Susana Sánchez Ropero, 1494978

APC

1-12-2022

Caso KAGGLE

Imagen que contiene Flecha

Descripción generada automáticamente

Índice

[Introducción 3](#_Toc121766020)

[Base de datos 4](#_Toc121766021)

[Data Cleaning and Pre-Processing 5](#_Toc121766022)

[Data Cleaning 5](#_Toc121766023)

[Valores nulos 5](#_Toc121766024)

[Redundancias 5](#_Toc121766025)

[Pre-Processing 5](#_Toc121766026)

[Matriz de correlación 5](#_Toc121766027)

[Atributo Listing Price 6](#_Toc121766028)

[Exploratory data analysis – EDA 7](#_Toc121766029)

[Descuentos Nike vs Adidas 7](#_Toc121766030)

[Número de artículos en venta Nike vs Adidas 8](#_Toc121766031)

[Nike vs Adidas 8](#_Toc121766032)

[Nike 8](#_Toc121766033)

[Adidas 9](#_Toc121766034)

[Precios Adidas vs Nike 10](#_Toc121766035)

[Valoraciones y reseñas 11](#_Toc121766036)

[Nike 11](#_Toc121766037)

[Adidas 11](#_Toc121766038)

[Conclusiones del estudio Nike vs Adidas 12](#_Toc121766039)

[Clasificación 13](#_Toc121766040)

[Atributo target Sale Price 13](#_Toc121766041)

[KNN 13](#_Toc121766042)

[Logistic Regression 13](#_Toc121766043)

[Random Forest Regressor 14](#_Toc121766044)

[Random Forest Classifier 14](#_Toc121766045)

[Atributo target Brand 15](#_Toc121766046)

[KNN 15](#_Toc121766047)

[Logistic Regression 15](#_Toc121766048)

[Random Forest Regressor 15](#_Toc121766049)

[Random Forest Classifier 16](#_Toc121766050)

[Atributo target Discount 16](#_Toc121766051)

[KNN 16](#_Toc121766052)

[Logistic Regression 16](#_Toc121766053)

[Random Forest Regressor 17](#_Toc121766054)

[Random Forest Classifier 17](#_Toc121766055)

[SVM 17](#_Toc121766056)

[Comparación de modelos 18](#_Toc121766057)

[Curva Precision Recall mejor modelo con Random Forest Classifier 19](#_Toc121766058)

[Curva ROC mejor modelo con Random Forest Classifier 19](#_Toc121766059)

[R2\_score Random Forest Regressor 20](#_Toc121766060)

[Feature importance Random Forest 20](#_Toc121766061)

[PCA 21](#_Toc121766062)

[Conclusiones 21](#_Toc121766063)

# Introducción

En esta tercera y última práctica de la asignatura de Aprendizaje Computacional usaremos una base de datos de la plataforma Kaggle.

Este proyecto constará de tres partes:

1. Una explicación detallada de los atributos más importantes de la base de datos y en especial del atributo objetivo, el que se va a predecir i/o clasificar.
2. Hacer una breve descripción del método de aprendizaje computacional aplicado, junto con los parámetros escogidos.
3. Y por último, una presentación de los resultados obtenidos.

En mi caso la base de datos con la que trabajaré trata sobre Adidas vs Nike. El un debate constante en la industria del deporte. Este conjunto de datos consiste en la información del producto de estas dos grandes empresas con información significativa.

Los datos de productos de Adidas y Nike se pueden utilizar para una serie de propósitos, como la investigación competitiva.

El link a kaggle es: [Adidas vs Nike | Kaggle](https://www.kaggle.com/datasets/kaushiksuresh147/adidas-vs-nike).

# Base de datos

Para empezar a adentrarme en la base de datos, primero analizo todos y cada uno de los atributos:

Pantalla de computadora con letras

Descripción generada automáticamente con confianza media

Ilustración 1: Base de datos

Como se puede observar en la imagen anterior, dispone de 10 atributos y de 3268 entradas.

