



**METODOLOGÍA CRISP-DM APLICADA AL ANÁLISIS DE LOS  
EFECTOS DE LA COVID-19 EN EL SERVICIO DE  
ATENCIÓN AL CLIENTE**

Tutora:  
Carmen López Martín

Autor:  
Daniel Rodríguez Fustes

Madrid, junio de 2022

# Índice de Contenidos

<b>1. INTRODUCCIÓN .....</b>	<b>2</b>
<b>2. ANTECEDENTES .....</b>	<b>4</b>
<b>3. METODOLOGÍA CRISP-DM. ....</b>	<b>8</b>
3.1 COMPRENSIÓN DEL NEGOCIO.....	9
3.2 COMPRENSIÓN DE LOS DATOS.....	10
3.3 PREPARACIÓN DE LOS DATOS.....	11
3.4 MODELOS TEÓRICOS. ....	15
3.4.1 MODELO LINEAL GENERALIZADO.....	16
3.4.2 MODELOS DE ÁRBOL DE DECISIÓN. ....	17
3.4.3 MODELOS REDES NEURONALES ARTIFICIALES. ....	18
3.4.4 MODELO NAÏVE BAYES. ....	18
3.5 EVALUACIÓN. ....	19
<b>4. DATOS Y RESULTADOS. ....</b>	<b>20</b>
4.1 ANÁLISIS DESCRIPTIVO DE LOS DATOS. ....	20
4.2 MODELIZACIÓN.....	28
4.2.1 MODELO LINEAL GENERALIZADO.....	28
4.2.2 MODELOS DE ÁRBOL DE DECISIÓN. ....	32
4.2.3 MODELOS REDES NEURONALES ARTIFICIALES. ....	35
4.2.4 MODELO NAÏVE BAYES. ....	37
4.3 EVALUACIÓN Y COMPARACIÓN ENTRE MODELOS.....	38
<b>5. DESPLIEGUE E IMPLEMENTACIÓN .....</b>	<b>40</b>
5.1 DESPLIEGUE. ....	40
5.2 IMPLEMENTACIÓN. ....	44
<b>6. CONCLUSIONES. ....</b>	<b>47</b>
<b>7. BIBLIOGRAFÍA.....</b>	<b>49</b>
<b>8. ANEXO .....</b>	<b>52</b>

## **1. INTRODUCCIÓN**

---

La pandemia de COVID-19 azotó al mundo entero desde finales de 2019 y alteró el mundo tal y como se conocía. Esta situación, diferente a la anterior, desembocó en la denominada “Nueva normalidad” que cambió para siempre la forma de relacionarse de los agentes económicos en las economías mundiales.

La pandemia provocó los confinamientos, que supusieron la casi paralización de la economía mundial debido a la detención de los procesos industriales, con una contracción tanto por el lado de la oferta como de la demanda en la mayoría de los mercados, a excepción de los asociados a servicios básicos como la alimentación, sanidad o la seguridad. Los confinamientos tuvieron muchos efectos muy variados sobre las relaciones interpersonales, y en concreto, impulsaron de una manera exponencial el uso de los canales digitales como forma de comunicación.

La situación de confinamientos, seguida por las restricciones en la comunicación entre personas de manera física, supuso la extensión masiva de las herramientas de comunicación digital en todo el mundo, y se estableció como la nueva forma de comunicación entre personas. A partir de este momento, toda la comunicación entre personas se digitalizó, lo que permitió su registro y archivo de forma extensiva.

Esta expansión global de la comunicación digital resulta en una oportunidad de adquisición de conocimiento sin precedentes. Esta oportunidad provoca que las empresas estén destinando más recursos a sus departamentos de servicio de atención al cliente, puesto que las técnicas de análisis de datos, tanto estructurados como no estructurados, aplicadas a este departamento, han demostrado su incalculable utilidad para, entre otras cosas, extraer relaciones, patrones de comportamiento, sentimientos, tendencias, ciclos estacionales, anomalías... todo ello con el fin de encontrar oportunidades/tendencias del negocio, mejorar la productividad, mejorar la competitividad, descubrir nuevas formas de disminuir costes y, en última instancia, incrementar los beneficios.

En un mercado de libre competencia globalizado, con escasez de chips semiconductores y fricción en el transporte internacional de mercancías, las empresas deben emplear todos los medios a su alcance para garantizar su viabilidad y supervivencia en el corto plazo. El análisis de datos aplicado al servicio de atención al cliente es completamente necesario para dar un servicio personalizado, omnicanal y dinámico a cada cliente de acuerdo a sus preferencias; así por ejemplo, el idioma, la región geográfica, la tecnología de comunicación o el tamaño del cliente son elementos clave que han de ser tomados en cuenta para dar un servicio adecuado, y, de esta manera, mantener o hacer crecer la relación comercial con cada uno de los clientes existentes además de aumentar las probabilidades de captar clientes potenciales.

En este entorno de incertidumbre, se hace prioritario analizar cuáles son las características más valoradas por los clientes en lo que a la gestión de pedidos se refiere. Para ello, este estudio presenta los antecedentes del servicio al cliente, su evolución en el tiempo y la importancia del empleo de algoritmos de *machine learning* en el análisis de los datos en la actualidad. Una vez desarrollados los antecedentes, se presenta la metodología que se considera más adecuada para la obtención de los resultados.

Para el desarrollo de la metodología, y debido al carácter cuantitativo del trabajo, se emplea el software RStudio, que es un lenguaje de programación gratuito que tiene una

capacidad excelente para la realización de procesos de data mining. Su capacidad para trabajar con *dataframes* (matrices con datos heterogéneos) lo hacen ideal para realizar las tareas de preprocesamiento, modelización y evaluación de grandes conjuntos de datos. Además, dispone de múltiples posibilidades para realizar estudios descriptivos sobre el conjunto de datos, apoyados por representaciones gráficas. Se proporciona todo el código empleado en el Anexo, estructurado por grupos de procesos y con comentarios aclaratorios.

Después de la obtención de los resultados numéricos, y con el empleo del conocimiento y la experiencia previa, se extrae la información relevante de los resultados numéricos, para en último lugar, presentar las conclusiones del estudio.

## 2. ANTECEDENTES

---

La forma de comunicarse de los departamentos de servicio al cliente ha ido cambiando acorde a los nuevos tiempos. Así, es posible diferenciar varias etapas por las que los departamentos de servicio al cliente han ido pasando. A continuación, se enumeran y se presentan en orden cronológico las etapas principales:

1. Etapa del correo postal. Las empresas contactan con sus clientes a través de correo postal. En esta fase se producen grandes problemas para que se desarrolle una comunicación fluida entre empresa y cliente lo cual perjudica de forma sustancial la relación económica.
2. Etapa del teléfono analógico. Las empresas contactan con sus clientes a través de llamadas telefónicas/fax. En esta fase se mejora mucho la velocidad de la comunicación entre empresa y cliente.
3. Etapa del correo electrónico. Las empresas contactan con sus clientes a través de correos electrónicos. En esta primera fase digital, la comunicación entre empresa y cliente es rápida y precisa.
4. Etapa de la autogestión. Las empresas otorgan herramientas a sus clientes para autogestionarse. En esta segunda fase digital, se provee a los clientes de las herramientas necesarias para su autogestión de forma que puedan agilizar aún más los tiempos de gestión.
5. Etapa de la robotización. Las empresas emplean bots para relacionarse con sus clientes. Es la fase de *customer experience*. En ella las empresas implementan bots reactivos que ayudan al cliente en la autogestión (Rall, 2021).

Figura 1. Evolución de las tecnologías asociadas al servicio al cliente.



Fuente: [www.blog.usu.com](http://www.blog.usu.com)

La Figura 1 muestra la evolución de servicio de atención al cliente desde el siglo XX hasta la actualidad.

Los canales de comunicación que se emplean en las distintas etapas no son excluyentes entre sí. Sin embargo, el correo postal quedó obsoleto como forma de comunicación hace muchos años, y la llamada telefónica y el correo electrónico serán gradualmente reemplazados por los bots en los próximos años. En última instancia, los departamentos de servicio al cliente pueden integrar todos los canales de comunicación, en lo que se conoce como omnicanal, de manera que las empresas tienen que poder contactar con sus clientes en su canal de comunicación preferido.

Esta última etapa, denominada *customer experience*, que recoge toda la experiencia del cliente con la empresa, desde el descubrimiento del producto o servicio, investigación sobre el mismo, su compra, su uso, su revisión y la postventa, sustituye al concepto tradicional de servicio al cliente, y tiene sus cimientos en la explotación de las bases de datos digitales a través de técnicas de análisis de datos. La recolección de datos por parte de las empresas para entender mejor el mercado y detectar oportunidades de negocio no es algo nuevo; sin embargo, hay un antes y un después, asociado a la digitalización de los datos. En la era analógica, la recolección y tratamiento de la información resultaban muy costosos y con un alto riesgo de errores humanos. Sin embargo, ya entonces, el análisis de los datos demostraba ser muy útil en los procesos de toma de decisiones. En la era digital, y gracias a las nuevas herramientas disponibles, la recolección de los datos por parte de las empresas resulta mucho más eficiente en términos de coste, cantidad y calidad de la información, y su uso se ha generalizado. De hecho, en la actualidad, los datos se consideran un activo más junto al capital y al trabajo.

Un ejemplo de análisis de datos en el área del comportamiento de los clientes se encuentra en el trabajo de Exenberger y Bucko (2020), en el que se realiza un desarrollo de una metodología CRISP-DM (Cross Industry Standard Process for Data Mining) para modelizar el comportamiento de los clientes mediante técnicas de *machine learning* y *clustering*. Su trabajo se centra en crear grupos de clientes homogéneos para enviar campañas de marketing personalizadas. Este tratamiento de los clientes mediante grupos es correcto, pero el futuro nos indica que el nivel de personalización demandado por los clientes seguirá creciendo, de manera que la solución debe ser específica para cada cliente. Es aquí donde se detecta que los esfuerzos a futuro deben dirigirse en la continua mejora en los algoritmos de *machine learning* que detecten las necesidades concretas de cada cliente, de forma que proactivamente se ofrezcan las soluciones exactas para cada cliente. Por supuesto, estas soluciones se deberán ir implementando de manera gradual a medida que los algoritmos y la tecnología lo permitan.

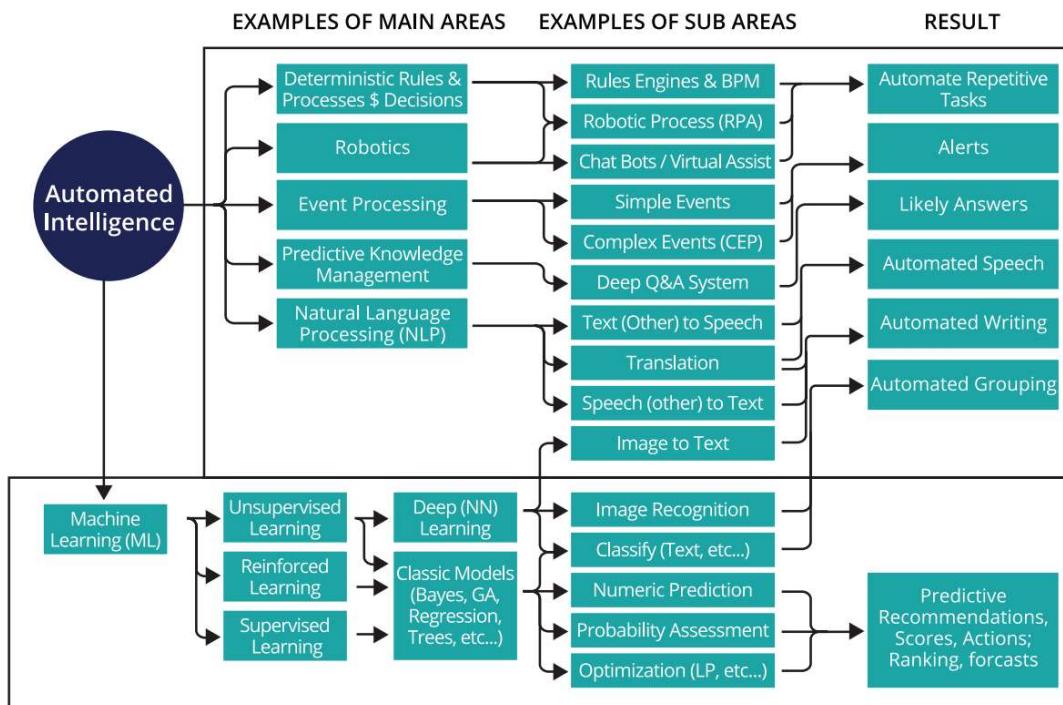
Debido a la facilidad de su implementación y al reducido coste en proporción a las oportunidades que brinda, la minería de datos se ha comenzado a usar de forma extensiva en los departamentos de servicio al cliente, en concreto, para analizar los datos de sentimientos de los clientes: datos cualitativos que contienen una opinión y que tienen un valor enorme para los procesos de toma de decisiones. En la actualidad, el mayor reto que tienen las empresas a la hora de extraer conocimiento de este tipo de datos, es la transformación de dichos datos no estructurados (imágenes, vídeos, textos etc...) en datos estructurados cuantificables, que permitan su incorporación en modelos y, consecuentemente, la obtención del conocimiento latente.

En lo referente a los procesos de minería de datos, existen dos grandes métodos: el SEMMA (Sample, Explore, Modify, Model, Asses) y el CRISP-DM. Si en este estudio se hubiesen incorporado variables con aspectos técnicos de los productos que produce la empresa, hubiese sido adecuado abordar el estudio desde una metodología DMME (Data Mining Methodology for Engineering Applications) (Huber et al. 2019) pero su inclusión habría complicado enormemente la modelización, por este motivo se desarrolla la metodología CRISP-DM por tener en cuenta tanto la comprensión del negocio como la implementación de las acciones necesarias, además de todas las etapas existentes en el método SEMMA.

El estudio de Schröer et al. (2021) analiza 24 investigaciones recientes en los que se empleó la metodología CRISP-DM. Este estudio permite observar cómo se está aplicando esta metodología y servirá de guía para desarrollar este trabajo con buenas prácticas. A modo de ejemplo, el trabajo de Solano et al. (2021) muestra la tendencia actual de emplear algoritmos de *machine learning* para modelizar y obtener resultados.

Una parte clave en los procesos de data mining se refiere al empleo de la inteligencia artificial aplicada a los datos y las posibilidades que brinda.

Figura 2. Técnicas de inteligencia artificial y áreas de uso.

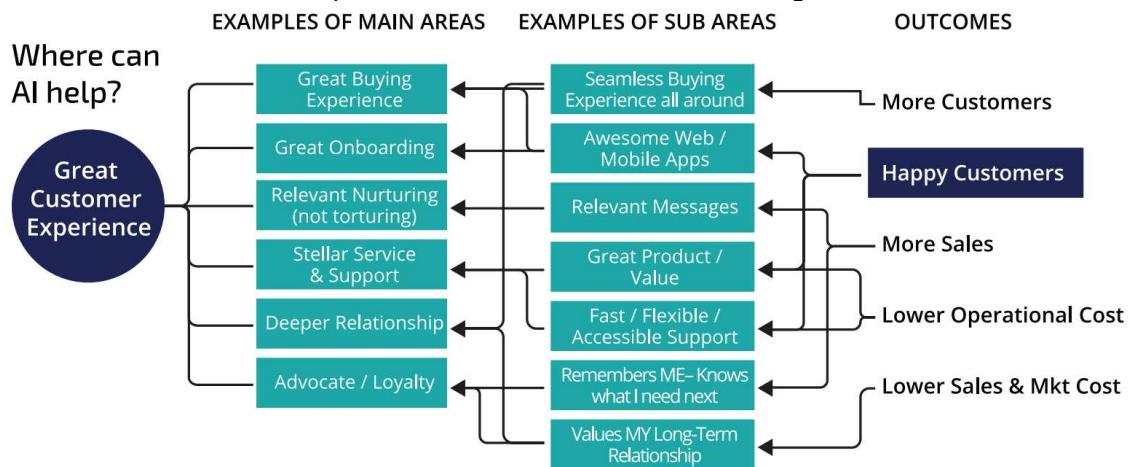


Fuente: [www.pega.com](http://www.pega.com)

En la Figura 2 se muestran algunos ejemplos de técnicas de inteligencia artificial disponibles en función del propósito que se persiga: desde la automatización de tareas o la estimación de modelos de regresión/clasificación hasta las técnicas más vanguardistas en economía como las redes neuronales artificiales para la extracción de información desde fuentes de datos no estructurados.

En concreto, y para el área de servicio al cliente, la inteligencia artificial puede emplearse en diversos campos. En la Figura 3 se muestran algunos ejemplos:

Figura 3. Ejemplos de áreas y subáreas del *customer experience* que mejoran con la implementación de herramientas de inteligencia artificial.



Fuente: [www.pega.com](http://www.pega.com)

Se puede decir, que el empleo de la inteligencia artificial es una herramienta imprescindible para lograr una gran *customer experience*. Así, por ejemplo, una empresa que centre sus esfuerzos en tener contenidos a sus clientes puede lograrlo mediante la implementación de algoritmos de inteligencia artificial en las siguientes subáreas:

- Página web. La inteligencia artificial aplicada en páginas webs y aplicaciones móviles mejora el *user experience*: contenido de calidad, fácil acceso y un entorno amigable muy intuitivo.
- Producto. La inteligencia artificial permite detectar las preferencias y tendencias del mercado, así como la minimización de costes para finalmente proporcionar un producto con una gran calidad-precio.
- Soporte. La inteligencia artificial proporciona herramientas que ayudan a los departamentos de soporte a prestar un servicio rápido, personalizado y flexible.

La inteligencia artificial potencia el *customer experience* para llevarlo a un nivel superior, pero en ningún caso se trata de una herramienta cuyo fin sea reemplazar a las personas (Davis, 2021).

Las empresas que no aplican la inteligencia artificial para lograr sus objetivos se hayan en desventaja con respecto a la competencia, al no emplear toda la tecnología a su alcance y al no adaptar sus herramientas productivas a las preferencias de los clientes.

En este trabajo, el autor, con más de 10 años de experiencia en el departamento de servicio al cliente, procede a realizar un estudio microeconómico completo partiendo de un conjunto de datos estructurados con más de 1.200 observaciones y 16 variables explicativas relacionadas con el departamento de servicio al cliente de una empresa multinacional que es fabricante de productos electrónicos de seguridad.

Siguiendo la metodología CRISP-DM se procede a recolectar, tratar y explotar los datos con el fin de obtener una valiosa información descriptiva y predictiva que permitan detectar las variables más relevantes que determinan una gestión de pedidos excelente desde el inicio de la pandemia y proponer acciones de mejora en el departamento de servicio al cliente de la empresa estudiada.

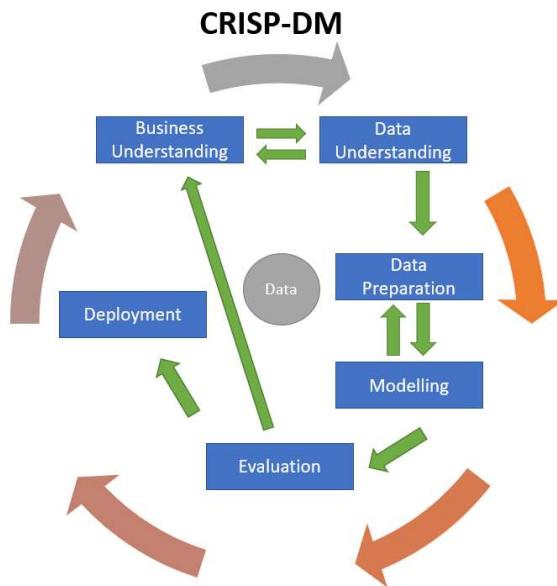
### 3. METODOLOGÍA CRISP-DM.

---

Con el fin de realizar las tareas de minería de datos de una forma ordenada y jerarquizada, se sigue una metodología CRISP-DM (SPSS, NCR y Daimler Chrysler, 2000). Esta metodología se ha empleado desde hace varios años y ha demostrado ser muy útil y eficaz para abordar cuestiones de negocio desde el análisis de datos.

La metodología CRISP-DM, tiene seis fases que se relacionan entre sí tal y como se muestra en la siguiente imagen:

Figura 4. Fases de la metodología CRISP-DM



Fuente: [www.towardsdatascience.com](http://www.towardsdatascience.com)

La Figura 4 es un esquema de las fases de un proyecto de CRISP-DM. A continuación, se describen brevemente cada una de estas fases:

- Fase 1. Comprensión del negocio. En esta fase se plantea la pregunta que quiere ser resuelta mediante el proceso de minería de datos.
- Fase 2. Comprensión de los datos. En esta fase se toma el conjunto de datos que se va a emplear para dar respuesta a la pregunta inicial. Se debe comparar el conjunto de datos disponible con la pregunta inicial para decidir si es válido.
- Fase 3. Preparación de los datos. En esta fase se prepara el conjunto de datos para la modelización. Se deben tratar datos anómalos, datos ausentes... para lograr una base de datos " limpia ". Se realiza además un análisis descriptivo.
- Fase 4. Modelización. En esta fase se especifica el modelo que se va a aplicar sobre el conjunto de datos y el porqué de su elección. Se realizan test técnicos para su validación. Si un modelo no resulta válido, debe ser revisado o sustituido por otro que sí resulte válido.
- Fase 5. Evaluación. En esta fase se comprueba si el modelo validado consigue dar respuesta a la pregunta inicial y se evalúa desde el plano del negocio para determinar su validez o no. Si no resulta válido, el proceso de minería de datos debe comenzar desde el principio.
- Fase 6. Despliegue. Esta es la fase de la explotación práctica de los resultados que acabará con la redacción de un informe o el desarrollo de un nuevo modelo de gestión.

### **3.1 Comprensión del negocio.**

La fase de comprensión del negocio es fundamental para el éxito del proceso de minería de datos. La explotación de los datos sin una perspectiva del negocio puede desembocar en conclusiones erróneas originadas por relaciones espurias entre las variables. Es por este motivo, que los datos en sí mismos, carecen de valor si no se explotan en un sentido basado en la experiencia y/o el conocimiento experto del investigador. En esta fase, se define la cuestión que se pretende resolver, cómo se aborda el proyecto y cuáles son los objetivos que se pretenden lograr. Se procede a desglosar esta fase en cuatro puntos:

**1. Determinación de la pregunta a responder.**

Se trata de la pregunta que establece el objetivo a alcanzar en el desarrollo de la metodología de data mining. En el siglo XXI las empresas tienen la obligación de establecer una inversión continua en materia de *customer experience* para no quedar obsoletas y en desventaja frente a la competencia. Además, los clientes finales demandan calidad, facilidad, rapidez y eficiencia al emplear los canales de información digitales prestados por su proveedor, es decir, el *user experience* es determinante a la hora de lograr el éxito o fracasar en la comunicación con el cliente. Con la intención de mejorar el *customer experience* brindado por la empresa objeto de estudio se realiza la pregunta: ¿cuáles son las variables más importantes a la hora de realizar una gestión de pedidos excelente?

