Aprenentatge computacional

Pràctica 1a

Profesores:
Pau Rodríguez
Jordi Gonzàlez

Alumnos:
Alejandro Garcia Carballo
1423957
Juan Plúa Gutiérrez
1358255

Índice

1. Introducción

Esta primera práctica consiste en resolver diferentes problemas utilizando datos reales aplicando modelos de regresión, donde se analizarán los atributos para seleccionar los más representativos y normalizarlos, evaluar correctamente el error del modelo utilizado, visualizar los datos y el modelo resultante, y, saber aplicar el proceso de descenso del gradiente.

A partir de una regresión, tendremos que poder predecir qué valores tomará una variable objetivo teniendo en cuenta otras, a partir de las cuáles el algoritmo aprende. Para decidir qué variable tomamos como objetivo, tendremos que entender qué datos son los que estamos tratando y saber cuáles serán de interés para la empresa o persona para la que estamos trabajando.

En este informe iremos documentando todo el proceso del estudio de los datos, sus respectivas representaciones en gráficas, los resultados obtenidos y las conclusiones a las que hemos llegado.

2. Estudio preliminar

En este apartado haremos un estudio del conjunto de datos que se nos ha proporcionado y cual es su significado.

Los datos los podemos separar en tres bloques, donde tendremos las fechas de los proyectos, las variables físicas y financieras del proyecto, que incluye el costo de construcción y los precios de venta actuales; y, las variables económicas correspondientes a los apartamentos residenciales de viviendas unifamiliares en la província de Teherán, Irán.

Antes para poder comprender mejor los datos hay que decir que en la moneda oficial en Irán es el Rial y que allí ahora mismo están en el año 1397 ya que siguen el calendario persa. A continuación mostramos a como está el cambio con el euro y una simple comparación del precio en el supermercado de un litro de leche en Irán y España actualmente.

1 EUR = 48.599,63 IRR
1 IRR = 0.0000205722 EUR

1L de leche en España: 0.78€
1L de leche en Irán: 0.58€

Las primeras variables que tenemos son las fechas de los proyectos, los años van de del 0 al 97, entendemos así que se refieren los años comprendidos entre el 1300 hasta el año actual, el 1397; y, los trimestres se comprende en cifras del 1 al 4, refiriéndose al que ocupa en su respectivo año:

FECHAS DE LOS PROYECTOS		
ID	**Descripción**	**Unidad**
F1	Año de inicio	N/A
F2	Trimestre de inicio	N/A
F3	Año de finalización	N/A
F4	Trimestre de finalización	N/A

En segundo lugar, las variables físicas y financieras del proyecto donde V9 y V10 utilizaremos como nuestras variables de salida:

VARIABLES FÍSICAS Y FINANCIERAS DEL PROYECTO		
ID	**Descripción**	**Unidad**
V1	Localidades del proyecto definidas por código postal	N/A (1)
V2	Área total de la construcción	m2
V3	Cantidad de área	m2
V4	Estimación preliminar del costo total de la construcción basado en los precios al inicio del proyecto	10M IRR
V5	Estimación preliminar del costo de la construcción basado en los precios al inicio del proyecto	10K IRRm
V6	Estimación preliminar del costo equivalente basado en los precios al inicio del proyecto	10K IRRm
V7	Duración de la construcción	N/A (2)
V8	Precio por unidad al comienzo del proyecto	10K IRR
	OUTPUTS	
V9	Precios de venta actuales	10K IRR
V10	Precio actual de construcción	10K IRR

(1) Unidades de 1 a 20

(2) Unidades random entre 2 y 23

Y, en tercer lugar, las variables económicas de las que tenemos cinco conjuntos de datos dependiendo de en el tiempo de retraso comprendidos entre 0 y 5:

VARIABLES ECONÓMICAS		
ID	**Descripción**	**Unidad**
V11	Número de permisos de construcción requeridos	N/A
V12	Índice de servicios de una construcción para un año preseleccionado base	N/A
V13	Índice de precios al por mayor de materiales de construcción para el año base	N/A
V14	Áreas totales del piso de los permisos de construcción emitidos por la ciudad / municipio	m2
V15	Liquidez cumulativa	10M IRRm
V16	Inversión del sector privado en nuevas construcciones	10M IRRm
V17	Índice de precios del terreno para el año base	10M IRRm
V18	Número de préstamos otorgados por los bancos en una resolución temporal	N/A
V19	El monto de los préstamos otorgados por los bancos en una resolución temporal	10m RRm
V20	La tasa de interés del préstamo en una resolución temporal	%
V21	El costo promedio de la construcción de edificios por el sector privado en el momento de la finalización de la construcción	10K IRRm/m2
V22	El promedio del costo de construcción de edificios por el sector privado al inicio de la construcción	10K IRR/m2
V23	Tipo de cambio oficial respecto a dólares	1 IRRm
V24	Tipo de cambio no oficial (mercado callejero) con respecto a dolares	1 IRRm
V25	(IPC) Índice de los precios al consumidor en el año base	N/A
V26	IPC de la vivienda, agua, combustible y energía en el año base	N/A
V27	Índice del mercado de valores	N/A
V28	V28 - Población de la ciudad	N/A
V29	Precio del oro por onza	1 IRRm

3. Análisis e interpretación de los datos

Analizando las variables que tenemos en la tabla, creemos que hay dos de ellas que son principalmente las que más interés pueden tener en el mundo real para una empresa, que serían el precio de venta y el costo de construcción. En nuestro caso, hemos optado por centrarnos en tener como variable objetivo el coste de construcción, ya que esto puede servir para que una constructora pueda predecir cuáles van a ser sus costes y decidir si llevar un proyecto a cabo o no.

