Grado en Estadística

TRABAJO FIN DE GRADO

Modelización estadística de ventas en el sector retail

Marta Venegas Pardo

Sevilla, Noviembre de 2021

Índice general

Prólogo

Escrito colocado al comienzo de una obra en el que se hacen comentarios sobre la obra o su autor, o se introduce en su lectura; a menudo está realizado por una persona distinta del autor.

También se podrían incluir aquí los agradecimientos.

Resumen

Resumen...

Aquí comienza mi resumen, que lo haré al final del estudio

Abstract

Abstract...

Índice de figuras

Índice de tablas

Introducción

Esta es la introducción de mi trabajo y se irá modificando conforme se vaya avanzando. FALTA UNA BREVE INTRODUCCIÓN Y PUESTA EN SITUACIÓN **

El concepto de retail es una orientación de la dirección del negocio que sostiene que las tareas clave de un minorista:

- Determinar las necesidades y deseos de su mercado objetivo
- Dirigir la empresa hacia la satisfacción de esas necesidades y deseos de forma más eficiente que sus competidores (Vigaray, 2005).

El comercio detallista o minorista es el último eslabón de la distribución comercial, siendo el intermediario que se dedica a la venta de productos, bienes o servicios a los consumidores o usuarios finales (Burruezo, 1999).

Este sector aglutina a comerciantes y empresas encargadas de la comercialización, ofreciendo de una gran variedad de productos y servicios a los consumidores. (Una tienda, un supermercado o una librería son claros ejemplos de lo que es el sector retail).

A continuación podemos destacar los siguientes objetivos de este estudio:

- Analizar y entender como varía la demanda de productos en función del tiempo y de otros factores
- Descubrir asociaciones y patrones entre productos aplicando un análisis de cesta de la compra
- Investigar y revisar diversos algoritmos de aprendizaje automático y aprendizaje estadístico para modelar las ventas de los diferentes productos
- Utilizar Rstudio, un entorno de desarrollo integrado para el lenguaje de programación R como soporte al estudio estadístico de los datos

Para ello, se llevará a cabo una revisión teórica de técnicas para la modelización de la variable *volumen de ventas* para posteriormente aplicarlas a un caso práctico real de ventas de productos lácteos, con el propósito final de construir un modelo que ayude a predecir la demanda futura, con una *precisión* mínima del 70 %. (*Nota para mi: no se si esto será posible, puede que modifique esto último de la precisión mínima*)

Dentro de las técnicas que se van a exponer a lo largo del trabajo, podemos distinguir dos vertientes: las técnicas puramente de aprendizaje estadístico y técnicas de aprendizaje automático (Machine Learning).

El término *Machine Learning* (ML, Aprendizaje Automático) se utiliza en el campo de la Inteligencia artificiál para referirse a algoritmos de predicción. Muchas de estas técnicas provienen del campo de la Estadística y por tanto, esta rama aplicada de las Matemáticas es la base de todos estos modelos para analizar datos. Por otro lado, desde el campo de la Estadística Computacional, se introdujo el término *Statistical Learning* (AE, Aprendizaje Estadístico) para referirse a este tipo de herramientas desde un punto de vista estadístico, es decir, se tiene en cuenta la incertidumbre debida a no disponer de toda la información.

Además, el ML no se preocupa del origen de los datos, siendo frecuente la consideración de un conjunto enorme de datos, lo que equivale a disponer toda la información (la población completa). Por el contrario, en el caso del AE, se trata de comprender la estructura de los datos y si son representativos de la población de interés.

Siguiendo esta línea, en el año 2001, Leo Breiman publica *Modelos Estadísticos*, donde diferencia dos objetivos en el análisis de datos, que el define como *información* y *predicción*. Cada uno de ellos da lugar a una cultura en el uso de modelos estadísticos para llegar a conclusiones a partir de los datos:

- *Modelización de datos*: se trata del desarrollo de modelos estocásticos que permitan ajustar los datos y realizar inferencia.
- *Modelización algorítmica* (en sentido predictivo): esta cultura está interesada en los algoritmos de predicción, no en los mecanismos que generan los datos, siendo el ML la base de esta cultura.

Para tratar de cumplir los objetivos, se ha estructurado el trabajo en diferentes capítulos. Los primeros cuatro capítulos consisten en la revisión teórica de todas las técnicas que posteriormente aplicaremos a un caso práctico. En primer lugar, un primer capítulo introductorio sobre el proceso de la ciencia de datos en el sector retail, a continuación, un segundo capítulo para la explicación de lo que se conoce como análisis de cesta de la compra, que tiene como objetivo conocer las asociaciones entre diversos productos. Los capítulos tres y cuatro explican respectivamente, las técnicas clásicas de modelado estadístico y el proceso de ciencia de datos completo con las correspondientes técnicas de aprendizaje automático. Un quinto capítulo donde se aplicarán todas las técnicas estudiadas a un caso práctico con datos reales. Por último, el capítulo sexto se exponen las conclusiones extraídas y se analizan los objetivos iniciales.

Frecuentemente la investigación estadísticas se ve enfrentada a manipular grandes cantidades de datos complejos que incluyen un gran número de variables, de los cuales es necesario obtener información, encontrar patrones y definir tendencias. (*Nota para mi: creo que este párrafo aquí sobra*)

Capítulo 1

La ciencia de datos en el sector retail

1.1. Modelización estadística de ventas en el sector retail

En la actualidad gran cantidad de empresas y organizaciones almacenan una cantidad creciente de datos de cualquier actividad que desenpeñen. Sin embargo, no todos los datos son luego procesados, es por ello que se busca implementar herramientas que ayuden a procesar estos datos y poder así transformarlos en información útil para el negocio que ayude en la toma de decisiones, ya que en muchas empresas, hasta hace unos años, estas decisiones eran tomadas de manera intuitiva o por conocimientos históricos que poseía la empresa debido a situaciones previas.

Se pueden aplicar diferentes técnicas para que empresas del sector retail se vean beneficiadas al conocer los patrones de compra de sus clientes generando así indicadores para poder llevar un control de la información y que esta sea de utilidad para la mejor toma de decisiones haciendo uso de herramientas Big Data como por ejemplo la Minería de Datos (*Data Mining*), que es una técnologia computarizada encargada de extraer, recoger, depurar y modelar grandes volúmenes de datos para así encontrar tendencias y patrones de coincidencia entre los datos que inicialmente eran desconocidos. Por tanto, la minería de datos es capaz de extraer información útil de los datos que poseen las empresas, y es por tal motivo que el usar esta herramienta con éxito proporcionará a la empresa una ventaja frente a sus competidores.

Por ejemplo, una tienda que ofrezca a sus clientes un servicio en línea, podrá tener acceso a todos los datos de los consumidores de sus servicios y/o productos. Esta información les servirá para conocer a sus clientes, sus preferencias y tendencias, pudiendo ofrecerle un servicio más adecuado o personalizado.

Entre las diferentes técnicas de la minería de datos, podemos encontrar: redes neuronales, regresión lineal, árboles de decisión, reglas de asociación, agrupamiento, análisis factorial, series temporales y pronóstico (forecasting). De todas ellas, las *reglas de asociación* son una herramienta muy potente, ya que encuentra los diferentes hechos que tienen en común un conjunto de datos y los asocia. Además, también es útil para encontrar asociaciones y/o correlaciones entre los datos. Esta técnica puede ser muy útil a la hora de poner precios a los productos y para desarrollar distintas herramientas de marketing para llegar al cliente con éxito.

La minería de datos es una heramienta aplicada por muchas empresas del sector retail, perteneciendo a muchos sectores como las finanzas, salud, bancos, sercicios públicos y seguros; ya que todos estos sectores tienen una gran cantidad de datos referidos a sus clientes, proveedores, productos y/o servicios.

Estos datos pueden ser de tipo demográfico (como edad, sexo o estado civil), datos económicos (carrera, puesto de trabajo, ingresos por familia) o datos geográficos (país, ciudad de residencia, dirección,...)