**2. Evaluación de la situación.**

La empresa analizada presta mucha atención a la calidad de su servicio de atención al cliente. Por ello tiene implementadas múltiples herramientas para dar soporte. En este momento, se encuentra implementando la gestión automática de pedidos por un RPA (Robotic Process Automation), de modo que los empleados que introducen manualmente los pedidos de los clientes en el sistema liberen su tiempo para realizar otras tareas de servicio al cliente más productivas. Actualmente la empresa tiene una valoración global del servicio de atención al cliente de 4,7 puntos sobre 5. Muchos de los clientes evalúan con 5 puntos el servicio recibido, pero otros tantos dan una puntuación menor.

Posibles desarrollos a futuro para la mejora del servicio al cliente (Marsden et al. 2022):

- Minería de opinión en redes sociales.
- Chatbox en tienda online y página web.
- Asistente de voz.
- Canal de comunicación con clientes vía redes sociales corporativas.
- Omnicanal de los diversos canales de comunicación.

**3. Determinación de los objetivos del proyecto.**

Se emplean desde técnicas económicas tradicionales hasta algoritmos de *machine learning*, todo ello para extraer la máxima información posible que contienen los datos y emplear el conocimiento adquirido en la mejora del servicio de gestión de pedidos para procurar una gran *customer experience*.

La detección de las variables más importantes para prestar un servicio de gestión de pedidos excelente permite proponer acciones de mejora sobre los puntos en que se detecten ineficiencias y de esta forma mantener a la empresa objeto de estudio en un nivel superior de servicio al cliente, que se traduzca en el mantenimiento de los clientes actuales y en la adquisición de otros nuevos.

#### 4. Plan del proyecto.

En primer lugar, se realiza el análisis en profundidad de los datos para la extracción del conocimiento.

En segundo lugar, con el conocimiento extraído se detectan las áreas de mejora.

En tercer lugar, se estudia caso a caso cada una de las áreas de mejora para la búsqueda de soluciones a los problemas existentes.

Por último, se evalúa la viabilidad y efectividad de cada una de las soluciones sugeridas y en caso afirmativo se procede a la implementación de las mismas.

### 3.2 Comprensión de los datos.

En esta fase se recopilan y analizan los datos disponibles para el desarrollo del proceso de minería de datos. Los datos deben contener información relacionada con la pregunta que se trata de resolver, si bien es cierto que pueden incorporarse variables de otras áreas del negocio. Lo importante es disponer de una buena muestra representativa de los aspectos más relevantes en el servicio al cliente.

A continuación, se desarrollan los cuatro puntos en que puede desglosarse esta fase:

#### 1. Recopilación de los datos.

La persona responsable del departamento global de servicio al cliente ha facilitado una base de datos que contiene una muestra de la experiencia de los clientes asociada a la gestión de pedidos. A esta base de datos, se han podido agregar nuevas variables procedentes de otras fuentes (tablas de SAP). Estas nuevas variables aportan información relacionada con las características de los clientes y de los pedidos, de manera que permiten un análisis más profundo.

#### 2. Descripción de los datos.

La base de datos principal (la muestra) es un archivo en formato Excel que contiene 1.201 observaciones y 22 variables. Las variables se presentan en la Tabla 1:

Tabla 1. Variables empleadas en el estudio y su descripción.

Variable	Descripción
Sales Order Number	Número de pedido en el sistema
Exclude Reporting	Observación no asociada al servicio comercial
#	Número de observación
Cancellation reason	Razón por la que se cancela el pedido
Region	Grupo de países
Order Creation Date	Fecha de creación del pedido en el sistema
Created Date	Fecha de envío de la encuesta
Created Date	Fecha de recepción de la encuesta
Account Number	Número de cuenta de cliente
Account: Account Name	Nombre de cuenta de cliente
Order Confirmation Email	Email del cliente
CS Survey Average	Media de la puntuación de la encuesta
Customer Service Availability	Disponibilidad del servicio al cliente
Customer Service Friendliness	Amabilidad del servicio al cliente
Customer Service Competence	Competencia del servicio al cliente
Customer Service Reliability	Fiabilidad del servicio al cliente

Product Availability	Disponibilidad del producto
Delivery Time	Plazos de entrega del producto
On-time Delivery	Pedido entregado a tiempo
Entire Order Process	Proceso completo de pedido
Comment	Comentarios del cliente
Comment why excluded	Comentarios del motivo de exclusión

A la base de datos principal se añaden 5 nuevas variables procedentes de tablas del ERP SAP, las variables añadidas se presentan en la Tabla 2:

Tabla 2. Variables adicionales y su descripción.

Variable	Descripción
Sales order value	Importe del pedido en euros
Customer Group	Grupo de cliente
Created by	Agente que generó el pedido en el sistema
Country	País del domicilio fiscal del cliente
Customer Revenue 2021	Facturación al cliente en 2021

### 3. Exploración de los datos.

Se comprueba que el conjunto de datos resultante es apropiado para dar respuesta a la pregunta planteada, ya que los variables contienen información adecuada para valorar el desempeño del servicio al cliente. La muestra de datos con 1.201 observaciones será suficiente para obtener respuestas insesgadas.

### 4. Verificación de la calidad de los datos.

Se tiene la seguridad de saber que el procedimiento de elaboración del conjunto de datos a explotar procede de una muestra representativa del conjunto de datos poblacional. El sistema que envía las encuestas chequea si se ha enviado alguna a un determinado cliente en un determinado plazo de tiempo. Si es así no envía nada. Si por el contrario ha pasado el tiempo suficiente, entonces envía una nueva encuesta. Por este motivo, resultan más representados los clientes que más pedidos envían, independientemente del importe de los mismos. Es sabido, que algunos clientes reciben las encuestas en buzones de correo automatizados que ignoran estos emails (ya que son detectados como spam). Debido a que todas las fuentes de datos proceden de sistemas informáticos, se descartan los errores tipográficos en los valores de las variables. Por lo tanto, se aprueba la calidad de los datos para continuar con el proceso de minería de datos.

## 3.3 Preparación de los datos.

En esta fase, denominada de preprocesamiento, se preparan los datos para ser analizados descriptivamente y posteriormente modelados. En el Anexo se puede consultar el código empleado en R para llevar a cabo la tarea de preprocesamiento. A continuación, se desarrollan los cuatro puntos en que puede desglosarse esta fase:

### 1. Selección de datos.

En este punto se seleccionan las observaciones y las variables que se emplearán en el estudio. Esta selección se realiza porque no todas las variables ni todas las

observaciones son válidas para el estudio. A continuación, se explican los motivos que llevan a desechar observaciones y variables:

Selección de observaciones: se parte de 1.201 observaciones, pero se van a excluir algunas de ellas por diversos motivos:

En primer lugar, se han detectado 37 observaciones duplicadas, triplicadas o más. El motivo es que el cliente ha contestado más de una vez la encuesta para un mismo pedido. Se suprimen del conjunto de datos.

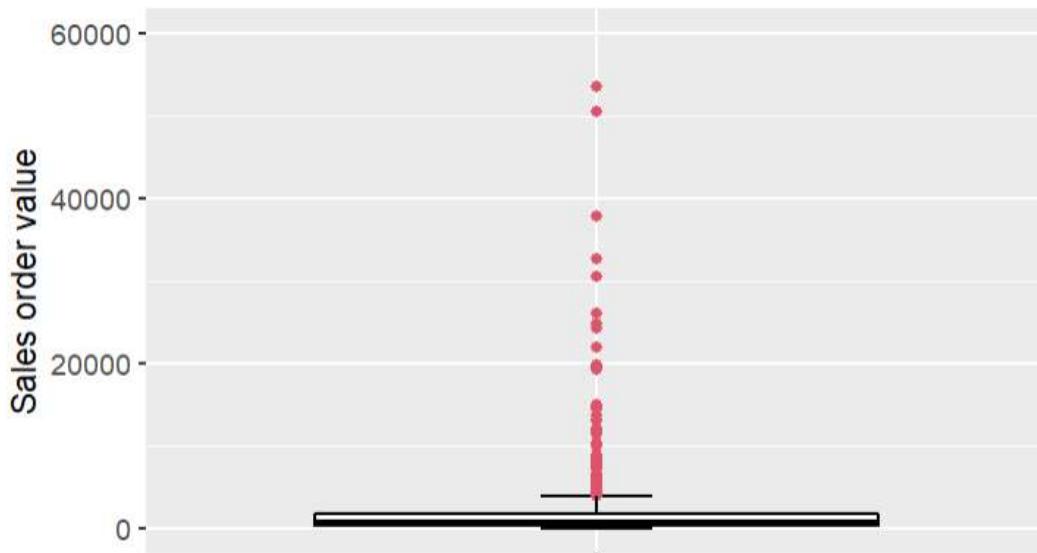
En segundo lugar, se suprimen 71 observaciones en las cuales la variable "Exclude Reporting" tiene el valor "True". Estas observaciones son aquellas en las que se recibió una calificación baja por parte del cliente, pero tras contactar con el cliente, éste aclara que el servicio al cliente es muy bueno y que su baja calificación se debe a incidencias no relacionadas con el servicio al cliente.

En tercer lugar, se suprimen 11 observaciones en las cuales la variable "Cancellation reason" toma el valor "Cancelled", ya que solo resulta de interés incluir en el estudio los pedidos que no fueron cancelados y que fueron gestionados con normalidad de principio a fin abordando toda la experiencia de compra.

En cuarto lugar, se suprimen 128 observaciones en las cuales la variable "Customer Group" toma el valor de "Not assigned", ya que, aunque se trata de una cantidad de observaciones nada despreciable, su eliminación permitirá una mejor clasificación de perfiles de cliente en base a esta característica.

En quinto lugar, se procede a comprobar la existencia de datos extremos u *outliers* en la variable numérica "Sales order value". Se detectan 99 observaciones. En este estudio, se van a conservar puesto que probablemente estén asociadas a grandes cuentas y su eliminación podría afectar a los resultados obtenidos. En la Figura 5 se presenta la distribución de las observaciones en un diagrama de caja que permite identificar los *outliers* con puntos rojos:

Figura 5. Diagrama de caja de la variable Sales Order Value.



El conjunto de datos queda finalmente con 953 observaciones.

Selección de variables: se parte de 27 variables explicativas, pero no todas resultan útiles de incorporar en el conjunto de datos al no aportar información adicional (su inclusión provocaría pérdidas de eficiencia en la estimación). Se suprimen las 11 variables mostradas en la Tabla 3:

Tabla 3. Variables suprimidas del conjunto de datos.

Variable	Descripción	Motivo de eliminación
Sales Order Number	Número de pedido en el sistema	No aporta información de utilidad
Exclude Reporting	Observación no asociada al servicio comercial	Tras la selección solo tiene un valor: "False"
#	Número de encuesta	No aporta información de utilidad
Cancellation reason	Razón por la que se cancela el pedido	Tras la selección solo tiene un valor: "Not cancelled"
Created Date	Fecha de envío de la encuesta	Se dispone de la variable "Order Creation Date"
Created Date	Fecha de recepción de la encuesta	Se dispone de la variable "Order Creation Date"
Account: Account Name	Nombre de cuenta de cliente	Se dispone de la variable "Account Number"
Order Confirmation Email	Email del cliente	No aporta información de utilidad
CS Survey Average	Media de la puntuación de la encuesta	Es combinación lineal de las siguientes 8 variables
Comment	Comentarios del cliente	Datos no estructurados
Comment why excluded	Comentarios del motivo de exclusión	Datos no estructurados

Por tanto, tras la selección de datos, el conjunto de datos resulta en un total de 953 observaciones y 16 variables explicativas.

## 2. Transformación de variables.

En ocasiones resulta apropiado crear nuevas variables a partir de las existentes. Por un lado, se dispone de la variable "Order Creation Date" que es una variable que contiene la fecha de la creación del pedido en el sistema. Esta variable es transformada a una variable nominal con únicamente dos categorías: "NO PANDEMIA" para las observaciones con fecha de pedido comprendida entre la observación más antigua y el 31 de Enero de 2020, y la categoría "PANDEMIA" para las observaciones con fecha de pedido comprendida entre el 1 de Febrero de 2020 y la observación más reciente del conjunto de datos. Esta transformación permitirá detectar los posibles cambios en la experiencia del cliente antes y desde la pandemia. En la Tabla 4 se muestra la transformación:

Tabla 4. Transformación de la variable nominal Order Creation Date en una variable nominal binaria.

Variable	Fecha	Categoría
Order Creation Date	Observaciones hasta el 31.01.2020	NO PANDEMIA
Order Creation Date	Observaciones desde el 01.02.2020	PANDEMIA

En segundo lugar, se procede a transformar la variable numérica "Customer Revenue 2021" en una variable nominal de tipo ordinal con 6 categorías según se muestra en la Tabla 5:

Tabla 5. Transformación de la variable cuantitativa Customer Revenue 2021 en una variable nominal con 6 niveles.

Variable	Valor numérico	Categoría
Customer Revenue 2021	Más de 500.000 €	OVER 500K
Customer Revenue 2021	De 200.000 a 500.000 €	200K TO 500K
Customer Revenue 2021	De 100.000 a 200.000 €	100K TO 200K
Customer Revenue 2021	De 50.000 a 100.000 €	50K TO 100K
Customer Revenue 2021	Desde 1 a 50.000 €	1 EUR TO 50K
Customer Revenue 2021	0 €	ZERO

Esta transformación permite segmentar los clientes en base al volumen total de negocio que tuvieron durante el año 2021; así es posible facilitar la interpretación de los siguientes análisis y ayudar a la implementación de medidas de mejora en el servicio al cliente según su volumen de negocio.

### 3. Formato de datos.

El conjunto de datos o *dataframe* contiene 9 variables numéricas y 7 variables nominales. Las variables numéricas no necesitan ninguna transformación, sin embargo, las variables nominales son transformadas a variables cuantitativas para poder ser empleadas en el proceso de modelización. Para realizar esta transformación, las variables cualitativas de tipo “texto” son transformadas en R a variables de tipo “factor”. Además, es importante distinguir entre las variables de tipo “factor” nominales (no pueden ser ordenadas) de las ordinales (pueden ser ordenadas). En la Tabla 6 se muestra la asignación por variable:

Tabla 6. Tipo y subtipo de las variables del conjunto de datos.

Variable	Tipo	Subtipo
Sales order value	Numérica	Decimal
Created by	Nominal	Texto
Region	Nominal	Texto
Order Creation Date	Nominal	Texto
Account Number	Nominal	Texto
Country	Nominal	Texto
Customer Group	Nominal	Texto
Customer Revenue 2021EUR	Nominal	Ordinal
1. Customer Service Availability	Numérica	Entero
2. Customer Service Friendliness	Numérica	Entero
3. Customer Service Competence	Numérica	Entero
4. Customer Service Reliability	Numérica	Entero
5. Product Availability	Numérica	Entero
6. Delivery Time	Numérica	Entero
7. On-time Delivery	Numérica	Entero
8. Entire Order Process	Numérica	Entero

### 3.4 Modelos teóricos.

Los resultados estadísticos descriptivos son útiles para obtener algunas conclusiones generales, pero insuficientes para extraer información adicional. Aquí entra en juego la microeconometría, que se presenta como una herramienta muy potente para la detección de relaciones más profundas entre variables de la muestra.

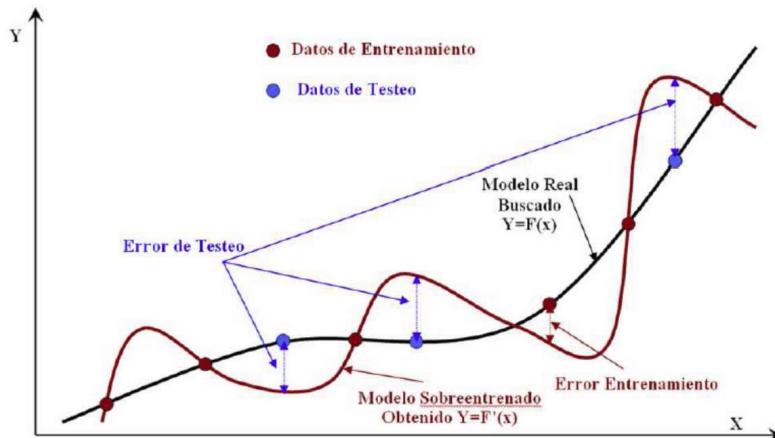
Un modelo microeconómico no es sino un modelo econométrico empleado sobre una muestra de datos microeconómicos, de manera que se estudia el comportamiento de los agentes involucrados de manera individual con el objetivo de cuantificar preferencias o gustos en la toma de decisiones entre varias alternativas, así como poder detectar efectos causales de unas variables sobre otras, aunque también pueden emplearse con fines predictivos.

Para contestar la pregunta de interés apoyándonos en información cuantitativa se tiene que especificar un modelo econométrico adecuado, para, posteriormente proceder a su estimación, y a la evaluación e interpretación de los resultados obtenidos.

Pero, antes de realizar las especificaciones y estimaciones de diferentes modelos sobre el conjunto de datos, se realizan una serie de actividades sobre el conjunto muestral de modo que se optimice el método de clasificación. Para ello, resulta muy conveniente emplear un conjunto de datos de entrenamiento y un conjunto de datos de test.

En este estudio se emplea el 80% de los datos muestrales para elaborar el conjunto de datos de entrenamiento, y se deja el 20% restante para el conjunto de datos de test. El conjunto de datos de entrenamiento sirve para estimar los parámetros del modelo mientras que el conjunto de datos de test se utiliza para comprobar el comportamiento de los modelos. Cada observación sólo aparece en uno de los conjuntos, y la división del conjunto muestral entre ambos subgrupos se realiza por muestreo aleatorio simple.

Figura 6. Ejemplo de comparación de un modelo real buscado con un modelo entrenado.

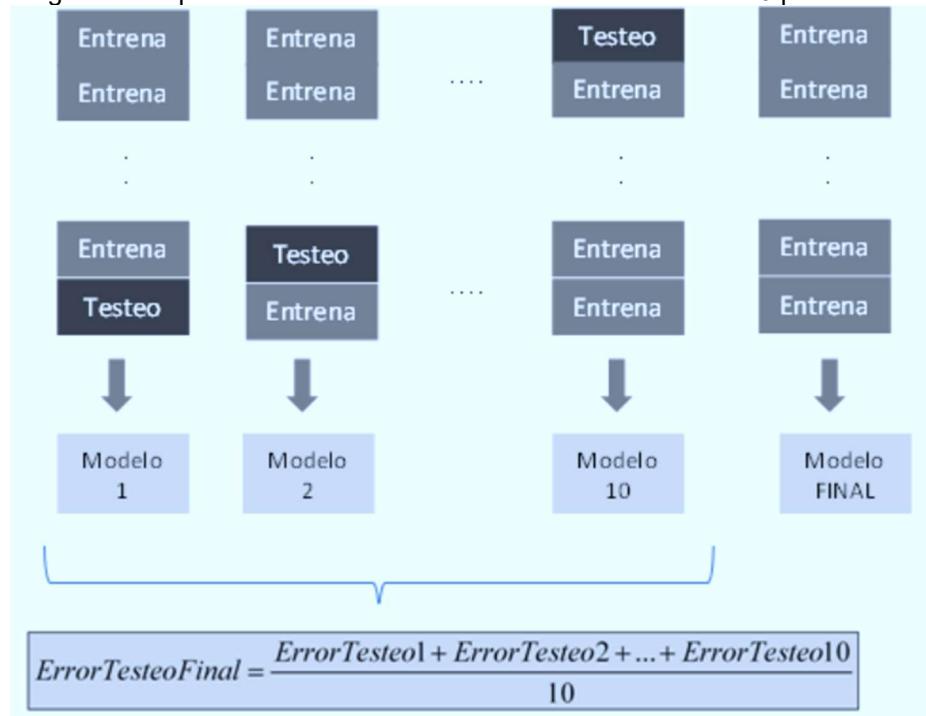


Fuente: Vicente et al., (2019).

La Figura 6 muestra como el modelo entrenado trata de aproximarse lo máximo posible al modelo real buscado. El conjunto de datos de test evalúa la eficiencia del modelo entrenado y permite calcular las diferencias entre la predicción del modelo y el valor real. El ejemplo representado en la Figura 6 muestra un modelo sobreentrenado que oscila demasiado sobre la función real a modelizar.

Lo ideal para especificar un modelo óptimo sería que se dispusiese de un conjunto de datos de entrenamiento independiente del conjunto de datos de test, pero en el caso que nos ocupa esto no es posible, ya que la fuente de datos es la misma para ambos subconjuntos. En esta situación, resulta muy adecuado emplear un esquema de validación cruzada, de manera que se procede a dividir el conjunto de entrenamiento en 5 particiones, de manera que el modelo se vaya entrenando y evaluando con las particiones restantes. El resultado final se obtiene por agregación de los resultados originales. La Figura 7 presenta un esquema teórico de validación cruzada con 10 particiones:

Figura 7. Esquema de un modelo de validación cruzada con 10 particiones.



Fuente: Vicente et al., (2019)

Una vez que se tienen preparados los conjuntos de datos entrenamiento y de test, se procede a la especificación del modelo de clasificación sobre el que se realizará la posterior estimación de varios modelos económicos para detectar las características que determinan si la experiencia del cliente en relación con la gestión de los pedidos es excelente (5 puntos sobre 5 en la variable Entire Order Process) o no lo es (valores entre 1 y 4). Es decir, la variable dependiente o explicada es la variable dicotómica “Entire Order Process”; siendo las variables independientes o explicativas el resto de variables disponibles en el conjunto de datos preprocesado a excepción de “Created by”, “Account Number” y “Country” por ser variables nominales con demasiados niveles que perjudicarían la estimación del resto de variables.

### 3.4.1 Modelo Lineal Generalizado.

El modelo lineal generalizado (Nelder y Wedderburn, 1972) es, tal y como su nombre indica, una generalización del modelo de regresión lineal simple. A diferencia del modelo lineal simple, el modelo lineal generalizado no necesita cumplir las hipótesis de distribución normal de los residuos, ni la hipótesis de varianza constante, así como

tampoco debe satisfacer la condición de relación lineal entre la variable dependiente y las independientes.

El uso de los modelos lineales generalizados resulta muy conveniente en la mayoría de las ocasiones, puesto que lo habitual es trabajar con conjuntos de datos que no cumplen las hipótesis del modelo lineal debido a la naturaleza de la información. Y, además, muchas veces las transformaciones monótonas de las variables no consiguen corregir la falta de normalidad, la heterocedasticidad o la no linealidad.

