UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

DESARROLLO DE UN MODELO DE PREDICCIÓN DE ASISTENCIA A SERVICIOS DE MANTENCIÓN AUTOMOTRIZ Y RECOMENDACIÓN DE ACCIONES PARA MEJORAR LA DEMANDA

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERA CIVIL INDUSTRIAL

CAMILA ANDREA JORQUERA LÓPEZ

PROFESOR GUÍA:
PABLO MARÍN VICUÑA

MIEMBROS DE LA COMISIÓN:
ALEJANDRA PUENTE CHANDÍA
LUIS SOLARI DÍAZ

SANTIAGO DE CHILE
2022

DESARROLLO DE UN MODELO DE PREDICCIÓN DE ASISTENCIA A SERVICIOS DE MANTENCIÓN AUTOMOTRIZ Y RECOMENDACIÓN DE ACCIONES PARA MEJORAR LA DEMANDA

El mercado de las empresas automotrices está en constante crecimiento, por lo que la gran cantidad de empresas existentes y las emergentes están en un constante intento de diferenciación del resto. En este contexto, una empresa automotriz reconocida, con un nivel bajo de uso de recursos computacionales está comenzando a integrar tecnologías que permitan mejorar el desempeño de sus áreas de trabajo y principalmente del área de post-venta, donde se ofrece el servicio de mantenciones automotrices.

El servicio de mantención automotriz es un servicio con costo que presta la empresa a todos sus clientes (dueños de vehículos de la marca) para que puedan mantener su vehículo en óptimas condiciones, pero los últimos años la asistencia a estas mantenciones ha ido decayendo poco a poco. En este marco, nace la necesidad de la empresa de intentar conocer el comportamiento de asistencia a las primeras 3 mantenciones de pauta recomendadas a sus clientes, con el fin de tener espacio a determinar un plan de trabajo de marketing que logre fomentar la demanda de estos servicios.

El presente trabajo tiene por objetivo identificar clientes idóneos a asistir a servicios de mantención vehicular y sus respectivos períodos de cumplimiento de pautas de mantención, y con esto, realizar recomendaciones a la empresa con el fin de aumentar la demanda del servicio. Para esto se pretende determinar la próxima fecha y propensión de asistencia de un cliente a la próxima mantención pauta correspondiente, por lo que se trabaja en la comparación de modelos de *Machine Learning* como por ejemplo Random Forest y Máquinas de soporte vectorial para determinar los de mejor desempeño.

Para determinar la próxima fecha de asistencia se trabaja con modelos de regresión que permiten predecir los días hasta la próxima mantención y los algoritmos que mejores resultados presentan son Bosques aleatorios y Máquinas de soporte vectorial, obteniendo en los mejores casos un error porcentual absoluto medio de 20 %. Para la propensión de asistencia al servicio se utilizan algoritmos de clasificación que permiten obtener una respuesta positiva o negativa de asistencia, y se obtiene que el algoritmo de mejor desempeño en cuanto a sensibilidad es Potenciación del gradiente, con un nivel de acierto cercano a 80 % en los casos de estudio.

Finalmente se concluye que si bien los modelos realizados presentan en algunos casos un bajo nivel de predicción, siguen implicando una mejora en comparación a los métodos de predicción utilizados actualmente por la empresa. Aún así, el gran obstáculo que se presentó en el desarrollo del trabajo es la escasa cantidad de datos de calidad existentes y el bajo nivel de información sociodemográfica disponible para estudio.

A mis padres Silvia y Marco,
gracias a ellos he llegado hasta aquí.

Agradecimientos

A mis padres Marco y Silvia les agradezco toda la paciencia, amor, contención y apoyo incondicional no solo en este proceso de titulación, sino que en toda la época universitaria y, en realidad, toda la vida, porque gracias a ustedes, a su constante esfuerzo porque nunca me faltara nada y su incentivo a lograr mis metas, he logrado llegar hasta aquí. Su apoyo y abrazos son un tesoro, son los mejores papás que existen, a Dios le doy gracias por eso. Espero poder retribuirles aunque sea parte de todo lo que ustedes me han dado. Los amo con todo mi corazón.

Al mejor hermano del mundo, mi hermano Diego, gracias por creer en mí y darme el empujón que faltaba para atreverme a elegir este tema de memoria y llevarlo a cabo. Gracias por siempre creer en mí y por todas las risas que siempre me sacas.Eres un ejemplo a seguir. Te quiero mucho.

A mi Mami y mi Manina, que siempre me llenaron de cariño y las recuerdo con mucho amor, siempre las he tenido presentes en todo y siempre vuelven en mis sueños. Muchos besitos al cielo.

A mi pololo Cristián, gracias por tu comprensión y apoyo incondicional todo este tiempo, y por creer en mí siempre que dudé lograr terminar con éxito este trabajo. Gracias por incentivarme a a mejorar, por ser mi compañero de viajes y por apañarme en todas; por ayudarme a distraerme y divertirme cuando estaba en el estado de *alma perdida* al estilo de la película Soul. Te amo muchísimo.

A mis amigos de la universidad que fueron un super apoyo gran parte de estos 7 años, sin su compañía, almuerzos, juegos y carretes hubiese sido mil más difícil el pasar por la U. En especial gracias a Javi B., Jose y Martín. A mis amigas del liceo, Fabi, Pame, Cata, Caro y nuevamente Javi, que a pesar de vernos poco siempre han sido un gran apoyo y espacio de seguridad; son las hermanas que me dió la vida. Se vienen muchas juntas. Los quiero mucho.

Y como no pueden faltar, a mi Thor, Pati y Pola, las mejores mascotas que siempre están y estuvieron para llenar de amor y cariño, y subir el ánimo con sus juegos, su compañía da alegría a la vida.

Al equipo de BI de Kia, Francisco M., Francisco R y Juan C., quienes me recibieron con mucha buena onda y me hicieron sentir parte del equipo, gracias por su excelente disposición a ayudarme en lo que necesitara y su comprensión. Son un gran equipo, feliz de haber realizado la memoria en su área.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Antecedentes

1.1. Caracterización de la empresa

1.1.1. La industria automotriz

Actualmente la industria automotriz en Chile cuenta con 61 marcas oficiales de automóviles. El ente regulador de esta industria es el Ministerio de Transportes y Telecomunicaciones, quien regula el comercio de estos automóviles, y lo relacionado al transporte en general. Las marcas que mayormente destacaron hacia fines del año 2020 en cantidad de ventas de vehículos acumulados a diciembre fueron Chevrolet con un 10,5 % de las ventas, Kia con 7,4 %, Suzuki con 7,2 %, Nissan con 7 % y Hyundai con 6,7 % [2]. El registro total se adjunta en el Anexo A.1, donde se observa que Kia es una de las empresas que lidera el mercado en cuanto a ventas de vehículos livianos el año 2020 luego de la marca líder Chevrolet, aunque sus diferencias son notorias ya que Kia realiza venta de $\frac{2}{3}$ de autos de los que vende Chevrolet.

1.1.2. La empresa

Kia Chile S.A. es una empresa perteneciente a la industria automotriz, la que realiza la compra de vehículos a la empresa koreana Kia y los importa a Chile para su venta, siendo la única representante de la marca en Chile.

La empresa declara su misión como:
"*La compañía cree firmemente que el éxito en cualquier negocio solo se puede lograr a través del placer del cliente y ejercemos el máximo cuidado para asegurarnos de brindar un servicio al cliente de muy alto nivel en cada interacción con nuestros valiosos clientes. Estamos igualmente dedicados a la tarea de crear un ambiente de trabajo saludable para nuestros empleados, quienes han sido el motor de nuestro éxito a lo largo de los años.*"
y su visión:
"*Establecer una sólida cultura de marketing que impregne todos los niveles de la organización y dentro de la cual se identifiquen y satisfagan las distintas necesidades de los clientes, seguirá siendo nuestra visión a corto, medio y largo plazo.*" [1]

Al año 2021 Kia Chile cuenta con alrededor de 600 empleados, cuya estructura de cargos es jerárquica tal como se representa en Figura 1.1, contando con una Gerencia

General y siete departamentos: Digital, Marketing, Planificación, Post Venta, Finanzas, Ventas/Producción y Desarrollo Organizacional. Dentro el departamento Digital se encuentran las áreas de Desarrollo TI y Business Intelligence.

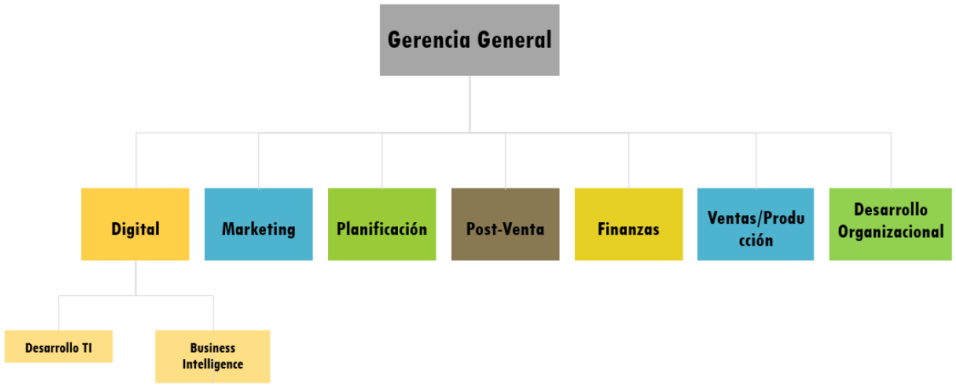


Figura 1.1: Organigrama general de KIA Chile permitido mostrar

Kia Chile, como representante de la marca koreana en Chile, realiza venta de automóviles de la marca a través de 20 empresas concesionarias denominadas "*dealers*" por la empresa, presentes a lo largo del país. A su vez, las empresas concesionarias venden estos vehículos a través de sus distintos *puntos de venta* que son sucursales en las que se pone en exposición los vehículos para ofrecerlos en venta a personas naturales o jurídicas, siendo estos el cliente final de la cadena de venta.

La relación existente entre la empresa productora koreana Kia y la chilena se basa en acuerdos de venta en los que la productora propone metas de venta de vehículos a la empresa chilena, por cuyo cumplimiento entrega bonos monetarios a la empresa chilena. Esta misma práctica se traduce en la relación entre la empresa chilena y los concesionarios asociados.

Además de la venta de vehículos, cada uno de los concesionarios de la marca ofrece los servicios de test-drive, venta de repuestos de auto, mantenciones y servicio de garantía. Las mantenciones son uno de los principales focos de trabajo del área de post-venta, y consisten en una revisión y diagnóstico del estado del automóvil, para luego realizar reparación, cambio o limpieza de algunas partes y piezas del mismo. Estas mantenciones se recomienda hacerlas cada cierta cantidad de períodos de uso o cantidad de kilómetros recorridos por el automóvil. El criterio de asistencia a las primeras mantenciones es como se indica a continuación.

1. Primera mantención: Asistencia sugerida luego de recorridos 1.000 kilómetros o 30 días desde la compra del automóvil.

2. Segunda mantención: Asistencia sugerida luego de recorridos 15.000 kilómetros o 1 año desde la compra del automóvil.

3. Tercera mantención: Asistencia sugerida luego de recorridos 30.000 kilómetros o 2 años desde la compra del automóvil.

4. Cuarta mantención: Asistencia sugerida luego de recorridos 45.000 kilómetros o 3 años desde la compra del automóvil.

Como se observa, se sugiere mantenciones cada 15.000 kilómetros recorridos o luego de un año desde la última mantención, existiendo así mantenciones que superan los 100.000 kilómetros de recorrido, pero los registros evidencian que la asistencia a mantenciones disminuye en la medida que aumenta el nivel de mantención correspondiente.

De la mano con la alta competencia existente en el mercado, se debe tener en consideración que cada concesionario que expone vehículos de la marca Kia, también expone automóviles de otras marcas en un mismo punto de venta. En consecuencia, los servicios de post-venta también cuentan con un mercado competitivo, dado que cada marca existente en el mercado automotriz ofrece su propio servicio de post-venta, además de existir otras empresas que ofrecen estos servicios. Según los registros del INE (2020), en el año 2019 habían 19.402 empresas registradas con subrubro de *Mantenimiento y reparación de vehículos automotores*.

De acuerdo a los registros de circulación al año 2021 existen 503.242 vehículos de la marca vigentes. Además, el año 2020 se registró un total de 19.187 automóviles marca Kia vendidos lo que corresponde a un 7,4 % de las ventas de mercado según la Asociación Nacional Automotriz de Chile[2]. La cantidad de ventas de la marca se puede desglosar según el tipo de vehículo, entre los que están los vehículos de pasajeros (vehículos más comunes), los vehículos tipo SUV (vehículos deportivos de utilidad) y los vehículos comerciales (utilizados para transportar cargas), de los que la cantidad vendida por la marca corresponde a un 13,6 %, 4,7 % y 8,7 % respectivamente, del total de ventas del mercado según tipo de vehículo, tal como se detalla en la Tabla 1.1.

Tabla 1.1: Ventas a público acumuladas de vehículos Kia del año 2020. Fuente:ANAC[3]

Tipo de automóvil	**Unidades**	**% del total de mercado**
Vehículo de pasajeros	12.269	13,6 %
SUV	4.559	4,7 %
Vehículo Comercial	2.359	8,9 %
Total	**19.187**	**7,4 %**

Las ventas de automóviles realizadas han ido en aumento desde el año 2016 hasta 2018, al igual que la tendencia de mercado, como se puede ver en la Figura 1.2, alcanzado su máximo nivel en el año 2018. A partir del año 2019, se puede observar que las ventas disminuyeron, lo que se puede explicar debido al estallido social y la pandemia experimentadas en el país desde ese año hasta la actualidad.

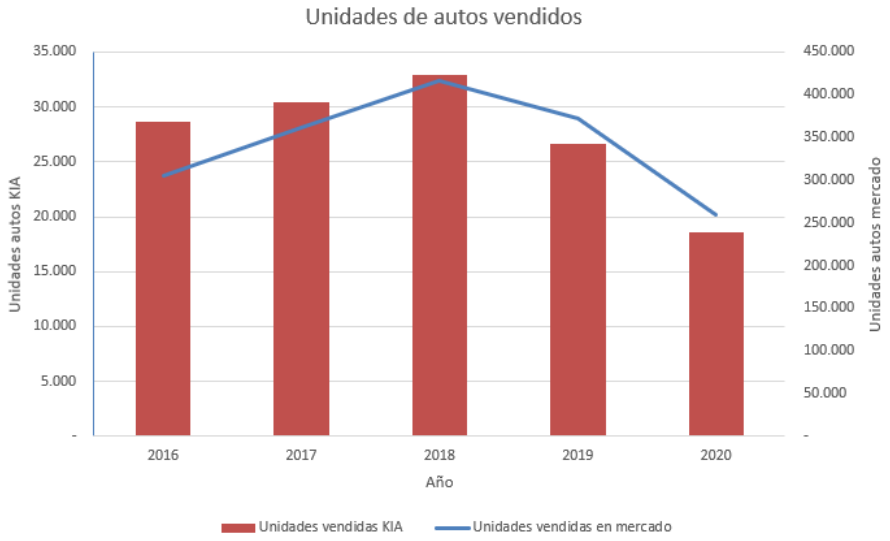


Figura 1.2: Evolución ventas realizadas. Fuente: Kia Chile

A continuación en la Figura 1.3 se observa la evolución de la cantidad de asistencias a los servicios de post-venta de la marca. La empresa estima que aproximadamente un 15 % de las asistencias realizadas no son registradas, por lo que el número total efectivo es mayor al que se expone pero es desconocido. Se puede observar que en general la concentración de asistencia se debe a las mantenciones vehiculares. En cuanto a la evolución de los servicios de post-venta, según el tipo de servicio existe una gran variación en la demanda de los servicios y productos, ya que algunos de ellos como desabolladuras y pintura tienden a tener un peak en los años 2012-2013, mientras que las mantenciones automotrices tienen su mayor demanda entre los años 2014 y 2019. Aún así, todos los servicios se enfrentaron a una disminución en su demanda debido a la situación social desde el año 2019 al presente.

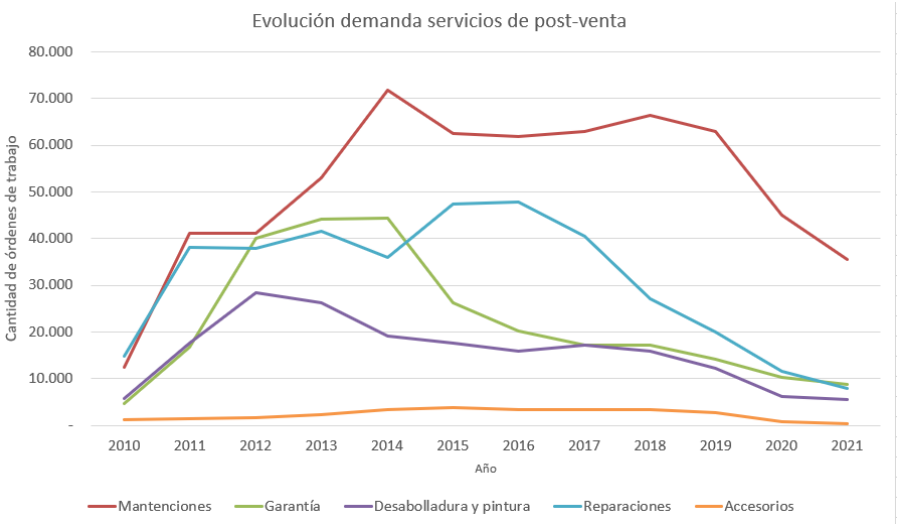


Figura 1.3: Evolución demanda de servicios de post-venta. Fuente: Kia Chile

Capítulo 2

Planteamiento del problema y justificación

2.1. Áreas relevantes de la empresa

2.1.1. Área Business Intelligence

El área de Business Intelligence en la empresa Kia nace en el año 2019 y es la encargada de realizar análisis de datos y marketing digital, con lo que permiten incorporar las bases de datos existentes en la empresa para generar reportes automatizados, diseño de KPIs y plantear estrategias digitales con el objetivo de aumentar ventas y asistencia a servicios de post-ventas mediante la comprensión del usuario de la marca.

Esta área está conformada por el Jefe de área Francisco Mallea, y dos Ingenieros analistas, Juan Calquin y Francisco Robles.

En esta área nace la iniciativa de mejorar la estimación de asistencia a los servicios de post venta, con el fin de generar incentivos más precisos y personalizados para aumentar la asistencia a los servicios de post venta de parte de los usuarios de vehículos de la marca.

De esta forma, el cliente del área de Business Intelligence son el resto de áreas de la empresa. En específico, en este caso el cliente es el área de Post-venta, a quién se le entregarán los resultados de este trabajo con el fin de mejorar el marketing directo orientado a incentivar la asistencia a los servicios de mantención pertenecientes al área de Post-venta.

2.1.2. Gerencia de Post venta

La gerencia de Post venta es la encargada de todo el servicio posterior a la venta de un automóvil. Su estructura está conformada como se puede observar en la Figura 2.1, por 4 áreas:

1. Planificación y precios: Área que tiene la función de gestionar incentivos, generación de data y análisis de información, desarrollo de proyectos comerciales y campañas

promocionales, gestión CSI, Mantenciones Pre-Pagadas y Seguros.

2. Comercial Postventa: Área responsable de coordinar y canalizar todas las necesidades comerciales entre la red de concesionarios y Kia Chile, incluyendo las visitas de Repuestos y Asistencia Técnica.

3. Asistencia Técnica: Área encargada de atender los problemas funcionales de los vehículos; dar la asistencia técnica, cumplir con las garantías y velar por la correcta comercialización de los vehículos mediante la certificación del cumplimiento de los estándares que se imponen en la empresa.

4. Abastecimiento: Se encargan de los insumos y herramientas necesarias para los servicios técnicos estén disponibles para su venta a público y uso en los servicios de mantención.

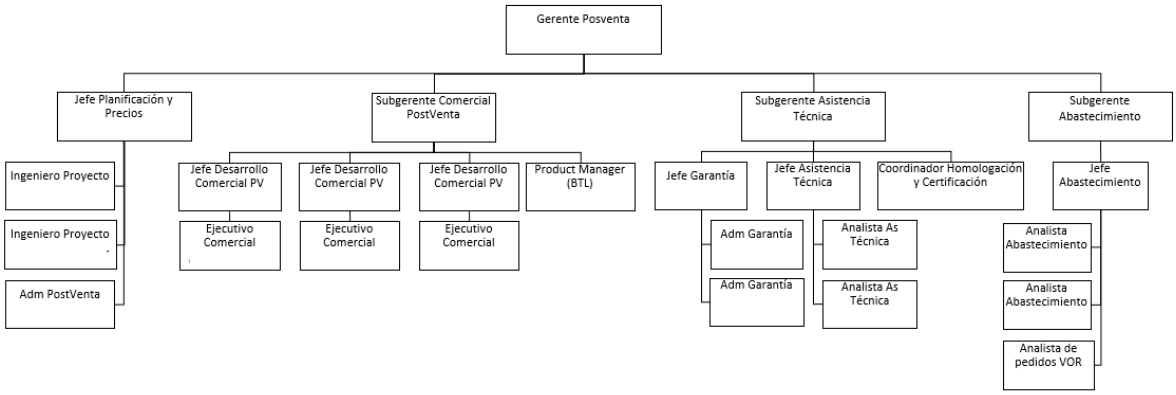


Figura 2.1: Organigrama área de Post venta Kia Chile

La gerencia de Post venta se encarga de que los concesionarios, que son quienes llevan a cabo los servicios de post-venta finalmente, cumplan con los estándares de calidad impuestos por la marca koreana, con el fin de dar cumplimiento a su misión y visión, enfocadas en entregar satisfacción mediante sus productos y servicios a los clientes.

La empresa en general, este último año se ha encontrado en una fase de cambio, intentando poner un foco aún mayor en los clientes y su satisfacción, por lo que se han realizado cambios estructurales dentro de la organización de la empresa, como unión de áreas y gerencias, y cambios en la forma de operación general.

De la mano con esto, es de interés de la empresa que los servicios posteriores a la venta del vehículo (mantención, garantía, reparaciones) sean contratados por los dueños de vehículos de la marca en los *servicios autorizados* por Kia, porque se tiene la noción que una persona dueña de un vehículo, luego de 3 años cambia su auto por uno nuevo, es decir, lo renueva. Por esta razón, y en conjunto con la alta competencia existente en el mercado, la empresa considera que si se mantiene a un cliente satisfecho con el producto y con el servicio que se le entrega en general, es más probable que vuelva a comprar un

vehículo en la misma marca.

Para incentivar a la asistencia a la marca, el área de post venta se encarga de contactar a los usuarios de la marca para avisarles que les corresponde un asistir a un servicio o entregarles información sobre la marca.

Como se mencionó anteriormente, Kia a través de sus concesionarios ofrece los servicios de post venta consistentes en:

- Garantías: Luego de vendido un automóvil, este contará con garantía por 36 meses o 100.000 kilómetros, la condición que primero ocurra. Este servicio sólo se mantiene vigente si sus mantenciones son realizadas en los servicios autorizados por la marca y que estén *al día*. Esta garantía se aplica en caso de que el auto presente una falla que sea atribuible a un defecto de producto, lo que debe ser corroborado por el servicio técnico autorizado por la marca.
- Mantenciones: La mantención vehicular es la revisión y diagnóstico del estado del vehículo para su posterior limpieza y reparación consistente en cambio, ajuste o limpieza de piezas. Las mantenciones son recomendadas luego de recorridos (dependiendo del modelo) los 10.000, 15.000, 20.000, 30.000 y 45.000 kilómetros o cada 6, 9, 12, 15 y 18 meses respectivamente. Estas mantenciones pueden ser prepagadas al momento de adquirir el vehículo o pagadas en el momento mismo de la mantención.

Con foco en el servicio de mantención, el cliente es toda persona que compró un vehículo de la marca y que asiste al servicio de mantención en un concesionario autorizado por Kia.

Dado lo anterior y como se mencionó anteriormente, a cada usuario de un vehículo de la marca se le contacta con la intención de incentivarlo a asistir a los servicios de mantención. Para que este contacto logre su fin, se realiza a modo de recordatorio cuando está próximo a la fecha en que le corresponde asistir al servicio, pero para esto es necesario conocer la fecha en que un cliente debería asistir.

Actualmente, el área de post venta genera contacto directo con los dueños de los vehículos Kia mediante correos y en determinadas fechas, en las que se le avisa que le corresponde asistir al servicio de mantención. Este contacto se genera en base a los periodos promedio estimados como equivalentes al tiempo en que se cumplen ciertos kilometrajes registrados a partir de la fecha de compra de un vehículo o su última mantención. Por esto, se genera contacto con el cliente luego de 15 días de la compra para que asista a una mantención preventiva gratuita a modo de control general del vehículo. Posterior a esto, como se desconoce exactamente cuándo un vehículo cumplirá con los 15.000 kilómetros recomendados para asistir a la mantención, o si decidirá ir al año cumplido, se le contacta cuando se cumplen 6, 9, 12 15 meses desde la compra, para invitarlo a asistir a la mantención correspondiente a los 15.000 kilómetros. Se tiene una metodología de contacto en forma de árbol de decisión tal como se detalla en el Anexo A.3, que corresponde a un extracto de la planificación de contacto.

Cuando un cliente no asiste a las mantenciones luego de los contactos realizados, luego de 15 meses cumplidos desde la compra o desde una fecha en que se ha realizado una mantención, se ofrece un “gancho”, que corresponde a una promoción para motivarlo a asistir a la mantención. Luego de esto, si el cliente no asiste, se da por inasistente a la mantención en cuestión, y se continúa con los recordatorios para asistir a las próximas mantenciones correspondientes.

Además, los distintos concesionarios que realizan la venta de vehículos, poseen la información de quienes han comprado vehículos en sus puntos de venta, por lo que tienen la opción de contactarlos también para invitarlos a asistir a las mantenciones, aunque se tiene noción de que esto no ocurre de forma usual. Aun así, en ciertas ocasiones el área de Post Venta de la empresa entrega una nómina a cada concesionario con los dueños de vehículo a quienes correspondería contactar para ofrecerles una promoción o incentivarlos a asistir al servicio.

Además, fuera de las promociones de “gancho” que se ofrecen para incentivar como última instancia, de forma usual se generan promociones y descuentos para el público en general.

Aún así, cabe destacar que actualmente no se tiene registro de qué personas han utilizado las promociones recibidas al asistir al servicio, ya que esto no se registra en las bases de datos de cada concesionario. Es debido a esto, que el área de post venta desconoce cuál es el efecto que tienen sus promociones en cada usuario de la marca, y por tanto, esto podría desencadenar en pérdidas monetarias para la empresa por ofrecer promociones a usuarios de la marca que con alta probabilidad asistirán al servicio de todas formas, y a la vez, no ofrecer promociones que motiven a asistir a los clientes que aumentarán su probabilidad de asistencia dada esta. Aún así, al segundo semestre del año 2021 se está trabajando en la implementación de una aplicación para teléfono móvil “MyKia Chile” donde se podría llevar registro de uso de promociones y asistencia a los servicios de post-venta de parte de los clientes de la marca, permitiendo recolectar información demográfica de los clientes que hoy no se posee.

A la vez, se tiene interés por saber qué tan probable es que una persona asista al servicio, para focalizar esfuerzos en quienes podrían aumentar su probabilidad de asistencia dada alguna promoción o contacto.

2.2. Identificación del problema

La empresa, al enfocar la mayoría de su negocio en la venta de un bien duradero como es un automóvil, debe mantenerse vigente importando los nuevos modelos de automóvil al mercado chileno para capturar el interés de las personas y clientes de la marca, y así lograr seguir realizando ventas de sus productos.

Con este fin, la empresa hace esfuerzos en capturar la atención de personas que no han sido clientes de la marca realizando publicidad mediante redes sociales, página web, radio, entre otros. Sin embargo, cuando una persona ya es cliente de la marca, la situación

cambia y es necesario generar contacto directo con ellos para mantenerlos interesados en la marca.

Con la idea de no generar tanto desgaste en contactar a los clientes de forma tan frecuente para asistir a los servicios de mantención, el área Comercial tiene interés por tener mayor claridad en cuándo y a quienes contactar para recordar y/o ofrecer promociones. Se considera que esto podría traer consecuencias positivas en los ingresos de la empresa por el posible aumento de demanda de los servicios y también un aumento de la recompra de vehículos en la misma marca. de forma que esto genere un aumento en la demanda de los servicio. Actualmente, la estimación de fechas de mantención se realiza por el período de tiempo pasado entre la venta y el tiempo promedio en que se estima que se cumple el kilometraje de pauta de cada mantención (cada 1 año en general). El problema recae en que los clientes pueden optar por asistir cuando cumplen el kilometraje recomendado para cada mantención pauta en lugar de las fechas, por lo que asistirá en una fecha distinta a la fecha estimada, lo que la empresa considera como *pérdida* de las personas que no se les logra contactar de forma cercana a la fecha en que efectivamente les correspondería asistir por kilometraje.

Como forma de fidelización, la empresa considera sumamente importante el entregar un buen servicio de post-venta, ya que estos servicios posteriores a la venta son la mayor fuente de contacto con el cliente, y si el cliente queda satisfecho, se considera que el cliente tendrá en mayor consideración a la marca para realizar una re-compra a futuro.

Se considera que estas instancias de post-venta son una oportunidad para mantener contento al cliente con la marca, y considerando que se recomienda realizar mantenciones vehiculares cada cierto tiempo para mantener funcionando un vehículo de forma óptima, el área Comercial de post-venta los últimos años se ha encargado de contactar a los usuarios de la marca para darles aviso cuando les corresponde asistir a realizar una mantención vehicular.

Es por esto, que en el área de Business Intelligence nace la iniciativa de poder mejorar las estimaciones de tanto de fechas para generar contacto con los usuarios de vehículos como la segmentación de clientes mediante la ciencia de datos. La segmentación de clientes sería útil para diferenciar a qué clientes se les debe ofrecer descuentos o promociones al momento de avisarles que les corresponde a asistir a una mantención para que efectivamente asistan, de los clientes que solo basta con avisarles que les corresponde asistir a mantención para que lo hagan.

Los servicios de mantención, como se mencionó anteriormente, corresponden a un servicio de post venta de la venta de un automóvil. Las mantenciones si bien no es una obligación realizarlas, sí es necesario mantener un vehículo en buen estado año a año para que pueda circular por el país, debido a que debe pasar los exámenes de revisión técnica de forma obligatoria para lograr obtener el permiso de circulación[5].

Además, es conveniente para el dueño del vehículo estar al día con las mantenciones, ya que puede evitar mayores costos a futuro[4]; por ejemplo, las fallas en el motor de un vehículo puede ser consecuencia de no realizar cambios de aceite, y el cambio de aceite

a largo plazo es más económico que el cambio de un motor. Por otro lado, año a año, alrededor de 1.300 accidentes automovilísticos son causados por fallas mecánicas en los vehículos, correspondiente al 2 % de los accidentes anuales (CONASET, 2019).

