Curso de Big Data

Trabajo Practico 3

Integrantes del grupo: Sebastián Einstoss, Sofía Marinkovic y Pilar Ruiz Orrico

Parte 1: Análisis de la base de hogares y cálculo de pobreza

Ejercicio 1

Algunas variables de la base de hogar que consideramos que serían muy útiles para predecir la pobreza por ingresos y perfeccionar el tp2 serían las relacionadas con las cuestiones estructurales del hogar. Algunos ejemplos son:

-acceso a agua (IV6)

-tipo de suelo del hogar (IV3)

-si tiene o no baño en el hogar (IV8).

Existe cierta evidencia en la literatura que la pobreza estructural y la pobreza por ingreso tienen cierta correlacion en el largo plazo, por eso la incorporacion.

Ejercicio 2

	CODUSU	ANO4	TRIMESTRE	NRO_HOGAR	REALIZADA	REGION	MAS_500
0	TQRMNOQUPHMKKUCDEOJAH00701956	2021	1	1	1	44	N
1	TQRMNOPWXHMOKRCDEGLDF00701361	2021	1	1	1	41	N

2 rows × 88 columns

	CODUSU	ANO4	TRIMESTRE	NRO_HOGAR	REALIZADA	REGION	MAS_500
15	TQRMNOQXUHJMMUCDEIJAH00693031	2021	1	1	1	1	S
21	TQRMNOSRPHLLNRCDEIJAH00651171	2021	1	1	1	1	S
24	TQRMNOPVUHKLMNCDEIIAD00655817	2021	1	1	1	1	S
36	TQRMNORYTHLOPMCDEIJAH00655933	2021	1	1	1	1	S
43	TQRMNOQPVHLMNNCDEIIAD00650542	2021	1	1	1	1	S

5 rows × 88 columns

	CODUSU	ANO4	TRIMESTRE	NRO_HOGAR	COMPONENTE	H15	REGION	M
0	TQRMNOPPQHLKLMCDEFIAH00646702	2021	1	1	1	1	43	
1	TQRMNOPPRHLKLMCDEFIAH00665104	2021	1	1	1	1	43	

2 rows × 177 columns

	CODUSU	ANO4	TRIMESTRE	NRO_HOGAR	COMPONENTE	H15	REGION
36799	TQRMNOPPRHKMLNCDEIIAD00655703	2021	1	1	1	1	
36800	TQRMNOPPRHKMLNCDEIIAD00655703	2021	1	1	2	1	
36801	TQRMNOPPWHKMNUCDEIIAD00655837	2021	1	1	1	1	
36802	TQRMNOPPWHKMNUCDEIIAD00655837	2021	1	1	2	1	
36803	TQRMNOPPWHKMNUCDEIIAD00655837	2021	1	1	3	1	

5 rows × 177 columns

Ejercicio 3

```
Longitud base hogares: 1474 ;Longitud base individual: 4082
```

```
4082
```

	CODUSU	ANO4	TRIMESTRE	NRO_HOGAR	COMPONENTE	H15	REGION	M
0	TQRMNOPPRHKMLNCDEIIAD00655703	2021	1	1	1	1	1	
1	TQRMNOPPRHKMLNCDEIIAD00655703	2021	1	1	2	1	1	
2	TQRMNOPPWHKMNUCDEIIAD00655837	2021	1	1	1	1	1	
3	TQRMNOPPWHKMNUCDEIIAD00655837	2021	1	1	2	1	1	
4	TQRMNOPPWHKMNUCDEIIAD00655837	2021	1	1	3	1	1	

5 rows × 242 columns

Ejercicio 4

Funciones para limpiar la base de datos

Para limpiar la base de datos de observaciones sin sentido vamos a usar la funcion "drop()" de pandas. Para ello vamos a ponerle diferentes condiciones, si no las cumple se dropea la fila.

Asimismo para rellenar los missing values usaremos de pandas la funcion "fillna()" y completaremos con lo que consieremos acrode (Por ejemplo la mediana).

