Curso de Big Data

Trabajo Practico 4

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

0 TQRMNOQUPHMKKUCDEOJAH00701956	2021	1	1	1	44
1 TQRMNOPWXHMOKRCDEGLDF00701361	2021	1	1	1	41
2 rows × 88 columns					

CODUSU ANO4 TRIMESTRE NRO_HOGAR REALIZADA REGION MAS_500

Ν

S S S S S

43

2

	CODUSU	ANO4	TRIMESTRE	NRO_HOGAR	REALIZADA	REGION	MAS_500
15	TQRMNQQXUHJMMUCDEIJAH00693031	2021	1	1	1	1	S
21	TQRMNOSRPHLLNRCDEIJAH00651171	2021	1	1	1	1	S
24	TQRMNOPVUHKLMNCDEIIAD00655817	2021	1	1	1	1	S

0	TQRMNOPPQHLKLMCDEFIAH00646702	2021	1	1	1	1	43	
	CODUSU	ANO4	TRIMESTRE	NRO_HOGAR	COMPONENTE	H15	REGION	
5 rov	ws × 88 columns							
43	TQRMNOQPVHLMNNCDEIIAD00650542	202	1	1	1 1		1	
36	TQRMNORYTHLOPMCDEIJAH00655933	3 202	1	1	1 1		1	

1 TQRMNOPPRHLKLMCDEFIAH00665104

36802 TQRMNOPPWHKMNUCDEIIAD00655837

2 rows >	177 columns						
	CODUSU	ANO4	TRIMESTRE	NRO_HOGAR	COMPONENTE	H15	REGION
36799	TQRMNOPPRHKMLNCDEIIAD00655703	2021	1	1	1	1	

1

1

2021

2021

2021

36800 TQRMNOPPRHKMLNCDEIIAD00655703 TQRMNOPPWHKMNUCDEIIAD00655837

36803	TQRMNOPPWHKMNUCE	DEIIAD00655837	2021	1	1	3	1
5 rows	× 177 columns						
Ejerc	icio 2						

4082

Ejercicio 3 En este inciso filtramos aquellos valores que no tiene sentido, como por ejemplo: Edades negativas y variables de ingreso

que tomen valores negatvos.

0 TQRMNOPPRHKMLNCDEIIAD00655703

1 TQRMNOPPRHKMLNCDEIIAD00655703

TQRMNOPPRHKMLNCDEIIAD00655703

TQRMNOPPRHKMLNCDEIIAD00655703

3 TQRMNOPPWHKMNUCDEIIAD00655837

4 TQRMNOPPWHKMNUCDEIIAD00655837

TQRMNOPPWHKMNUCDEIIAD00655837

Ejercicio 4 - Construcción de variables

código se aclara cuales son. Finalmente para realizar el analisis convertimos las variables categoricas en dummies.

También reemplazamos los valores faltantes por la mediana y dropeamos variables que no consideramos relevantes. En el

CODUSU ANO4 TRIMESTRE NRO_HOGAR COMPONENTE H15 REGION M/

1

1

2

2

3

Unnamed:

NaN

NaN

NaN

0.35

NaN

NaN

NaN

0.35

Unnamed:

Mujeres

0.35

0.37

32

32

32

32

32

32

32

32

32

32

MAS_500 AGLOMERADO CH04 CH06

32

32

32

32

32

32

32

32

32

32

1

2

1

2

1

2

1

2

true

417.0

407.0

439.0

430.0

negative

modelo

Análisis de

Lineal

Discriminante

Vecinos cercanos

Arbol de decisión

Regresión logística

true

178.0

174.0

98.0

193.0

positive

2

2

2

1

1

1

2

2

2

29

29

53

22

19

CH06 PP(

29

29

53

22

19

PP0

29

29

66

27

65

CH06 PP0

29

29

66

27

65

Mujeres

Unnamed:

Varones

Unnamed:

NaN

NaN

NaN

NaN

NaN

2

NaN

NaN

NaN

0.35

Varones

Unnamed:

Varones

0.35

0.37

Unnamed:

Unnamed:

NaN

NaN

Edad

según sexo y edad

NaN

NaN

NaN

0.35

NaN

NaN

Edad

NaN

Edad

1año

Menor de 1 año

según sexo y edad

Menor de 1 año

Mujeres

1

1

1

2 rows × 242 columns 3308

1

1

Generamos dos variables para tener en cuenta a la hora de analizar la pobreza. La primera es si hay niños en el hogar (los cuales representan "una boca más que alimentar" y podría pensarseque hay una relación entre la cantidad de hijos y la

1

CODUSU ANO4 TRIMESTRE NRO_HOGAR COMPONENTE H15 REGION M

2021

2021

2021

2021

2021

2021

2021

5	ō ro	ws × 242 columns							
		CODUSU	NRO_HOGAR	REGION	MAS_500	AGLOMERADO	CH04	СН06	PI
	0	TQRMNOPPRHKMLNCDEIIAD00655703	1	1	1	32	1	29	
	1	TQRMNOPPRHKMLNCDEIIAD00655703	1	1	1	32	2	29	

pobreza), y la segunda es si se trata de un hogar del tipo monoparental con jefe de hogar femenino (intentando captar un efecto de la feminización de la pobreza)

(210, 1035)