En síntesis, los modelos lineales generalizados son una extensión de los modelos lineales que permiten utilizar y explotar la información incluso cuando no se cumplen las hipótesis básicas y obtener resultados válidos.

#### Modelo Lineal Generalizado de tipo logit.

El modelo logit (Berkson, 1944) resulta muy adecuado cuando se pretende estimar un modelo econométrico con variable dependiente binaria. A diferencia del modelo de regresión lineal, el modelo logit emplea la función de distribución acumulada logística. Esta función se encuentra acotada entre 0 y 1; además, presenta un crecimiento no lineal con mayores incrementos en la parte central (Matilla et al., 2017).

#### Modelo Lineal Generalizado de tipo probit.

El modelo probit (Fechner, 1860) también resulta adecuado cuando se pretende estimar un modelo econométrico con variable dicotómica. El modelo probit es muy parecido al modelo logit. Su diferencia se haya en que emplea la función de densidad acumulada de una normal tipificada, por lo que esta función tiene las colas algo más anchas que la función de distribución logística, por lo que las probabilidades para los valores extremos (próximos a 0 y 1) son algo mayores (Matilla et al., 2017).

### **3.4.2 Modelos de árbol de decisión.**

Los algoritmos que se emplean en los modelos de árbol no están basados en la estimación de parámetros de una ecuación especificada. Su funcionamiento a la hora de clasificar se apoya en la sucesión de particiones (binarias o más) en los valores de una variable cada vez. Estas escisiones maximizan las diferencias de la variable dependiente, de forma que se van formando grupos homogéneos dentro de la población.

Los árboles de decisión constan de nodos, ramas y hojas. Los nodos representan a las variables de entrada, las ramas son los posibles valores de la variable dependiente, y las hojas son los valores de salida. Además, el nodo principal o raíz es la variable independiente más relevante en el proceso de clasificación ya que desde este nodo se van produciendo las escisiones (ramas) de manera decreciente en relevancia hasta llegar a los valores de salida finales (hojas).

#### Árbol CART (Classification and Regression Trees).

El árbol CART (Breiman, 1984). Con este algoritmo se generan árboles de decisión binarios, es decir, cada nodo se divide en dos ramas. El algoritmo utiliza la medida de impureza (grado de homogeneidad entre variables). Cada rama se subdivide cada vez en dos ramas, colocando a la izquierda los resultados que cumplen la condición y a la derecha los resultados que no cumplen la condición. De esta forma se va construyendo el árbol hasta que no hay diferencias significativas entre el resto de las variables, de forma que se conforman grupos homogéneos.

### Árbol C5.0

El modelo de árbol C5.0 (Quinlan, 1992) se basa en un algoritmo para generar árboles de decisión. Este modelo construye el árbol de decisión utilizando los datos de entrenamiento bajo el concepto de entropía de la información. El criterio para dividir eficazmente el conjunto de datos es el normalizado para ganancia de información (diferencia de entropía), de modo que la variable con mayor información se establece como nodo principal o raíz, desde el cual se generan las siguientes ramas en función nuevamente de la ganancia de información que aportan.

### Árbol random forest.

El árbol random forest (Breiman, 2001) es un algoritmo que se basa en la combinación de árboles de decisión independientes que utilizan diferentes vectores de muestreo aleatorio procedentes del mismo conjunto de datos que tienen la misma distribución para todos los árboles empleados en el algoritmo.

### **3.4.3 Modelos redes neuronales artificiales.**

Una red neuronal artificial (Cybenko, 1989; Funahashi, 1989; Hornik et al., 1989) no es sino un modelo que imita al cerebro humano de manera simplificada. Su funcionamiento se basa en un número elevado de unidades de procesamiento que se hayan interconectadas y que gestionan los inputs recibidos en la capa de entrada para ser tratados en la capa oculta y devolver un output en la capa de salida. Las redes neuronales se caracterizan por trabajar bien con bases de datos con ruido o incompletas, así como por tener una alta tolerancia a fallos.

#### Red neuronal perceptrón multicapa.

El modelo de red neuronal de perceptrón multicapa (Rumelhart et al., 1986) es de propósito general, y es capaz de organizar una representación interna del conocimiento en las capas ocultas a fin de aprender la relación entre los conjuntos de datos de entrada y de salida.

#### Red neuronal de función de base radial.

En las redes neuronales de función de base radial (Moody y Darken, 1989) el output disminuye o aumenta de forma monótona con la distancia a un punto fijo llamado centroide. La diferencia más importante con respecto al modelo de perceptrón multicapa reside en cómo se procesa la información en la capa oculta de la red, de modo que la función de base radial calcula la distancia entre el input y el centroide, y a esa diferencia le aplica una función radial con forma gaussiana.

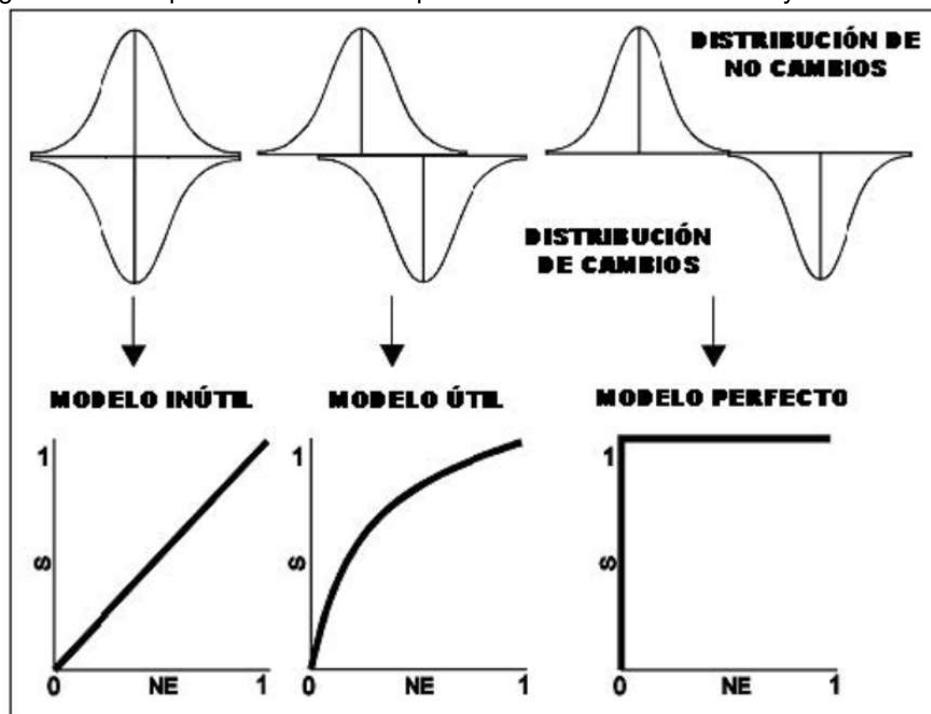
### **3.4.4 Modelo Naïve Bayes.**

El algoritmo Naïve Bayes (Bayes, 1763) es un modelo clasificador que parte de la suposición de que todas las variables son independientes conocido el valor de la variable dependiente.

### 3.5 Evaluación.

Para comparar los modelos entre sí, y decidir cuál es el que presenta un mejor desempeño, debe utilizarse la misma métrica comparativa para todos ellos. En cualquier proceso de minería de datos resulta fundamental poder estimar los niveles de calidad de los modelos estimados, esto es encontrar el modelo que cometa el menor error de predicción. El área bajo la curva ROC (Receiver Operating Characteristic) es una buena métrica para la comparación de modelos de clasificación ya que muestra la distribución de las fracciones de verdaderos positivos y la fracción de falsos positivos. Es decir, el AUC (Area Under Curve) puede tomar dos valores extremos, uno de ellos es 0,5 y este valor se produce cuando la capacidad predictiva es nula ya que clasifica correctamente el 50% de las observaciones, que es la misma probabilidad que si la clasificación se realizase al azar. El otro caso es que el AUC tenga un valor exactamente de 1, esta situación es la idónea puesto que el modelo es capaz de clasificar correctamente el 100% de las observaciones del conjunto muestral.

Figura 8. Correspondencia entre solapamiento de las distribuciones y la curva ROC.



Fuente: Vicente et al., (2019).

La Figura 8 muestra las distintas formas que puede tener una curva ROC siendo el AUC el dato numérico que proporciona el desempeño global del modelo. El criterio del área AUC permite comparar entre sí todos los modelos estimados y elegir el que logra el mejor desempeño (mayor valor de AUC).

## **4. DATOS Y RESULTADOS.**

---

En este punto, se dispone de un conjunto de datos adecuado (sin datos anómalos, sin datos ausentes, con las variables transformadas convenientemente y con un número de observaciones suficiente) para comenzar con el análisis descriptivo para la obtención de resultados. En el Anexo se puede consultar el código empleado en R.

### **4.1 Análisis descriptivo de los datos.**

En la Tabla 7 se muestra un resumen de la información descriptiva por variable:

Tabla 7. Sumario de datos estadísticos del conjunto de datos.

Data summary										
Name								CX		
Number of rows								953		
Number of columns								16		
Column type frequency:										
factor								7		
numeric								9		
Group variables								None		
Variable type: factor										
skim_variable	n_missing	complete_rate	ordered		n_unique	top_counts				
Created by	0	1	FALSE		33	BAC: 186, WEB: 143, SPI: 91, ZOO: 66				
Region	0	1	FALSE		5	DAC: 443, Nor: 152, Sou: 139, UK: 119				
Order Creation Date	0	1	FALSE		2	NO : 499, PAN: 454				
Account Number	0	1	FALSE		295	100: 94, 100: 38, 100: 31, 150: 26				
Country	0	1	FALSE		44	DE: 242, IE: 109, AT: 83, SE: 73				
Customer Group	0	1	FALSE		9	Dir: 334, Sec: 269, Sie: 197, VAP: 72				
Customer Revenue 2021	0	1	FALSE		6	1 E: 454, OVE: 159, 200: 135, ZER: 73				
Variable type: numeric										
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Sales order value	0	1	1893.67	4596.33	6.6	275.96	649.95	1728.89	68164.58	
Customer Service Availability	0	1	4.69	0.58	1.0	4.00	5.00	5.00	5.00	
Customer Service Friendliness	0	1	4.80	0.48	1.0	5.00	5.00	5.00	5.00	
Customer Service Competence	0	1	4.73	0.52	1.0	5.00	5.00	5.00	5.00	
Customer Service Reliability	0	1	4.73	0.52	1.0	5.00	5.00	5.00	5.00	
Product Availability	0	1	4.31	1.00	1.0	4.00	5.00	5.00	5.00	
Delivery Time	0	1	4.44	0.94	1.0	4.00	5.00	5.00	5.00	
On-time Delivery	0	1	4.55	0.83	1.0	4.00	5.00	5.00	5.00	
Entire Order Process	0	1	4.58	0.74	1.0	4.00	5.00	5.00	5.00	

De la Tabla 7 se puede extraer mucha información descriptiva. Así, de las variables de tipo factor, se observa que en la muestra están representados 33 agentes de introducción de pedidos, 5 regiones de negocio, 2 momentos temporales (antes y

después del inicio de la pandemia), 295 clientes, 44 países, 9 grupos de clientes en base a su área de negocio y 6 tipos de clientes en función de su facturación durante el año 2021. En los que respecta a las variables de tipo numérico se observa que el pedido medio (variable “Sales order value”) tiene un valor de 1.894 euros y que todas las puntuaciones de las variables de la encuesta tienen una media superior a 4,3 puntos sobre 5. En lo que respecta a la mediana o valor central, que coincide con el percentil 50, se observa un valor de pedido de 650 euros, y, en las variables de la encuesta, una mediana de 5 puntos que además coincide con el valor de la moda (valor más frecuente).

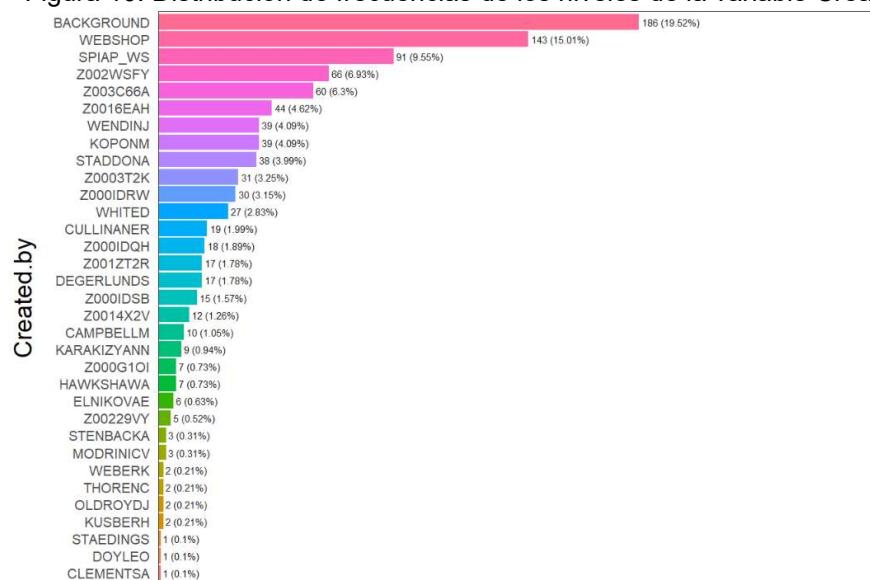
A continuación, en la Figura 9, se muestran los histogramas o distribución de frecuencias de las variables numéricas y se confirman los datos estadísticos obtenidos y resaltados en color en la Tabla 7:

Figura 9. Histogramas de las variables numéricas del conjunto de datos.



En la Figura 10 se muestra la distribución de frecuencias de la variable de tipo factor “Created by”:

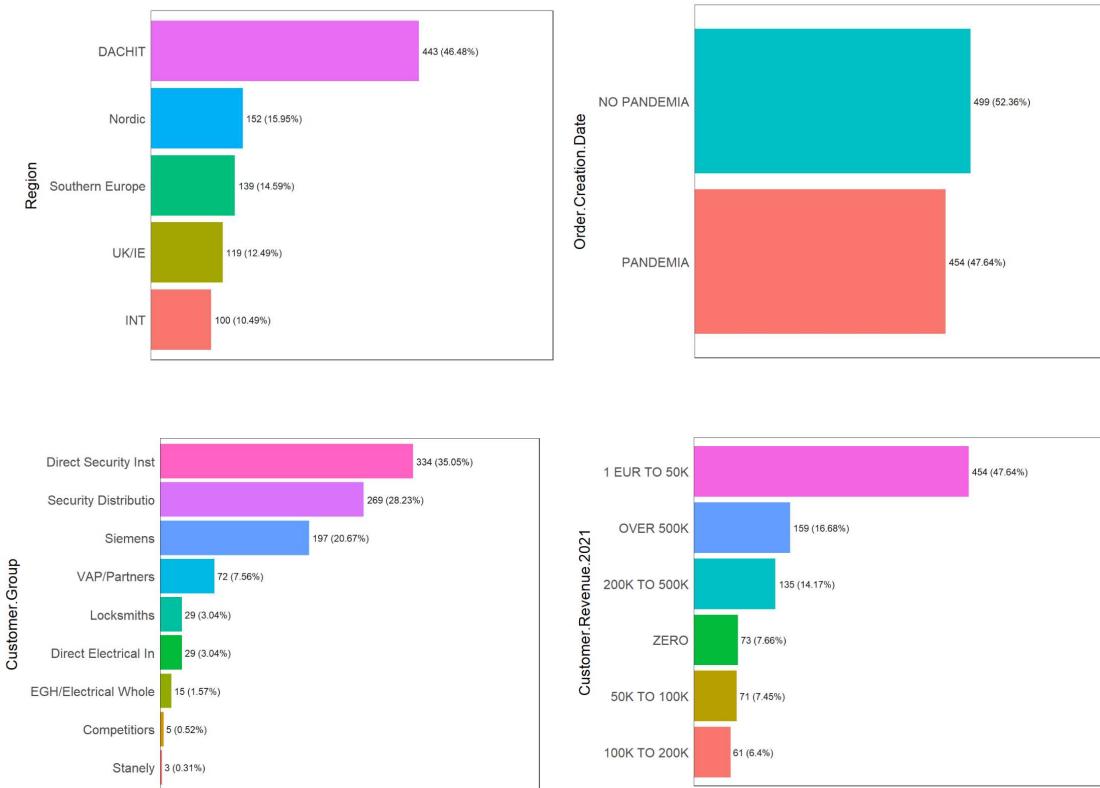
Figura 10. Distribución de frecuencias de los niveles de la variable Created by.



Se observa que prácticamente el 20% de las observaciones se corresponden con “Background” (pedidos introducidos de forma automatizada desde el ERP del cliente hacia el ERP de la empresa), y que un 15% de los pedidos se han introducido en el ERP de forma automatizada a través de la tienda online “Webshop”. Resaltar también que hay muy pocas observaciones de pedidos introducidos mediante el RPA “Staedings”, puesto que se está implementando gradualmente desde finales de 2021 y lamentablemente, el tener solo una observación no permite obtener conclusiones con respecto a su desempeño de cara a la satisfacción de los clientes.

En la Figura 11 se muestran las distribuciones de frecuencias de las variables de tipo factor “Region”, “Order Creation Date”, “Customer group” y “Customer Revenue 2021”:

Figura 11. Distribución de frecuencias de las variables Region, Order Creation Date, Customer group y Customer Revenue 2021.



Se observa que casi el 50% de las observaciones pertenecen a la región “DACHIT”; con respecto al momento temporal se observa que la distribución de la muestra es próxima al 50% para cada uno de los dos momentos (No pandemia y pandemia); en lo que se refiere al “Customer Group”, se observa que un 35% de los clientes son “Direct Security Installers”, un 28% son “Security Distributors” y un 21% son “Siemens”. Con respecto al volumen de facturación que tuvieron los clientes en 2021 se observa que el 47% de las observaciones se asocian a pequeños clientes que facturaron entre 1 y 50.000 euros; curiosamente, el siguiente grupo de clientes con más observaciones registradas se corresponde al grupo de clientes más grandes que facturaron más de 500.000 euros. Resaltar también que un 7% de las observaciones se relacionan con clientes que no facturaron en 2021.

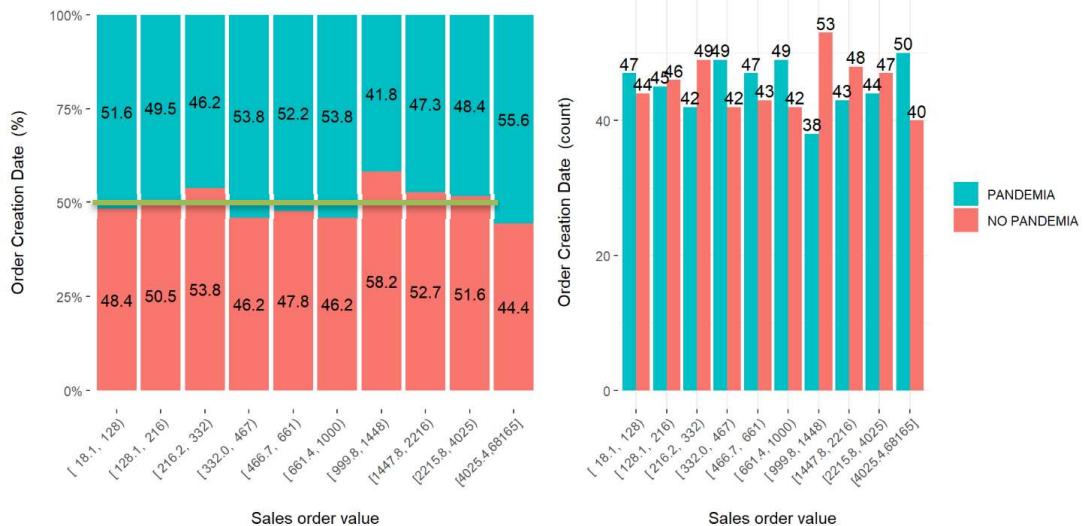
Hasta aquí se ha presentado el análisis descriptivo general de los datos de la muestra. Esto es muy útil para obtener una idea general de la distribución de los datos y poder abordar, a continuación, un estudio descriptivo más detallado. En primer lugar, se procede a realizar un equilibrado de la muestra en relación con la variable factor

“Order Creation Date”, que tiene 2 niveles: “Pandemia” y “No Pandemia”. El equilibrado, permitirá clasificar correctamente la evolución de las variables en función del tiempo, y así detectar cambios de patrón o nuevas tendencias en los datos como consecuencia de la pandemia. El conjunto muestral preprocesado contiene 953 observaciones, de las cuales 499 tienen como valor “No Pandemia”, y el resto, 454 observaciones, tienen el valor “Pandemia”. Para equilibrar la muestra, se reduce el número de observaciones con valor “Pandemia” de 499 observaciones a 454. Esta reducción se realiza eliminando 45 observaciones con valor “No Pandemia” de forma aleatoria.

Los rangos de valores en que las variables se encuentran en un porcentaje de frecuencia próximo al 50% para ambos momentos temporales indican que las variables no han sufrido un cambio debido a la nueva situación por la pandemia; por otro lado, los rangos de valores que tienen un reparto menos equilibrado (por ejemplo 45% – 55% o 60% – 40%), sugieren que se han producido significativos en la distribución de las variables.

Se analiza la diferencia entre los dos períodos en algunas de las variables:

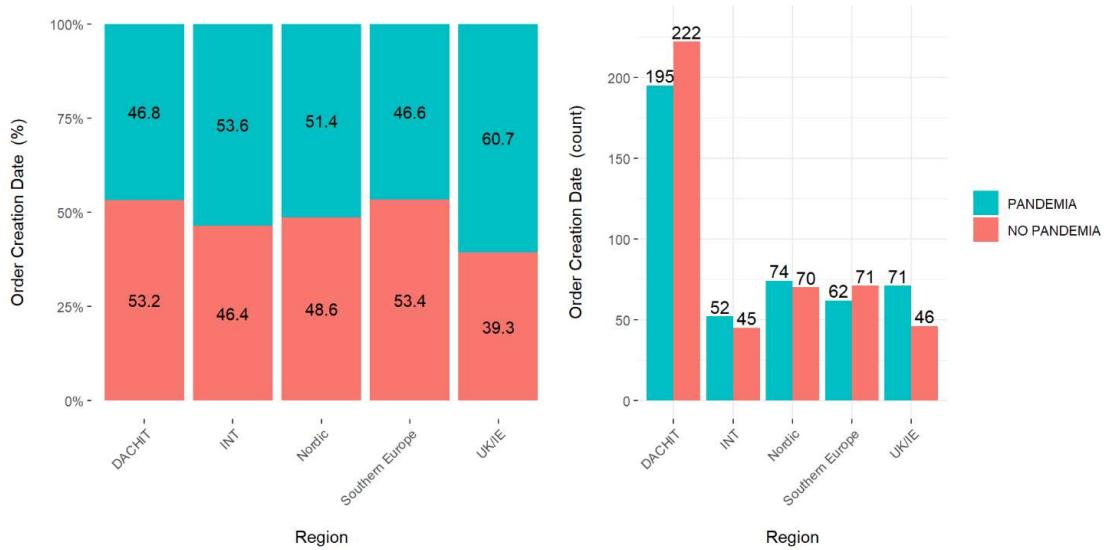
Figura 12. Distribución de frecuencias de la variable Sales Order Value en función de los niveles de la variable Order Creation Date.