Los atributos son:

* Product Name: El nombre del producto.
  + Tipo: object.
* Product ID: Un ID único para cada producto.
  + Tipo: object.
* Listing Price: Precio de lista.
  + Tipo: int.
  + Rango: 0 – 29.999
* Sale Price: Precio de venta.
  + Tipo: int.
  + Rango: 449 – 36.500.
* Discount: El descuento que se le puede aplicar al producto.
  + Tipo: int.
  + Rango: 0 – 60.
* Brand: La marca del producto.
  + Tipo: object.
* Description: Una breve descripción del producto.
  + Tipo: object.
* Rating: Valoración del producto.
  + Tipo: float.
  + Rango: 0 – 5.
* Reviews: El número de reseñas que tiene el producto.
  + Tipo: int.
  + Rango: 0 - 223
* Last Visited: La fecha concreta de la última visita que tuvo el producto.
  + Tipo: object.

El atributo objetivo lo estableceremos más adelante, ya que no viene preestablecido.

# Data Cleaning and Pre-Processing

Una vez sé de qué trata la base de datos, que atributos contiene y cuál es el atributo objetivo que voy a predecir/clasificar, procedo a limpiar los datos y a preprocesarlos.

## Data Cleaning

### Valores nulos

Primero de todo, miro si el dataset contiene valores nulos. Y efectivamente tiene, pero únicamente 3 entradas los contienen. Por esta razón he decidido eliminarlas directamente sin hacer ningún tipo de tratamiento especial, como por ejemplo sustituir el valor nulo por la media, ya que considero que no influirá en los resultados, no se pierde información.

### Redundancias

Al indagar en la base de datos, me he dado cuenta de que en el atributo Brand hay dos marcas, en concreto 'Adidas ORIGINALS', 'Adidas ORIGINALS', que se refieren a la misma, es información redundante. Por lo que he decidido agruparlas en una sola marca: 'Adidas ORIGINALS'.

## Pre-Processing

### Matriz de correlación

Gráfico, Gráfico de barras

Descripción generada automáticamente

Ilustración 2: Matriz correlación

Analizando la matriz de correlación, vemos que los atributos con más correlación son Sale Price y Listing Price con 0.31 y Reviews y Discount con 0.31 también.

**Atributo Last Visited**

El atributo Last Visited indica la fecha en la que un cliente ha visitado ese producto. Pero la información está en formato object.

He transformado este atributo object en datetime64 para que más adelante sea más cómodo trabajar con él.

Me he dado cuenta de que este atributo únicamente contiene fechas con 0 días y 00:36:43 de diferencia. Así que hay información redundante.

Para esto, he creado dos atributos nuevos llamados Minute y Second. Así solo tendré que consultar estos datos.

### Atributo Listing Price

Al hacer dataset.describe() y analizarlo, me he dado cuenta que el mínimo de este es de 0. Y no concuerda que el precio sea de 0 euros, debe de ser un error. Por este motivo, lo reemplazo por el precio de venta.

Esto lo hago de la siguiente forma:

*boolean\_condition = dataset['Listing Price'] == 0*

*Column\_Name ='Listing Price'*

*new\_value = dataset['Sale Price']*

*dataset.loc[boolean\_condition, Column\_Name] = new\_value*

# Exploratory data analysis – EDA

En este apartado analizaré la información de la base de datos. Y de esta forma averiguaré, únicamente centrándome en los datos de esta base de datos, que marca prefiere la clientela, si Nike o Adidas.

## Descuentos Nike vs Adidas

Los descuentos son una forma de llamar la atención de los clientes y animarlos a comprar. Es un punto clave para saber que prefiere el comprador.

He mostrado en esta gráfica los descuentos que ofrecen las dos marcas, veamos que obtenemos:

Gráfico, Gráfico circular

Descripción generada automáticamente

Ilustración 3: descuentos Nike vs Adidas

Claramente se ve en el gráfico que Nike no ofrece ningún tipo de descuento. A diferencia de las tres marcas de Adidas que sí lo hacen. En concreto, Adidas CORE / NEO, con un 36,6% de descuentos. Aunque las tres submarcas de Adidas ofrecen alrededor del 30% de descuento.

## Número de artículos en venta Nike vs Adidas

### Nike vs Adidas

Gráfico, Gráfico de barras

Descripción generada automáticamente

Ilustración 4: Nº artículos

Las tres marcas de Adidas tienen a la venta muchos más artículos que Nike, más de la mitad.

Esto puede hacer que el estudio no sea fiable, ya que tenemos las muestras de las diferentes marcas muy desbalanceadas.

