Supervised and unsupervised machine learning: case study "fetal deaths" using the open database of the "INEC - Ecuador".

Facultad de Energía, Carrera de Ingeniería en Sistemas

Steeven Armijos-Bravo

Universidad Nacional de Loja — 1859

Carrera de Ingeniería en Sistemas / Computación

Abstract

*In the present project, three Supervised Learning (Classification) algorithms were applied: **RandomTree, RandomForest, J48**, in order to predict the main cause of fetal deaths in the year 2020: **Clustering** (SimpleKMeans), to group the most likely disease causing fetal deaths according to common characteristics, and **Association Rules** (Apriori) to obtain dataset-based rules to aid in decision making, for example in medical diagnostics. This involved a process of **problem definition, data collection** (obtained from the INEC Open Data Bank, .csv files were downloaded for the years 2015-2018), data preparation (deletion of check marks, special characters, outliers), data division (80% for training and 20% for testing), **training** (the 3 classification algorithms mentioned above were trained) and **model validation** (data were collected in a table with the predictive accuracy of each classification algorithm). Finally, it was obtained that the main cause of fetal deaths is caused by the disease: intrauterine hypoxia, unspecified.*
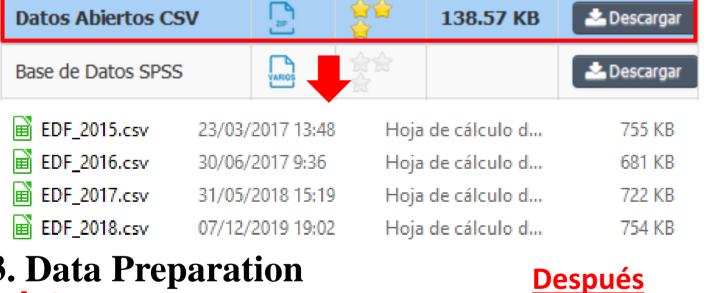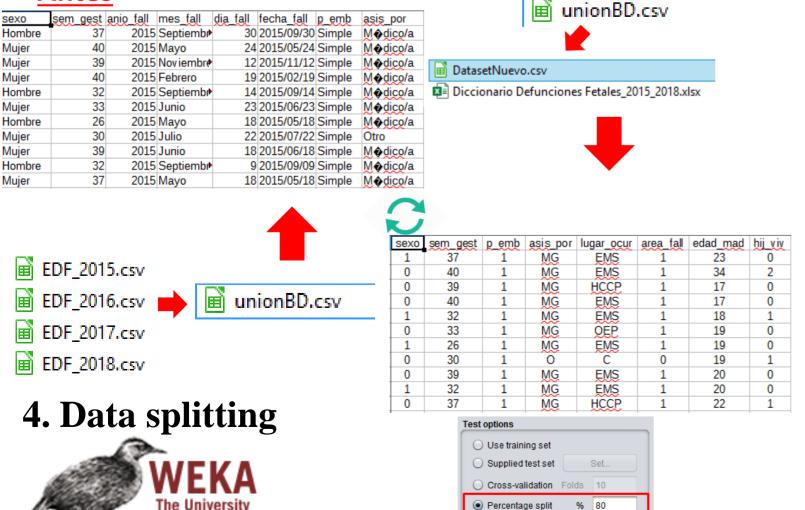
1. Problem