En la Figura 12 se observa la distribución de frecuencias y en porcentaje de la variable “Sales order value”. En general no se aprecia un nuevo patrón o tendencia en el tamaño de los pedidos de los clientes, ya que, si se observan los valores desde 18 hasta 4.025 euros, la distribución por tamaño de pedidos oscila alrededor del 50% (se representa una línea horizontal al nivel del 50%). Sin embargo, se percibe una marcada expansión en el rango entre 4.025 y 68.000 euros durante la pandemia (con un 55,6% de observaciones en el periodo de “Pandemia” frente a un 44,4% en el periodo de “No Pandemia”), que pueden deberse a diversos motivos. Algunas hipótesis del autor son:

- Mayor planificación y acopio de material por parte de las empresas más grandes.
- Recuperación económica y reinicio de actividad empresarial asociada a instalaciones o proyectos de tamaño medio o grande.
- Agrupación de pedidos por acumulación de trabajos(instalaciones por realizar).

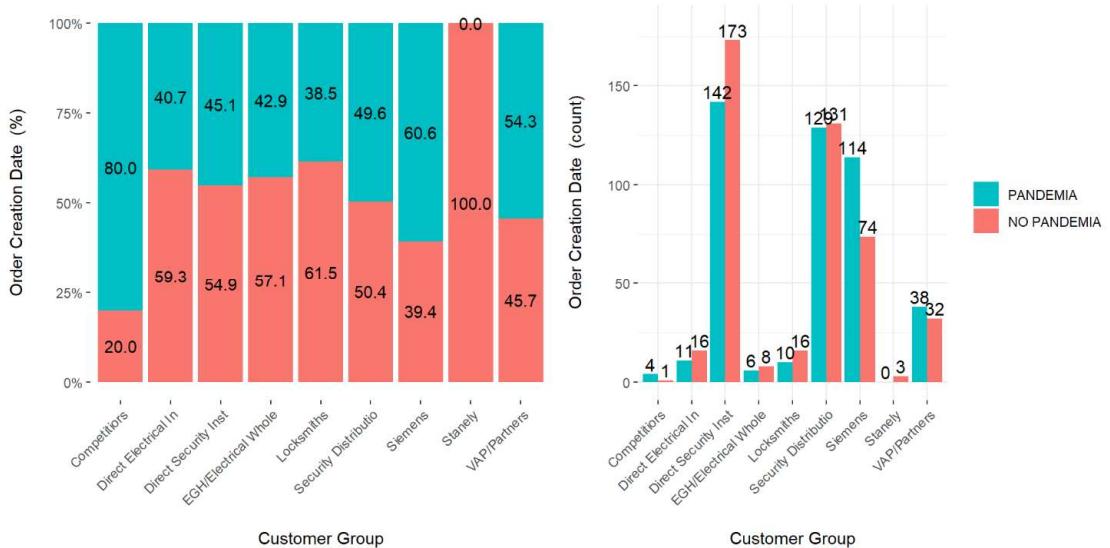
Figura 13. Distribución de frecuencias de la variable Region en función de los niveles de la variable Order Creation Date.



En la Figura 13 se muestra la distribución de los pedidos por “Region”. Es remarcable la contracción de la región DACHIT con un reparto del 53,2% de observaciones antes de la pandemia y un 46,8% de observaciones tras la pandemia, así como la expansión de los pedidos en la región UK/IE (Reino Unido e Irlanda) desde la pandemia con un reparto del 39,3% de observaciones antes de la pandemia y un 60,7% de observaciones tras la pandemia.

A continuación, se muestra la evolución en la variable “Customer Group”:

Figura 14. Distribución de frecuencias de la variable Customer Group en función de los niveles de la variable Order Creation Date.



En la Figura 14 destacan la contracción de los “Direct Security Inst” con un reparto de 54,9% de observaciones antes de la pandemia y un 45,1% de observaciones tras la pandemia y la expansión de “Siemens” con un reparto de 39,4% de observaciones antes de la pandemia y un 60,6% de observaciones desde la pandemia. Los clientes dentro del grupo “Security Distribution” se mantienen estables.

A continuación, se muestra la evolución en la variable “Customer Revenue 2021”:

Figura 15. Distribución de frecuencias de la variable Customer Revenue 2021 en función de los niveles de la variable Order Creation Date.



La Figura 15 muestra la evolución de la variable “Customer Revenue 2021”, y en ella destaca las 59 observaciones con valor “ZERO”; esto son clientes que tuvieron facturación en los años anteriores a 2021, pero que en 2021 no facturaron. Posiblemente sean empresas que no pudieron transformar sus procesos productivos a la nueva situación mundial provocada por la pandemia o no tuvieron capacidad económica para superar la crisis, aunque también podría deberse a pérdida de clientes por no cumplir expectativas. Con el objetivo de indagar cuales son las posibles causas de la pérdida de clientes se genera la Tabla 8 que muestra la variable “Customer Group” al que pertenecen estas 59 observaciones y las medias de sus valoraciones en la encuesta al cliente:

Tabla 8. Recuento de observaciones y valoración media recibida del grupo Zero.

Customer Group	Count	CS_Av	CS_Fr	CS_Co	CS_Re	Prod_Av	Del_Time	Ontime_del	Entire_Pr
Direct Electrical In	4	4,25	4,75	4,50	4,00	3,75	4,50	4,25	4,25
Direct Security Inst	28	4,39	4,54	4,50	4,46	4,43	4,50	4,50	4,57
Locksmiths	6	4,17	4,67	4,83	4,83	3,33	3,50	4,50	4,17
Security Distributio	11	4,45	4,64	4,73	4,55	4,91	4,73	4,73	4,55
VAP/Partners	10	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00

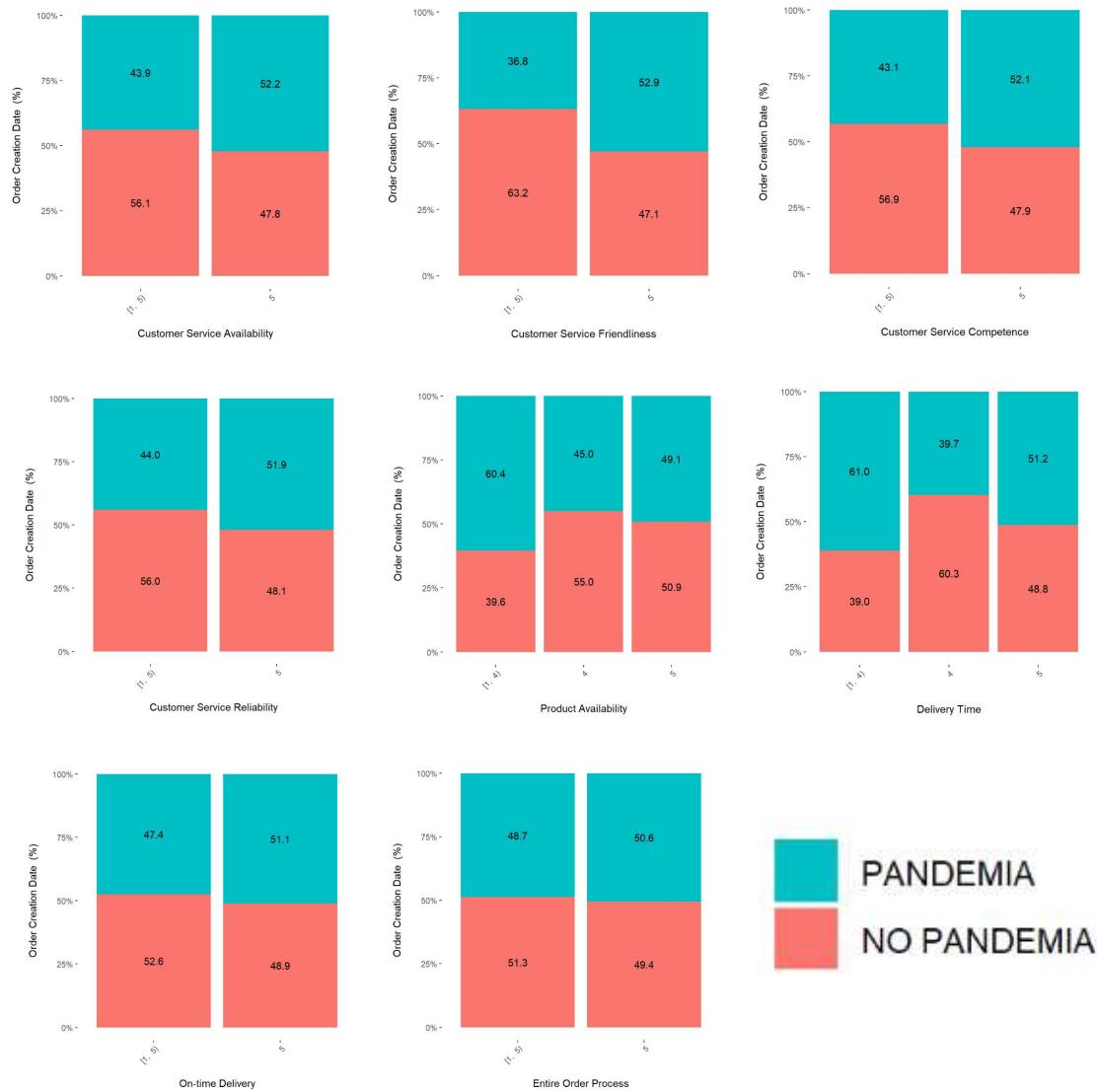
Nota: CS Av (Customer Service Availability), CS Fr (Customer Service Friendliness), CS Co (Customer Service Competence), CS Re (Customer Service Reliability), Prod Av (Product Availability), Del Time (Delivery time), Ontime del (On-time delivery) y Entire Pr (Entire Order Process).

Se observa que no hay relación entre la no facturación en el año 2021 y las valoraciones de la encuesta de satisfacción al cliente recibida en los años anteriores, ya que, en general, todas las puntuaciones son superiores a 4 puntos sobre 5, y únicamente se han detectado algunas puntuaciones inferiores a 4 puntos (sombreadas en la Tabla 8) que se relacionan con la disponibilidad del producto y con el tiempo de entrega. Por tanto, es bastante improbable que estos clientes se hayan perdido por algún motivo asociado a la calidad del servicio al cliente recibido.

La Figura 16 muestra las 8 variables relacionadas con la encuesta de satisfacción del cliente. Se observa que las variables de “Customer Service Availability”, “Customer Service Friendliness”, “Customer Service Competence” y “Customer Service Reliability” del equipo de servicio al cliente experimentan una mejora en la puntuación desde el inicio de la pandemia. En cuanto a las variables “Product Availability” y “Delivery time” han experimentado una contracción en la puntuación recibida; en concreto, para el rango de puntuaciones entre 1 y 3 puntos sobre 5, se tiene un reparto aproximado del

40% de observaciones antes de la pandemia y un 60% de observaciones desde la pandemia. Esto se debe a la gran crisis de los semiconductores (escasez mundial de chips) que impacta directamente en los niveles de producción y por tanto en la cadena de suministro de los productos del portfolio de la empresa estudiada. En cualquier caso, el número de instancias con el rango de valor entre 1 y 3 de estas dos variables son pequeñas en relación con el número total de casos, por lo que la variación no resulta grande en términos absolutos.

Figura 16. Distribución de frecuencias de las variables.

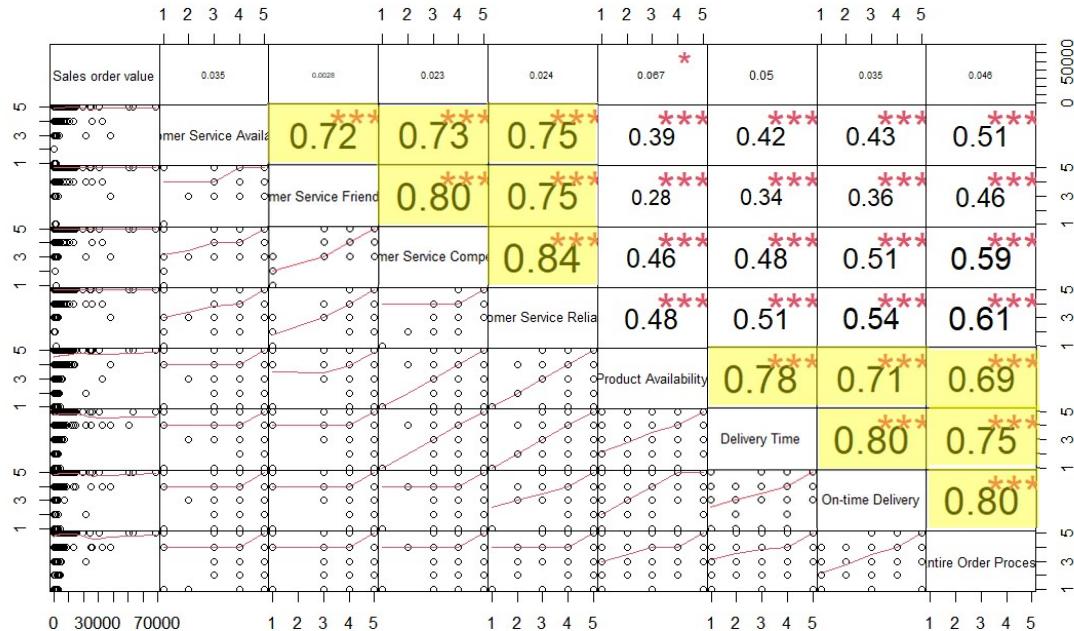


Finalmente, la variable “Entire Order Process” ha experimentado una pequeña mejora desde el inicio de la pandemia (50,6 a 49,4). Pero como el conjunto de datos no es muy grande y la variación es muy pequeña, no se puede garantizar que la *customer experience* haya mejorado desde el inicio de la pandemia.

En la Tabla 9 se presentan las diferentes correlaciones entre las variables numéricas analizadas. Como se puede observar, no hay relación significativa entre la variable “Sales order value” y las variables de la encuesta de satisfacción del cliente, y es que, además, los valores de correlación son próximos a cero (no hay relación positiva ni negativa). Por el contrario, se observa que hay una relación positiva entre todas las variables de la encuesta de satisfacción al cliente. En concreto, las variables con una

correlación superior a 0,7 tienen una fuerte asociación (resaltadas en color amarillo). Así pues, se diferencian dos grandes grupos de variables correlacionadas.

Tabla 9. Matriz de correlaciones entre variables numéricas.



Nota: (\*\*\*) nivel de significación entre las variables del 99%, (\*\*) nivel de significación entre las variables del 95%, (\*) nivel de significación entre las variables del 90%, () Variables no significativas.

El primer grupo está formado por las cuatro variables Customer Service: Availability, Friendliness, Competence y Reliability. La correlación entre estas variables es grande, positiva y significativa, y además no están correlacionadas fuertemente con la variable “Entire Order Process” (0,51, 0,46, 0,59 y 0,61 respectivamente) lo cual significa que, aun teniendo relación positiva con la valoración final de todo el proceso de compra, no son las variables con mayor influencia sobre la puntuación global o “Entire Order Process”.

El segundo grupo está formado por las variables “Product Availability”, “Delivery Time” y “On-time Delivery”. La correlación entre estas variables es grande, positiva y significativa, y además están correlacionadas fuertemente con la variable “Entire Order Process” (0,69, 0,75 y 0,8 respectivamente) lo cual significa que son las variables con mayor influencia sobre la puntuación global o “Entire Order Process”.

Es decir, la valoración o puntuación final (“Entire Order Process”) que asignan los clientes en la encuesta está relacionada con todas las puntuaciones que han asignado al resto de variables de la encuesta, pero, está especialmente representada por las variables “Product Availability”, “Delivery Time” y “On-time Delivery”. En concreto, la entrega de los productos en el plazo determinado (“On-time Delivery”) tiene la correlación más alta de todas las variables con “Entire Order Process”, con un valor de 0,8. Esto significa que la característica más importante para los clientes en el proceso de compra es cumplir con las fechas de entrega proporcionadas en las confirmaciones de pedido.

## 4.2 Modelización.

Antes de estimar los modelos de clasificación resulta muy conveniente equilibrar los niveles que tiene la variable dependiente “Entire Order Process”. En el conjunto de datos se recuentan 656 casos en los que la variable dependiente toma el valor “EXCELENTE” y 297 casos en los que toma el valor “NO\_EXCELENTE”. Esta diferencia provoca que los clasificadores estén sesgados a predecir un porcentaje más elevado a la clase con más observaciones. Por ello, se realiza un submuestreo de los datos con valor “EXCELENTE” para dejar el conjunto muestral equilibrado.

### 4.2.1 Modelo Lineal Generalizado.

#### Modelo Lineal Generalizado de tipo logit.

La Tabla 10 presenta los resultados obtenidos al considerar un modelo logit para estimar la variable dicotómica “Entire Order Process” en función del resto de variables consideradas.

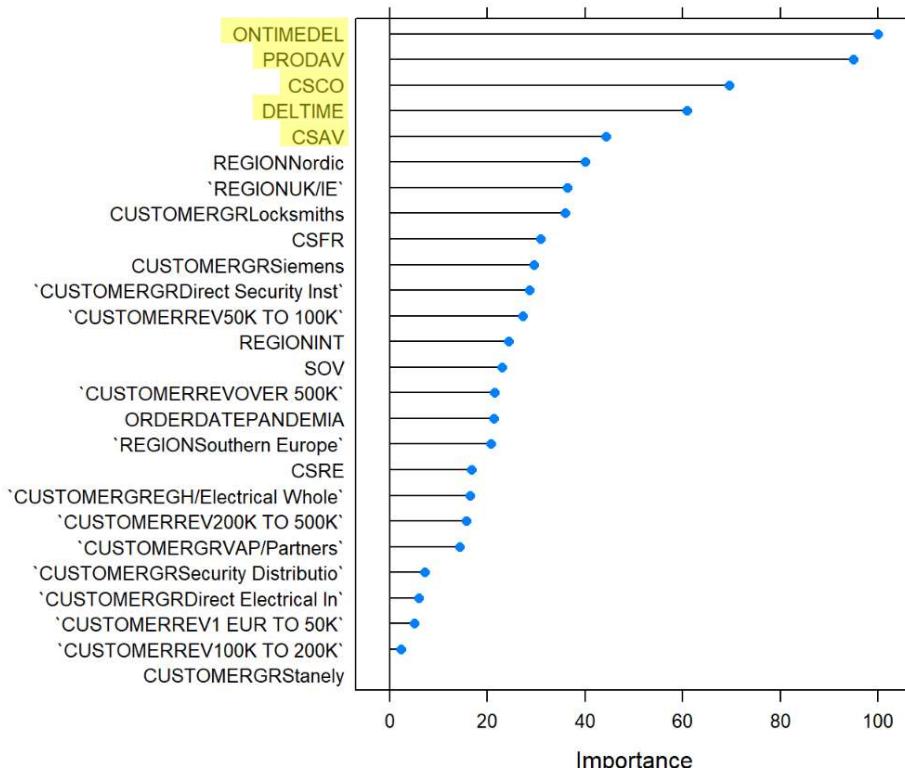
Tabla 10. Estimación del modelo y resultados de un MLG de tipo logit.

```
Call:  
glm(formula = ENTIRE ~ ., family = binomial("logit"), data = cx.balanceado[,  
  c(-2, -5, -6)])  
  
Deviance Residuals:  
    Min      1Q  Median      3Q     Max  
-2.7460  -0.2222   0.0662   0.3418   3.7197  
  
Coefficients:  
              Estimate Std. Error z value Pr(>|z|)  
(Intercept) -2.725e+01  3.508e+00 -7.770 7.86e-15 ***  
SOV          -1.298e-05  4.021e-05 -0.323 0.746918  
REGIONINT     4.624e-01  6.809e-01  0.679 0.497046  
REGIONNordic  -8.997e-01  6.238e-01 -1.442 0.149233  
REGIONSouthern Europe -3.483e-01  4.905e-01 -0.710 0.477576  
REGIONUK/IE    -3.568e-01  7.726e-01 -0.462 0.644170  
ORDERDATEPANDEMIA 3.433e-01  3.456e-01  0.993 0.320602  
CUSTOMERGRDirect Electrical In -4.956e-02  2.725e+00 -0.018 0.985487  
CUSTOMERGRDirect Security Inst -7.745e-01  2.464e+00 -0.314 0.753250  
CUSTOMERGRGREGH/Electrical Whole -7.215e-01  2.834e+00 -0.255 0.799013  
CUSTOMERGRLocksmiths -1.745e+00  2.675e+00 -0.652 0.514232  
CUSTOMERGRSecurity Distributio -2.890e-01  2.499e+00 -0.116 0.907921  
CUSTOMERGRSiemens -8.014e-01  2.476e+00 -0.324 0.746199  
CUSTOMERGRStanely 7.976e-02  4.263e+00  0.019 0.985072  
CUSTOMERGRVAP/Partners -8.695e-02  2.550e+00 -0.034 0.972797  
CUSTOMERREV1 EUR TO 50K -4.466e-01  7.430e-01 -0.601 0.547787  
CUSTOMERREV50K TO 100K -9.715e-01  9.479e-01 -1.025 0.305450  
CUSTOMERREV100K TO 200K -7.233e-01  9.850e-01 -0.734 0.462752  
CUSTOMERREV200K TO 500K -1.620e-01  9.247e-01 -0.175 0.860902  
CUSTOMERREVOVER 500K -5.074e-01  9.540e-01 -0.532 0.594782  
CSAV          9.369e-01  4.274e-01  2.192 0.028388 *  
CSFR          -7.694e-01  6.919e-01 -1.112 0.266097  
CSCO          2.094e+00  6.354e-01  3.296 0.000981 ***  
CSRE          9.485e-01  5.804e-01  1.634 0.102229  
PRODAV        7.152e-01  2.072e-01  3.452 0.000557 ***  
DELTIME       8.040e-01  3.022e-01  2.661 0.007799 **  
ONTIMEDEL    1.529e+00  3.558e-01  4.298 1.72e-05 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 823.46 on 593 degrees of freedom  
Residual deviance: 280.40 on 567 degrees of freedom  
AIC: 334.4
```

Number of Fisher Scoring iterations: 7

El resultado de la estimación del modelo lineal generalizado de tipo logit muestra que solamente son significativas las siguientes variables: Customer Service Availability, Customer Service Competence, Product Availability, Delivery Time y On-time Delivery. Se observa que todas estas variables tienen signo positivo y afectan positivamente la probabilidad de que la variable dependiente (Entire Order Process) tome el valor “Excelente” (5 puntos sobre 5).