En los servicios de mantención de Kia, se definen las tasas de retención como *cantidad de mantenciones realizadas luego de cumplido el período en que correspondería asistir a una mantención desde la venta*, que es de 1 mes para la primera mantención, 1 año para la segunda mantención, 2 años para la tercera mantención y 3 años para la cuarta mantención; por sobre la *cantidad de ventas de vehículos correspondientes a la fecha de inicio considerada*, como por ejemplo

$$\text{Tasa retención} = \frac{\text{N}^{\text{o}} \text{ asistencias a segunda mantención registradas en mes de febrero de 2021}}{\text{Cantidad de vehículos vendidos en febrero de 2020}}$$

Los últimos años, las tasas de retención según la data con la que se cuenta son las siguientes:

Tabla 2.1: Tasas de retención de clientes en el servicio de mantención vehicular de Kia

Año	1era mantención	2da mantención	3era mantención	4ta mantención
2017	55 %			
2018	64 %	32 %		
2019	71 %	34 %	19 %	
2020	52 %	29 %	15 %	3 %

Como se puede ver, la tasa de retención de la primera revisión a los 1.000 kilómetros que corresponde a un chequeo general gratuito para comprobar que el vehículo funciona bien y no tenga fallas de fábrica, posee el mayor porcentaje de asistencia, lo que se justifica porque según los clientes de la marca *es conveniente asistir por ser gratuito*, según una encuesta realizada por Kia Chile a sus clientes. Aún así, la asistencia ha ido a la baja los últimos dos años debido al cierre temporal de sucursales provocado por cuarentenas debido a la pandemia. En las siguientes mantenciones se puede ver que la asistencia va en disminución a medida que aumenta el kilometraje por el que se debe asistir. Además, el área considera como máximo la asistencia al servicio de mantención de 60.000 kilómetros, puesto que se tiene la noción en base a datos que luego de 3 años de uso una persona suele cambiar su vehículo por uno nuevo.

Por otra parte, uno de los factores que podría estar afectando la retención del servicio son los precios. Los precios de las mantenciones no son equivalentes en todos los concesionarios; sino que estos son fijados por cada uno de los concesionarios en cuestión. Kia Chile realiza una recomendación de precios al consumidor basados en los gastos de insumos y mano de obra que conllevan por cada modelo y tipo de mantención, pero no puede imponer un precio fijo para todos, debido a que pasaría a llevar la libre competencia cometiendo el delito de colusión[6].

Es por esto que Kia tiene interés por aumentar la asistencia al servicio de mantención mediante una mejora en la personalización de ofertas y contacto con los usuarios de la

marca en el momento preciso. Esto se alinea con la misión de la empresa, ya que con esto se busca entregar un servicio de alto nivel a los clientes de la empresa.

2.3. Impacto del cambio propuesto

Las ganancias por mantención para Kia dependen del modelo y tipo de mantención; por ejemplo, el cobro base recomendado por la mantención de 15.000 kilómetros (segunda mantención) para el modelo *Morning*, corresponde a $188.000 (cada concesionario añade un margen extra de cobro al cliente para obtener ganancias) y KIA margina aproximadamente un 10 % de esto (el resto corresponde a gasto de implementos utilizados), siendo este margen un ingreso “bajo” para la empresa. A pesar de esto, la oportunidad radica en que si se mantiene al cliente interesado en los servicios de la marca, este aumentará la posibilidad de realizar una recompra en Kia.

Las utilidades que podría entregar la recompra de vehículos por fidelización, se podrían calcular considerando que la tasa de retención anual de las mantenciones correspondientes a la cuarta mantención en el año 2020 fue de aproximadamente un 3 % de un total de 32.957 vehículos. Se tiene la noción (según la empresa) de que un 50 % de los asistentes a la cuarta mantención harán recompra de un vehículo en la marca, por tanto, si se logra aumentar en un 1 % (por dar un ejemplo) la tasa de retención de la cuarta mantención, y tomando las cifras de 2020, se aumentaría en 330 vehículos asistes a la cuarta mantención, y por ende, 165 recompradores de vehículos de la marca. Con esto, considerando que el vehículo más económico y más demandado de la marca tiene un valor de venta base de $8.190.000, del que la empresa margina un 10 %, se tendrían un aumento de utilidades por recompra de 165*$819.000 = $135.135.000 de forma anual.

Capítulo 3

Objetivos

3.1. Objetivo General

Identificación de perfiles de individuos más propensos a asistir a servicios de mantención vehicular y sus respectivos períodos de cumplimiento de pautas de mantención, con el fin de tener una base de información para aumentar la demanda del servicio mediante técnicas de marketing de parte de la empresa.

3.2. Objetivos Específicos

1. Identificar y seleccionar los datos relevantes relacionados a características de un vehículo y al conductor del mismo que permitan determinar cuándo cumplirá con los kilometrajes en que se realizan mantenciones y que tan probable es que asista a los servicios de mantención.

2. Evaluar distintos modelos de predicción de fechas en que se cumplen los kilometrajes que corresponden a mantenciones vehiculares y de predicción de propensión de asistencia a estos servicios, y determinar cuál es el de mejor desempeño.

3. Entregar recomendaciones a partir del trabajo realizado y propuestas de trabajo a futuro.

Capítulo 4

Alcances

Este trabajo tiene por fin entregar una estimación de cuándo asistirá una persona a una mantención y qué clientes serán propensos a asistir. Esto se realiza para la segunda, tercera y cuarta mantención de pauta, a los que la empresa recomienda asistir una vez cumplidos los 15.000 kilómetros o 1 año desde la compra, 30.000 kilómetros o 2 años desde la venta, y 45.000 kilómetros o 3 años desde la venta, respectivamente.

El enfoque del estudio está orientado a clientes correspondientes a personas naturales que han comprado un auto nuevo (*"primera mano"*) de la marca. De esta forma se excluyen clientes empresas y clientes de segunda mano, a la vez que se excluye del estudio el modelo de automóvil comercial "Frontier".

Los modelos a aplicar para el desarrollo del trabajo se diseñarán sólo con la información disponible en la empresa. Actualmente, la empresa posee escasa información de sus clientes debido a que solo desde 2016 se comenzó a digitalizar las bases de datos de ventas y servicios, y además no existía mayor recopilación de información demográfica asociada a sus clientes. Por esto, algunos modelos aplicados en el desarrollo excluyen ciertos modelos de autos, debido a la escasa variabilidad de los mismos por baja cantidad de observaciones.

El modelo de predicción de fecha de asistencia entregará una cifra en días que se sumará a la última fecha de asistencia a mantención o de venta, en caso de que correspondiese.

El modelo de propensión de asistencia entregará como resultado una respuesta binaria: "asistirá" o "no asistirá" a la mantención que corresponda.

Este trabajo no aborda los problemas desde el punto de vista operacional debido a que la empresa posee limitaciones en este aspecto, puesto que los servicios prestados tanto como los precios impuestos en general son manejados por los concesionarios, y la empresa solo les entrega recomendaciones. Aún así, la empresa puede realizar promociones o descuentos a los clientes haciéndose cargo de los costos asociados.

Capítulo 5

Resultados esperados

Como resultado del trabajo de título se espera obtener en primera instancia 2 modelos: un Modelo de predicción de días hasta la próxima mantención que permita determinar en qué fecha le correspondería asistir a servicio de mantención a un vehículo registrado de la marca, y otro modelo de propensión de asistencia del público a los servicios de mantención autorizados por Kia, que segmenten entre asistentes y no asistentes a la mantención. De forma más específica, se pretende aplicar estos modelos para la segunda, tercera y cuarta mantención pauta.

Se espera que los resultados de estos modelos sean lo suficiente precisos y acertados para implementarlos en la empresa, y en el caso del modelo de predicción de fecha de asistencia, que este obtenga resultados con un error más bajo que los obtenidos actualmente por la empresa. A partir de los algoritmos aplicados, se espera determinar cuáles son los que mejor desempeño presentan y a partir de esto, poder recomendarlos a la empresa en estudio, y también realizar recomendaciones sobre estos modelos y sobre trabajo futuro a realizar que sería conveniente desempeñar en la empresa.

Capítulo 6

Marco conceptual

Para el entendimiento del trabajo a describir en este informe, a continuación se presenta el marco teórico y una descripción de los algoritmos y modelos a utilizar en el desarrollo del trabajo de título.

Los modelos predictivos son modelos que buscan obtener relaciones de datos variables con respecto a una variable de interés, lo que permite predecir comportamientos futuros dadas las condiciones de las variables relacionadas. Este tipo de modelos son utilizados en diversos ámbitos: en medicina ha sido aplicado para detectar diagnósticos de cáncer [7], hasta en empresas financieras, intentando predecir patrones de fugas de clientes[8].

A continuación se presentan los modelos de regresión utilizados en el desarrollo de este trabajo.

6.1. Modelo de regresión lineal

Los modelos de regresión son modelos matemáticos que describen las relaciones entre una variable dependiente que es lo que buscamos predecir, y variables independientes o explicativas, que son las características o atributos en los que se basará el modelo para entregar una predicción.

Para el caso de la predicción de kilometraje o de días hasta que se cumpla cierto kilometraje, se puede utilizar una regresión lineal con el fin de regresionar como variable dependiente el kilometraje o los días hasta cumplir el kilometraje esperado, utilizando variables independientes que permitan caracterizar a cada individuo.

Este tipo de modelo, utilizando mínimos cuadrados ordinarios (MCO) ha sido usado en otra ocasión por Li & Kockelman (2019) para la predicción de distancia recorrida diaria por vehículos en Washington, utilizando historial de más de un año de registros obtenidos por medio de GPS's.

En el desarrollo de un modelo de regresión lineal, las variables independientes utilizadas deben cumplir con cuatro supuestos básicos para que los resultados sean confiables:

1. Normalidad: Variables siguen la Ley Normal.

2. Independencia: Errores en medición de variables explicativas independientes entre sí.

3. No colinealidad: Correlación entre variables independientes nula.

4. Homocedasticidad: Errores de la variable dependiente tengan varianza constante.

5. Linealidad: Relación entre las variables lineal.

6.2. Árboles de clasificación y regresión

Los árboles de decisión son una técnica de aprendizaje no paramétrica basada en la categorización de variables, cuyo nombre proviene de su similitud con la estructura de un árbol y se basan en la clasificación de los datos de acuerdo a los valores que toman las variables observadas. Su estructura se compone de nodos y ramas que forman el árbol, y la terminología se define como sigue:

- Nodo de decisión: Nodo en que es necesario tomar una decisión para continuar con un proceso. Este se identifica con un cuadrado.
- Nodo de probabilidad: Nodo en que ocurre un evento aleatorio en el proceso. Es la probabilidad de que ocurra un evento probable como resultado de una decisión. Se suele representar por un círculo.
- Nodo terminal: Nodo en que todos los casos posible tienen el mismo valor para la variable dependiente. Es homogéneo, no requiere ninguna división adicional porque es “puro”.
- Rama: Son las conexiones que muestran los distintos caminos que se pueden seguir al tomar una decisión o bien ocurrir algún evento aleatorio. Serían el resultado de las posibles interacciones entre las alternativas de decisión y los eventos.

Este modelo permite realizar tanto clasificación como regresión (como lo dice su nombre) según se requiera dependiendo de si la variable de interés (dependiente) es categórica o numérica, respectivamente. Para lograr implementar un árbol de decisión existen diversos algoritmos de división, entre los que están principalmente CHAID (*Chi-square automatic interaction detector*), CHAID exhaustivo, Árboles de clasificación y regresión (*CART-Classificacion and regression trees*), QUEST (*Quick, unbiased, efficient, statical tree*), C5.0, entre otros. De estos, se utilizará CART que permite realizar regresiones con respecto a una variable continua.[23]

El algoritmo CART el más usado en la actualidad para implementar árboles de decisión. Esto se debe a que posee como ventajas el ser robusto frente a outliers, tener invarianza en la estructura de sus árboles en caso de existir transformaciones monótonas de las variables independientes, y que posee alta interpretabilidad.
El método que utiliza CART consiste en 3 principales pasos. En primer lugar, construir un árbol saturado, es decir, con todas las ramificaciones necesarias posibles. En segundo lugar se determina el tamaño que debe tener el árbol, y por último, se clasifican los nuevos datos a partir del árbol que se ha construido.

En ocasiones, los árboles de decisión tienden crear gran cantidad de ramificaciones, por lo que suele ocurrir un sobreajuste de los modelos. Para evitar esto, se realiza *poda* del árbol, donde se cortan ciertas ramas del árbol, poniendo cotas que permiten disminuir el número de divisiones del árbol, permitiendo que cada nodo obtenga un valor mayor de observaciones.

6.3. Random Forest (Bosques Aleatorios)

Random Forest es un algoritmo de aprendizaje supervisado de tipo bagging, cuya estructura se basa en el desarrollo de una gran cantidad de modelos débiles de los que se obtienen finalmente una predicción

que se compone de un conjunto de árboles de decisión, por lo que corresponde a los llamados métodos de ensemble[1]. Este algoritmo puede ser usado tanto para regresión o clasificación. Se suele considerar un algoritmo de mejor desempeño que los árboles de decisión porque permite reducir el sobreajuste que se da en los últimos.

Este algoritmo se basa en la creación de un conjunto de entrenamiento y un conjunto de testeo a partir de la base de datos original, desde los que se desarrolla un árbol de decisión. Esto lo realiza reiteradas veces tomando distintos valores de la muestra para entrenamiento y testeo, aumentando así la variabilidad de los datos. A partir de los resultados de todos los árboles de decisión realizados, se determina la clase más "popular" para cada registro mediante *votos* obtenidos de cada árbol, a partir de los que se obtiene la clasificación de cada registro.

Este algoritmo busca reducir la varianza por medio de la creación de diversos árboles profundos y ramificados. Los hiperparámetros que se pueden manejar usualmente en el diseño de este algoritmo son los siguientes:

- Cp: Parámetro de complejidad α de la poda del árbol, donde el valor 1 es un árbol sin divisiones y 0 es un árbol de profundidad máxima. El algoritmo evita hacer nuevas divisiones en caso de que la proporción de reducción del error sea inferior al valor Cp.
- Maxdepth: Es la profundidad máxima del árbol (cuántos nodos consecutivos máximo posea el árbol).
- Minsplit y Minbucket: mínimo de observaciones en un nodo intermedio y en uno terminal para particionarlo, respectivamente.
- xval: Número de grupos (folds) para validación cruzada.

El modelo de Random Forest ha sido utilizado en la predicción de abandono en la demanda de cerveza de parte de locales de venta a un distribuidor, pertenecientes a una región de España, el que fue comparado frente a otros algoritmos como redes neuronales, XGBoost y Regresión logística, siendo el que ha presentado mejor desempeño.

[1] Métodos que combinan múltiples modelos en uno para lograr obtener un equilibrio entre el sesgo y varianza, buscando mejorar las predicciones de cualquier modelo individual.

6.4. Máquinas de soporte vectorial (SVM)

El modelo de Máquinas de Soporte Vectorial (Support Vector Machines) es una técnica de aprendizaje supervisado útil para modelar problemas de clasificación y regresión.

La formulación del algoritmo para problemas de clasificación mapean los datos en un espacio alto-dimensional de características, donde se genera un hiperplano de separación. El hiperplano se genera a partir de la maximización de la distancia de los patrones más cercanos, maximizando así el margen. En los casos en que los hiperplanos están lejos de las fronteras de las clases de los objetos existen márgenes de separación mayores. Cuando los hiperplanos aciertan en la asignación de objetos a las clases correspondientes realmente existe un menor error de clasificación. Por tanto, para encontrar el óptimo se debe maximizar el margen de separación y minimizar el error de clasificación, lo que en ocasiones puede ser complicado de cumplir en paralelo.

Si se busca encontrar el óptimo hiperplano en un problema de dos clases, se propone un problema de minimización cuadrático convexo de optimización, el que generaría el *hiperplano óptimo de separación* (HOS). En los casos en que no es posible la separación lineal de clases, se realiza aplicando funciones *kernel* o núcleo que transforma de forma implícita el espacio de entrada en un espacio de características de alta dimensión, o aplicando variables de pérdida o slacks.[8]

El kernel o núcleo corresponde a una función que evalúa el producto interno entre dos puntos, aún cuando se desconoce el mapeo de los puntos. Algunas de las funciones de kernel se muestran en la Figura 6.1.

$\theta(u)$	$K(u,v)$
Degree d polynomial	$(u \cdot v + 1)^d$
Radial Basis Function Machine	$\exp -\frac{\|u-v\|^2}{2\sigma}$
Two-Layer Neural Network	$sigmoid(\eta(u \cdot v) + c)$

Figura 6.1: Funciones de kernel más utilizadas para *SVM*. Fuente: [9]

Este algoritmo ha sido utilizado para predicción de fugas de clientes de una institución financiera como problema de clasificación, mostrando superioridad frente a un modelo basado en redes neuronales (Weber, R. et al. 2005)[8]

Las SVM entregan una habilidad alta de generalización, por lo que son una herramienta robusta tanto para regresión como clasificación en ambientes de datos complejos y ruidosos, pudiendo extraer información relevante para construir algoritmos rápidos para grandes cantidades de datos. Como desventaja, este tipo de algoritmo requiere de gran capacidad computacional y los resultados podrían ser de difícil interpretación.

Como algoritmo de regresión tienen la ventaja de ser muy flexibles y son relativamente

robustas frente a valores atípicos. Como desventaja, al ser un modelo

6.5. Redes neuronales artificiales

Las redes neuronales artificiales (RNA) son un sistema de procesamiento de datos que se basa en las redes neuronales biológicas para su estructura tanto como su operación. Estos modelos corresponden a modelos de aprendizaje de máquinas e inteligencia artificial, es decir, es un modelo que logra aprender por sí solo a tarea que se impone. Para esto, al igual que en otros modelos, es necesario entregarle un porcentaje de los datos como set de entrenamiento (generalmente un 80 % de los datos) para que pueda descubrir patrones o relaciones entre los datos, y luego se le entrega un set de testeo, donde se evalúa el modelo generado por las redes y se compara con el resultado real del set de datos. Este modelo permite trabajar con variables continuas o discretas.[12]

Las *neuronas* son elementos sencillos que se encargan de procesar la información, y transmiten señales mediante *conexiones* o *enlaces de comunicación*. Estas conexiones tienen un peso asociado, el que permite a una neurona que recibe el impulso de la anterior, ponderar la señal recibida y mediante una función de activación entrega un flujo de salida, el que actúa como flujo de entrada a otra neurona y resto de la red.

Las neuronas se agrupan en capas, existiendo capas de entrada, capas ocultas y capas de salida. La capa de entrada posee neuronas que envían datos a una segunda capa, siendo nodos pasivos que simplemente transmiten información.

La *segunda capa* o *capa oculta* realiza un filtro de los patrones relevantes que contendrían la información de importancia, dejando fuera los que no lo son. Para esto, los nodos de esta capa combinan los datos de entrada, lo que realizan multiplicándolos por el peso asociado a la conexión, y luego estos son delimitados por una *función de activación* que suele ser sigmoide o logística que mejora la eficiencia.

La *capa de salida* o *tercera capa* repite lo realizado por la segunda capa y vuelve a combinar los datos ponderados por los respectivos pesos, obteniendo así valores de salida. Esta estructura se puede observar en la Figura 6.2. Para determinar cuándo se detendrá el modelo en el entrenamiento se define una condición de detención, la que normalmente es que el cálculo del error cuadrado sobre todos los ejemplos alcance el mínimo o cuando el error observado sea inferior a un umbral determinado[11].

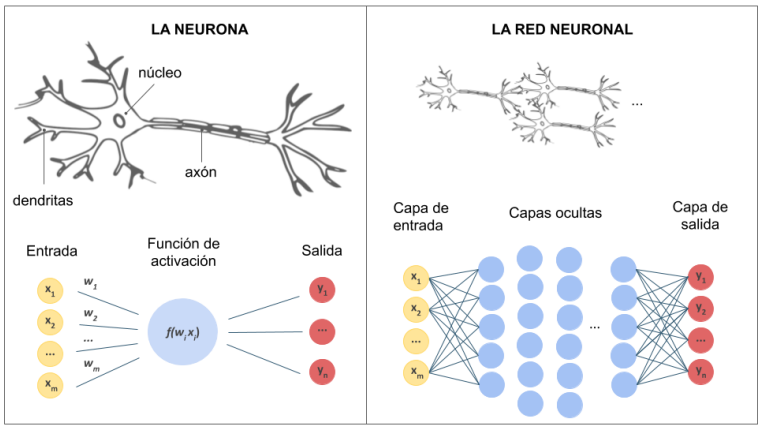


Figura 6.2: Estructura de red neuronal artificial.[10]

El algoritmo de Back-propagation es el mayormente utilizado para entrenar redes neuronales. Este consiste en un algoritmo de retroalimentación, donde al comienzo del entreno se asignan valores de peso de las conexiones neuronales aleatorios y pequeños, los que dan como resultado un out-put final, el que se compara con los valores reales mediante la diferencia cuadrática. Mientras mas pequeño sea este valor, mejor es la predicción. Por tanto, el algoritmo reitera el proceso hasta cientos de veces, hasta que el error deja de cambiar y no disminuye más. [13] Las redes neuronales artificiales destacan por la capacidad que tienen para identificar patrones y dar sentido a datos incompletos, ambiguos y que en ocasiones pueden ser hasta contradictorios, lo que ha llegado a aumentar el uso de este tipo de modelos.

Existen diversos trabajos que utilizan redes neuronales para realizar modelado, por ejemplo, para pronosticar las variaciones del nivel de las mareas en la desembocadura del río Chao Phraya en Tailandi por Supratid, S.(2003) [14], [16]

6.6. Potenciación del Gradiente (GBM)

Potenciación del Gradiente (más conocido en inglés como *Gradient Boost Machines*) es una generalización del modelo Boosting Machine (modelo de combinación de múltiples modelos sencillos, *ensemble*) que permite aplicar el método de descenso de gradiente para optimizar cualquier función de coste durante el ajuste del modelo. Los modelos sencillos en los que se basa GBM se suelen llamar weak learners, los que generalmente son árboles de decisión.

El valor que predice un modelo GBM es la agregación de lo que predicen los modelos individuales que forman la combinación de modelos o *ensemble*. En problemas de clasificación, este valor predicho es la moda, mientras que en modelos de regresión es la media.

Al igual que en RNA's, es posible utilizar *Early Stopping* (Detención temprana del entrenamiento) con el fin de evitar el sobreajuste que puede causar un alto número de árboles utilizados para el algoritmo. Para esto, es necesario encontrar el número óptimo de árboles que eviten el sobre-ajuste, lo que se logra aplicando ciertos parámetros límites frente a los que se debe detener el algoritmo. Por ejemplo, se puede indicar como parámetro de

detención el porcentaje mínimo de mejora entre dos iteraciones consecutivas o cuando se supere un tiempo máximo de ejecución de una iteración.

Además de *Early Stopping*, el algoritmo utiliza una muestra aleatoria de observaciones para el entrenamiento de cada árbol, es decir, tiene un *comportamiento estocástico*, lo que ayuda a evitar el sobreajuste del modelo.

Este algoritmo, a diferencia de Bosques Aleatorios, busca reducir el error a través de disminuir el sesgo, creando árboles débiles y poco profundos. Este modelo recientemente ha sido utilizado en variados estudios; buscando predecir la quiebra de entidades bancarias europeas, Momparler, Carmona, y Climent (2016) obtuvieron resultados bastante favorables con esta técnica [17].

6.7. Métricas de desempeño

Para los problemas que se intentan resolver con este trabajo, se utilizarán métricas de desempeño tanto para el problema de regresión como de clasificación. Para los problemas de regresión se utilizan métricas como error cuadrático medio, R-cuadrado, F1, entre otras, mientras que para los problemas de clasificación se acostumbra a utilizar métricas como Accuracy, Sensibilidad, precisión entre otros. A continuación se detallan las consideradas más importantes para este trabajo.

6.7.1. Métricas de desempeño modelo de clasificación

Las métricas de desempeño de un modelo de clasificación se basan en los valores de la matriz de confusión:

		Real	
		0	1
Predicho	0	TN	FN
	1	FP	TP

En esta, se indican la cantidad de casos predichos con respecto a los reales, tal que:

1. Verdaderos positivos (“True Positives (TP”): Casos predichos positivos que son positivos en realidad.
2. Falsos positivos (“False Positives (FP”): Casos predichos positivos que son negativos en realidad.
3. Verdaderos negativos (“True Negative (TF”): Casos predichos negativos que son negativos en realidad.
4. Falsos negativos (“False Negative (FN”): Casos predichos negativos que son positivos en realidad.

A partir de estos, se pueden calcular las diversas métricas.

- Precisión (*Accuracy*): Casos que fueron correctamente predichos sobre el total de casos.

$$\frac{TP + TF}{TP + TF + FP + FN}$$

- Sensibilidad (*Recall*): Casos predichos como positivos del total de positivos (reales).

$$\frac{TP}{TP + FN}$$

- Precisión (*Precision*): Casos predichos correctamente como positivos del total de casos predichos como positivos.

$$\frac{TP}{TP + FP}$$

- Especificidad (*Specifity*): Casos predichos como negativos del total de negativos.

$$\frac{TN}{TN + FP}$$

- Puntuación F1: Es la media armónica entre Recall y Precisión.

$$\frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$

En el problema de clasificación que se presenta en este trabajo, existe mayor interés por saber qué cliente tendrá mayor tendencia a asistir al servicio de mantención puesto que así se podría aumentar o asegurar la asistencia misma. Es por esto, que la métrica de mayor interés es la sensibilidad, es decir, se determinará el mejor modelo en base a la mayor sensibilidad presentada.

6.7.2. Métricas de desempeño modelos de regresión

Raíz del Error Cuadrático Medio (RMSE)

La raíz del error cuadrático medio o *root mean squared error (RMSE)* 6.1 es una métrica utilizada para comparar problemas de regresión, y mide la raíz del promedio de los errores al cuadrado. El valor resultante tiene la misma unidad que los datos ocupados, lo que permite percibir la magnitud del error con respecto a los datos reales de forma sencilla.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \tag{6.1}$$

Error Absoluto Medio (MAE)

El error absoluto medio o *mean absolute error (MAE)* 6.2 es una métrica utilizada para comparar problemas de regresión, y mide el promedio de la diferencia absoluta entre los valores predichos frente a los valores reales, permitiendo obtener una mejor noción del error general existente al ser absoluto, y también al encontrarse en la misma escala que

los datos estudiados.

$$MAE = \frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{n} \tag{6.2}$$

Error Porcentual Absoluto Medio (MAPE)

El error porcentual absoluto medio o *porcentual mean absolute error (MAPE)* 6.3 es una métrica utilizada para comparar problemas de regresión, y mide el valor absoluto del promedio del porcentaje de la diferencia absoluta entre los valores predichos y los valores reales, del valor real, es decir, el promedio del error del porcentaje del valor real. Este permite saber tener una noción del cuán precisas son las predicciones.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|Y_i - \hat{Y}_i|}{Y_i} \tag{6.3}$$

6.8. Correlación

Existen distintos coeficientes de correlación, que son una medida que permite determinar el grado de asociación entre dos variables. En este trabajo se utiliza la correlación de Pearson, que es independiente de la escala de medida de las variables numéricas y permite obtener la relación lineal entre estas. El coeficiente de Pearson se calcula según la ecuación 6.4. Este puede tomar valores entre -1 y 1, donde coeficientes cercanos a 1 o -1 representan alta correlación, positiva y negativa respectivamente, y valores cercanos a 0 indican baja correlación. En la ecuación, x e y son los valores reales de las variables en estudio, σ la covarianza y ρ es el coeficiente de Pearson.

$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y} \tag{6.4}$$

También se utiliza la medida de correlación mediante el coeficiente V de Cramer, que mide el grado de correlación entre variables categóricas. Este se basa en la fórmula 6.5, que V es el coeficiente de Cramer, χ^2 es el estadístico *Chi-cuadrado*, N es el número de observaciones y m es definido como *min(f-1,c-1)*, es decir, el menor valor de “filas - 1” y “columnas - 1”.

$$V = \sqrt{\frac{\chi^2}{Nm}} \tag{6.5}$$

Capítulo 7

Metodología

Para el desarrollo del trabajo se utiliza una metodología de tipo CRISP-DM[21], ya que es menos complejo que una metodología de tipo KDD, y posee una pauta de trabajo más elaborada (ver Figura B.1) .

La metodología se compone de las siguientes etapas:

1. Comprensión del negocio: Se realizan entrevistas a trabajadores de la empresa con el fin de conocer la situación actual de la misma y cuál es el problema u oportunidad a abordar. A través de esto, se realizan entrevistas a distintas áreas de la empresa, detectando que el problema son la baja asistencia a los servicios de mantención, y se conversa con distintas entidades de la empresa con el fin de comprender de dónde se puede obtener información y datos valiosos para realizar el trabajo.

2. Conocimiento de datos: La empresa cuenta con diversas bases de datos en que existe registro del parte del parque automotriz, asistencia a servicios de post-venta, información demográfica de los clientes, evaluación de los servicios por parte de los usuarios o dueños de vehículos y contacto realizado de parte de la marca a los clientes. La base de datos de registro de asistencia a los servicios de mantención provienen de la unión de las bases de datos de los distintos concesionarios autorizados por Kia, cuyos datos son ingresados de forma manual por los vendedores, por tanto, estos datos registran bastantes datos incorrectos, mal ingresados, incoherentes o vacíos. De la mano con esto, la información demográfica registrada en estos concesionarios cuenta con datos vacíos y posee poca información relevante. Es por esto que se hace necesario para la empresa comprar información relacionada a los dueños de vehículos de la marca, con el fin de lograr una identificación más clara de los sujetos de estudio.

3. Preparación de los datos: Para el trabajo con los datos, se adecuan las variables dependiendo del modelo que se utilizará. En general, para la predicción de kilometraje de asistencia, la base de registros de datos se dividirá en bases según el tipo de mantención pauta que se desea estudiar y luego según asistencia previa; por ejemplo, para el modelo enfocado en la tercera mantención, se utilizará una base de datos con todos los vehículos que posean registro de asistencia a la tercera mantención y esta se dividirá en sub-sets según asistencia a las mantenciones previas. Para el caso de la propensión de asistencia, se crearán variables de asistencia a una mantención además de la variable

recency correspondiente a la frescura.

Se realiza un estudio de las variables mismas en los modelos, por medio de clasificación de variables a través de estudio de correlación.

4. Modelamiento: Con los datos preparados, se crearán modelos utilizando Rstudio que permitan predecir los días hasta la asistencia a la próxima mantención de cada vehículo a través de modelos de regresión y machine learning. Posteriormente, se diseñarán modelos que permitan clasificar a los clientes entre propensos a asistir y no propensos a asistir a una mantención, para lo que se utilizarán modelos como Máquinas de soporte vectorial, Random Forest, entre otros modelos de clasificación.

5. Evaluación de modelos:Tanto los modelos generados para la predicción de asistencia como para propensión de asistencia, se evaluarán en base a las métricas de desempeño indicadas en el marco conceptual. Para los modelos de fecha de predicción de asistencia se utilizarán las medidas de error, mientras que para la propensión de asistencia se considerarán las medidas de desempeño de modelos de clasificación. En ambos casos, la evaluación de modelos se aplicará sobre parte de los datos que se considerarán como datos de prueba o *testeo* de los modelos generados.

