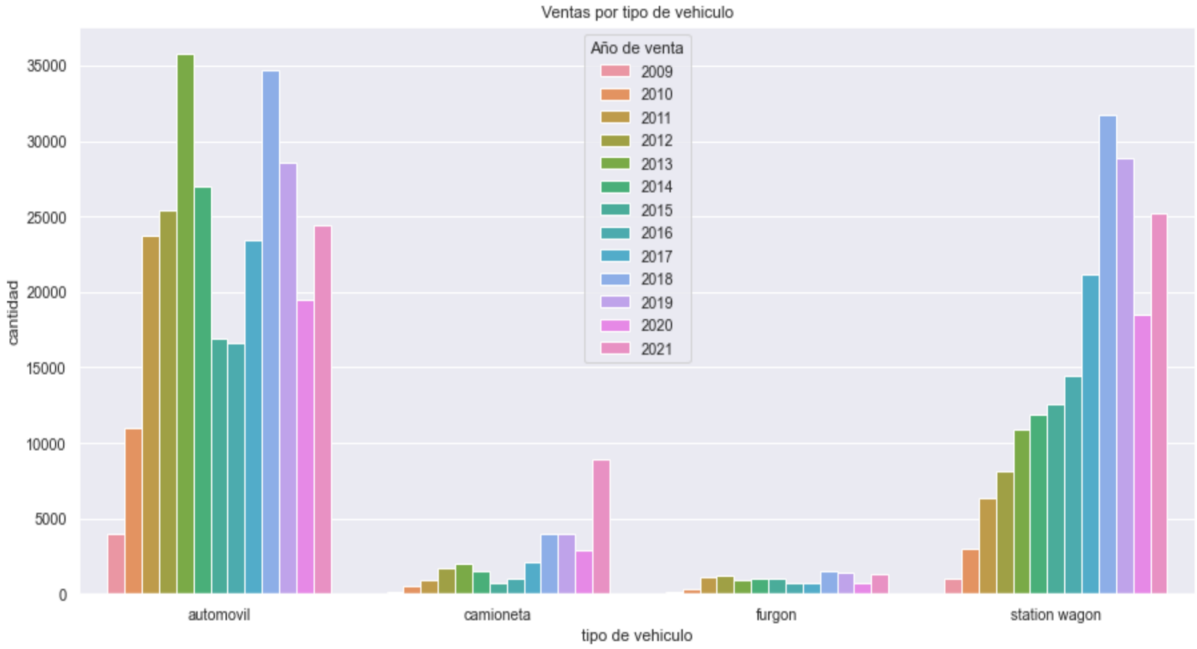
Tal como se puede observar en la figura 7, los vehículos japoneses son los que predominan en ventas dentro del periodo de estudio. En el caso del país de origen Corea, con la marca Renault-Samsung y con el país de origen Francia con Renault y su unión entre el grupo Renault-Nissan-Samsung, se prefiere no juntar las ventas de la primera marca hacía Renault, debido a que los vehículos Renault-Samsung son construidos en Corea, pero sus motores y Chasises son provenientes desde fábricas de Renault.

Distribución de Ventas por tipo de vehículo

El siguiente gráfico muestra las ventas de cada tipo de vehículos comercializado por la automotora, donde cada barra corresponde a la venta de un año en particular del periodo en análisis.

Figura 8

Ventas anuales por tipo de vehículo



Nota. Distribución de las ventas por tipo de vehículo desde el año 2009 a 2021.