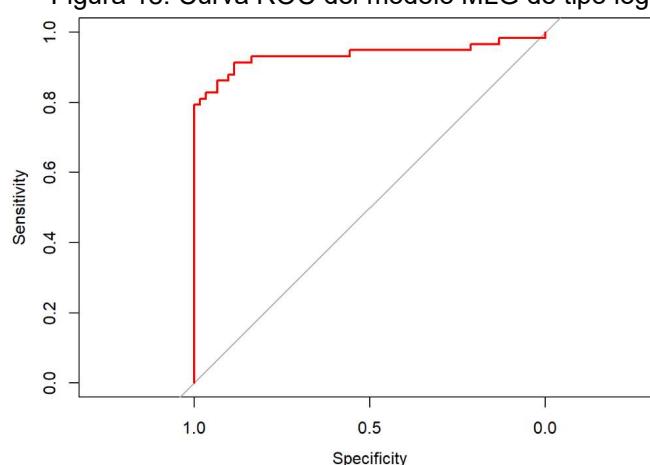
Figura 17. Listado de variables más importantes según el MLG de tipo logit.



En la Figura 17 se muestra en orden de importancia las variables que más afectan al valor que toma la variable dependiente “Entire Order Process”. Se resaltan en amarillo las variables que han resultado significativas. Se observa que las variables “On-time delivery” y “Product Availability” son en magnitud las dos más relevantes, seguidas de “Customer Service Competence”, “Delivery time” y “Customer Service Availability”.

El valor del área bajo la curva ROC del modelo lineal generalizado tipo logit es de 0,931. Se muestra la curva en la Figura 18:

Figura 18. Curva ROC del modelo MLG de tipo logit.



### Modelo Lineal Generalizado de tipo probit.

La Tabla 11 presenta los resultados obtenidos al considerar un modelo probit para estimar la variable dicotómica “Entire Order Process” en función del resto de variables considerada.

Tabla 11. Estimación del modelo y resultados de un MLG de tipo probit.

```

Call:
glm(formula = ENTIRE ~ ., family = binomial("probit"), data = CX.balanceado[, 
c(-2, -5, -6)])

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-2.6268 -0.2411  0.0580  0.3447  3.7428

Coefficients:
                                         Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -1.475e+01  1.604e+00 -9.195 < 2e-16 ***
SOV                                 -1.769e-06  2.279e-05 -0.078 0.938119
REGIONINT                           3.031e-01  3.451e-01  0.878 0.379842
REGIONNordic                        -4.530e-01  3.253e-01 -1.393 0.163729
REGIONSouthern Europe                -2.348e-02  2.582e-01 -0.091 0.927520
REGIONUK/IE                          -3.052e-02  4.085e-01 -0.075 0.940435
ORDERDATEPANDEMIA                   1.875e-01  1.786e-01  1.050 0.293731
CUSTOMERGRDirect Electrical In     5.241e-02  1.222e+00  0.043 0.965775
CUSTOMERGRDirect Security Inst    -2.289e-01  1.075e+00 -0.213 0.831403
CUSTOMERGREGH/Electrical whole     -3.996e-01  1.259e+00 -0.317 0.750977
CUSTOMERGRLocksmiths               -6.916e-01  1.214e+00 -0.570 0.568799
CUSTOMERGRSecurity Distributio   -1.118e-01  1.096e+00 -0.102 0.918745
CUSTOMERGRSiemens                  -4.184e-01  1.081e+00 -0.387 0.698748
CUSTOMERGRStanely                  3.131e-03  1.879e+00  0.002 0.998671
CUSTOMERGRVAP/Partners            1.744e-02  1.130e+00  0.015 0.987689
CUSTOMERREV1 EUR TO 50K           -1.641e-01  3.777e-01 -0.435 0.663909
CUSTOMERREV50K TO 100K             -3.664e-01  4.849e-01 -0.756 0.449864
CUSTOMERREV100K TO 200K            -1.124e-01  5.004e-01 -0.225 0.822232
CUSTOMERREV200K TO 500K            1.192e-01  4.688e-01  0.254 0.799297
CUSTOMERREVOVER 500K              -2.018e-01  4.895e-01 -0.412 0.680169
CSAV                               4.850e-01  2.285e-01  2.123 0.033787 *
CSFR                               -3.525e-01  3.509e-01 -1.004 0.315167
CSCO                               1.168e+00  3.277e-01  3.566 0.000362 ***
CSRE                               4.070e-01  3.088e-01  1.318 0.187561
PRODAV                            4.358e-01  1.148e-01  3.797 0.000146 ***
DELTIME                            4.049e-01  1.665e-01  2.432 0.015019 *
ONTIMEDEL                         7.955e-01  1.968e-01  4.041 5.31e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 823.46 on 593 degrees of freedom
Residual deviance: 285.49 on 567 degrees of freedom
AIC: 339.49

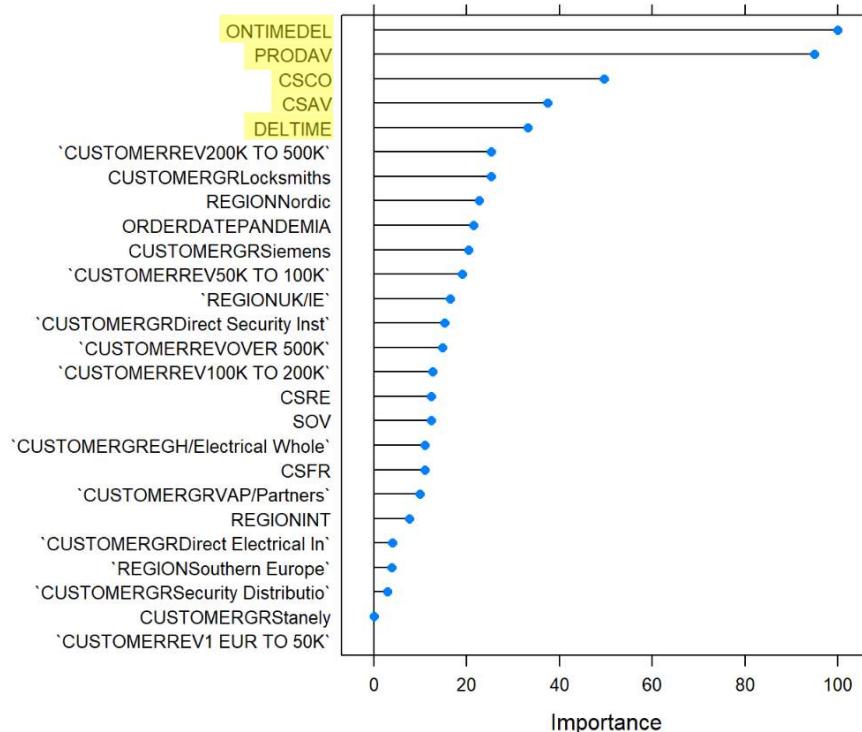
Number of Fisher scoring iterations: 9

```

El resultado de la estimación del modelo lineal generalizado de tipo probit muestra que solamente son significativas las siguientes variables: Customer Service Availability, Customer Service Competence, Product Availability, Delivery Time y On-time Delivery.

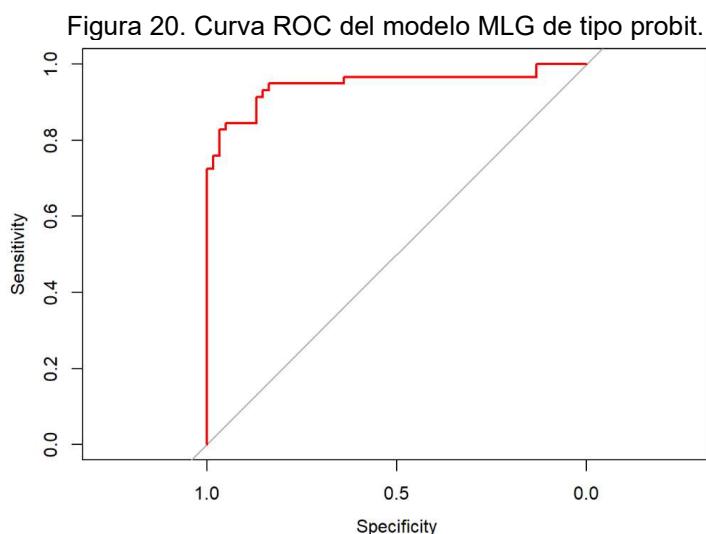
Se observa que todas estas variables tienen signo positivo y afectan positivamente la probabilidad de que la variable dependiente (Entire Order Process) tome el valor “Excelente” (5 puntos sobre 5). A continuación, se muestra el listado de la importancia de las variables según el modelo lineal generalizado de tipo probit:

Figura 19. Listado de variables más importantes según el MLG de tipo probit.



En la Figura 19 se muestra en orden de importancia las variables que más afectan al valor que toma la variable dependiente “Entire Order Process”. Se resaltan en amarillo las variables que han resultado significativas. Se observa que las variables “On-time delivery” y “Product Availability” son en magnitud las dos más relevantes, seguidas a cierta distancia de “Customer Service Competence”, Customer Service Availability” y “Delivery time”.

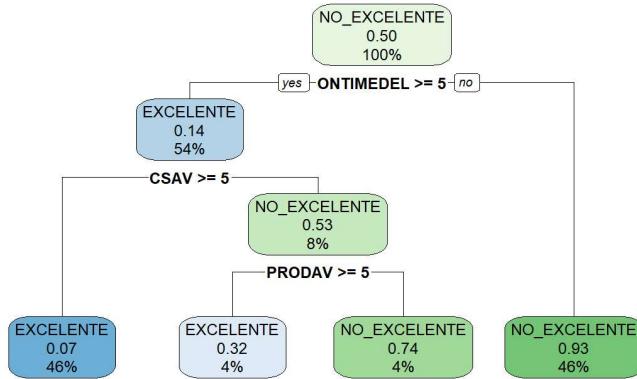
El valor del área bajo la curva ROC del modelo lineal generalizado tipo probit es de 0,952. Se muestra la curva en la Figura 20:



#### 4.2.2 Modelos de árbol de decisión.

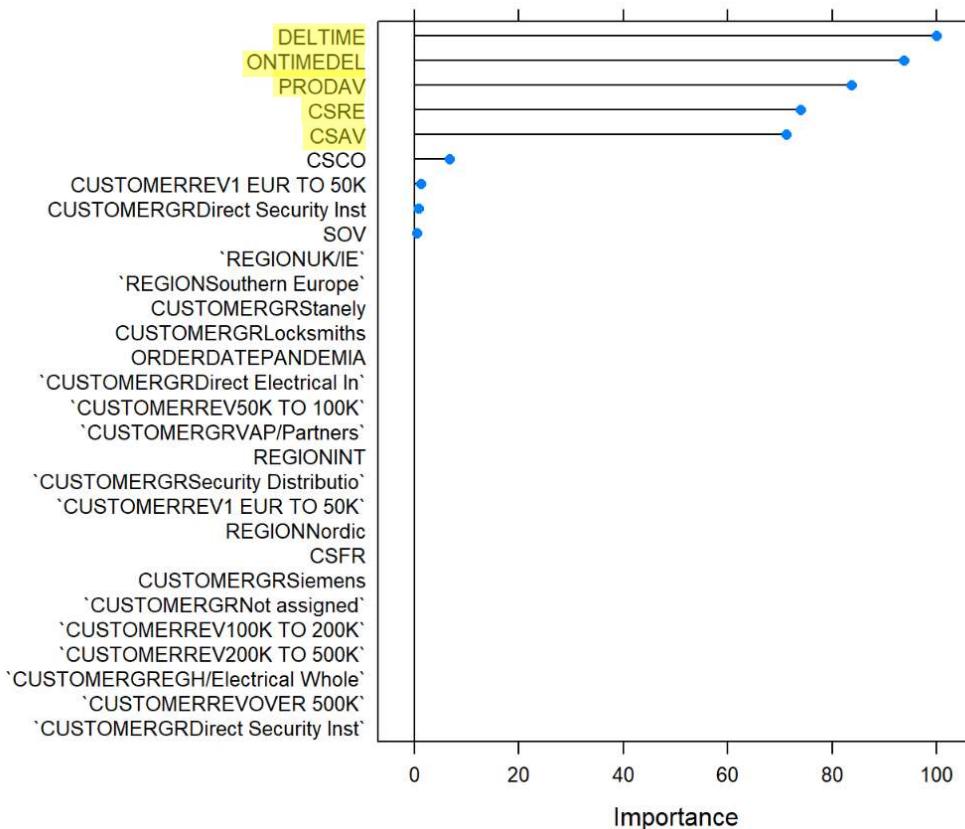
##### Árbol CART (Classification and Regression Trees).

Figura 21. Resultado del modelo de árbol empleno un algoritmo CART.

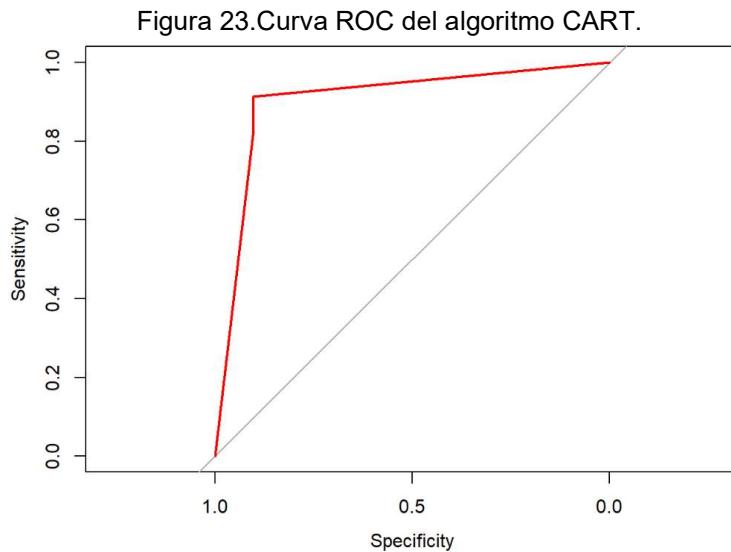


En la Figura 21 se muestra el árbol CART, en el que la variable más importante para determinar una valoración de 5 puntos en la encuesta de satisfacción del cliente es “On-time Delivery”, siendo éste el nodo principal y teniendo la escisión binaria en las dos ramas más importantes del árbol dependiendo de si se cumple o no la condición “On-Time Delivery  $\geq 5$ ”. En el caso que nos ocupa el modelo CART ha considerado que tres variables son suficientes para explicar la distribución de probabilidades de la variable dependiente según los valores que toman tres de las variables explicativas, estando las demás variables del conjunto de datos representadas en las hojas finales del árbol.

Figura 22. Listado de variables más importantes según el algoritmo CART.



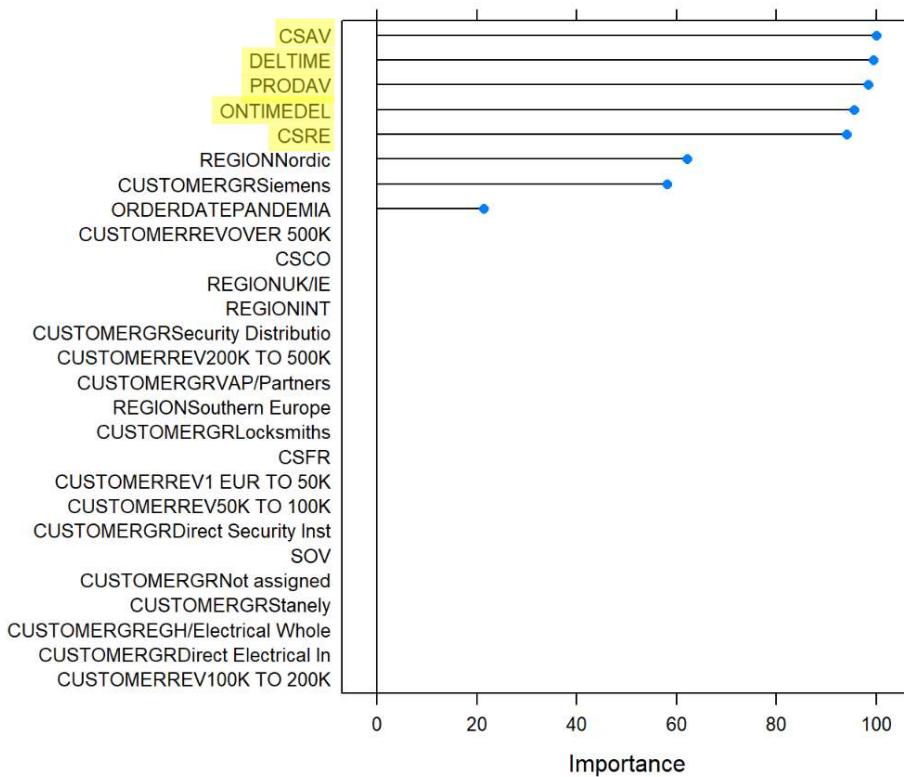
El valor del área bajo la curva ROC del modelo de árbol CART es de 0,901. La Figura 23 muestra la curva ROC:



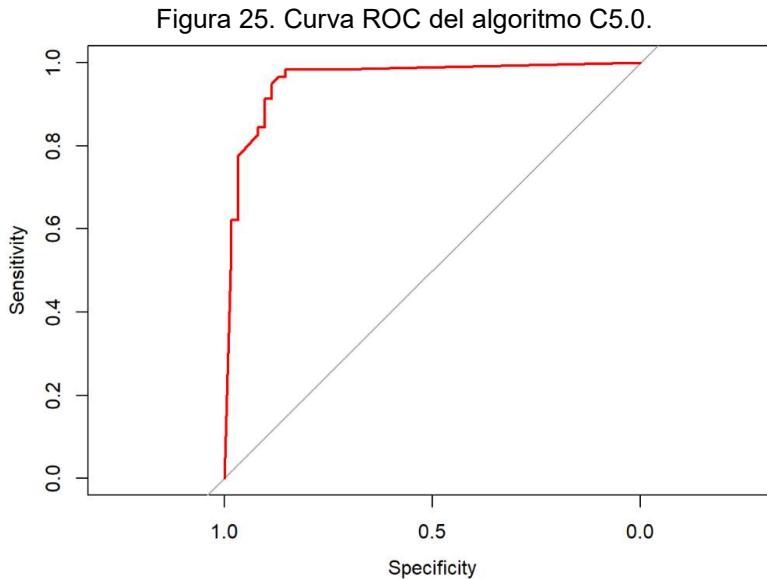
### Árbol C5.0

El modelo de árbol C5.0 considera que las variables más importantes son “Customer Service Availability”, seguida de “Delivery time”, “Product Availability”, “On-time delivery” y “Customer Service Reliability”.

Figura 24. Listado de variables más importantes según el algoritmo C5.0.



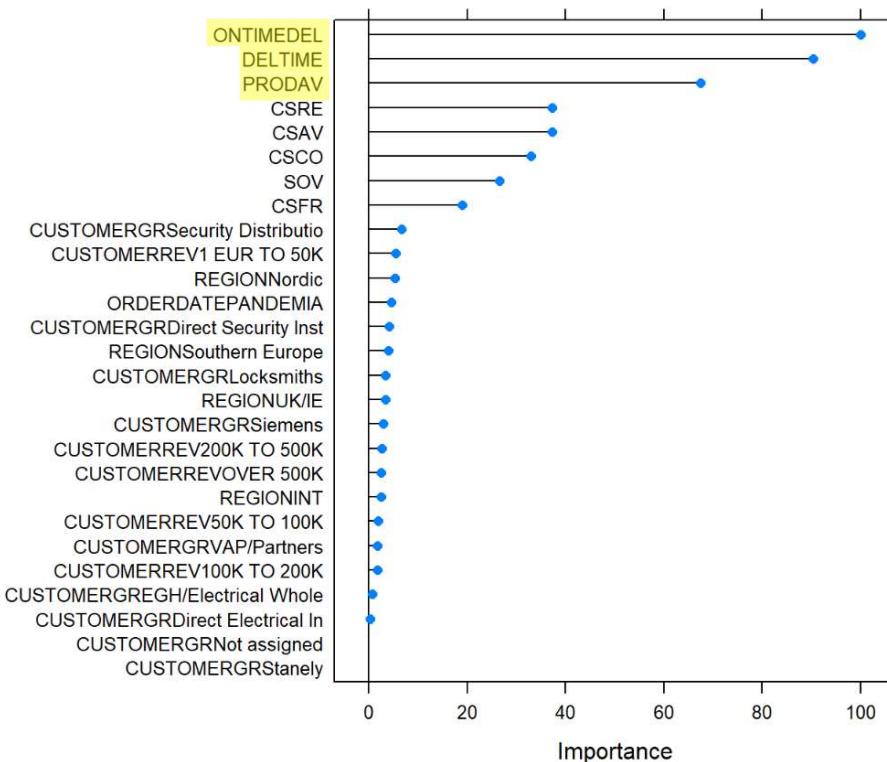
El valor del área bajo la curva ROC del modelo de árbol C5.0 es de 0,959. Se muestra la curva en la Figura 25:



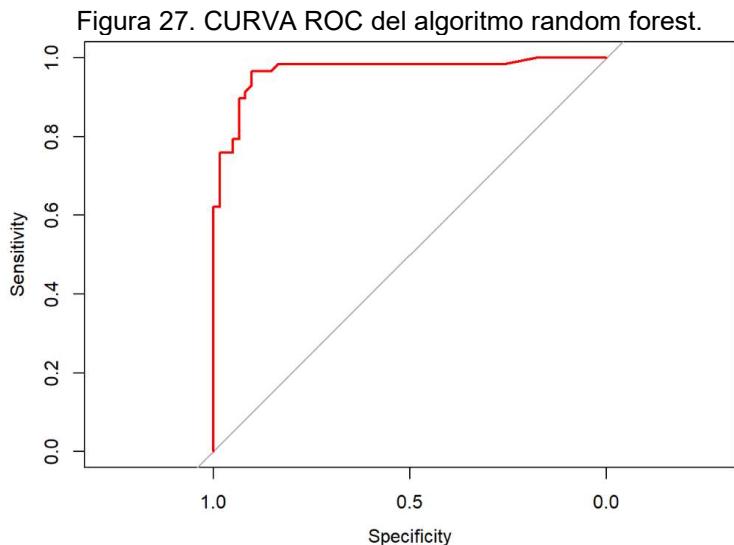
### Árbol random forest.

El modelo de árbol random forest considera que las variables más importantes son “On-time Delivery”, “Delivery time” y “Product Availability”.