Para convertir variables categoricas en dummies usaremos del paquete numpy la funcion "where()".

Por último vamos a eliminar las variables sin sentido para el análisis. Para ello usaremos nuevamente la función "drop()" de pandas.

Ejercicio 5

En este inciso filtramos aquellos valores que no tiene sentido, como por ejemplo: Edades negativas, variables de ingreso que tomen valores negatvos.

También reemplazamos los valores faltantes por la mediana y dropeamos variablees que no consideramos relevantes. En el código se aclara cuales son.

Finalmente para realizar el analisis convertimos las variables categoricas en dummies.

	CODUSU	ANO4	TRIMESTRE	NRO_HOGAR	COMPONENTE	H15	REGION	M
0	TQRMNOPPRHKMLNCDEIIAD00655703	2021	1	1	1	1	1	
1	TQRMNOPPRHKMLNCDEIIAD00655703	2021	1	1	2	1	1	

2 rows × 242 columns

```
3308
```

	CODUSU	ANO4	TRIMESTRE	NRO_HOGAR	COMPONENTE	H15	REGION	M
0	TQRMNOPPRHKMLNCDEIIAD00655703	2021	1	1	1	1	1	
1	TQRMNOPPRHKMLNCDEIIAD00655703	2021	1	1	2	1	1	
3	TQRMNOPPWHKMNUCDEIIAD00655837	2021	1	1	2	1	1	
4	TQRMNOPPWHKMNUCDEIIAD00655837	2021	1	1	3	1	1	
5	TQRMNOPPWHKMNUCDEIIAD00655837	2021	1	1	4	1	1	

5 rows × 242 columns

	CODUSU	NRO_HOGAR	REGION	MAS_500	AGLOMERADO	CH04	CH06	PP0
0	TQRMNOPPRHKMLNCDEIIAD00655703	1	1	1	32	1	29	
1	TQRMNOPPRHKMLNCDEIIAD00655703	1	1	1	32	2	29	

2 rows × 962 columns

Ejercicio 6

5 Variables relevantes de la encuesta de hogares

IV1['4'] = Tipo de vivienda, pieza en hotel / pensión

IV8 = Baño/letrina

V5 = Inresos de subsidio o ayuda social(en dinero)del gobierno, iglesias

IV3['3'] = Tipo de suelo, ladrillo suelto / tierra

IV6['2'] = Acceso al agua, fuera de la vivienda pero dentro del terreno

```
              IV1_4          IV8           V5        IV3_3        IV6_2
count   3308.000000  3308.000000  3308.000000  3308.000000  3308.000000
mean       0.003023     0.998186     0.189541     0.002116     0.013603
std        0.054907     0.042556     0.391997     0.045959     0.115855
min        0.000000     0.000000     0.000000     0.000000     0.000000
25%        0.000000     1.000000     0.000000     0.000000     0.000000
50%        0.000000     1.000000     0.000000     0.000000     0.000000
75%        0.000000     1.000000     0.000000     0.000000     0.000000
max        1.000000     1.000000     1.000000     1.000000     1.000000
```

Ejercicio 7

	Tabla de equivalencias de necesidades energéticas. Unidades de adulto equivalente, según sexo y edad	Unnamed: 1	Unnamed: 2	Unnamed: 3
0	NaN	NaN	NaN	NaN
1	NaN	NaN	NaN	NaN
2	Edad	Mujeres	Varones	NaN
3	NaN	NaN	NaN	NaN
4	Menor de 1 año	0.35	0.35	NaN

	Tabla de equivalencias de necesidades energéticas. Unidades de adulto equivalente, según sexo y edad	Unnamed: 1	Unnamed: 2
0	NaN	NaN	NaN
1	NaN	NaN	NaN
2	Edad	Mujeres	Varones
3	NaN	NaN	NaN
4	Menor de 1 año	0.35	0.35