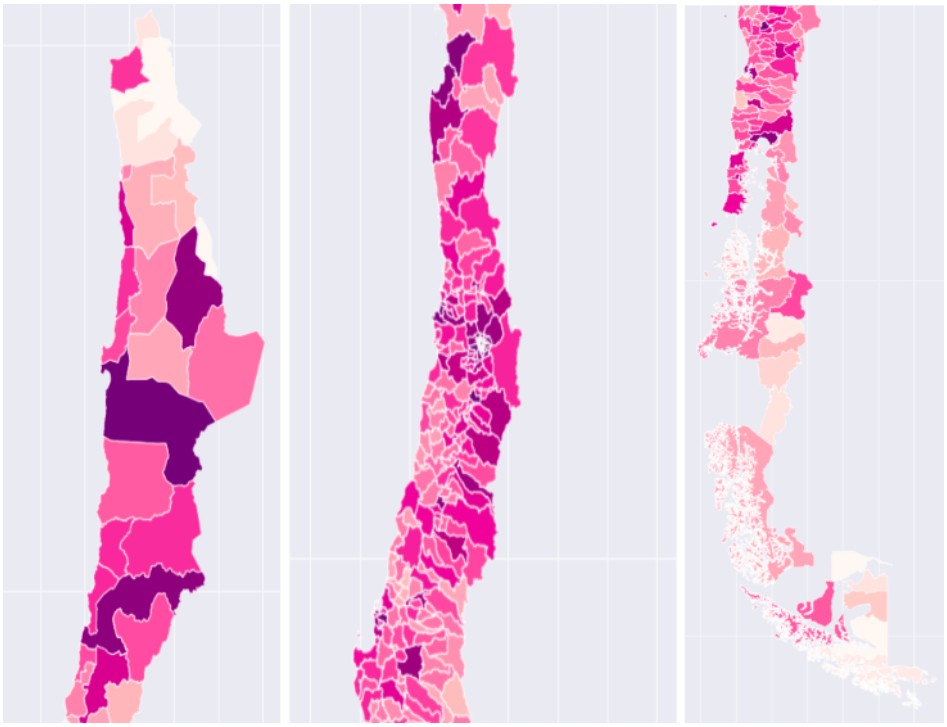
Tal como se puede observar en la figura 8, los vehículos de tipo automóvil, son los que tienen venta predominante en las preferencias de compra de los vehículos comercializados por la concesionaria dentro del periodo de estudio.

Distribución de ventas por comuna

En la siguiente figura se muestra cada comuna del territorio nacional, donde las zonas más oscuras representan aquellas que concentran una mayor cantidad de las ventas en relación con la población de dicha comuna.

Figura 9

Cantidad de ventas a nivel nacional



Nota. Distribución de las ventas de vehículos de la concesionaria en mapa de Chile.

Tal como podemos observar en la figura 9, la automotora tiene una presencia relativamente alta en todo el territorio, a excepción del extremo sur-austral, donde la población es mucho menor en comparación al resto del territorio nacional.

4.2.4. Implementación modelos de clasificación

En esta sección nos enfocaremos en la generación de clústers en base a la información de venta a través de la implementación de modelos de aprendizaje no supervisado de clasificación. Las variables que estudiaremos son las siguientes:

- marca (categórica)
- año de venta (categórica)
- comuna de residencia del comprador (categórica)
- modelo de vehículo (categórica)
- país de origen de la marca (categórica)
- región de origen de la marca (categórica)

- tipo de automóvil (categórica)
- precio de lista (numérica)

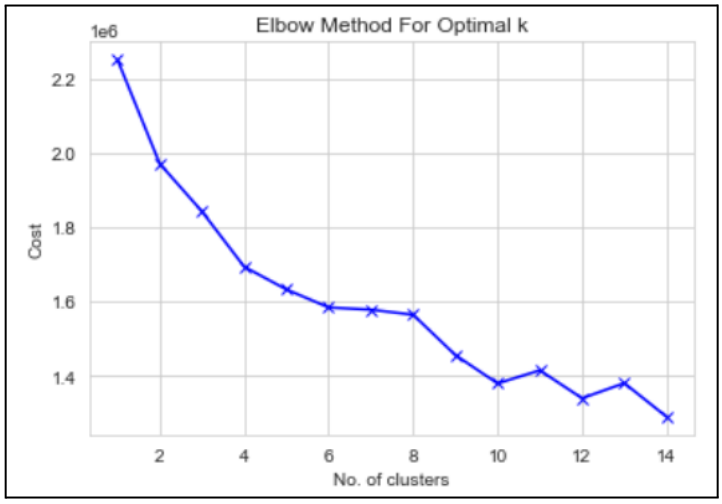
Teniendo en cuenta las variables descritas, nuestro objetivo es encontrar patrones en base a los modelos de clasificación, utilizando estas variables mediante la implementación de modelos de aprendizaje no supervisado de clasificación. Por otro lado, las variables de estudio corresponden a variables numéricas y categóricas por lo que nos enfocaremos en la implementación de dos tipos de modelos que operan de forma eficiente con este tipo de datos: Modelo de aprendizaje no supervisado **K-Modes** sólo para las variables categóricas y **K-Prototypes** para las variables categóricas y numéricas.