A continuación, vamos a exponer a modo de aplicación dos ejemplos de empresas del sector Retail que aplican la Minería de Datos:

- **Wal Mart**

  Se trata de corporación multinacional de tiendas de origen estadounidense, que vende al por menor y funciona como una cadena de hipermercados, almacenes grandes de descuento y almacenes de comestibles.

  Esta multinacional es una de las pioneras en el uso de la minería de datos y la gestión de sus datos. Toma datos sobre transacciones que se realizan en sus 2900 tiendas, almacenándolos en una base de datos con una capacidad de 1.5 terabytes. La empresa ofrece a sus proveedores información sobre los productos, para que éstos pueden identificar los patrones de compra de sus cliente, logrando así la gestión de los inventarios en el almacén.

- **Papas "Chips"**

  Es una empresa que distribuye productos como golosinas y refrescos. La empresa utiliza la minería de datos para ofrecer a sus clientes un buen servicio, obteniendo así la fidelización de los mismos, lo cual repercute positivamente en los ingresos de la empresa.

  La empresa aplica técnicas de minería de datos para llevar un registro de ventas, conociendo así los períodos con un volumen de ventas mayor para un producto o productos determinados. De esta forma, se logra abastecer a los clientes teniendo siempre productos en stock.

1.2. La Minería de Datos y el Sector Retail

Como ya sabemos, Retail hace referencia al detalle de productos, por tanto, se trata de un sector empresarial enfocado a productos que se venden de manera masiva, por lo que es un sector con una gran cantidad de clientes, dado que es el contacto directo con el consumidor final de cada producto ofrecido en el mercado.

La Minería de Datos y el Sector Retail están relacionados dependiendo del tipo de la empresa en cuestión, su tamaño o el tipo de cliente (consumudores, familias, minoristas, supermercados, centros comerciales, bancos, pequeños establecimientos de venta como tiendas,...), por tanto, esta herramienta es útil para analizar todo tipo de bases de datos con el objetivo de obtener información sobre los clientes, segmentarlos en función de sus necesidades, tendencias de compra, ...

Al aplicar la minería de datos en este sector, se pueden obtener resultados como los patrones de compra, las preferencias de los clientes o asociaciones entre productos que no

conocíamos, abriendo la posibilidad a la creación de distintas estrategias como ofertas o packs.

Por tanto, como resultado de la aplicación de la minería de datos en el sector retail, obtendremos información relevante para cada empresa, ya que nos mostrará el comportamiento de cada tipo de cliente que tenga cada empresa mediante la aplicación de diferentes técnias y herramientas estadísticas.

Capítulo 2

Análisis de cesta de la compra (Market basket análisis)

En este capítulo revisaremos el método de análisis de cesta de la compra con el fin de identificar patrones o asociaciones entre diversos grupos de productos. Además de la validación de distintas asociaciones que se pueden considerar obvias debido a una reiterada compra conjunta, esta metodología pretende encontrar relaciones entre productos cuya asociación no es tan evidente. La principal fuente de información para llevar a cabo este tipo de análisis serán datos que recogen las transacciones de los clientes en cada compra.

El concepto de reglas de asociación fué conocido en el año 1993 tras la publicación de un artículo de Agrawal, R; Imielinski, T.; Swami, A. *Mining association rules between sets of items in large databases*. En este artículo se introducen las reglas de asociación para descubrir relaciones entre las transacciones de productos registrados por los sistemas de punto de venta de los supermercados.

A continuación vamos a definir una serie de conceptos para entender mejor este método. Una *asociación* es una concurrencia de dos o más cosas. Por ejemplo, los perritos calientes pueden estar asociados con los refrescos, los pepinillos o el ketchup. Se dice que existe una asociación *positiva* si la presencia de algunos productos implica la presencia de otros dentro de la misma transacción. También existe la asociación *negativa*, la cual consiste en que la presencia de algunos productos haga muy improbable la presencia de otros elementos en la misma transacción.

Se conoce como *reglas de asociación* a la agrupación de productos en función de la afinidad existente entre ellos. la utilidad de estas reglas se encuentra en la identificación de oportunidades y el diseño de diversos grupos de productos que puedan ser atractivos para los consumidores. Estas reglas tienen como finalidad el descubrimiento de las relaciones implícitas en los datos.

Cuando aplicamos un análisis de cesta de la compra a unos datos, habitualmente podemos obtener tres posibles resultados generados a partir de las reglas de asociación:

- **Resultados factibles**: Contienen información útil y de calidad, ya que los patrones identificados son factibles. Estas reglas nos podrían ayudar en la toma decisiones como por ejemplo, saber a que productos se aplicar una promoción con el objetivo de impulsar las ventas de los mismos y de sus productos relacionados.

- **Resultados triviales**: Estos con conocidos por aquellos que están familiarizados con el negocio en cuestión. Este tipo de reglas nos muestra productos en los que no se plantea la compra de un producto sin otro, por ejemplo, la compra de pintura requiere la necesidad de comprar también pinceles.
- **Resultados imposibles**: Estos resultados implican la obtención de reglas incongruentes, ya que nos proporciona un información que no facilita el entender el comportamiento del consumidor, sino que se trata de casos puntuales en un momento determinado.

2.1. Introducción

2.2. Definición

2.3. Desarrollo

Capítulo 3

Modelos estadísticos clásicos

A continuación se exponen los modelos estadísticos clásicos de cara a predecir la demanda de los productos. Esta varaible depende de varios factores, como el período del año, el precio del producto, el precio de los productos competidores o los gustos de cada consumidor, entre otros. Se trata de una variable cuantitativa discreta, ya que el número de productos que se venden será un valor entero no negativo $y = 0, 1, \ldots$.

3.1. Modelo de Regresión Lineal General

El objetivo es encontrar un modelo estadístico que describa la situación real de ventas de productos a través de un Modelo Lineal General (MLG), donde una variable de interés (variable objetivo) pueda ser descrita por un conjunto de variables explicativas (variables independientes).

Para ello, debemos estimar los parámetros que caracterizan al modelo, es decir, medir el efecto de cada variable explicativa sobre la variable objetivo, y con este fin, es necesario definir una serie de hipótesis del modelo de regresión lineal general:

- Independencia lineal entre las variables explicativas: Esto significa que cada variable explicativa contiene información adicional sobre la variable objetivo, ya que si hubiera información repetida, habría variables explicativas dependientes linealmente de otras.
- Los MLG suponen que existe una función g, llamada función link, que relaciona la media de la variable respuesta, μ con el resto de variables explicativas de la siguiente forma:

$$E[Y] = \mu = g^{-1}(\eta) = g^{-1}(X^t\beta)$$

Siendo:

- Y la variable objetivo
- E(Y) es el valor esperado de la variable Y
- $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = X^t\beta$ es el predictor lineal, se trata de una combinación lineal de parámetros desconocidos
    - $X_1, \ldots, X_p$ son las variables explicativas

- $\beta = (\beta_o, \beta_1, \ldots, \beta_p)$ representan el efecto de cada variable independiente sobre la variable objetivo
- g es la función link, monótona y diferenciable

3.1.1. Componentes del Modelo Lineal Generalizado

En este tipo de modelización estadística podemos diferenciar tres componentes: la componenete aleatoria, la sistemática y la función link o enlace. Será la combinación de estas tres componentes la que defina por completo un Modelo Lineal Generalizado.

3.1.1.1. Componente aleatoria

Esta componente es la que identifica la variable respuesta y su distribución de probabilidad.

Sea Y la variable aleatoria objetivo o variable respuesta objeto de estudio y sean las n variables aleatorias independientes e idénticamente distribuidas $Y_1, \ldots, Y_n$ la muestra aleatoria procedente de Y. Siendo Y la componente aleatoria cuya distribución pertenece a la familia exponencial de distribuciones.