6. Implementación: Como trabajo futuro queda el implementar los mejores modelos de para cada problema en los programas de la empresa, pero este punto queda fuera del alcance de este trabajo de memoria.

Capítulo 8

Desarrollo Metodológico

8.1. Conocimiento de los datos

La base de datos utilizada en este trabajo está conformada por la integración de diversas bases de datos de las que dispone la empresa, cuya información proviene en su mayoría de los concesionarios asociados como se mencionó anteriormente. A causa de esto, existe mucha información duplicada en las bases de datos, y también existe información incoherente, lo que se reduce al máximo con la limpieza de datos posterior. Las bases existentes se pueden clasificar a grandes rasgos por:

1. Datos de ventas de vehículos: Registros de fechas de venta y modelo del automóvil.
2. Registros de órdenes de trabajos de post venta: Registros de fecha de asistencia, tipo de trabajo realizado, kilometraje del vehículo registrado. Esta base cuenta con un total de 617.056 registros de trabajo.
3. Datos de patentes vehiculares e información demográfica del cliente: Base de datos perteneciente al Registro Civil, donde se indica el VIN (código identificador) de un vehículo asociado a un RUT del dueño de este, el año de registro y modelo del auto.
4. Información de nivel socio-económico: Obtenida de la empresa financiera asociada, corresponde al grupo socio-económico al que pertenece el dueño del vehículo.
5. Registro de la información técnica de cada modelo de vehículo: Contiene datos sobre los modelos de automóvil como peso, rendimiento, especificaciones, entre otros.

Para unificar estas tablas en una base de datos se juntan las observaciones según el VIN del automóvil (llave primaria) y el RUT del dueño del automóvil (llave secundaria), teniendo una fila por cada VIN existente que ha asistido a alguna mantención, comenzando así con 33.846 datos. La base de datos, posee registros de las mantenciones a las que ha asistido el cliente, las que van desde la primera a la décima mantención (o mayor), siendo cada mantención cada 15.000 kilómetros o cada un año extra de posesión del automóvil desde la compra del mismo. Para el trabajo con modelos de predicción se consideran solo los registros hasta la cuarta mantención, ignorando si existen o no registros posteriores, con el fin de lograr obtener predicciones para las primeras cuatro mantenciones vehiculares, que son las mantenciones pauta que mayor asistencia presentan en general en la empresa, por lo que el interés del estudio se enfoca en ellas. Las variables que contiene

el set de datos se encuentran en la tabla 8.1.

Tabla 8.1: Variables disponibles en bases de datos.

Variable	Descripción	Nombre variable	Tipo de variable
Edad	Edad del cliente	EDAD	Numérica. Entero entre 18 y 90
Grupo Socioeconómico	Nivel socioeconómico que posee un cliente. De acuerdo al ingreso monetario, corresponde una de las categorías: ABC1, C2, C3, D y E, con ABC1 el nivel de mayores ingresos y E con menores ingresos.	GSE	Variable categórica. Niveles: ABC1, C2, C3, D y E.
Género	Rol social de un cliente de acuerdo a su sexo de nacimiento.	Genero	Binaria: 1/0. 0 = Mujer 1 = Hombre
Región de pertenencia	Región en la que aparece registrado el domicilio del cliente. Corresponde a una de las 13 regiones pertenecientes al territorio chileno.	REGION_PART	Categórica. Nombre de la región.
Ruralidad	Tipo de comuna rural o urbana a la que pertenece el dueño del auto. Se determina según la comuna de procedencia y datos de ruralidad del INE.	Tipo_comuna	Categórica. *Rural* o *Urbana*
Modelo	Modelo del auto perteneciente a la marca. Existen 17 modelos en la base de datos.	MODELO_PPU	Categórica. Nombres de modelos.
Fecha de venta	Fecha de venta del automóvil.	F_FACT_I	Fecha. Fecha a partir de 01/01/2016
Kilómetros registrados	Kilómetros que registra un vehículo al asistir a una mantención de las indicadas en pauta. Existe una variable por cada mantención pauta.	M_{1K} M_{15K} M_{30K} M_{45K}	Numérica. Enteros mayores a 0.
Fecha de asistencia a mantención	Fecha en que un cliente asistió a una mantención pauta. Existe una variable por cada mantención pauta. Registros desde 2016-01-01 hasta 2021-06-01.	F_1K_ F_15K_ F_30K_ F_45K_	Fecha formato “yyyy-mm-dd”.

Además de estas variables se tiene el VIN del vehículo como el RUT del cliente, las que son útiles como llaves primaria y secundaria para lograr identificar el vehículo y asociar los datos socio-demográficos del cliente.

De la unificación de las bases de datos se seleccionan aquellas observaciones que poseen registros con información completa de las variables socio-demográficas del cliente a excepción de la variable edad.

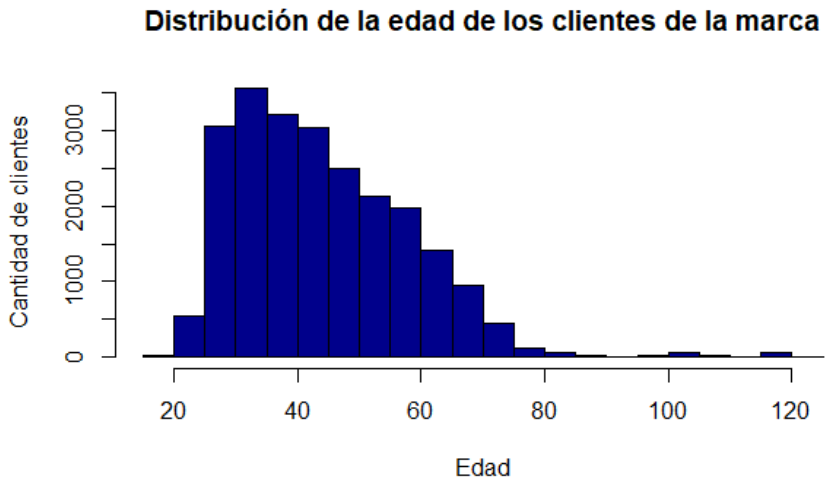


Figura 8.1: Distribución de edad de clientes

En la Figura 8.1 se puede observar la distribución de la edad de los clientes, observándose que la mayoría de los clientes (80 % aproximadamente) poseen entre 25 a 60 años de edad, decreciendo la cantidad de clientes a medida que aumenta la edad desde los 35 años. Se observa además que existen pocos registros cuya edad es superior a 100 años, por lo que se decide excluir estos registros de la base de datos a utilizar por ser probablemente incorrectos y corresponder a *outliers*. Al estudiar la variable edad se logra identificar que es una variable que posee errores en sus registros observando que existen datos que difieren con respecto a la edad aproximada tomando la fecha de nacimiento del cliente, además de existir un 8 % de los registros vacíos.

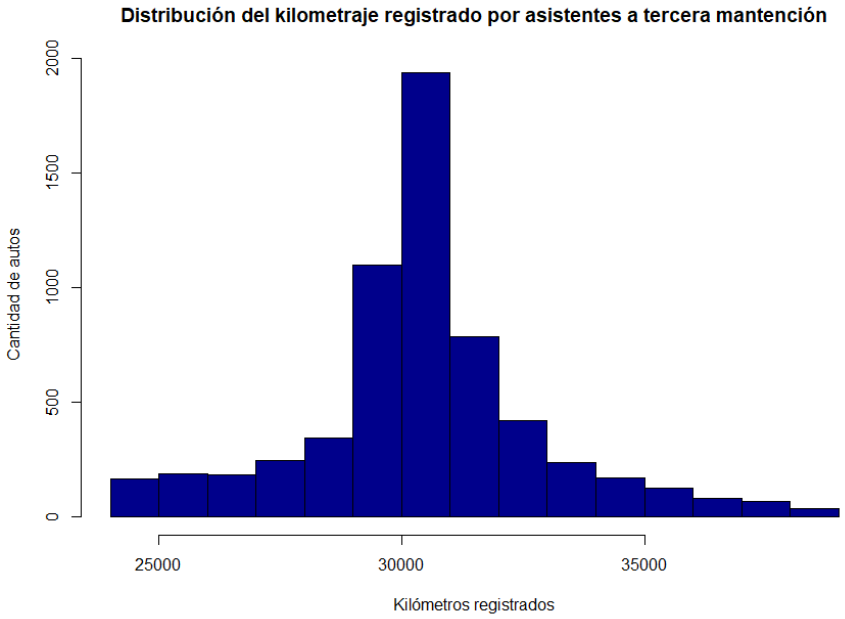


Figura 8.2: Distribución de kilometraje en tercera mantención

Se estudian las variables M_{1K}, M_{15K}, M_{30K} y M_{45K}. Estas poseen una distribución con alta tendencia a la normal, y la mayor asistencia a mantención ocurre en el punto del kilometraje recomendado para asistir y posterior a cumplido este, por lo que se deduce que la conducta más marcada es asistir a las mantenciones luego de cumplido el kilometraje recomendado para asistir una mantención pauta, como se puede ver por ejemplo en la Figura 8.2, donde se observa lo ocurrido con la tercera mantención.

En cuanto a los modelos de automóvil, los más comprados son los modelo *Morning* y *Río* en sus versiones 4 y 5, según la Tabla 8.2. En la misma, se puede apreciar que los modelos *Cadenza*, *Mohave*, *Río 3*, *Quoris*, *Koup*, y *Optima Híbrido* poseen una cantidad de registros bajo 20 observaciones.

Tabla 8.2: Cantidad de autos según modelo

Modelo	**Cantidad VIN**	**Modelo**	**Cantidad VIN**
Cadenza	5	Quoris	1
Carnival	653	Rio 3	20
Cerato	2.785	Rio 4	5.264
Cerato 5	82	Rio 5	5.408
Koup	2	Seltos	473
Mohave	20	Soluto	1.597
Morning	7.719	Sonet	227
Niro	39	Sorento	1.923
Optima	35	Soul	981
Optima Hibrido	10	Sportage	5.528

Estudiando la cantidad de personas por género se observa que las cifras son bastante parejas, siendo 12.004 automóviles correspondientes a mujeres y 11.908 a hombres, por lo que no existe una tendencia en cuanto a género. El nivel socio-económico está claramente marcado por una preferencia de la marca del segmento D, correspondiente al 43 % de los datos, seguido por el segmento C3 (23 %), luego el C2 (21 %) y finalmente, con cifras menores, el segmento ABC1 (8 %) y E (5 %). Esto puede mostrarnos la clara preferencia por la marca de parte de la llamada *clase media*.

La asistencia a las mantenciones por fecha del año, similar a lo ocurrido con la cantidad de ventas por año, poseen una baja en 2020, lo que fue provocado porque algunos concesionarios detuvieron su servicio por cuarentenas y probablemente porque los clientes dejaron de asistir como forma de tomar precauciones para evitar contagios.

La cantidad de autos según región de procedencia dispuesta en la Tabla 8.3, es marcada por la alta cantidad de automóviles en la región Metropolitana de Santiago. Esto es natural considerando la concentración de población en el centro del país. Se puede identificar que las grandes ciudades son las que poseen mayor cantidad de autos de la marca, como lo es Valparaíso por ser un puerto principal. Además, se observa que existen solo 23.912 registros en que se identifica la región de procedencia, quedando 8.860 registros

sin información.

Tabla 8.3: Cantidad de autos según Región de procedencia

Región	Cantidad VIN
Aysen Del General Carlos Ibanez	25
De Antofagasta	647
De Arica Y Parinacota	55
De Atacama	307
De Coquimbo	923
De La Araucania	283
De Los Lagos	958
De Los Rios	359
De Magallanes Y Antartica Chilena	45
De Ñuble	279
De Tarapaca	73
De Valparaíso	3.384
Del Bio Bio	1.131
Del Libertador Bernardo Ohiggins	1.113
Del Maule	757
Metropolitana De Santiago	13.573
Total	**23.912**

Cabe considerar que existen también algunos registros de ciudad y comuna particular del cliente pero menos cantidad que los de región particular. Se decide excluir estos datos del estudio debido a que estas variables poseen alta cardinalidad (57 ciudades y 298 comunas distintas) y muchas de estas ciudades o comunas poseen muy pocas observaciones asignadas a ellas, por lo que se requeriría eliminar estas observaciones para tener predicciones con menos errores.

Además, por su posterior utilidad se estudia la correlación de pearson entre las variables numéricas existentes. Para esto se seleccionan las filas de datos que poseen un kilometraje registrado en cada mantención mayor a cero, debido a que de lo contrario, se estarían considerando observaciones en que un cliente no asistió a una mantención y por tanto no se tiene registro del kilometraje que debía registrar en ese momento. Con esto, el estudio de correlación se realiza a 624 datos, lo que se observa en 8.3. Como se ve, las variables edad y género (considerada como binaria) poseen casi nula correlación con las variables de kilometraje. Además, los kilometrajes registrados en las diferentes mantenciones poseen muy baja correlación, pero aún así esta relación es positiva, es decir que a medida que el registro de kilometraje aumenta o disminuye en una mantención, en otra mantención se observará el mismo comportamiento aunque no en la misma escala.

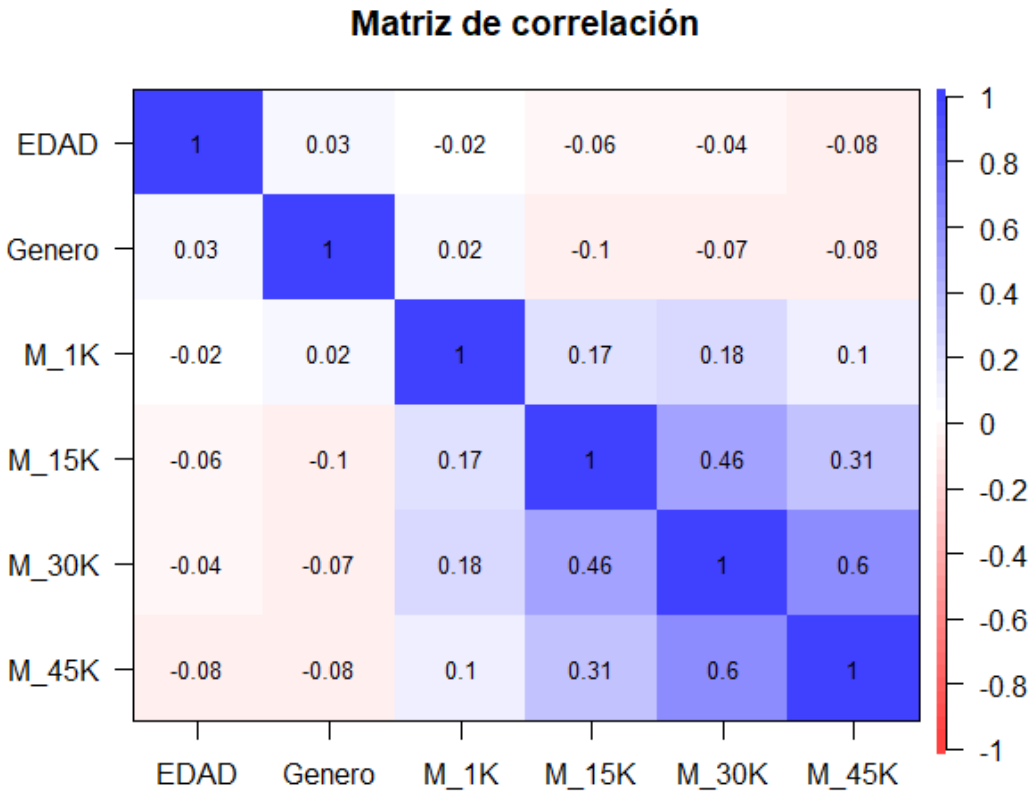


Figura 8.3: Matriz de correlación de variables numéricas

Para estudiar la correlación de las variables categóricas se calcula el parámetro V de Cramer, cuyo coeficiente indica correlación si es cercano a 1 y baja correlación si el valor es cercano a 0. En este caso, en la Tabla 8.4, se observa que existe una correlación media entre la región de procedencia del cliente y el tipo de comuna (o ruralidad). El resto de variables presenta una baja correlación entre sí, lo que permite considerar a estas variables como independientes.

Tabla 8.4: Correlación V de Cramer para variables categóricas

Variables	Modelo	Género	GSE	Región
Modelo	-	-	-	-
Género	0.119	-	-	-
GSE	0.122	0.047	-	-
Región	0.048	0.136	0.136	-
Ruralidad	0.048	0.008	0.101	0.614

8.2. Preparación de los datos

La preparación de los datos, como se mencionó anteriormente, consta en limpieza de datos incorrectos, vacíos y también transformación y creación de variables que pueden ser de ayuda para mejorar los resultados esperados. Para comenzar, se realiza una preparación de datos generales, y luego se realiza preparación de datos para cada problema a trabajar, por lo que esta sección se divide en tres partes a modo de tener una mayor

comprensión del proceso.

El primer paso de la preparación de los datos consta en eliminar los registros que corresponden a clientes que son empresas, puesto que se entiende que su comportamiento difiere del cliente persona natural usual, que es el tipo del cliente en quién quiere la empresa sea aplicado el estudio. Esto se realiza excluyendo de los datos a quienes poseen RUT mayor a 30 millones.

En segundo lugar, se excluyen de los datos el modelo de automóvil Frontier, por ser un modelo de vehículo de carga utilizado por empresas con fines de laborales, siendo su comportamiento distinto del cliente promedio como ocurre con los vehículos registrados por empresas. Sumado a esto, se excluyen los modelos Cadenza, Mohave, Río 3, Quoris, Koup, y Optima Híbrido, indicados en la sección anterior por poseer menos de 20 registros por modelo, lo que podría perjudicar las métricas de rendimiento de las predicciones debido a tener pocos *casos* relacionados a las variables en los que basar el estudio. Debido a este mismo motivo, se decide excluir del estudio las regiones de Aysén del General Carlos Ibañez, Arica y Parinacota, Magallanes y la Antartica Chilena y Tarapacá, debido a que poseen una cantidad inferior al 1 % de los datos.

Continuando con la limpieza de datos, se eliminan los registros que poseen datos vacíos de grupo socio-económico, género, región de procedencia, modelo del auto, RUT del cliente, VIN del vehículo y fecha de venta, debido a que estos serían datos útiles y necesarios para poder trabajar con modelos de predicción. Las variables de kilómetros registrados tanto como la fecha de asistencia a una mantención no son motivo de eliminación de filas de inicio, ya que estas variables se considerarán de distintas formas dependiendo del problema al que busque resolver. Con esta limpieza de datos quedan disponibles un total de 23.407 observaciones.

Para lograr obtener datos completos y correctos de la variable *Edad* se actualizan realizando el cálculo a diciembre de 2021 con respecto a la fecha de nacimiento registrada del cliente. Además, en los casos en que no se posee la fecha de nacimiento, se calcula el año de nacimiento con la fórmula de regresión lineal[22].

$$Rut \cdot 3{,}34 + 1932{,}26$$

Donde el RUT se considera un número decimal tomando los millones como unidades, permitiendo así calcular la edad de los clientes. Se excluyen de los datos los casos en que la edad es mayor a 100 años debido a la poca cantidad de observaciones presentes con estos datos (como se observó en la Figura , y en que figura como valor negativo, los que serían provocados por el margen de error de la fórmula utilizada para el cálculo de la edad.

Se trabaja en resolver dos problemas de diferente naturaleza como lo es encontrar la fecha de asistencia a la próxima mantención y por otra parte, la propensión de asistencia a una mantención, para lo que se crean sub-sets de datos en que se seleccionan las variables y se transforman las que sean necesarias de acuerdo a los requerimientos que presenta cada problema. A continuación se presenta el estudio general por problema de

los set de datos con los que se trabaja.

8.3. Modelo de estimación de fecha de asistencia

8.3.1. Preparación de los datos

Para lograr predecir la próxima fecha de asistencia a mantención de un cliente se determina como buena medida de estimación los días hasta la próxima mantención (debido a que con fechas directamente no es posible trabajar). Los días obtenidos de la predicción serían sumados a la última fecha de asistencia o de venta del vehículo, según sea el caso.

En los casos en que un cliente no asiste a una mantención, los días a esa mantención se desconocen, por lo que se hace necesario dividir la base de datos en sub-sets de datos para cada mantención pauta a la que se quiere predecir fecha de asistencia. Además, estos set de datos se dividen según asistencias a mantenciones pauta previas, con el fin de obtener filas con datos completos para cada modelo de estimación.

Para evitar sobre-ajuste de los algoritmos y que los resultados sean confiables, se decide excluir por cada sub-set de datos algunos *Modelos* y *Regiones* en los casos en que existe una cantidad muy baja de observaciones según estos. (Para más detalle están los Anexos C.1 y C.2). La creación de los sub-sets de datos luego de aplicada esta limpieza se observa en la Tabla 8.5.

Tabla 8.5: División de sub-sets de datos para predicción de fecha asistencia a mantenciones

Mantención	**Asistencia a mantención previa *(Código modelo)***	**Cantidad de datos**	**% del total**
Segunda Mantención (15.000 kms.)	Primera mant. (1.000 kms.) *(D15M1)*	5.801	29,8 %
	No asistió a mantención previa *(D15M0)*	5.932	30,5 %
Tercera Mantención (30.000 kms.)	Primera y segunda (de 1.000 y 15.000 kms.) *(D30M3)*	2.398	12,3 %
	Primera mant. solamente (de 1.000 kilómetros) *(D30M1)*	385	2 %
	Segunda mant. solamente (de 15.000 kms.) *(D30M15)*	2.310	11,9 %
	no asistió a otras mantenciones previas *(D30M0)*	808	4,1 %
Cuarta Mantención (45.000 kms.)	Segunda y tercera mant. (de 15.000 y 30.000 kms.) *(D45M4)*	1.167	6 %
	Segunda mant. (de 15.000 kms.) *(D45M15)*	206	1,1 %
	Tercera mant. (de 30.000 kms.) *(D45M30)*	290	1,5 %
	sin asistencia a segunda ni tercera mant. *(D45M0)*	179	0,9 %

Como se indica en la Tabla 8.5 se obtienen 10 sub-sets de datos. En el caso de la cuarta mantención se considera la asistencia a las dos últimas mantenciones previas a la cuarta mantención, debido a que la información proporcionada por la primera mantención es probable que sea redundante y a la vez, se generarían muchas configuraciones de sets de datos extras para la cuarta mantención, perdiendo validez los modelos considerando la baja cantidad de datos ya presente en algunos de los sub-sets propuestos. la cantidad de datos para algunos set de datos es inferior a 400 observaciones, por lo que se estima que los resultados de predicción que otorguen estas observaciones podrían ser

poco confiables.

Dado que se busca predecir los días que faltan hasta la próxima mantención pauta, se crea nuevas variables que corresponden a los días de diferencia hasta asistir a una mantención con respecto a la fecha de la última mantención a la que se registró asistencia de un cliente, o en su defecto, a la fecha de venta del vehículo. Por ejemplo, para el caso en que un cliente asistió a la tercera y segunda mantención, los días hasta la tercera mantención se calculan como la diferencia entre la fecha de asistencia a la tercera mantención y la fecha de asistencia a la segunda mantención. En los casos en que no ha asistido a una mantención previa, se calcula la diferencia con respecto a la fecha de venta del automóvil. En la Tabla 8.6 se observa a más detalle las nuevas variables creadas.

Tabla 8.6: Creación de variables Días entre mantenciones

Modelo	**Variables**
D15M1	D1: Diferencia de días entre venta y primera mantención del vehículo. D15: Diferencia de días entre primera y segunda mantención.
D15M0	D15: Diferencia de días entre venta y segunda mantención.
D30M3	D1: Diferencia de días entre venta y primera mantención del vehículo. D15: Diferencia de días entre primera y segunda mantención. D30: Diferencia de días entre segunda y tercera mantención.
D30M15	D15: Diferencia de días entre venta y segunda mantención. D30: Diferencia de días entre segunda y tercera mantención.
D30M1	D1: Diferencia de días entre venta y primera mantención. D30: Diferencia de días entre primera y tercera mantención.
D30M0	D30: Diferencia de días entre venta y tercera mantención.
D45M4	D15: Diferencia de días entre venta y segunda mantención. D30: Diferencia de días entre segunda y tercera mantención. D45: Diferencia de días entre tercera y cuarta mantención.
D45M30	D30: Diferencia de días entre venta y tercera mantención. D45: Diferencia de días entre tercera y cuarta mantención.
D45M15	D15: Diferencia de días entre venta y segunda mantención. D45: Diferencia de días entre segunda y cuarta mantención.
D45M0	D45: Diferencia de días entre venta y cuarta mantención.

Además, para la cuarta mantención se crea la variable *asistió a primera mantención*, representada por "*A1*". Se crea también la variable "*Numfec*", correspondiente a un número mayor a cero relativo a la fecha de venta de un vehículo. Por términos prácticos, se considera la fecha 01-01-2016 como día 0, al que se le suma una unidad por cada día siguiente. Así, según la fecha de venta se asigna un número correlativo que aumenta a medida que más reciente es la fecha de venta del mismo.

8.3.2. Conocimiento de los datos

Como se añaden nuevas variables y se acota el tamaño de los set de datos utilizados para modelar cada situación, se realiza nuevamente un estudio de los datos. Como ejemplo se presenta el estudio del set de datos para el modelo *D30M3*.

La correlación de Pearson del sub-set, en la Figura 8.4 muestra que existe alta relación entre las variables de kilometraje registrado y los respectivos días hasta la asistencia a la mantención correspondiente. Debido a esto se decide excluir del estudio las variables de kilometraje y en lugar de estas, utilizar las variables de días hasta una mantención, debido a que lo que se espera obtener son los días hasta la próxima mantención correspondiente a un cliente, y basarse en variables similares con parámetros similares permite obtener una predicción más precisa que con las variables de kilometraje.

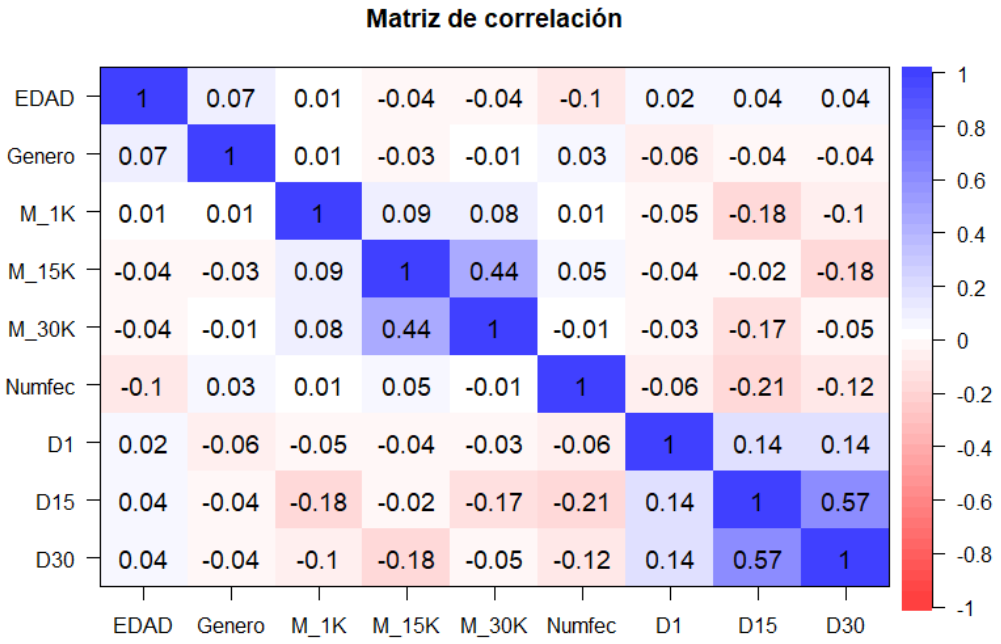


Figura 8.4: Matriz de correlación de variables numéricas sub-set D30M3

Además, en la Figura 8.5 se puede observar que las variables de kilometraje de segunda y tercera mantención y la distribución de los días hasta una mantención poseen una distribución normal.

Figura 8.5: Matriz de distribución de variables numéricas sub-set D30M3

8.3.3. Aplicación de modelos

Con las variables a disposición los modelos aplicados se representan por una regresión como la siguiente, para el caso del modelo M30D, por ejemplo:

$$D30 \sim Edad + Genero + GSE + Region + Tipo\_comuna + Modelo + Numfec + D15 + D1 + M\_15K + M\_1K \quad (8.1)$$

donde D30 es la mantención correspondiente a la que se quiere regresionar los días, y las variables independientes, a excepción de *D15* son las utilizadas para todos los modelos de regresión como regresores. Las variables de *días hasta la una mantención* indicadas en la Tabla 8.6 y asistencia a la primera mantención *A1* son las que varían de un algoritmo a otro según apliquen al caso.

Para comparar los distintos algoritmos utilizados en este trabajo se divide cada set de datos en set de entrenamiento y testeo con proporciones 80:20 , permitiendo obtener los modelos proporcionando los datos de entrenamiento a cada algoritmo y observando los resultados realizando las predicciones sobre el sub-set de testeo, comparándolo con los datos reales.
Para cada sub-set de datos se realiza aplicación de regresión múltiple, árbol de regresión, bosques aleatorios, máquinas de soporte vectorial y potenciación del gradiente, utilizando el programa R studio. Los modelos se configuran para todos los casos de la siguiente forma:

- **Regresión múltiple**: Se utiliza una regresión múltiple básica con la función *lm* y la

ecuación 8.1 (a modo de ejemplo), con la variable dependiente y las variables de días independientes asignadas según lo indicado en la Tabla 8.6. Para aplicar una regresión múltiple se verifica que se cumplan los supuestos indicados en la sección 6.1.

- **Árbol de decisión CART**: Se utiliza el árbol tipo *CART* para realizar una regresión. Se utiliza la función *rpart* del paquete del mismo nombre, y configura la herramienta para que utilice el método anova, que compara la varianza de las medias de cada grupo creado por el árbol de decisión.

- **Bosques aleatorios (RF)**: Se convierten las variables categóricas a factores, debido a que este algoritmo (al igual que SVM y GBM) sólo trabajan con variables numéricas y factores, excluyendo las variables categóricas. Así, cada variable categórica considera números correspondientes a niveles según sea la cantidad de valores únicos de cada variable categórica. Se utiliza la función *randomForest* perteneciente al paquete con el mismo nombre. Se aplica también búsqueda de rejilla (grid search) que permite optimizar los hiperparámetros de la función, permitiendo obtener los mejores resultados que puede entregar el algoritmo.

- **Máquinas de soporte vectorial (SVM)**: Al igual que en bosques aleatorios se convierten las variables categóricas a factores y se utiliza la función *svm* del paquete *e1071*, otorgándole las variables correspondientes en cada caso, el set de entrenamiento y se configura la herramienta para que escale los datos (considerando que SVM tiende a considerar la magnitud de los valores para otorgarle importancia a las variables) de forma que se consideren con similares pesos cada variable, independiente de la escala que posean. Se utiliza la configuración “eps-svr” que indica a la herramienta que se debe realizar una regresión, lo mismo que el kernel *rbfdot*, asignando una función de costo de 1 y epsilon de 0.1, como se asigna por defecto en la función.