1. K-Modes

Es un algoritmo especializado para trabajar con grandes cantidades de datos categóricos. Opera mediante la obtención del valor *K* el cual representa la moda de cada conjunto y le asigna a cada valor del conjunto de datos la moda más cercana. Para implementar este algoritmo debemos determinar el número óptimo de clústers, para lo cual utilizaremos el método del codo:

Figura 10

Método del codo en K-Modes



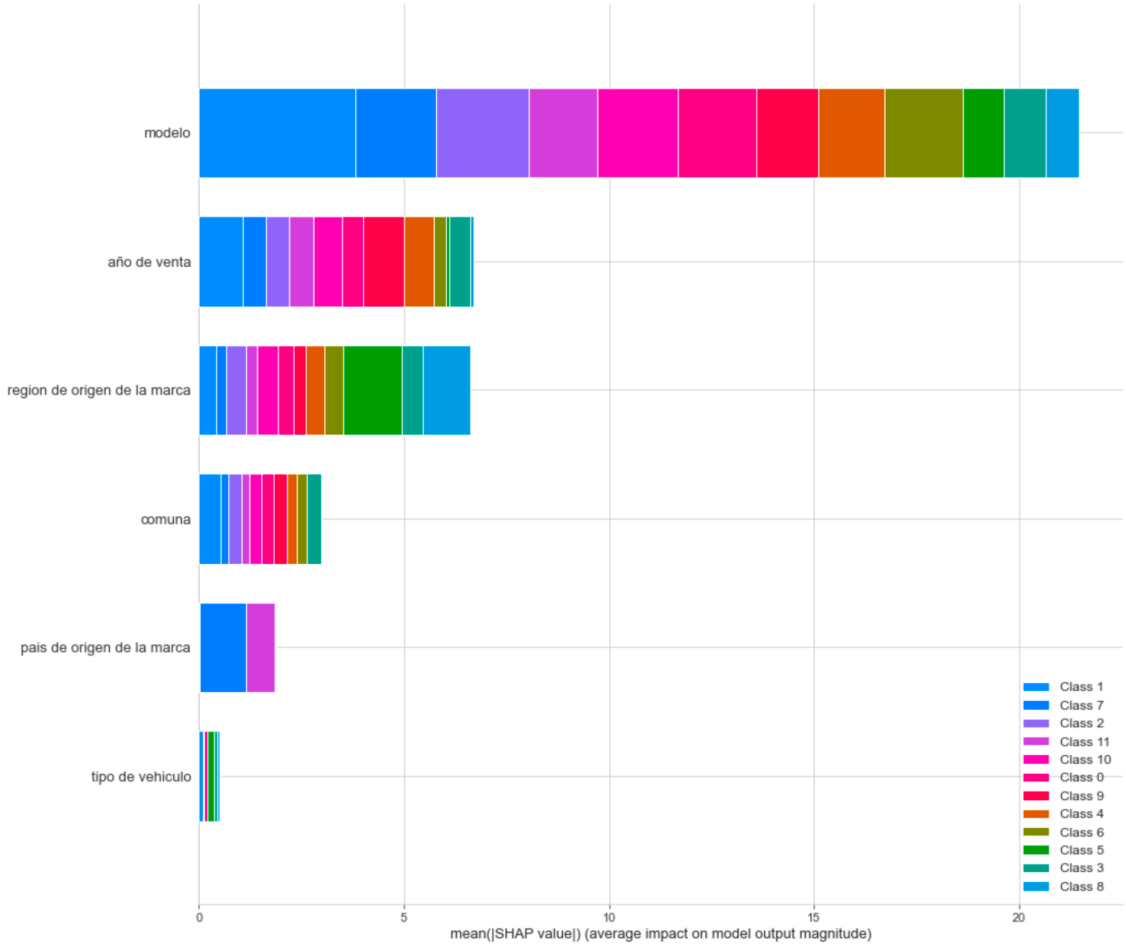
Nota. Para los datos, la cantidad óptima de conjuntos es 12.

Podemos observar en la Figura 10, que la varianza explicada comienza a disminuir de forma lineal en 12, por lo que esta será la cantidad de clústers que utilizaremos para el modelo de clasificación K-Modes.