300000

250000

200000

150000

2 rows × 1033 columns

Ejercicio 5 - Gráfico

100000 50000 0 0

hog_mono_fem_hijos

0 _ 1

100000

80000

60000

40000

20000

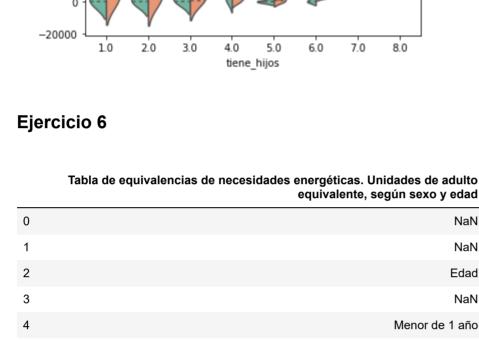


Tabla de equivalencias de necesidades energéticas. Unidades de adulto equivalente,

Tabla de equivalencias de necesidades energéticas. Unidades de adulto equivalente,

hog_mono_fem_hijos

2 4

0

1

2

3

5

8

0

1

3

4

0

1 2

0

0

6			
7			
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
•		0.10	

1

1

ad_equiv_hogar

1.67

3.10

1.76

2.53

3.92

1

1

CODUSU

1

CODUSU NRO_HOGAR REGION MAS_500 AGLOMERADO CH04

1

1

TQRMNOPPRHKMLNCDEIIAD00655703

TQRMNOPPRHKMLNCDEIIAD00655703

TQRMNOPPWHKMNUCDEIIAD00655837

TQRMNOPPWHKMNUCDEIIAD00655837

TQRMNOPPWHKMNUCDEIIAD00655837

TQRMNOPPQHJMQLCDEIJAH00702455

TQRMNOPPQHJNMQCDEIJAH00698190

TQRMNOPPQHJONQCDEIJAH00693114

TQRMNOPPQHKMOSCDEIJAH00656008

TQRMNOPPQHLMPPCDEIJAH00701610

TQRMNOPPRHKMLNCDEIIAD00655703

TQRMNOPPRHKMLNCDEIIAD00655703

TQRMNOPPWHKMNUCDEIIAD00655837

TQRMNOPPWHKMNUCDEIIAD00655837

TQRMNOPPWHKMNUCDEIIAD00655837

5 rows × 1037 columns

Ejercicio 7

Ejercicio 8

2474

821

5 rows × 1036 columns

0.55

sexo value

0.35 0.37

0.46

0.51

0.55

2

2

0.55

4 años

Edad

Menor de 1 año Mujeres

4 años

CH06 adulto_equiv CH04

0.35 0.37

0.46

1año Mujeres 2 años Mujeres

3 años Mujeres

Mujeres

	CODUSU	NRO_HOGAR	REGION	MAS_500	AGLOMERADO	CH04	CH06	PP	
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 × 1038 columns									

NRO_HOGAR REGION

1

1

1

1

1

1

1

1

1

1

1

1

NRO_HOGAR REGION MAS_500 AGLOMERADO CH04

CODUSU

CODUSU

0 TQRMNOPPRHKMLNCDEIIAD00655703

TQRMNOPPRHKMLNCDEIIAD00655703

TQRMNOPPXHLLLNCDEIIAD00690221

TQRMNOPPXHLLLNCDEIIAD00690221

TQRMNOPPXHLLLNCDEIIAD00690221

TQRMNOPPRHKMLNCDEIIAD00655703

TQRMNOPPXHLLLNCDEIIAD00690221

TQRMNOPPXHLLLNCDEIIAD00690221

TQRMNOPPXHLLLNCDEIIAD00690221

Parte 2: Construcción de funciones

Parte 3: Clasificación y Regularización

1 TQRMNOPPRHKMLNCDEIIAD00655703

5 rows × 1039 columns

pobre 0 2385188 1356105 Name: PONDIH, dtype: int64

36.24696060960743

Diccionario de modelos

Ejercicio 1

Ejercicio 2

accurancy

1001

1000

5 rows × 1039 columns

1868

680

Ejercicio 1: función evalua_metodo
Ejercicio 2: función cross_validation
Ejercicio 3: función evalua_config
Ejercicio 4: función evalua_multiples_metodos

Ejercicio 3

El modelo que mejor predice es el de Bagging.

modelo que arroja el menor error cuadrático medio (0,148).

Mejoramos nuestras predicciones respecto al TP3. Si bien también habíamos elegido como modelo a Bagging, como aora

mejora.

Para este caso en particular se usan los hiperparámetros base de la función.

negative positive 0.800808 0.769149 0.199192 97.0 0.781965 0.751193 0.218035 101.0

0.722746 0.647199 0.277254

	,	0.031932	0.617995	0.140040	80.0
6	3	0.811575	0.766449	0.188425	112.0
7	7	0.829071	0.790835	0.170929	98.0

Support vector 0.808883 0.776309 0.191117 96.0 46.0 422.0 179.0 svm machines (SVM) 0.851952 0.817995 0.148048 86.0 24.0 444.0 189.0 bagg Bagging 28.0 Random Forests 440.0 163.0 29.0 bst Boosting 439.0 177.0

false

51.0

61.0

29.0

38.0

hiperparametro

lasso con

lambda =0 LDA con

n_components

knn con k =11

 $max_depth = 16$

Cuenta no solo con la mejor precisión (0,852), sino tambien con la mayor área bajo la curva de ROC (0,818). A su vez, es el

cart con

false

177.0

82.0

introducimos dos nuevas variables que parecen ser relevantes las metricas de las nuevas predicciones son mejores. En concreto, el ECM del TP3 era de 0,16016 y ahora es de 0,148. Para el resto de las métricas también se observa una

Ejercicio 5

Ejercicio 4

262 (821,)[31.91230207] La tasa que predice el modelo se ubica 31,91% contra 36,25% que encontramos en la base respondieron.