Una vez tenemos claro cuál va a ser nuestro objetivo, debemos comenzar a pensar en qué variables del conjunto de entrada son las que vamos a escoger para predecir. Primero, probamos de hacer una predicción con todos los datos para poder ir comparando las mejoras.

	MSE	**Root MSE**
Mínimo	1423.579634073532	37.730354279724594
Máximo	25754.801130362594	160.48302443050665

Tabla 1: MSE y RMSD haciendo regresión con todos el dataframe

Para realizar esta prueba, hemos separado los datos de entrenamiento de los datos de test. Hemos utilizado un 80% de los datos para el entrenamiento y un 20% para el test. A continuación hemos creado una regresión con todos los datos, y hemos calculado el MSE y el RMSD (Root-mean-square deviation, lo que viene a ser la raíz cuadrada del MSE) de la predicción. Si tenemos en cuenta que cada unidad de coste son 10.000 IRR, el MSE máximo puede ser un problema, ya que el error es muy grande.

¿Cuáles son los atributos más importantes para hacer una buena predicción?

Para poder ver mejor las relaciones entre las distintas variables, hemos decidido hacer una matriz de correlación, de esta manera podremos centrar mejor nuestra atención en cuáles son las variables de interés y cuáles son aquellas que podemos descartar.

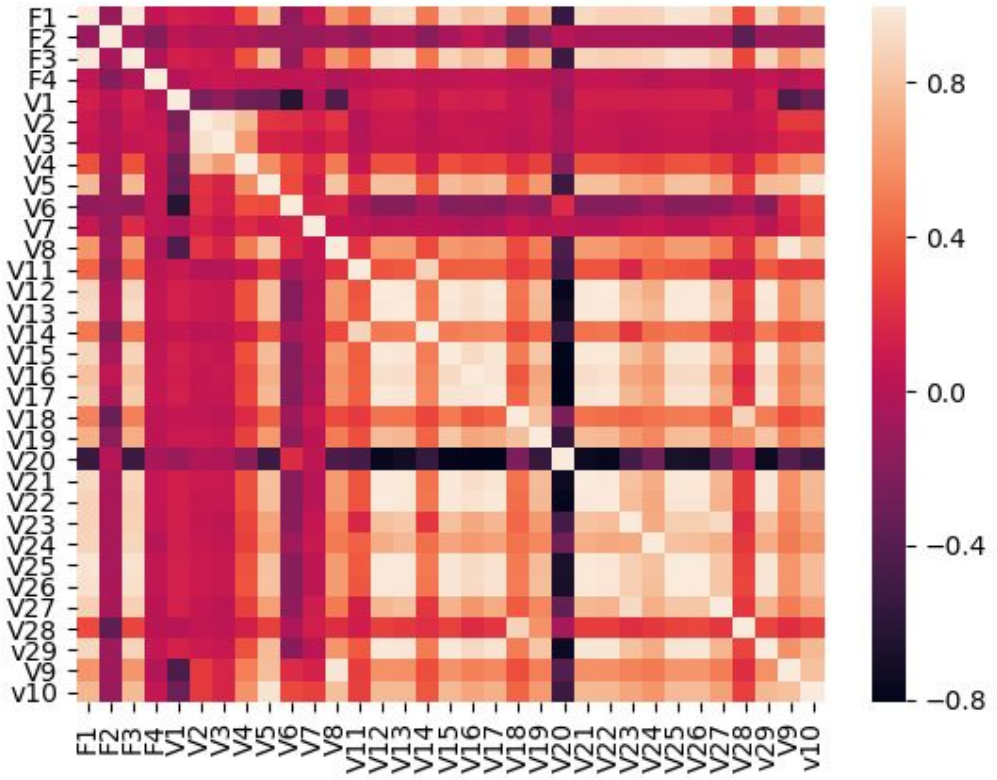


Ilustración 1: Matriz de correlación entre todas las variables

Algo que nos ha llamado la atención al ver estas correlaciones, es que la V2 y V3 (que son variables relacionadas con el área de construcción) no tienen casi influencia en el coste y el precio del apartamento. Por lo tanto, tanto estas variables como todas las que no tengan influencia en el coste de la construcción, no las tendremos en cuenta a la hora de predecir.

Si vemos la matriz, podemos ver que la variable target (V10, que es el coste actual) tiene una correlación alta con V5 (que es la previsión de costes basada en los precios). Esto tiene sentido, y significa que si esa variable sirve para hacer una predicción los estudios de previsión que se han ido haciendo hasta el momento eran acertados.

Primero probaremos de hacer regresiones con valores individuales, para ver como de acertados son, y luego probaremos combinaciones de 2 y de más valores. Comenzaremos probando con V5:

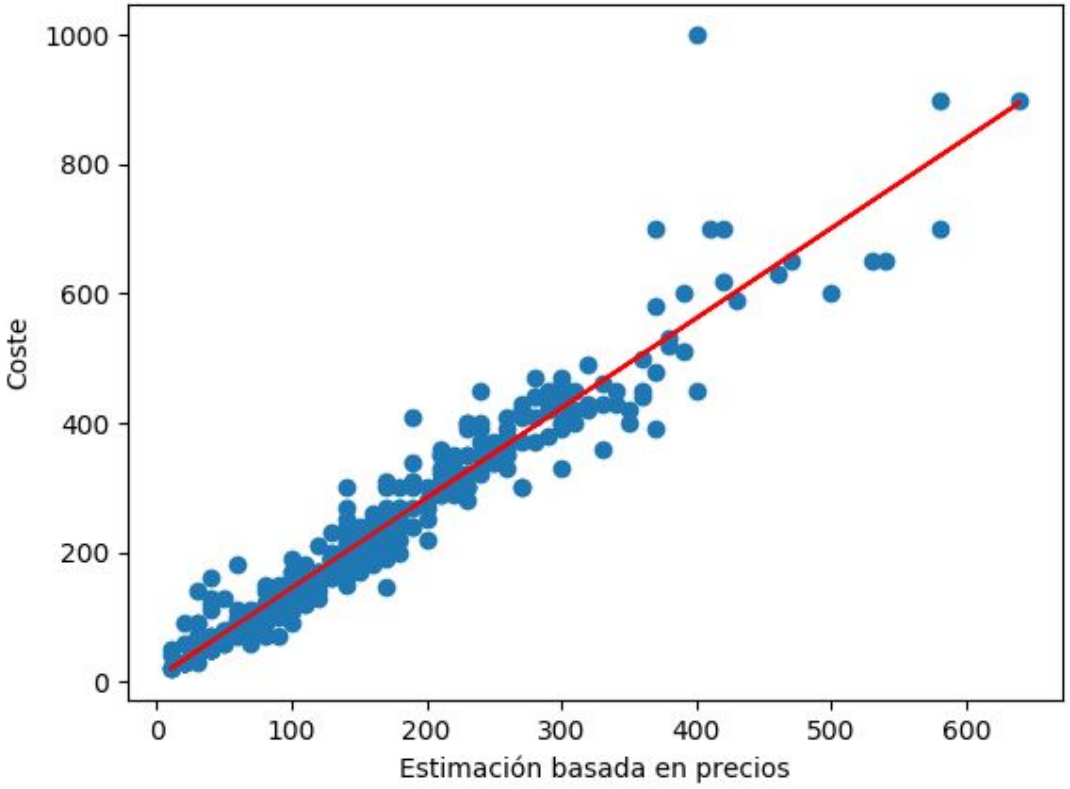


Ilustración 3: Regresión lineal coste - estimación precios

Como podemos observar en la Ilustración 3, gráficamente ya se puede observar que la línea se ajustará bastante a los datos, pero tenemos un punto suelto (coordenada cercada a 400 – 1000) que lo único que hace es bajar el MSE, ya que es un valor poco común. Teniendo en cuenta esto, los datos que tenemos son los siguientes:

MSE	Root MSE	R2
2093.303172972442	45.752630230101985	0.9225869813602681

Tabla 2: MSE, Root MSE y R2 de una predicción basada en V5

Tenemos un R2 bastante elevado, lo que indica que hacer las previsiones de coste basadas en la estimación de los precios está siendo eficaz. Vamos a intentar conseguir bajar el error y subir el R2 eliminando ese dato inusual.

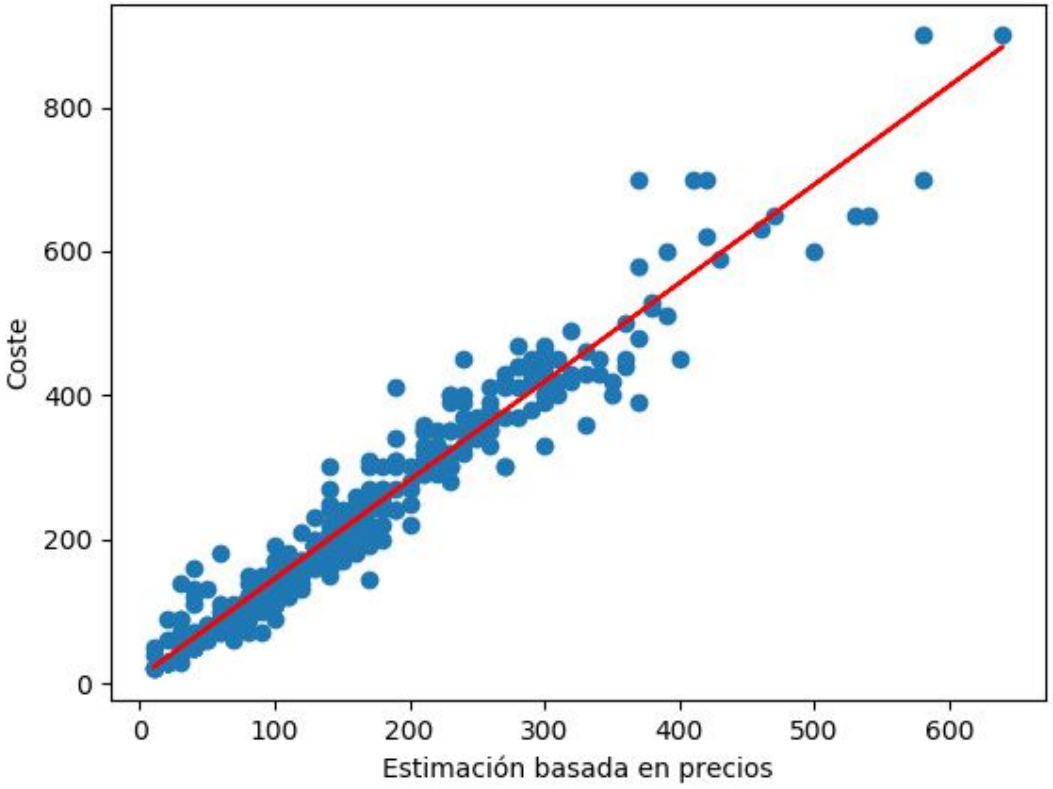


Ilustración 4: Misma gráfica pero con un valor inusual eliminado

Como podemos ver en la gráfica, ahora el máximo de los datos en coste no llega a 1000. Vamos a analizar el error producido con esta mejora:

MSE	Root MSE	R2
1279.26193616677876	35.76677139703537	0.9449061799238512

Tabla 3: MSE, Root MSE y R2 con dato inusual eliminado

Podemos comprobar que haciendo esta modificación en el dataframe, estamos consiguiendo una mejora del R2 y estamos reduciendo el MSE y RMSD. Parece que va a ser difícil mejorar este valor, pero quizá a la empresa le interesa poder determinar el coste sin tener en cuenta esta previsión de precios. Haremos alguna combinación de valores para ver si podemos dar mejores resultados o alguna alternativa a utilizar la variable V5.

Otro valor que podría influenciar podría ser el coste por unidad. Vamos a probar qué resultados obtenemos entrenando el modelo a partir de V12 (Índice de servicios de una construcción "B" para un año preseleccionado base "A"):

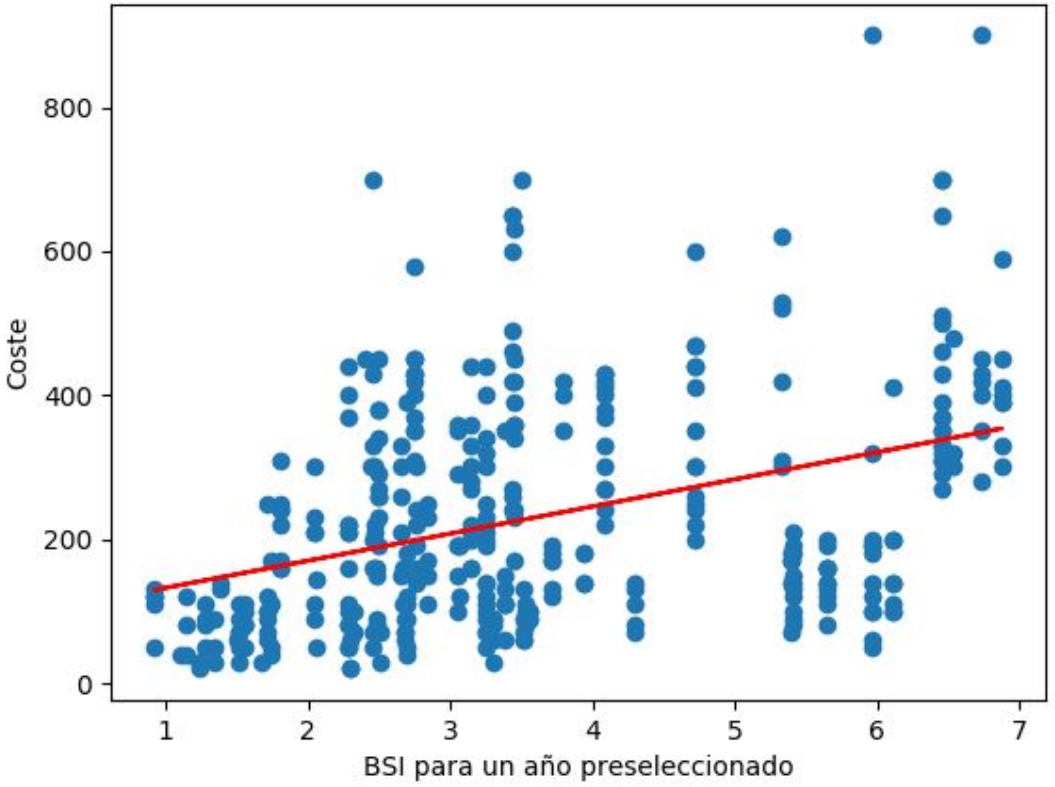


Ilustración 5: Regresión lineal a partir de V12

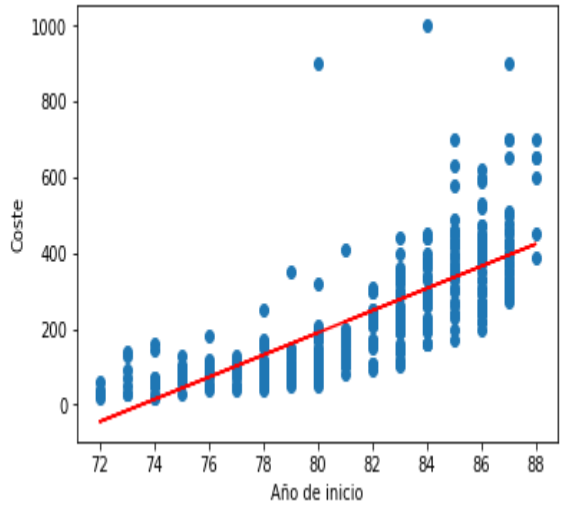
Si observamos la gráfica, vemos que el modelo obtenido tiene unos datos más dispersos.