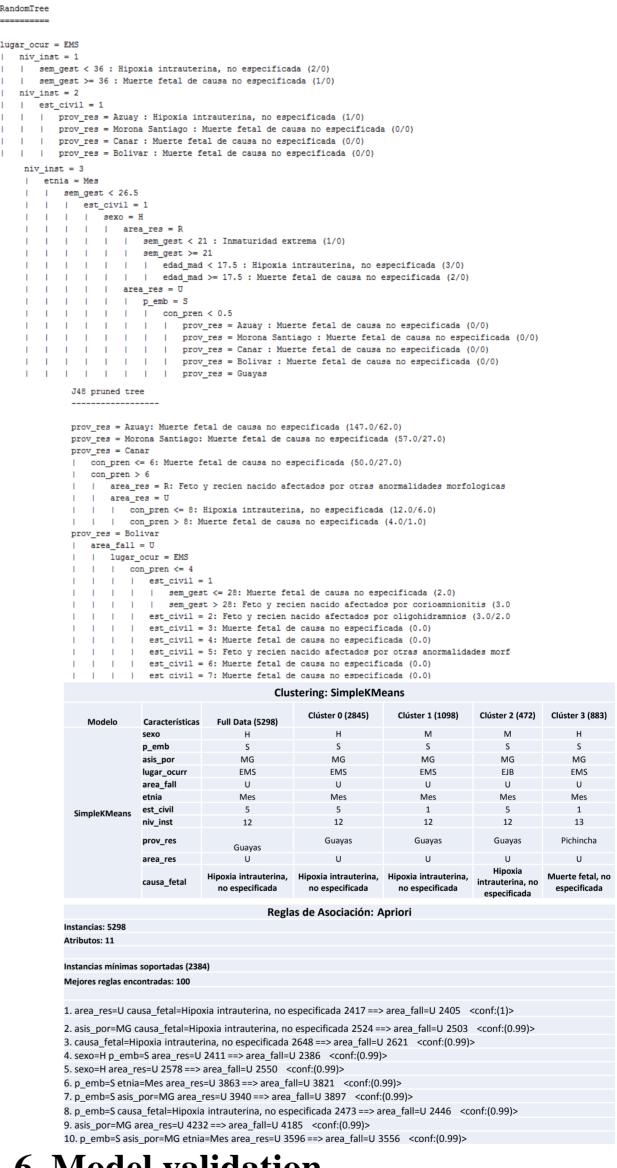What is the main cause of fetal deaths in the year 2020?

Público objetivo

INSTITUTO ECUATORIANO DE SEGURIDAD SOCIAL (IESS) · TRABAJO SOCIAL · CLÍNICA DE GINECOLOGÍA Y CONTROL DEL EMBARAZO · Obstetricia · CLINICA · Universidad Nacional de Loja · Ministerio de Salud Pública

Datos disponibles — BANCO DE DATOS ABIERTOS

Tipo de problema

Machine Learning → Aprendizaje Supervisado → Clasificación; Aprendizaje no Supervisado → Clustering, Reglas de Asociación

2. Data collection

http://aplicaciones3.ecuadorencifras.gob.ec/BIINEC-war/index.xhtml

| Datos Abiertos CSV | ZIP | ⭐⭐ | 138.57 KB | Descargar |
| Base de Datos SPSS | VARIOS | ☆☆ | | Descargar |

EDF_2015.csv	23/03/2017 13:48	Hoja de cálculo d...	755 KB
EDF_2016.csv	30/06/2017 9:36	Hoja de cálculo d...	681 KB
EDF_2017.csv	31/05/2018 15:19	Hoja de cálculo d...	722 KB
EDF_2018.csv	07/12/2019 19:02	Hoja de cálculo d...	754 KB

3. Data Preparation

Antes

sexo	sem_gest	anio_fall	mes_fall	dia_fall	fecha_fall	p_emb	asis_por
Hombre	37	2015	Septiembre	30	2015/09/30	Simple	Médico/a
Mujer	40	2015	Mayo	24	2015/05/24	Simple	Médico/a
Mujer	39	2015	Noviembre	12	2015/11/12	Simple	Médico/a
Mujer	40	2015	Febrero	19	2015/02/19	Simple	Médico/a
Hombre	32	2015	Septiembre	14	2015/09/14	Simple	Médico/a
Mujer	33	2015	Junio	23	2015/06/23	Simple	Médico/a
Hombre	26	2015	Mayo	18	2015/05/18	Simple	Médico/a
Mujer	30	2015	Julio	22	2015/07/22	Otro	
Mujer	39	2015	Junio	18	2015/06/18	Simple	Médico/a
Hombre	32	2015	Septiembre	9	2015/09/09	Simple	Médico/a
Mujer	37	2015	Mayo	18	2015/05/18	Simple	Médico/a

EDF_2015.csv · EDF_2016.csv · EDF_2017.csv · EDF_2018.csv → unionBD.csv

Después

unionBD.csv → DatasetNuevo.csv · Diccionario Defunciones Fetales_2015_2018.xlsx

sexo	sem_gest	p_emb	asis_por	lugar_ocur	area_fall	edad_mad	hij_viv
1	37	1	MG	EMS	1	23	0
0	40	1	MG	EMS	1	34	2
0	39	1	MG	HCCP	1	17	0
0	40	1	MG	EMS	1	17	0
1	32	1	MG	EMS	1	18	0
0	33	1	MG	OEP	1	19	0
1	26	1	MG	EMS	1	18	0
0	30	1	O	C	0	19	1
0	39	1	MG	EMS	1	20	0
1	32	1	MG	EMS	1	20	0
0	37	1	MG	HCCP	1	22	1

4. Data splitting

WEKA — The University of Waikato

Test options
- ○ Use training set
- ○ Supplied test set — Set...
- ○ Cross-validation — Folds 10
- ● Percentage split — % 80
- More options...

