DATOS PERSONALES		FIRMA
Nombre: DNI:		
Apellidos:		
ESTUDIO	**ASIGNATURA**	**CONVOCATORIA**
MÁSTER UNIVERSITARIO EN INGENIERÍA MATEMÁTICA Y COMPUTACIÓN (PLAN 2016)	4391020006.- TÉCNICAS MULTIVARIANTES	Ordinaria
FECHA	**MODELO**	**CIUDAD DEL EXAMEN**
09-11/07/2021	Modelo - C	
Etiqueta identificativa		

INSTRUCCIONES GENERALES

1. Ten disponible tu documentación oficial para identificarte, en el caso de que se te solicite.
2. Rellena tus datos personales en todos los espacios fijados para ello y lee atentamente todas las preguntas antes de empezar.
3. Las preguntas se contestarán en la lengua vehicular de esta asignatura.
4. Si tu examen consta de una parte tipo test, indica las respuestas en la plantilla según las características de este.
5. Debes contestar en el documento adjunto, respetando en todo momento el espaciado indicado para cada pregunta. Si este es en formato digital, los márgenes, el interlineado, fuente y tamaño de letra vienen dados por defecto y no deben modificarse. En cualquier caso, asegúrate de que la presentación es suficientemente clara y legible.
6. Entrega toda la documentación relativa al examen, revisando con detenimiento que los archivos o documentos son los correctos. El envío de archivos erróneos o un envío incompleto supondrá una calificación de “no presentado”.
7. Durante el examen y en la corrección por parte del docente, se aplicará el Reglamento de Evaluación Académica de UNIR que regula las consecuencias derivadas de las posibles irregularidades y prácticas académicas incorrectas con relación al plagio y uso inadecuado de materiales y recursos.

Puntuación

Examen

- 10 10.00 puntos

El examen consta de cuatro preguntas.

Las puntuaciones son:

Pregunta 1: 1.5 puntos.

Pregunta 2: 1.5 puntos.

Pregunta 3: 2 puntos.

Pregunta 4: 5 puntos.

Responde a las preguntas en el espacio indicado entre las páginas 3 y 15.

Encontrarás las preguntas del examen a partir de la página 16.

¡Suerte!

1. Pregunta 1 (Responder en 1 caras)

2. Pregunta 2 (Responder en 1 caras)

3. Pregunta 3 (Responder en 2 caras)

4. Pregunta 4 (Responder en 5 caras)

(1) **(1.5 puntos)** Si en un modelo que se estima con 120 datos y que tiene 4 regresores, incluyendo la constante, se obtiene un coeficiente de determinación corregido igual a 0.75, determina qué porcentaje de variación de la variable endógena queda explicado por la regresión.

(2) **(1.5 puntos)** Tras 85 semanas en estudio con un grupo de 95 trabajadores de los cuales 79 (controles) no estuvieron sometidos a un determinado agente, supuestamente nocivo, y 16 de ellos sí lo estuvieron, se realizó examen médico para detectar cuáles tenían cierto parámetro desestabilizado. Los resultados vienen en la tabla siguiente:

	Exposición NO	Exposición SÍ	Total
Parámetro Desestabilizado NO	74	12	86
Parámetro Desestabilizado SÍ	5	4	9
Total	79	16	95

Utiliza la regresión logística para establecer si es significativo o no el factor exposición al agente.

(3) **(2 puntos)** La siguiente tabla contiene 8 casos bidimensionales a partir de los cuales queremos agrupar los diferentes ejemplos en tres grupos.

Caso	X1	X2
1	1	1
2	2	4
3	3	2
4	3	5
5	4	4
6	4	7
7	6	4
8	6	6

¿Cómo quedarían clasificados estos ejemplos de acuerdo al algoritmo k-means? Detalla cada uno de los pasos que harías aplicados a este ejemplo.

¿Se te ocurre otro algoritmo de aprendizaje automático que podrías usar para este caso? ¿Cuál y por qué?