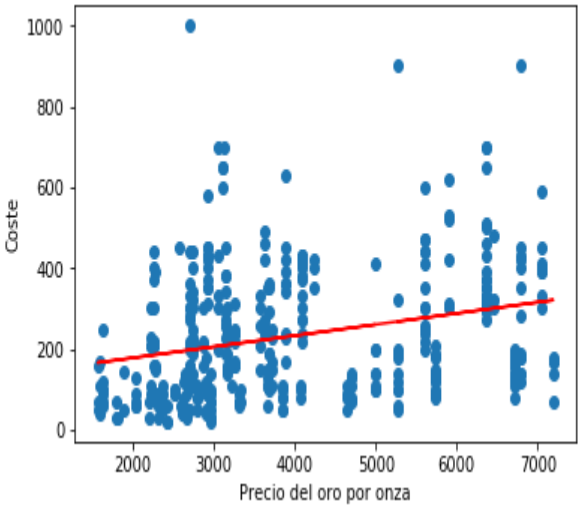
MSE	Root MSE	R2
21305.851484033283	145.96524067062433	0.13843583279956873

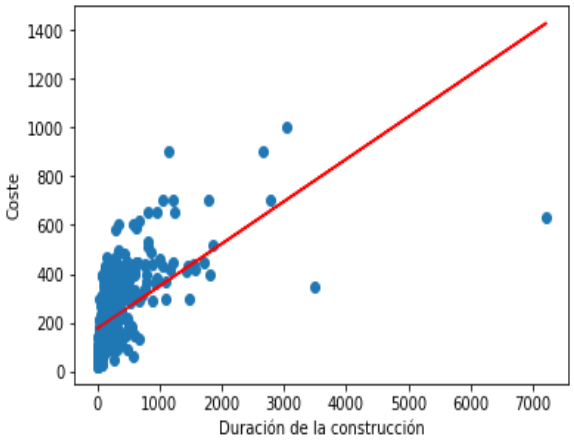
Tabla 4: MSE, Root MSE y R2 de una predicción basada en V12

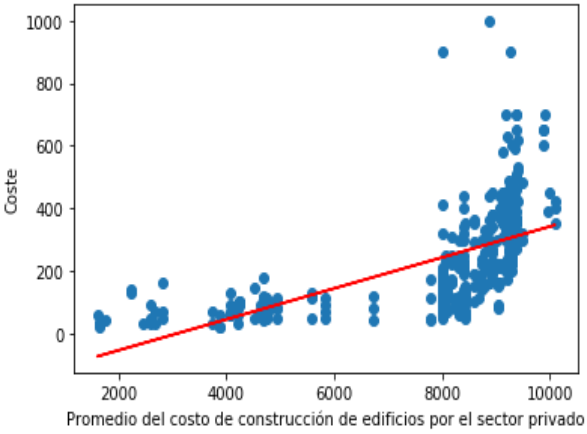
Y el resultado en error y R2 es peor que el que hemos obtenido al utilizar V5. Por lo tanto, este dato no lo escogeremos, al menos como dato individual.

Otras gráficas que también hemos visualizado con valores que no han conseguido un mejor resultado:









	MSE	Root MSE	R2
F1	9188.35	95.85	0.611
F2	24891.48	157.77	0.007
F3	10656.85	103.23	0.592
F4	26846.67	163.84	-0.001
V1	23176.20	152.23	0.071
V2	22777.78	150.92	0.045
V3	26942.09	164.14	0.002

V4	16470.37	128.33	0.326
V5	1504.85	38.79	0.942
V6	21733.40	147.42	0.054
V7	24970.71	158.02	0.048
V8	9969.63	99.84	0.579
V11	22780.75	150.93	0.074
V12	10475.07	102.34	0.593
V13	9696.58	98.47	0.583
V14	19650.29	140.17	0.161
V15	9733.72	98.65	0.600
V16	12661.28	112.52	0.487
V17	11636.90	107.87	0.538
V18	22110.71	148.69	0.162
V19	14730.33	121.36	0.387
V20	16251.32	127.48	0.309
V21	8326.53	91.24	0.656
V22	9941.61	99.70	0.594
V23	13807.77	117.50	0.457
V24	16777.55	129.52	0.367
V25	8937.43	94.53	0.638
V26	7803.41	88.33	0.673
V27	13023.22	114.11	0.444
V28	23633.84	153.73	0.081
V29	9292.62	96.39	0.613

Tabla 5: Tabla resumen MSE, Root MSE y R2 todos los datos

Podemos observar en la ilustración 9 que no hay ningún dato que de manera individual consiga un mejor resultado que V5.