### Nike

Nike ofrece la siguiente lista de productos:

Gráfico, Gráfico de barras, Gráfico de cajas y bigotes

Descripción generada automáticamente

Ilustración 5: Productos ofrecidos por Nike

Nike ofrece 5 tipos diferentes de productos.

### Adidas

Adidas ofrece la siguiente lista de productos:

Gráfico, Gráfico de barras

Descripción generada automáticamente

Ilustración 6: Productos ofrecidos por Adidas

Adidas ofrece 5 tipos de productos, igual que Nike.

## Precios Adidas vs Nike

Gráfico, Histograma

Descripción generada automáticamenteVamos a indagar en los precios de las dos marcas, tanto de venta como de lista.

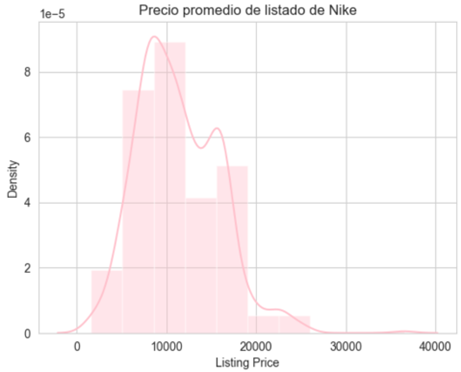


Ilustración 7: Precio listado

### 

Gráfico

Descripción generada automáticamenteGráfico, Gráfico de líneas

Descripción generada automáticamente

Ilustración 8: Precio de venta

Si observamos los gráficos del precio promedio de listado, en el caso de Nike ronda entre los 7.500 y 10.000. En Adidas va de los 5.000 y 7.500.

Si ahora nos fijamos en los de precio de venta, vemos que los precios de Nike rondan los 5.000 – 7.500. Y los de Adidas también.

## Valoraciones y reseñas

### Nike

Pantalla de computadora con letras

Descripción generada automáticamente con confianza mediaLas valoraciones de Nike tienen una media de 2.7 sobre 5, no son muy buenas, los clientes no están satisfechos con los productos que ofrecen.

Ilustración 9: Valoraciones y reseñas Nike

Pero si nos fijamos en la desviación estándar de las reseñas vemos que es bastante significativa, lo que sugiere popularidad.

### Adidas

Pantalla de computadora con letras

Descripción generada automáticamente con confianza mediaSi estudiamos las valoraciones de Adidas vemos que de media son un 3.3 sobre 5, mejor que Nike. Esto podría deberse al hecho de que los productos Adidas son bastante baratos y ofrecen grandes descuentos en sus productos.

Ilustración 10: Valoraciones y reseñas Adidas

La desviación estándar para las reseñas es nuevamente buena en comparación con Nike, la posible razón de esto puede ser que Adidas ofrece una amplia gama de productos, mientras que los productos Nike son muy limitados.

Es de mencionar que Nike tiene 643 valoraciones y reseñas, mientras que Adidas tiene 2.625. Considero que la muestra no esta balanceada y esto puede hacer que los resultados no sean del todo fiables.

# Clasificación

Una vez hecho el preprocesamiento de la base de datos y la EDA, comenzamos con la clasificación según el atributo target.

En este punto me di cuenta de que no sabía cuál era el atributo a predecir. Así que decidí probar con el que más obvio parecía, el atributo Sale Price.

Recalcar que todos los accuracy ’s son sobre el conjunto de validación.

## Atributo target Sale Price

Una vez escogido el target que me parecía el idóneo, probé con diferentes métodos de clasificación:

### KNN

Imagen que contiene texto

Descripción generada automáticamente

Ilustración 11: Clasificación con KNN

Como se puede observar en los resultados después de entrenar un modelo con KNN he obtenido resultados muy malos, ninguna de las K’s obtiene un buen resultado y estas clasificaciones no son correctas.

Cómo no sabía cuál era la causa de esto, decidí ir probando con diferentes modelos de clasificación para comprobar si los resultados mejoraban o, por otro lado se quedaban igual.

### Logistic Regression

Texto

Descripción generada automáticamente

Ilustración 12: Clasificación con Logistic Regression

Podemos ver que la predicción sigue siendo errónea como en el caso del KNN.