(4) **(5 puntos)** La siguiente tabla contiene información acerca de una colección de libros. Se conoce del peso total de cada libro, el volumen que tiene y el tipo de tapas (duras (D) o blandas (B)):

Peso	849	850	950	1640	150	750	600	875
Volumen	890	805	1001	1525	39	701	641	1028
Tapas	D	D	D	D	D	D	D	D
Peso	975	450	1050	335	525	850	760	1212
Volumen	1492	519	1110	505	834	944	920	1350
Tapas	B	B	B	B	B	B	B	B

Se quiere generar un modelo lineal múltiple que permita predecir el peso de un libro en función de su volumen y del tipo de tapas.

Responde a las siguientes preguntas, justificando todas las respuestas:

(a) Comprueba, mediante análisis gráfico y correlación, si existe una relación lineal significativa entre la variable peso y la variable volumen.

(b) Comprueba, mediante un boxplot, si la variable Tapas puede influir de forma significativa en el peso.

(c) Como consecuencia, ¿pueden ambas variables volumen y tapas ser buenos predictores en un modelo lineal múltiple para la variable dependiente peso?

(d) Genera un modelo lineal múltiple.

(e) ¿Qué porcentaje de la variabilidad observada en el peso de los libros es capaz de explicar el modelo?

(f) ¿Son útiles los predictores?

(g) ¿Es significativo el modelo en su conjunto?

(h) Comprueba las siguientes condiciones para la regresión múltiple lineal:

- Relación lineal entre los predictores numéricos y la variable dependiente. ¿Hay algún dato atípico?
- Distribución normal de los residuos. En el caso que haya valores atípicos, ¿es diferente si los excluimos?
- Variabilidad constante de los residuos.
- Multicolinealidad.
- Tamaño de la muestra.

Examen

EJERCICIO 1

El modelo cuenta de 4 regresores, lo cual significa que tiene 4 variables de regresion. Al tener un coeficiente de determinacion corregido R2, se representa la proporcion de la varianza de la variable a predecir, o de la variable de la cual queremos obtener la regresion. Este coeficiente se puede representar como un cociente de la suma de cuadrados de la regresion, respecto a la suma total de estos mismos. Este coeficiente toma valores entre 0 y 1, por lo cual la medida de la variabilidad de los valores que se ajustan a la media muestral, es muy cercana a la media de la misma variabilidad de la muestra de la variable a predecir.

Un valor de R muy cercano a cero, indicaria que muy pocos valores se estan pegando a la respuesta de la prediccion, o en otras palabras, a la regresion. Un valor de R cercano a 1, indicaria un posible overfitting.

