Universidad de Cádiz
Escuela Superior de Ingeniería

Aprendizaje Computacional

Curso 2020/2021

# Práctica Final - Supercai

AUTOR 1: Andrés Moreno Bernal
AUTOR 2: Álvaro Torné Zambrano
AUTOR 3: Allae Ghayat Zayoun
AUTOR 4: Pablo Ortiz Gutiérrez

Puerto Real, 2020

# Índice

# 1. Propuesta de servicios

El objetivo de nuestro estudio es ofrecer un servicio que, mediante el uso de redes neuronales, pueda hacer una estimación, lo más exacta posible, del consumo eléctrico de un supermercado o centro comercial. Dicho estudio podrá ser extrapolado a cualquier área en la cual se trabaje en una superficie con necesidad de consumo eléctrico.

Este trabajo ha sido elegido debido a que en estos locales, los supermercados, se tiene que pagar cada cierto tiempo por el consumo de electricidad que se hará por adelantado. Por tanto, hay casos, en los que se estima una cantidad mayor a la que realmente hará falta, malgastando de esta forma tanto los recursos como el capital, el cual podría haber sido empleado en cualquier otra área.

En el lado opuesto, aunque apenas ocurre, encontramos el caso de una estimación menor a la que va a hacer falta, por lo que se podrían estropear productos y tener que desecharlos. Esto lleva otra vez a la pérdida de dinero.

Mediante nuestro servicio se podrán hacer predicciones de lo más exacta posible, con el fin de reducir al máximo la pérdida de dinero que este desfase genera.

Aparte de lo ya explicado anteriormente, el estudio realizado nos va a permitir ofrecer otro servicio distinto pero igualmente beneficioso, destinado esta vez al área de marketing. Mediante datos sobre la edad, los ingresos anuales y el gasto en el supermercado o centro comercial en cuestión de los clientes, podremos realizar mediante técnicas de clustering una clasificación de estos con el fin de promover campañas de marketing especializadas para cada tipo de cliente.

Los clientes se dividirán en 6 grupos diferentes, cada uno con unas características específicas que ayudará al departamento de marketing a tener una segmentación de sus clientes, ver que grupo predominan y poder ver de que grupo son cada uno de los clientes que nos ha aportado para hacer el estudio.

De esta forma, se pretende dar la opción de mejorar la experiencia al consumidor a la hora de buscar los productos que mejor se adapten a ellos. Al tener que dar los datos dichos previamente esta medida es opcional para el cliente, pudiéndose darse de baja en cualquier momento. De esta forma se le da seguridad al consumidor sobre la protección de sus datos e intimidad a la hora de realizar compras.

A nuestro cliente, gerente del Supermercado SuperCai también se le entregará un estudio sobre la edad media de los clientes, porcentaje según géneros, la relación entre los datos que nos ha aportado de cada cliente...

## 2. Infraestructura requerida

Para la primera parte, la enfocada en el ahorro del consumo eléctrico no es necesaria ninguna infraestructura especial aparte de sensores que midan la temperatura de la zona, así como medidores para el viento y la presión del ambiente.

En cuanto al segundo servicio que ofrecemos solo se necesitará una página web en la que los clientes puedan suscribirse o darse de baja. Este sitio web puede ser sustituido por una simple base de datos gestionada por un empleado, que será el encargado de las solicitudes de los consumidores, aunque quitaría algo de transparencia al proceso, al no ser el mismo cliente el que se encarga.

## 3. Cronograma del desarrollo

### 3.1. Fases Predicciones de consumo eléctrico

En este proyecto hemos pasado por numerosas fases. Al principio hemos empezado la implementación con el lenguaje de Matlab pero debido a ciertos problemas que nos acarreaba las predicciones con la red decidimos probar la implementación con Python, ya que este es un lenguaje más versátil y con multitud de librerías a mano del usuario para su uso.