Figura 26. Listado de variables más importantes según el algoritmo random forest.



El valor del área bajo la curva ROC del modelo de árbol random forest es de 0,967. Se muestra la curva en la Figura 27:



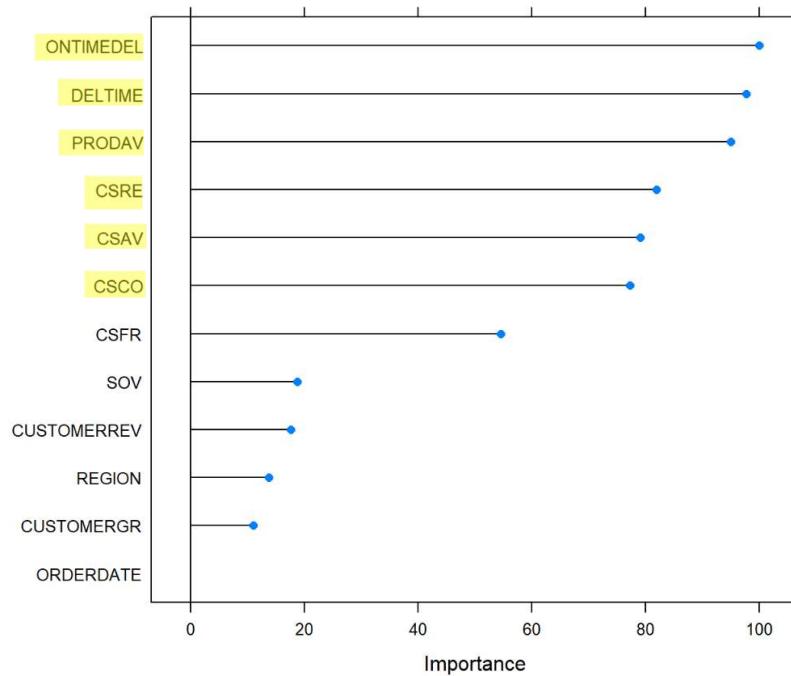
#### 4.2.3 Modelos redes neuronales artificiales.

A continuación, se presentan los resultados para los distintos modelos de redes neuronales implementados. Se ha utilizado el método de pesos en la capa oculta para determinar la importancia de las variables (Gevrey et. al., 2003).

##### Red neuronal perceptrón multicapa.

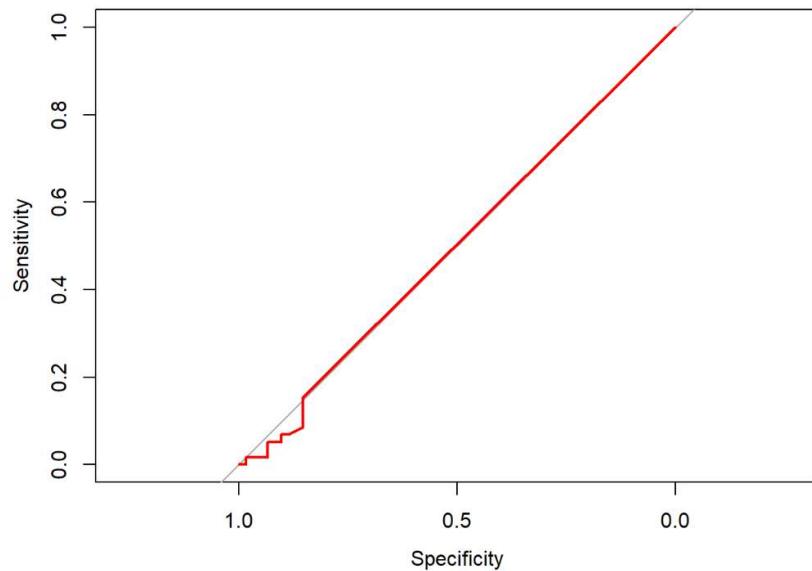
Como se muestra en la Figura 28, el modelo de red neuronal perceptrón multicapa considera que las variables más importantes son “On-time Delivery”, “Delivery time”, “Product Availability”, Customer Service Reliability”, “Customer Service Availability” y “Customer Service Competence”.

Figura 28. Listado de variables más importantes según la RNA de perceptrón multicapa.



El valor del área bajo la curva ROC de la red neuronal de tipo perceptrón multicapa es de 0,501. La curva ROC obtenida en este caso se muestra en la Figura 29:

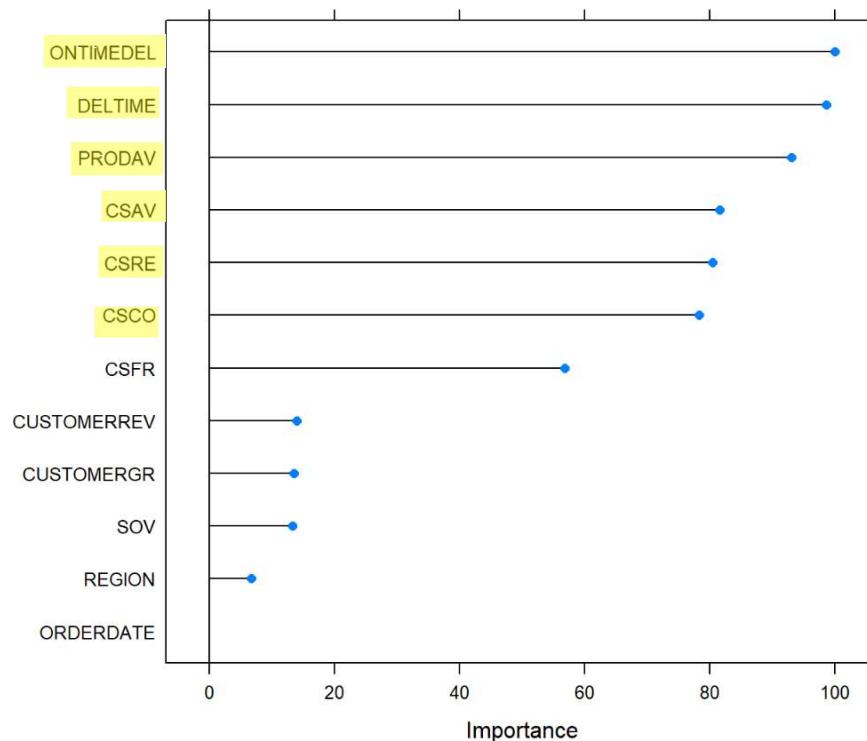
Figura 29. Curva ROC del modelo RNA perceptrón multicapa.



#### Red neuronal de función de base radial.

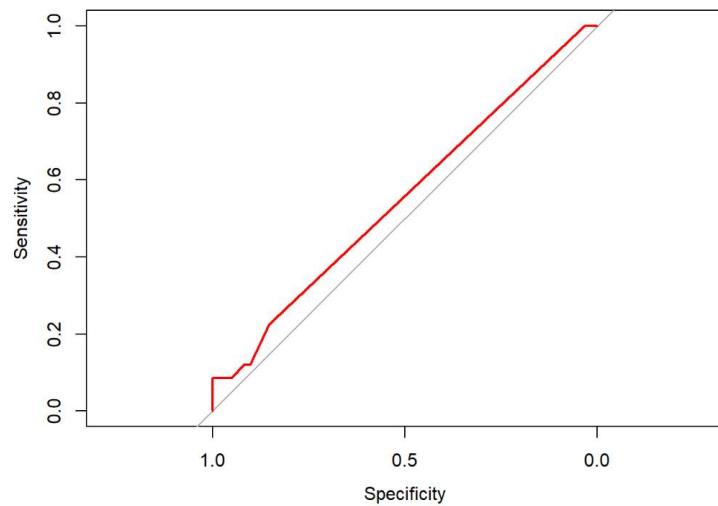
El modelo de red neuronal de función de base radial considera que las variables más importantes son “On-time Delivery”, “Delivery time”, “Product Availability”, “Customer Service Availability”, “Customer Service Reliability” y “Customer Service Competence”.

Figura 30. Listado de variables más importantes según la RNA de base radial.



El valor del área bajo la curva ROC de la red neuronal de tipo función de base radial es de 0,548. Se muestra la curva en la Figura 31:

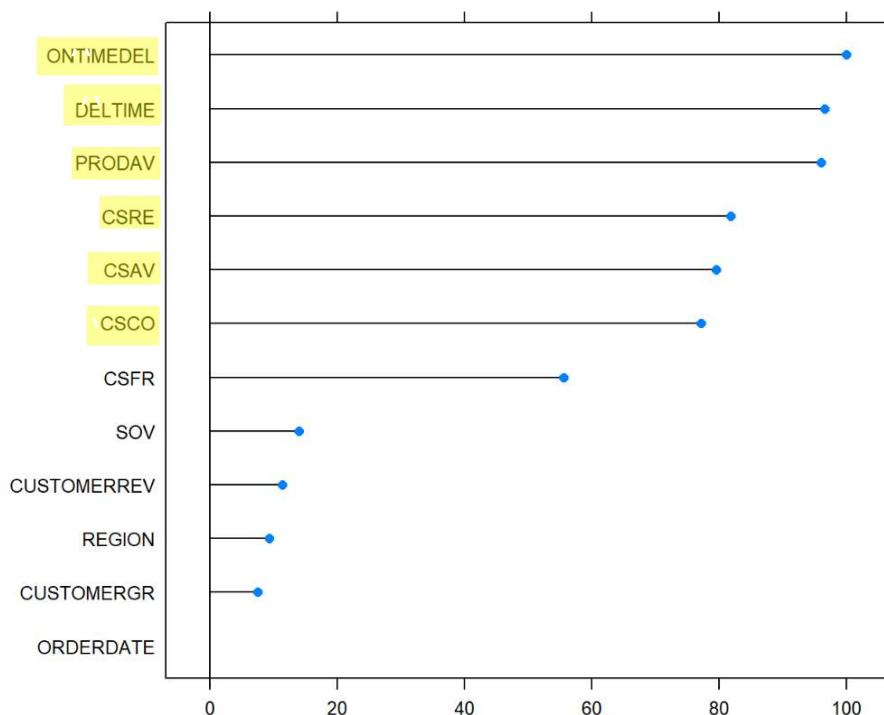
Figura 31. Curva ROC del modelo RNA de función de base radial.



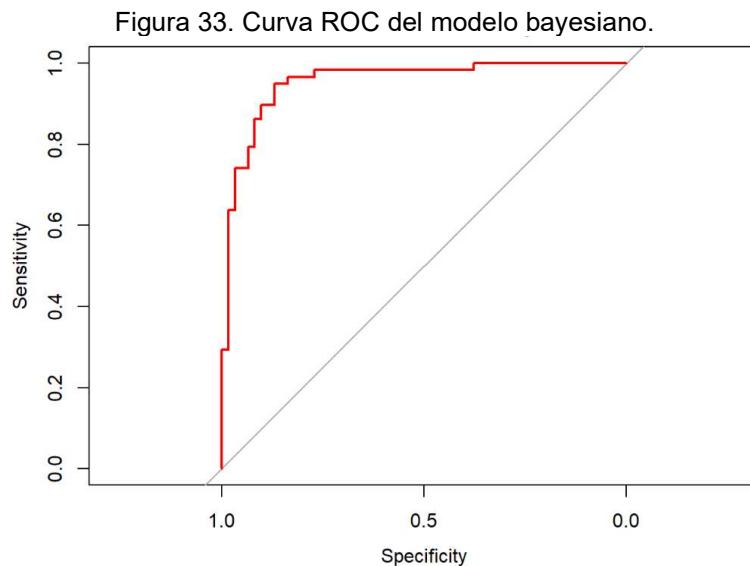
#### 4.2.4 Modelo Naïve Bayes.

El modelo de Bayes considera que las variables más importantes son “On-time Delivery”, “Delivery time”, “Product Availability”, Customer Service Reliability”, “Customer Service Availability” y “Customer Service Competence”. En la Figura 32 se presentan la importancia de cada una de estas variables para explicar la variable dependiente considerada.

Figura 32. Listado de variables más importantes según el modelo bayesiano.



El valor del área bajo la curva ROC del modelo bayesiano es de 0,955. Se muestra la curva en la Figura 33:



### 4.3 Evaluación y comparación entre modelos.

Los modelos estimados son una selección variada de diferentes técnicas de predicción: desde los modelos estadísticos clásicos, pasando por modelos de árbol de decisión, un modelo de multiclasicadores, un modelo de estadística bayesiana y dos modelos de redes neuronales artificiales. En la modelización, por tanto, se han empleado técnicas muy distintas que responden al mismo fin y que calculan sus estimaciones por diferentes caminos.

Por una parte, se han explotado dos modelos lineales generalizados no lineales: estos son los modelos logit y probit, que utilizan la función de distribución logística y la función de probabilidad respectivamente para calcular la probabilidad de que el servicio al cliente sea evaluado como excelente contra la probabilidad de que no lo sea. Estos dos modelos han obtenido un resultado bastante bueno, con un AUC de 0.931 en el caso del logit y un 0.952 para el caso del probit. Por tanto, son válidos para el empleo de odds ratios con el fin de cuantificar el impacto en términos de probabilidad de cada una de las variables significativas sobre la variable dependiente. En cualquier caso, ninguno de ellos resultó ser el modelo con la mayor puntuación de AUC por lo que no se profundiza en este aspecto.

En segundo lugar, se tienen los modelos de árbol tipo CART, C5.0 y random forest. El algoritmo CART ha obtenido un AUC de 0.901 lo cual es un buen resultado y, además, proporciona una interpretación visual sencilla y estructurada de las variables más importantes a la hora de clasificar en términos de probabilidad los valores que puede tomar la variable dependiente. Por otra parte, se tiene el algoritmo C5.0 que con un AUC de 0.959 ha resultado el segundo mejor modelo tras el random forest que ha obtenido el AUC más elevado de todos los modelos, con un valor de 0.967 y que pone en evidencia la superioridad de un algoritmo multiclasicador sobre los algoritmos clasificadores normales.

En tercer lugar, se tienen los modelos de redes neuronales. No han obtenido un resultado AUC elevado, más bien todo lo contrario. Un AUC inferior al 0.6 es en cualquier caso un resultado pobre, que no permite clasificar correctamente la variable dependiente con los datos del conjunto muestral.

En cuarto lugar, se tiene el modelo de estadística bayesiana, que ha obtenido un AUC de 0.956 que es un resultado muy bueno, y, por tanto, la estimación de la variable dependiente se realiza con mucha precisión. Este resultado pone de relieve la validez de los modelos bayesianos en la actualidad.

En la Tabla 12 se muestra un resumen con los resultados por modelo.

Tabla 12. Sumario de resultados de los valores AUC obtenidos por modelo.

MODELO	AUC
GLM logit	0.931
GLM probit	0.952
Árbol CART	0.901
Random Forest	0.967
C5	0.959
Perceptrón Multicapa	0.501
Función de base radial	0.548
Bayesiano	0.956

Este resultado no es sorprendente puesto que, entre todos los modelos evaluados, el random forest es el único multiclásificador, es decir, un algoritmo que genera múltiples clasificadores (árboles CART) que se encuentran no correlacionados para después ser promediados. De esta manera, se obtienen resultados más robustos y eficientes que cuando se emplea un único clasificador. Retomando los resultados obtenidos por el random forest en la Figura 26, se observan en color amarillo aquellas variables que tienen una importancia mayor al 50% a la hora de determinar la excelencia en el servicio al cliente (5 puntos sobre 5 en la variable dependiente “Entire Order Process”). Estas variables son, en orden descendente de importancia, “On-time Delivery”, “Delivery time” y “Product Availability” y son las variables sobre las que se deberán tomar medidas en la fase de despliegue del proceso CRISP-DM.

## **5. DESPLIEGUE E IMPLEMENTACIÓN**

---

### **5.1 Despliegue.**

La fase de despliegue se divide en varios puntos. En primer lugar, se analizan brevemente las variables más importantes estimadas por nuestro algoritmo ganador. En este estudio, ha sido el modelo random forest, que ha seleccionado las variables “On-time delivery”, “Delivery time” y “Product availability” como las más determinantes a la hora de clasificar la valoración de la gestión de pedidos por parte de los clientes. Una vez presentadas, se estudiarán cuáles son los posibles orígenes de las ineficiencias en dichas variables. Conocido el origen de las ineficiencias se podrán presentar y desarrollar algunas soluciones. Por último, se propondrán una serie de acciones para implementar las soluciones en el proceso productivo e incorporarlas en la empresa.

Siguiendo el esquema presentado en el párrafo anterior, se procede, en primer lugar, a analizar cada una de las variables más importantes detectadas por nuestro algoritmo de modelización de una gestión de pedidos excelente.

On-time delivery. La variable “On-time delivery” se refiere a la entrega de los productos en la fecha confirmada a los clientes en el momento de la compra. La variable hace referencia a la precisión en la fecha proporcionada y que en muchas ocasiones resulta crítica puesto que las empresas realizan su planificación en base a los tiempos previstos. Cualquier cambio supone un estrés importante sobre las necesidades no cubiertas o los compromisos adquiridos por el cliente, de modo que puede ser un motivo importante para mantener o perder un cliente.

“Delivery time”. La variable “Delivery time” hace referencia al plazo de entrega confirmado al cliente, es decir, el tiempo que se prevé que tardará en suministrarse el material.

“Product Availability”. La variable “Product Availability” se refiere a la disponibilidad de un producto. Es decir, si el producto se puede suministrar o no en el momento de la compra. Cuando un producto no se halla en stock o se encuentra retenido por revisión técnica la variable “Delivery time” se dilata en el tiempo, y el cliente observa plazos de entrega más largos de los habituales.

En segundo lugar, se procede a buscar el origen de errores en la precisión del “On-time Delivery” así como la causa de largos períodos de “Delivery time”. La variable “Product Availability” afecta directamente a “Delivery time”. Cuando se falla en la fecha prometida al cliente se deben buscar los motivos que provocan esta ineficiencia en la gestión. Podemos diferenciar tres grandes grupos en los cuales puede haber un problema:

1. Incidencias en el inventario. Se producen cuando no hay stock para el envío del material.
2. Incidencias en la preparación del envío. Se producen cuando el tiempo de preparación de los materiales en sus cajas con sus etiquetas lleva más tiempo del esperado y el transporte no comienza en el momento previsto.
3. Incidencias en la entrega. Se producen cuando el transporte, aun saliendo a reparto en la fecha prevista, no entrega en la fecha prometida al cliente.

En la empresa objeto de estudio no hay ineficiencias destacables en cuanto a incidencias de entrega ni tampoco en la preparación del envío. Estas tareas las realiza un operador logístico subcontratado que realiza estos trabajos y, salvo en contadas ocasiones, los paquetes se preparan en tiempo y forma, y el envío de los paquetes no suele tener incidencias más allá de situaciones extraordinarias (tales como huelgas, incidencias meteorológicas etc...)

Sin embargo, sí que existen incidencias en el inventario. Estas incidencias pueden deberse a dos motivos:

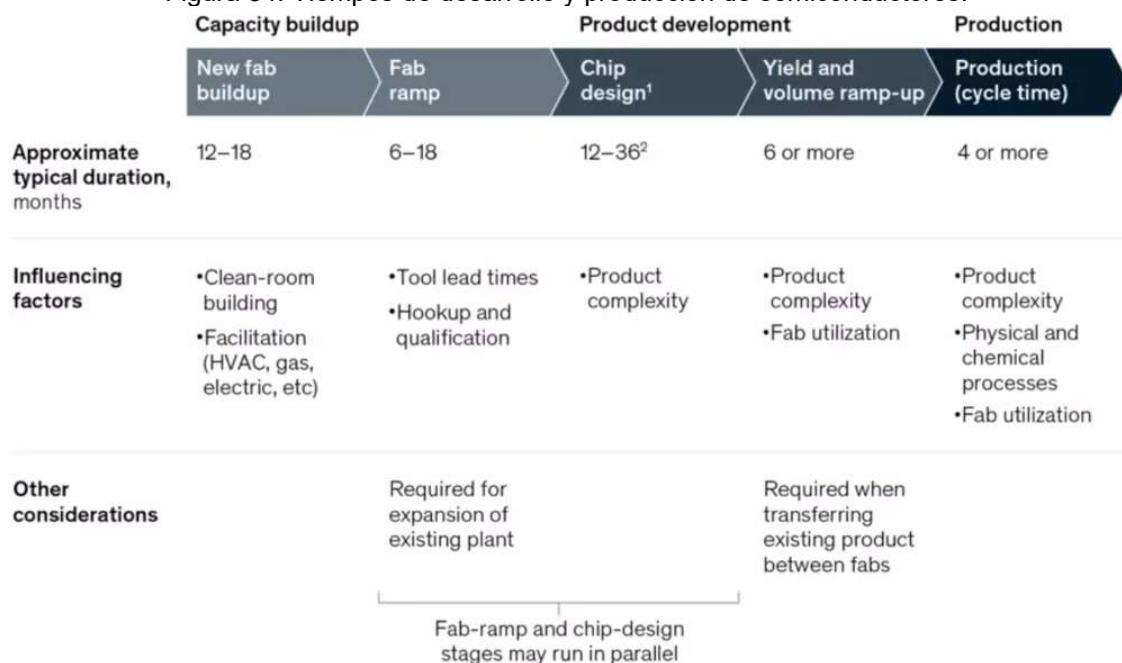
1. Se falla en el “On-time Delivery” porque se envió una confirmación de entrega que no se puede cumplir porque el material no está disponible para ser enviado cuando se tenía previsto.
2. Se proporciona un “Delivery time” muy largo al cliente porque no hay stock disponible y la previsión de envío es lejana en el tiempo.

Ambas incidencias desembocan en el mismo resultado, una dilatación en el tiempo de la fecha de entrega del material, pero hay una diferencia importante, y es que un “Delivery time” largo puede ser aceptado con resignación por el cliente, que mantiene la capacidad de planificación y previsión de cuándo va a recibir su paquete, y por lo tanto puede organizarse en consecuencia. Pero cuando se incumple la fecha de entrega confirmada al cliente en el “On-time delivery” el problema es mayor, puesto que la situación es inesperada para el cliente, que esperaba recibir la mercancía en una fecha concreta que no se cumple. Esto no es un problema (o no suele serlo) si el cliente recibe el material antes de lo previsto, pero cuando no se cumple y la fecha se retrasa varios días o semanas, entonces el problema es grave, puesto que el cliente falla en sus compromisos adquiridos y toda su planificación se incumple, pudiendo generar un efecto dominó sobre otros procesos productivos e incluso afectando a otros clientes.