Con esta información, implementamos el algoritmo de clasificación y utilizando el algoritmo de inteligencia artificial explicable SHAP, podemos determinar cuánto aporta cada variable a los clústers clasificados por el algoritmo:

Figura 11

Explicabilidad de modelo K-Modes con SHAP



Nota.: Distribución de las variables para el modelo K-Modes, en base a análisis SHAP.

Para evaluar la efectividad de la clasificación realizada por el algoritmo K-Modes, calculamos la métrica F1-Score la cual resulta en 99,08% a través de un modelo LightGBM. Tal como muestra la figura 11, el modelo del vehículo, es la variable que más aporta al modelo de clasificación, en segundo lugar el año de venta y en tercer lugar el origen de la marca.

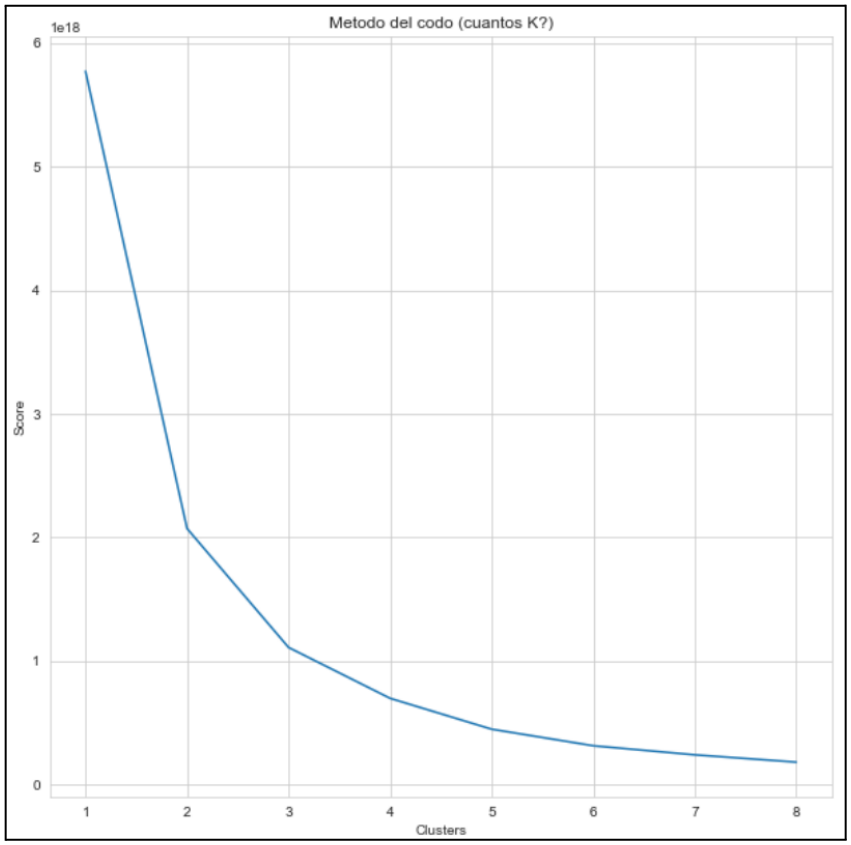
2. K-Prototypes

Es un algoritmo utilizado en conjuntos de datos mixtos que combinan datos categóricos y numéricos. Este algoritmo calcula la distancia entre la data categórica y los clústers para luego asignarlos al más cercano.

Al igual que con K-Modes, debemos determinar el número óptimo de clústers para el algoritmo, por lo que utilizamos nuevamente el método del codo:

Figura 12

Método del codo en K-Prototypes



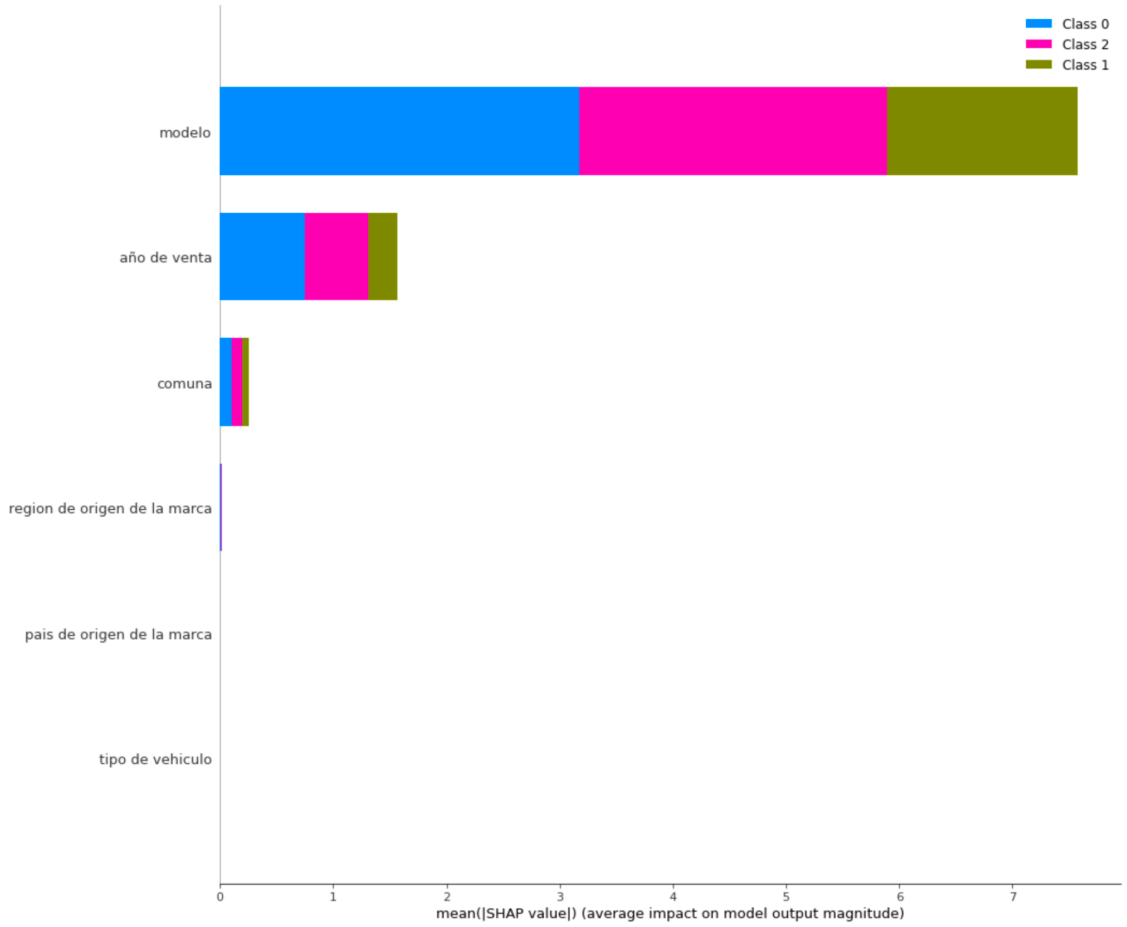
Nota. Para los datos, la cantidad óptima de conjuntos es 3.

Podemos observar en la figura 12, que la varianza explicada comienza a disminuir de forma lineal en 3, por lo que esta será la cantidad de clústers que utilizaremos para el modelo de clasificación K-Prototypes.

Con esta información, implementamos el algoritmo de clasificación y utilizando el algoritmo de inteligencia artificial explicable SHAP, podemos determinar cuánto aporta cada variable a los clústers clasificados por el algoritmo:

Figura 13

Explicabilidad de modelo K-Prototypes con SHAP



Nota. Para los datos, la cantidad óptima de conjuntos es 3 con K-Prototypes.

Para evaluar la efectividad de la clasificación realizada por el algoritmo K-Prototypes, calculamos la métrica F1-Score la cual resulta en 84,21% a través de un modelo LightGBM. Tal como muestra la figura 13, el modelo del vehículo, es la variable que más aporta al modelo de clasificación, en segundo lugar el año de venta y en tercer lugar el origen de la marca.

