APRENDIZAJE ESTADÍSTICO

Boletín 1: Evaluación y Selección de Modelos (KNN)

ANDRÉS CAMPOS CUIÑA

FECHA DE ENTREGA: 19/11/2021

ÍNDICE

1 EJERCICIO 1

Suponiendo que se quiere hacer la predicción de la variable de salida para X1=0, X2=0, X3=0 mediante KNN.

a. Computar la distancia entre cada observación y el punto de test.

Distancia entre el punto [0 3 2] y el punto de test [0 0 0] = 3.605551275463989

Distancia entre el punto [3 0 3] y el punto de test [0 0 0] = 4.242640687119285

Distancia entre el punto [0 3 -1] y el punto de test [0 0 0] = 3.1622776601683795

Distancia entre el punto [3 0 0] y el punto de test [0 0 0] = 3.0

Distancia entre el punto [1 2 1] y el punto de test [0 0 0] = 2.449489742783178

Distancia entre el punto [2 1 0] y el punto de test [0 0 0] = 2.23606797749979

b. ¿Cuál es la predicción para K=1? ¿Por qué?

La predicción para [0 0 0] es 0 y el 1 vecino más cercanos es:

El vecino [2 1 0] a una distancia de 2.23606797749979 con clase 0

c. ¿Cuál es la predicción para K=3? ¿Por qué?

La predicción para [0 0 0] es 1 y los 3 vecino más cercanos son:

El vecino [2 1 0] a una distancia de 2.23606797749979 con clase 0

El vecino [1 2 1] a una distancia de 2.449489742783178 con clase 1

El vecino [3 0 0] a una distancia de 3.0 con clase 1

2 EJERCICIO 2

Dado el problema de clasificación Blood Transfusion Service Center:

a. Analiza las características del conjunto de datos: número y tipo de variables de entrada y salida, número de instancias, número de clases y distribución de las mismas, correlación entre las variables, valores perdidos, etc.

	Recency	Frequency	Monetary	Time	IsMarchDonor
0	2	50	12500	98	1
1	0	13	3250	28	1
2	1	16	4000	35	1
3	2	20	5000	45	1
4	1	24	6000	77	0
...	...	...	...	...	...
743	23	2	500	38	0
744	21	2	500	52	0
745	23	3	750	62	0
746	39	1	250	39	0
747	72	1	250	72	0

748 rows × 5 columns

Tipo de dato de cada columna del Dataframe :