- **Potenciación del gradiente (GBM)**: Al igual que en los dos últimos algoritmos, se convierten las variables categóricas a factores, y se aplica la función *gbm* del paquete con el mismo nombre. Se configura la herramienta para que utilice una función gaussiana con 500 árboles de decisión y se dejan el resto de hiperparámetros por defecto en la herramienta.

Para cada uno de los algoritmos se utiliza un set de entrenamiento para generar cada modelo, y luego se evalúa mediante la aplicación de cada modelo sobre el set de testeo (20 %). Para obtener las métricas de comparación se calculan el error cuadrático medio (RMSE), el error absoluto medio (MAE) y el error porcentual absoluto medio (MAPE). Además, para intervalos de $\pm 15, \pm 30, \pm 45$y $\pm\ 60$ días se calcula el porcentaje de datos correctamente identificados.

8.3.4. Análisis de resultados

Debido a que se trabaja en el obtener el modelo con mejor rendimiento para 10 subsets de datos, los resultados de estos se adjuntan en el Anexo E.

8.3.4.1. Segunda mantención

Al aplicar los algoritmos de predicción de fecha de asistencia a la segunda mantención, los modelos que mejor desempeño presentan para cada sub-set de datos se indican en la Tabla 8.7, siendo en ambos modelos el algoritmo de *máquinas de soporte vectorial* el de mejor rendimiento. Los resultados indican que con estos modelos se logra obtener un error porcentual medio (*MAE*) de aproximadamente un 65 % y 42 % para los sets D15M1 y D15M0 respectivamente.

Tabla 8.7: Resultados modelos mejor desempeño segunda mantención

Sub-set	Modelo	RMSE	MAE	MAPE
D15M1	SVM	138.8	97.9	65 %
D15M0	SVM	273.74	217.88	30 %

Estudiando más a fondo los resultados, al calcular el nivel de coincidencia mediante intervalos de días, es interesante notar que para ambos casos, considerando alrededor de un mes de margen de error (30 días), solo un 20 % de los datos es predicho de forma correcta, mientras que a 60 días se alcanza entre un 35 % y 40 % de precisión. Considerando que la idea de este trabajo es mejorar la precisión de la predicción de fechas de asistencia, este nivel de acierto se considera bastante bajo para la magnitud de movimiento de ventas y transacciones en la empresa. Aún así, considerando que el nivel de predicción actual para la segunda mantención es el adjunto en la Tabla 8.8 y que el comportamiento general es a asistir posterior a cumplir el kilometraje recomendado en la mantención, los resultados obtenidos en este trabajo aplicando algoritmos de máquinas de soporte vectorial se consideran una mejoría del nivel de predicción existente. Se logra mejorar en aproximadamente un 12 % las predicciones para un rango de error de predicción de 60 días.

Tabla 8.8: Comparación resultados obtenidos con existentes para la segunda mantención (D15M1)

Subset	Modelo	Dif $\pm$ 15 días	Dif $\pm$ 30 días	Dif $\pm$ 40 días	Dif $\pm$ 60 días	N datos
Modelo previo	8 %	16 %	21 %	30 %	70 %	1.612
Máquinas de soporte vectorial	13 %	23 %	30 %	42 %	58 %	1.16

8.3.4.2. Tercera mantención

Al trabajar en la predicción de la fecha de asistencia a la tercera mantención, los algoritmos que mejor resultado presentaron en general fueron *bosques aleatorios* y *máquinas de soporte vectorial*. El resumen de los mejores modelos obtenidos se presenta en la Tabla 8.9, donde en este caso el error porcentual absoluto medio (*MAPE*) es bastante bajo para los casos que se considera la asistencia previa a la segunda mantención, siendo del orden de 27 %, mientras que para los set de datos en que no existe asistencia a la segunda mantención el error aumenta considerablemente. Esto puede ser ocasionado por la cantidad de observaciones utilizadas en cada modelo (como se observa en la Tabla 8.5), donde los casos en que la cantidad de observaciones es menor a 1.000 observaciones, los resultados presentan muy baja precisión.

Tabla 8.9: Resultados modelos mejor desempeño tercera mantención

Sub-set	Modelo	RMSE	MAE	MAPE
D30M3	RF	102.83	75.03	0.27
D30M15	SVM	109.59	80.29	0.27
D30M1	RF	222.80	167.84	0.43
D30M0	SVM	226.09	173.58	0.36

Al estudiar el nivel de precisión según intervalos de días, acorde a lo indicado anteriormente sobre el nivel de error en debido a las observaciones, coincide con que los modelos D30M3 y D30M15 presentan un nivel de acierto de alrededor de un 50 % en un intervalo de 60 días, y para los modelos D30M1 y D30M0, el nivel de error en este mismo intervalo es cercano a 25 %.

Nuevamente, ocurre que al comparar con los niveles de precisión de predicciones actuales en la empresa, el set de datos que es posible comparar, que es *D30M3* (ya que no existen más resultados en la empresa), presentaría una mejoría si se aplicaran el algoritmo de máquinas de soporte vectorial. Aún así, a diferencia del caso de la segunda mantención, las diferencias con respecto al nivel de precisión actual de la empresa son mínimas, siendo de alrededor de un 5 % en intervalos de 60 días y no existe diferencia de nivel de error en intervalos de 30 días.

Tabla 8.10: Comparación precisión de predicción días hasta tercera mantención con previa asistencia (Modelo D30M3)

Modelo	Dif $\pm$ 15 días	Dif $\pm$ 30 días	Dif $\pm$ 40 días	Dif $\pm$ 60 días	Dif >60 días	N datos
RF	16 %	27 %	36 %	52 %	47 %	480
Modelo actual	15 %	27 %	34 %	47 %	53 %	1.040

8.3.4.3. Cuarta mantención

En el caso de la predicción de días para la cuarta mantención, los algoritmos que mejores resultados presentaron son *bosques aleatorios* (RF) y *regresión múltiple*. Los resultados se presentan en la Tabla 8.11, donde se puede observar que nuevamente, para los casos en que se tiene información de la asistencia a la mantención inmediatamente previa a la mantención en cuestión, es decir, de la tercera mantención en este caso, el nivel de error es de aproximadamente un 25 % (modelos D45M4 y D45M30) lo que es una buena estimación aproximada. Por otra parte, para los casos en que no se tiene información de la tercera mantención, el nivel de error ronda cercano a 35 % (MAPE). Al observar los errores según intervalos de días, los sets de datos con información de la tercera mantención alcanzan alrededor de un 50 % de acierto en un intervalo de error de 60 días, lo que se considera relativamente bueno; mientras que para los casos en que no se tiene información de la tercera mantención, el nivel baja significativamente a alrededor de un 20 % de acierto.

Tabla 8.11: Resultados modelos mejor desempeño cuarta mantención

Sub-set	Modelo	RMSE	MAE	MAPE
D45M5	RF	108.79	78.41	0.28
D45M30	RF	150.65	128.75	0.23
D45M15	RF	127.24	97.67	0.40
D45M0	RL	270.56	208.02	0.23

Una vez más, comparando frente a los resultados de las predicciones actuales de la empresa, en este caso para el set de datos D45M4 (Tabla 8.12, el nivel de acierto de la empresa es menor en alrededor de un 2 % con respecto a los resultados obtenidos por el mejor modelo para este set de datos.

Tabla 8.12: Comparación precisión de predicción días hasta cuarta mantención con previa asistencia (Modelo D45M4)

Subset	Modelo	Dif $\pm$ 15 días	Dif $\pm$ 30 días	Dif $\pm$ 40 días	Dif $\pm$ 60 días	N Datos
Modelo previo	14 %	29 %	37 %	49 %	51 %	334
Random Forest	16 %	27 %	36 %	52 %	48 %	227

8.3.5. Conclusiones modelo predicción fecha de asistencia

Como se aprecia, los modelos que poseen mejor rendimiento para la predicción de días hasta la asistencia a una mantención son bosques aleatorios y máquinas de soporte vectorial en general. Aún así, en la mayoría de los casos el error absoluto medio ronda e incluso supera el valor de 100 días lo que corresponde a mas de 3 meses, lo que se considera un error de predicción muy alto si se busca predecir la fecha de asistencia en que avisar a un cliente que le correspondería asistir a una mantención. Esto claramente es afectado por baja cantidad de datos para entrenamiento y también la poca cantidad de información (variables) disponibles, puesto que no se logra detectar claramente el comportamiento de los clientes, lo que es afectado también por el hecho de que las mantenciones sean realizadas en intervalos de tiempo tan extensos como un año.

Por otra parte, considerando que las predicciones realizadas actualmente por la empresa presentan un porcentaje de error mayor que las obtenidas con estos modelos (en los casos en que se tiene información para comparar), se podría considerar una mejora a las predicciones el aplicar estos algoritmos a los datos de la empresa. Es claro que estos podrían mejorar poco a poco a futuro a medida que se tenga mayor cantidad de registros de mantenciones.

En cuanto a la importancia de las variables para la predicción de días, es notable que frente a la ausencia de variables de asistencia previa a mantenciones disminuye la precisión de las predicciones, pero esto también ocurre porque se tiene una cantidad menor de datos a medida que existe ausencia a mantenciones previas. Por otra parte, estas variables de asistencia previas, tanto días como kilometraje, son las que más importancia tienen para los modelos, siguiendo con la información de modelo del vehículo y región de residencia, mientras que las variables que en todos los casos entregan muy poco o nulo aporte son las de género del cliente y ruralidad (Tipo comuna), mostrando que en el

comportamiento de asistencia no influye en realidad si el dueño del vehículo es hombre o mujer.

8.4. Modelos de propensión de asistencia

8.4.1. Preparación

El modelo de propensión de asistencia consiste en un modelo de clasificación que permite determinar si un individuo asistirá o no a una mantención. Con el fin de obtener una predicción más precisa, se trabaja con sub-sets de datos para cada mantención, es decir, para la asistencia a la segunda, tercera y cuarta mantención pauta, correspondientes a las mantenciones de 15.000 o 1 año desde la compra, 30.000 o 2 años desde la compra, y 45.000 kilómetros o 3 años desde la compra, respectivamente. Para esto, se utiliza todos los datos disponibles de vehículos que hasta la fecha han realizado una mantención y/o ya debiesen haber cumplido con asistir considerando el tiempo que ha pasado desde la compra del auto o última mantención asistida. De forma más específica, los filtros para cada sub-set son los indicados en la Tabla 8.13.

Tabla 8.13: Filtros creación sub-sets de datos para modelo de propensión

Mantención	Filtros	Cantidad de datos
Segunda mantención	1) Segunda mantención realizada. 2) Segunda mantención no realizada y que cumpla a agosto de 2021: - Más de 640 días desde la fecha de primera mantención. - Más de 725 días desde la fecha de venta.	21.817
Tercera mantención	1) Tercera mantención realizada. 2) Tercera mantención no realizada y que cumpla a agosto de 2021: - Más de 650 días desde la segunda mantención. - Más de 1.030 días desde la primera mantención. - Más de 1.125 días desde la fecha de venta.	15.520
Cuarta mantención	1) Cuarta mantención realizada. 2) Cuarta mantención no realizada y que cumpla a agosto de 2021: - Más de 615 días desde la tercera mantención. - Más de 1.030 días desde la segunda mantención. - Más de 1.380 días desde la primera mantención. - Más de 1.450 días desde la fecha de venta.	8.089

Los subset de datos indicados en la Tabla 8.13 se desarrollan considerando los casos en que se registra asistencia a una mantención, y como no existe información certera de inasistencias porque esta información no la proporciona el cliente ni se poseen registros de esto mas que la suposición de que ya ha pasado tiempo suficiente como para decir que no asistió a la mantención, se decide fijar una cota inferior de días que se deben cumplir como mínimo para considerar que es un caso de inasistencia a una mantención. Esta cota inferior se determina considerando el percentil 97 de la distribución de los días de diferencia entre dos mantenciones consideradas. Por ejemplo, para calcular el percentil 97 del caso de asistencia a la tercera mantención con respecto a un cliente que asistió a la primera mantención solamente, se calculan los días de diferencia para los casos en que clientes asistieron efectivamente a la tercera mantención y a la primera mantención

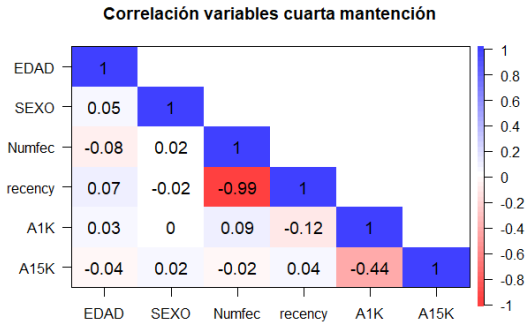
(excluyendo casos en que hubo asistencia a la segunda mantención) y se calcula el percentil 97 de estos datos. Se decide que sea el percentil 97 debido a que la distribución de los datos posterior a este percentil posee una desviación estándar muy alta con respecto al resto de los datos, por lo que en este punto se logra cubrir los casos mas usuales y mayoritarios.

Adicionalmente a lo anterior, se crea la variable *recency* (frescura). Para cada caso, en que se estudia la propensión de asistencia a una mantención "Y", la frescura (*recency*) se define como: (a) si existe registro de asistencia a la mantención "Y" estudiada, entonces la variable se define como los días de diferencia entre la fecha de mantención a la que asistió previamente a la estudiada y la fecha de asistencia a la mantención "Y"; y en el caso (b) en que no existe asistencia registrada a la mantención "Y", entonces se toma la diferencia de días entre la fecha de la última mantención a la que asistió el cliente previa a "Y" y la fecha de actual (del día de hoy). Para el desarrollo de este trabajo, como se trabaja con datos estáticos hasta el día 31 de agosto de 2021, se considera esta última fecha en lugar de la fecha actual.

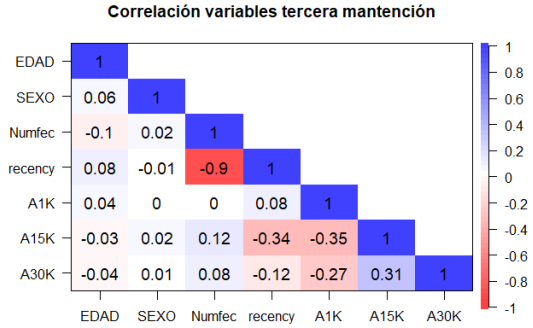
También se crea la variable de asistencia a una mantención para cada mantención pauta, donde el nombre de cada variable viene representada por *"AXK"* con X el valor de miles de kilómetros de la mantención correspondiente, siendo así las variables nombradas como "A1", "A15", "A30" y "A45". Esta variable se crea como variable binaria, tomando valores 0 o 1, correspondiendo a no asistencia y asistencia, respectivamente.

Por fines prácticos, para el desarrollo de modelos de propensión de asistencia no se utilizan las variables de kilometraje ni días hasta una mantención, debido a que lo importante es saber si un cliente asistirá o no a una mantención y para esto es necesario comparar situaciones en que se posea información completa para ambos casos. Para el caso de las variables de kilometraje y días hasta una mantención esto no podría cumplirse, dado que los casos de no asistencia carecerían de estos datos, por lo que no sería posible compararlos con respecto a los casos de asistencia de forma equitativa.

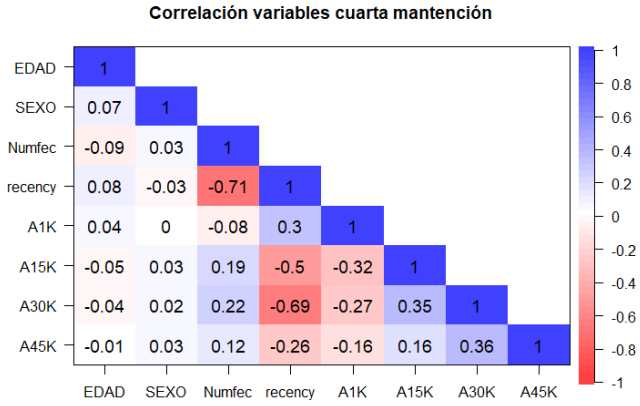
Para el desarrollo de estos modelos, al igual que con los modelos anteriores, se excluyen algunos modelos de vehículo y región del cliente dependiendo de las observaciones existentes.



(a) Variables segunda mantención



(b) Variables tercera mantención

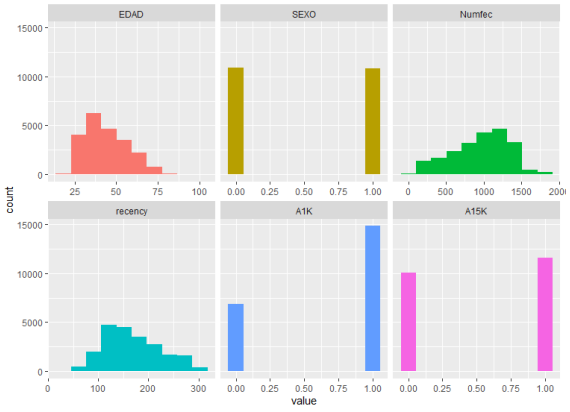


(c) Variables cuarta mantención

Figura 8.6: Matrices de correlación sets de datos para modelos de propensión de asistencia

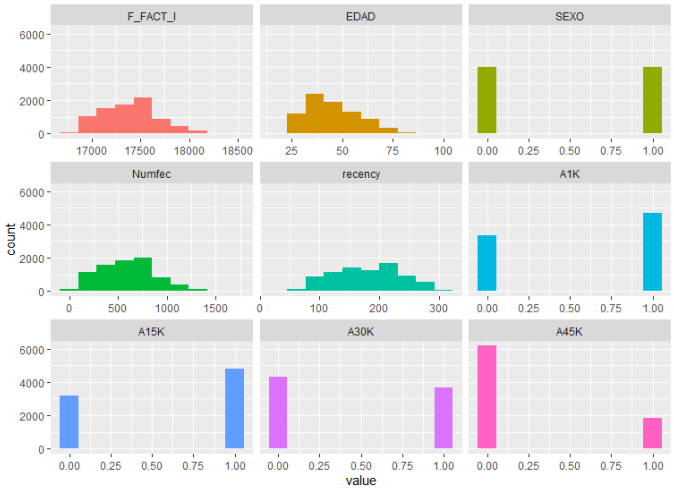
Con estos datos, se estudia la correlación entre variables, la que se presenta en las Figuras 8.6. Se observa que en los 3 sets de datos las variables *recency* con *Numfec* están altamente correlacionadas de forma negativa, por lo que es necesario excluir del estudio la variable *Numfec* dado que generalmente es usada la variable *recency* para este tipo de problemas. La variable *A30K* también posee correlación media-alta con la variable *recency*, lo que puede deberse a que el ser de datos de la cuarta mantención posee un alto grado de observaciones en que la variable de asistencia a la tercera mantención es 0. El resto de variables presentan correlación baja y hasta nula, por lo que se puede asumir independencia entre las variables.

Además, en las figuras 8.7 se observa que las variables numéricas poseen una distribución generalmente con tendencia a normal. Para el set de la cuarta mantención se observa que la variable dependiente (*A45*) posee un desbalance de los datos, siendo las observaciones de asistencia la mtiad de la cantidad de observaciones de inasistencia a la cuarta mantención.



(a) Variables segunda mantención

(b) Variables tercera mantención



(c) Variables cuarta mantención

Figura 8.7: Distribución sets de datos para modelos de propensión de asistencia

8.4.2. Aplicación de modelos

En la búsqueda del mejor modelo de propensión de asistencia para cada subset de datos se prueban los algoritmos de regresión logística, bosques aleatorios, máquinas de soporte vectorial, potenciación de gradiente y redes neuronales artificiales. Para cada algoritmo se utiliza la variable dependiente de asistencia a la mantención vehicular en cuestión. Las variables independientes seleccionadas para utilizar son todas las variables socio-demográficas disponibles. Por ejemplo, para predecir la probabilidad de asistencia a la mantención de 30.000 kilómetros, se utilizan las variables socio-demográficas del cliente más la asistencia a la mantención de 1.000 kilómetros, asistencia a mantención de 15.000 kilómetros, la 'frescura' ("recency") según cada caso. En resumen, las fórmulas vienen dadas como sigue:

Modelo propensión de asistencia a segunda mantención

$$A15K \sim A1K + Edad + Genero + GSE + Region + Tipo\_comuna + Modelo + Modelo + recency \quad (8.2)$$

Modelo propensión de asistencia a tercera mantención

$$A30K \sim A1K + A15K + Edad + Genero + GSE + Region + Tipo\_comuna + Modelo + recency \quad (8.3)$$

Modelo propensión de asistencia a cuarta mantención

$$A45K \sim A1K + A15K + A30K + Edad + Genero + GSE + Region + Modelo + Tipo\_comuna + Modelo + recency \quad (8.4)$$

Se utilizan los algoritmos mencionados anteriormente, a los que se introduce las ecuaciones presentadas en 8.2, 8.3 y 8.4. Al igual que en el problema de predicción de fecha de asistencia, se divide cada set de datos en set de entrenamiento y testeo, en proporciones 80:20. Como se mencionó anteriormente, existe un desbalance en la cantidad de observaciones con respuesta positiva con respecto a la negativa, lo que puede provocar un sesgo en los resultados de los modelos. Para evitar esto, se realiza undersampling, es decir, un sub-muestreo de los datos (se utiliza la función *ovunsample* del paquete *ROSE*), con la idea de dejar el set de entrenamiento de los datos con cantidades equilibradas de observaciones. Se realiza un balance 50 % y 50 %, ya que los modelos como Random Forest y los otros modelos utilizados, se desempeñan mejor cuando poseen una cantidad equivalente de observaciones por clase.

Cada algoritmo se configura como sigue:

- **Regresión Logística**: Se realiza una regresión logística múltiple básica utilizando la función *glm*, la que se configura aplicando las ecuaciones para cada caso, indicando en la configuración de la función que la variable de respuesta es binomial.
- **Bosques aleatorios (RF)**: Se convierten las variables categóricas a factores. Así, cada variable categórica considera números correspondientes a niveles según sea la cantidad de valores únicos de cada variable categórica. Se utiliza la función *randomForest* perteneciente al paquete con el mismo nombre. Se aplica también búsqueda de rejilla (grid search) que permite optimizar los hiperparámetros de la función, permitiendo obtener los mejores resultados que puede entregar el algoritmo.
- **Máquinas de soporte vectorial (SVM)**: Se convierten las variables categóricas a factores y se utiliza la función *svm* del paquete *e1071*, otorgándole las variables correspondientes en cada caso, el set de entrenamiento y se configura la herramienta para que escale los datos (considerando que SVM tiende a considerar la magnitud de los valores para otorgarle importancia a las variables) de forma que se consideren con similares pesos cada variable, independiente de la escala que posean.
- **Potenciación del gradiente (GBM)**: Al igual que en los dos últimos algoritmos, se convierten las variables categóricas a factores, y se aplica la función *gbm* del paquete con el mismo nombre. Se configura la herramienta para que utilice una función

gaussiana con 500 árboles de decisión y se dejan el resto de hiperparámetros por defecto en la herramienta. Se configura con 500 árboles y distribución *bernoulli*.

- **Redes Neuronales Artificiales**: Para este algoritmo se utiliza el entorno *h2o*, con la función *deeplearning*, utilizada para aplicar redes neuronales artificiales. Se aplican 100 capas y función de activación “rectifier”.

Luego de entrenar cada modelo, se calcula la matriz de confusión sobre el set de datos de testeo, y a partir de esta, se calculan las métricas de *Accuracy*, sensibilidad, precisión, especificidad y puntuación F1. Además se grafica la curva ROC (sensibilidad versus) como forma de estudiar el rendimiento de cada algoritmo.

8.4.3. Análisis de resultados

Luego de aplicados los algoritmos a cada set de datos, se obtiene los resultados para la propensión de asistencia a la segunda mantención en la Tabla 8.14, a la tercera mantención en la Tabla 8.15 y a la cuarta mantención en la Tabla 8.16. Debido a que en este caso es de interés de la empresa conocer con precisión qué clientes tienen mayor probabilidad de asistir a la mantención para intentar asegurar la asistencia mediante recordatorios y contacto con el cliente, y en segundo lugar, conocer quienes tienen menor tendencia a asistir, para evaluar planes de motivación mediante recursos de marketing u otros métodos (por planificar a futuro por la empresa), las métricas de rendimiento más importantes a evaluar en este caso son sensibilidad y luego especificidad, y por ende, accuracy.

8.4.3.1. Segunda mantención

Al desarrollar los algoritmos indicados para este tipo de problema de clasificación, algoritmos como LOGIT, bosques aleatorios y potenciación de gradiente entregan análisis de la importancia de las variables para el modelo en estudio. Por esto, en el Anexo F.1, en la Tabla F.3 se presenta los resultados de la regresión logística, el que entrega que las variables ingresadas al modelo general son significativas, a excepción de el nivel socio-económico (GSE). Por otra parte, potenciación de gradiente tanto como los bosques aleatorios indican (como los mismos resultados) que las variables menos significativas son Tipo_comuna (ruralidad) y el género del cliente (indicado como la variable *SEXO* en la imagen 8.8), lo que se debería a que el género de los clientes se encuentra distribuido de forma más bien uniforme dentro de los datos, y la variable de tipo comuna toma mayormente el valor de urbano (en más del 80 % de los datos iniciales).

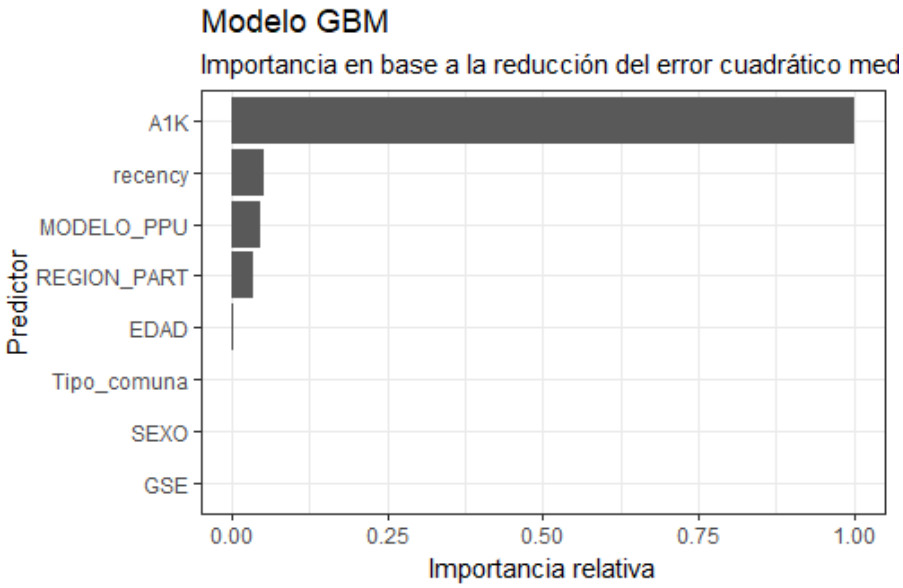


Figura 8.8: Importancia variables según modelos GBM

Por otra parte, las variables de mayor importancia en los 3 casos son la variable de asistencia previa a mantención y *recency*, lo que hace sentido considerando que son las variables que poseen mayor dinamismo a través del tiempo, a diferencia del modelo del automóvil o la región de pertenencia del cliente, que es muy probable que sean estáticas en el tiempo, pero pueden ser útiles para segmentar a los clientes. La asistencia previa, a pesar de que se suele considerar que la asistencia a la primera mantención es incentivada mayormente por ser gratuita y tener solo un mes de nuevo el vehículo, en este caso presenta una gran influencia en la asistencia a la segunda mantención, por lo que fomentar la asistencia a la primera mantención podría incidir en la asistencia a posteriores mantenciones, si se consideran estos resultados como relevantes.

Además, como se puede ver el nivel socio-económico no posee mucha relevancia, aún cuando por ejemplo, en el estudio aplicado por la empresa[15], los clientes señalaron que uno de los principales motivos de inasistencia a las mantenciones son que consideran que poseen un precio muy elevado.

Tabla 8.14: Resultados modelos de propensión de asistencia a 2da mantención.

Modelo	Accuracy	Sensibilidad	Especificidad	Precisión	Puntuación F1
LOGIT	0,68	0,54	0,85	0,81	0,65
Random Forest	0,69	0,60	0,79	0,76	0,67
SVM	0,68	0,49	0,89	0,83	0,62
ANNs	0,65	0,82	0,47	0,63	0,71
GBM	0,64	0,85	0,41	0,61	0,71

Para el caso de los resultados de la segunda mantención que se presentan en la Tabla 8.14, se puede observar que el modelo que mejor desempeño presenta en cuanto a sensibilidad es el de potenciación del gradiente (GBM), estando casi al mismo nivel que

las redes neuronales, tal como se observa en la curva ROC de la Figura 8.9. Las máquinas de soporte vectorial (*SVM*) como la regresión logística (LOGIT) son los modelos que peores resultados presentan al momento de identificar los casos reales positivos del total, y a la vez, se puede observar que los modelos con mayor sensibilidad poseen menor especificidad y viceversa. En cuanto a accuracy, todos los algoritmos presentaron valores entre 60 y 70, por lo que el rendimiento en este aspecto es similar entre los modelos. Como nuestro mayor interés es la sensibilidad el modelo más adecuado en este caso a utilizar sería *potenciación de gradiente*, ya que permitiría reconocer con mayor seguridad los casos de alta probabilidad de asistencia.

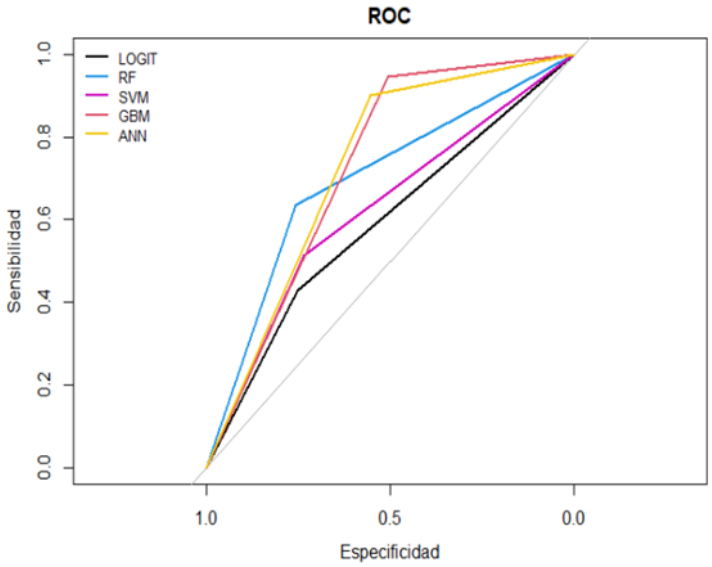


Figura 8.9: Curva ROC

El hecho de que el algoritmo de Potenciación del gradiente sea el que mejor resultados entrega ya que el proceso de *boosting* que realiza, permite ir mejorando los errores de forma secuencial, llegando cada vez a mejores resultados; es decir, sería como realizar una búsqueda de grilla pero ocurre de forma automática dentro de este algoritmo.