En la empresa objeto de estudio, el inventario se gestiona mediante un ERP que tiene automatizado el flujo de los productos con compras de reposición programadas a futuro en el tiempo en base al consumo medio mensual. De forma que se mantienen unos niveles de stock bajos debido a la alta rotación de los productos. La rotación de los productos en el almacén sigue un esquema FIFO (first input first output). Lo cierto, es que la cadena de suministro ha funcionado relativamente bien durante el tiempo previo a la pandemia, y salvo contadas disrupciones por el lado de la oferta o la demanda de productos, el stock ha tenido siempre unidades disponibles para garantizar el flujo de mercancías hacia los clientes. Ha sido la pandemia junto a la crisis de oferta de los chips semiconductores y el colapso del transporte internacional, los factores que han provocado una desestabilización de la cadena de suministro, afectando a los tiempos de entrega y a la precisión en las fechas de entrega confirmadas a los clientes.

Por tanto, en relación con la incidencia de inventario, se buscan soluciones a los problemas derivados de la escasez mundial de semiconductores y el colapso del transporte internacional. Para la mejor comprensión del alcance de la crisis de oferta de los chips semiconductores se muestra una aproximación del mercado de semiconductores para evaluar la situación actual. En la Figura 34 se muestran los tiempos de producción actuales de chips semiconductores:

Figura 34. Tiempos de desarrollo y producción de semiconductores.



<sup>1</sup>Chip design can be driven independent of fab manufacturing capacity.

<sup>2</sup>Eg, ~12-month product lifecycle for mobile phones; 24–36-month development time for automotive microcontroller units.

Source: McKinsey analysis

Fuente: [www.weforum.org](http://www.weforum.org)

La Figura 34 muestra cuales son los meses de duración aproximada de producción de semiconductores en función de la madurez de la tecnología y de las instalaciones existentes.

Son varias las posibilidades en términos de tiempo de producción:

- Production (cycle time): Si un fabricante no cambia su lugar de fabricación y ya se encuentra empleando al 100% sus recursos productivos, entonces los tiempos de espera son de 4 meses para semiconductores con una línea de fabricación bien establecida.
- Yield and volume ramp up: Si un fabricante decide incrementar su producción y mover su lugar de producción a otro que permita producir más capacidad entonces el periodo de producción añade otros 6 meses adicionales. En este caso, lo más eficiente es “copiar exactamente el lugar de producción” Varian (2010), puesto que la más mínima diferencia del diseño de las fábricas, como los procedimientos de limpieza o la longitud de las mangueras de refrigeración, podrían influir significativamente en el proceso productivo. Solamente se debe incorporar un cambio sobre el modelo base si reporta enormes beneficios.
- Product development: Si se busca cambiar de fabricante de chips semiconductores, y éste no dispone de la tecnología específica para la producción de un chip específico, entonces, debe crear los wafers (plantillas circulares y delgadas que se emplean como recipientes para la fabricación de chips semiconductores) y diseñar el chip a producir (estas etapas pueden hacerse de forma simultánea), por lo que el periodo de fabricación añade al menos un año antes de empezar a producir.

Por tanto, es de esperar que la oferta de semiconductores vaya incrementándose gradualmente hasta lograr el punto de equilibrio con la demanda, al mismo que tiempo que se vaya descongestionando el apartado logístico. El ajuste del mercado se está

realizando vía precios y vía cantidad: el alza de los precios está frenando le exceso de demanda a la par que los fabricantes están incrementando gradualmente sus niveles de producción.

Una vez que la situación se estabilice se puede esperar una situación parecida a la anterior a la crisis, pero probablemente no vuelva a ser igual. Por ello, y con una visión a corto y medio plazo, se sugiere invertir en la flexibilidad en la cadena de suministro como opción de estabilidad ante futuras fluctuaciones en la demanda o en la oferta.

La flexibilidad en el área de la cadena de suministro se refiere a la diversificación, o lo que es lo mismo, la posibilidad de obtener un mismo producto desde diversos proveedores, que además no deben estar relacionados, es decir, que no dependan a su vez de los mismos sub-proveedores, que además se encuentren ubicados en diferentes zonas geográficas y como requisito adicional que se encuentren lo más próximo posible al almacén central de la empresa, de modo que no exista una dependencia de los modos de transporte por mar o aire, es decir, flexibilidad en el transporte.

El objetivo de esta flexibilidad en la cadena de suministro es garantizar el reabastecimiento desde múltiples proveedores y diversas ubicaciones, que permita sobrepasar las posibles dificultades que se puedan plantear con un proveedor habitual en un momento determinado.

Por otra parte, según nuestro modelo seleccionado, hay un segundo grupo de variables que también explican un servicio al cliente excelente, aunque no han resultado tan importantes como el grupo de variables anterior (asociadas al producto y su transporte). Son las variables “Customer Service Reliability”, “Customer Service Availability” y “Customer Service Competence”. Todas estas variables, reciben unas puntuaciones elevadas en las encuestas de satisfacción al cliente. Tal y como se vió en el análisis descriptivo, todas ellas tienen una media superior a 4,3 puntos sobre 5, y la mediana y la moda toman el valor 5. Además, desde el periodo de pandemia, estas variables han mejorado ligeramente en comparación con el periodo de tiempo previo a la pandemia. Por tanto, la empresa se encuentra en este momento con un equipo de personas capacitado (Maslavi, 2022) y preparado con las herramientas necesarias para otorgar un servicio al cliente próximo a la excelencia.

En cualquier caso, y debido a la velocidad de los cambios en la actualidad, se hace necesario seguir mejorando procesos e implementando nuevas tecnologías que permitan mantener o hacer crecer el nivel de *customer experience* de los clientes.

Con el fin de seguir avanzando hacia un servicio al cliente de calidad excelente, se sugiere la implementación de un chatbox en la página web de la empresa estudiada. Los chatbox son empleados para simular conversaciones humanas, que mediante la inteligencia artificial son capaces de interpretar la información que reciben y responder de forma lógica.

Los chatbox ofrecen a los clientes la posibilidad de utilizar un nuevo canal de comunicación con la empresa. Canal que está siempre disponible, sin horarios y con respuesta inmediata. En relación con el departamento de servicio al cliente, puede facilitar el estado de un pedido, el plazo de entrega previsto o la disponibilidad de un material entre otras muchas cosas. Y por supuesto, con la implementación de otros algoritmos específicos, también puede ser una herramienta magnífica para los equipos de marketing y de ventas.

Tabla 13. Comparación de características de mayor eficiencia entre humanos y máquinas.

<b>Bots and humans excel at different tasks</b>	
Where humans win	Where chatbots win
<ul style="list-style-type: none"> <li>• Answering a variety of questions</li> <li>• Dealing with complex situations</li> <li>• Understanding human emotion</li> </ul>	<ul style="list-style-type: none"> <li>• Answering common questions quickly</li> <li>• Reducing hold times</li> <li>• Quick routing to the right place</li> </ul>

Fuente: [www.intercom.com](http://www.intercom.com)

La Tabla 13 muestra los diferentes aspectos donde los humanos y los chatbox otorgan un servicio al cliente de mayor calidad. Los chatbox no reemplazan a los equipos de personas de servicio al cliente, sino que se ocupan de las tareas simples, frecuentes y repetitivas que liberan a las personas de las tareas monótonas que pueden ser automatizadas. Además, un chatbox bien entrenado redundaría finalmente en una mejor *user experience* por parte de los clientes, ya que la experiencia es inmediata y útil, lo que es apreciado por los clientes que tienen una herramienta adicional de comunicación con la empresa. Una *user experience* positiva ayuda enormemente a lograr un gran *customer experience* que fideliza clientes; a la vez que puede captar nuevos clientes que demanden este tipo de comunicación inmediata y disponible en cualquier momento.

Para aprovechar al máximo las capacidades del chatbox, se sugiere desarrollar algoritmos de análisis de sentimientos u *opinión mining* que se encuentran, además, en las redes sociales de la empresa. Tal y como se realizó en el trabajo de Zambrano et al. (2020), que siguiendo un proceso CRISP-DM recopilaron, trajeron y analizaron el conjunto de datos no estructurados de las opiniones reflejadas por los internautas ( criterio positivo, negativo o neutro). En su estudio, los datos no estructurados se trajeron en la etapa de preprocesamiento mediante la eliminación de caracteres innecesarios. Después se llevó a cabo una etapa de selección en la que se eliminan términos poco frecuentes. En la tercera etapa hacen recuento de los términos más frecuentes. En cuarto lugar, realizaron la selección de los términos más frecuentes y les añadieron una etiqueta que reflejase el sentimiento relacionado (positivo, negativo o neutro).

El mayor reto del análisis de este tipo de datos es la implementación de algoritmos de *machine learning* que permitan su interpretación y transformación en datos estructurados explotables.

## 5.2 Implementación.

En esta fase se sugiere un proceso por pasos para implementar tanto la flexibilidad en la cadena de suministro como el chatbox en la página web de la empresa.

- Flexibilidad de la cadena de suministro mediante *multi-sourcing*.

Pasos a seguir para la implementación:

- Realización de un estudio en profundidad de todos los posibles proveedores que puedan ofrecer los chips necesarios para la producción de cada producto.
- Proceso de calificación: Comunicación con cada uno de ellos para evaluar la capacidad de producción y suministro, así como los precios orientativos.
- Evaluación y nivel de riesgo: se usan métricas para cuantificar las características que le interesan a la empresa objeto de estudio.
- Selección y segmentación: se filtran los proveedores que cumplen con las características deseadas y se ordenan en base a las métricas empleadas en la fase anterior. En esta fase, resulta de gran importancia la selección de proveedores muy diferentes entre sí: que no dependan del mismo subproveedor, que se encuentren en diferentes países y a ser posible en distintos continentes. Se valorará muy positivamente aquellos proveedores muy próximos a la ubicación geográfica del almacén de la empresa estudiada.
- Incorporación e información de gestión: se recolecta, almacena y gestiona toda la información relativa a cada proveedor de forma que facilite la relación comercial.
- Gestión del desempeño: se miden con métricas el desempeño que está teniendo el proveedor.
- Desarrollo del proveedor: a través de contacto regular con el objetivo de crear una fuerte relación. Comunicación con *feedback* constructivo.
- Mitigación de riesgos inesperados: se realiza un plan de contingencia para situaciones en las que el proveedor principal no puede enviar material de modo que se pueda contar con otro proveedor alternativo de los seleccionados.
- Gestión de la relación: cultivando la lealtad se estrechan los lazos de relación con el proveedor, de modo que en situaciones imprevistas se pueda contar con su apoyo de forma natural en lugar de forzada.

En la Figura 35 se muestran los pasos a seguir para la incorporación de nuevos proveedores a la empresa:

Figura 35. Proceso de incorporación de nuevos proveedores en una empresa.



Fuente: [www.frevvo.com](http://www.frevvo.com)

- Chatbox en la página web de la empresa.

Pasos a seguir para la implementación:

- Creación de un listado con las características que se esperan del chatbox y los enrutamientos avanzados demandados.
- Revisión de los chatbox existentes en el mercado y comparación entre los mismos mediante algunas métricas asociadas a precios y características.
- Elección del chatbox a integrar en la página web de la empresa.
- Fase de entrenamiento del chatbox para implementar las tareas mínimas y necesarias que se esperan del mismo como, por ejemplo:
  1. Estado de un pedido.
  2. Estado de una entrega.
  3. Acceso al precio de un producto.
  4. Acceso a la disponibilidad de un producto.
  5. Explicación de procedimientos tales como: reparaciones o devoluciones de material.
  6. Prestación de URLs importantes tales como: distribuidores o instaladores cercanos o características y prestaciones de un determinado producto.
- Fase de testeo del chatbox y corrección de errores.
- Puesta en productivo del chatbox con una métrica de evaluación de la *customer experience* de los clientes en referencia al mismo.
- Comunicación a los clientes de su nueva herramienta de comunicación con la empresa y capacidades que tiene.
- Análisis de datos de las evaluaciones de los clientes.
- Revisión de ineficiencias y mejora de las características del chatbox.

## **6. CONCLUSIONES.**

---

Los confinamientos provocaron la disminución de la oferta de semiconductores eléctricos (por la paralización temporal y/o temporal de la producción) y un incremento de la demanda de los mismos. Esta demanda de semiconductores ya estaba creciendo de forma no lineal desde los últimos años debido a la naturaleza de la Industria 4.0 en la que el *Internet de las cosas* (IoT) y la *Producción de sistemas ciberfísicos* (CPPS) producen una cantidad de datos masiva mediante el uso de las últimas tecnologías que integran y conectan sensores; sensores que son elaborados con semiconductores. Uno de los grandes retos de la Industria 4.0 pasa por adaptarse a una situación en la que existe un crecimiento exponencial en la demanda de sensores junto a una oferta limitada de los mismos.

La Industria 4.0 incorpora los datos como un activo más en los procesos de producción siendo cada vez mayor su importancia. Es por ello, que, desde este momento, el análisis de los datos adquiere un protagonismo total en la adquisición del conocimiento. Conocimiento que, en materia económica, finalmente redundará en la detección de oportunidades de negocio y/o detección de inefficiencias, siendo este punto clave para entender por qué las empresas están invirtiendo cada vez más recursos en la recolección y explotación de los datos.

La extensión del uso de herramientas de comunicación digitales, ha provocado a su vez, que las empresas presten especial atención a los canales de información de que disponen de cara a sus clientes. Son los clientes los que deciden cual es la forma de comunicarse con las empresas, por lo tanto, son las empresas las que deben tener todos los canales de comunicación posibles disponibles; pero, además, no se espera que el cliente realice la demanda de este servicio a la empresa, sino que, las empresas deben proactivamente desarrollar estos nuevos canales de comunicación basados en algoritmos de *machine learning*.

Estos canales de comunicación emergentes, basados en la inteligencia artificial, deben estar correctamente integrados, esto es, deben ser una herramienta útil que ayude al cliente y que le haga tener una experiencia de comunicación eficiente y gratificante. Eficiente en el sentido de la inmediatez y la calidad, y gratificante en el sentido de la sencillez de uso. Estas dos propiedades, determinan el *user experience* de las apps, y su correcto desarrollo determina actualmente su éxito o su fracaso, incidiendo tal experiencia directamente sobre la *customer experience*. Por dicho motivo, resulta necesario la implementación de todos los canales de comunicación posibles, que sean coherentes con la actividad económica realizada por la empresa, y que aporten un valor real a los clientes. De nada sirve, la implementación de un canal de comunicación que no brinda herramientas o facilidades de gestión a los clientes.

Adicionalmente, se espera de los nuevos canales de comunicación automatizados, la proactividad de los mismos. Mediante algoritmos avanzados con capacidad de aprendizaje de *machine learning* el futuro pasa por entrenar modelos que sean capaces de anticipar las necesidades de los clientes. De esta forma, la comunicación resulta bidireccional, de modo que cliente y empresa, comparten información. La información en forma de datos que recibe la empresa a través de los canales de información es analizada, de manera que la empresa es capaz de ofrecer soluciones a las necesidades potenciales que pueda tener un cliente en un momento determinado.

Esta proactividad de las empresas, especialmente desde el área de servicio al cliente, debe tener una aproximación personalizada según la actividad de la empresa y el perfil

del cliente. Tal y como se ha mencionado antes, el canal de comunicación debe prestar un servicio útil al cliente, lo cual, a día de hoy, significa que no se puede establecer un mecanismo general para todas las empresas, sino que, debe desarrollarse una solución para cada caso.

De forma interna, las empresas ven como se multiplican los canales de información a medida que pasa el tiempo. Por ello, es importante crear una red de canales perfectamente sincronizada que centralice toda la información obtenida desde los distintos canales. Esta sincronización entre las redes, se conoce como omnicanal, y permite controlar los flujos de información de manera unificada. Esta unificación permite controlar mejor la información con el cliente, y, no menos importante, agrupar de forma organizada todo el conjunto de datos, para su posterior análisis.

Es el análisis de la información recopilada desde los distintos canales, la que permite, la detección de oportunidades de negocio o detección de necesidades no cubiertas de los clientes, y de esta manera, y mediante el canal de comunicación adecuado, la empresa informa al cliente de manera proactiva, proyectando el conocimiento obtenido del conjunto de datos en forma de información relevante para el cliente con la consiguiente mejora de la *customer experience*.

La diversidad en la naturaleza del conjunto de datos digital genera importantes desafíos para la obtención del conocimiento de los mismos. Así pues, las empresas que se propongan analizar los conjuntos de datos deben partir de lo más sencillo, como puede ser el análisis de datos estructurados: numéricos o nominales factorizables; para posteriormente, ir agregando modelos de *machine learning* que permitan interpretar y transformar en información explotable, datos no estructurados como imágenes, vídeos o textos. Siendo especialmente útiles para la explotación de datos no estructurados las redes neuronales artificiales, por su gran flexibilidad y tolerancia a fallos.

La cantidad de datos que recopilan las empresas crecerá en los siguientes años, y su correcta explotación supondrá la diferencia entre el éxito o el fracaso empresarial. Es por ello, que es fundamental enfocar cada análisis empleando las herramientas adecuadas, así pues, tal y como se ha demostrado en este estudio, las redes neuronales no han sido un modelo adecuado para modelizar con datos estructurados la excelencia en el servicio al cliente. Siendo mucho más eficientes modelos econométricos tradicionales como lo son los modelos lineales generalizados o el modelo bayesiano. En cualquier caso, resulta adecuado tratar de extraer el conocimiento desde las máximas fuentes posibles, y elegir la fuente más confiable en base a métricas que comparan los modelos bajo un mismo paradigma, de forma que la elección se base en un criterio científico y no en una decisión subjetiva. Es importante resaltar, la necesidad de explotar el conjunto de datos empleando el conocimiento experto en la materia, para poder llegar a conclusiones correctas. La explotación de los datos bajo la perspectiva matemática/estadística llevará a soluciones puramente matemáticas que no tienen por qué coincidir con las soluciones reales que se plantean resolver.

Igualmente, importante resulta la capacidad de presentar los resultados obtenidos del análisis de datos de una forma clara y concisa, que sea fácilmente entendible por las personas que, sin ser expertas en la materia, puedan extraer el conocimiento. Por esta razón, es importante la transformación de los resultados en esquemas, histogramas o cualquier otra presentación gráfica que facilite su comprensión. Esta transformación de los datos puramente numéricos en presentaciones gráficas es tan importante como el correcto desarrollo de las diferentes técnicas para el análisis de datos, ya que, de nada sirve la obtención de unos resultados correctos si no es posible la interpretación de la información que contienen para la transmisión del conocimiento.

## **7. BIBLIOGRAFÍA**

---

- Bayes, T. (1763): "An essay towards solving a problem in the Doctrine of Chances", *Royal Society*.
- Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American statistical association*, 39(227), 357-365.
- Breiman, L., Friedman,J., Stone, C.J. & Olshen, R.A. (1984): "Classification and Regression Trees", Ed. Taylor & Francis.
- Breiman. L. (2001) "Random forests. Machine learning", Ed. Robert E. Schapire.
- Chapman, P., Clinton, J., Kerber, R., Khazaba,T., Reinartz, T., Shearer, C. y Wirth, R. (2000) "CRISP-DM 1.0 Step by step Data Mining guide. CRISP-DM, SPSS
- Cybenko. G. (1989) "Approximation by Superpositions of a Sigmoidal Function" *Mathematical control, signals and systems*, 2, 303-314.
- Exenberger, E.; Bucko, J. (2020) "Analysis of Online Consumer Behavior - Design of CRISP-DM Process Model" *AGRIS On-line Papers in Economics and Informatics*, Tomo 12, N.<sup>o</sup> 3, (13-22).
- Fechner, G. T. (1860) "Elemente der Psychophysik". Leipzig: Breitkopf und Härtel.
- Funahashi, K. (1989) "On the approximate realization of continuous mappingby neural networks", *Neural Networks*, 2,183:192.
- Gevrey, M., Dimopoulos, I., & Lek, S. (2003). "Review and comparison of methods to study the contribution of variables in artificial neural network models" *Ecological Modelling*.
- Hornik, K., Maxwell, S., White, Halbert (1989) "Multilayer feedforward networks are universal approximators", *Neural Networks*, 2:331-409.
- Huber, S., Wiemer, H., Schneider, D. & Ihlenfeldt, S. (2019): "DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model" *Procedia CIRP* Vol.79, 403-408.
- Matilla García, M., Pérez Pascual, P. y Sanz Carnero, B. (2017): "Econometría y Predicción", Ed. UNED.
- Moody, J., Darken, C.J. (1989) "Fast learning in networks of locally-tuned processing units", *Neural Computation*, 1:281-294.
- Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3), 370-384.
- Quinlan, J. (1992). "Learning with continuous classes". *Proceedings of the 5th Australian Joint Conference On Artificial Intelligence*, 343-348.
- Rumelhart, D., Hinton, G., Williams, R. (1986) "Learning representations by back-propagating errors", *Nature*, 323, 533–536.

Schröer, C., Kruse, F. & Gómez, J. M. (2021): "A Systematic Literature Review on Applying CRISP-DM Process Model" *Elsevier B.V* pp. 527-533.

Solano, J. A., Lancheros C., D.J., Umaña Ibáñez, S.F., Coronado-Hernández, J. R. (2022): "Predictive models assessment based on CRISP-DM methodology for students performance in Colombia - Saber 11 Test" *Elsevier B.V*.

Varian, H.R. (2010): "Microeconomía intermedia, un enfoque actual", Ed. Antoni Bosch. 8<sup>a</sup> Edición, pp.358.

Vicente Vírseda, J.A., González Arias, J., Parra Rodríguez, F.J. y Beltrán Pascual, M. (2019): "Métodos de data science aplicados a la economía y a la dirección y administración de empresas", Ed. UNED.

Zambrano, F.J.R., Flores, B.F.V., Mendoza, W.J.P., Zambrano, R.A.R., Rivadeneira, S. M.C. (2020) "Aplicación de minería de texto para el análisis de sentimientos del servicio de telefonía móvil en el Ecuador" *HOLOS*, Tomo 36, N.<sup>o</sup> 7, pp.1-16.

## Webgrafía

"Artificial Intelligence and Improving the Customer Experience" (2017) DISPONIBLE EN: <https://www.pega.com/system/files/resources/pdf/ai-and-improving-cx-en.pdf?rid=YToxOntzOjc6ImNvbnRfaWQiO3M6OToiQ09OVC02ODY3lit9>  
Consultado: 18/05/2022

Caret documentation (2019) "varImp: Calculation of variable importance for regression and classification model." Disponible en: <https://rdrr.io/cran/caret/man/varImp.html>  
Consultado: 21/06/2022.

Davis, J. (2021): "How Automation Improves EX and CX, Not Replaces Humans" How Automation Improves EX and CX, Not Replaces Humans (usu.com) Consultado: 20/05/2022.