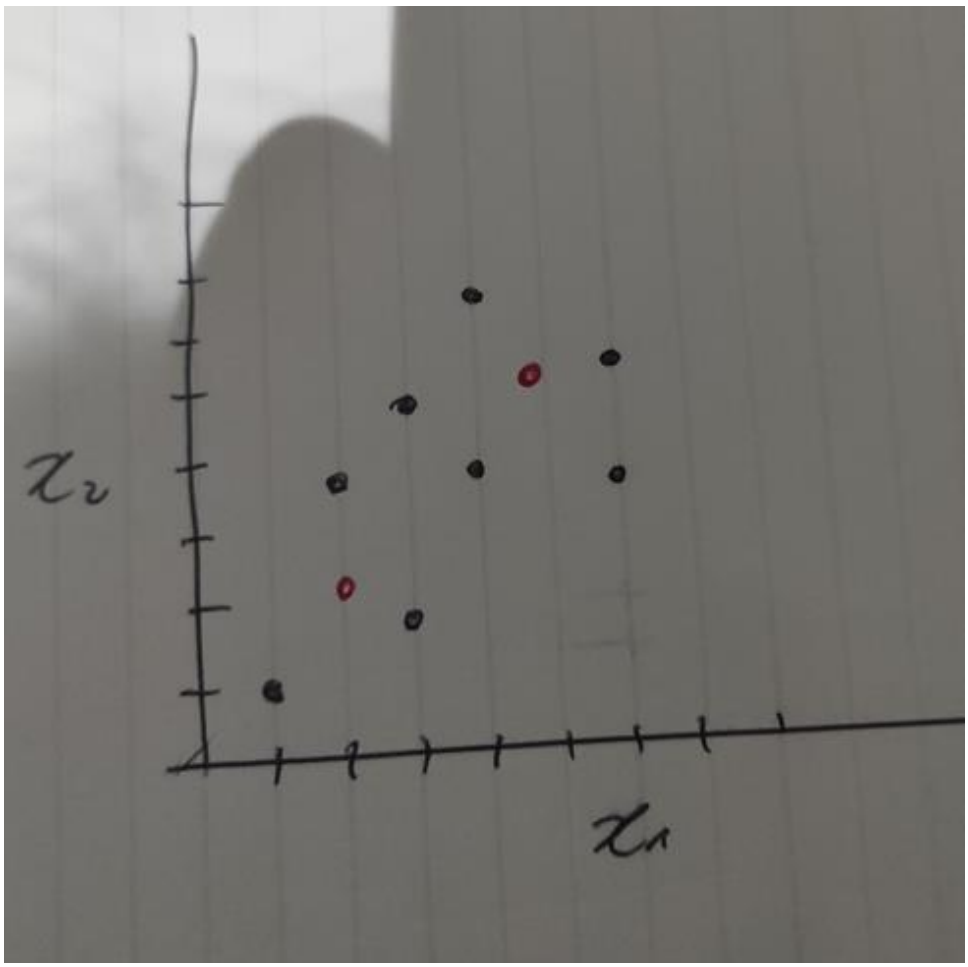
EJERCICIO 3

El metodo de creacionde clusters K Means es de tipo no supervisado, por lo cual por cada dato agregado se recalcula de nuevo el centroide.
Si queremos clasificar en 3 grupos, debemos configurar 3 centroides, por lo cual cada centroide se obtendra der cada conjunto de datos obtenido.

K-means funciona separando los datos en clusters, por lo cual, la inercia obtenida es en relacion a “la distancia” que se encuentra el centroide de los datos. Entre mas cerca esten los centroides a un conjunto de datos, la inercia baja poco a poco. Para determinar el valor optimo de categorias, se obtiene el “Codo” visualizado en un grafico.

Para crear los clusters, partimos de centroides, o bien separando regiones siguiendo el proceso:
Generar K centroides, los cuales determinaran cuantas categorias tendremos. Luego asignar a cada observacion(o muestra, o fila)
, el centroide mas cercano. Se calculan de nuevo los centroides por cada dato agregado dentro del cluster(Esto tiende a mover un poco la posicion del centroide, y aproximarlo a una region mas precisa), y por ultimo, se vuelven a generar centroides, asignandolos a cada observacion, hasta que las observaciones ya no varian del cluster.

Examen



A consecuencia del tiempo del examen, no me ha dado tiempo de graficar k-means, pero los datos quedan distribuidos de la manera en que se observa en la fotografia. Se podria decir que los centroides posiblemente estan aproximados o cercanos a donde se encuentran los puntos rojos

Se podria utilizar otro algoritmo como KNN, solo que en este caso es supervisado, y hay que tener en cuenta que depende de la seleccion de K, es la precision del algoritmo.

EJERCICIO 4

El codigo del ejercicio 4, he utilizado el codigo desarrollado en clase, con explicacion en el notebook y separandolo en el notebook. Sin embargo se ha desarrollado con las mismas variables utilizadas anteriormente de “housing” por el poco tiempo del examen. Se han desarrollado todos los pasos del notebook a pesar de no pedirse de la misma manera por el poco tiempo del examen. Gracias de antemano por su comprension al tema..

A) Si existe una relacion entre la variable peso y la variable volumen. El valor de correlacion obtenido es alto, cercano a 1, y el metodo empleado tiene como valor maximo 1, por lo cual se estima que la correlacion entre ambas variables es alta.

B) La variable Tapas puede influir significativamente en el peso, pero es mas significativo el volumen del libro hacia el peso correlativamente.

Código de examen: 175778

C) Ambas variables pueden ser buenos predictores, sin embargo en el caso de elegir un solo predictor, se optaria por una prediccion con el volumen del libro.
D) El modelo generado de regresion lineal multiple se ha hecho con los ejempos de clase, por lo cual las variables tienen el mismo nombre para ahorrar tiempo en el desarrollo.