8.4.3.2. Tercera mantención

Para la tercera mantención, las variables que toman mayor importancia en el modelo de potenciación de gradiente y bosques aleatorios, son las variables de asistencia a mantenciones previas, luego *recency*, seguido de la región de vivienda del cliente y el modelo del vehículo. El resto de variables presentan casi nula importancia con respecto a las variables anteriormente mencionadas; dentro de estas se encuentra nuevamente el género del cliente, el nivel socio-económico, la edad y el Tipo de comuna.

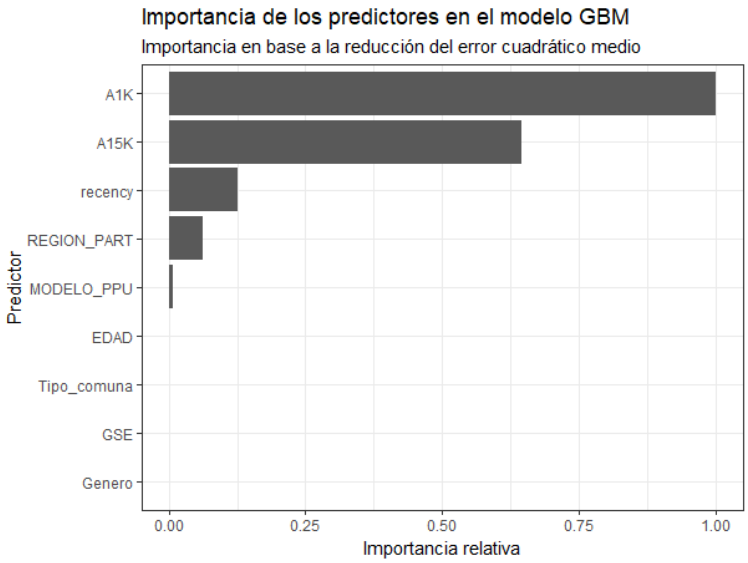


Figura 8.10: Importancia variables en modelos tercera mantención

A partir de los resultados de la aplicación de algoritmos, se obtiene la curva ROC de sensibilidad versus especificidad (Figura 8.11), observando que la curva que mayor área abarca es la de potenciación de gradiente (GBM) y redes neuronales (ANN).

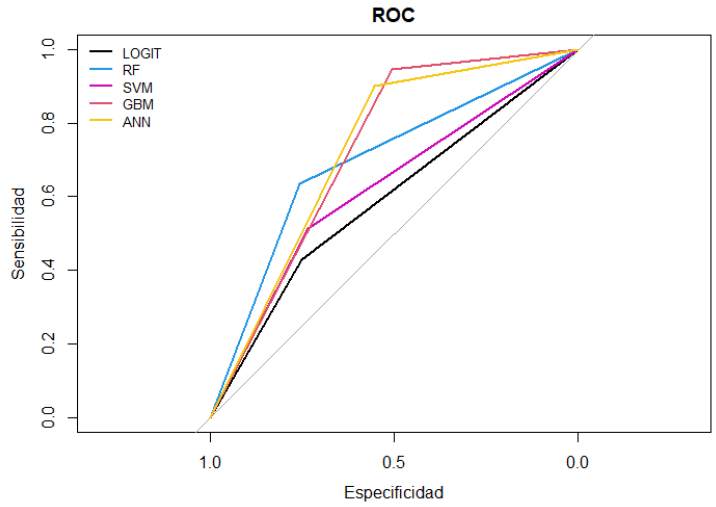


Figura 8.11: Curva ROC en modelos de tercera mantención

Esto se confirma con los resultados numéricos presentados en la Tabla 8.15, donde se observa que el modelo de mejor desempeño en cuanto a sensibilidad, al igual que en la segunda mantención, es *Potenciación del gradiente*, seguido por las redes neuronales y superando con creces al resto de algoritmos utilizados. El nivel de sensibilidad, al ser del 94 % es muy alto, por lo que se cree que podría existir cierto sobreajuste del modelo, considerando que esta es una de las falencias del algoritmo de boosting. Aún así, y al igual que en la segunda mantención, el nivel de especificidad es el mas bajo para los modelos con mayor sensibilidad, por lo que se debe optar por uno de estos parámetros para determinar cuál es el más adecuado. Los modelos presentan un *Accuracy* promedio entre 60 y 70 %, al igual que en el caso de la segunda mantención. Por tanto, y de la mano

con los resultados del parámetro de puntuación F1, el modelo con mejor desempeño alineado a los intereses de la empresa sería *potenciación del gradiente*.

Tabla 8.15: Resultados Modelo propensión asistencia a tercera mantención

Modelo	Accuracy	Sensibilidad	Especificidad	Precisión	Puntuación F1
LOGIT	0.62	0.43	0.75	0.53	0.48
Random Forest	0.71	0.63	0.75	0.63	0.63
SVM	0.65	0.51	0.74	0.56	0.54
ANNs	0.69	0.91	0.55	0.57	0.70
GBM	0.69	0.94	0.52	0.56	0.71

8.4.3.3. Cuarta mantención

Al analizar las variables que mejor importancia le entregan los modelos de bosques aleatorios y potenciación de gradiente, en la Figura 8.12, se observa que solo son 4 las variables que poseen importancia significativa para la clasificación, siendo estas *A30K, recency, A1K y MODELO_PPU*. El resto de variables es considerada con importancia nula para este set de datos. Incluso, se esperaba que tuviese mayor incidencia de importancia la variable de asistencia a la segunda mantención (A15K) que la variable de asistencia a la primera mantención (A1), pero estos resultados indican lo contrario. Al igual que en los sets de datos anteriores, las variables de mayor importancia son las que se relacionan con la asistencia a las mantenciones y no las variables socio-demográficas presentes.

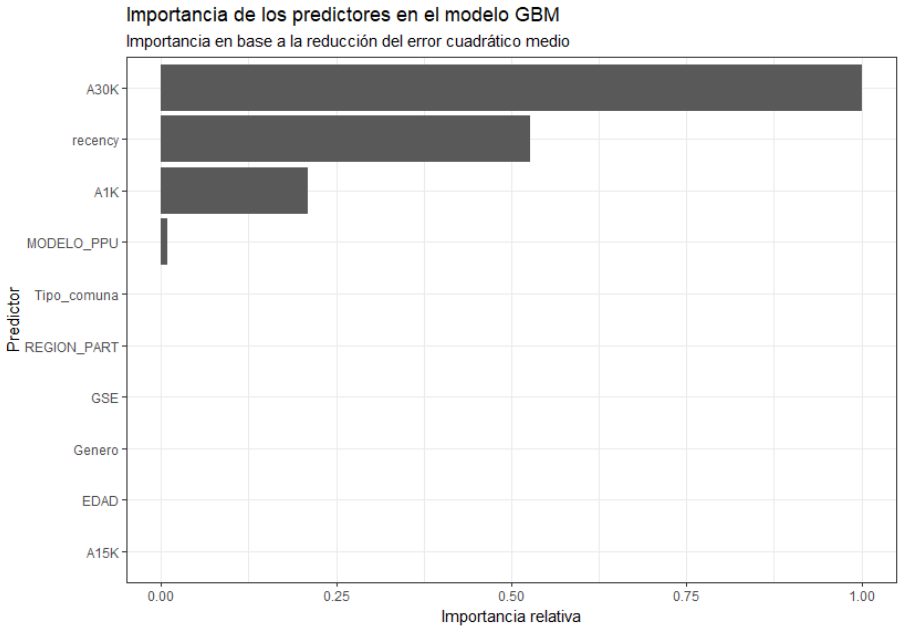


Figura 8.12: Importancia variables modelos cuarta mantención

En el caso de la cuarta mantención, además se presentar muy buenos resultados tanto potenciación del gradiente como redes neuronales, destaca el modelo de bosques aleatorios, el que, como se puede observar en la Figura 8.13 presenta una curva muy

cercana a los otros dos modelos. Se observa también que en este caso los resultados de los parámetros en estudio son bastante parejos, a diferencia de lo ocurrido en los casos presentados para la segunda y tercera mantención. Aún así, en este caso el que mejor resultados presenta, en la Tabla 8.16, nuevamente es potenciación del gradiente, y en cuanto a especificidad el modelo de bosques aleatorios. Es interesante observar que ambos modelos trabajan en base a árboles de decisión, por un lado los bosques aleatorios que se basan en la creación de múltiples árboles de forma paralela, y por otro, potenciación del gradiente crea árboles de forma secuencial basándose en los errores de los árboles previos.

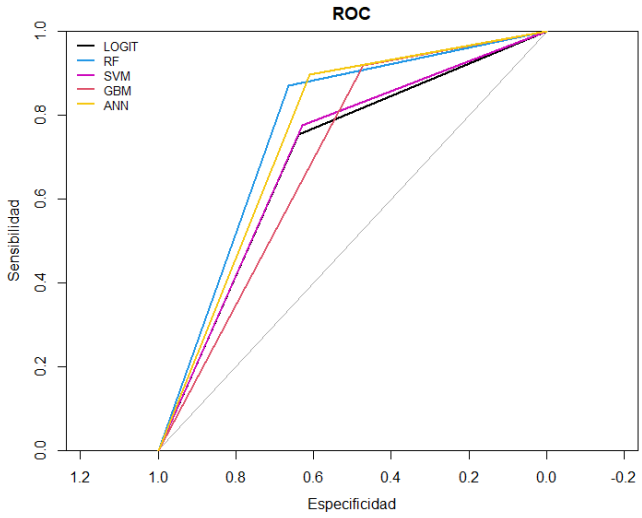


Figura 8.13: Curva ROC modelos cuarta mantención

Tabla 8.16: Resultados Modelo propensión asistencia a cuarta mantención

Modelo	Accuracy	Sensibilidad	Especificidad	Precisión	Puntuación F1
LOGIT	0.67	0.77	0.64	0.37	0.50
Random Forest	0.71	0.86	0.67	0.42	0.56
SVM	0.66	0.78	0.63	0.37	0.50
ANNs	0.60	0.88	0.53	0.34	0.49
GBM	0.69	0.90	0.63	0.40	0.55

8.4.4. Conclusiones modelo de propensión de asistencia

Observando los resultados obtenidos en las tres mantenciones estudiadas, se logra concluir que a pesar de cada uno de estos modelos presentan distintos sets de datos y disponibilidad de variables, finalmente predomina el algoritmo de potenciación de gradiente por sobre el resto de algoritmos aplicados con respecto al desempeño en cuanto a sensibilidad. El desarrollo de este algoritmo es bastante reciente, por lo que su diseño puede ser considerado como una mejora a los modelos de clasificación ya existentes.

Por otra parte, se considera que si bien los modelos consideran las variables de asistencia previa como las de mayor importancia en general, el poder caracterizar a los clientes según su información socio-económica sigue siendo muy relevante y a la vez, también

aporta información útil a los modelos.

Se considera que los resultados obtenidos por estos modelos son suficientemente buenos como para ser utilizados por la empresa, ya que están por sobre el 60 % de precisión general. Además, como se busca predecir la asistencia de los clientes a las mantenciones con la idea de asegurar la asistencia de quienes tienen cierta tendencia a asistir, estos modelos se pueden considerar muy eficientes al presentar una sensibilidad por sobre el 80 % en los tres casos estudiados. Además, considerando que actualmente no existe ningún estudio sobre la propensión de asistencia a mantenciones de parte de la empresa, el integrar estos modelos significaría una mejora general para esta, sobretodo para el área de marketing y post-venta, ya que se podrá enfocar los esfuerzos en asegurar a los clientes que tienen mayor propensión a asistir, y a la vez, se podría hacer esfuerzos en aumentar la probabilidad de asistencia en los clientes que se predice que tienen tendencia a no asistir a las mantenciones. A la vez, cabe destacar que tener mayor claridad en el foco de trabajo proporciona ahorro de tiempo e incluso, puede implicar ahorros en términos económicos también.

Además, en caso de existir intenciones de identificar quienes son propensos a no asistir a una mantenciones, sería recomendable aplicar los modelos que presentar mayor nivel de especificidad, los que serian máquinas de soporte vectorial en el caso de la segunda mantención y random forest para la tercera y cuarta mantención considerando también para estos casos el nivel de accuracy (ya que tanto random forest, como máquinas de soporte vectorial y regresión logística presentan resultados de especificidad similares). Esto sería útil para aplicar políticas de atracción de clientes como promociones que sean muy llamativas para quienes en general tienden a no asistir a las mantenciones.

Capítulo 9

Recomendaciones

Sobre los modelos de predicción de fecha de asistencia se recomienda aplicar aquellos en que se obtuvo un nivel de predicción a intervalos de 60 días cercano o mayor a 50 %, debido a que para el resto está basado en bases de datos muy pequeñas que no generan la suficiente confianza para asegurar su validez. Los modelos de propensión de asistencia se recomienda aplicarlos, ya que presentan resultados confiables y con buenos niveles de predicción en general (cercanos al 80 % en sensibilidad) si la idea es intentar definir quiénes serán quienes asistirán a las mantenciones con mayor probabilidad. Esta recomendación se realiza considerando que la empresa posee el programa *Alteryx* que permite aplicar herramientas de predicción de forma accesible y relativamente fácil si se tiene conocimiento del programa y una leve noción de los algoritmos utilizados, por lo que la aplicación de modelos de predicción mas complejos no requieren de incorporación de nuevas tecnologías o costos extra para la empresa.

En términos momento de aplicación, se recomienda ejecutar cada modelo una vez al mes idealmente con la idea de actualizar los resultados y poder estimar la próxima fecha de asistencia de quienes debieron cumplir en fechas cercanas alguna mantención y haya cambiado la mantención próxima a asistir. Además, considerando que el margen de predicción de fechas de asistencia es más de 60 días en la mayoría de los casos, una buena idea sería contactar al cliente cuando falten entre 60 a 90 días de la fecha predicha de asistencia aproximadamente, considerando que para asistir a una mantención se debe tener en cuenta el gasto que implica asistir al servicio. En cuanto al modelo de propensión, este se debe ejecutar en conjunto con el modelo de predicción de fecha de asistencia, permitiendo así obtener la fecha y probable asistencia o no a la mantención, lo que permitiría tomar acciones como ofrecer promociones a un cliente, o enviarle recordatorios, entre otros.

Para ambos modelos, en los casos en que exista un nuevo modelo de vehículo integrado al mercado, se recomienda no incluirlo en las predicciones a menos que la cantidad de registros de cada modelo superen alrededor de 50 observaciones, de modo que no perjudique la predicción del resto de modelos de auto; lo mismo se debe aplicar para los registros por región.

Aplicados estos modelos, se recomienda tomar acción para diseñar planes de marketing en que se contacte al cliente cuando corresponda, pudiendo ofrecer promociones

y descuentos con el fin de asegurar la asistencia de los clientes (como espera realizar la empresa), y posteriormente evaluar los resultados obtenidos una vez aplicados los modelos. Una vez revisados estos resultados se recomienda crear políticas de atracción de clientes que tienen propensión a no asistir a mantenciones, pero para esto se recomendaría realizar predicciones mediante los algoritmos que presentaron mayor nivel de especificidad, con el fin de aplicar promociones o descuentos para captar clientes.

De la mano con los resultados obtenidos y las cifras observadas, también se recomienda aplicar mayor esfuerzo en la atracción de clientes a las primeras mantenciones, ya que en la mayoría de los casos, quienes tienden a llegar a ir a la cuarta mantención son quienes han asistido a todas las mantenciones previas. Esto es posible que sea provocado debido a la validez de las garantías vehiculares, las que en general se pierden si un cliente no asiste a las mantenciones correspondientes, por lo que se puede perder el incentivo a asistir a una mantención.

Capítulo 10

Trabajo futuro

Como se ha mencionado en este informe, existe mucho trabajo por realizar de parte de la empresa para lograr conocer a su cliente promedio, comenzando por reunir mayor información del cliente e intentar conocer sus preferencias y el motivo de su asistencia o inasistencia, según sea el caso a mantenciones vehiculares u otros servicios de post-venta.

Debido a esto, como trabajo futuro para la empresa queda lo siguiente:

- Mejorar la recopilación de datos sobre asistencia a mantenciones. Como se mencionó anteriormente, el hecho de que los datos registrados en las instancias de mantención (kilometraje registrado, modelo del vehículo, tipo de mantención a la que asiste, entre otros) sean registrados manualmente por los operarios de los concesionarios, provoca un gran nivel de error en la información existente en las bases de datos, por lo que una buena idea sería aplicar formularios uniformes a cada concesionario que permita seleccionar los datos que tengan opciones acotadas y que el sistema de registro verifique la cantidad de cifras de cada registro para que tengan mayor coherencia en general. Esto facilitaría la limpieza de los datos y mejoraría el material para realizar modelos predictivos y el estudio de comportamiento del cliente.

- Recopilar y actualizar información socio-demográfica de los clientes. Los datos que se poseen actualmente por la empresa, en muchos casos datan de casi 4 o 5 años, lo que podría afectar en la aplicación de modelos predictivos al no ser correctos. Además, es necesario añadir un sistema de recopilación de esta información de clientes, lo que puede realizarse por medio del mismo registro de clientes en la nueva aplicación que se está intentando implementar (*MyKia*), pero además podría ser útil el recopilar esta información en las instancias en que el cliente se acerca a los servicios de post-venta o incluso mediante instancias como concursos o encuestas en los que generalmente se recopila información interesante.

- Crear un canal de comunicación más directo con el cliente. Como actualmente se está trabajando en el mejorar el servicio y cambiando el enfoque de la empresa a uno centrado en el cliente, es necesario tener información sobre distintos medios de contacto con los clientes para poder tener canales que permitan avanzar en el estudio de datos de la empresa. Esto daría espacio a realizar estudios sobre el canal de comunicación más adecuado para cada cliente, segmentación de clientes, estudio

de respuesta diferencial (*uplift*), entre otros, los que generalmente permiten mejorar el manejo de recursos de un empresa a la vez de aumentar su demanda.

- Con la idea de mejorar el estudio realizado de propensión de asistencia a mantenciones, se recomienda aplicar a futuro **modelos de respuesta incremental o uplift**. Estos modelos van más allá que los modelos de propensión de asistencia [18][19], permitiendo obtener la *probabilidad* de asistencia de un cliente a un servicio (por ejemplo, una mantención) frente a una acción de marketing (como contacto al cliente o promoción). Así, se crea un grupo de *tratamiento* en que se aplica esta acción de marketing mencionada,y un grupo de *control*: a quienes no se aplica el tratamiento; y luego se observa la *respuesta* del cliente en cuanto a asistencia al servicio. A partir de esto, se puede determinar cuatro segmentos de clientes: *no molestar*, *causa perdida*, *influenciables* y *comprará igualmente*, tal como en el ejemplo señalado en la Figura 10.1.

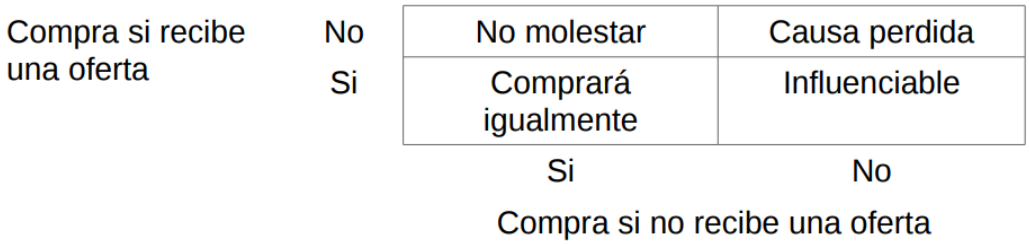


Figura 10.1: Clasificación de uplift frente a campaña publicitaria. Fuente: Espino, C.(2017). [20]

Para estos modelos, como se indica, es **necesario** conocer información sobre el tratamiento aplicado a un cliente, ya que en esta variable y la respuesta obtenida es en que se basa la segmentación de clientes indicada. Considerando el desarrollo de la aplicación de la empresa mencionada anteriormente, no se considera lejana la idea de obtener esta información de forma fácil para la empresa.

Capítulo 11

Conclusiones generales

De los resultados del modelo de predicción de fechas de asistencias a las mantenciones, se pudo ver que los niveles de predicción tienen un error porcentual aproximado promedio de más de 50 %, lo que podría ser considerado relativamente bajo, pero dado el caso de que estas predicciones serían aplicadas a mantenciones que en promedio son una vez cada un año, se considera que los resultados son buenos al tener un margen de error de 4 meses en general, ya que se tiene el espacio para aplicar alguna política de atracción del cliente con el fin de asegurar la asistencia a la mantención (al reservar una hora a mantención), y a la vez, permite a la empresa tener cierta noción de la demanda cada dos meses. Además, al realizar estimaciones que presentan un error cercano a 5 % mayor que los obtenidos por la empresa previamente en cuanto a aciertos en intervalos de días, este trabajo presenta una mejora para la empresa. Además, es muy probable que esta relativamente baja calidad de los modelos diseñados sea producto de la baja cantidad y calidad de los datos e información disponible en la empresa, ya que trabajar con aproximadamente 10 variables presenta poco espacio de creación y optimización del trabajo. Además, también se debe considerar el factor de que producto de la pandemia muchas sucursales en que se realiza mantenciones cerraron , lo que afecta en las fechas de asistencia y kilometraje registrado por quienes asistieron a estas mantenciones.

En cuanto a los algoritmos utilizados para la predicción de fechas de asistencia, se concluye que los que pueden realizar una mejor estimación frente a una baja cantidad de variables son los bosques aleatorios y las máquinas de soporte vectorial, superando a la estimación mediante proporciones utilizada por la empresa.

Para la propensión de asistencia, los modelos poseen un desempeño relativamente bueno al presentar una sensibilidad de 80 % en su mayoría. Estos podrían ser útiles para el desarrollo de campañas de marketing focalizadas que permitan aumentar la asistencia a las mantenciones. Si bien los resultados no son tan precisos para el caso de inasistencia a mantenciones, actualmente no se posee ningún método de predicción de asistencia, y por lo tanto, se considera que en este aspecto cualquier mejora puede ser bien recibida por la empresa.

En cuanto a los modelos realizados para la propensión de asistencia, a pesar de poseer poca información socio-económica del cliente se pudo ver que esto no afectó mayormente, ya que la mayor importancia fue asignada a las variables de asistencia previa

a mantenciones en su mayoría. Además, si bien se obtiene como respuesta una variable binaria de asistencia o no asistencia, sería mucho más útil el poder estimar la probabilidad de asistencia a las mantenciones, lo que permitiría diferenciar entre clientes que no es necesario contactar para que asistan al servicio de la marca, de los que sería muy necesario realizar alguna acción de contacto o *promoción* para que decidan asistir finalmente a las mantenciones.

Se concluye que se logra cumplir con mejorar el sistema de predicción de fecha de próxima asistencia a una mantención, pero aún así se considera relativamente bajo el nivel de predicción entregado, existiendo un gran espacio de mejora a futuro en este ámbito. Además, se considera un logro la predicción de la propensión de asistencia, por lo que su implementación serviría como comienzo para mejorar la captación de clientes y con ello el aumento de demanda del servicio de mantenciones.

Además, como se mencionó anteriormente, siempre existe un margen de mejora en cuanto a la modificación del mismo servicio ofrecido, ya sea de precios o de forma de contacto con el cliente que podría mejorar el nivel de demanda existente. Incluso, el hecho de que los modelos identificados en este trabajo sean efectivos a la hora de estimar la propensión de asistencia depende también del factor económico presente en todo proceso comercial, el que en ocasiones evita que los clientes opten por contratar un servicio.

Bibliografía

[1] Kia Korea. 2021. Misión y visión. [en línea] <https://www.kia.com/kw/experience/local-company/missionvision.html> [consulta: 14 de junio de 2021]

[2] Asociación Nacional Automotriz de Chile A.G. (ANAC). 2021. Conferencia de prensa ANAC 2021. Resultados año 2020 y proyecciones de mercado 2021. [en línea] <https://www.anac.cl/wp-content/uploads/2021/02/CONFERENCIA-DE-PRENSA-ANAC-2021-VF.pdf> [consultado 24 de junio 2021].

[3] Asociación Nacional Automotriz de Chile A.G. (ANAC). 2021. Informe del Mercado Automotor. <https://www.anac.cl/wp-content/uploads/2021/02/12-ANAC-Mercado-Automotor-Diciembre-2020-VF.pdf>

[4] KIA Perú. 2020. Qué es el mantenimiento preventivo de autos y por qué es importante realizarlo. Perú. [en línea] <https://www.kia.com/pe/util/news/que-es-mantenimiento-preventivo-autos-importancia.html> [consulta: 4 de mayo de 2021]

[5] CHILE. Ministerio del Interior. 1979. Ley 3.063: Sobre rentas municipales. Título IX, Artículo 12: De los impuestos municipales.

[6] CHILE. Congreso Nacional. 1496. Ley Sobre rentas municipales. Título IX, Artículo 12: De los impuestos municipales.

[7] Ni Zhang, Yi-Xin Cai, Yong-Yong Wang, Yi-Tao Tian, Xiao-Li Wang, Benjamin Badami, Skin cancer diagnosis based on optimized convolutional neural network, Artificial Intelligence in Medicine, Volume 102, 2020, 101756, ISSN 0933-3657, <https://doi.org/10.1016/j.artmed.2019.101756>.

[8] Weber, Richard, and Miranda, Jaime, and Rey, Pablo. 2005. Predicción de Fugas de Clientes para una institución financiera mediante Support Vector Machines. Revista Ingeniería de Sistemas. Volumen XIX. Páginas 49 - 68. Chile, Universidad de Chile, Departamento de Ingeniería Industrial.

[9] Bennett, K. P. and Campbell, C. (2000). Support vector machines: Hype or hallelujah? SIGKDD Explorations, 2(2). <http://www.acm.org/sigs/sigkdd/explorations/issue2-2/bennett.pdf>.

[10] Barragán, A. 2020. ¿Qué es una red neuronal artificial?. Bélgica. [en línea] <https://cebebelgica.es/es_ES/blog/10/que-es-una-red-neuronal-artificial.html>

[11] Matich, D. 2001. Redes Neuronales: Conceptos Básicos y Aplicaciones. Cátedra: Informática Aplicada a la Ingeniería de Procesos. Argentina, Universidad Tecnológica Nacional, Departamento de Ingeniería Química.

[12] I.A Basheer, M Hajmeer, Artificial neural networks: fundamentals, computing, design, and application, Journal of Microbiological Methods, Volume 43, Issue 1, 2000, Pági-

nass 3-31, ISSN 0167-7012, https://doi.org/10.1016/S0167-7012(00)00201-3. [en línea] <https://www.sciencedirect.com/science/article/pii/S0167701200002013>

[13] Krogh, Anders. (2008). What are artificial neural networks?. Nature biotechnology. 26. 195-7. DOI: 10.1038/nbt1386.

[14] Supratid, Seree. (2003). Application of neural network model in establishing a stage-discharge relationship in tidal river. Hydrological Processes. 17. 3085 - 3099. 10.1002/hyp.1278.

[15] CADEMO. (2021). Costumer Journey Indumotora: Informe Fase Cuantitativa. 198 p.

[16] Gervilla García, Elena, and Jiménez López, Rafael, and Montaño Moreno, Juan José, and Sesé Abad, Albert, and Cajal Blasco, Berta, and Palmer Pol, Alfonso (2009). La metodología del Data Mining. Una aplicación al consumo de alcohol en adolescentes. Adicciones, 21(1),65-80.[fecha de Consulta 25 de Noviembre de 2021]. ISSN: 0214-4840. Disponible en: <https://www.redalyc.org/articulo.oa?id=289122882009>

[17] Momparler, Alexandre and Carmona, Pedro and Climent, Francisco. (2016). Banking failure prediction: a boosting classification tree approach. Revista Española de Financiación y Contabilidad. 45. 1-29. 10.1080/02102412.2015.1118903.

[18] Siegel, E. 2016 Predictive Analytics: The power to predict who will click, buy, lie or die. Hoboken, Nueva Jersey. John Wiley and Sons, Inc.

[19] Lo, Víctor. 2014. The True Lift Model - A Novel Data Mining Approach to Response Modeling in Database Marketing. [en línea] Boston, Estados Unidos. <https://www.researchgate.net/publication/220520042_The_True_Lift_Model_-_A_Novel_Data_Mining_Approach_to_Response_Modeling_in_Database_Marketing> [23 de junio 2021].

[20] Espino T., Carlos. 2017. Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso. Trabajo de fin de Grado en Ingeniería Informática. España, Universitat Oberta de Catalunya. 65p.

[21] Moine, J., Haedo, A., Gordillo S. 2011. Estudio comparativo de metodologías para minería de datos. Argentina, Grupo de investigación en Minería de Datos, UTN Rosario. Facultad de Ciencias Exactas, Universidad Nacional de Buenos Aires. Facultad de Informática, Universidad Nacional de La Plata.