### Random Forest Regressor

Decidí probar con Random Forest, y en este caso sí obtuve mejores resultados que con los clasificadores hechos con KNN y Regresión Logística.

Texto

Descripción generada automáticamente

Ilustración 13: Clasificación con Random Forest Regressor

En este caso obtengo accuracy’s de 0.6, 0.66 y 0.71. Muchísimo mejor que en los dos apartados anteriores.

### Random Forest Classifier

Texto

Descripción generada automáticamente

Ilustración 14: Clasificación Random Forest Classifier

En este caso vemos una mejoría en comparación con el Random Forest Regressor.

Aún sin estar contenta del todo con los resultados, aunque con el último clasificador he obtenido una mejora significativa, decidí probar con otro atributo a predecir: el atributo Brand.

## Atributo target Brand

### KNN

Un conjunto de letras blancas en un fondo blanco

Descripción generada automáticamente con confianza media

Ilustración 15: Clasificación KNN

Al cambiar el atributo a predecir, inmediatamente se nota una mejoría en la clasificación. Se puede observar que los errores en la clasificación han menguado drásticamente con KNN.

A ver si en los demás modelos de clasificación ha pasado lo mismo:

### Logistic Regression

Texto

Descripción generada automáticamente

Ilustración 16: Clasificación Logistic Regression

En este caso también notamos una gran mejoría, antes obteníamos un accuracy del 0.15 y ahora del 0.75.

### Random Forest Regressor

Texto

Descripción generada automáticamente

Ilustración 17: Clasificación con Random Forest Regressor

Con este target y un modelo de clasificación con Random Forest obtenemos un 94% de accuracy. Un muy buen resultado, el mejor hasta el momento.

### Random Forest Classifier

Texto

Descripción generada automáticamente

Ilustración 18: Clasificación con Random Forest Classifier

Esta a la par con el Random Forest Regressor.

## Atributo target Discount

El otro atributo con mayor correlación es Discount, por lo tanto, hay que estudiar los modelos prediciéndolo:

### KNN

Un conjunto de letras blancas en un fondo blanco

Descripción generada automáticamente con confianza media

Ilustración 19: Clasificación KNN

Podemos ver que la predicción tiene pocos errores, podría ser un candidato.

### Logistic Regression

Texto

Descripción generada automáticamente

Ilustración 20: Clasificación Logistic Regression

El accuracy de este modelo es del 0.85, muy bueno.

### Random Forest Regressor

Texto

Descripción generada automáticamente

Ilustración 21: Clasificación Random Forest Regressor

Este modelo Random Forest clasificando el atributo Discout es el que mayor accuracy me ha dado, 0.97. Por lo tanto, este va a ser el taregt que voy a escoger finalmente.

### Random Forest Classifier

Texto

Descripción generada automáticamente

Ilustración 22: Clasificación Random Forest Classifier

En este caso sigue dando mejores resultados el Random Forest Regressor.

### SVM

Una vez he escogido el target, quería probar el clasificador SVM.

El mejor resultado me ha dado con los parámetros: C=10.000 y kernel=’poly’, y son los siguientes:

*Correct classification SVM 0.5 % of the data: 0.9253365973072215*

*Correct classification SVM 0.7 % of the data: 0.9398572884811417*

*Correct classification SVM 0.8 % of the data: 0.9541284403669725*

Calendario

Descripción generada automáticamente con confianza media

Ilustración 23: SVM

Este modelo clasifica mejor que los anteriores aunque, tarda más que el Random Forest anterior. Así que es mejor coger este último.

## Comparación de modelos

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sale Price | Tiempo | C | Kernel | Accuracy 0.8 | Best K |
| KNN | 1.3 s | - | - | 426/654 errores | 1 |
| Logistic Regression | 15.3 s | - | - | 12.84% | - |
| Random Forest Regressor | 3.4 s | - | - | 68.04% | - |
| Random Forest Classifier | 8.3 s | - | - | 80.42% | - |
| SVM | 27 min 51 s | 10.000 | poly | 33.94% | - |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Brand | Tiempo | C | Kernel | Accuracy 0.8 | Best K |
| KNN | 1.3s | - | - | 72/654 errores | 1 |
| Logistic Regression | 0.7 s | - | - | 73.09% | - |
| Random Forest Regressor | 3.9 s | - | - | 94.34% | - |
| Random Forest Classifier | 2.4 s | - | - | 94.80% | - |
| SVM | 19 min 55 s | 10.000 | poly | 79.61% | - |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Discount | Tiempo | C | Kernel | Accuracy 0.8 | Best K |
| KNN | 1.19 s | - | - | 114/654 errores | 1 |
| Logistic Regression | 1.8 s | - | - | 87.16% | - |
| Random Forest Regressor | 3.4 s | - | - | 96.94% | - |
| Random Forest Classifier | 6.1 s | - | - | 94.03% | - |
| SVM | 3 min 56 s | 10.000 | poly | 92.81% | - |