4.2.5. Implementación de modelos predictivos

En esta última sección tomaremos los clústers creados con K-Modes y K-Prototypes para evaluar cuatro alternativas de modelos predictivos con el fin de poder predecir y validar las clasificaciones que han resultado de los modelos de aprendizaje no supervisado de clasificación K-Modes y K-Prototypes:

- Regresión Lineal
- Redes neuronales
- XGBoost
- Random Forest

Para la implementación de estos modelos predictivos, se utiliza la variable que resulta de los modelos de clasificación, indicando a qué clúster pertenece cada registro y luego, esta variable formará parte de las variables independientes que alimentarán los modelos. Las variables independientes que utilizaremos son las siguientes:

Tabla 4

Variables independientes

Variable	Dataset de origen
Marca primer vehículo (0)	tabla_journey_javier
Año venta primer vehículo (1)	tabla_journey_javier
Comuna primer vehículo (2)	tabla_journey_javier
Modelo primer vehículo (3)	tabla_journey_javier
País marca primer vehículo (4)	juicio experto
Origen marca primer vehículo (5)	juicio experto
Tipo primer vehículo (6)	patentes_historico_02
Precio primer vehículo (7)	tabla_journey_javier
Clúster primer vehículo (8)	A partir de modelo de aprendizaje no supervisado K-Modes y K-Prototypes

Luego utilizaremos la variable correspondiente a la marca del próximo vehículo que comprará un cliente. Esta variable constituirá la variable dependiente a predecir en todos los modelos.

Para cada modelo utilizamos 34.779 registros de entrenamiento y 14.906 de prueba. Por último, calculamos las principales métricas de clasificación, con las cuales determinamos que el modelo predictivo con mejor rendimiento es **Random-Forest** para los dos métodos de agrupación K-Modes, K-Prototypes y la predicción sobre cómo se comportan las variables, cuando se predice en base a tomar en cuenta las los clusters resultantes del capítulo 4.2.4.

1. Modelo predictivo con cluster K-Prototypes

A continuación en la figura 14, se muestran métricas principales del modelo construido a partir de un Random-Forest incluido el clúster (variable creada a partir de los modelos de clasificación antes propuestos) con K-Prototypes. En este caso, la variable que mejor predice el modelo, es el cluster_id clasificado con K-Prototypes del primer vehículo con un f1-score de 91.7%. La precisión general de este modelo predictivo con el cluster K-Prototypes, tiene un f1-score de 85.91%

Figura 14

Principales métricas para modelo Random Forest

	precision	recall	f1-score	support
marca del primer automovil	0.580	0.776	0.663	441
año de venta del primer automovil	0.547	0.472	0.507	123
comuna cliente	0.579	0.655	0.615	862
modelo automovil	0.600	0.590	0.595	188
pais de origen de marca de primer automovil	0.657	0.673	0.665	1131
origen de marca de primer automovil	0.881	0.950	0.914	2432
tipo primer automovil	0.919	0.821	0.867	1606
precio de lista primer automovil	0.764	0.674	0.716	573
cluster_id kprototypes primer automovil	0.931	0.904	0.917	7550
accuracy			0.850	14906
macro avg	0.718	0.724	0.718	14906
weighted avg	0.856	0.850	0.852	14906

Nota. Random-Forest con cluster K-Prototypes

2. Modelo predictivo con cluster K-Modes

A continuación en la figura 15, se muestran métricas principales del modelo construido a partir de un Random-Forest incluido el clúster (variable creada a partir de los modelos de clasificación antes propuestos) con K-Modes. En este caso, la variable que mejor predice el modelo, es también el cluster_id clasificado con K-Modes del primer vehículo con un f1-score de 95.8%. La precisión general de este modelo predictivo con el cluster K-Modes, tiene un f1-score de 90.3%

Figura 15

Principales métricas para modelo Random Forest

	precision	recall	f1-score	support
marca primer auto	0.638	0.791	0.706	441
año venta primer auto	0.580	0.415	0.483	123
comuna primer auto	0.722	0.749	0.735	862
modelo clean primer auto	0.582	0.527	0.553	188
pais origen marca primer auto	0.731	0.679	0.704	1131
origen marca primer auto	0.959	0.978	0.968	2432
tipo vehiculo primer auto	0.966	0.923	0.944	1606
precio de lista primer auto	0.823	0.724	0.771	573
cluster id kmodes primer auto	0.952	0.963	0.958	7550
accuracy			0.903	14906
macro avg	0.773	0.750	0.758	14906
weighted avg	0.903	0.903	0.902	14906