"How customer service chatbots are redefining customer engagement" (2022)  
Disponible en: <https://www.intercom.com/blog/customer-service-chatbots/#:~:text=A%20customer%20service%20chatbot%20is,what%20is%20your%20pricing%3F%E2%80%9D>. Consultado: 14/06/2022

Marsden, T., Gormley, G., Hyken, S., Munch-Andersen, S., Yang, J., Gelhaar, L., Miron, U. y Condell, M. (2022): "Customer service trends guide 2022" Disponible en [https://www.ultimate.ai/guides/customer-service-trends-2022?hsa\\_ad=&hsa\\_grp=1341404977468154&hsa\\_tgt=kwd-83838702535700:loc-170&hsa\\_acc=9916259511&hsa\\_ver=3&hsa\\_src=o&hsa\\_net=adwords&hsa\\_kw=customer%20experience%20stats%202020&hsa\\_cam=8489195445&hsa\\_mt=b&msclkid=56c6997013bc1ef9eed4fbf4bfe599be&utm\\_source=bing&utm\\_medium=cpc&utm\\_campaign=Search\\_Trends-2022\\_General-TOF\\_null\\_EU\\_Ebook&utm\\_term=customer%20experience%20stats%202020&utm\\_content=Ebook%201%20-%20Customer%20Service%20Trends](https://www.ultimate.ai/guides/customer-service-trends-2022?hsa_ad=&hsa_grp=1341404977468154&hsa_tgt=kwd-83838702535700:loc-170&hsa_acc=9916259511&hsa_ver=3&hsa_src=o&hsa_net=adwords&hsa_kw=customer%20experience%20stats%202020&hsa_cam=8489195445&hsa_mt=b&msclkid=56c6997013bc1ef9eed4fbf4bfe599be&utm_source=bing&utm_medium=cpc&utm_campaign=Search_Trends-2022_General-TOF_null_EU_Ebook&utm_term=customer%20experience%20stats%202020&utm_content=Ebook%201%20-%20Customer%20Service%20Trends) Consultado: 18/05/2022.

Maslavi, S. (2022): "Eight Tips For Providing Excellent Customer Service" Disponible en: <https://www.forbes.com/sites/forbesbusinesscouncil/2022/02/02/eight-tips-for-providing-excellent-customer-service/?sh=356b1bfa259c> Consultado: 11/05/2022.

"Multi-sourcing: How to Make Your Procurement Strategy a Competitive Advantage" (2022). Disponible en: <https://www.frevvo.com/blog/multi-sourcing/> Consultado: 23/06/2022

Rall, C. (2021): "5 Trends Transforming Customer Service in 2019" Disponible en: <https://blog.usu.com/en-us/customer-service-trends-2019> Consultado: 10/04/2022.

"Use On-Time Delivery Metrics to Improve Customer Satisfaction" Disponible en: <https://optimoroute.com/on-time-delivery-metric/#what-is-it> Consultado: 05/06/2022.

"When the chips are down: How the semiconductor industry is dealing with a worldwide shortage" (2022) Disponible en: <https://www.weforum.org/agenda/2022/02/semiconductor-chip-shortage-supply-chain/> Consultado: 23/06/2022

## 8. ANEXO

---

### Código R implementado para el análisis estadístico

#### PREPROCESAMIENTO

En primer lugar, se carga la base de datos original y se convierte en el dataframe “CX”.

```
suppressWarnings(suppressPackageStartupMessages(library(readxl)))
CX <- read_excel("TABLA_DATOS_4.xlsx",
  col_types = c("numeric", "text", "text",
    "text", "text", "text",
    "text", "text", "text", "numeric", "numeric", "numeric", "numeric",
  "numeric", "numeric", "numeric", "numeric"))
CX <- as.data.frame(CX, stringsAsFactors = FALSE)
```

Se convierten en factor las variables que son de clase carácter (variables nominales).

```
for (columna in 1:17) {
  if(is.character(CX[,columna])){
    CX[,columna]<-as.factor(CX[,columna])
  }
}
```

Se ordenan los niveles del factor Customer Revenue 2021

```
CX$`Customer Revenue 2021`<-factor(CX$`Customer Revenue 2021`,levels=levels(
  X$`Customer Revenue 2021`)[c(6,1,4,2,3,5)])
```

A continuación, se realiza una exploración general de los datos del dataframe para detectar posibles datos ausentes (NA).

```
sapply(CX,function(x) sum(is.na(x)))
```

Se eliminan las 11 observaciones de la variable “Cancellation reason” con valor distinto a “Not cancelled”.

```
CX<-CX[CX$`Cancellation Reason`=="Not cancelled",]
```

Se eliminan las observaciones de la variable “Cancellation reason” con valor distinto a “Not cancelled”.

```
CX<-CX[CX$`Customer Group`!="Not assigned",]
```

Se elimina la variable “Cancellation reason” por tener un solo valor.

```
CX<-CX[,c(1:2,4:17)]
```

A continuación, se procede a detectar visualmente si hay datos extremos (outliers) mediante los gráficos de caja.

```
suppressWarnings(suppressPackageStartupMessages(library(ggplot2)))
ggplot(CX,aes(x="",y=CX[,1]))+ stat_boxplot(geom = "errorbar",width = 0.15)+
  geom_boxplot(color=1,outlier.colour = 2,)+labs(x="",y="Sales order value") +c
  oord_cartesian(ylim = c(0, 60000))
```

## ANÁLISIS DESCRIPTIVO DEL CONJUNTO DE DATOS

Información descriptiva de las variables.

```
suppressWarnings(suppressPackageStartupMessages(library(skimr)))
skim(CX)
```

A continuación, se aborda el análisis descriptivo gráfico.

Se ve la distribución de los valores de las variables numéricas de manera gráfica.

```
suppressWarnings(suppressPackageStartupMessages(library(funModeling)))
plot_num(CX)
```

También se visualizan las frecuencias de las variables categóricas.

```
freq(CX)
```

Se equilibra con una submuestra aleatoria la variable “Order Creation Date” para distinguir periodo de “PANDEMIA” y de “NO PANDEMIA”

```
set.seed(123)
subset_no_pand<-sample(rownames(CX[CX$`Order Creation Date` == "NO PANDEMIA",]),size=454,replace = F)
subset_pand<-rownames(CX[CX$`Order Creation Date` == "PANDEMIA",])
subset_final<-c(subset_no_pand,subset_pand)
```

Cruzamos variables con la submuestra equilibrada

```
cross_plot(CX_pand[,c(1:4,6:16)],target="Order Creation Date",auto_binning = T)
```

A continuación, se emplean tablas para obtener información más específica en relación a las cuentas que no tuvieron facturación en 2021

```
suppressWarnings(suppressPackageStartupMessages(library(tidyverse)))
suppressWarnings(suppressPackageStartupMessages(library(kableExtra)))
kable(CX_pand %>%
  filter(`Customer Revenue 2021`=="ZERO", `Order Creation Date`=="NO PANDEMIA")) %>%
  group_by(`Customer Group`)%>%
  summarise(Count=n(),CS_Av=mean(`Customer Service Availability`),CS_Fr=mean(`Customer Service Friendliness`),CS_Co=mean(`Customer Service Competence`),CS_Re=mean(`Customer Service Reliability`),Prod_Av=mean(`Product Availability`),Del_Time=mean(`Delivery Time`),Ontime_del=mean(`On-time Delivery`),Entire_Pr=mean(`Entire Order Process`)))
```

Correlaciones. Se genera el fichero “var\_num” con las variables numéricas y sobre este archivo se calculan las correlaciones entre ellas.

```
var_num <- CX[,c(1,9,10,11,12,13,14,15,16)]
suppressWarnings(suppressPackageStartupMessages(library(PerformanceAnalytics)))
chart.Correlation(var_num, histogram = F, pch = 19)
```

## EQUILIBRADO DE LA MUESTRA

Se especifica un modelo de clasificación sobre la variable dependiente Entire Order Process. Las observaciones excelentes con puntuación 5 se les asigna el valor "EXCELENTE" y, a las puntuaciones entre 1 y 4 se les asigna el valor "NO\_EXCELENTE"

```
CX[CX$`Entire Order Process` != 5, 16] <- "NO_EXCELENTE"
CX[CX$`Entire Order Process` == 5, 16] <- "EXCELENTE"
CX$`Entire Order Process` <- as.factor(CX$`Entire Order Process`)
```

Equilibrado de la muestra. Se inspecciona la variable dependiente. Se recuentan más de 656 casos en los que la variable toma el valor "EXCELENTE" y 297 casos en los que toma el valor "NO\_EXCELENTE". La diferencia es muy grande por lo que se hace necesario un equilibrado de la muestra. Se realiza un balanceo a nivel de datos. Se realiza un submuestreo de los datos con valor "EXCELENTE" mediante el método del cubo.

```
CX_SI <- subset(CX, CX$`Entire Order Process` == "EXCELENTE")
# Creamos las variables indicadores para cada una de las variables de equilibrio
suppressWarnings(suppressPackageStartupMessages(library(sampling)))
X1 <- disjunctive(CX_SI$`Created` by`)
sumario<-summary(CX_SI$`Created` by`)
sumario<-as.vector(names(sumario))
# Se suprimen dos niveles de factor sin observaciones en la submuestra
sumario<-sumario[c(-3,-19)]
colnames(X1) <- levels(sumario)

X2 <- disjunctive(CX_SI$Region) levels(CX_SI$Region)
colnames(X2) <- levels(CX_SI$Region)

X3 <- disjunctive(CX_SI$`Order` Creation Date`)
colnames(X3) <- levels(CX_SI$`Order` Creation Date`)

X4 <- disjunctive(CX_SI$`Customer` Group`)
sumario4<-summary(CX_SI$`Customer Group`)

X5 <- disjunctive(CX_SI$`Customer Revenue 2021`)
colnames(X5) <- levels(CX_SI$`Customer Revenue 2021`)

# Creamos también una variable que vale 1 en todos los registros
# (para comprobar la estimación del tamaño poblacional)
UNO = rep(1, dim(CX_SI)[1])

# Construimos la matriz de equilibrio a partir de estas variables
X <- cbind(UNO, X1, X2, X3, X4, X5)

# Calculamos las probabilidades de inclusión.
# En este caso se trata de un m.a.s. con tamaño muestral de nB = 297
# Por lo tanto, la prob. de inclusión de cada individuo es 297/nA;
# donde nA es el tamaño de la población.
nB = 297
nA = nrow(CX_SI)
pik = rep(nB/nA, nA)

# Seleccionamos la muestra con la matriz de equilibrio X
# Order=1; los datos son ordenados aleatoriamente
# method=2; fase de aterrizaje mediante supresión de variables
s = samplecube(X, pik, order=1, comment = FALSE, method = 2)
muestra.SI = cbind(CX_SI, s)
```

```

muestra.SI <- subset(muestra.SI, s == 1)
muestra.SI$s <- NULL

# Data frame que contiene la población ENTIRE ORDER PROCESS = NO EXCELENTE
muestra.NO<-subset(CX,CX$`Entire Order Process`=="NO_EXCELENTE")

```

Se combinan la submuestra "muestra.SI" con la muestra "muestra.NO"

```
CX.balanceado <- rbind( muestra.SI, muestra.NO )
```

### PREPARACION MUESTRAS DE ENTRENAMIENTO Y DE TEST.

Se cambian los nombres de las var. Indep. Por otros más cortos.

```

colnames(CX.balanceado)<-c("SOV", "CREATEDBY", "REGION", "ORDERDATE", "CLIENTE", "COUNTRY", "CUSTOMERGR", "CUSTOMERREV", "CSAV", "CSFR", "CSCO", "CSRE", "PRODAV", "DEL TIME", "ONTIMEDEL", "ENTIRE")

```

#Quitamos la columna de CREADO POR, pais y cliente

```
CX.balanceado <- CX.balanceado[,c(-2,-5,-6)]
```

# Preparamos las muestras de entrenamiento y de test. Se define una semilla para no tener distintos resultados.

```
set.seed(99)
```

# Se define en 80% el tamaño de la muestra de entrenamiento.

```
train_sample<-sample(nrow(CX.balanceado),0.8*nrow(CX.balanceado))
```

```
entrena<-CX.balanceado[train_sample,]
```

```
test<-CX.balanceado[-train_sample,]
```

# Se carga la librería caret

```
suppressWarnings(suppressPackageStartupMessages(library(caret)))
```

#Se crea la función fiveStats que devolverá formateadas las 5 métricas

```
fiveStats<-function(...){twoClassSummary(...),defaultSummary(...)}
```

#Se configura el traincontrol con validación cruzada, 5 iteraciones de remezcla y 3 repeticiones.

```
control<-trainControl(method="repeatedcv",number=5,repeats=3, classProbs=T,summaryFunction=fiveStats)
```

## ESTIMACIONES DE MODELOS

### MODELOS LINEALES GENERALIZADOS

Se muestra a continuación el resultado de las métricas obtenidas por un modelo GLM de familia logit con los datos de entrenamiento:

```
# Se realiza la estimación de una clasificación logística mediante un modelo GLM
```

```
GLM<-train(ENTIRE~,data = entrena,method="glm",family=binomial("logit"), metric="ROC",trControl=control)
```

```
kable(GLM$results)%>%kable_styling(bootstrap_options = "striped",full_width = F, position = "left")
```

Se muestran las variables más importantes según el modelo GLM logit:

```
plot(varImp(GLM))
```

Se calcula el ROC del modelo GLM logit con los datos de test y se presenta la curva ROC:

```
# Se calculan las predicciones con la muestra de test
```

```
GLM.test<-predict(GLM,test,type = "prob")
```

```
# Se calcula el AUC de la muestra de test
```

```
suppressWarnings(suppressPackageStartupMessages(library(pROC)))
```

```
GLM.test.ROC<-roc(test$ENTIRE,GLM.test[, "EXCELENTE"])
```

```
# Grafico de la curva ROC con los datos de test
```

```
plot.roc(GLM.test.ROC,col = "red", main="Curva ROC del modelo GLM")
```

Se muestra a continuación el resultado de las métricas obtenidas por un modelo GLM de familia probit con los datos de entrenamiento:

```
# Se realiza la estimación de una clasificación mediante un modelo GLM tipo probit
```

```
GLM.probit<-train(ENTIRE~.,data = entrena,method="glm",family=binomial("probit"), metric="ROC",trControl=control)
```

```
kable(GLM.probit$results)%>%kable_styling(bootstrap_options = "striped",full_width = F, position = "left")
```

Se muestran las variables más importantes según el modelo GLM probit:

```
plot(varImp(GLM.probit))
```

Se calcula el ROC del modelo GLM probit con los datos de test y se presenta la curva ROC:

```
# Se calculan las predicciones con la muestra de test
```

```
GLM.probit.test<-predict(GLM.probit,test,type = "prob")
```

```
# Se calcula el AUC de la muestra de test
```

```
suppressWarnings(suppressPackageStartupMessages(library(pROC)))
```

```
GLM.probit.test.ROC<-roc(test$ENTIRE,GLM.probit.test[, "EXCELENTE"])
```

```
# Grafico de la curva ROC con los datos de test
```

```
plot.roc(GLM.probit.test.ROC,col = "red", main="Curva ROC del modelo GLM probit")
```

## MODELOS DE ÁRBOLES DE DECISIÓN

Se realiza a continuación la estimación de un modelo CRT con el conjunto de datos de entrenamiento y se muestran las métricas obtenidas:

```

# Después de realizar varias pruebas sobre la configuración del grid se realiza la estimación a través de un modelo de árbol de decisión tipo CRT con 4 predictores.

CRT.grid<-expand.grid(cp=seq(0,0.05,0.005))

CRT<-train(ENTIRE~,data=entrena,method="rpart",metric="ROC",trControl=control,tuneGrid=CRT.grid)

rpart.plot(CRT$finalModel)

```

Se muestran las variables más importantes del modelo CRT:

```
plot(varImp(CRT))
```

Se calcula el ROC del modelo CRT con los datos de test y se presenta la curva ROC:

```

# Se calculan las predicciones con la muestra de test

CRT.test<-predict(CRT,test,type = "prob")

# Se calcula el AUC de la muestra de test

suppressWarnings(suppressPackageStartupMessages(library(pROC)))

CRT.test.ROC<-roc(test$ENTIRE,CRT.test[, "EXCELENTE"])

# Grafico de la curva ROC con los datos de test

plot.roc(CRT.test.ROC,col = "red", main="Curva ROC del modelo CRT")

```

Se prueba, un árbol C5.0, se realiza su estimación usando el conjunto de datos de entrenamiento. Se muestran las métricas obtenidas:

```

# Después de varias pruebas sobre la configuración del grid se realiza la estimación a través de un modelo de árbol de decisión tipo C5.0 con 10 trials, modelo de reglas y con winnow.

suppressWarnings(suppressPackageStartupMessages(library(C50)))

suppressWarnings(suppressPackageStartupMessages(library(plyr)))

C5.grid<-expand.grid(trials=5,model="rules",winnow=T)

C5<-train(ENTIRE~,data=entrena,method="C5.0",metric="ROC",trControl=control,tuneGrid=C5.grid)

# Se imprimen los resultados

kable(C5$results)%>%kable_styling(bootstrap_options = "striped",full_width = F, position = "left")

```

Se muestran las variables más importantes del modelo C5:

```
plot(varImp(C5))
```

Se calcula el ROC del modelo C5 con los datos de test y se presenta la curva ROC:

```

# Se calculan las predicciones con la muestra de test

C5.test<-predict(C5,test,type = "prob")

# Se calcula el AUC de la muestra de test

```

```

suppressWarnings(suppressPackageStartupMessages(library(pROC)))

C5.test.ROC<-roc(test$ENTIRE,C5.test[, "EXCELENTE"])

# Grafico de la curva ROC con los datos de test

plot.roc(C5.test.ROC,col = "red", main="Curva ROC del modelo C5")

```

Se calcula el ROC del modelo RF con los datos de test y se presenta la curva ROC:

```

# Se calculan las predicciones con la muestra de test

RF.test<-predict(RF,test,type = "prob")

# Se calcula el AUC de la muestra de test

suppressWarnings(suppressPackageStartupMessages(library(pROC)))

RF.test.ROC<-roc(test$ENTIRE,RF.test[, "EXCELENTE"])

# Grafico de la curva ROC con los datos de test

plot.roc(RF.test.ROC,col = "red", main="Curva ROC del modelo RF")

```

## MODELOS DE REDES NEURONALES

Se realiza a continuación la estimación de un modelo Perceptron Multicapa con 30 neuronas usando el conjunto de datos de entrenamiento. Se muestran las métricas obtenidas:

```

# Después de realizar varias pruebas sobre la configuración del grid se realiza la estimación a través de un modelo de RNA tipo Perceptron Multicapa con 30 neuronas.

PM.grid<-expand.grid(size=30)

PM<-train(ENTIRE~.,data=entrena,method= "mlp", metric="ROC", trControl=control,
tuneGrid=PM.grid)

# Se imprimen los resultados

kable(PM$results)%>%kable_styling(bootstrap_options = "striped",full_width =
F, position = "left")

```

Se muestran las variables más importantes del modelo Perceptron Multicapa:

```
plot(varImp(PM))
```

Se calcula el ROC del modelo Perceptron Multicapa con los datos de test y se presenta la curva ROC:

```

# Se calculan las predicciones con la muestra de test

PM.test<-predict(PM,test,type = "prob")

# Se calcula el AUC de la muestra de test

suppressWarnings(suppressPackageStartupMessages(library(pROC)))

PM.test.ROC<-roc(test$ENTIRE,PM.test[, "EXCELENTE"])

# Grafico de la curva ROC con los datos de test

plot.roc(PM.test.ROC,col = "red", main="Curva ROC del modelo PM")

```

Se realiza a continuación la estimación de un modelo RNA Función de base radial con 30 neuronas usando el conjunto de datos de entrenamiento. Se muestran las métricas obtenidas:

```
# Despues de realizar varias pruebas sobre la configuración del grid se realiza la estimación a través de un rna tipo Funcion de base radial con 30 neuronas.
```

```
RBF.grid<-expand.grid(size=30)
```

```
RBF<-train(ENTIRE~,data=entrena,method= "rbf", metric="ROC", trControl=control, tuneGrid=RBF.grid)
```

```
# Se imprimen los resultados
```

```
kable(RBF$results)%>%kable_styling(bootstrap_options = "striped",full_width = F, position = "left")
```

Se muestran las variables más importantes del modelo función de base radial:

```
plot(varImp(RBF))
```

Se calcula el ROC del modelo Función de base radial con los datos de test y se presenta la curva ROC:

```
# Se calculan las predicciones con la muestra de test
```

```
RBF.test<-predict(RBF,test,type = "prob")
```

```
# Se calcula el AUC de la muestra de test
```

```
RBF.test.ROC<-roc(test$ENTIRE,RBF.test[, "EXCELENTE"])
```

```
# Grafico de la curva ROC con los datos de test
```

```
plot.roc(RBF.test.ROC,col = "red", main="Curva ROC del modelo RBF")
```

## MODELO BAYESIANO

Se realiza a continuación la estimación de un modelo Bayesiano usando el conjunto de datos de entrenamiento. Se muestran las métricas obtenidas:

```
# se realiza la estimación a través de un modelo tipo Bayesiano.
```

```
suppressWarnings(suppressPackageStartupMessages(library(naivebayes)))
```

```
BY<-train(ENTIRE~,data=entrena,method= "naive_bayes", metric="ROC", trControl=control)
```

```
BY$results
```

Se muestran las variables más importantes del modelo Bayesiano:

```
plot(varImp(BY))
```

Se calcula el ROC del modelo Bayesiano con los datos de test y se presenta la curva ROC:

```
# Se calculan las predicciones con la muestra de test
```

```
BY.test<-predict(BY,test,type = "prob")
```

```
# Se calcula el AUC de la muestra de test
```

```

BY.test.ROC<-roc(test$ENTIRE,BY.test[, "EXCELENTE"])

# Grafico de la curva ROC con los datos de test

plot.roc(BY.test.ROC,col = "red", main="Curva ROC del modelo bayesiano")

```

**COMPARACIÓN DE RESULTADOS ENTRE MODELOS**

```

# Se crea un vector con los valores AUC obtenidos en cada modelo"

ROC<-c(round(GLM.test.ROC$auc,4),round(GLM.probit.test.ROC$auc,4),round(RF.te
st.ROC$auc,4),round(C5.test.ROC$auc,4),round(PM.test.ROC$auc,4),round(RBF.tes
t.ROC$auc,4),round(BY.test.ROC$auc,4))

# Se establece un vector con los nombres de los modelos

names.models<-c("GLM logit","GLM probit","Random Forest","C5","Perceptrón Mul
ticapa","Función de base radial","Bayesiano")

#Se crea una matriz con los datos obtenidos y los nombres de las filas y colu
mnas

ROC.modelos<-matrix(data=ROC,ncol=1,dimnames = list(names.models,"ROC"))

# Se imprime la tabla con los resultados

ROC.modelos%>%
  kbl(caption = "Tabla de resultados") %>%
  kable_paper(full_width = F)%>%
  kable_styling(bootstrap_options = c("striped", "hover"),full_width = F)

```