Ya con Python también tuvimos demasiados problemas con el preprocesado de la base de datos ya que nos daba predicciones demasiado altas a las que debería. Se intentó todo tipo de preprocesado y validación para solucionar el problema, como por ejemplo: entrenar con menos características de las necesarias, entrenar con 80 % de los datos y testear con el resto...

Gran parte del tiempo consumido para el desarrollo del trabajo ha sido para solucionar este problema ya que hemos estado trabajando con una base de datos real de un supermercado y es a lo que nos enfrentaremos en la vida fuera de la Universidad.

Finalmente nos dimos cuenta que la base de datos con la que entrenábamos tenía 23 días cada mes, por lo que cuando entrenábamos por ejemplo con los primeros 10 días, cogíamos 10 días de la base de datos para entrenar y cogíamos 10 datos de la base de datos para testear, pero esta base de datos tenía los días del 24 al 31 mas 4 días del siguiente mes por lo que la validación en este caso no tenía mucho sentido.

Así que para comprobar el funcionamiento de las predicciones hemos cogido los consumos de los días 23 y 24 del primer mes y hemos predicho el consumo para el día 24 para compararlo con el consumo del día 23 que ya lo teníamos en el fichero de entrenamiento. Aún así, nos hemos dado cuenta de que las predicciones seguían desvariando un poco ya que en este caso concreto, teníamos

tomado como características la temperatura, la humedad y el viento, este último era muy alto con respecto al del día 23 y este diferencia tan alta era la que hacía que no nos diese predicciones tan exactas. Excluyendo estos casos concretos, los resultados de la red son bastante aceptable para realizar las predicciones del consumo mensual de un supermercado.

## 3.2. Base de datos usada

La base de datos son las que están en los ficheros train.csv y test.csv . Como bien su nombre indica, una base de datos es usada para entrenar y la otra para testear y asi poder realizar las predicciones. En la base de datos train.csv tenemos in ID, un campo datatime compuesto por la fecha y la hora, este campo se deberá de dividir posteriormente en un preprocesador para dividir en campos individuales(dia, mes, año, hora, minuto...), también tiene un campo de presión, temperatura, velocidad del viento, y consumo eléctrico. Hay dos campos más que han sido omitidos llamados *var1 y var2* que no sabemos que indican, por lo tanto no los hemos usado y de todas formas no eran importante para el entrenamiento de la red ya que las características importantes son la fecha, la temperatura, la presión y la velocidad del viento.

## 3.3. Metodología y funcionamiento

En primer lugar realizamos un preprocesado de los datos para extraer las características, escalarlas y normalizar los datos para que nuestra red neuronal de la que hablaremos posteriormente pueda trabajar correctamente con ellos.
Hemos implementado un modelo de regresión para predecir los consumos de electricidad a partir de todas las características que tenemos. Para ello usamos una red neuronal Deep Feed-Forward con una capa de entrada de 11 neuronas (11 características tiene el input) 2 capas ocultas de 11 y 5 neuronas, ambas con función de activación relu, mientras que en la capa de salida tenemos solo una neurona y con función de activación sigmoide. El número de epochs ha sido establecido a 100 y el batch size (número de filas de datos consideradas antes de actualizar los pesos en cada epoch) a 10 pues así lo hemos visto conveniente en el aspecto duración del entrenamiento-calidad de la predicción tras un proceso de pruebas.

La red ha sido implementada mediante la clase Sequential y capas Dense de la libreria Keras, las cuales pueden modelar cualquier función matemática y son muy comunes de utilizar.
Para el entrenamiento usamos las características de los primeros 23 días de cada mes y hemos hecho un K Cross Validation con K=3 para que no tarde demasiado tiempo pues recordemos que tenemos 100 epochs y 10 de batch size. Una vez hemos entrenado con los datos de train.csv, usaremos las características de test.csv para predecir y más tarde comparar esas predicciones con los datos de consumo reales que nos vienen en test.csv.

Finalmente observamos de la mejor manera a nuestro alcance, que es mediante gráficas, la calidad de estas predicciones.