He elegido estos parámetros de SVM ya que eran los que más accuracy daban. He probado C = 0.1, 1, 100, 1000, 10000 y kernel = poli, rbf.

## Curva Precision Recall mejor modelo con Random Forest Classifier

Imagen que contiene Gráfico

Descripción generada automáticamente

Ilustración 24: Curva PR mejor modelo Random Forest Classifier

Para entender un poco mejor las curvas, especificaré a que corresponden cada una de las 6 clases de los gráficos.

* Class 0 🡪 50% descuento
* Class 1 🡪 40% descuento
* Class 2 🡪 60% descuento
* Class 3 🡪 0% descuento
* Class 4 🡪30% descuento
* Class 5 🡪 20% descuento

En general, podemos ver que casi todas las clases tienen un área bastante cercana al 1. En concreto, los descuentos del 0% y 30% son los más cercanos a 1, siendo los descuentos del 50% y 40% de 1.

El que tiene una área más baja es el 20%

Curva ROC mejor modelo con Random Forest Classifier

Gráfico

Descripción generada automáticamente

Ilustración 25: Curva ROC mejor modelo Random Forest Classifier

En el caso de las curvas ROC, volvemos a ver unas curvas muy cercanas a 1, de hecho, los descuentos del 50% y 40% son igual a 1, y los restantes son prácticamente 0.99.

Son muy buenos resultados los de las dos gráficas.

## R2\_score Random Forest Regressor

Tenemos un R2\_socre de 0.965 con el modelo hecho con Random Forest Regressor. Es muy buen valor.

## Gráficas Random Forest Regressor

### Validation – Predicted

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Ilustración 26: Validation Vs Predicted

### Test – Predicted

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Ilustración 27: Test Vs Predicted

## Feature importance Random Forest Regressor

Gráfico, Gráfico de barras

Descripción generada automáticamente

Ilustración 28: Feature Importance RF regressor

Como podemos ver en este gráfico de barras, Random Forest Regressor tiene mucho más en cuenta el atributo Product ID en el modelo. Pero también tiene mucho en cuenta las Descripciones de los productos y Last Visited.

Estos son los tres atributos que más importancia tienen en este modelo.

## Feature importance Random Forest Classifier

Gráfico, Gráfico de barras

Descripción generada automáticamente

Ilustración 28: Feature Importance RF classifier

Como podemos ver en este gráfico de barras, Random Forest Classifier tiene mucho más en cuenta el atributo Description en el modelo. Pero también tiene mucho en cuenta las Descripciones de los Sale Price y Last Visited.

Estos son los tres atributos que más importancia tienen en este modelo.

# PCA

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Si observamos el grafico 3D de la derecha que nos muestra el PCA, podemos ver que no sigue ningún tipo de patrón. Esto concuerda con los resultados no muy buenos del modelo de Regresor Logístico.

Así que este gráfico no nos da ningún tipo de información.

Ilustración 29: PCA

# 

# Conclusiones

Lo primero a comentar es que Nike está compitiendo con tres submarcas de Adidas y las muestras están muy desbalanceadas. No es un estudio justo.

Pero si me baso en los datos que tengo, veo que Adidas tiene unos precios más bajos y ofrece descuentos a sus clientes, a diferencia de Nike. Por esto, los compradores parecen contentos y lo expresan en las valoraciones y reseñas de la marca.

He encontrado un modelo Random Forest Regressor que obtiene más de un 96% de accuracy prediciendo el atributo Discount que he establecido como target.

El hecho de fijar un target me costó. Probé los que me parecieron más lógicos y me quedé con el que mejores resultados obtenía.