	Tabla de equivalencias de necesidades energéticas. Unidades de adulto equivalente, según sexo y edad	Unnamed: 1	Unnamed: 2
2	Edad	Mujeres	Varones
4	Menor de 1 año	0.35	0.35
5	1año	0.37	0.37
6	2 años	0.46	0.46
7	3 años	0.51	0.51

2	Edad	Mujeres	Varones
4	Menor de 1 año	0.35	0.35
5	1año	0.37	0.37
6	2 años	0.46	0.46
7	3 años	0.51	0.51
8	4 años	0.55	0.55

	Edad	sexo	value
0	Menor de 1 año	Mujeres	0.35
1	1año	Mujeres	0.37
2	2 años	Mujeres	0.46
3	3 años	Mujeres	0.51
4	4 años	Mujeres	0.55

	CH06	adulto_equiv	CH04
0	0	0.35	2
1	1	0.37	2
2	2	0.46	2
3	3	0.51	2
4	4	0.55	2
5	5	0.60	2
6	6	0.64	2
7	7	0.66	2
8	8	0.68	2
9	9	0.69	2

	CODUSU	NRO_HOGAR	REGION	MAS_500	AGLOMERADO	CH04	CH06	PP0
0	TQRMNOPPRHKMLNCDEIIAD00655703	1	1	1	32	1	29	
1	TQRMNOPPRHKMLNCDEIIAD00655703	1	1	1	32	2	29	
2	TQRMNOPPWHKMNUCDEIIAD00655837	1	1	1	32	2	53	
3	TQRMNOPPWHKMNUCDEIIAD00655837	1	1	1	32	2	22	
4	TQRMNOPPWHKMNUCDEIIAD00655837	1	1	1	32	1	19	

5 rows × 963 columns

	ad_equiv_hogar
CODUSU	
TQRMNOPPQHJMQLCDEIJAH00702455	1.67
TQRMNOPPQHJNMQCDEIJAH00698190	3.10
TQRMNOPPQHJONQCDEIJAH00693114	1.76
TQRMNOPPQHKMOSCDEIJAH00656008	2.53
TQRMNOPPQHLMPPCDEIJAH00701610	3.92

	CODUSU	NRO_HOGAR	REGION	MAS_500	AGLOMERADO	CH04	CH06	PP0
0	TQRMNOPPRHKMLNCDEIIAD00655703	1	1	1	32	1	29	
1	TQRMNOPPRHKMLNCDEIIAD00655703	1	1	1	32	2	29	
2	TQRMNOPPWHKMNUCDEIIAD00655837	1	1	1	32	2	53	
3	TQRMNOPPWHKMNUCDEIIAD00655837	1	1	1	32	2	22	
4	TQRMNOPPWHKMNUCDEIIAD00655837	1	1	1	32	1	19	

5 rows × 964 columns

Ejercicio 8

```
2474
```

```
821
```

```
0.24916540212443095
```

	CODUSU	NRO_HOGAR	REGION	MAS_500	AGLOMERADO	CH04	CH06	PP0
0	TQRMNOPPRHKMLNCDEIIAD00655703	1	1	1	32	1	29	
1	TQRMNOPPRHKMLNCDEIIAD00655703	1	1	1	32	2	29	
7	TQRMNOPPXHLLLNCDEIIAD00690221	1	1	1	32	1	66	
8	TQRMNOPPXHLLLNCDEIIAD00690221	1	1	1	32	1	27	
9	TQRMNOPPXHLLLNCDEIIAD00690221	1	1	1	32	2	65	

5 rows × 965 columns

Ejercicio 9

	CODUSU	NRO_HOGAR	REGION	MAS_500	AGLOMERADO	CH04	CH06	PP0
0	TQRMNOPPRHKMLNCDEIIAD00655703	1	1	1	32	1	29	
1	TQRMNOPPRHKMLNCDEIIAD00655703	1	1	1	32	2	29	
7	TQRMNOPPXHLLLNCDEIIAD00690221	1	1	1	32	1	66	
8	TQRMNOPPXHLLLNCDEIIAD00690221	1	1	1	32	1	27	
9	TQRMNOPPXHLLLNCDEIIAD00690221	1	1	1	32	2	65	