Nota. Random-Forest con cluster K-Modes

A continuación, se expone la importancia de las variables dentro del modelo predictivo creado con todas las variables, incluido el cluster generado con el modelo de clasificación K-Modes, dando como resultado, que la variable correspondiente al modelo de clasificación (8), es la variable más importante en la predicción de resultados:

Tabla 5

Variable	Importancia en modelo
Clúster ID K-Modes	0.195490
Precio de lista primer vehículo	0.177285
Comuna primer vehículo	0.176695
Marca primer vehículo	0.156218
País de origen vehiculo	0.114203
Comuna propietario vehículo	0.064498
Año venta vehiculo	0.053042
Región origen marca vehículo	0.032657
Tipo de vehículo	0.029912

3. Modelo predictivo sin clusters generados con modelos de aprendizaje no supervisado de clasificación (sólo variables independientes)

A continuación en la figura 16, se muestran métricas principales del modelo construido a partir de un Random-Forest, en esta ocasión, sin incluir las variables correspondientes a los clusters K-Modes y K-Prototypes. En este caso, la variable que mejor predice el modelo, es el país de origen de la marca del primer vehículo con un f1-score de 90.9%. La precisión general de este modelo predictivo sin clusters, tiene un f1-score de 85.13%.

Figura 16

Principales métricas para modelo Random Forest

	precision	recall	f1-score	support
marca_primer_auto	0.605	0.734	0.663	455
year_Venta_primer_auto	0.485	0.463	0.474	108
tj_comuna_primer_auto	0.574	0.678	0.622	848
modelo_clean_primer_auto	0.612	0.467	0.530	199
pais_origen_marca_primer_auto	0.654	0.678	0.666	1129
origen_marca_primer_auto	0.886	0.934	0.909	2380
v_tipo_primer_auto	0.918	0.833	0.873	1577
precio_de_lista_primer_auto	0.752	0.681	0.715	609
micro avg	0.772	0.790	0.781	7305
macro avg	0.686	0.684	0.682	7305
weighted avg	0.779	0.790	0.782	7305

Nota. Random-Forest con cluster K-Modes

A continuación, se expone la importancia de las variables dentro del modelo predictivo creado con todas las variables, incluido el cluster generado con el modelo de clasificación K-Modes, dando como resultado, que la variable correspondiente al modelo de clasificación (8), es la variable más importante en la predicción de resultados:

Tabla 6

Variable	Importancia en modelo
Precio lista primer vehículo	0.25247001
Modelo primer vehículo	0.21350277
País origen marca vehículo	0.16696872
País origen marca vehículo	0.11325435
Comuna propietario vehículo	0.09984172
Año venta vehiculo	0.06052205
Origen marca vehículo	0.05322039
Tipo de vehículo	0.04021998

5. Resultados

A partir de los resultados expuestos en las métricas de los modelos de predicción Random-Forest con los clústers resultantes de los modelos de clasificación K-Modes, K-Prototypes y un modelo predictivo con las variables de los clusters, se puede interpretar que el modelo predictivo en base al modelo de clasificación K-Modes, es el que mejor puede predecir la marca que comprara un cliente de esta concesionaria en cuestión de estudio, con un porcentaje de predicción de los datos f1-score de 90.3%, respecto del estudio de los clusters en base a K-Prototypes, el cual tiene una capacidad de predicción de los datos f1-score de 85.91%.

Respecto del modelo predictivo sin las variables correspondientes a los clusters resultantes con K-Modes y K-Prototypes, este tiene una capacidad predictiva f1-score de 85.13%, lo cual revela que el cluster modelo predictivo desarrollado con K-Prototypes, con respecto al que no tiene cluster, es bastante parecido, pudiendo inferir que la variable del cluster, no aporta considerablemente al modelo, como se puede observar en la tabla de importancia de features del modelo predictivo con cluster K-Prototypes:

Tabla 7

Variable	Importancia en modelo
Modelo primer vehículo	0.215083
Precio lista primer vehículo	0.209017
Marca primer vehículo	0.161216
País origen marca vehículo	0.115879
Comuna propietario vehículo	0.099841
Año venta vehiculo	0.060522
Origen marca vehículo	0.053220
Tipo de vehículo	0.040219

6. Conclusiones

Con la realización del análisis preliminar de la información de venta, podemos observar que la automotora se especializa en la comercialización de vehículos de origen asiático, con énfasis en las marcas Suzuki y Mazda. Las ventas van aumentando progresivamente hasta el año 2018, luego bajan los años 2019 y 2020, para luego volver a aumentar el año 2021. El periodo 2019 y 2020 comprende los años en los cuales ocurrieron sucesos tales como el estallido social (suceso local) y pandemia COVID-19 (suceso global), por lo que podemos atribuir este cambio en la tendencia de venta a estos sucesos. Esto constituye un detalle a considerar, dado que es un periodo de tiempo atípico dentro del conjunto de datos.