## 3.4. Fases segmentación clientes( Clustering)

Esta parte ha sido mucho más fácil de implementar que la red neuronal. La implementación ha sido casi inmediata y no hemos tenido tantos problemas. Prácticamente pudimos terminarla en un día. Su implementación al igual que la red neuronal también se ha realizado en Python.

Se clasificarán los clientes en 6 grupos (que serán los centroides seleccionados) atendiendo a sus características citadas en la siguiente sección de *Base de datos usada*. Una vez clasificados los clientes se les ofertarán productos o se realizarán eventos publicitarios hacia ese cliente para incitarlo a realizar la compra.

Además, se realizarán gráficos mostrando el porcentaje de clientes que hay en cada grupo, que grupo gasta más, etc.

## 3.5. Base de datos usada

En este caso usamos una base de datos diferente ya que esta vez queremos clasificar los clientes en grupos. Esta base de datos contiene 5 campos, Customer ID único para cada cliente , el sexo del cliente, la edad, ingresos anuales del cliente en miles de dólares y por último una puntuación de gasto que va de 1-100.

## 3.6. Metodología

- Estudio de los datos aportados por la base de datos
  - Visualización primeros registros
  - Cantidad de datos aportados
  - Tipo de los campos de cada registro
  - Detección de registros con campos nulos
- Histogramas de cada característica
- Cantidad de clientes según característica
- Relación entre variables (Correlación)
- Distribución de las característica según el género del cliente
- Clustering utilizando diferentes número de características y clusters (Algoritmo K-Means )

## 3.7. Funcionamiento

Se realiza el estudio de los datos aportados por la base de datos. En nuestro caso tenemos una base de datos bastante limpia, en la cual todos los datos son numéricos excepto el género que viene indicado con caracteres ("Male" y "Female"). No tenemos ningún registro con campo nulo. En total contamos con 200 registros , cada uno con 5 características (Id del cliente, Género, Edad, Renta anual (ingresos anuales), y puntuación según cuanto gaste en el supermercado).

Posteriormente, realizamos un estudio sobre los valores de las características observando los valores máximos, mínimos, media y varianza de cada una de las características.

Una vez tenemos clara la distribución y naturaleza de los datos con los que contamos vamos a probar diferentes posibilidades de clustering con estos datos. Para cada una de las posibilidades veremos el valor de la "inertia" usando desde 1 a 15 clusters.

- **Edad y Puntuación de gasto**: Estimamos que según la inertia un valor correcto de clusters para esta combinación son 4. Observamos que se segmenta correctamente visualizando gráficamente la división que se realiza en los datos con 4 clusters.
- **Renta anuales y Puntuación de gasto**: Estimamos que según la inertia un valor correcto de clusters para esta combinación son 5. Observamos que se segmenta correctamente visualizando gráficamente la división que se realiza en los datos con 5 clusters.
- **Edad, Renta anual y Puntuación de gasto**: Estimamos que según la inertia un valor correcto de clusters para esta combinación son 6. Esta combinación es la que decidimos finalmente para presentarle a la empresa ya que creemos que cuenta con una cantidad suficiente de características a tener en cuenta para realizar la segmentación de clientes.

Se ha decidido usar la combinación de Edad, Renta anual y Puntuación de gasto al estimar que es una cantidad suficiente y que hace una buena segmentación. Pero si la empresa decidiera que también quiere tener en cuenta el género de los clientes o dividir los clientes en una cantidad mayor o menor de grupos también se podría realizar. Esto será comunicado a la empresa en la presentación final de nuestra propuesta.

Una vez tenemos realizada la segmentación aportaremos a la empresa a que grupo pertenece cada uno de los clientes que se encuentran en la base de datos que nos han aportado para realizar el estudio, al igual que le indicaremos la cantidad de clientes que pertenecen a cada grupo.

También indicamos a la empresa diferentes posibilidades/estrategias de marketing que podría realizar con los resultados obtenidos, aunque la decisión final de realizar o no estas estrategia será responsabilidad del departamento de marketing de la empresa de nuestro cliente.

# 4. Presupuesto

El presupuesto total del proyecto será la suma del proyecto de las predicciones de electricidad y la segmentación de clientes.