En los MLG se supone que la variable respuesta Y se distribuye de tal forma que su función de probabilidad, en el caso de estar modelizando una variable discreta o función de densidad para el caso continuo viene dada por la sigiente expresión general, conocida como forma canónica:

$$f(y; \theta, \phi) = a(y, \phi) \cdot e^{\left(\frac{y\theta - k(\theta)}{\phi}\right)}$$

Donde

- θ es el parámetro canónico
- $k(\theta)$ es la función cumulante
- $\phi > 0$ se trata del parámetro de dispersión
- $a(y, \phi)$ es una constante normalizadora
- El soporte no depende de θ ni de ϕ

Además, la media de y es función del parámetro canónico θ, por tanto, se tiene que:

$$E(Y) = \mu = \frac{\partial}{\partial\theta}k(\theta), Var(Y) = \sigma^2 = \phi\frac{\partial^2}{\partial\theta^2}k(\theta) = \phi\frac{\partial}{\partial\theta}\left(\frac{\partial}{\partial\theta}k(\theta)\right) = \phi\frac{\partial}{\partial\theta}\mu > 0.$$

Es decir, μ es una función estrictamente creciente de θ, por lo que estos dos parámetros mantienen una relación biyectiva.

A

$$V(\mu) = \frac{\partial\mu}{\partial\theta}$$

se le denomina función varianza, por lo que se tiene que:

$$V(Y) = \phi Var(\mu)$$

3.1.1.2. Componente sistemática

Se trata de la componente que especifica las variables predictoras utilizadas en la función predictora lineal en forma de efectos fijos de un modelo lineal y recoge la variabilidad de la respuesta Y expresada a través de p variables explicatibas $X_1, \ldots, X_p$, que denotamos por X, y de los correspondientes parámetros desconocidos $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$.

Esta componente, tambien conocida como predictor lineal, viene representada por η y es una combinación lineal de las variables explicativas, que viene dada por:

$$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = X^t\beta = X^t\beta$$

3.1.1.3. Función link o función enlace

La *función link*: se trata de una función del valor esperado de la variable respuesta $E[Y]$, como una combinación lineal de las variables predictoras. Sin embargo, en muchos casos reales esta relación no es adecuada, por lo que es necesario la introducción de una función con el objetivo de relacionar el valor esperado con las variables explicativas. Por ello, introducimos la función link o función enlace, $g(\cdot)$ que relaciona μ con el predictor lineal de la siguiente forma:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

En problemas reales, pueden existir varias funciones link, por lo que se elegirá aquella que facilite la interpretación del modelo óptimo obtenido. En particular, para cada elemento de la familia exponencial existe una función enlace denominada función canónica, que permite relacionar el parámetro canónico con el predictor lineal.

$$\theta_i = \theta(\mu_i) = \eta_i = X_i^t\beta \quad g(\mu_i) = \theta(\mu_i)$$

3.1.2. Estimación de parámetros

Tras la construcción de los modelos, se estiman los parámetros desconocidos del predictor lineal, $\beta = (\beta_o, \beta_1, \ldots, \beta_p)'$ por $\hat{\beta} = (\hat{\beta}_o, \hat{\beta}_1, \ldots, \hat{\beta}_p)'$ y el valor del parámetro de dispersión ϕ por $\hat{\phi}$. Posteriormente, se valora la precisión de las estimaciones con el objetivo de seleccionar un modelo óptimo.

Generalmente, la estimación de los parámetros se lleva a cabo por el método de la *Máxima Verosimilitud* o el método de *Mínimos Cuadrados Ordinarios*. Una vez desarrollados los modelos, se realizará una comparación de los mismos con el objetivo de seleccionar el mejor. En el caso del modelado con fines predictivos, se selecionará el modelo que explique el mayor porcentaje de variabilidad de la respuesta.

Para ello, emplearemos el **Criterio de información de Akaike (AIC)**, medida relativa de la calidad de un modelo estadístico. Este criterio trata de proporcionar una compensación entre la bondad de ajuste del modelo y la complejidad del mismo, es decir, el criterio penaliza al número de parámetros.

Dado un conjunto de modelos candidatos para los datos, el modelo preferido es aquel que tiene mínimo valor del AIC.

En el caso general, el AIC viene dado por la siguiente expresión:

$$AIC = 2k - 2ln(\hat{L})$$

Siendo:

- k el número de parámetros del modelo
- $\hat{L}$ es el máximo valor de la función de verosimilitud para el modelo estimado

Otro criterio en el que nos basaremos es en el criterio de bondad de ajuste, destacando el cálculo del **coeficiente de determinación** R^2, que es una medida del grado de fiabilidad o bondad de ajuste del modelo ajustado a un conjunto de datos. Se trata de una medida acotada por definición, siendo sus límites $0 \leq R^2 \leq 1$. Un coeficiente de determinación igual a 1 indica un ajuste lineal perfecto, y por tanto, la variación total de la variable Y es explicada por el modelo de regresión. Por el contrario, el valor 0 indica que el modelo no explica nada de la variación total de la variable Y.

Para la bondad de ajuste, otra medida interesante es el RMSE, raíz del error cuadrático medio. Representa la raíz cuadrada de la distancia promedio entre el valor real y el pronosticado e indica el ajuste absoluto del modelo a los datos, es decir, cómo de cerca están los puntos observados de los valores predichos del modelo. Valores más bajos de RMSE indican un mejor ajuste.

En muchos casos la variable respuesta es de tipo conteo, como lo es la variable que queremos modelizar, demanda de productos. Se denominan variables de recuento o variables de tipo conteo, a aquellas que determinan el número de sucesos que ocurren en una misma unidad de observación en un intervalo espacial o temporal definido. Esta variable Y, puede tomar infinitos números de valores y su probabilidad va en descenso a medida que sea mayor el valor de la variable.

Para este caso, los modelos que tienen especial interés y que podemos formalizar a través de modelización lineal son el modelo de *Poisson* y el modelo de *Binomial negativa*. Estos modelos permiten analizar el comportamiento de variables de conteo frente a los valores del conjunto de variables explicativas.

3.1.3. Modelo de Regresión Poisson

Se trata del modelo más simple y es el modelo de referencia para variables respuesta de tipo conteo. Este modelo asume que la variable respuesta Y sigue una distribución de Poisson, por lo que en el caso de la modelización de ventas, se define como el número de ventas que ocurren en un intervalo de tiempo, cuya ocurrencia es aleatoria.

Esta distribución se caracteriza por que la media y varianza coinciden:

$$E(Y) = Var(Y) = \mu$$

Se tiene que la distribución de probabilidad de Poisson, y en nuestro caso, la probabilidad de observar y ventas es:

$$P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, \quad y = 0, 1, \ldots; \mu > 0$$

Y por tanto, la forma canónica o componente aleatoria para esta distribución es la siguiente:

$$f(y;\mu) = e^{-\mu} \cdot \frac{\mu^y}{y!} = \frac{1}{y!} e^{ylog(\mu)-\mu}, \quad y \in \{0, 1, \dots\}$$

Es decir, el modelo de Posión se obtiene tomando como función enlace el parámetro canónico.

donde:

- $\theta = log(\mu)$ es el parámetro canónico
- $k(\theta) = \mu = e^{\theta}$ es la función cumulante
- $\phi = 1$ el parámetro de dispersión
- $a(y, \phi) = 1/y!$ la constante normalizadora

En este caso se tiene que: $g(\mu_i) = X_i^t \beta$ y una elección usual de la función link g el parámetro canónico, $g(x) = log(x)$, lo que equivale a:

$$\mu_i = exp(\beta_o) \cdot \exp(x_{i1}\beta_1) \dots exp(\beta_p x_{ip}) = exp(\beta_o + \beta_1 x_{i1} + \dots + \beta_p x_{ip}) \acute{o} log(\mu_i) = \eta_i = X_i \beta$$

así si x_i se incrementa en una unidad, entonces μ_i se multiplica por $exp(\beta_i)$. Por tanto, si $\beta_i > 0$, μ_i crece cuando x_i aumenta y si $\beta_i < 0$, μ_i decrece cuando x_i aumenta.

Este modelo se ha desarrollado suponiendo que la media y la varianza de los datos coinciden (equidispersión). Sin embargo, suele ocurrir lo que se conoce como sobredispersión, es decir, que la varianza es mayor que la media. Lo habitual es que esta situación se de debido a la existencia de heterogeneidad entre las observaciones. Cuando esto ocurra, recurriremos al modelo binomial negativa.