5 rows × 966 columns

	CODUSU	NRO_HOGAR	REGION	MAS_500	AGLOMERADO	CH04	CH06	PP0
0	TQRMNOPPRHKMLNCDEIIAD00655703	1	1	1	32	1	29	
1	TQRMNOPPRHKMLNCDEIIAD00655703	1	1	1	32	2	29	
7	TQRMNOPPXHLLLNCDEIIAD00690221	1	1	1	32	1	66	
8	TQRMNOPPXHLLLNCDEIIAD00690221	1	1	1	32	1	27	
9	TQRMNOPPXHLLLNCDEIIAD00690221	1	1	1	32	2	65	

5 rows × 966 columns

Ejercicio 10

```
1868
```

```
680
```

```
pobre
0    2385188
1    1356105
Name: PONDIH, dtype: int64
```

```
36.24696060960743
```

La tasa de pobreza para hogares de GBA del primer trimestre 2021 se ubica en 36,25%.

Si bien da un resultado similar al que publicó el INDEC (37,1%) nunca nos va a dar exactamente lo mismo dado que no sabemos en que mes fueron tomados los datos de los diferentes hogares y por lo tanto estamos comparando los ingresos contra la canasta básica promedio del trimestre y no con la que le correspondería a cada familia (según el mes de relevamiento de los datos).

Parte 2: Construcción de funciones

Ejercicio 1: función evalua_metodo

Ejercicio 2: función cross_validation

Ejercicio 3: función evalua_config

Ejercicio 4: función evalua_multiples_metodos

Diccionario de modelos

valores que pueden tomar "penalty" y "C"

Parte 3: Clasificación y Regularización

Ejercicio 1

```
928
```

```
927
```

Ejercicio 2

	accurancy	auc	ecm	hiperparametro	modelo
0	0.794078	0.761558	0.205922	1.0	Regresión logística
1	0.775236	0.745101	0.224764	NaN	Análisis de Discriminante Lineal
2	0.736205	0.684876	0.263795	NaN	3 vecinos cercanos
3	0.804845	0.771605	0.195155	NaN	Arbol de decisión
4	0.810229	0.779627	0.189771	NaN	Support vector machines (SVM)
5	0.837147	0.800245	0.162853	NaN	Bagging
6	0.804845	0.753609	0.195155	NaN	Random Forests
7	0.826380	0.784949	0.173620	NaN	Boosting

Ejercicio 3

Para elegir el λ por validación cruzada hay que seguir una serie de pasos.

Para un λ dado se realizan los siguientes pasos:

En primer lugar partimos los datos de la base en k partes iguales. Luego, ajustamos el modelo dejando afuera una de las k partes. Computamos el error de predicción para los datos no utilizados. Este procedimiento lo repetimos para todos los k (en cada caso se deja una parte k afuera del ajuste del modelo). Luego, se promedian todos los errores de predicción.

Este procedimiento lo repetimos para los diferentes valores de λ.

Por último comparamos los errores de predicción promedio calculados con cada λ y nos quedamos con el λ que genera un menor resultado. Ese será nuestro λ elegido.

¿Por qué se deja afuera una parte de testeo?

Nuestro objetivo es medir el el error de pronóstico fuera de la muestra. Para no separar arbitrariamente la base en etrenamiento y test se utiliza este método que permite separar la muestra en k partes e ir estimando con k-1 partes. Siempre hay una parte k que no se usa en la estimación porque es la que se va a usar luego para medir el error de predicción. De esta forma, se puede ver esto "fuera de los datos de entrenamiento" ya que se dejo una parte por fuera del ajuste de modelo. En concreto se estima todo k veces y en cada estimación se deja afuera una de las partes.

Ejercicio 4

Por un lado, si k es chico maximiza los datos para estimar pero es sensible a los valores particulares de la muestra.