[22] Villena, F. (marzo, 2019). *Obtener la Edad de un Individuo desde su RUT*. Link: https://fabianvillena.cl/blog/obtener-la-edad-de-un-individuo-desde-su-rut/

[23] Berlanga, V., Rubio, M., Vilà, R. 2013. Cómo aplicar árboles de decisión en SPSS. [En línea] REIRE, Revista d'Innovació i Recerca en Educació, 6 (1), 65-79. [en línea] <http://www.ub.edu/ice/reire.htm>

[24] Modelos de árboles de decisión. 2021. IBM. Recuperado el 5 octubre de 2021 de <https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=trees-decision-tree-models>

[25] Piñeiro A., José. 2011. Modelo de propensión al abandono en el canal Horeca. Memoria para optar al Título de Máster Técnicas Estadísticas. Coruña, Universidad de Vigo, Universidad de Coruña y Universidad de Santiago de Compostela 79p. <http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_1679.pdf>

Anexos

Anexo A. Antecedentes

A.1. Participación de mercado automotriz

	Vehículo de Pasajeros		SUV		Camioneta		Vehículo Comercial		Uni	%
	Uni	%	Uni	%	Uni	%	Uni	%		
1 CHEVROLET	15.362	17,1%	5.155	5,3%	4.948	11,0%	1.800	6,8%	27.265	10,5%
2 KIA	12.269	13,6%	4.559	4,7%			2.359	8,9%	19.187	7,4%
3 SUZUKI	15.353	17,1%	3.108	3,2%			280	1,1%	18.741	7,2%
4 NISSAN	5.050	5,6%	8.228	8,5%	4.736	10,5%	172	0,6%	18.186	7,0%
5 HYUNDAI	8.192	9,1%	5.391	5,5%			3.776	14,3%	17.359	6,7%
6 TOYOTA	3.258	3,6%	4.859	5,0%	6.044	13,4%	60	0,2%	14.221	5,5%
7 PEUGEOT	3.198	3,6%	4.315	4,4%	5	0,0%	4.265	16,1%	11.783	4,6%
8 MG	3.958	4,4%	6.829	7,0%					10.787	4,2%
9 MITSUBISHI	90	0,1%	2.486	2,6%	7.808	17,3%			10.384	4,0%
10 FORD	207	0,2%	4.671	4,8%	4.603	10,2%	760	2,9%	10.241	4,0%
11 MAZDA	2.925	3,3%	4.909	5,0%	1.741	3,9%			9.575	3,7%
12 VOLKSWAGEN	4.665	5,2%	2.246	2,3%	2.324	5,1%	138	0,5%	9.373	3,6%
13 CHERY	1.030	1,1%	6.048	6,2%					7.078	2,7%
14 CHANGAN	186	0,2%	5.198	5,3%	192	0,4%	1.150	4,3%	6.726	2,6%
15 CITROEN	2.213	2,5%	1.467	1,5%			2.072	7,8%	5.752	2,2%
16 RENAULT	2.485	2,8%	1.709	1,8%	651	1,4%	568	2,1%	5.413	2,1%
17 JAC	338	0,4%	3.339	3,4%	1.104	2,4%	607	2,3%	5.388	2,1%
18 SUBARU	1.218	1,4%	3.217	3,3%					4.435	1,7%
19 GREAT WALL	781	0,9%	1.654	1,7%	1.951	4,3%			4.386	1,7%
20 RAM					2.052	4,5%	1.613	6,1%	3.665	1,4%
21 MAXUS					2.742	6,1%	907	3,4%	3.649	1,4%
22 MERCEDES BENZ	857	1,0%	901	0,9%	383	0,8%	1.447	5,5%	3.588	1,4%
23 SSANGYONG	72	0,1%	859	0,9%	2.134	4,7%			3.065	1,2%
24 HONDA	486	0,5%	1.893	1,9%	184	0,4%			2.563	1,0%
25 BRILLIANCE			2.270	2,3%			3	0,0%	2.273	0,9%
26 BMW	981	1,1%	1.249	1,3%					2.230	0,9%
27 FIAT	1.647	1,8%	114	0,1%	3	0,0%	372	1,4%	2.136	0,8%
28 JEEP			1.859	1,9%	4	0,0%			1.863	0,7%
29 MAHINDRA	652	0,7%	468	0,5%	565	1,3%			1.685	0,7%
30 DFM	333	0,4%	1.170	1,2%			1	0,0%	1.504	0,6%

Figura A.1: Ventas a público de vehículos livianos acumuladas por marca Año 2020

		Vehículo de Pasajeros		SUV		Camioneta		Vehículo Comercial		Uni	%
		Uni	%	Uni	%	Uni	%	Uni	%		
31	HAVAL			1.470	1,5%					1.470	0,6%
32	FOTON					30	0,1%	1.425	5,4%	1.455	0,6%
33	KYC							1.405	5,3%	1.405	0,5%
34	DFSK			711	0,7%			597	2,3%	1.308	0,5%
35	OPEL	276	0,3%	523	0,5%			283	1,1%	1.082	0,4%
36	VOLVO	266	0,3%	633	0,7%					899	0,3%
37	AUDI	393	0,4%	473	0,5%					866	0,3%
38	SEAT	252	0,3%	452	0,5%					704	0,3%
39	BAIC			702	0,7%			1	0,0%	703	0,3%
40	FAW	2	0,0%	309	0,3%			124	0,5%	435	0,2%
41	GEELY	211	0,2%	205	0,2%					416	0,2%
42	DODGE	7	0,0%	365	0,4%			1	0,0%	373	0,1%
43	SKODA	251	0,3%	97	0,1%					348	0,1%
44	ZXAUTO					300	0,7%			300	0,1%
45	LAND ROVER	1	0,0%	249	0,3%					250	0,1%
46	MINI	186	0,2%	47	0,0%					233	0,1%
47	PORSCHE	39	0,0%	158	0,2%					197	0,1%
48	DS	9	0,0%	148	0,2%					157	0,1%
49	LEXUS	19	0,0%	127	0,1%					146	0,1%
50	LIFAN			123	0,1%			10	0,0%	133	0,1%
51	FUSO							123	0,5%	123	0,0%
52	JAGUAR	34	0,0%	68	0,1%					102	0,0%
53	ZNA	2	0,0%			90	0,2%			92	0,0%
54	ALFA ROMEO	27	0,0%	35	0,0%					62	0,0%
55	IVECO							53	0,2%	53	0,0%
56	MASERATI	6	0,0%	27	0,0%					33	0,0%
57	FERRARI	8	0,0%							8	0,0%
58	JETOUR			5	0,0%					5	0,0%
59	GAC GONOW							3	0,0%	3	0,0%
60	CHRYSLER			1	0,0%			1	0,0%	2	0,0%
61	BENTLEY	1	0,0%	1	0,0%					2	0,0%
OTROS		98	0,1%	230	0,2%	550	1,2%	91	0,3%	969	0,4%
TOTAL ACUMULADO		89.894	100%	97.330	100%	45.144	100%	26.467	100%	258.835	100%

Figura A.2: Ventas a público de vehículos livianos acumuladas por marca Año 2020 - parte 2

A.2. Extracto de metodología de contacto con el cliente

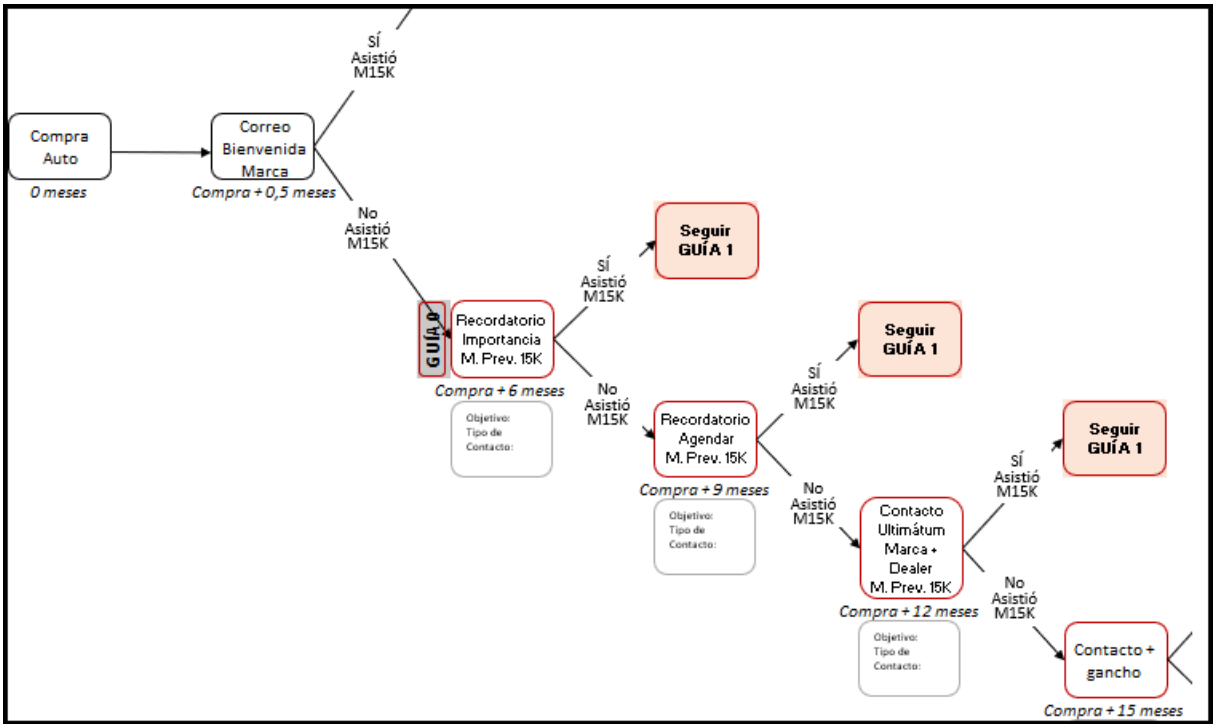


Figura A.3: Extracto de Metodología de contacto con el cliente

Anexo B. Metodología

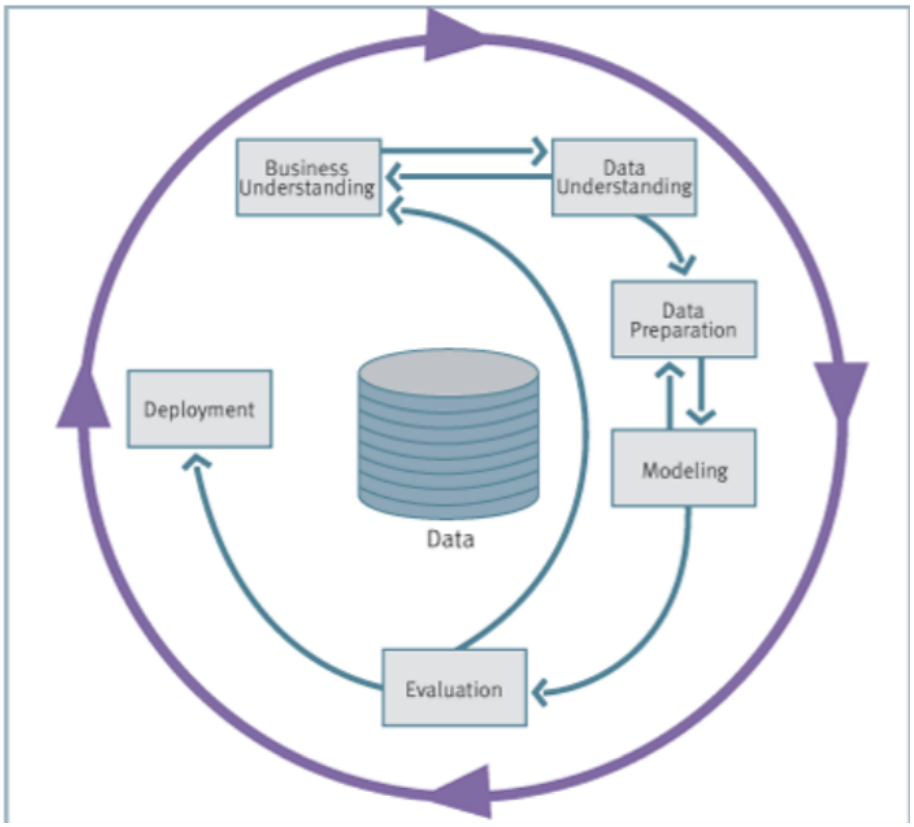


Figura B.1: Metodología CRISP-DM

Figura B.2: Paso a paso metodología CRISP-DM

Anexo C. Preparación de datos

Región\Data	D30M3	D30M15	D30M1	M30M0	D15M0	D15M1	D45M4	D45M30	D45M15	D45M0
Aysen Del General Carlos Ibanez	No	No	No	No	No	No	No	No	No	No
De Antofagasta	Sí	Sí	No	Sí	Sí	Sí	Sí	No	Sí	No
De Arica Y Parinacota	No	No	No	No	No	No	No	No	No	No
De Atacama	Sí	Sí	No	No	No	No	No	No	No	No
De Coquimbo	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No	No	No
De La Araucania	Sí	Sí	No	No	Sí	Sí	Sí	No	No	No
De Los Lagos	Sí	Sí	Sí	No	No	Sí	Sí	Sí	No	No
De Los Rios	Sí	Sí	No	No	Sí	Sí	Sí	No	No	No
De Magallanes Y Antartica Chilena	No	No	No	No	No	No	No	No	No	No
De Ñuble	No	Sí	No	Sí	Sí	Sí	Sí	Sí	No	No
De Tarapaca	No	No	No	No	No	Sí	No	No	No	No
De Valparaiso	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No
Del Bio Bio	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí
Del Libertador Bernardo Ohiggins	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No
Del Maule	Sí	Sí	Sí	No	Sí	Sí	Sí	No	Sí	No
Metropolitana De Santiago	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí

Figura C.1: Selección de regiones de Chile utilizadas por subset

Modelo\ Subset	D15 M0	D15 M1	D30 3M	D30 15M	D30 1M	D30 0M	D45 4M	D45 30M	D45 15M	D45 0M
Cadenza	No	No	No	No	No	No	no	no	No	No
Carens	No	No	No	No	No	No	no}	no	No	No
Carnival	Sí	Sí	Sí	Sí	No	Sí	Sí	Sí	Sí	Sí
Cerato	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí
Cerato 5	Sí	Sí	No	No	No	No	No	No	No	No
Koup	No	No	No	No	No	No	No	No	No	No
Mohave	No	No	No	No	No	No	No	No	No	No
Morning	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí
Niro	Sí	Sí	No	no	No	No	No	No	No	No
Optima	Sí	Sí	No	No	No	No	No	No	No	No
Optima Hibrido	No	No	No	No	No	No	No	No	No	No
Quoris	No	No	No	No	No	No	-	no	No	No
Rio 3	No	No	No	No	No	No	no	no	No	No
Rio 4	Sí	Sí	Sí	Sí	Sí	Sís	Sí	Sí	Sí	Sí
Rio 5	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí
Seltos	Sí	Sí	No	no	No	No	No	No	sí	No
Soluto	Sí	Sí	No	Sí	No	No	No	No	No	No
Sonet	Sí	No	No	no	No	No	No	No	No	No
Sorento	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí
Soul	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No	Sí
Sportage	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí

Figura C.2: Selección de modelos de auto utilizadas por subset

Anexo D. Exploración datos

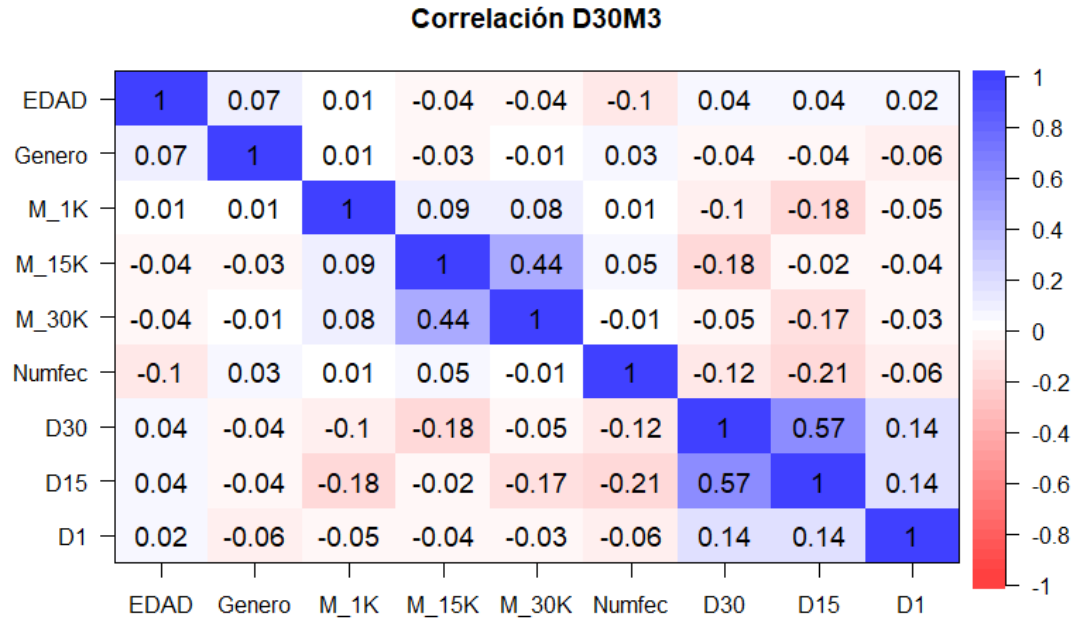


Figura D.1: Correlación variables set de datos D30M3

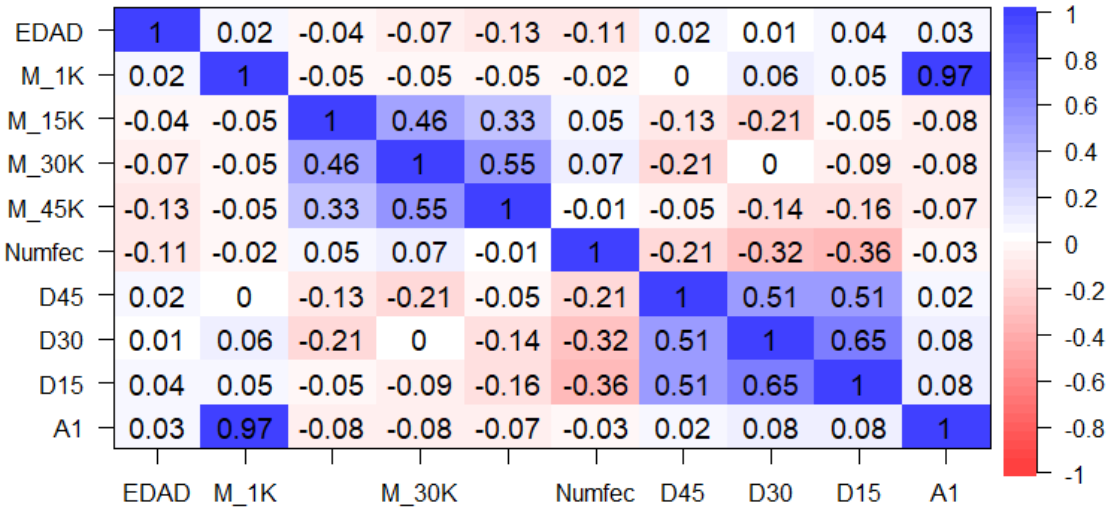


Figura D.2: Correlación variables set de datos D45M4

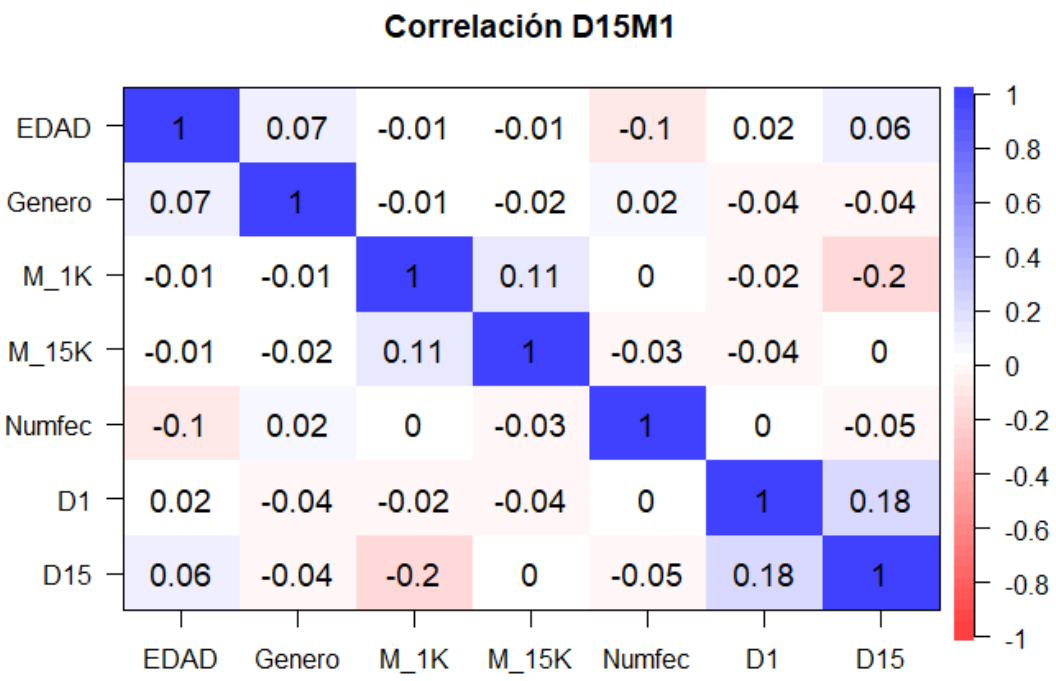


Figura D.3: Correlación variables set de datos D15M1

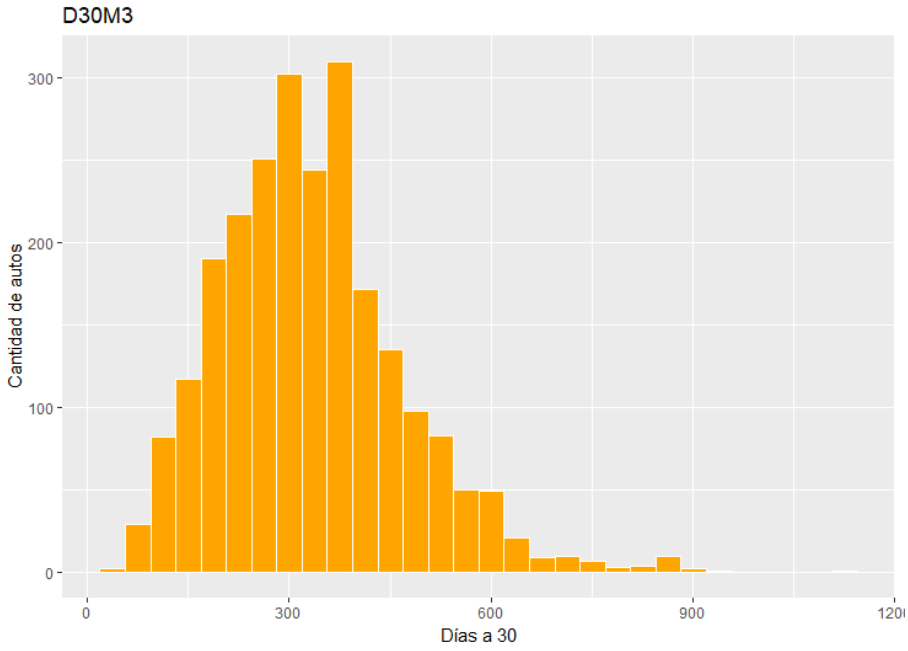


Figura D.4: Distribución días a tercera mantención set de datos D30M3

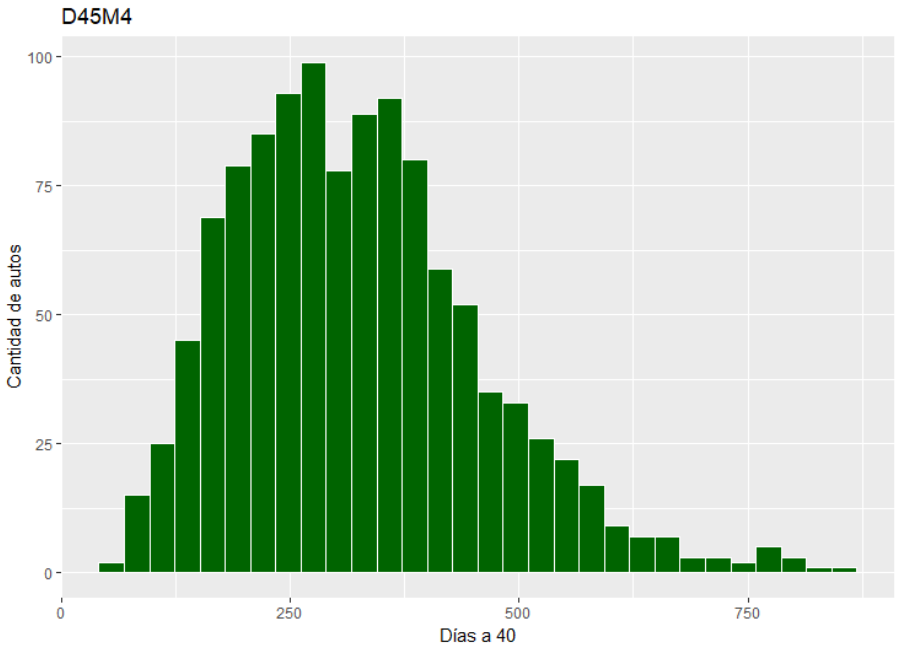


Figura D.5: Distribución días a cuarta mantención set de datos D45M4

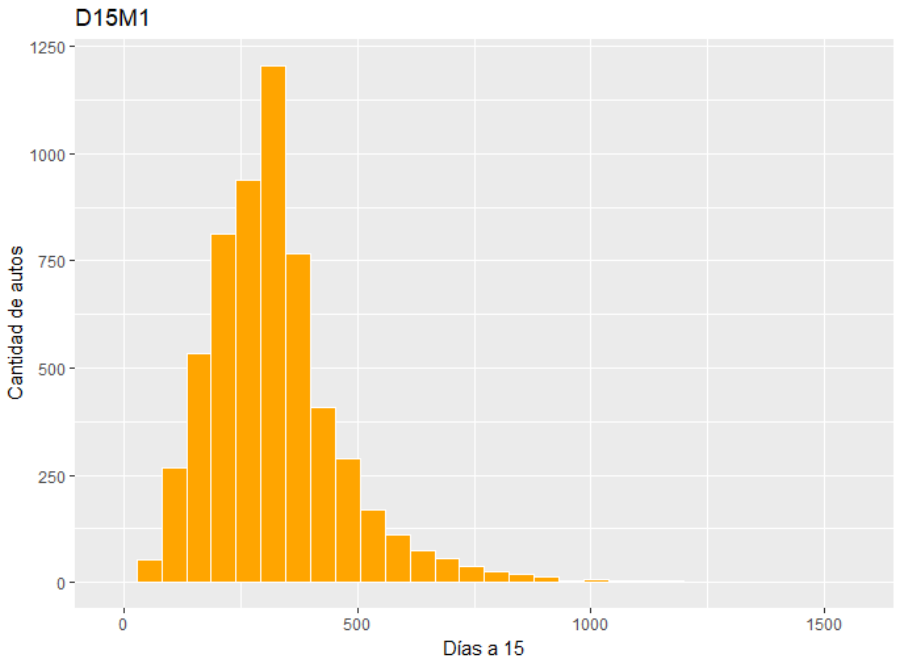


Figura D.6: Distribución días a segunda mantención set de datos D15M1

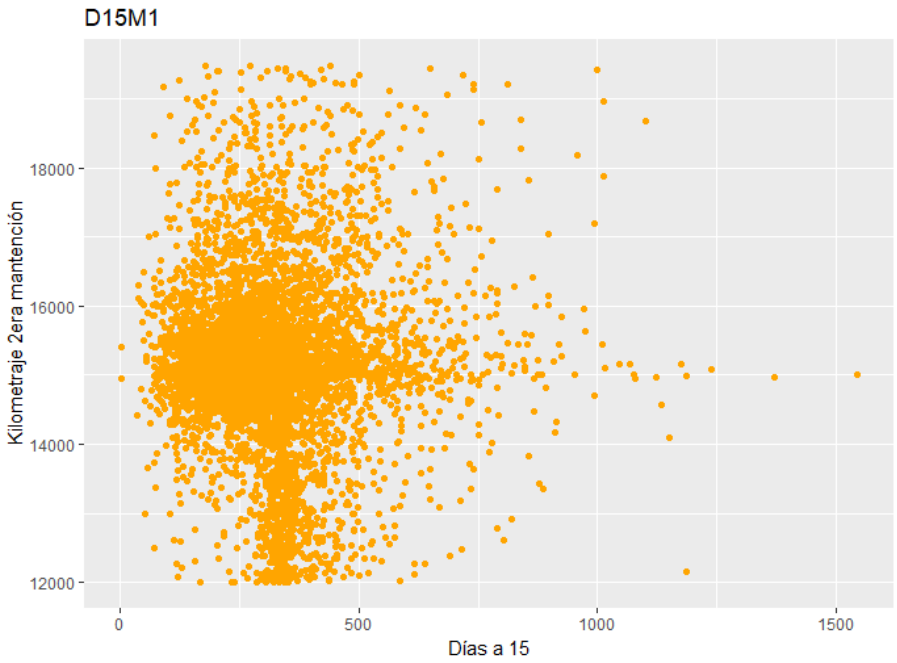


Figura D.7: Días vs kilometraje a segunda mantención set de datos D15M1

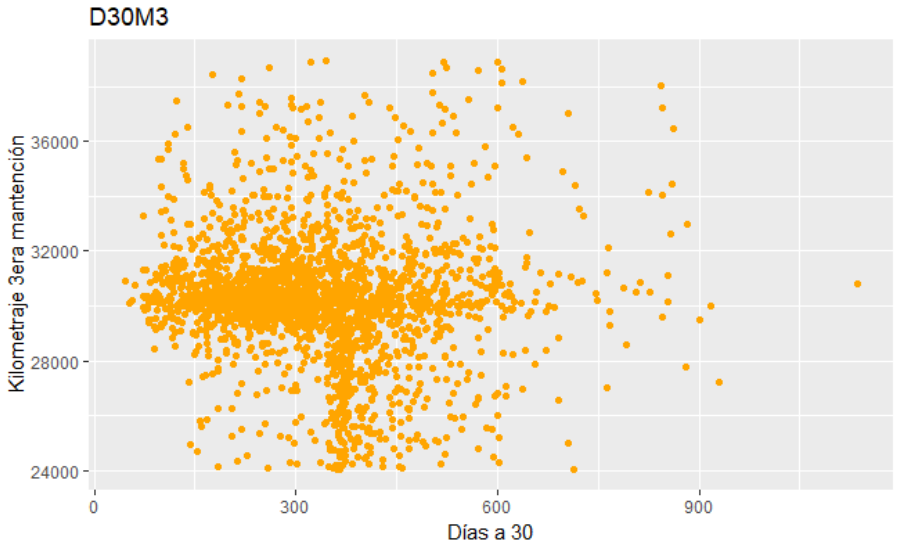


Figura D.8: Días vs kilometraje a tercera mantención set de datos D30M3

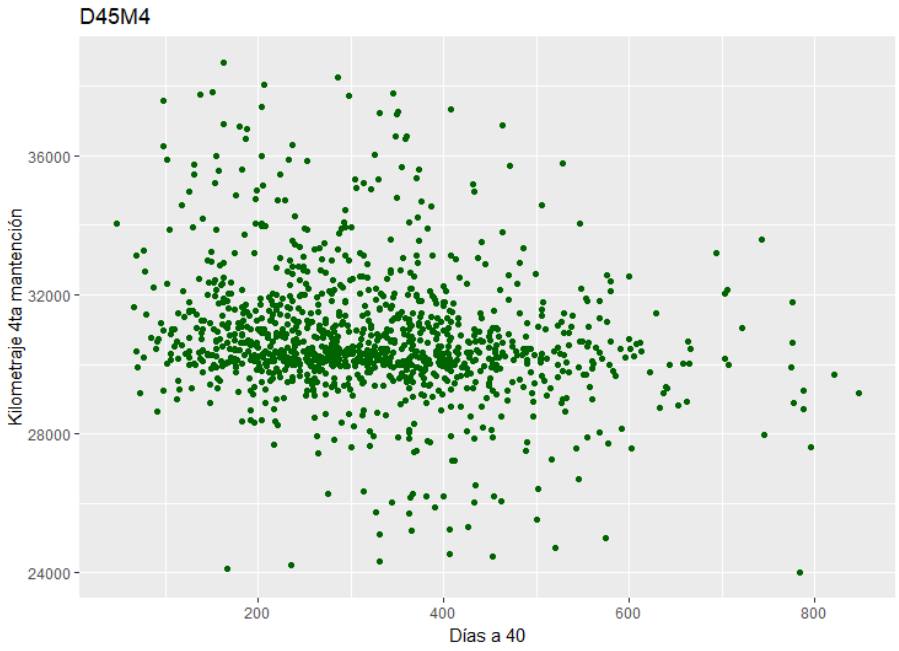


Figura D.9: Días vs kilometraje a cuarta mantención set de datos D45M4

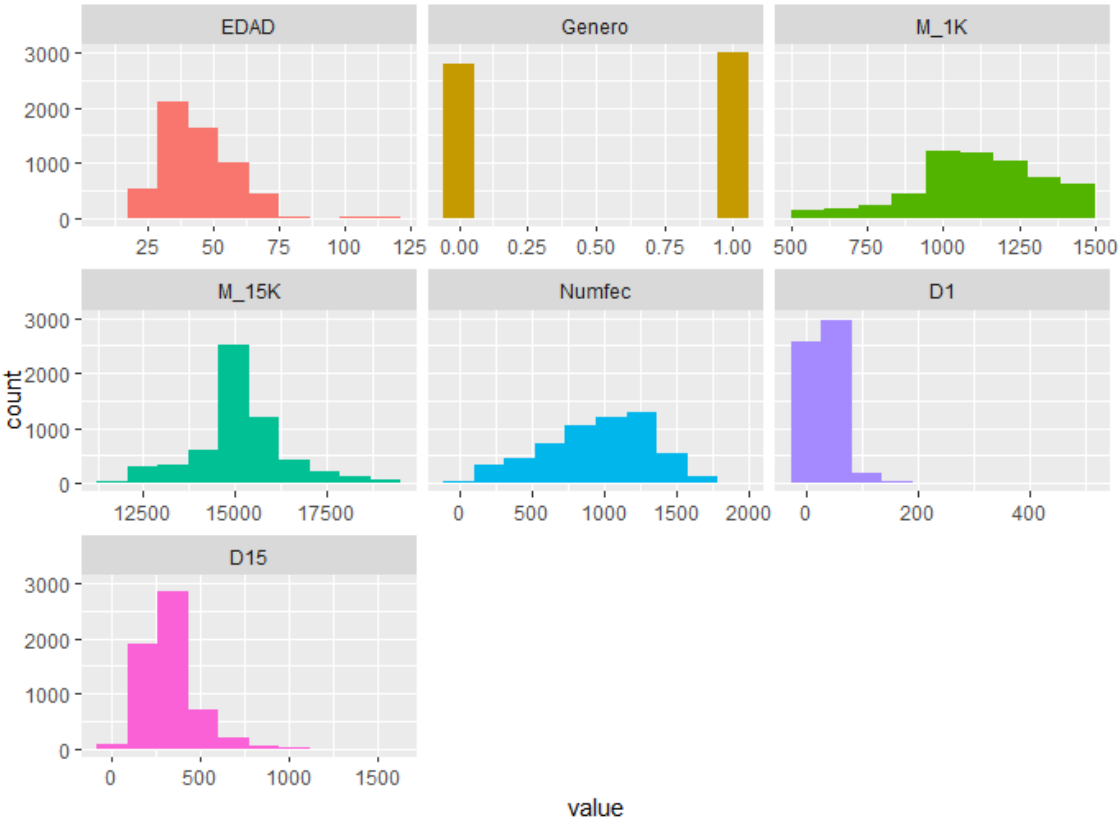


Figura D.10: Distribución variables D15M1

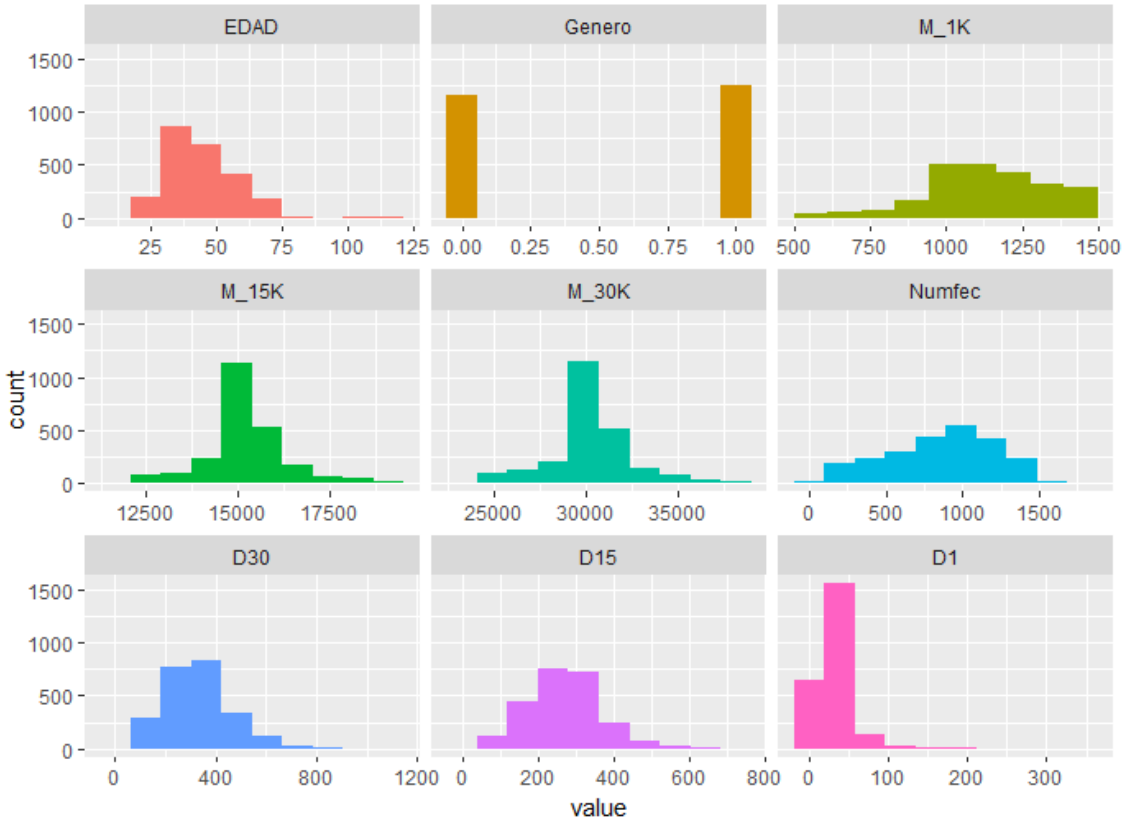


Figura D.11: Distribución variables D30M3

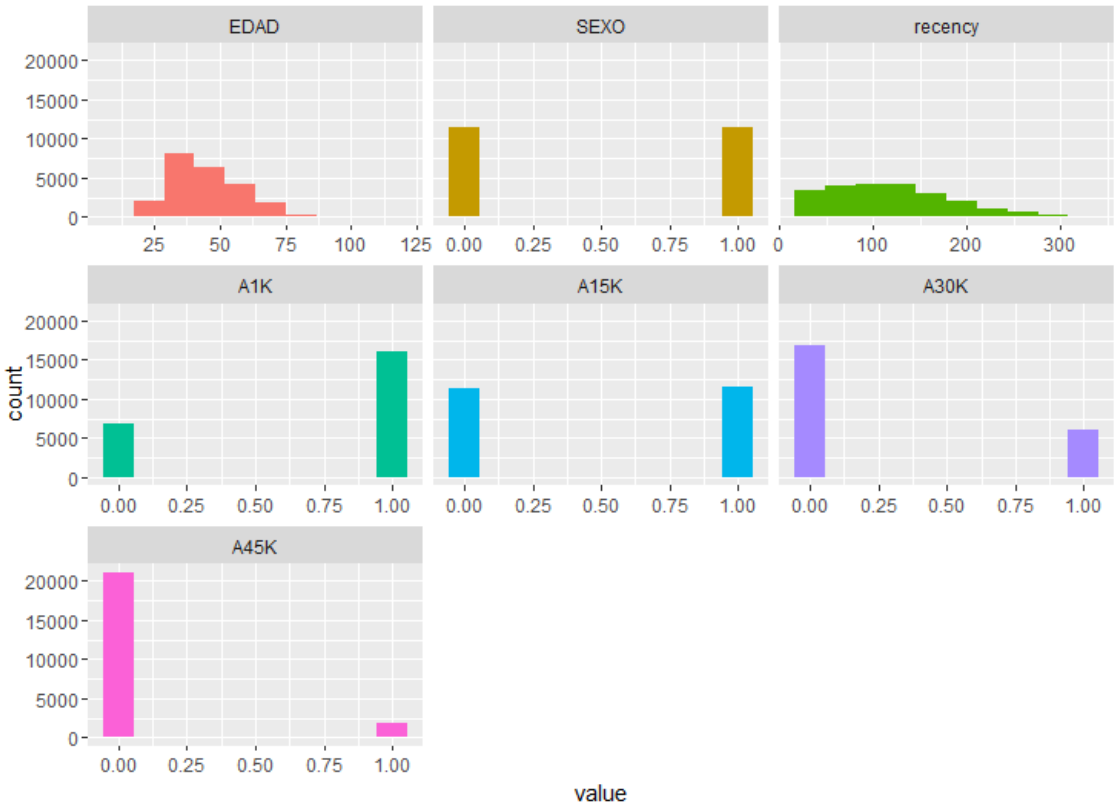


Figura D.12: Distribución variables D45M4

Anexo E. Resultados modelos de predicción de fecha de asistencia

E.1. Resultados Regresión lineal para predicción de fecha de asistencia

Tabla E.1: Resultados regresión modelo D30M3

Coeficientes	Estimate	Std. Error	t value	Pr(>|t|)	Sign.
(Intercept)	465.427199	50.341639	9.245	<2e-16	***
D1	0.367626	0.103543	3.55	0.000394	***
M_1K	0.019607	0.012392	1.582	0.11377	
D15	0.809481	0.026701	30.317	<2e-16	***
M_15K	-0.023036	0.002212	-10.413	<2e-16	***
REGION_PARTDE ATACAMA	-4.963595	38.570105	-0.129	0.897616	
REGION_PARTDE COQUIMBO	-21.642245	29.303102	-0.739	0.460263	
REGION_PARTDE LA ARAUCANIA	14.287783	40.411932	0.354	0.723713	
REGION_PARTDE LOS LAGOS	-16.785444	28.861274	-0.582	0.560912	
REGION_PARTDE LOS RIOS	-16.895957	34.315381	-0.492	0.622513	
REGION_PARTDE VALPARAISO	-26.402045	26.751093	-0.987	0.323793	
REGION_PARTDEL BIO BIO	-63.311656	33.084873	-1.914	0.055821	.
REGION_PARTDEL LIBERTADOR BERNARDO OHIGGINS	-24.732032	27.70318	-0.893	0.372104	
REGION_PARTDEL MAULE	-38.485304	28.704058	-1.341	0.180159	
REGION_PARTMETROPOLITANA DE SANTIAGO	-37.062618	26.147498	-1.417	0.156518	
MODELO_PPUCERATO	-30.796337	15.931294	-1.933	0.053377	.
MODELO_PPUMORNING	-23.561935	15.355425	-1.534	0.12509	
MODELO_PPURIO 4	-26.049763	15.368057	-1.695	0.090229	.
MODELO_PPURIO 5	-6.183846	15.572536	-0.397	0.691339	
MODELO_PPUSORENTO	-29.728339	16.270836	-1.827	0.067843	.
MODELO_PPUSOUL	7.281951	18.710703	0.389	0.697182	
MODELO_PPUSPORTAGE	-6.519205	14.972624	-0.435	0.663316	
Tipo_comunaUrbana	15.651032	10.947641	1.43	0.15299	
GSEC2	-8.341507	10.634972	-0.784	0.432935	
GSEC3	-6.569377	10.659993	-0.616	0.537794	
GSED	-14.793641	10.145663	-1.458	0.144972	
GSEE	-21.974333	14.897498	-1.475	0.14037	
Genero	-1.191229	4.963682	-0.24	0.810365	
EDAD	0.162219	0.187241	0.866	0.386401	
Numfec	0.00457	0.007291	0.627	0.530868	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
FMx					
Residual standard error: 110.9 on 1890 degrees of freedom					
Multiple R-squared: 0.3613, Adjusted R-squared: 0.3522					
F-statistic: 39.6 on 27 and 1890 DF, p-value: <2.2e-16					

Tabla E.2: Resultados regresión modelo D30M15

Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	**Sign.**
(Intercept)	172.90	32.69	5.29	0.000	***
D15	0.71	0.03	28.32	<2e-16	***
REGION_PARTDE ATACAMA	-61.98	35.15	-1.76	0.078	.
REGION_PARTDE COQUIMBO	-33.97	28.41	-1.20	0.232	
REGION_PARTDE LA ARAUCANIA	-32.66	35.86	-0.91	0.363	
REGION_PARTDE LOS LAGOS	-61.63	26.39	-2.34	0.020	*
REGION_PARTDE LOS RIOS	-50.76	34.84	-1.46	0.145	
REGION_PARTDE NUBLE	-76.69	29.22	-2.63	0.009	**
REGION_PARTDE VALPARAISO	-42.30	23.32	-1.81	0.070	.
REGION_PARTDEL BIO BIO	-61.28	25.03	-2.45	0.014	*
REGION_PARTDEL LIBERTADOR BERNARDO OHIGGINS	-61.20	26.02	-2.35	0.019	*
REGION_PARTDEL MAULE	-51.66	27.94	-1.85	0.065	.
REGION_PARTMETROPOLITANA DE SANTIAGO	-42.33	21.62	-1.96	0.050	.
MODELO_PPUCERATO	-3.40	15.83	-0.22	0.830	
MODELO_PPUMORNING	-0.11	15.12	-0.01	0.994	
MODELO_PPURIO 4	-1.94	15.16	-0.13	0.898	
MODELO_PPURIO 5	18.07	15.01	1.20	0.229	
MODELO_PPUSOLUTO	-67.58	34.91	-1.94	0.053	.
MODELO_PPUSORENTO	15.13	15.32	0.99	0.324	
MODELO_PPUSOUL	-5.80	20.82	-0.28	0.781	
MODELO_PPUSPORTAGE	14.71	14.26	1.03	0.302	
Tipo_comunaUrbana	0.17	14.54	0.01	0.990	
GSEC2	4.02	10.43	0.39	0.700	
GSEC3	-11.03	10.62	-1.04	0.299	
GSED	-20.01	9.86	-2.03	0.043	*
GSEE	-42.61	16.15	-2.64	0.008	**
Genero	-4.74	5.63	-0.84	0.400	
EDAD	-0.04	0.21	-0.21	0.836	
Numfec	-0.01	0.01	-1.14	0.253	
—					
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					
Residual standard error: 118.1 on 1803 degrees of freedom					
Multiple R-squared: 0.3704, Adjusted R-squared: 0.3606					
F-statistic: 37.89 on 28 and 1803 DF, p-value: <2.2e-16					

Tabla E.3: Resultados regresión modelo D30M1

Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	**Sign.**
(Intercept)	775.22	100.16	7.74	1.89E-13	***
D1	1.79	0.39	4.58	7.16E-06	***
REGION_PARTDE LOS LAGOS	-109.04	73.02	-1.49	0.1365	
REGION_PARTDE VALPARAISO	-9.05	69.66	-0.13	0.8968	
REGION_PARTDEL BIO BIO	20.67	80.54	0.26	0.7976	
REGION_PARTDEL LIBERTADOR BERNARDO OHIGGINS	-41.13	85.23	-0.48	0.6298	
REGION_PARTDEL MAULE	-83.73	88.89	-0.94	0.3471	
REGION_PARTMETROPOLITANA DE SANTIAGO	-7.03	63.61	-0.11	0.9121	
MODELO_PPUMORNING	5.87	44.69	0.13	0.8956	
MODELO_PPURIO 4	23.81	46.42	0.51	0.6084	
MODELO_PPURIO 5	55.83	47.31	1.18	0.2389	
MODELO_PPUSORENTO	27.77	71.06	0.39	0.6962	
MODELO_PPUSOUL	56.41	63.16	0.89	0.3725	
MODELO_PPUSPORTAGE	74.31	42.54	1.75	0.0818	.
Tipo_comunaUrbana	-38.83	65.56	-0.59	0.5541	
GSEC2	25.59	51.60	0.50	0.6203	
GSEC3	-9.54	55.88	-0.17	0.8646	
GSED	-7.96	49.47	-0.16	0.8723	
GSEE	52.31	83.17	0.63	0.5299	
Genero	-33.55	25.16	-1.33	0.1834	
EDAD	-0.44	1.00	-0.44	0.6616	
Numfec	-0.16	0.04	-4.21	3.45E-05	***
—					
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					
Residual standard error: 206.3 on 276 degrees of freedom					
Multiple R-squared: 0.1655, Adjusted R-squared: 0.102					
F-statistic: 2.606 on 21 and 276 DF, p-value: 0.0002175					

Tabla E.4: Resultados regresión modelo D30M0

Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	**Sign.**
(Intercept)	847.15	100.50	8.43	2.47E-16	***
REGION_PARTDE COQUIMBO	135.87	84.71	1.60	0.10926	
REGION_PARTDE NUBLE	-96.69	86.45	-1.12	0.26381	
REGION_PARTDE VALPARAISO	-10.97	69.07	-0.16	0.87386	
REGION_PARTDEL BIO BIO	-35.77	62.79	-0.57	0.56909	
REGION_PARTDEL LIBERTADOR BERNARDO OHIGGINS	79.91	78.91	1.01	0.31157	
REGION_PARTMETROPOLITANA DE SANTIAGO	4.48	58.12	0.08	0.93853	
MODELO_PPUCERATO	11.21	46.09	0.24	0.80795	
MODELO_PPUMORNING	34.31	43.69	0.79	0.43262	
MODELO_PPURIO 4	-43.79	45.45	-0.96	0.33573	
MODELO_PPURIO 5	47.22	43.62	1.08	0.27944	
MODELO_PPUSORENTO	-31.41	44.65	-0.70	0.48208	
MODELO_PPUSOUL	74.88	56.57	1.32	0.18609	
MODELO_PPUSPORTAGE	19.40	42.09	0.46	0.64492	
Tipo_comunaUrbana	40.54	56.13	0.72	0.47044	
GSEC2	-11.84	31.97	-0.37	0.71125	
GSEC3	-94.90	34.29	-2.77	0.00581	**
GSED	-42.12	31.48	-1.34	0.18145	
GSEE	-71.87	55.95	-1.29	0.19942	
Genero	-32.42	19.13	-1.70	0.09056	.
EDAD	-0.53	0.73	-0.73	0.467	
Numfec	-0.19	0.03	-7.07	4.26E-12	***
—					
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					
Residual standard error: 237.2 on 615 degrees of freedom					
Multiple R-squared: 0.156, Adjusted R-squared: 0.1271					
F-statistic: 5.412 on 21 and 615 DF, p-value: 2.611e-13					

E.2. Predicción fecha de asistencia a segunda mantención

Modelo *D15M1*

Para este set de datos se observa que las variables de mayor importancia son los registros de la asistencia previa, a la primera mantención, mientras que las variables de ruralidad (tipo comuna) y género son las que menos influyen en la estimación de los días a la mantención.

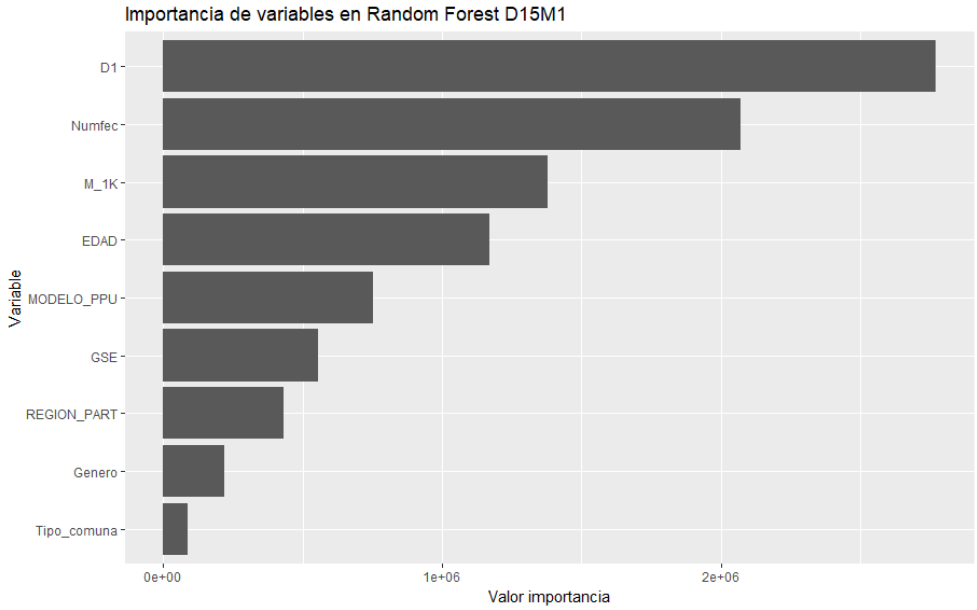


Figura E.1: Importancia variables modelo D15M1 según Random Forest

Como resultado se obtiene lo indicado en la Tabla E.5. Se observa que a pesar de la alta cantidad de observaciones presentes en el set de datos, la estimación de días hasta la asistencia a la segunda mantención posee un error de alrededor de 3 meses y medio con respecto a los valores reales. El modelo que mejor rendimiento presenta es el de Potenciación del Gradiente en cuanto a error absoluto medio, y Máquinas de soporte vectorial en cuanto a menor error medio absoluto porcentual.

Tabla E.5: Resultados regresión segunda mantención modelo D15M1

Modelo	**RMSE**	**MAE**	**MAPE**
Regresión múltiple	142.8	103.7	71 %
Árbol podado	139.5	100.1	70 %
Random Forest	137.6	98.4	75 %
Gradient Boosting	136.2	97.6	68 %
Máquinas de soporte	138.8	97.9	65 %

Tabla E.6: Nivel de precisión de predicción de días por intervalos.

Modelo	**± 15 días**	**± 30 días**	**± 40 días**	**± 60 días**	**>60 días**
Regresión múltiple	11 %	22 %	28 %	40 %	60 %
Árbol podado	10 %	21 %	28 %	40 %	60 %
Random Forest	12 %	23 %	29 %	40 %	60 %
Gradient Boosting	11 %	22 %	28 %	40 %	60 %
Máquinas de soporte	13 %	23 %	30 %	42 %	58 %

Modelo *D15M0*

En este set de datos, las variables de mayor importancia según el algoritmo de bosques aleatorios E.2se tornan la antigüedad del vehículo, la edad del cliente y luego el modelo del auto. Igual que en el set de datos anterior las variables de menor importancia son género y tipo comuna.

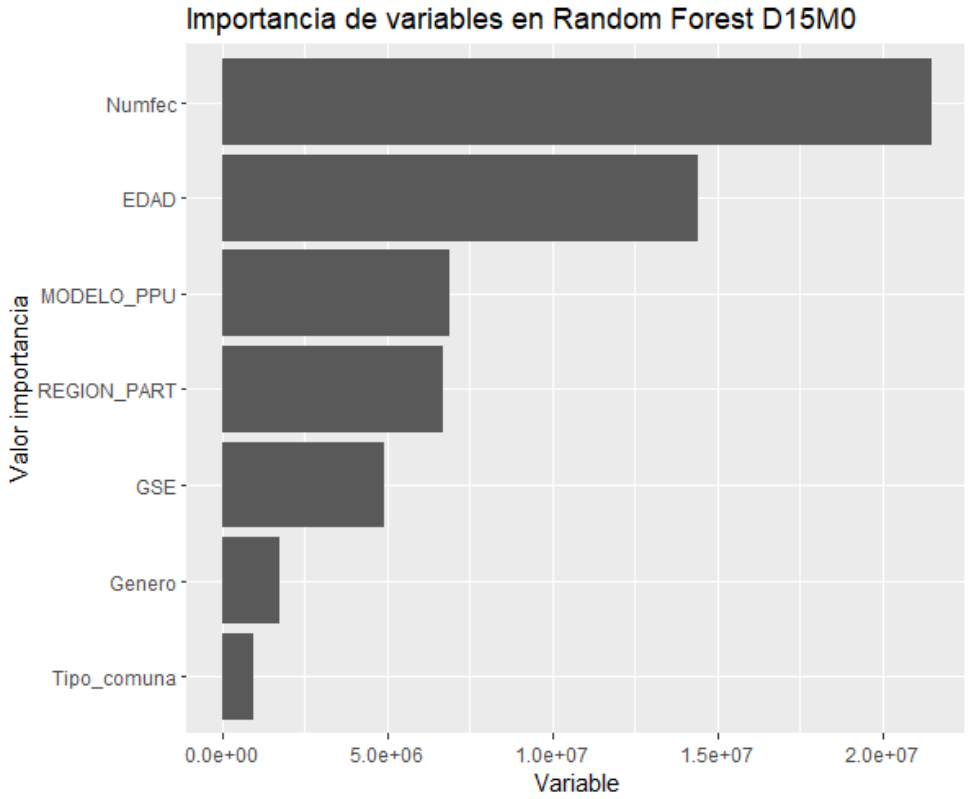


Figura E.2: Importancia variables modelo D15M0 según Random Forest

Se observa en la Tabla E.7 los resultados obtenidos, donde nuevamente destaca la gran cantidad media de días de error en las predicciones. Aún así, se observa que el error absoluto porcentual medio es inferior al 50 % en todos los casos, lo que se podría considerar como buenos resultados en general. Nuevamente los dos modelos que destacan en cuanto a mejor desempeño son Máquinas de Soporte Vectorial y Potenciación del Gradiente.

Tabla E.7: Resultados modelos de regresión, sub-set D15M0

Modelo	RMSE	MAE	MAPE
RL	155.24	115.74	0.45
Árbol de regresión	157.50	116.98	0.47
Random Forest	157.00	116.94	0.46
GBM	153.82	114.19	0.45
SVM	157.45	114.41	0.42

Tabla E.8: Nivel de precisión de predicción de días por intervalos sub-set D15M0.

Modelo	± 15 días	± 30 días	± 40 días	± 60 días	>60 días
Regresión múltiple	10 %	19 %	25 %	34 %	66 %
Árbol podado	10 %	20 %	26 %	35 %	65 %
Random Forest	8 %	17 %	23 %	35 %	65 %
Gradient Boosting	9 %	19 %	25 %	35 %	65 %
Máquinas de soporte	10 %	20 %	27 %	36 %	64 %

E.3. Predicción fecha de asistencia a tercera mantención

Modelo *D30M3*

Como resultado del modelo de árbol de decisión se obtiene que la importancia de las variables es alta para D15, Numfec y D1, pero para el resto de variables es casi insignificante. El modelo de Random Forest a su vez, considera las variables más importantes D15, D1, $Modelo\_ppu$ y Edad, mientras que el resto de variables como GSE, Género y Tipocomuna poseen una significancia muy baja, lo mismo que se aprecia con los resultados de Random Forest.

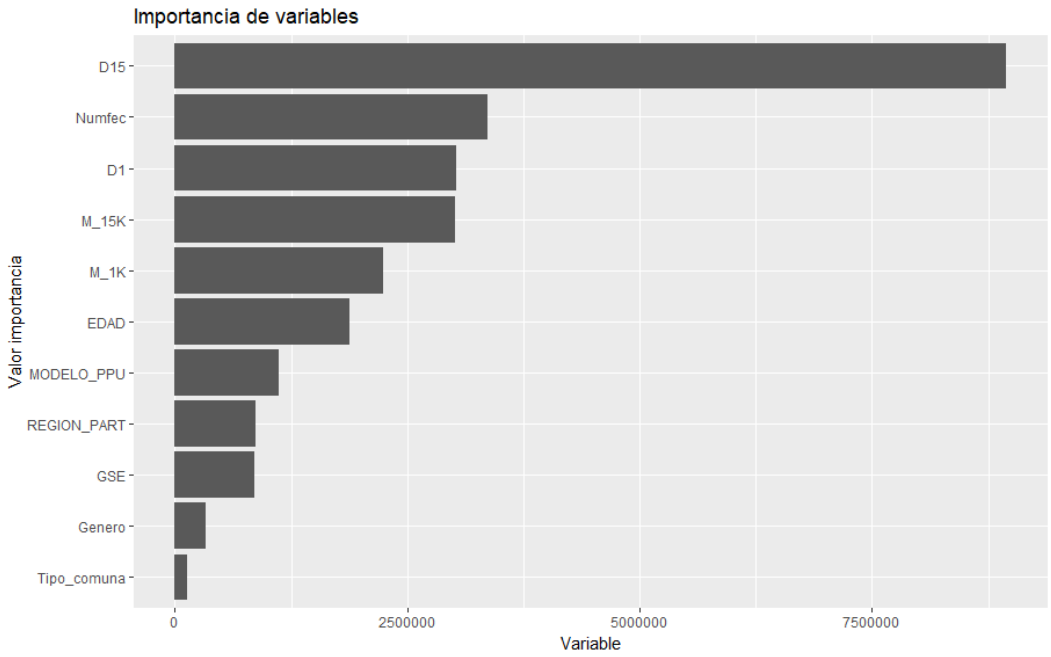


Figura E.3: Importancia variables modelo D30M3 según Random Forest

Se obtienen las métricas de rendimiento de error de los modelos aplicados, adjuntas en la Tabla E.9. Se observa que el modelo de mejor desempeño en cuanto a menor error absoluto medio es Random Forest, presentando un error de 75 días promedio en la predicción de fechas, siendo aproximadamente 2 meses y medio.

Tabla E.9: Resultados predicción fecha de asistencia modelo D30M3

Modelo	RMSE	MAE	MAPE
Regresión lineal	113.57	83.91	0.30
Árbol de regresión	108.67	82.35	0.30
Random Forest	102.83	75.03	0.27
GBM	109.14	79.82	0.28
SVM	108.52	77.50	0.26

Para visualizar mejor la situación actual de predicción de fechas de la empresa contrastado con los resultados obtenidos por el mejor modelo de este caso (Random Forest) se presenta la precisión en rangos de días de diferencia entre las predicciones y valores reales en la Tabla 8.10. Se presenta una leve mejoría con respecto a los métodos de cálculo usados actualmente; solo mejor en un 5 % el intervalo que contiene $\pm$60 días de error.

Tabla E.10: Nivel de precisión de predicción de días por intervalos sub-set D30M3

Modelo	$\pm$ 15 días	$\pm$ 30 días	$\pm$ 40 días	$\pm$ 60 días	>60 días
Regresión múltiple	12 %	26 %	33 %	47 %	53 %
Árbol podado	11 %	25 %	31 %	44 %	56 %
Random Forest	13 %	26 %	37 %	52 %	48 %
Gradient Boosting	14 %	28 %	35 %	48 %	52 %
Máquinas de soporte	12 %	26 %	33 %	47 %	53 %

Modelo *D30M15*

Para analizar la influencia de las variables independientes sobre la variable de interés, se realiza una regresión múltiple, obteniéndose los resultados presentes en la Tabla E.2. Se observa que tanto el intercepto como D15 son significativos al 95 %, mientras que algunos niveles de GSE como algunas regiones también poseen cierto nivel de significancia.

Al estudiar la importancia de las variables, *step AIC* indica que tanto los días hasta la segunda mantención (D15), el nivel socio-económico (GSE) y la fecha correlativa de compra (Numfec) son variables importantes para el modelo. El árbol de regresión indica que las variables consideradas para el modelo son principalmente D15 y Numfec, y en un porcentaje muy bajo GSE, Región, Modelo y Género del cliente. Random Forest a diferencia del árbol de regresión considera la edad también como una variable de importancia, y el resto de variables en una menor medida (Figura E.4).