3.1.4. Modelo de Regresión de Binomial Negativa

Este modelo es empleado para variables de tipo conteo cuándo existe sobredispersión, es decir, la media condicional es menor que la varianza condicional (no coinciden). Existen diferentes modelos binomiales negativos en función de la variable que se trate de modelar, pero en este trabajo nos centraremos en el caso de datos de tipo conteo.

La distribución binomial negativa estudia la probabilidad de observar un número determinado de fracasos y (no se producen ventas) , antes del r-ésimo éxito (se venden r unidades) en una serie de experimentos Bernoulli independientes, siendo r un número positivo. Se tiene que esta distribución pertenece a la familia exponencial si el parámetro de dispersión ϕ es una constante.

Se dice que la variable aleatoria de conteo (número de ventas) Y_i, con $i = 1, \dots, n$ sigue una distribución Binomial Negativa de parámetros r y p, y se representa como $Y_i \sim BN(r, p)$,si su función de probabilidad viene dada por:

$$P[Y_i = y_i] = \binom{y_i + r - 1}{r - 1} p^r (1-p)^{y_i}$$

donde

- $0 < p < 1$
- $r > 0$

- $y_i = 0, 1, 2, \ldots$

Y en este caso, la forma canónica o componente aleatoria para esta distribución es la siguiente:

$$f(y;\mu) = exp\left\{ y \cdot ln(1-p) + rln(p) + ln\left(\frac{y_i + r - 1}{r-1}\right)\right\}$$

donde

- $0 < p < 1$
- $r = 0, 1, 2, \ldots$
- $y_i = 0, 1, 2, \ldots$
- $\theta = log(1-p)$ es el parámetro canónico
- $k(\theta) = -rln(p) = -rln(1-e^{\theta})$ es la función cumulante

En este caso la función link es de tipo logarítmico y viene dada por:

$$g(\mu_i) = \theta(\mu_i) = ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) = X_i^t\beta = \eta_i$$

3.2. Análisis de series temporales

Aplicaremos este modelo de predicción para tratar de identificar los patrones de la demanda anterior a lo largo del tiempo y luego proyectar (predecir) los patrones en el futuro.

Se define una serie temporal como una sucesión de datos ordenados en el tiempo que corresponden a una misma variable. Los datos son suelen ser tomados en intervalos regulares de tiempo.

Nuestro objetivo dentro del análisis de series temporales será identificar el proceso estocástico que ha sido capaz de generar la serie de estudio.

(No se si añadirlo) Se dice proceso estocástico a una colección o familia de variables aleatorias $\{X_t,\ \text{con}\ t \in T\}$ que siguen la misma ley de distribución y están relacionadas entre sí, pudiendo por este motivo, describir la información de estas variables en términos de medias, variaciones y covarianzas.

A continuación encontramos las cuatro etapas en un análisis descriptivo de series temporales para elegir un modelo que se adecue a nuestros datos:

- **Representación gráfica de la serie**. Para tener así una primera aproximación del comportamiento de la serie y la existencia de posibles tendencias.
- **Modelización**: Se trata de encontrar el modelo que mejor se ajuste a los datos.
- **Validación de los modelos**: Es necesario saber si el modelo ajustado es adecuado o no, por lo que es muy importante el estudio de los residuos.
- **Predicciones**: Una vez construido y validado un modelo, realizaremos estimaciones del futuro con nuevas observaciones.

En un enfoque clásico de series temporales, asumiremos que el comportamiento de la variable con respecto al tiempo se compone de cuatro componentes:

1. **Tendencia**: Se trata del movimiento suave y regular de la serie a largo plazo. La tendencia existe cuando hay un aumento o disminución a largo plazo de los datos. Puede ser lineal (ajuste mediante una recta) o no lineal (aproximación mediante una curva, como por ejemplo logarítmica o exponencial)
2. **Ciclo**: Componente de tipo oscilante caracterizada por movimientos recurrentes en torno a la tendencia de la serie y que se repiten cada año pero sin una frecuencia fija.
3. **Componente estacional**: Se trata de movimientos regulares dentro de la serie con una periodicidad menor a un año, es decir, aquello que ocurre generalmente y con la misma intensidad año tras año en los mismos períodos, por ejemplo, en la misma época del año o día de la semana. Vamos a denotar por L al número de estaciones.
4. **Componente irregular**: Se trata de las variaciones de la serie sin un comportamiento sistemático y que no son explicadas por las otras tres componentes

Existen diferentes modelos de combinación de las componentes. Para describir los modelos necesitamos primero una nomenclatura básica. Denotando por X_t al valor de la variable en el instante t, se tiene:

$$X_t = f(T_t, E_t, I_t)$$

donde:

- T_t: Valor de la tendencia en el instante t
- E_t: Valor de la componente estacional en el instante t
- I_t: Valor de la componente irregular en el instante t (ruido).

Por tanto, los modelos que puede adoptar la función f son los siguientes:

- **Modelo multiplicativo**: La composición de la serie se realiza mediante el producto de sus componentes.
$$X_t = T_x \times E_t \times I_t$$
- **Modelo aditivo**: Las componentes se agregan para formar la serie temporal.
$$X_t = T_x + E_t + I_t$$
- **Modelo mixto**: La composición de la serie de la parte irregular viene de forma aditiva y la parte regular de forma multiplicativa.
$$Xt = Tx \times Et + It$$

Tras haber detectado el modelo mas adecuado, podremos conocer el comportamiento de la serie a largo plazo.

El siguiente paso realizar una estimación de la tendencia, T_t, habiendo eliminado previamente la componente estacional para impedir que estas oscilaciones perturben la identificación de la tendencia.

Para estimar T_t, debemos hacer una hipótesis sobre su forma:

- **Tendencia determinista**: Se supone que la tendencia es una función determinística del tiempo:
$$T_t = a + bt \quad a, b \in \mathbb{R}$$

Siendo a y b constantes, que se estimarán mediante un modelo de regresión lineal.

Sin embargo, el método que aplicaremos será el que exponemos a continuación:

- **Tendencia evolutiva (método de medias móviles)**: Este método consiste en definir la tendencia como una serie suavizada. Sunpondremos que la tendencia de la serie es una función que evoluciona lentamente y que podremos aproximar función simple del tiempo, suponiendo así una recta.

Una vez identificada la tendencia, procedemos a hacer un análisis de la estacionalidad de la serie, con el objetivo de:

- **Desestacionalizar la serie**, es decir, eliminar las oscilaciones periodicas que se repiten a lo largo de los años, haciendo así que los datos de distintas estaciones sean comparables. La serie desestacionalizada la conseguimos diferenciando la serie.
- **Realizar predicciones**, ya que si nuestros datos están afectados por una componente estacional, necesitaremos una estimación de esta de cara a realizar una predicción

Para desestacionalizar la serie, emplearemos los índices de variación estacional asociados a cada estación, ya que se suponen constantes año a año. Con esta técnica, se evidencian las diferencias en cada período, por ejemplo, podemos ver la diferencia del volumen de ventas en función de la época del año (mes, día de la semana, estación,...) Estos índices reflejan la cantidad fija o proporción en la que se modifica la tendencia en cada estación.

Una vez calculados estos índices, se desestacionaliza la serie, eliminando así el efecto de cada estación.

Por último, procedemos a realizar las predicciones. Para ello, necesitamos que se cumpla la condición de estacionariedad, es decir, la media y la varianza permanecen constantes en el tiempo (no tiene raíces unitarias). En el caso de no imponer esta condición de estacionariedad, predeciríamos carasterísticas que no serán las mismas en el futuro que en el pasado.

Se tiene que todo proceso lineal es estacionario, por tanto, obtendremos trabajaremos con series estacionarias, y de lo contrario, podremos aplicar los mismos métodos a series no estacionarias realizando las transformaciones pertinentes para conseguir la estacionariedad.

En nuestro caso, aplicaremos la metodología Box-Jenkis como método predictivo.