Por otro lado, la utilización de los modelos de aprendizaje no supervisado para la clasificación de los registros, en base a modelos de tipo K-Modes y K-Prototypes, nos permitió generar clústers robustos en base a los datos, como también a identificar que la variable de mayor relevancia para el proceso de clasificación es la variable marca del vehículo en ambos algoritmos.

Por último, la utilización de la variable generada a través de modelos de aprendizaje no supervisados K-Modes y K-Prototypes, que identifique el clúster al que pertenece cada registro del dataset, aporta positivamente a los modelos desarrollados en base a modelos de aprendizaje supervisado que determinan la marca del vehículo que comprara un cliente, modelos tales como Linear-Regression, XGBoost, Neural-Networks, y Random-Forest, esto dado los valores de la métrica f1 score obtenidas en cada modelo.

Bibliografía

- ANAC. (s/f). Anac.cl. Asociación Nacional Automotriz de Chile. https://www.anac.cl/.
- Abdul-Rahman, S., Arifin, N. F. K., Hanafiah, M., & Mutalib, S. (2021). Customer segmentation and profiling for life insurance using K-modes clustering and decision tree classifier. International Journal of Advanced Computer Science and Applications: IJACSA, 12(9). https://doi.org/10.14569/ijacsa.2021.0120950
- Ai, L. (2022, enero 18). Explicabilidad de modelos de Machine Learning con SHAP. Squeeze the Data. https://www.squeezethedata.com/explicabilidad-de-modelos-de-machine-learning-con-shap-2/
- Kumar, S. (2021, mayo 7). Clustering Algorithm for data with mixed Categorical and Numerical features. Towards Data Science. https://towardsdatascience.com/clustering-algorithm-for-data-with-mixed-categorical-and-numerical-features-d4e3a48066a0
- Huang, Z. (s/f). Extensions to the k-means algorithm for clustering large data sets with categorical values. Edu.hk. Recuperado el 5 de diciembre de 2022, de https://cse.hkust.edu.hk/~qyang/Teaching/537/Papers/huang98extensions.pdf
- Zhang, L., Priestley, J., DeMaio, J., Ni, S., & Tian, X. (2021). Measuring customer similarity and identifying cross-selling products by community detection. Big Data, 9(2), 132–143. https://doi.org/10.1089/big.2020.0044
- Ruberts, A. (2020, mayo 16). K-prototypes - customer clustering with mixed data types. Well Enough. https://antonsruberts.github.io/kproto-audience/

Coevaluaciones y Autoevaluaciones

Autoevaluación Nicolas Pastor Bueno
Esta última parte del proyecto ha sido bastante desafiante tanto en la parte técnica, como en la parte personal. Creo que mi desempeño en esta última parte fue la apropiada al contexto.
Nota autoevaluación: 7.

Autoevaluación Maria Alejandra Hermosilla Urriola
Respecto de esta última parte, fue bastante compleja debido a que me he encontrado con situaciones laborales complejas que han quitado foco a esto. De igual manera he intentado poner toda la ayuda necesaria para lograr este objetivo
Nota autoevaluación: 7.

Coevaluacion Nicolas Pastor Bueno a Maria Alejandra Hermosilla Urriola
Maria Alejandra ha tenido muy buena disponibilidad en colaborar para esta parte del proyecto que desarrollamos juntos, aportando en la idea más general del proyecto y poniendo foco en detalles. Por este motivo la co-evaluación a Maria Alejandra en la escala 0 a 7 es 7

Autoevaluación Maria Alejandra Hermosilla Urriola a Nicolas Pastor Bueno
Nicolas ha podido desarrollar la parte técnica de este proyecto con gran compromiso, aportando con sus competencias técnicas. Nota coevaluación: 7.