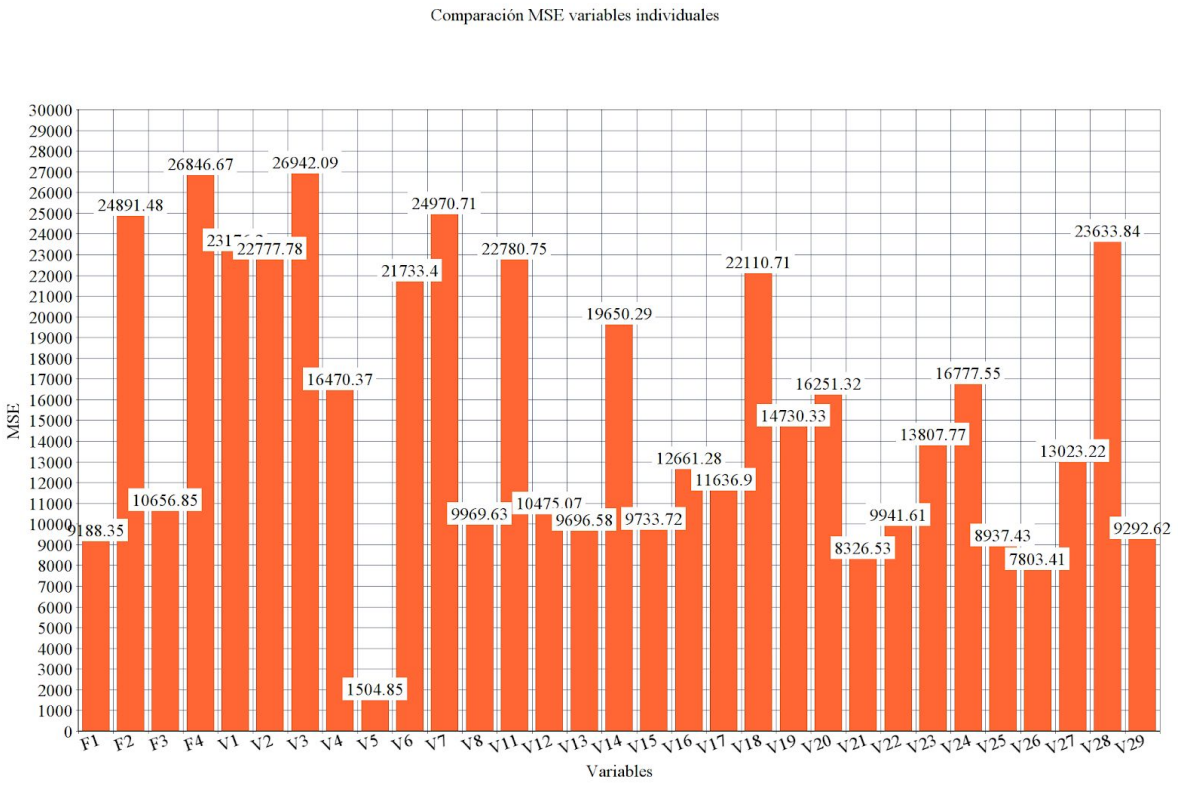


Ilustración 6: Histograma comparativo de MSE

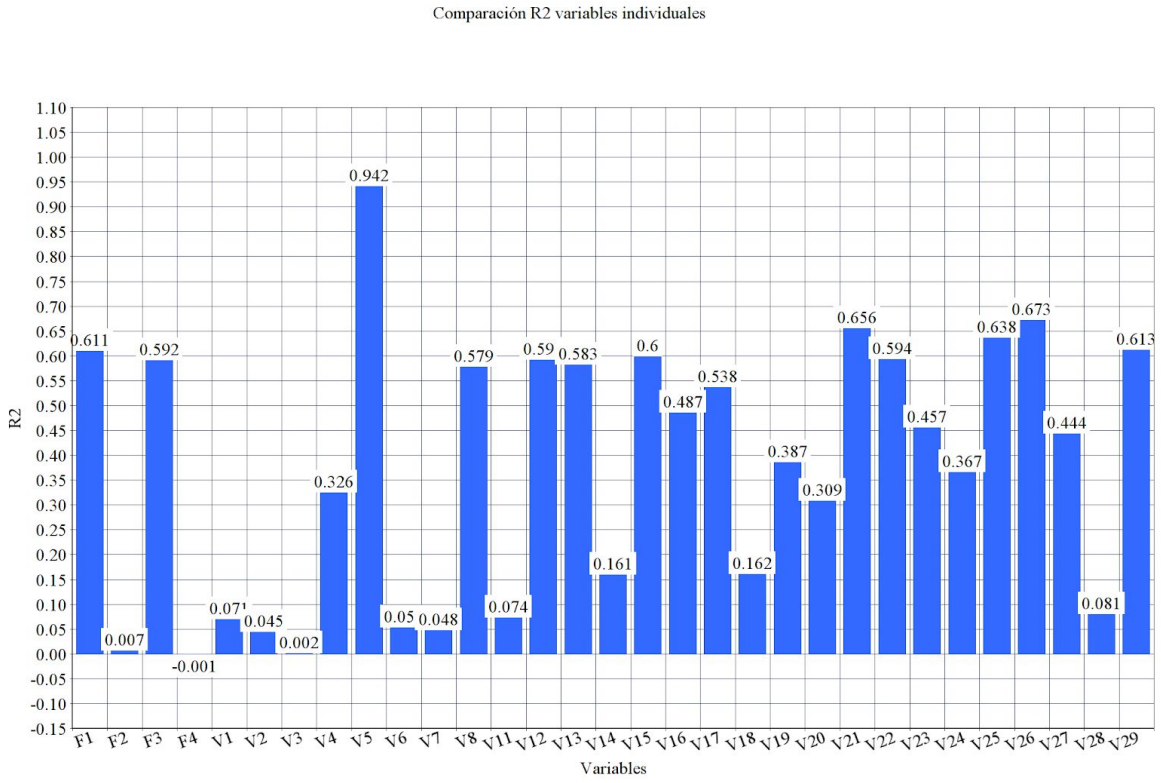


Ilustración 7: Histograma comparativo de R2

Como conclusión podemos decir que las variables que nos permitirán hacer una buena predicción serán aquellas que influyan más a nuestra variable objetivo, y que por lo tanto nos den un MSE lo más bajo posible y un R2 lo más cercano a 1 posible.

¿Cómo influye la normalización en la regresión?