3.2.1. Metodología Box-Jenkis

Esta metodología tiene en cuenta la dependencia existente entre los datos, es decir, cada observación en el instante t será modelada a partir de los valores pasados. Los modelos se conocen con el nombre de ARIMA (modelos integrados autorregresivos de medias móviles), que deriva de las siguientes componentes: AR (Autorregresivo) , I (integrado), MA(Medias móviles)

El siguiente paso es identificar el modelo más adecuado a través del estudio de la función de autocorrelación (FAC) y la función de autocorrelación parcial (FAP).

Nota: el método recomienda como mínimo 50 observaciones en la serie temporal.

Fases de la metodología Box-Jenkis:

1. Identificar el la estructura ARIMA que sigue la serie a través del estudio de la función de autocorrelación simple (FAS) y la función de autocorrelación parcial (FAP). Determinar el modelo arima consiste en identidicar los órdenes p y q de su estructura autoregresiva y de medias móviles
2. Estiamción de parámetros: Una vez tenemos identificado el modelo, estimamos los parámetros AR y MA del modelo por el método de máxima verosimilitud, obteniendo el error estándar y los resíduos del modelo

   Nota: Es muy importante comprobar que las estimaciones son significativamente no nulas.
3. Diagnosis del modelo: Comprobamos que los resíduos sigan un proceso de ruido blanco mediante el Test de Ljung-Box.

   Si hemos identificado varios modelos y todos ellos pasan la diagnosis, nos quedaremos con uno de ellos según el criterio del menor AIC
4. Predicción: una vez identificado y validado el mejor modelo, se realizan las predicciones con éste.

Capítulo 4

Proceso de ciencia de datos (Data science process)

4.1. Introducción

La ciencia de datos es la combinación de múltiples campos, como la estadística, la inteligencia artificial (IA), el análisis de datos,... con el objetivo de extraer información de valor de esos datos. La ciencia de datos, abarca las siguientes etapas: recolección de los datos, limpieza, análisis exploratorio, construcción y validación de modelos y predicciones.

Una parte importante de la ciencia de datos es el Aprendizaje Automático o Machine Learning (ML). Se trata de un subcampo dentro de la ciencia de datos, concretamente, una subcategoría de la inteligencia artificial. Está basada en algoritmos, y consiste en que éstos descubran de manera autónoma patrones recurrentes del conjunto de datos. Los algoritmos de ML al detectar patrones en los datos, aprenden y mejoran el rendimiento en la ejecución de una tarea o al hacer predicciones. Una vez entrenado y validado el modelo, el algoritmo podrá encontrar patrones en nuevos datos (predicciones)

Para la correcta explicación de las técnicas que se van a describir, es necesario la definición del aprendizaje estadístico supervisado.

El aprendizaje estadístico supervisado es una de las principales herramientas del aprendizaje automático y consiste en una serie de técnicas para deducir una función a partir de una serie de datos de entrenamiento. El objetivo es crear o estimar una función capaz de predecir el valor deseado después de haber visto una serie de ejemplos. Para ello, tiene que generalizar a partir de los datos presentados anteriormente a las nuevas situaciones no vistas previamente. La salida de la función puede ser un valor numérico (como en problemas de regresión) o una etiqueta de clase (como en los de clasificación).

4.2. Etapas del proceso de ciencia de datos

A continuación se van a exponer las diferentes etapas que es necesario completar para la correcta realización de un proceso de ciencia de datos (DSP; Data Science Process)

4.2.1. Conocimiento del negocio (Knowledge of Bussiness)

En esta primera etapa, es fundamental la definición del problema que nos ocupa, la definición de unos objetivos claros y la metodología para cumplirlos.

Esto implica la comprensión de los requisitos del proyecto desde el punto de vista de negocio, utilizando las perspectivas de negocio para determinar a que problemas podemos dar respuesta mediante el uso de la minería de datos.

4.2.2. Adquisición de los datos (Collect the data)

Explicación de los datos, fuente, explicación de las variables,...

Consiste en explicar como se ha llevado a cabo la adquisición de los datos, la identificación de las distintas fuentes y la explicación de los mismos.

4.2.3. Preparación de los datos (data preparation)

Raramente encontraremos los datos preparados para su análisis, ya que normalmente es necesario la limpieza y la transformación de los mismos. Para ello, es necesario llevar a cabo un paso previo llamado pre-procesamiento de los datos.

Fases de la preparación de los datos (data cleaning):

- Eliminación de duplicados (filas y columnas)
- Datos erróneos (ej: precios negativos)
- Detección de valores faltantes: decidir si eliminar esos registros o imputarlos
- Detección de outliers (decidir si mantener, quitar o tratar a parte)
- Unificación de variables (unificación de unidades,...)
- Creación de variables a partir de otras ya existentes si fuera necesario

Preparación de los datos

- Reformateo de variables, por ejemplo, formatos horarios.
- Categorización,...
- Selección de variables (Feature selection): elegir las mejores variables que alimenten nuestros algoritmos dictarán la máxima calidad que podemos conseguir, ya que no todas las variables explican el problema que queremos modelar. Podemos resumir esto con la siguiente frase: "Garbage in, garbage out", es decir, si entra basura saldrá basura. Refiriendonos con basura a ruido en los datos o información pobre.

4.2.4. Análisis exploratorio de datos (EDA)

El análisis exploratorio se utiliza para ver lo que nos pueden ofrecer los datos antes de la etapa del modelado y se lleva a cabo para resumir las principales características del conjunto de datos a través de diferentes tareas:

- Estudio descriptivo de los datos: La estadística descriptiva es la parte de la estadística dedicada a la ordenación y tratamiento de la información por medio de gráficas y tablas, además de la obtención de parámetros útiles para explicar la información
- Visualizaciones de los datos:

  - Análisis univariante: Empleado para observar diferentes características de interés, tratar de identificar patrones en los datos o ver la distribución de las variables . Algunos ejemplos serían los gráficos de caja y bigote o histogramas
  - Análisis multivariante: Donde tratamos de ver la asociación o relación que pueden tener las distintas variables de interés. Encontramos los gráficos de barras o gráficos de dispersión entre los ejemplos de representaciones multivariantes.
- Relación entre las variables

Este tipo de análisis permite obtener medidas descriptivas de un conjunto de datos para poder extraer conclusiones referentes a una muestra o población.

4.2.5. Modelado

En la etapa de modelado aplicaremos algoritmos de aprendizaje automático. Para llevar a cabo esta fase y con el objetivo de obtener mayor robustez en los modelos, aplicaremos la técnica conocida como *hold out*.

El *hold out* es una técnica en la que dividimos los datos en dos partes mutuamente excluyentes (no superpuestas), utilizando una de ellas para el entrenamiento de los modelos y la otra para el testeo.

La traducción literal para hold-out es *retención* y esta técnica recibe este nombre porque reservamos una parte de los datos para probar el modelo en datos nuevos.

Esta técnica se emplea para evitar el sobreajuste. Este aparece cuando un modelo que se adapta perfectamente a los datos de entrenamiento obteniendo unas métricas muy buenas pero que luego es incapaz de generalizar con datos nuevos, y por tanto, existe una sobrevaloración de la capacidad predictiva de los modelos obtenidos.

Por tanto, dividimos los datos en el conjunto de datos de entrenamiento, validación y testeo. Para ello, generamos un conjunto de entrenamiento y otro de testeo a partir del conjunto de datos muestral. A continuación, volvemos a dividir los datos de entrenamiento en datos de entrenamiento y validación, obteniendo así tres conjuntos de datos: entrenamiento, validación y testeo.