El cliente podrá optar a los dos servicios ofertados, podrá contratar uno de los dos o los dos, como el cliente desee.

## 4.1. Presupuesto Predicción de consumo eléctrico

El presupuesto de este proyecto es variable ya que dependerá de las dimensiones del supermercado. Constará de dos pagos. El primero está orientado a la venta del proyecto en sí, es decir, las predicciones anuales cuyo precio rondará sobre los 200€ y segundo pago será por la instalación del equipamiento requerido para el desarrollo del proyecto.
Este pago no tiene una cuantía fija ya que dependerá de las dimensiones del supermercado, si es más grande necesitará más recursos para la toma de datos. El equipamiento requerido son los siguiente:

- Termómetro profesional.
- Anemómetro (Medidor de velocidad del viento).
- Barómetro (Medidor de presión atmosférica).

El coste del mantenimiento será de 75€ mensuales.

## 4.2. Presupuesto Segmentación de clientes

En este presupuesto no tenemos ninguna instalación ya que el cliente comprará un servicio online, donde lo único que debe de disponer es de una página web. Por lo tanto, solo pagará por el mantenimiento del servicio que tiene un coste mensual de 75€.

# 5. Anexo Gráficas

## 5.1. Predicción electricidad

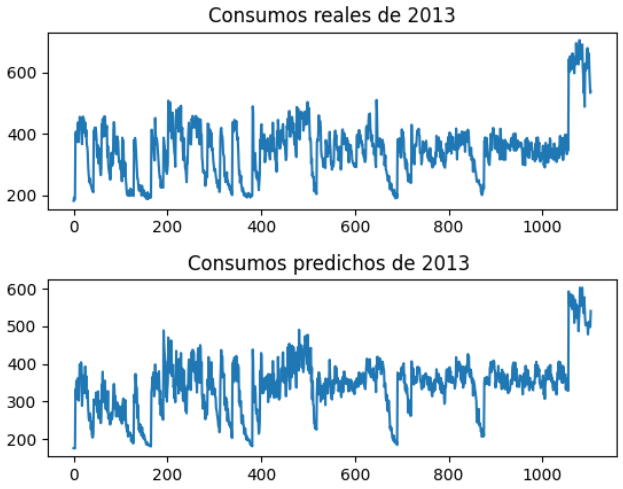


Figura 1: Consumo eléctrico real (arriba) y consumo predicho (abajo) durante 2013

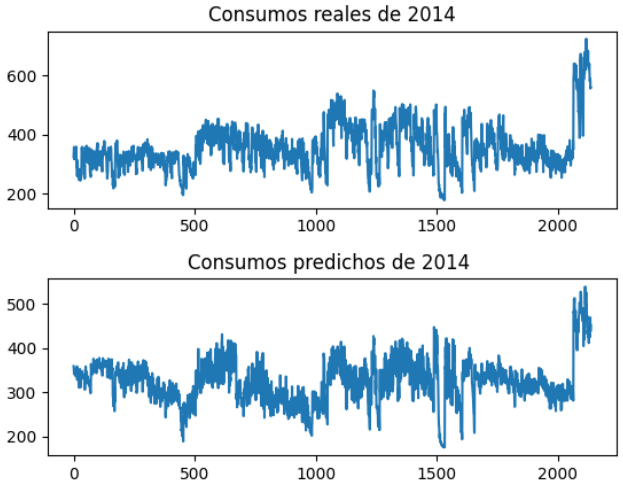


Figura 2: Consumo eléctrico real (arriba) y consumo predicho (abajo) durante 2014