Vamos a probar si conseguimos reducir el error y mejorar el R2 si trabajamos con los datos estandarizados:

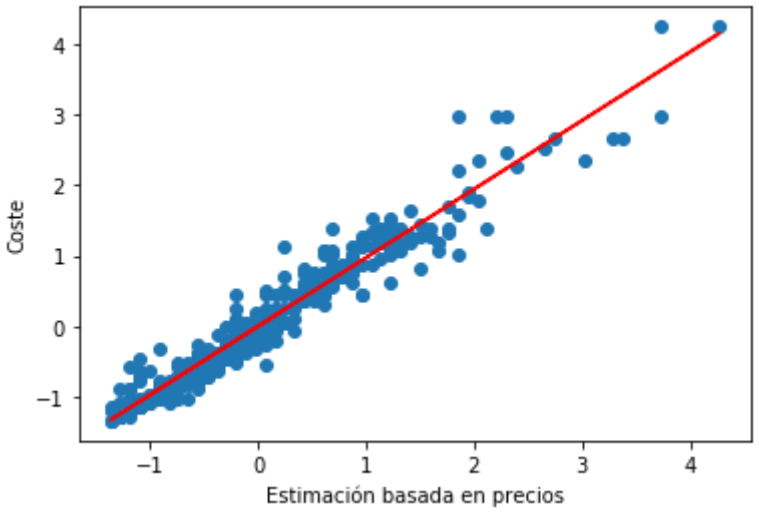


Ilustración 8: Regresión lineal con V5 y datos estandarizados

A primera vista, podemos ver que la gráfica no ha cambiado prácticamente, esto indica que los datos sin normalizar ya estaban en unas escalas que no eran demasiado distintas. Vamos a ver cómo ha afectado a los resultados de errores:

MSE	Root MSE	R2
0.0560481	0.23674479629829667	0.9423597312519516

Tabla 6: MSE, Root MSE y R2 de una predicción basada en V5 con datos normalizados

Como vemos, los resultados son prácticamente iguales a los resultados que habíamos obtenido sin estandarizar los datos. Viendo esta gráfica:

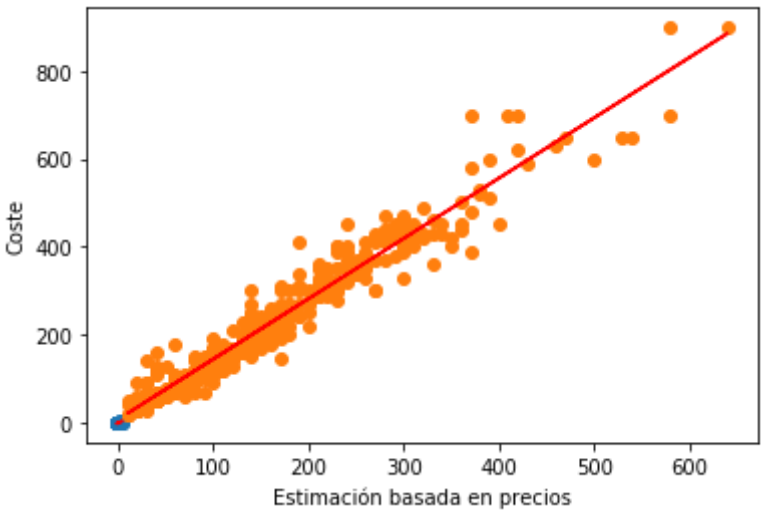


Ilustración 9: Comparación datos estandarizados y sin estandarizar

Teniendo en cuenta que los datos azules son los estandarizados y los naranjas sin estandarizar, comprobamos que si los interponemos, los datos estandarizados se acercan un poco más a cero, pero no cambian de los datos sin estandarizar.

¿Cómo mejora la regresión cuando se filtran aquellos atributos de las muestras que no contienen información?

Ahora vamos a hacer pruebas teniendo en cuenta más de una variable en conjunto a la V5 que ha sido la que mejor resultado nos ha dado anteriormente. Comenzaremos haciendo una prueba con V5 y V8, ya que V8 tiene unas unidades similares a V5 y al ser un precio unitario puede complementar a V5.

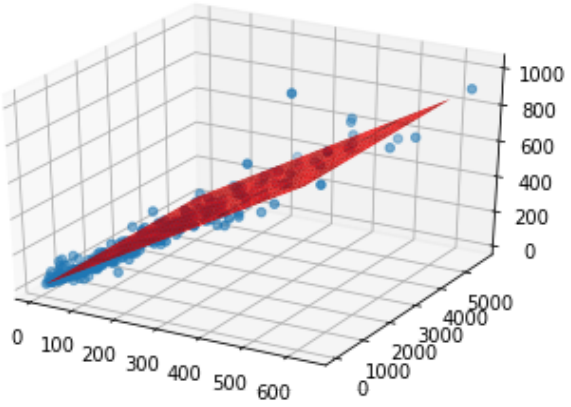


Ilustración 10: Regresión teniendo en cuenta V5 y V8

	MSE	Root MSE	R2
V5 y V8	2075.496914974071	45.55762191965326	0.9233745958578743
V5	1279.26193616677876	35.76677139703537	0.9449061799238512

Tabla 7: Comparación resultados predicción con V5 y V8 y solo V5

Como comprobamos en los resultados, el hecho de añadir la variable V8 para realizar la predicción está influyendo de manera negativa en el error y en el R2.

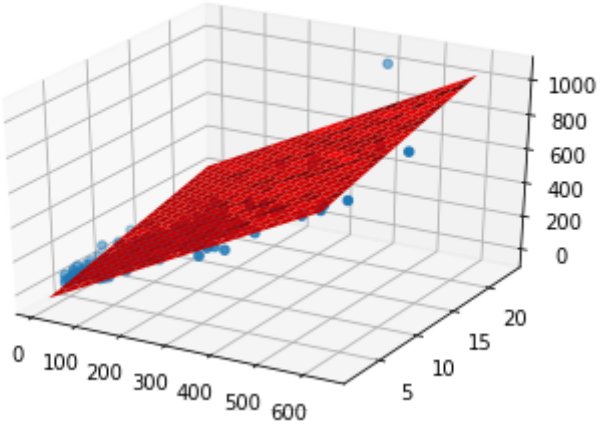


Ilustración 11: Modelo con V5 y V12

Este modelo consigue unos resultados muy similares a los de utilizar únicamente V5, pero tampoco los mejora. Lo que sí podemos observar en este caso es una pequeña mejora en el error a la hora de trabajar con los datos normalizados en vez de tratarlos sin normalizar.