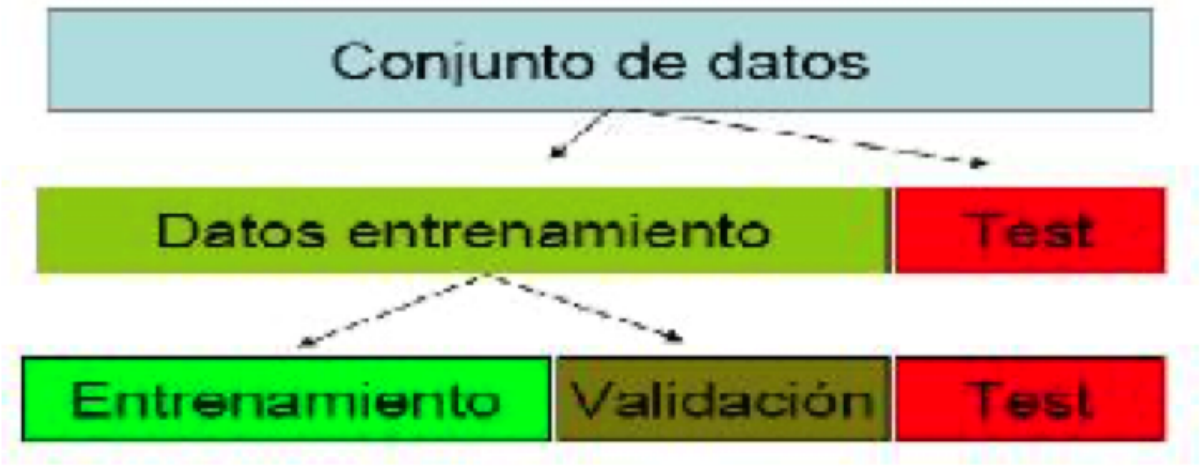

Figura 4.1: División de datos muestrales para entrenamiento, validación y testeo

El conjunto de *datos de entrenamiento* es aquel que utilizamos para probar diferentes hiperparametrizaciones de cada modelo para ver cual es la más óptima. La hiperparametrización variará en función de los parámetros aplicables a cada algoritmo utilizado.

Una vez hayamos entrenado los modelos, pasamos a la fase de validación, donde aplicaremos a los *datos de validación* los diferentes algoritmos con la configuración de parámetros que mejor haya funcionado en el conjunto de datos de entrenamiento.

El modelo con el que obtengamos las mejores métricas será el que posteriormente apliquemos a los *datos de testeo*, ofreciéndonos el error real cometido con el modelo seleccionado. Es decir, este último conjunto de datos se utiliza para estiamr el error de generalización del modelo, ya que nuestro objetivo es obtener un error de generalización pequeño evitando el sobreajuste.

A continuación vamos a exponer tres algoritmos de aprendizaje automático que posteriormente aplicaremos a nuestros datos de ventas de productos.

4.2.5.1. Máquinas de vector soporte (Support Vector Machines SVMs)

Las máquinas de vector soporte son un conjunto de algoritmos de aprendizaje estadístico supervisado pertenecientes a la familia de los clasificadores lineales. Este algoritmo, más conocido como SVM fue desarrollado en los laboratorios AT&T Bell por Vapnik y otros autores a mediados del 1960, inicialmente para problemas de clasificación binaria, basados en la idea de separar los datos mediante hiperplanos. Actualmente existen extensiones dentro de esta metodología pra clasificación con más de dos categorías, para regresión y también para la detección de datos atípicos. La idea fundamental es la utilización de vectores que hacen de soporte con el fin de maximizar la separación entre los datos y el hiperplano.

Suponiendo que tenemos ejemplos de sólo dos categorías y sin pérdida de generalidad, una SVM construye un hiperplano en un espacio de dimensionalidad muy alta. Este hiperplano separa de forma óptima los puntos de una clase de la otra. La característica fundamental de estos algoritmos es el concepto de "separación óptima", ya que se busca el hiperplano que tenga la máxima distancia con los puntos que estén más cerca de él mismo al tiempo que clasifica correctamente tantos puntos de entrenamiento como sea posible. Los algoritmos SVM representan el hiperplano óptimo con vectores de soporte.

En nuestro caso al ser la variable volumen de ventas una variable numérica, vamos a centrarnos en la variante SVM para regresión, tambien conocida como SVR (support vector regressor). El caso del problema de regresión es una generalización del problema de clasificación, en la que el modelo devuelve un valor continuo, es decir, un modelo de regresión estima una función multivariante de valor continuo.

4.2.5.1.1. Descripción del algoritmo

Dado un conjunto de ejemplos de entrenamiento $S = \Big\{(x_1, y_1), \ldots, (x_n, y_n)\Big\}$, donde $x_i \in \mathbb{R}^d$ e $y_i \in \mathbb{R}$, en el que se asume que todos los valores y_i de todos los ejemplos de S pueden ser ajustados mediante un hiperplano, nuestro objetivo será encontrar los parámetros $w = (w_1, \ldots, w_d)$ que permitan definir el hiperplano de regresión

$$y = f(x) = (w_1 x_1 + \cdots + w_d x_d) + b = \langle w, x \rangle + b \ , \ b \in \mathbb{R}$$

La generalización de SVM a SVR se logra introduciendo una región insensible a ϵ alrededor de la función. Esta región se conoce como tubo épsilon. Este tubo reformula el problema de optimización para encontrar el tubo que mejor se aproxime a la función al tiempo que equilibra el error de predicción, es decir, se formula un problema de optimización definiendo una función de pérdida a minimizar insensible a ϵ y encontrando el tubo más plano que contiene a la mayoría de instancias de entrenamiento.

Se dice **rudio, perturbación aleatoria o tubo épsilon** y se representa por $\epsilon \sim N(0, \sigma^2)$, al error en la medición del valor y, por tanto, $y = f(x) + \epsilon$

El valor de ϵ determina el ancho del tubo, y un valor más pequeño indica menor tolerancia al error, cuando más pequeño sea el valor de ϵ, el límite del tubo se desplaza hacia dentro, habiendo más puntos de datos alrededor del límite, lo que indica más vectores de soporte.

Se define la **función de pérdida lineal** $\epsilon-$ insensible, y se representa como L_ϵ a una función lineal en el que la función de pérdida toma valor nulo y viene definida de la siguiente forma:

$$L_\epsilon = \begin{cases} 0 & \text{si} |y - f(x)| \leq \epsilon, \\ |y - f(x)| - \epsilon & \text{en caso contrario} \end{cases}$$

Por tanto, el problema fué planteado por Vapnik como el siguiente probema de optimización:

$$Min_{w,b} \frac{1}{2} ||w||^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

$$s.a = \begin{cases} y_i - w \cdot x_i - b & \leq \epsilon, + \xi_i \\ w \cdot x_i + b - y_i & \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq 0 \forall i \end{cases}$$

Cuando el error es menor que ϵ, las variables de holgura valen 0. Para resolverlo, podemos recurrir al problema dural y al uso de funciones base para trabajar en espacios de mayor dimensión.

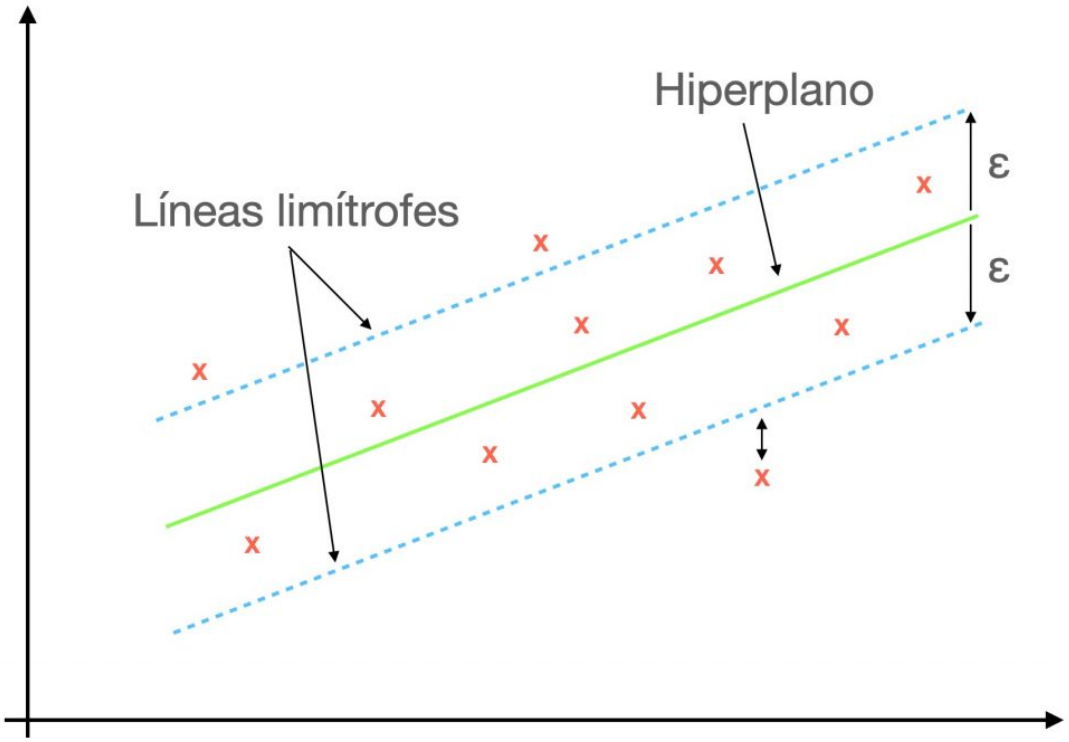

Figura 4.2: Vectores de Soporte de Regresión.