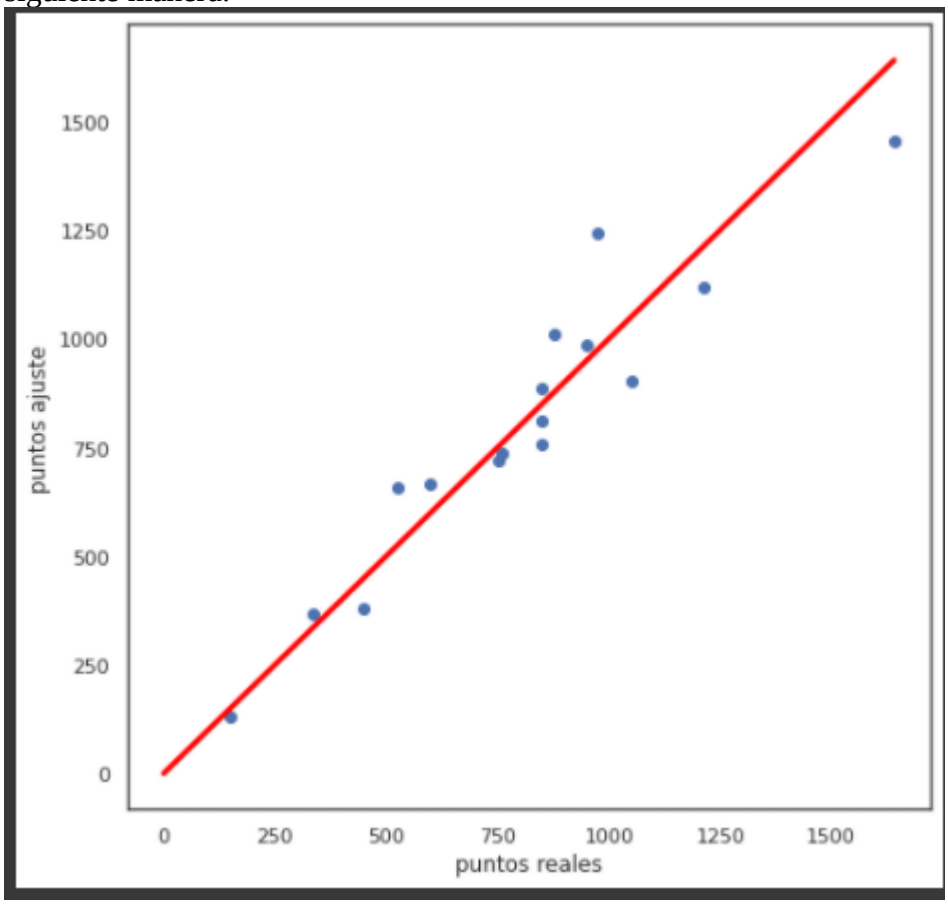
OLS Regression Results

Dep. Variable:	Peso	R-squared:	0.894
Model:	OLS	Adj. R-squared:	0.878
Method:	Least Squares	F-statistic:	54.99
Date:	Sat, 10 Jul 2021	Prob (F-statistic):	4.53e-07
Time:	00:38:54	Log-Likelihood:	-98.140
No. Observations:	16	AIC:	202.3
Df Residuals:	13	BIC:	204.6
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	711.5904	44.121	16.128	0.000	616.272	806.909
x1	328.2053	31.445	10.437	0.000	260.272	396.139
x2	179.4442	62.891	2.853	0.014	43.577	315.311

Omnibus:	1.996	Durbin-Watson:	1.738
Prob(Omnibus):	0.369	Jarque-Bera (JB):	0.880
Skew:	-0.571	Prob(JB):	0.644
Kurtosis:	3.126	Cond. No.	2.67

Se observa que la variable categorica tiene un p valor aceptado menor a 0.05, sin embargo es mucho mejor la variable cuantitativa del volumen relacionado al peso. La regresion Lineal ha quedado de la siguiente manera:



F) Si son utiles los predictores. En el grafico de la regresion lineal se observa que ningun dato tiene outliers, se puede decir que cerca del 1000 en puntos reales de los datos, se tiene algo parecido a un outlier, pero no esta muy alejado del valor predicho. En este caso, puede ser posible que la estadistica robusta no funcione bien, sin embargo, son muy pocos valores para asegurar eso.

G) Si es significativo

H) En caso de excluir valores atipicos, se estan dejando de considerar muestras, por lo cual si es diferente si se dejan de considerar. Puede no ser tan notoria la variacion si nos basamos en estadistica robusta.