	MSE	Root MSE	R2
Sin normalizar	2026.8708363247833	45.02078227135534	0.9205771517070892
Normalizado	0.0583874837307511	35.76677139703537	0.9449061799238512

Tabla 8: Comparación resultados predicción con V5 y V12 normalizados y sin normalizar

Si hacemos un modelo que tenga en cuenta todas las variables excepto aquellas que hemos visto por los resultados anteriores que no eran buenas para predecir, obtenemos este resultado:

MSE	Root MSE	R2
0.4306048	0.6562048467825735	0.5914157152983761

Tabla 9: Resultados de modelo que utiliza todas las variables menos las que no aportan información

Como conclusión a este apartado, podemos decir que el hecho de tener en cuenta sólo aquellas variables que nos aportan información mejora mucho respecto a las que no aportan nada, comparando el error cometido con las variables que se han descubierto que son útiles respecto a las que no lo son.

¿Cómo afecta la división de la base de datos en aprendizaje y test en la correcta evaluación de un regresor?

Ahora vamos a probar el modelo con una distribución de datos de entrenamiento y de test distinta. Hasta ahora, lo hemos probado todo con una porcentaje de 80% de datos de entrenamiento y un 20% de datos de validación. Vamos a probar nuestro modelo con un 50% de datos de entrenamiento y un 50% de datos de validación.

	MSE	Root MSE	R2
50% T 50% V	1612.711875914822	40.15858408752507	0.9438958011672498
80% T 20% V	1279.2619361667787	35.76677139703537	0.9449061799238512

Tabla 10: Comparación resultados de V5 considerando 50% training 50% validación y 80% training 20% validación

El MSE en los datos entrenados con un 80% de datos de entrenamiento consigue un mejor error que los que están entrenados con un 50%. También se puede observar si se hacen varias pruebas, que el modelo entrenado con un 50% de datos, es más inestable, es decir, da unos resultados más variantes cada vez que se ejecuta, y por lo tanto hace que este modelo sea menos fiable. La división de los datos nos permite evaluar si un regresor funciona correctamente o no por el hecho de que al partir los datos en dos grupos, podemos utilizar el segundo conjunto para probar el regresor con valores que quizá no ha tratado nunca, y de esta manera comprobar que responde bien a cualquier dato y que no ha aprendido solo a tratar los datos de entrenamiento. Cuando no usamos esta separación, el MSE nos da un valor constante, cuando separamos los datos, el MSE cambia cada vez que lo ejecutes.

¿Tiene sentido el modelo encontrado cuando se visualiza sobre los datos?

Sí, ya que se puede observar como la recta de regresión pasa a una distancia lo más pequeña posible del conjunto de datos, para conseguir el menor error. En nuestro caso, los datos de V5 tienen una forma bastante lineal, por lo que es perfecto para aplicar la recta de regresión.

¿Ayuda la visualización para identificar aquellas muestras para las que el regresor obtiene los peores resultados de predicción?

Sí, se puede observar sobretodo en gráficas donde se observa que los datos son muy dispersos, y no hay una recta que pueda conseguir un buen error entre unas muestras tan dispares. Además, igual que en nuestro caso, puedes observar si hay algún valor inusual que está empeorando tu error y eliminarlo para mejorar tu error.

4. Dificultades

Durante el proceso de realización de este proyecto hemos tenido bastantes dudas para saber si lo que estábamos haciendo tenía algún sentido o si era correcto, y esa ha sido una de las partes más complicadas, ya que debíamos darle un sentido a los datos a través del análisis de estos mismos. Algo que nos ha extrañado bastante es el hecho de que aplicando diferentes técnicas explicadas como la normalización o utilizar varias variables, no han mejorado nuestro resultado de manera notable. Y de igual manera, el hecho de trabajar con un 50% de datos de entrenamiento y un 50% de test, hay una pequeña diferencia pero creíamos que debía influir más.

5. Conclusiones

Hemos logrado hacer los apartados C y B de la práctica aunque los resultados obtenidos no nos hayan parecido del todo correctos, hemos escogido los casos más evidentes para poder realizar los ejercicios y poder plasmarlo en este informe.

La experiencia de programar el código en los programas de lenguaje python Jupyter Notebook y Spyder ha resultado sencillo y cómodo. Pero si hemos notado complicaciones a la hora de poder aplicar todos los conceptos dados en teoría en la práctica debido a su complejidad.

Hemos conseguido un modelo que tienen un error bastante bajo y con un R2 cercano a 1, pero solo ha hecho falta la utilización de una variable para ello. Esa variable era la estimación del coste a través de los precios individuales, por lo tanto, tiene sentido que esa variable se acerque al resultado, y eso significa que esa estimación se está haciendo correctamente.

Por último, nos hubiera gustado contar con más tiempo para poder realizar la práctica en todos los apartados. Pero, aun así, consideramos que lo aprendido a lo largo de esta primera parte del proyecto ha sido muy positivo para la aplicación en problemas de la vida real.