4.2.5.2. K-Nearest Neighbor Regression (KNN)

El algoritmo de K-vecinos más cercanos, más conocido como KNN, fué desarrollado en el año 1951 por los matemáticos Evelyn Fix y Andrew Hodges.

El algoritmo KNN es un método de aprendizaje supervisado que está basado en criterios de vecindad, por lo que es necesario establecer cierta medida de distancia entre los diferentes elementos de la representación. La ventaja de la aplicación de técnicas basadas en la vecindad es la siguiente: el valor de salida que se otorgará a una nueva instancia se calculará en función de los valores de los puntos más cercanos a ella. Se trata de un método local, que asume que la salida de un nuevo dato depende exclusivamente de los k vecinos de entrenamiento más próximos.

4.2.5.2.1. Descripción del algoritmo

Este algoritmo puede ser utilizado para modelos de clasificación y de regresión, ocupándonos en este trabajo la segunda opción. En el caso de la clasificación, se determinará la clase a la que pertenecerá la nueva instancia en función de la clase mayoritaria de los vecinos más cercanos del conjunto de entrenamiento; y en regresión, el modelo debe determinar el valor del nuevo dato como el valor medio de los k ejemplos de entrenamiento más cercanos, siguiendo la siguiente ecuación del valor de la nueva instancia de entrada:

$$Valor(Inst_{\text{entrada}}) = \frac{1}{K}\sum_{i=1}^{k} Valor(P_i)$$

Como ya habíamos avanzado antes, para determinar como de cercanos se encuentran unas instancias de otras, es necesario definir una medida de similitud o distancia para todos los datos del conjunto muestral. Definiremos esta medida de similitud a través de una función, como puede ser la distancia Manhattan, la distancia Minkow o las más utilizada, la distancia Euclídea, que es la que se va a utilizar, y viene dada por:

$$d(p,1) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Una vez definida esta medida, procedemos a la descripción del algoritmo:

- Se almacena el conjunto de datos de entrenamiento compuesto por un vector de entrada y otro de salida
- Se establece el valor del parámetro k
- Se presenta una nueva instancia j teniendo en cuenta únicamente el vector de entrada de esta nueva instancia
  - Se calcula la distancia euclídea de la nueva instancia con todos lso datos del conjunto de entrenamiento
  - Se calcula la salida del este nuevo dato como la media de las salidas de los k datos más cercanos a él
- Se repite el paso anterior para todas las instancias del conjunto de datos

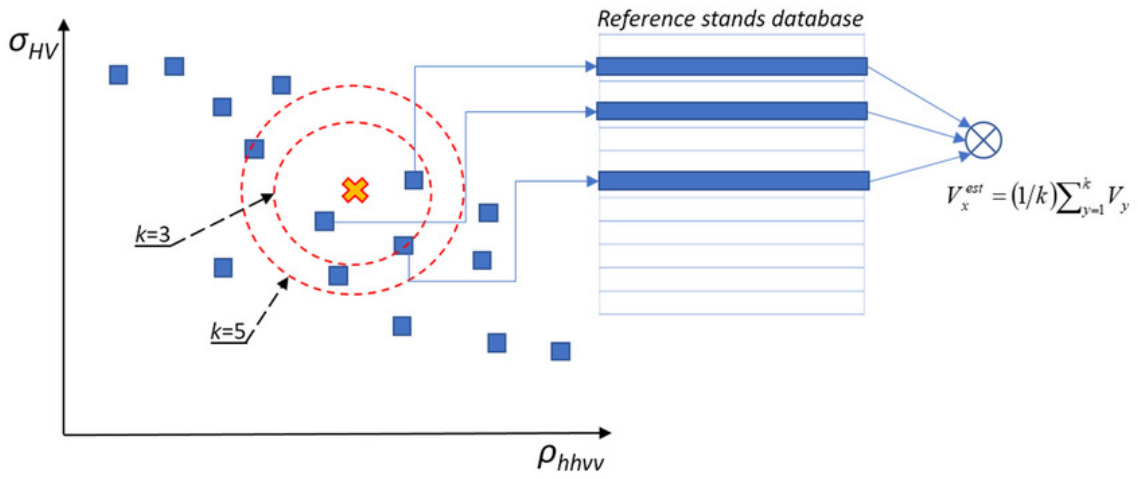

Figura 4.3: Esquema conceptual del algoritmo KNN de regresión.

En la figura 4.3 podemos ver un esquema conceptual del algoritmo, donde el valor de salida que el modelo le dará al nuevo punto marcado con una x será la media de los valores de los puntos vecinos. Habiéndose seleccionado a modo de ejemplo el valor de k=3 y k=5.

4.2.5.2.2. Elección del parámetro k

Es necesario seleccionar el valor que se le va a dar al parámetro k, es decir, el número de vecinos con los que se realizará la media para obtener el valor de salida de la nueva instancia. Si este valor es muy grande, la idea de vecinos que están lejos podrían influir con

la nueva instancia sin tener relación, pero si este valor es demasiado pequeño, el algoritmo será muy sensible a valores extremos.

4.2.5.3. Árboles de decisión (XGBoost Model)

Los árboles de regresión conducen a dividir o segmentar el espacio predictor en regiones más simples de tal forma que la predicción de una instancia se hará a través de la media (o moda) de la región a la que pertenece.

En el caso de los árboles de decisión, el conjunto de reglas empleadas para la segmencación del espacio de predicción se puede resumir en un árbol.

El análisis de árboles de clasificación y regresión, generalmente consiste en tres fases:

- **Construcción del árbol máximo**: este árbol se cosntruye empleando un procedimiento de partición binaria, comenzando en la raíz del árbol. Este primer árbol describe el conjunto de entrenamiento y contiene gran cantidad de niveles (sobreajuste) y nodos, pudiendo ser demasiado complejo. Cada grupo es categorizado por la media (regresión) o la distribución (clasificación) de la variable respuesta, el tamaño del nodo y los valores de las variables explicativas que lo definen.