5. Training a model

```
RandomTree
==========

lugar_ocur = EMS
|  niv_inst = 1
|  |  sem_gest < 36 : Hipoxia intrauterina, no especificada (2/0)
|  |  sem_gest >= 36 : Muerte fetal de causa no especificada (1/0)
|  niv_inst = 2
|  |  est_civil = 1
|  |  |  prov_res = Azuay : Hipoxia intrauterina, no especificada (1/0)
|  |  |  prov_res = Morona Santiago : Muerte fetal de causa no especificada (0/0)
|  |  |  prov_res = Canar : Muerte fetal de causa no especificada (0/0)
|  |  |  prov_res = Bolivar : Muerte fetal de causa no especificada (0/0)
|  niv_inst = 3
|  |  etnia = Mes
|  |  |  sem_gest < 26.5
|  |  |  |  est_civil = 1
|  |  |  |  |  sexo = H
|  |  |  |  |  |  area_res = R
|  |  |  |  |  |  |  sem_gest < 21 : Inmaturidad extrema (1/0)
|  |  |  |  |  |  |  sem_gest >= 21
|  |  |  |  |  |  |  |  edad_mad < 17.5 : Hipoxia intrauterina, no especificada (3/0)
|  |  |  |  |  |  |  |  edad_mad >= 17.5 : Muerte fetal de causa no especificada (2/0)
|  |  |  |  |  |  area_res = U
|  |  |  |  |  |  |  p_emb = S
|  |  |  |  |  |  |  |  con_pren < 0.5
|  |  |  |  |  |  |  |  |  prov_res = Azuay : Muerte fetal de causa no especificada (0/0)
|  |  |  |  |  |  |  |  |  prov_res = Morona Santiago : Muerte fetal de causa no especificada (0/0)
|  |  |  |  |  |  |  |  |  prov_res = Canar : Muerte fetal de causa no especificada (0/0)
|  |  |  |  |  |  |  |  |  prov_res = Bolivar : Muerte fetal de causa no especificada (0/0)
|  |  |  |  |  |  |  |  |  prov_res = Guayas
```

```
J48 pruned tree
------------------

prov_res = Azuay: Muerte fetal de causa no especificada (147.0/62.0)
prov_res = Morona Santiago: Muerte fetal de causa no especificada (57.0/27.0)
prov_res = Canar
|  con_pren <= 6: Muerte fetal de causa no especificada (50.0/27.0)
|  con_pren > 6
|  |  area_res = R: Feto y recien nacido afectados por otras anormalidades morfologicas
|  |  area_res = U
|  |  |  con_pren <= 8: Hipoxia intrauterina, no especificada (12.0/6.0)
|  |  |  con_pren > 8: Muerte fetal de causa no especificada (4.0/1.0)
prov_res = Bolivar
|  area_fall = U
|  |  lugar_ocur = EMS
|  |  |  con_pren <= 4
|  |  |  |  est_civil = 1
|  |  |  |  |  sem_gest <= 28: Muerte fetal de causa no especificada (2.0)
|  |  |  |  |  sem_gest > 28: Feto y recien nacido afectados por corioamnionitis (3.0
|  |  |  |  est_civil = 2: Feto y recien nacido afectados por oligohidramnios (3.0/2.0)
|  |  |  |  est_civil = 3: Muerte fetal de causa no especificada (0.0)
|  |  |  |  est_civil = 4: Muerte fetal de causa no especificada (0.0)
|  |  |  |  est_civil = 5: Feto y recien nacido afectados por otras anormalidades morf
|  |  |  |  est_civil = 6: Muerte fetal de causa no especificada (0.0)
|  |  |  |  est civil = 7: Muerte fetal de causa no especificada (0.0)
```