Recency: int64

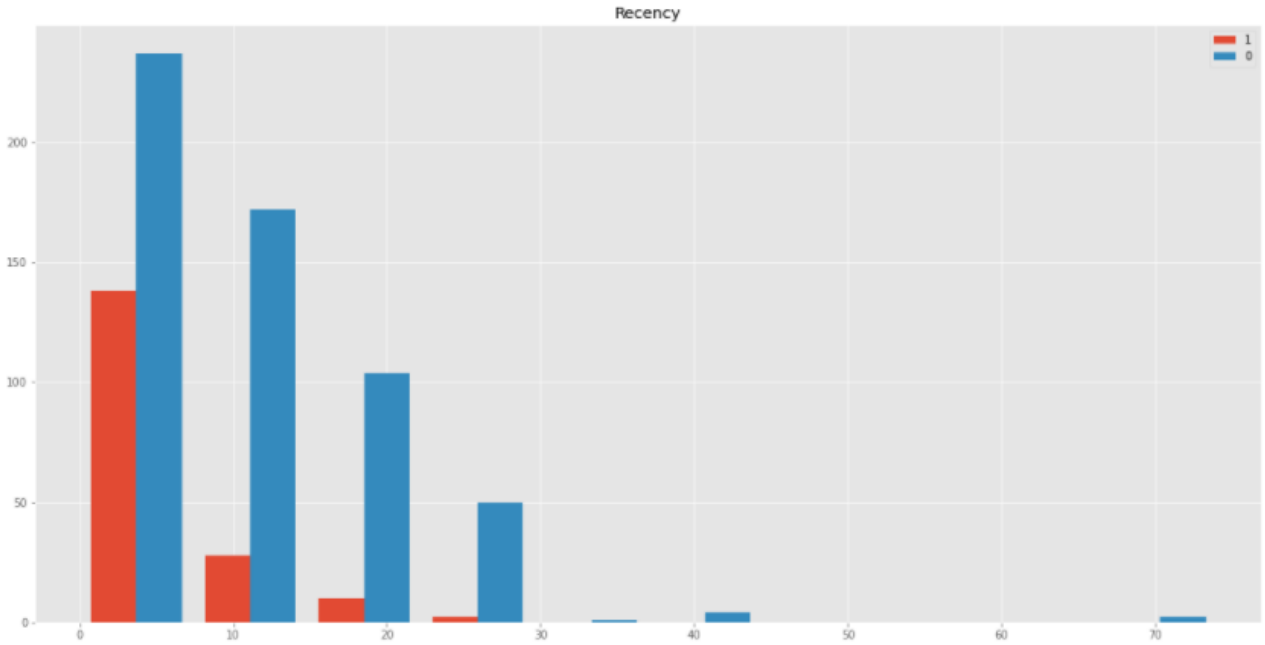
Frequency: int64

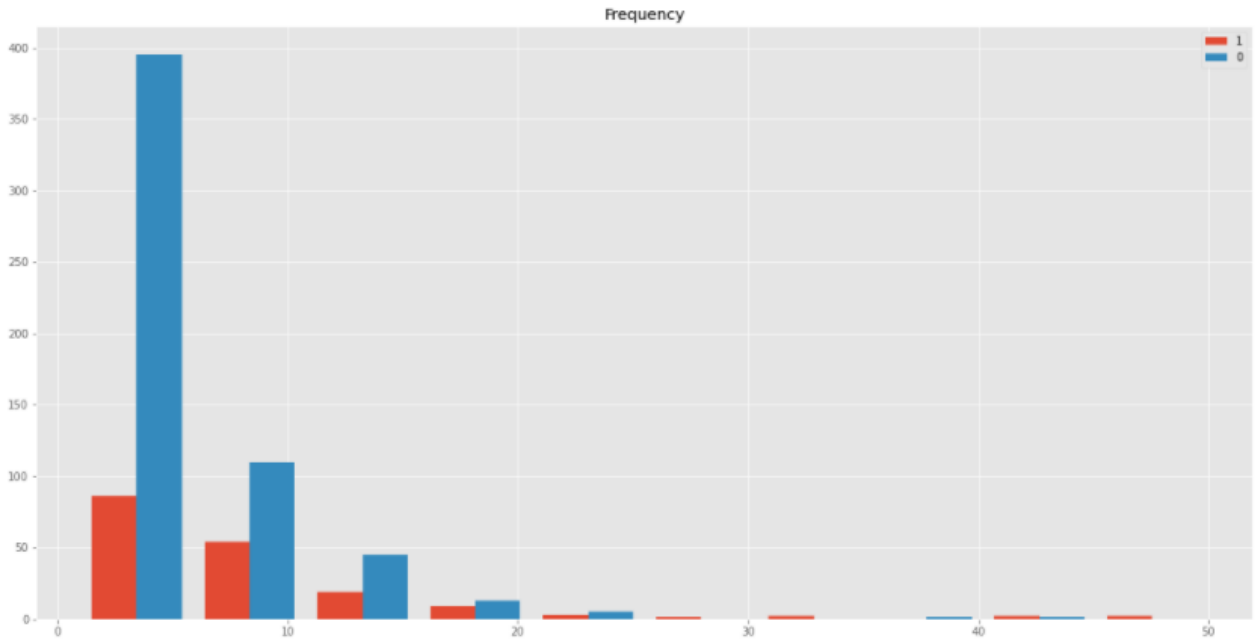
Monetary: int64

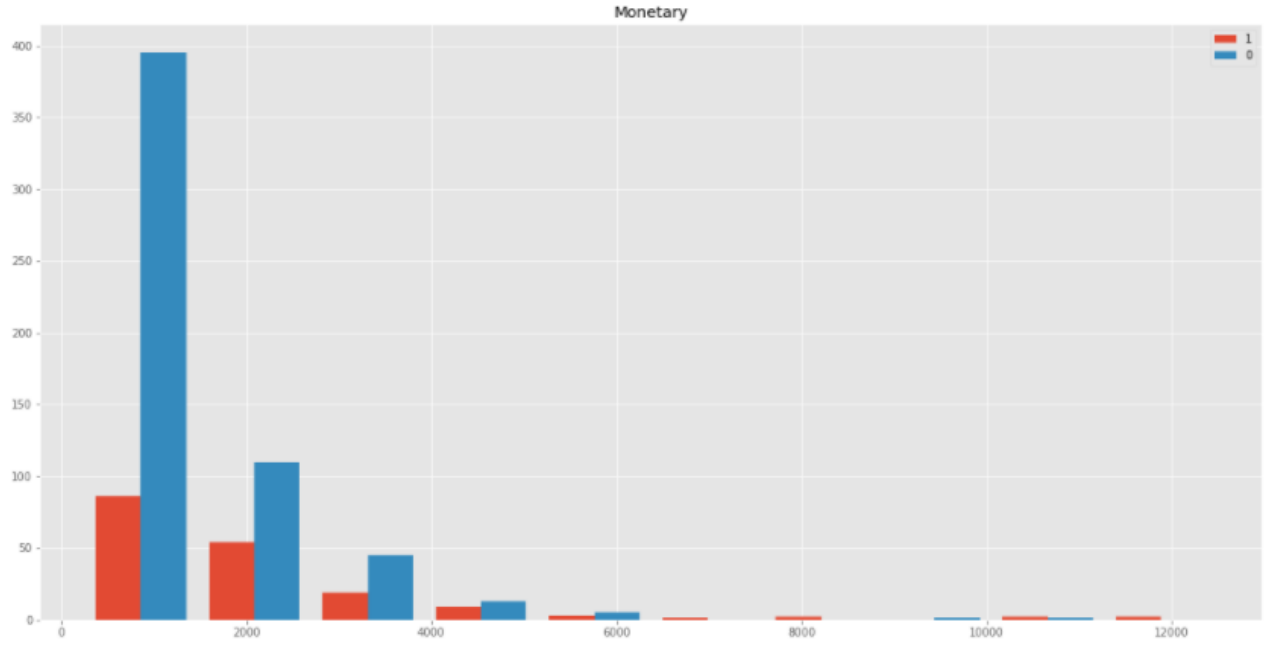
Time: int64

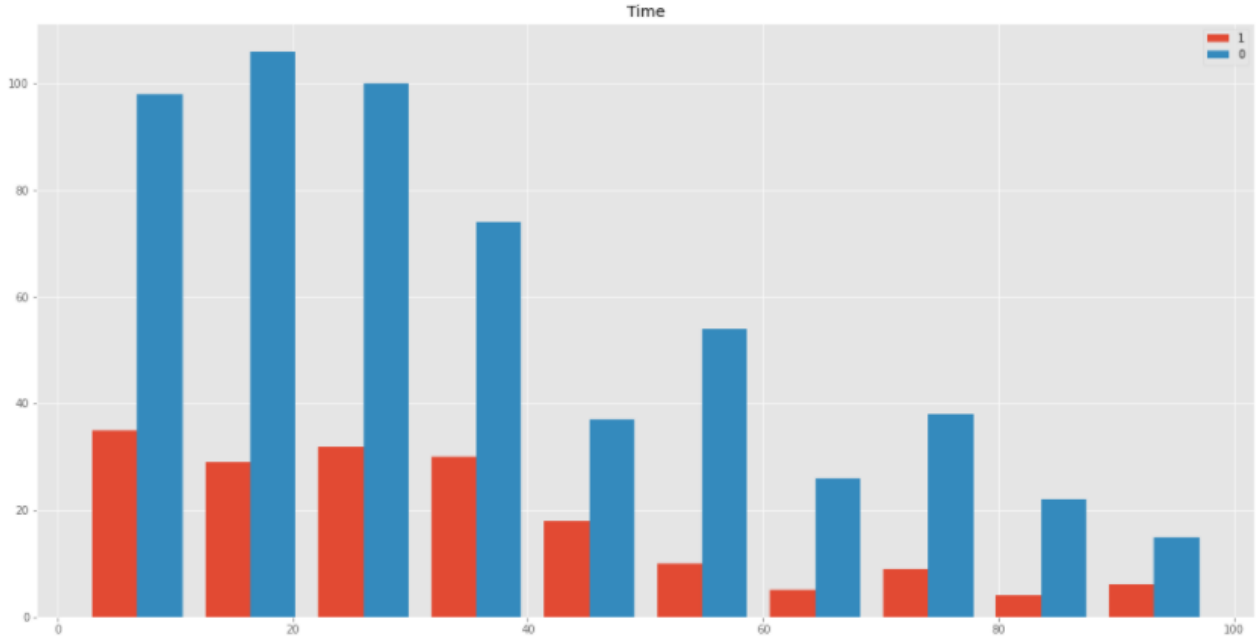
IsMarchDonor: int64

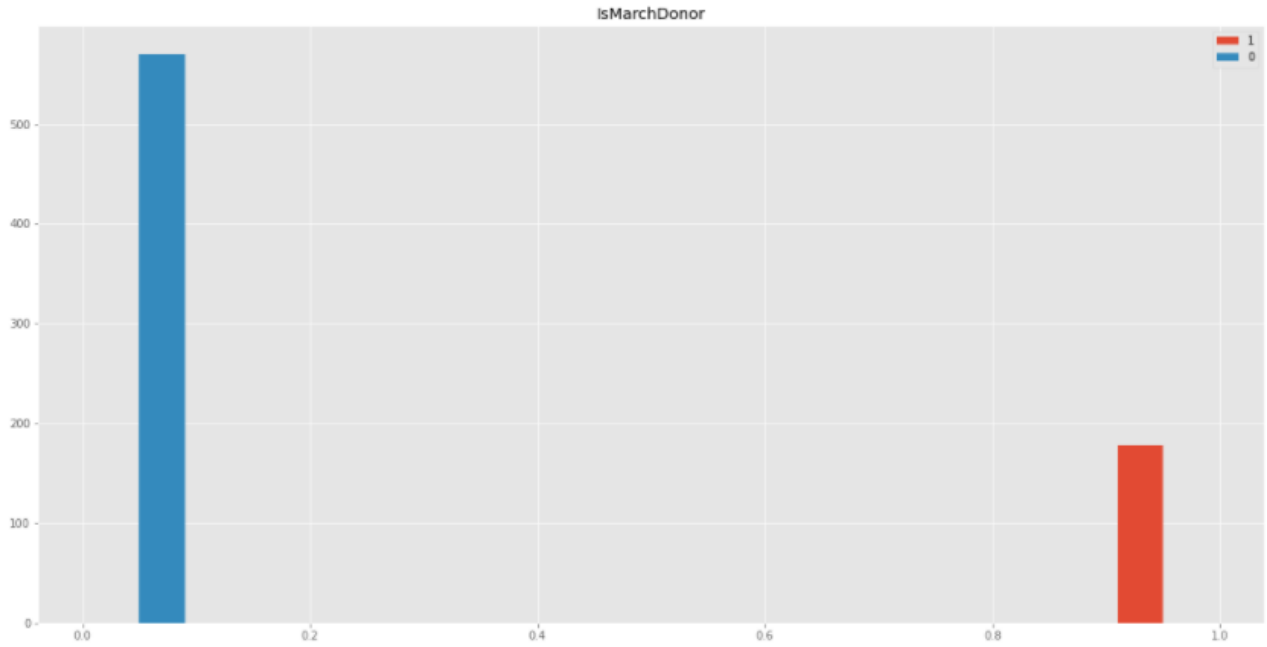
Número de clases y distribución de las mismas:



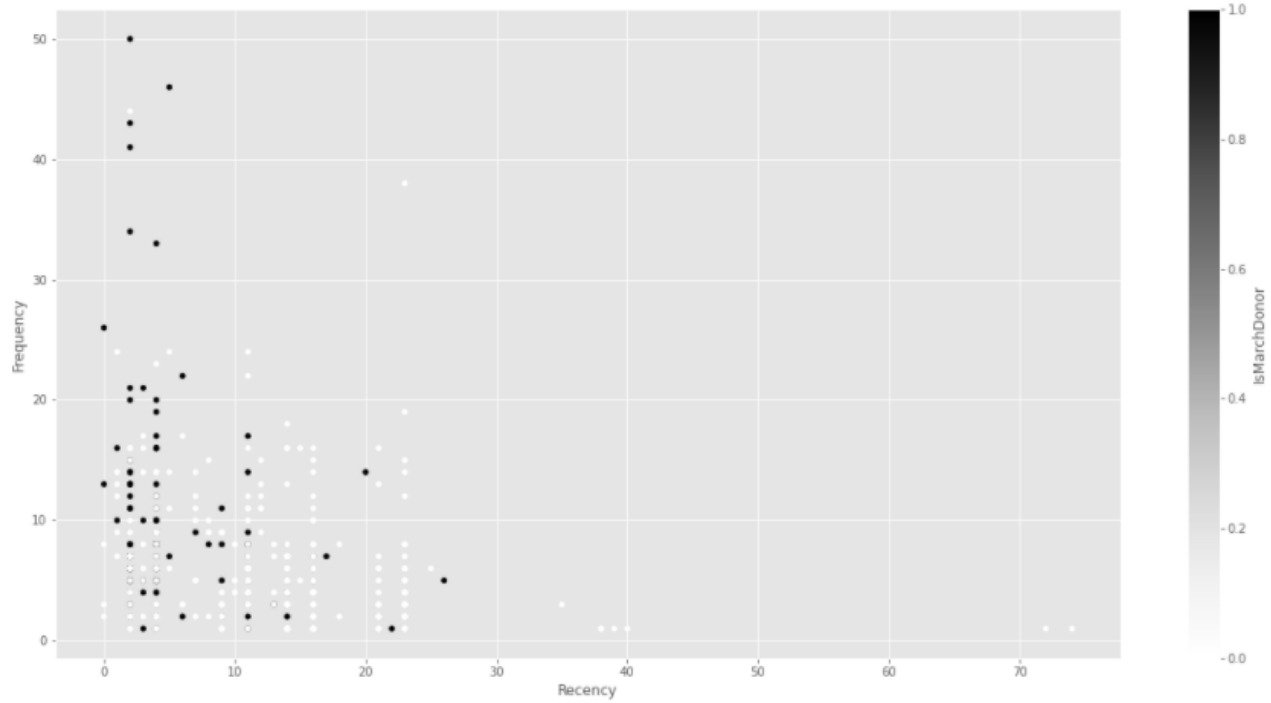


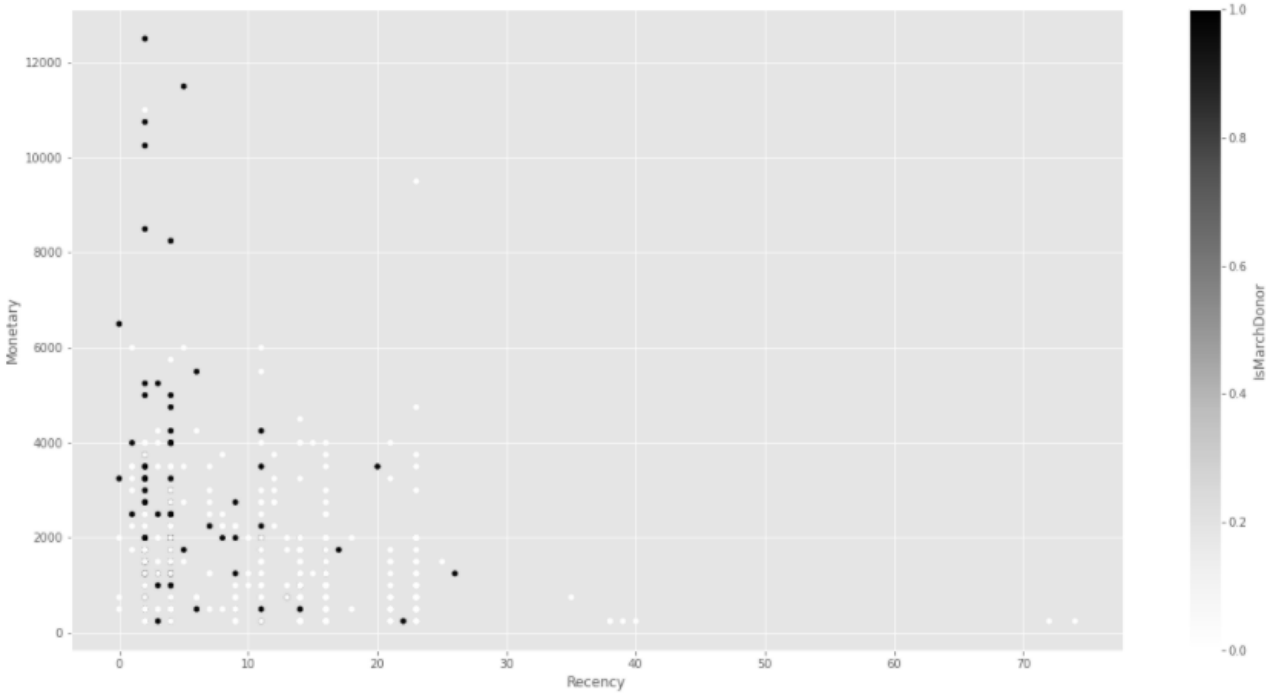

Monetary
1
0
400
350
300
250
200
150
100
50
0
0
2000
4000
6000
8000
10000
12000

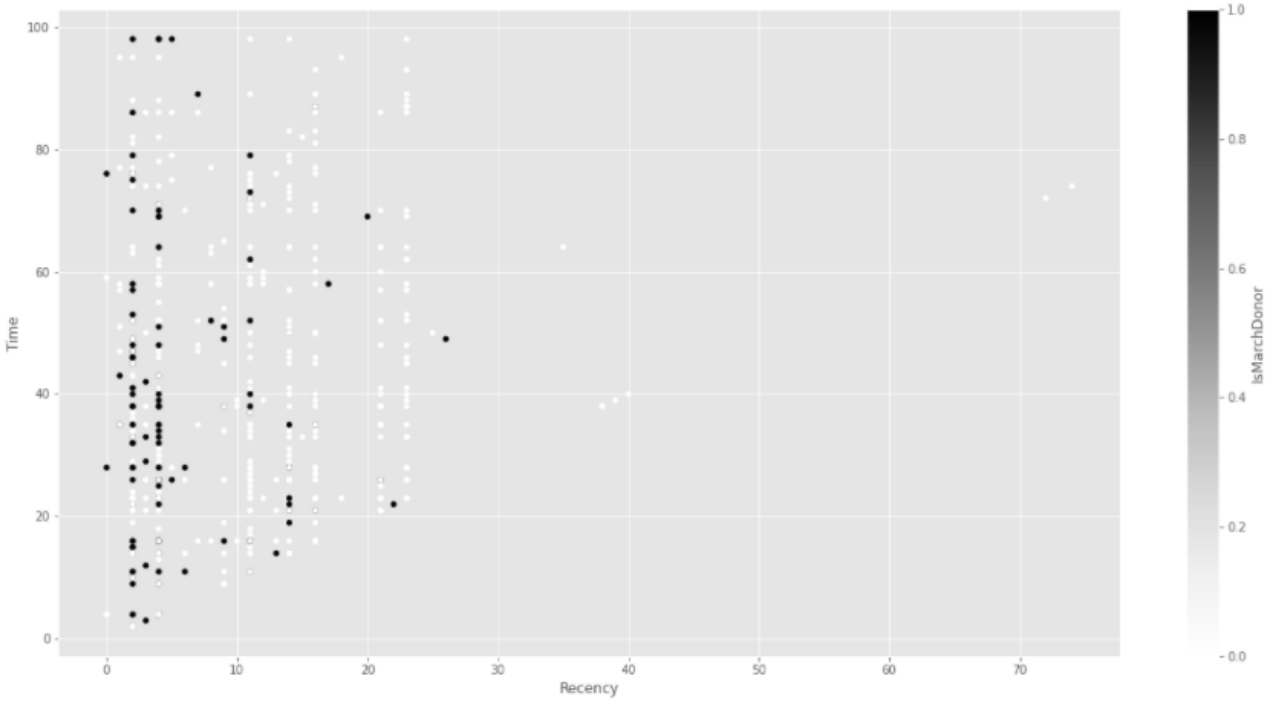