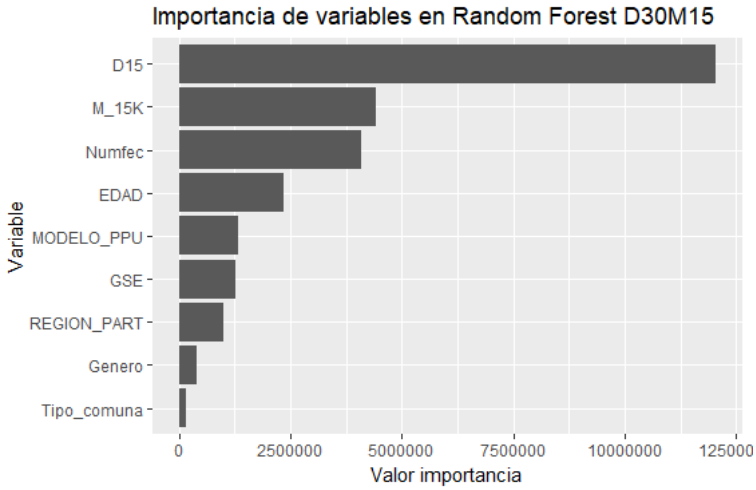


Figura E.4: Importancia variables modelo D30M15 según Random Forest

Los resultados de los modelos evaluados muestran que el nivel de predicción es muy similar entre ellos, lo que indica que en realidad la relativamente baja precisión vendría dada por las variables con las que se está trabajando. Aún así, el modelo que mejor desempeño tiene en este set de datos es las máquinas de soporte vectorial, con un error medio absoluto de 80.3 días, según indica la Tabla E.11. Se obtienen los porcentajes de acierto de este modelo por intervalos de días de error, los que se indican en la Tabla E.12.

Tabla E.11: Resultados predicción fecha de asistencia modelo D30M15

Modelo	**RMSE**	**MAE**	**MAPE**
Regresión lineal	111.71	83.11	0.31
Árbol de regresión	110.35	83.16	0.31
Random Forest	108.39	80.77	0.30
GBM	109.72	82.08	0.30
SVM	109.59	80.29	0.27

Tabla E.12: Nivel de precisión de predicción de días por intervalos sub-set D30M15

Modelo	**± 15 días**	**± 30 días**	**± 40 días**	**± 60 días**	**>60 días**
Regresión múltiple	7 %	15 %	21 %	30 %	70 %
Árbol podado	12 %	23 %	28 %	44 %	56 %
Random Forest	11 %	24 %	33 %	46 %	54 %
Gradient Boosting	11 %	20 %	28 %	47 %	53 %
Máquinas de soporte	13 %	26 %	35 %	50 %	50 %

Modelo *D30M1*

Para el desarrollo de este modelo, al igual que en los casos anteriores se desarrolla un modelo de regresión lineal múltiple y se observa que D1 y Numfec son las únicas

variables significativas del modelo, además del intercepto. Los días hasta la primera mantención afectarían de proporcionalmente la cantidad de días hasta la primera mantención, lo que era de esperar ya que mientras más demoren en realizar la primera mantención, más probable es demorar en realizar las siguientes mantenciones. El modelo posee un ajuste muy bajo, aún cuando es significante al 95 %.

Al aplicar *backward step AIC* se obtiene que el modelo que minimiza el error es el compuesto de las variables D1 y Numfec. Lo mismo indica el árbol de regresión aplicado. Random Forest en la Figura E.5 también considera como las variables más importantes D1 y Numfec. GBM entrega el mismo orden de importancia que Random Forest, aunque considera que la variable Tipo_comuna no posee información relevante al modelo.

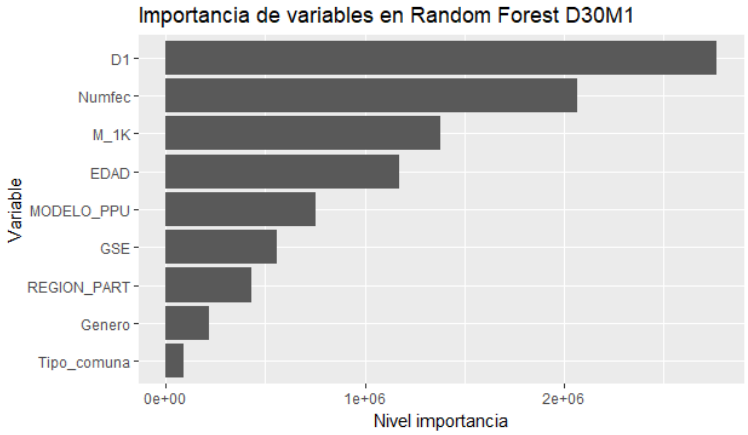


Figura E.5: Importancia variables modelo D30M1 según Random Forest

Los resultados entregados por los modelos utilizados se adjuntan en la Tabla E.13. El rendimiento de los modelos es bastante similar, aunque destaca por una pequeña diferencia nuevamente Random Forest, tanto en error cuadrático medio, error absoluto medio y error porcentual. Aún así el predecir bien un 50 % de los datos deja espacio a mucho error.

Tabla E.13: Resultados modelos de regresión, sub-set D30M1

Modelo	**RMSE**	**MAE**	**MAPE**
Regresión lineal	223.23	171.14	0.50
Árbol de regresión	235.11	183.61	0.48
Random Forest	222.80	167.84	0.43
GBM	222.91	174.84	0.48
SVM	225.16	174.87	0.49

El modelo de mejor rendimiento, en este caso Random Forest, presenta los porcentajes de predicción señalados en la tabla E.14 según cada intervalo de días considerados.

Tabla E.14: Nivel de precisión de predicción de días por intervalos sub-set D30M1

Modelo	± 15 días	± 30 días	± 40 días	± 60 días	>60 días
Regresión múltiple	4 %	11 %	15 %	29 %	71 %
Árbol podado	8 %	15 %	23 %	27 %	73 %
Random Forest	5 %	11 %	17 %	31 %	69 %
Gradient Boosting	11 %	15 %	19 %	25 %	75 %
Máquinas de soporte	7 %	17 %	19 %	25 %	75 %

Modelo *D30M0*

Al realizar la regresión lineal, en la Tabla E.4 se observa que solo la variable Numfec y el nivel socio-económico C3 poseen un grado de significancia para el modelo, además del intercepto.

Se estudia la importancia de las variables seleccionadas para este modelo, donde según el modelo Random Forest, en esta ocasión toman gran importancia las variables de fecha de venta relativa del automóvil, la edad del cliente y el modelo del vehículo principalmente.

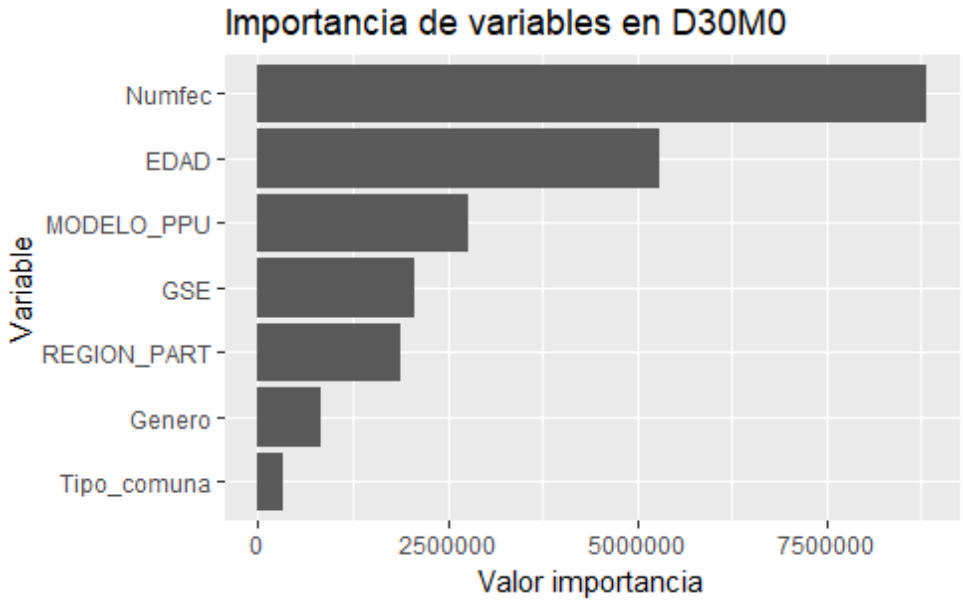


Figura E.6: Importancia variables modelo D30M0 según Random Forest

En cuanto a los resultados de las predicciones, estas fueron bastante similares entre los modelos, al igual que con los sub-sets de datos anteriores. En esta ocasión el modelo que presenta mejores resultados es SVM considerando el error absoluto medio y el error absoluto porcentual medio, al igual que ocurre con el set de datos D30M15.

Tabla E.15: Resultados modelos de regresión, sub-set D30M0

Modelo	RMSE	MAE	MAPE
RL	229.65	180.90	0.38
Árbol de regresión	224.63	175.08	0.38
Random Forest	230.50	180.78	0.39
GBM	225.50	174.76	0.37
SVM	226.09	173.58	0.36

Tabla E.16: Nivel de precisión de predicción de días por intervalos sub-set D30M0

Modelo	$\pm$ 15 días	$\pm$ 30 días	$\pm$ 40 días	$\pm$ 60 días	>60 días
Regresión múltiple	4 %	11 %	15 %	22 %	78 %
Árbol podado	5 %	11 %	16 %	24 %	76 %
Random Forest	9 %	14 %	18 %	25 %	75 %
Gradient Boosting	9 %	16 %	18 %	25 %	75 %
Máquinas de soporte	5 %	11 %	17 %	27 %	73 %

De la Tabla E.16 se aprecia que el nivel de predicción es bastante bajo a menor número de días por intervalo. Incluso, la predicción con certeza de intervalos de 2 meses absolutos no alcanza a ser ni un 30 % de los datos.

E.4. Predicción fecha de asistencia a cuarta mantención

Modelo *D45M4*

Las variables de mayor importancia en este modelo son las variables relacionadas a las mantenciones previas del vehículo y en menor medida las relacionadas a la descripción del cliente, según el modelo de bosques aleatorios en la Figura E.7. Además, tal como se esperaba, la variable correspondiente a la primera mantención posee muy poca importancia para el modelo, probablemente por corresponder a un dato *antiguo* con respecto a la fecha de la mantención estudiada.

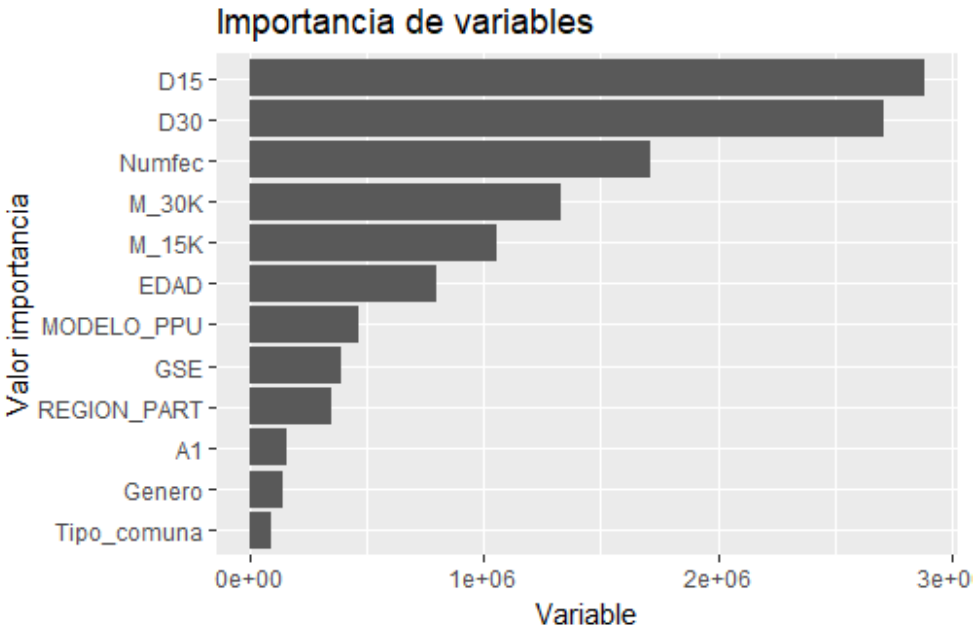


Figura E.7: Importancia variables modelo D45M4 según Random Forest

Se realizan los modelos de regresión para este set de datos, obteniendo un error de aproximadamente 80 días de diferencia con los valores reales según lo entregado por la métrica error absoluto medio. Aún así, como se aprecia en la tabla E.17 existe poca diferencia entre los modelos de predicción aplicados. El que mejor métrica presenta en cuanto a error absoluto medio es el algoritmo de Potenciación del Gradiente, presentando también el menor error porcentual.

Tabla E.17: Resultados modelos de regresión, sub-set D45M4

Modelo	**RMSE**	**MAE**	**MAPE**
RL	115.87	83.73	0.29
Árbol de regresión	110.68	79.23	0.27
Random Forest	108.79	78.41	0.28
GBM	108.05	77.89	0.27
SVM	113.52	80.44	0.26

Tabla E.18: Nivel de precisión de predicción de días por intervalos sub-set D45M4

Modelo	**± 15 días**	**± 30 días**	**± 40 días**	**± 60 días**	**>60 días**
Regresión múltiple	10 %	27 %	32 %	50 %	50 %
Árbol podado	14 %	25 %	35 %	50 %	50 %
Random Forest	13 %	26 %	38 %	53 %	47 %
Gradient Boosting	12 %	24 %	33 %	48 %	52 %
Máquinas de soporte	10 %	23 %	34 %	50 %	50 %

Modelo *D45M15*

Nuevamente las variables de mayor importancia en este modelo son las variables relacionadas a las mantenciones previas del vehículo y las relacionadas a la descripción del cliente aumentan su importancia relativa con respecto al modelo que posee datos de la tercera mantención, como se puede ver en la Figura E.8.

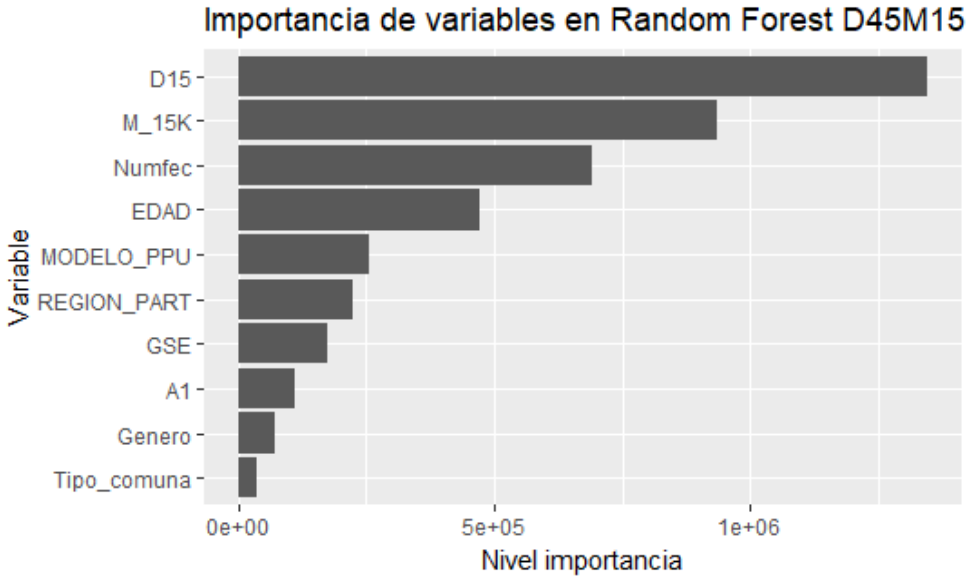


Figura E.8: Importancia variables modelo D45M15 según Random Forest

width=9cm En la Tabla E.19 se presentan los resultados obtenidos de los modelos aplicados a la predicción de fecha de asistencia a la cuarta mantención con asistencia previa a la segunda mantención. Se puede ver que, a diferencia de otros estudios, este dataset obtiene mayores diferencias de predicción entre modelos, y la predicción que entrega no es muy buena en general, porque el error es de aproximadamente 140 días en promedio, lo que equivale a 5 meses y esto no es muy útil si se busca una predicción más precisa que la existente. Esta diferencia entre predicciones es probable que sea provocada por la baja cantidad de datos que genera un sobreajuste sobre el set de entrenamiento que no permite predecir de forma correcta el sobre el set de testeo.

Tabla E.19: Resultados modelos de regresión, sub-set D45M15

Modelo	RMSE	MAE	MAPE
RL	176.37	149.92	0.25
Árbol de regresión	151.55	126.48	0.23
Random Forest	150.65	128.75	0.23
GBM	177.27	150.21	0.27
SVM	154.76	133.59	0.24

Tabla E.20: Nivel de precisión de predicción de días por intervalos sub-set D45M15

Modelo	± 15 días	± 30 días	± 40 días	± 60 días	>60 días
Regresión múltiple	8 %	14 %	14 %	17 %	83 %
Árbol podado	3 %	14 %	19 %	31 %	69 %
Random Forest	3 %	19 %	25 %	28 %	72 %
Gradient Boosting	3 %	6 %	11 %	19 %	81 %
Máquinas de soporte	11 %	14 %	19 %	19 %	81 %

Modelo *D45M30*

Las variables de mayor importancia en este modelo son las variables relacionadas a las mantenciones previas del vehículo y en menor medida las relacionadas a la descripción del cliente, según el modelo de bosques aleatorios en la Figura E.9.

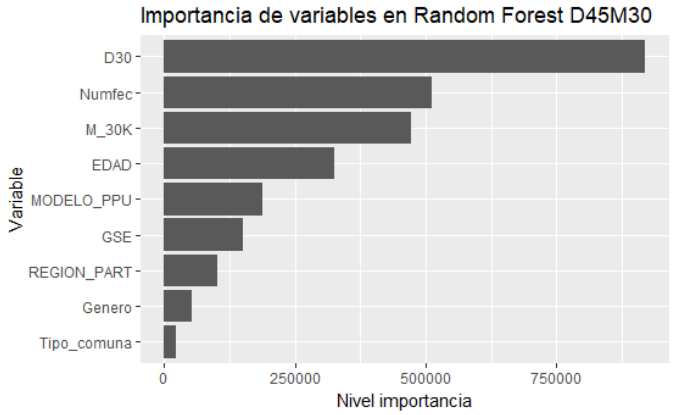


Figura E.9: Importancia variables modelo D45M30 según Random Forest

Los resultados para el modelo de predicción de días a cuarta mantención con asistencia previa a tercera mantención se observan en la Tabla E.21. Los modelos presentan un desempeño similar en cuanto al *RMSE* a excepción del modelo de árbol de regresión, el que presenta el peor desempeño también en error absoluto medio. El modelo que menor error absoluto presenta es las Máquinas de soporte vectorial.

Tabla E.21: Resultados modelos de regresión, sub-set D45M30

Modelo	RMSE	MAE	MAPE
RL	127.04	94.83	0.37
Árbol de regresión	136	100.5	0.31
Random Forest	127.24	97.67	0.40
GBM	125.84	96.80	0.39
SVM	121.04	91.12	0.35

Tabla E.22: Nivel de precisión de predicción de días por intervalos sub-set D45M30

Modelo	± 15 días	± 30 días	± 40 días	± 60 días	>60 días
Regresión múltiple	14 %	24 %	27 %	41 %	59 %
Árbol podado	14 %	25 %	35 %	45 %	55 %
Random Forest	16 %	29 %	37 %	49 %	51 %
Gradient Boosting	8 %	25 %	31 %	49 %	51 %
Máquinas de soporte	12 %	29 %	31 %	47 %	53 %

Modelo *D45M0*

En este modelo, debido a que no existen variables relacionadas a mantenciones previas a excepción de la binaria de asistencia a la primera mantención (A1), las variables relacionadas a características del cliente toman la importancia general, como se observa en la Figura E.10.

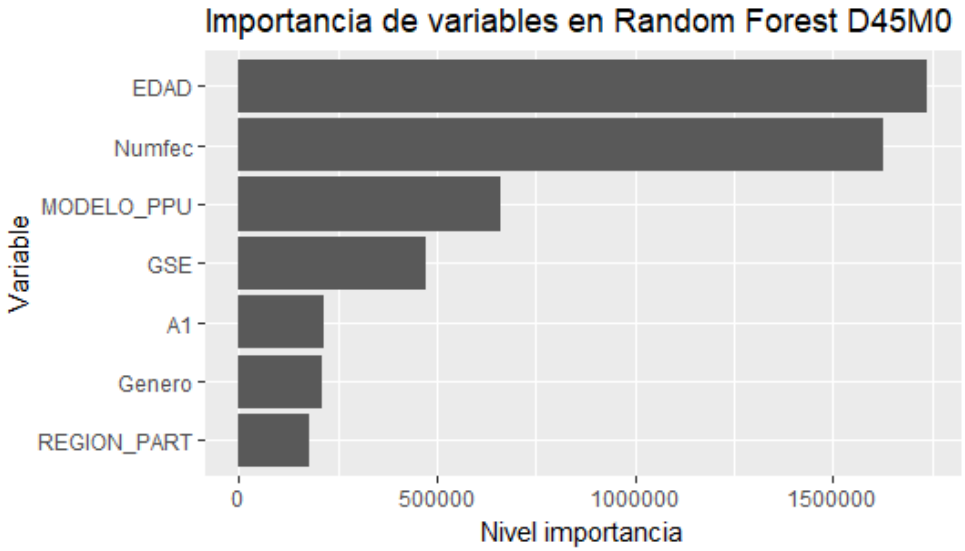


Figura E.10: Importancia variables modelo D45M0 según Random Forest

width=9cm Los resultados para el modelo de predicción de días a cuarta mantención sin asistencia previa a mantenciones se presentan en la Tabla E.23. Este subset de datos es el que peores resultados presenta en cuanto a predicción. Esto se debe a que posee la menor cantidad de observaciones de los set de datos, y por tanto, como ocurre en otros casos, el sobreajuste del set de entrenamiento impide el lograr predecir con bajo error los días hasta la cuarta mantención.

Tabla E.23: Resultados modelos de regresión, sub-set D45M0

Modelo	RMSE	MAE	MAPE
RL	270.56	208.02	0.23
Árbol de regresión	271.82	227.54	0.31
Random Forest	258.40	209.42	0.28
GBM	276.10	228.08	0.30
SVM	273.74	217.88	0.30

Tabla E.24: Nivel de precisión de predicción de días por intervalos sub-set D45M0

Modelo	$\pm$ 15 días	$\pm$ 30 días	$\pm$ 40 días	$\pm$ 60 días	>60 días
Regresión múltiple	4 %	11 %	11 %	19 %	81 %
Árbol podado	4 %	4 %	4 %	7 %	93 %
Random Forest	4 %	4 %	4 %	7 %	93 %
Gradient Boosting	4 %	4 %	7 %	11 %	89 %
Máquinas de soporte	4 %	4 %	7 %	7 %	93 %

Anexo F. Propensión de asistencia

F.1. Propensión de asistencia a segunda mantención

Tabla F.1: Resultados regresión logística segunda mantención

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	Sign.
(Intercept)	1,546	0,204	7,589	0,000	***
EDAD	-0,006	0,001	-5,067	0,000	***
SEXO	0,097	0,035	2,791	0,005	**
GSEC2	0,138	0,073	1,893	0,058	.
GSEC3	0,104	0,073	1,422	0,155	
GSED	0,054	0,069	0,785	0,432	
GSEE	0,048	0,103	0,469	0,639	
REGION_PARTDE ATACAMA	0,370	0,188	1,968	0,049	*
REGION_PARTDE COQUIMBO	0,451	0,144	3,143	0,002	**
REGION_PARTDE LA ARAUCANIA	0,248	0,194	1,277	0,202	
REGION_PARTDE LOS LAGOS	1,006	0,143	7,060	0,000	***
REGION_PARTDE LOS RIOS	0,876	0,189	4,640	0,000	***
REGION_PARTDE NUBLE	0,687	0,224	3,061	0,002	**
REGION_PARTDE TARAPACA	0,492	0,345	1,426	0,154	
REGION_PARTDE VALPARAISO	0,811	0,116	6,962	0,000	***
REGION_PARTDEL BIO BIO	0,171	0,135	1,263	0,207	
REGION_PARTDEL LIBERTADOR BERNARDO OHIGGINS	0,805	0,134	6,023	0,000	***
REGION_PARTDEL MAULE	0,826	0,143	5,770	0,000	***
REGION_PARTMETROPOLITANA DE SANTIAGO	0,753	0,110	6,849	0,000	***
Tipo_comunaUrbana	0,060	0,082	0,727	0,467	
MODELO_PPUCERATO	-0,165	0,134	-1,234	0,217	
MODELO_PPUCERATO 5	-0,096	0,341	-0,281	0,779	
MODELO_PPUMORNING	-0,559	0,127	-4,391	0,000	***
MODELO_PPURIO 4	-0,295	0,129	-2,282	0,022	*
MODELO_PPURIO 5	-0,364	0,129	-2,833	0,005	**
MODELO_PPUSORENTO	0,061	0,138	0,442	0,659	
MODELO_PPUSOUL	-0,192	0,152	-1,266	0,205	
MODELO_PPUSPORTAGE	-0,047	0,128	-0,370	0,712	
A1K	-2,225	0,045	-49,404	<2e-16	***
recency	-0,001	0,000	-2,291	0,022	*
—					
Signif. codes: 0 '***' 0.001 '**' 0.01 '*'	0.05 '.' 0	.1 ' ' 1			
(Dispersion parameter for binomial family t	aken to be	1)			
Null deviance: 23397 on 16924 degrees	of freedom				
Residual deviance: 19695 on 16895 degrees	of freedom				
AIC: 19755					
Number of Fisher Scoring iterations: 4					

F.2. Propensión de asistencia a tercera mantención

Tabla F.2: Resultados regresión logística tercera mantención

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	Sign.
(Intercept)	-0,730	0,240	-3,030	0,000	**
EDAD	-0,010	0,000	-4,020	0,000	***
Genero	0,020	0,040	0,600	0,550	
GSEC2	0,010	0,080	0,170	0,860	
GSEC3	0,010	0,080	0,170	0,860	
GSED	-0,010	0,080	-0,140	0,890	
GSEE	-0,000	0,120	-0,030	0,970	
REGION_PARTDE ATACAMA	0,430	0,240	1,780	0,080	.
REGION_PARTDE COQUIMBO	0,770	0,190	4,120	0,000	***
REGION_PARTDE LA ARAUCANIA	0,240	0,240	0,990	0,320	
REGION_PARTDE LOS LAGOS	1,020	0,180	5,540	0,000	***
REGION_PARTDE LOS RIOS	0,850	0,240	3,620	0,000	***
REGION_PARTDE NUBLE	0,870	0,230	3,700	0,000	***
REGION_PARTDE TARAPACA	1,000	0,380	2,590	0,010	**
REGION_PARTDE VALPARAISO	0,690	0,160	4,400	0,000	***
REGION_PARTDEL BIO BIO	0,500	0,170	2,880	0,000	**
REGION_PARTDEL LIBERTADOR BERNARDO OHIGGINS	0,900	0,170	5,150	0,000	***
REGION_PARTDEL MAULE	0,880	0,180	4,780	0,000	***
REGION_PARTMETROPOLITANA DE SANTIAGO	0,990	0,150	6,710	0,000	***
Tipo_comunaUrbana	-0,010	0,100	-0,130	0,890	
MODELO_PPUCERATO	-0,490	0,130	-3,760	0,000	***
MODELO_PPUCERATO 5	-0,280	0,320	-0,860	0,390	
MODELO_PPUMORNING	-0,740	0,130	-5,920	0,000	***
MODELO_PPURIO 4	-0,520	0,130	-4,120	0,000	***
MODELO_PPURIO 5	-0,540	0,130	-4,230	0,000	***
MODELO_PPUSORENTO	-0,110	0,130	-0,800	0,430	
MODELO_PPUSOUL	-0,490	0,150	-3,260	0,000	**
MODELO_PPUSPORTAGE	-0,240	0,120	-1,970	0,050	*
A15K	1,110	0,050	23,400	<2e-16	***
A1K	-0,740	0,040	-16,910	<2e-16	***
recency	-0,000	0,000	-2,200	0,030	*
—					
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’	0.05 ‘.’ 0	.1 ‘ ’ 1			
(Dispersion parameter for binomial family taken to b 1)					
Null deviance: 16321 on 12226 degrees of fredoom					
Residual deviance: 14497 on 12196 degrees of freedom					
AIC: 14559					
Number of Fisher Scoring iterations: 4					

F.3. Propensión de asistencia a cuarta mantención

Tabla F.3: Resultados regresión logística segunda mantención

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	Sign.
(Intercept)	1,546	0,204	7,589	0,000	***
EDAD	-0,006	0,001	-5,067	0,000	***
SEXO	0,097	0,035	2,791	0,005	**
GSEC2	0,138	0,073	1,893	0,058	.
GSEC3	0,104	0,073	1,422	0,155	
GSED	0,054	0,069	0,785	0,432	
GSEE	0,048	0,103	0,469	0,639	
REGION_PARTDE ATACAMA	0,370	0,188	1,968	0,049	*
REGION_PARTDE COQUIMBO	0,451	0,144	3,143	0,002	**
REGION_PARTDE LA ARAUCANIA	0,248	0,194	1,277	0,202	
REGION_PARTDE LOS LAGOS	1,006	0,143	7,060	0,000	***
REGION_PARTDE LOS RIOS	0,876	0,189	4,640	0,000	***
REGION_PARTDE NUBLE	0,687	0,224	3,061	0,002	**
REGION_PARTDE TARAPACA	0,492	0,345	1,426	0,154	
REGION_PARTDE VALPARAISO	0,811	0,116	6,962	0,000	***
REGION_PARTDEL BIO BIO	0,171	0,135	1,263	0,207	
REGION_PARTDEL LIBERTADOR BERNARDO OHIGGINS	0,805	0,134	6,023	0,000	***
REGION_PARTDEL MAULE	0,826	0,143	5,770	0,000	***
REGION_PARTMETROPOLITANA DE SANTIAGO	0,753	0,110	6,849	0,000	***
Tipo_comunaUrbana	0,060	0,082	0,727	0,467	
MODELO_PPUCERATO	-0,165	0,134	-1,234	0,217	
MODELO_PPUCERATO 5	-0,096	0,341	-0,281	0,779	
MODELO_PPUMORNING	-0,559	0,127	-4,391	0,000	***
MODELO_PPURIO 4	-0,295	0,129	-2,282	0,022	*
MODELO_PPURIO 5	-0,364	0,129	-2,833	0,005	**
MODELO_PPUSORENTO	0,061	0,138	0,442	0,659	
MODELO_PPUSOUL	-0,192	0,152	-1,266	0,205	
MODELO_PPUSPORTAGE	-0,047	0,128	-0,370	0,712	
A1K	-2,225	0,045	-49,404	<2e-16	***
recency	-0,001	0,000	-2,291	0,022	*
—					
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’	0.05 ‘.’ 0	.1 ‘ ’ 1			
(Dispersion parameter for binomial family t	aken to be	1)			
Null deviance: 23397 on 16924 degrees	of freedom				
Residual deviance: 19695 on 16895 degrees	of freedom				
AIC: 19755					
Number of Fisher Scoring iterations: 4					