Clustering: SimpleKMeans

Modelo	Características	Full Data (5298)	Clúster 0 (2845)	Clúster 1 (1098)	Clúster 2 (472)	Clúster 3 (883)
SimpleKMeans	sexo	H	H	M	M	M
	p_emb	S	S	S	S	S
	asis_por	MG	MG	MG	MG	MG
	lugar_ocurr	EMS	EMS	EMS	EJB	EMS
	area_fall	U	U	U	U	U
	etnia	Mes	Mes	Mes	Mes	Mes
	est_civil	5	5	1	5	1
	niv_inst	12	12	12	12	13
	prov_res	Guayas	Guayas	Guayas	Guayas	Pichincha
	area_res	U	U	U	U	U
	causa_fetal	Hipoxia intrauterina, no especificada	Hipoxia intrauterina, no especificada	Hipoxia intrauterina, no especificada	Hipoxia intrauterina, no especificada	Muerte fetal, no especificada

Reglas de Asociación: Apriori

Instancias: 5298
Atributos: 11

Instancias mínimas soportadas (2384)
Mejores reglas encontradas: 100

1. area_res=U causa_fetal=Hipoxia intrauterina, no especificada 2417 ==> area_fall=U 2405   <conf:(1)>
2. asis_por=MG causa_fetal=Hipoxia intrauterina, no especificada 2524 ==> area_fall=U 2503   <conf:(0.99)>
3. causa_fetal=Hipoxia intrauterina, no especificada 2648 ==> area_fall=U 2621   <conf:(0.99)>
4. sexo=H p_emb=S area_res=U 2411 ==> area_fall=U 2386   <conf:(0.99)>
5. sexo=H area_res=U 2578 ==> area_fall=U 2550   <conf:(0.99)>
6. p_emb=S etnia=Mes area_res=U 3863 ==> area_fall=U 3821   <conf:(0.99)>
7. p_emb=S asis_por=MG area_res=U 3940 ==> area_fall=U 3897   <conf:(0.99)>
8. p_emb=S causa_fetal=Hipoxia intrauterina, no especificada 2473 ==> area_fall=U 2446   <conf:(0.99)>
9. asis_por=MG area_res=U 4232 ==> area_fall=U 4185   <conf:(0.99)>
10. p_emb=S asis_por=MG etnia=Mes area_res=U 3596 ==> area_fall=U 3556   <conf:(0.99)>

6. Model validation

		Instancias : 1060: 80% de entrenamiento y 20% test					
		RandomForest		RandomTree		J48	
		Correctos	Incorrectos	Correctos	Incorrectos	Correctos	Incorrectos
		563	497	461	599	539	521
	Precisión	53.11%	46.89%	43.49%	56.51%	50.85%	49.15%

```
Predictions on test split === RandomForest

inst#   actual  predicted error prediction
  1 3:'Hipoxia intrauterina, no especificada' 3:Hipoxia intrauterina, no especificada   0.755
  2 1:'Muerte fetal de causa no especificada' 1:Muerte fetal de causa no especificada   0.74
  3 30:'Malformacion congenita, no especificada' 3:Hipoxia intrauterina, no especificada +  0.533
  4 3:'Hipoxia intrauterina, no especificada' 3:Hipoxia intrauterina, no especificada   0.92
  5 3:'Hipoxia intrauterina, no especificada' 3:Hipoxia intrauterina, no especificada   0.92

Predictions on test split === RandomTree

inst#   actual  predicted error prediction
  1 3:'Hipoxia intrauterina, no especificada' 3:Hipoxia intrauterina, no especificada   1
  2 1:'Muerte fetal de causa no especificada' 1:Muerte fetal de causa no especificada   1
  3 30:'Malformacion congenita, no especificada' 3:Hipoxia intrauterina, no especificada +  1
  4 3:'Hipoxia intrauterina, no especificada' 3:Hipoxia intrauterina, no especificada   1
  5 3:'Hipoxia intrauterina, no especificada' 3:Hipoxia intrauterina, no especificada   1

Predictions on test split === J48

inst#   actual  predicted error prediction
  1 3:'Hipoxia intrauterina, no especificada' 3:Hipoxia intrauterina, no especificada   0.785
  2 1:'Muerte fetal de causa no especificada' 1:Muerte fetal de causa no especificada   0.612
  3 30:'Malformacion congenita, no especificada' 3:Hipoxia intrauterina, no especificada +  0.785
  4 3:'Hipoxia intrauterina, no especificada' 3:Hipoxia intrauterina, no especificada   0.785
  5 3:'Hipoxia intrauterina, no especificada' 3:Hipoxia intrauterina, no especificada   0.785
```