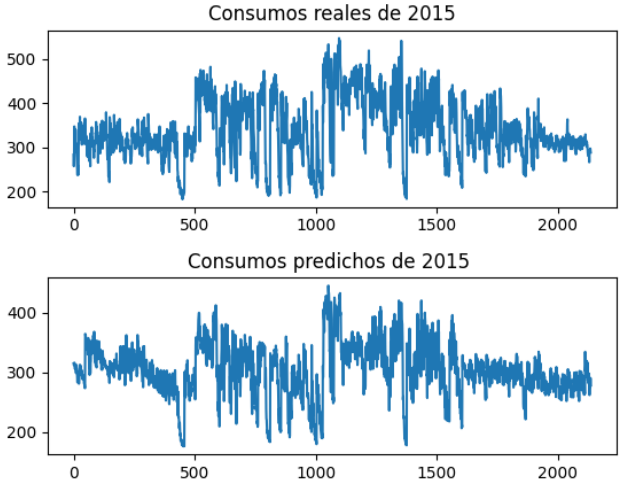


Figura 3: Consumo eléctrico real (arriba) y consumo predicho (abajo) durante 2015

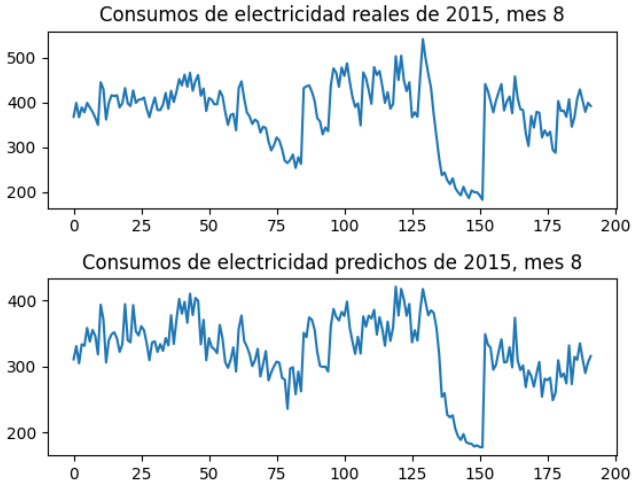


Figura 4: Consumo eléctrico real (arriba) y consumo predicho (abajo) durante Agosto de 2015

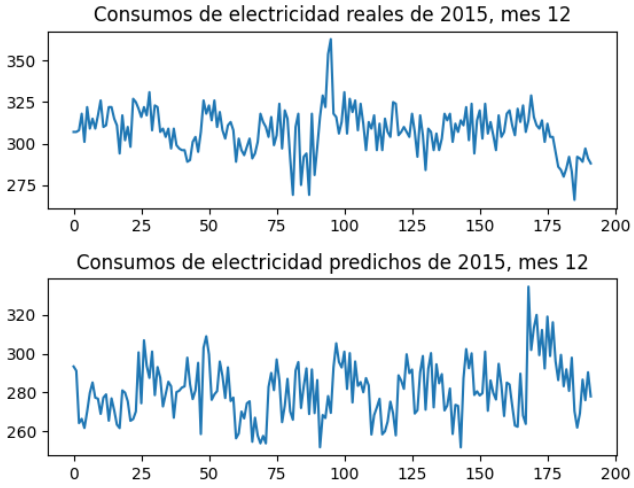


Figura 5: Consumo eléctrico real (arriba) y consumo predicho (abajo) durante Diciembre de 2015