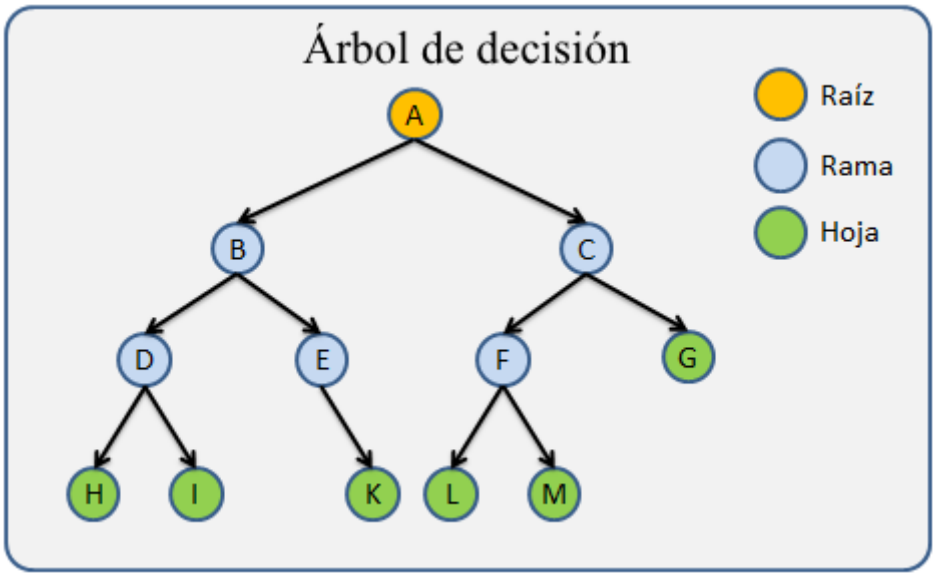

Figura 4.4: Árbol de decisión. Partes

En la figura anterior, podemos ver la representación de un árbol de decisión desde el nodo raíz. Entre sus componentes encontramos: *las ramas*, que son los segmentos del árbol que conectan los nodos; *los nodos internos*, puntos a lo largo del árbol donde se va dividiendo el espacio predictor y *los hojas* (nodos terminales).

- **Poda del árbol**: El árbol máximo está generalmente sobreajustado y por tanto, un árbol más pequeño con menos divisiones podría conducir a una menor varianza. Por este motivo, procede a la poda de éste cortando ramas hasta encontrar el tamaño "adecuado" del árbol.

Una forma de resolver el problema es generar una serie de árboles anidados (árboles de secuencia anidada) de tamaños decrecientes, seleccionando para cada tamaño el mejor de

todos. Posteriormente, se comparan para determinar el óptimo mediante el criterio de coste-complejidad.

Para cada árbol T, se define la función de costo-complejidad, y se representa como $R_\alpha(T)$, como: $R_\alpha(T) = R(T) + \alpha|\widetilde{T}|$,

donde

- R(T) indica el promedio de la suma de cuadrados entre los nodos
- $|\widetilde{T}|$ indica la complejidad del árbol, que se define como el número total de nodos terminales
- α es el parámetro de complejidad, valores altos de este parámetro indican árboles pequeños
- **Selección del árbol óptimo mediante validación cruzada**: El objetivo es seleccionar uno de los árboles de todos los árboles podados como el árbol óptimo, que será el árbol solución. El método de selección consistirá en asociar una medida de error a cada árbol y elegir el que tenga asociado un menor error.

El parámetro de complejidad definido es el que controla la compensación entre la complejidad (tamaño) del árbol y el ajuste a los datos de entrenamiento. Cuando $\alpha = 0$, el subarbol T es el árbol máximo. Sin embargo, a medida que aumenta α, las ramas del árbol se podan de forma anidada, siendo sencillo conseguir una secuencia completa de subárboles en función del valor del parámetro de complejidad. Podemos seleccionar un valor de α a través de un conjunto de validación cruzada.

Para la descripción del algoritmo *Extreme Gradient Boosting*, también conocido como XGBoost, es necesario la definición de varios conceptos, los *bosques aleatorios* y el método de aprendizaje estadístico *Boosting*.

Los bosques aleatorios (*Random Forest*) son una extensión de los árboles de clasificación. El modelo Random Forest es una técnica utilizada tanto para clasificación como para regresión basada en un conjunto de árboles de decisión. Este método selecciona submuestras del conjunto de datos inicial, asegurando así el uso de todas las variables y datos para construir el modelo, haciéndolo además idóneo para trabajar con grandes conjuntos de datos.

El método *boosting* tiene como propósito la reducción del sesgo. Se trata de un proceso iterativo que en lugar de ajustar un árbol de decisión, aplica la técnicas repetidas veces de forma secuencial, y por tanto, el algoritmo va aprendiendo lentamente. Este método no se aplica sobre los datos, sino sobre los resíduos.

Dado un árbol que ha sido previamente ajustado, se aplica un nuevo árbol para los resíduos del modelo, permitiendo así el reajuste del modelo. Este nuevo árbol de decisión construido con los resíduos se añade dentro de la función ajustada con el fin de mejorar el algoritmo en cada iteracción y no de forma global.

Para aplicar esta técnica, es necesario fijar una serie de parámetros:

- Tamaño del árbol **d**, es decir, el número de nodos terminales
- El número de árboles **B**. Nota: un valor muy alto podría llevar a sobreajuste
- El parámetro de regularización λ, que puede ser interpretado como una proporción de aprendizaje, es decir, la velocidad a la que aprende el algoritmo. Se trata de un parámetro acotado entre 0 y 1, siendo habitual elegir $\lambda = 0.01$ o $\lambda = 0.001$.

Este tipo de técnicas es aplicable tanto a problemas de regresión como de clasificación, sin embargo, nosotros nos centraremos en las de regresión, especialmente en el modelo *Extreme Gradient Boosting Algorithm*, tambien conocido como XGBoost.

Este algoritmo se encuentra dentro del marco de algoritmos de aprendizaje supervisado, y fué propuesto en el año 2016 por Chen y Guestrin y presenta las siguientes características:

- Consiste en la agregación de árboles de manera secuencial con el objetivo de aprender el resultado de los árboles previos y corregir el error producido por éstos, hasta que no se pueda reducir más el error (gradiente descendente)
- Para evitar el sobreajuste, realia un procesamiento en paralelo, la poda de árboles, el control de los valores perdidos y la optimización que penaliza la complejidad de los modelos.

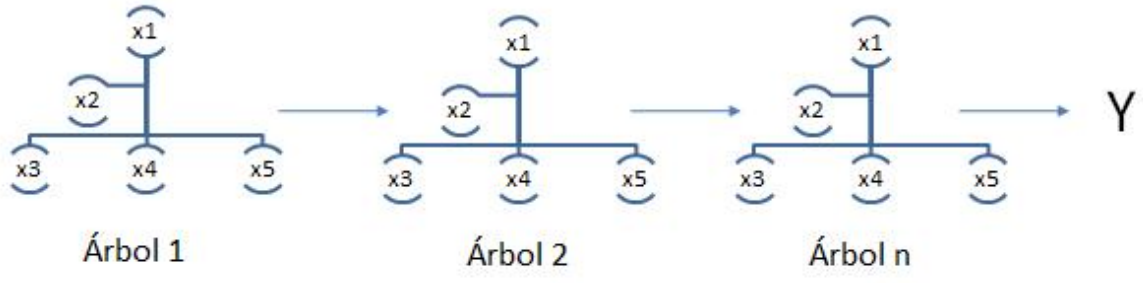

Figura 4.5: Algoritmo XGBoost

4.2.5.3.1. Descripción del algoritmo

EL funcionamiento del algoritmo se puede resumir en cuatro pasos:

Dada una miestra inicial de aprendizaje $\{(x_i, y_i) : i = 1, \ldots, n\}$.

- En primer lugar, se obtiene un árbol inicial con d divisiones, T_o para predecir la variable respuesta Y, asociando el resultado a un valor residual r_i
- Se obtiene un nuevo árbol R que se ajusta al error del paso anterior
- Los resultados de los árboles T_o y R se combinan para obtener un árbol T_1, donde el error cuadrático medio será menor que el del árbol inicial
- Se continúa el proceso de forma iterativa hasta que el error es minimizado lo máximo posible

4.2.5.4. Evaluación y presentación de resultados (+análisis del error)

- Predicciones con el mejor modelo
- Final de la historia de una forma ordenada y resumida
- Señalar posibles mejoras y recomendaciones para proyectos futuros

Capítulo 5

Caso práctico con datos reales

Conclusiones

Conclusiones de mi trabajo

\nocite{Luque2017,Luque2019,DataMiningIntro_TFG,WikiIntroWalmart,MestaTFG,CampoTFG,Def
ST_Art,PEREZ-MARTINEZ2021_CV,IntroFeatureSelection, EDA,SVM_CampoTFG,
SVM_SVR, ECII_AP, KNN_R_TFG,XGBoost_TFG,XGBoost_Casal,XGBoost_CasalII,XGBoostAlg