Time
1
0
100
80
60
40
20
0
0
20
40
60
80
100

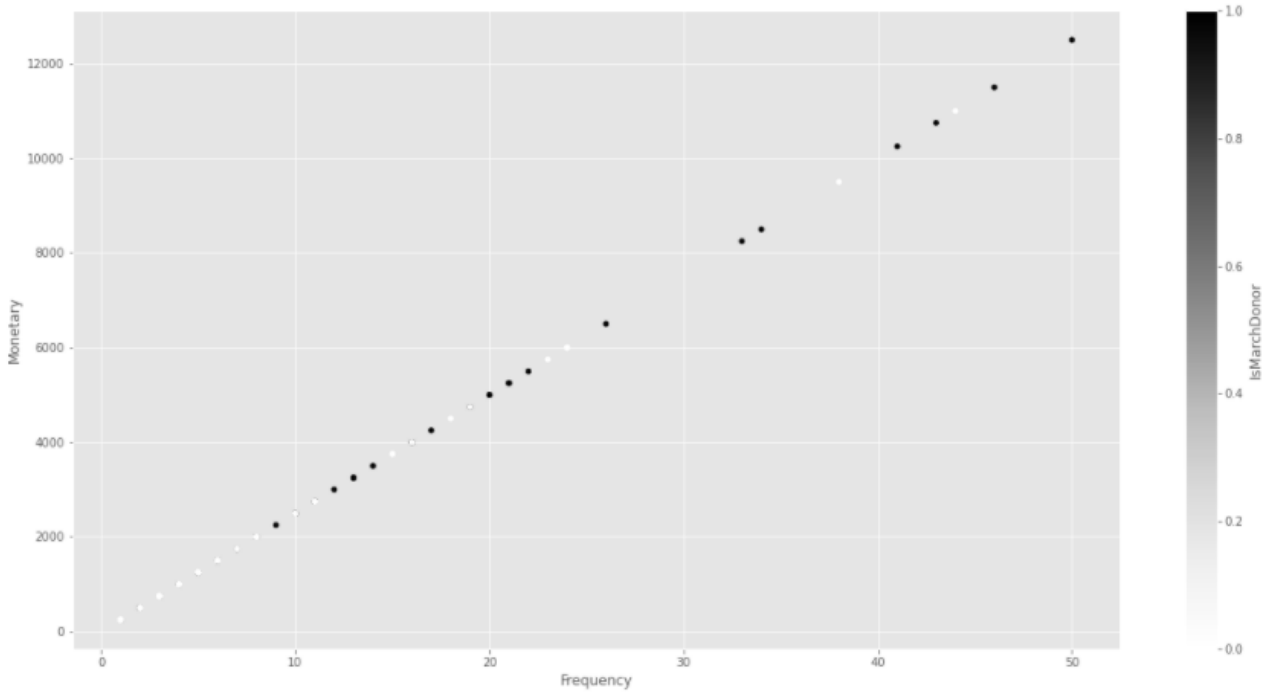


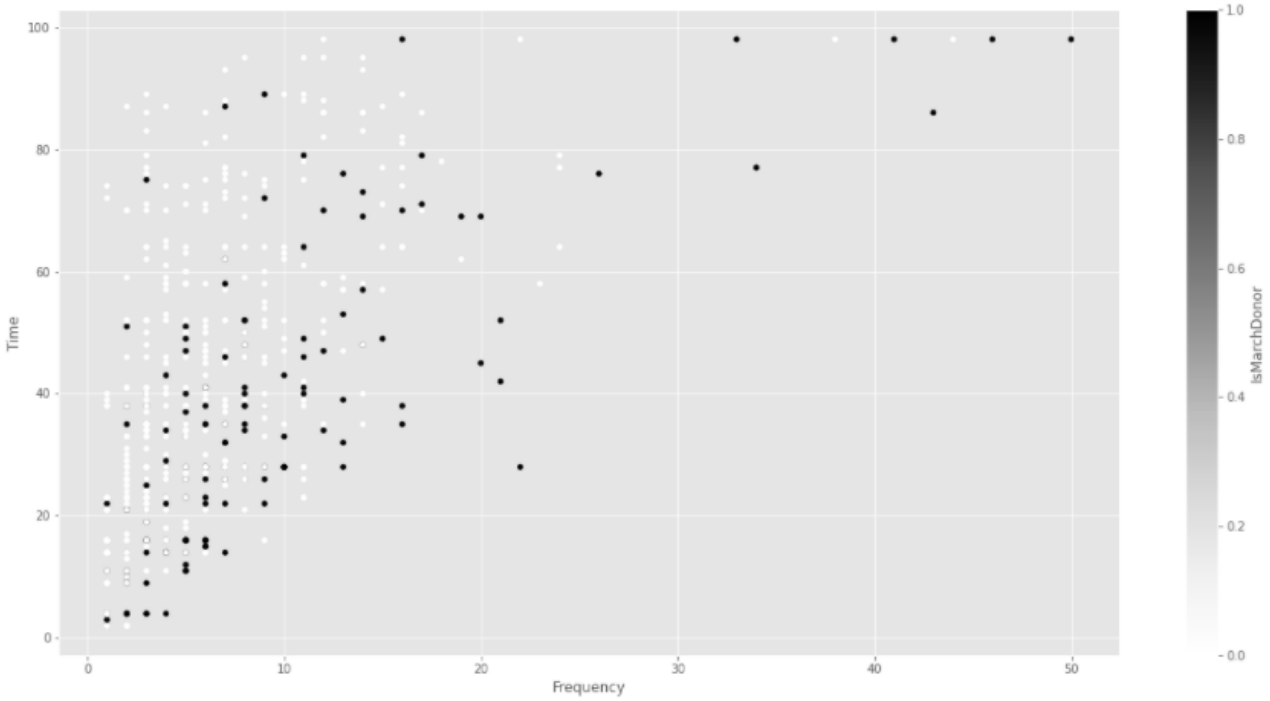
La correlación entre las variables:

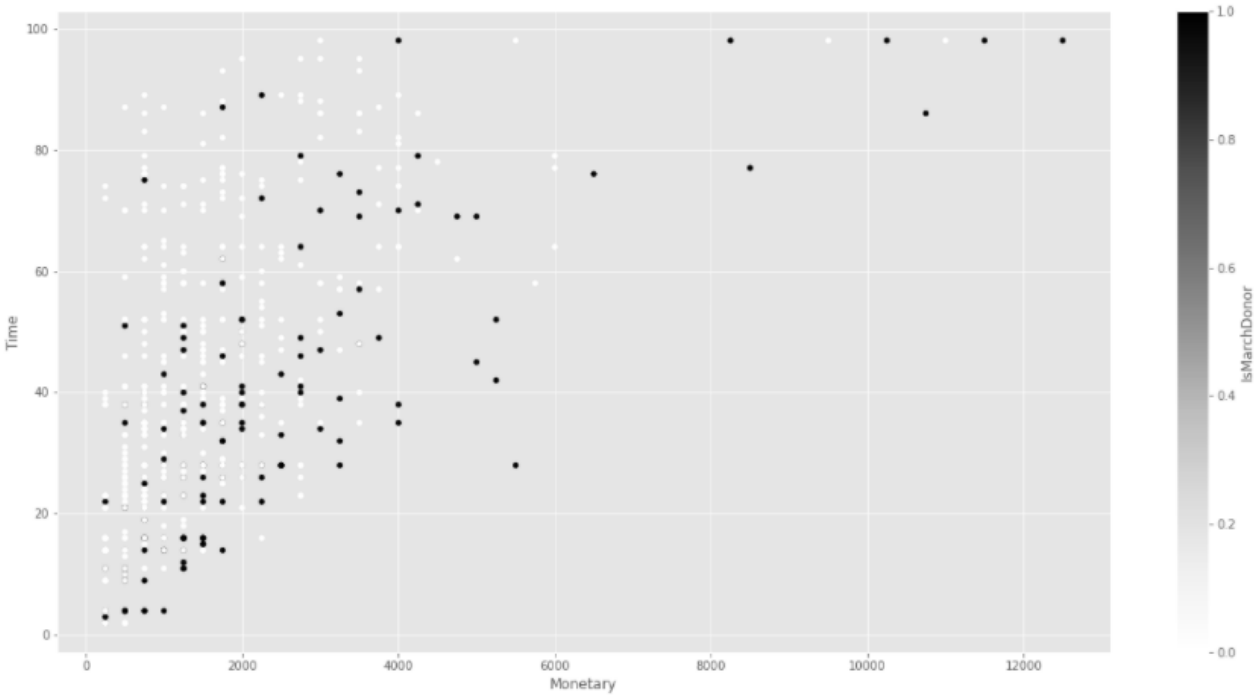



Monetary
Recency
IsMarchDonor


Time
Recency
IsMarchDonor


Monetary
Frequency
IsMarchDonor
0
2000
4000
6000
8000
10000
12000
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0


Time
Frequency
IsMarchDonor
0
20
40
60
80
100
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0



b. Una de las clases que implementa el algoritmo KNN en *scikit-learn* es *sklearn.neighbors.KNeighborsClassifier*. Revisa los parámetros y métodos que tiene.

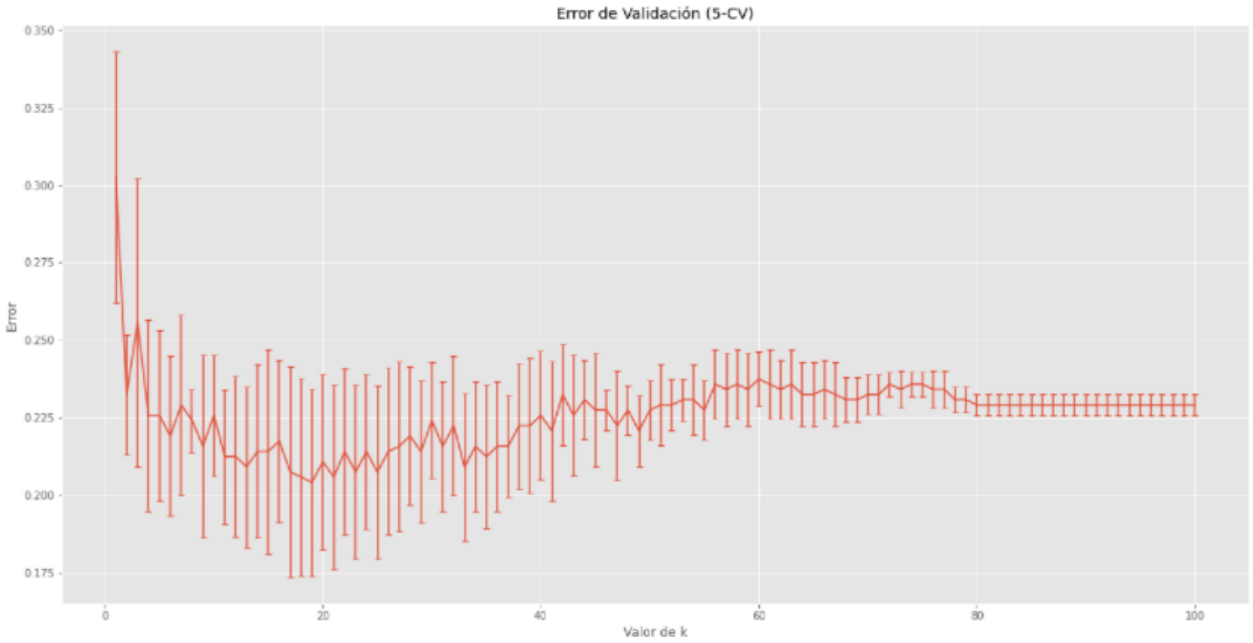
Revisado.

c. Divide los datos en entrenamiento (80%) y test (20%).

Hecho.

d. Realiza la experimentación con KNN (*KNeighborsClassifier*) usando como hiper-parámetro el número de vecinos.

Muestra la gráfica del error de entrenamiento con validación cruzada (5-CV) frente al valor del hiper-parámetro. ¿Cuál es el menor error de validación cruzada, su desviación estándar y el valor del hiper-parámetro para el que se consigue? ¿Cuál es el valor del hiperparámetro si se aplicase la regla de una desviación estándar?



Menor error de validación cruzada, su desviación estándar y el valor del hiper-parámetro para el que se consigue:

	param_n_neighbors	mean_test_score	std_test_score	rank_test_score
18	19	0.79591	0.030314	1

El error real es: 0.2040896358543417

El valor del hiper-parámetro si se aplicase la regla de una desviación estándar:

param_n_neighbors: 55

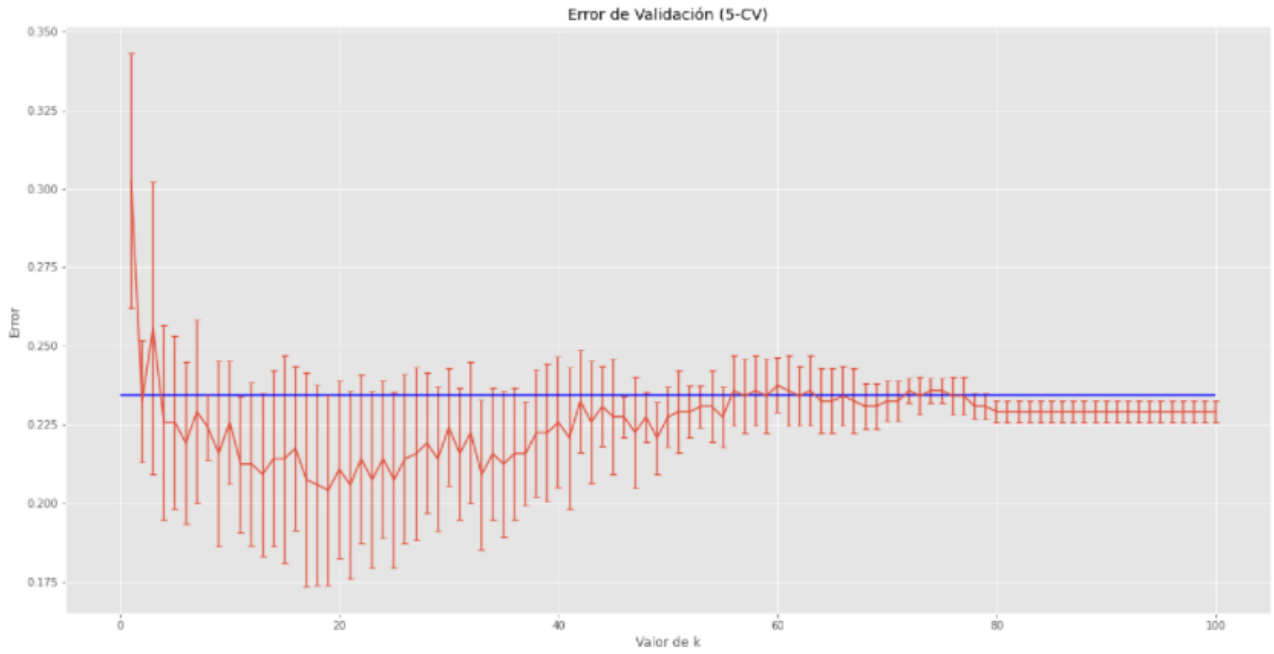
mean_test_score: 0.772577

std_test_score: 0.009657

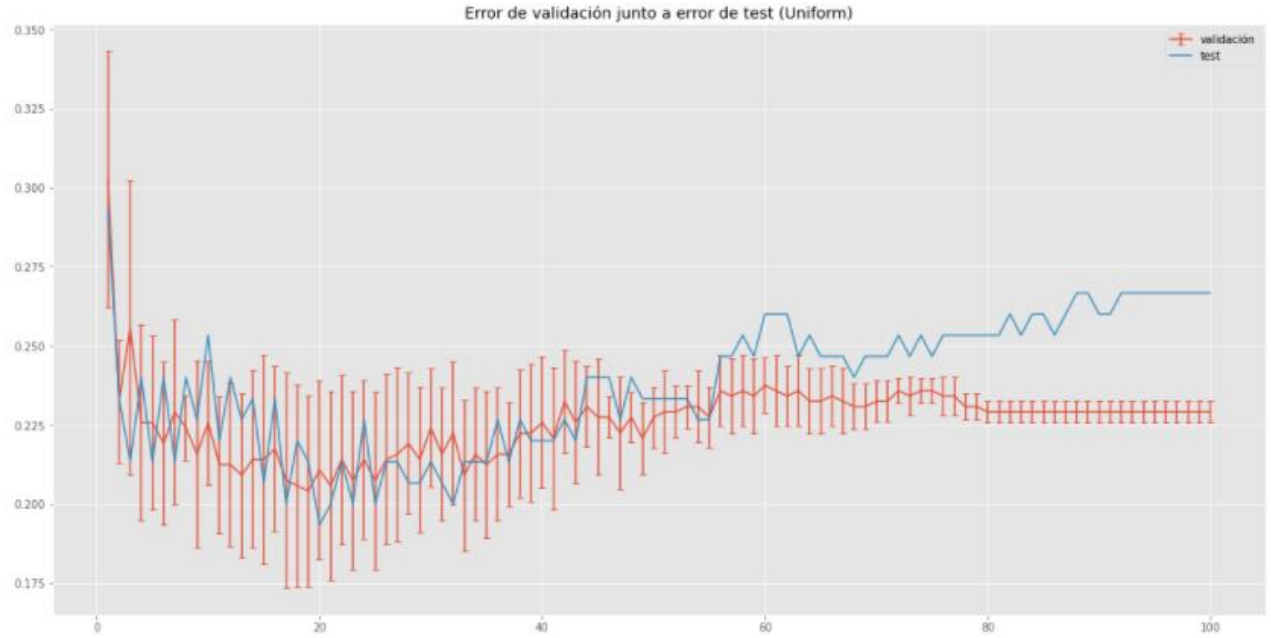
rank_test_score: 41

Error real: 0.22742296918767513

La gráfica de la selección de este valor es la siguiente:



Muestra la gráfica del error de test frente al valor del hiper-parámetro, y valora si la gráfica del error de entrenamiento con validación cruzada ha hecho una buena estimación del error de test. ¿Cuál es el menor error de test y el valor del hiper-parámetro para el que se consigue? ¿Cuál es el error de test para el valor del hiper-parámetro seleccionado por la validación cruzada? ¿Cuál es el error de test para el valor del hiper-parámetro seleccionado por la validación cruzada mediante la regla de una desviación estándar?



El menor error de test y el valor del hiper-parámetro para el que se consigue:

	param_n_neighbors	mean_test_score	std_test_score	rank_test_score
19	20	0.806667	0.0	1

Error real: 0.19333333333333336

El error de test para el valor del hiper-parámetro seleccionado por la validación cruzada:

0.21333333333333337

El error de test para el valor del hiper-parámetro seleccionado por la validación cruzada mediante la regla de una desviación estándar:

0.22666666666666668

3 EJERCICIO 3

Dado el problema de regresión Energy Efficiency:

a. Analiza las características del conjunto de datos: número y tipo de variables de entrada y salida, número de instancias, número de clases y distribución de las mismas, correlación entre las variables, valores perdidos, etc.

	X1	X2	X3	X4	X5	X6	X7	X8	Y
0	0.98	514.5	294.0	110.25	7.0	2	0.0	0	21.33
1	0.98	514.5	294.0	110.25	7.0	3	0.0	0	21.33
2	0.98	514.5	294.0	110.25	7.0	4	0.0	0	21.33
3	0.98	514.5	294.0	110.25	7.0	5	0.0	0	21.33
4	0.90	563.5	318.5	122.50	7.0	2	0.0	0	28.28
...	...	...	...	...	...	...	...	...	...
763	0.64	784.0	343.0	220.50	3.5	5	0.4	5	21.40
764	0.62	808.5	367.5	220.50	3.5	2	0.4	5	16.88
765	0.62	808.5	367.5	220.50	3.5	3	0.4	5	17.11
766	0.62	808.5	367.5	220.50	3.5	4	0.4	5	16.61
767	0.62	808.5	367.5	220.50	3.5	5	0.4	5	16.03

768 rows × 9 columns

Tipo de dato de cada columna del Dataframe :

X1 float64

X2 float64

X3 float64

X4 float64

X5 float64

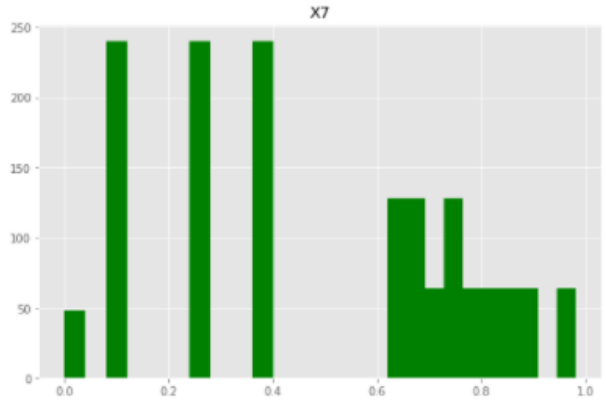
X6 int64

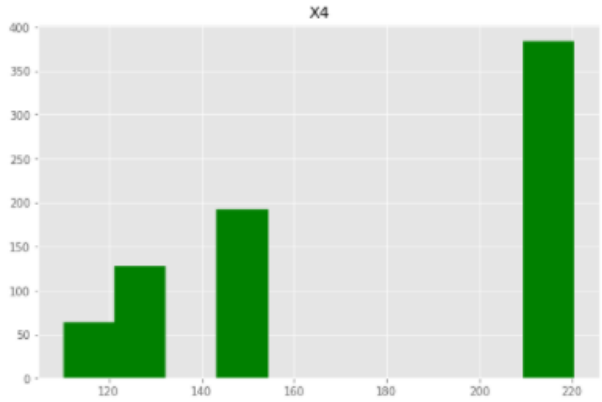
X7 float64

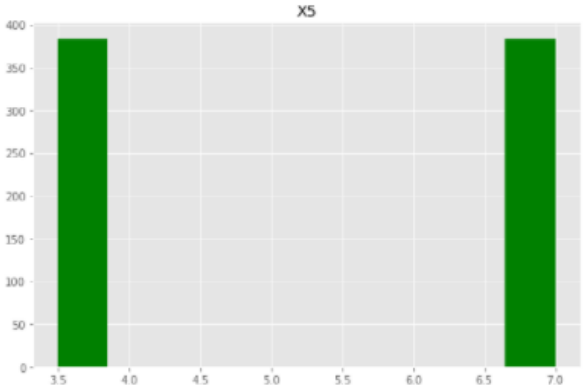
X8 int64

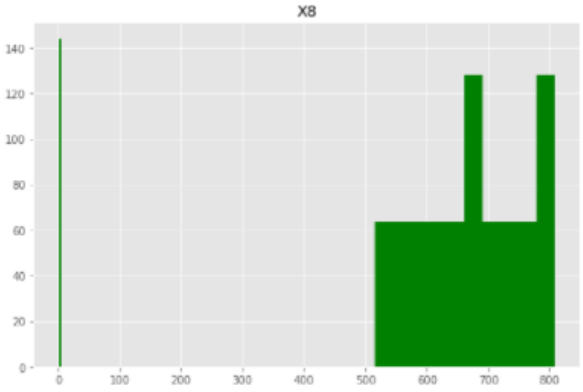
Y float64

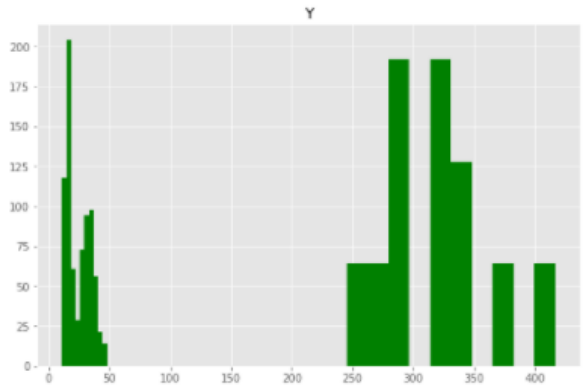
Número de clases y distribución de las mismas:

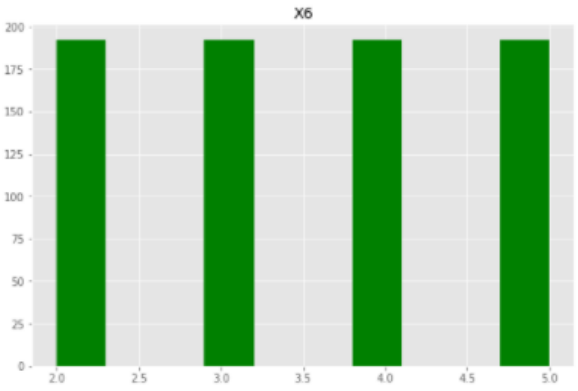




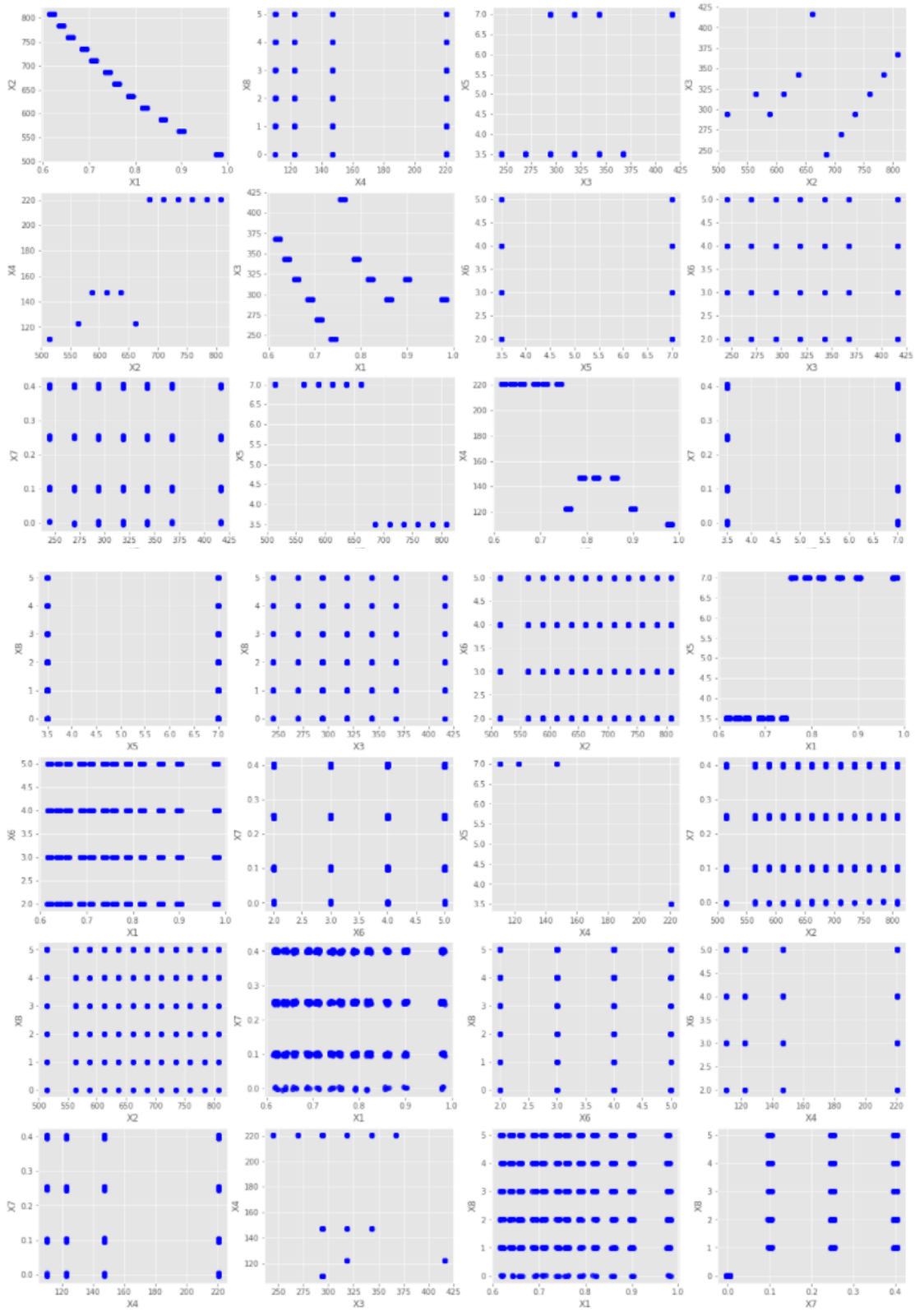








La correlación entre las variables:



b. Una de las clases que implementa el algoritmo KNN en *scikit-learn* es *sklearn.neighbors.KNeighborsRegressor*. Revisa los parámetros y métodos que tiene.

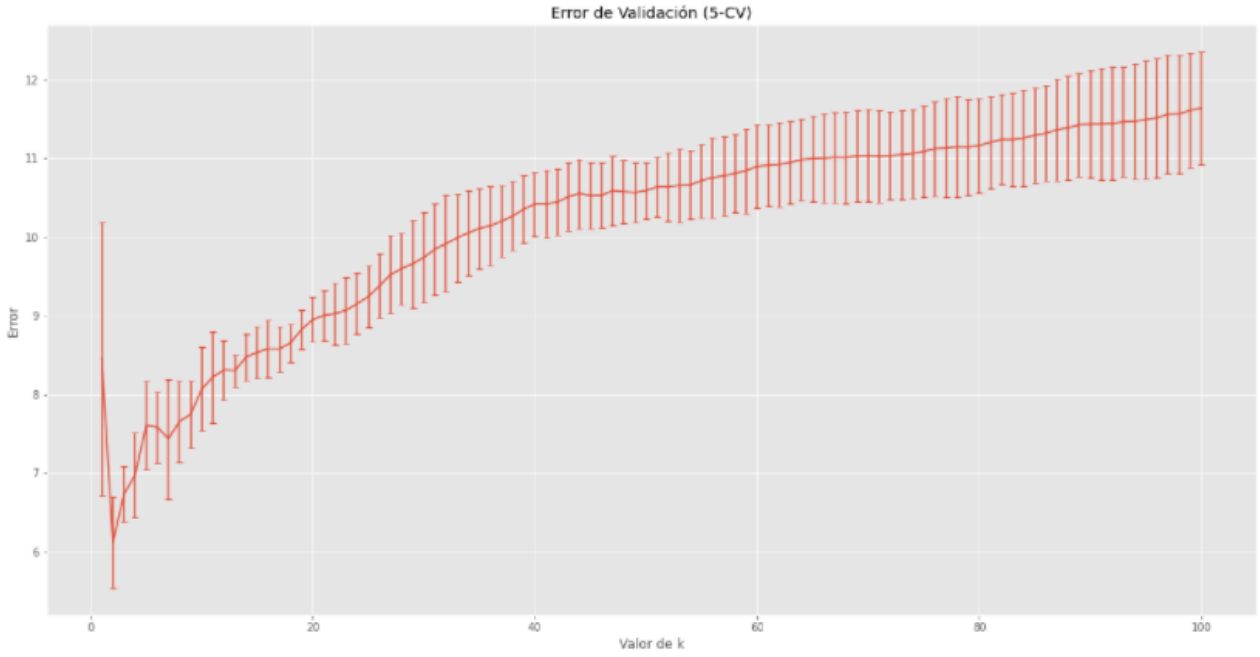
Revisado.

c. Divide los datos en entrenamiento (80%) y test (20%).