## 5.2. Segmentación clientes

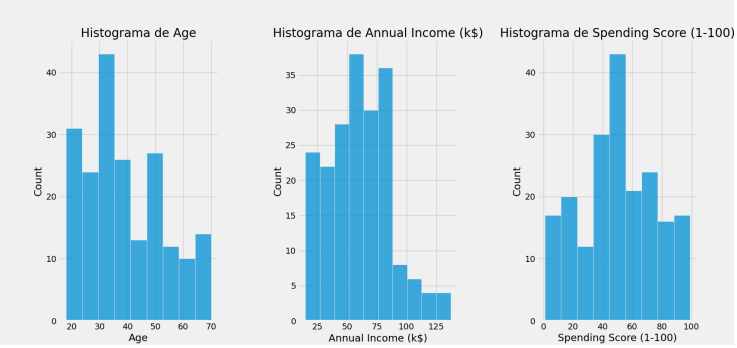


Figura 6: Histogramas de los valores de cada característica de los clientes

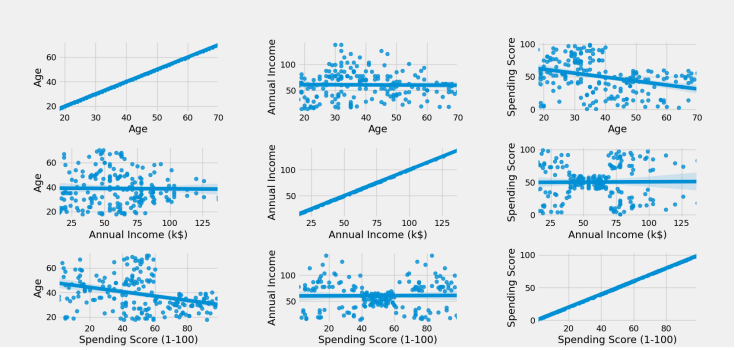


Figura 7: Regresión entre las características. (Correlación)

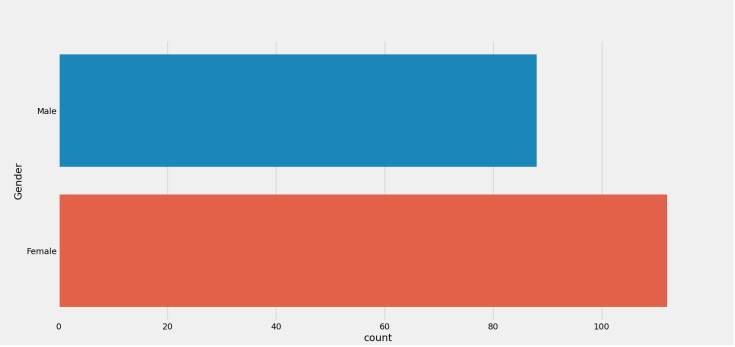


Figura 8: Cantidad de hombres y mujeres en los datos aportados por nuestro cliente

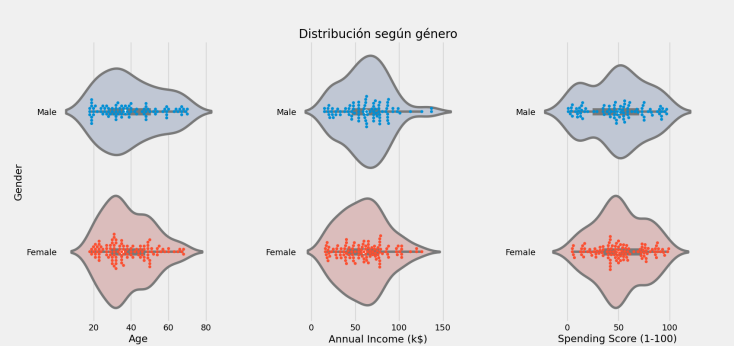


Figura 9: Distribución que siguen las características según sea hombre o mujer

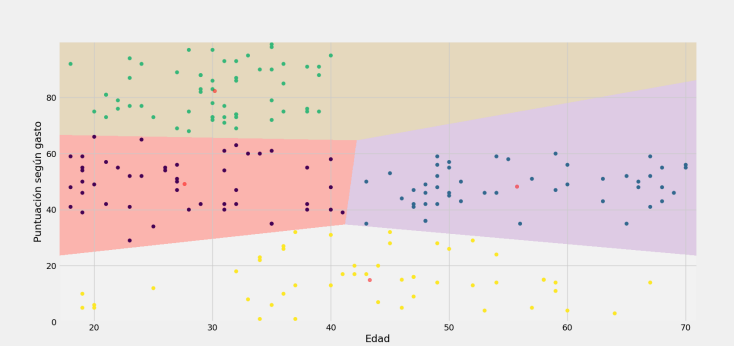


Figura 10: Visualización gráfica del clustering utilizando puntuación del gasto y edad con 4 clusters