Por el otro lado, si k es grande maximiza los datos para evaluar pero el modelo estimado es menos preciso.

En el caso particular de k=n el modelo se estima n veces con n-1 datos.

Como "regla general" se usa un k=5 o k=10

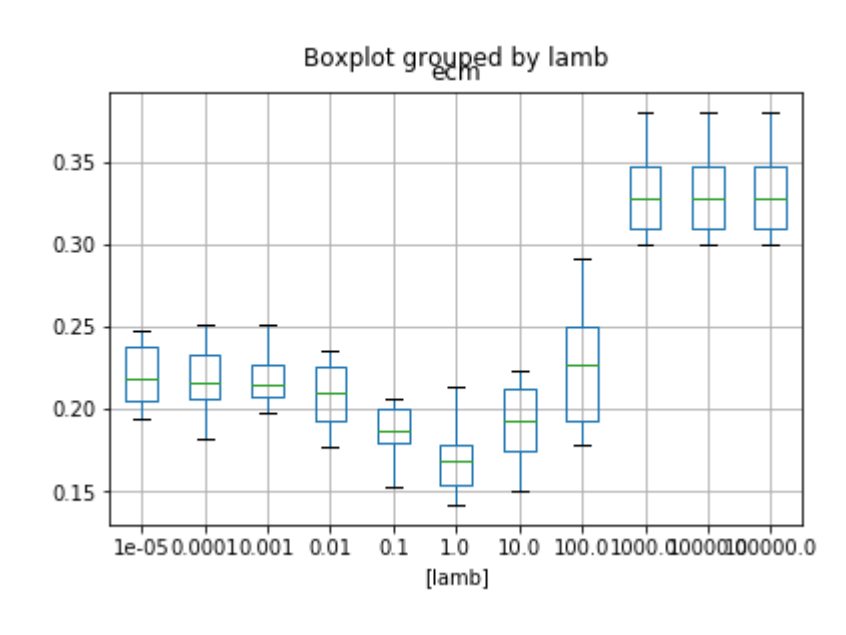
Ejercicio 5

```
[99999.99999999999,
 10000.0,
 1000.0,
 100.0,
 10.0,
 1.0,
 0.1,
 0.01,
 0.001,
 0.0001,
 1e-05]
```