Hecho.

d. Realiza la experimentación con KNN (*KNeighborsRegressor*) usando como hiper-parámetro el número de vecinos.

Muestra la gráfica del error de entrenamiento con validación cruzada (5-CV) frente al valor del hiper-parámetro. ¿Cuál es el menor error de validación cruzada, su desviación estándar y el valor del hiper-parámetro para el que se consigue? ¿Cuál es el valor del hiperparámetro si se aplicase la regla de una desviación estándar?



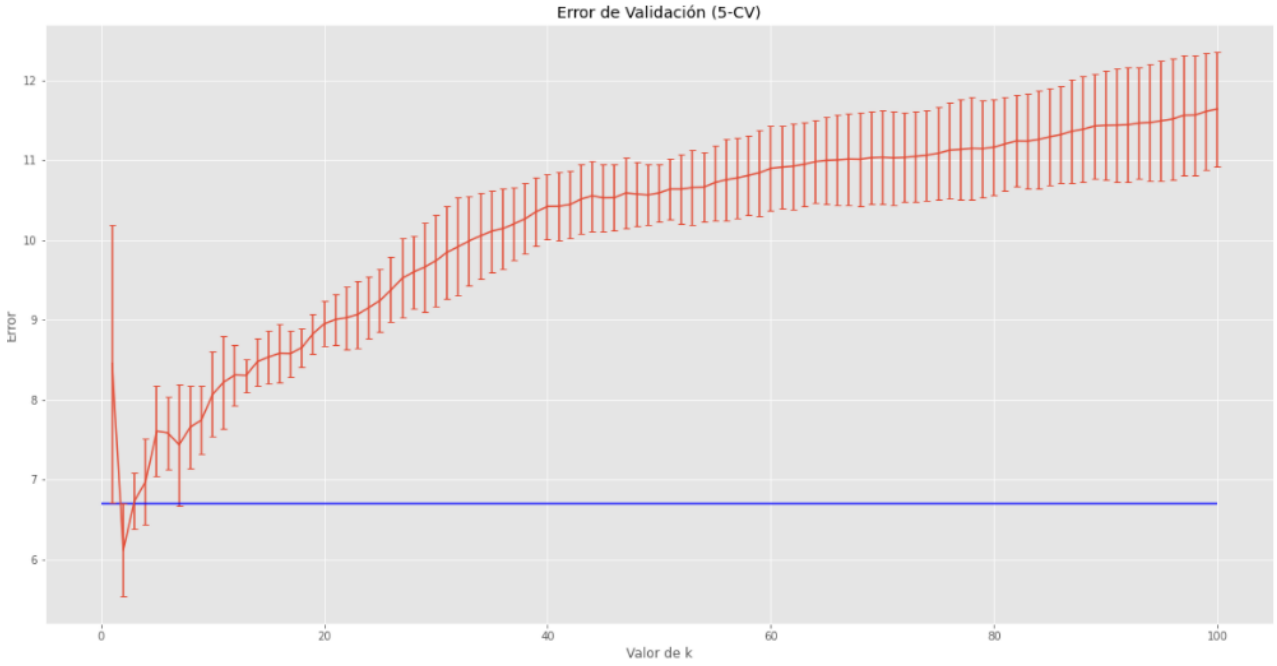
Menor error de validación cruzada, su desviación estándar y el valor del hiper-parámetro para el que se consigue:

	param_n_neighbors	mean_test_score	std_test_score	destandardized_mean_test_score	destandardized_std_test_score	rank_test_score
1	2	-0.068429	0.006466	6.115019	0.577814	1

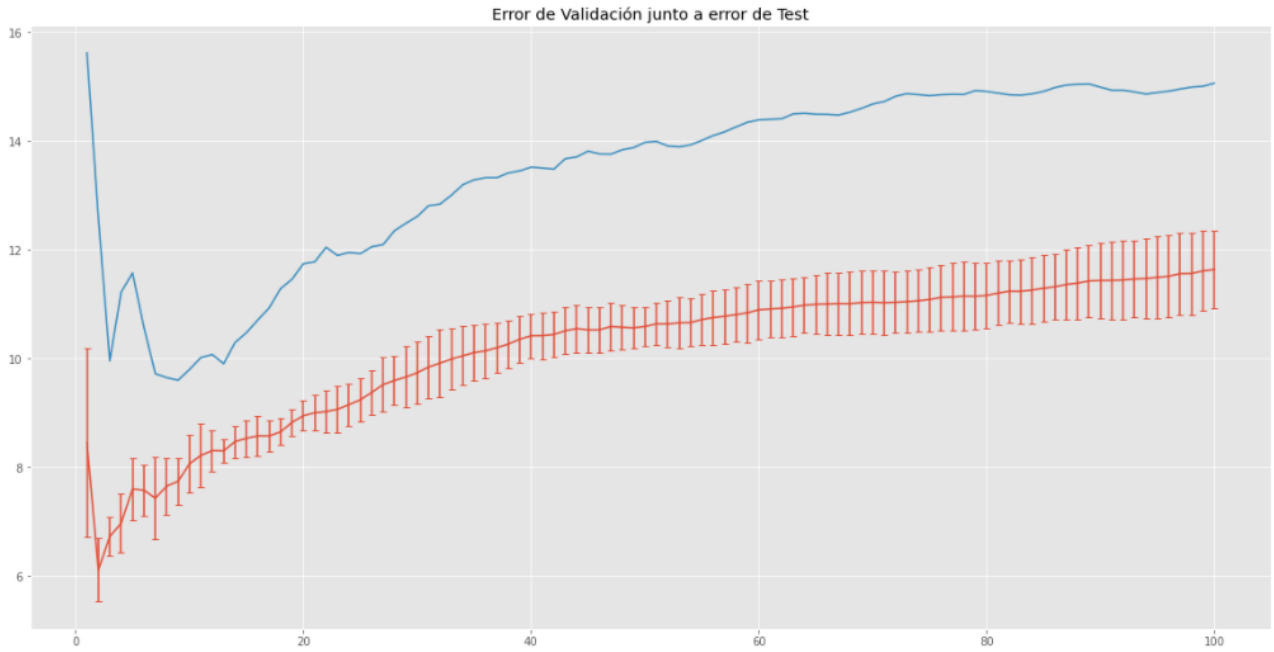
El valor del hiper-parámetro si se aplicase la regla de una desviación estándar:

	param_n_neighbors	mean_test_score	std_test_score	destandardized_mean_test_score	destandardized_std_test_score	rank_test_score
1	2	-0.068429	0.006466	6.115019	0.577814	1

La gráfica de la selección de este valor es la siguiente:



Muestra la gráfica del error de test frente al valor del hiper-parámetro, y valora si la gráfica del error de entrenamiento con validación cruzada ha hecho una buena estimación del error de test. ¿Cuál es el menor error de test y el valor del hiper-parámetro para el que se consigue? ¿Cuál es el error de test para el valor del hiper-parámetro seleccionado por la validación cruzada? ¿Cuál es el error de test para el valor del hiper-parámetro seleccionado por la validación cruzada mediante la regla de una desviación estándar?



El menor error de test y el valor del hiper-parámetro para el que se consigue:

	param_n_neighbors	mean_test_score	std_test_score	destandardized_mean_test_score	destandardized_std_test_score	rank_test_score
8	9	-0.107463	0.0	9.603207	0.0	1

El error de test para el valor del hiper-parámetro seleccionado por la validación cruzada:

12.588656331168831

El error de test para el valor del hiper-parámetro seleccionado por la validación cruzada mediante la regla de una desviación estándar:

12.588656331168831