Esta es la introducción de mi trabajo y se irá modificando conforme avance.

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, una librería, son claros ejemplos de lo que es el sector retail.)

Objetivo general

Los objetivos principales del trabajo consistirán en analizar y entender como varía la demanda de productos en función del tiempo así como la aplicación de diversos algoritmos de aprendizaje automático para modelarla.

Para ello, se expondrán de manera teórica una serie de técnicas para la modelización de la variable número de ventas para posteriormente aplicarlas a un caso real de ventas de productos lácteos, con el propósito final de poder predecir las ventas de cara al futuro.

Dentro de las técnicas estudiadas, 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 este motivo, 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 estaístico, 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 ello, se ha estructurado el trabajo en diferentes secciones. En primer lugar, una primera sección introductoria sobre el proceso de la ciencia de datos en el sector retail, a continuación, una segunda sección para la explicación de lo que se conoce como análisis de cesta de la compra. Las secciones tres y cuatro explican las técnicas clásicas de modelado estadístico y el proceso de ciencia de datos completo con las correspondoentes técnicas de aprendizaje automático, respectivamente. Una quinta sección donde se aplicarán todas las técnicas estudiadas a un caso práctico con datos reales mediante Rstudio, un entorno de desarrollo integrado para el lenguaje de programación R. Por último, en la sección sexta se exponen las conclusiones extraidas 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.