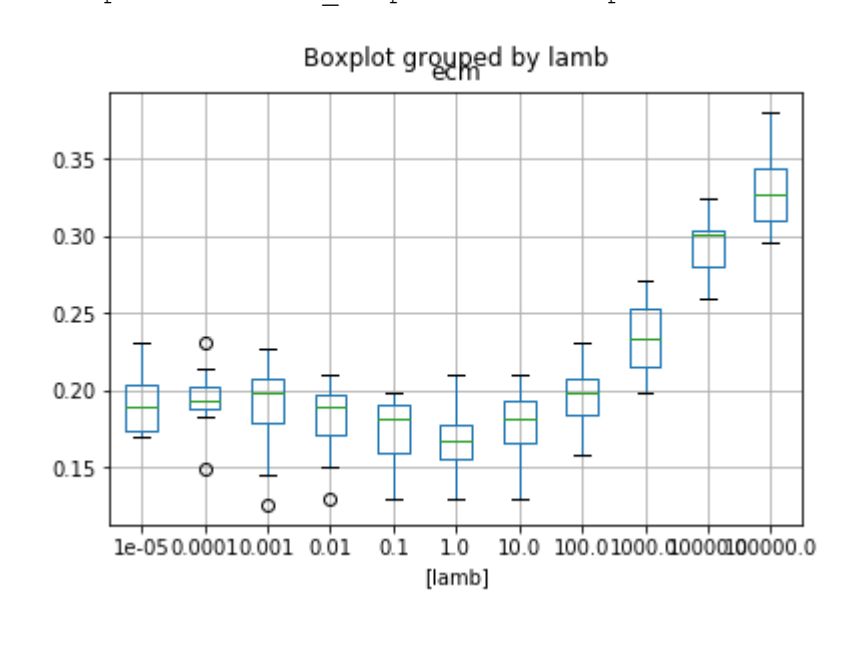
	λ Lasso	λ Ridge
0	1.0	1.0

	lamb	num	ecm
0	0.00001	0	0.193548
1	0.00001	1	0.217742
2	0.00001	2	0.233871
3	0.00001	3	0.217742
4	0.00001	4	0.238866
...	...	...	...
105	100000.00000	5	0.299595
106	100000.00000	6	0.299595
107	100000.00000	7	0.331984
108	100000.00000	8	0.307692
109	100000.00000	9	0.380567

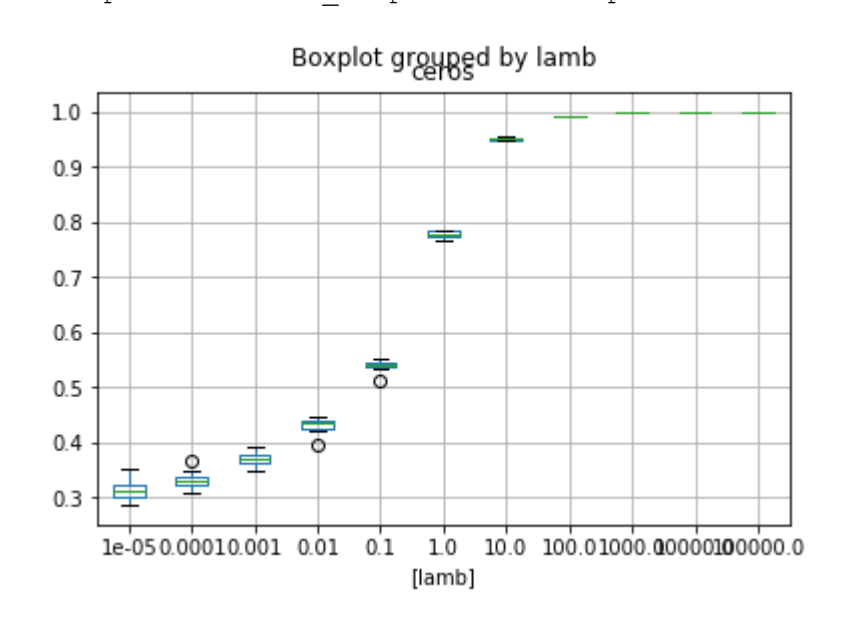
110 rows × 3 columns




```
<matplotlib.axes._subplots.AxesSubplot at 0x1a9e0fb4c88>
```




```
<matplotlib.axes._subplots.AxesSubplot at 0x1a9db4bdfc8>
```



Ejercicio 6

	predictor	coef
1	CH06	-0.009912
3	PP02C2	0.160747
4	PP02C3	0.091649
5	PP02C4	0.046124
6	PP02C5	0.060252
...	...	...
905	PP11L1_1.0	-0.057510
909	PP11M_3.0	-0.038399
911	PP11N_2.0	-0.075439
919	PP11O_8.0	-0.094207
926	PP11Q_9.0	0.012794

314 rows × 2 columns

Ejercicio 7

```
ECM mejor lasso 0.20457604306864063
ECM mejor ridge 0.19515477792732167
```

Ridge fue mejor método de regularización ya que obtuvo un menor ECM.

Ejercicio 8

El método que mejor predijo nuestra variable explicada ("pobre") fue Bagging. Para este modelo no tuvimos que elegir los hiperparámetros con cross validation. En todas las métricas fue el que obtuvo los mejores resultados:

La precisión fue de 0.839838 (máximo valor en comparación con el resto de los métodos).

El área bajo la curva (auc) alcanzó 0.800132, siendo el mayor valor reportado.

Por último, obtuvo el menor ECM, 0.160162.

Ejercicio 9

```
310
```

```
(821,)
```

```
[37.75883069]
```

La tasa que predice el modelo se ubica 37% contra 36,25% que encontramos en la base respondieron.