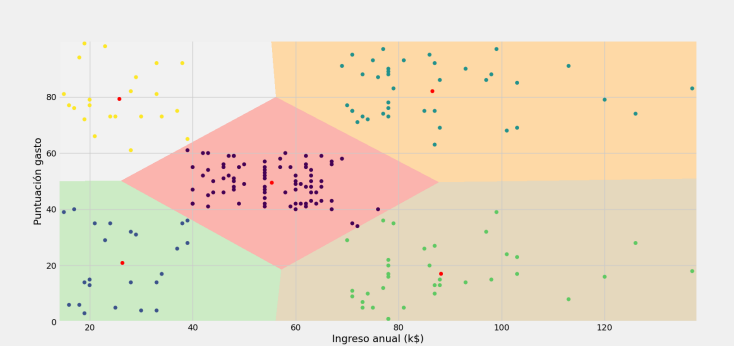


Figura 11: Visualización gráfica del clustering utilizando puntuación del gasto y renta anual con 5 clusters

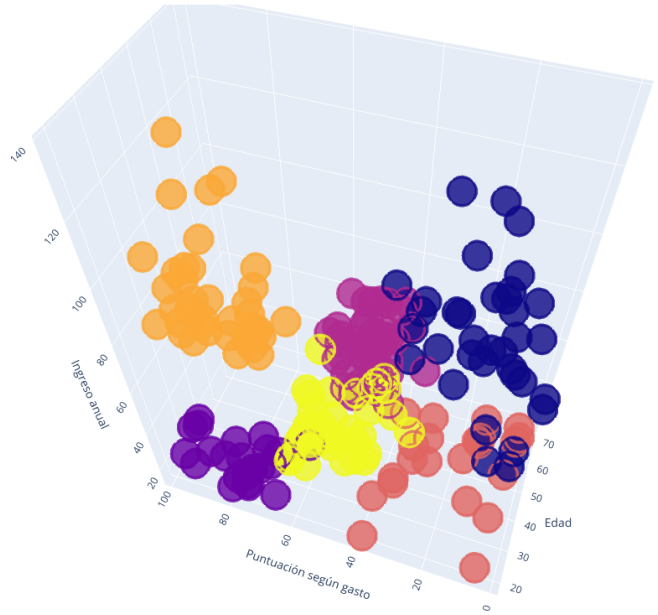


Figura 12: Visualización gráfica del clustering utilizando puntuación del gasto,renta anual y edad con 6 clusters

| | EDAD (18-70) | RENTA (15-137) | GASTO(1-100) |
|---|---|---|---|
| Grupo 1 | 42 | 88 | 17 |
| Grupo 2 | 56 | 53 | 49 |
| Grupo 3 | 25 | 25 | 79 |
| Grupo 4 | 32 | 86 | 82 |
| Grupo 5 | 27 | 56 | 49 |
| Grupo 6 | 44 | 25 | 19 |

Figura 13: Valores representativos de cada característica para cada grupo de la segmentación

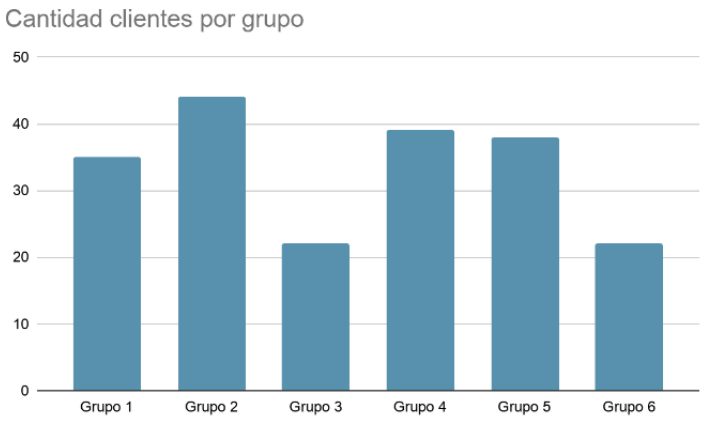


Figura 14: Cantidad de clientes por grupo