INTRODUCIR UN TÍTULO

Objetivos

Preprocessing

Análisis
Recaída

Árbol
Decisiones

Conclusiones

Referencias

Pablo Martín Escobar, Jorge Sáenz de Miera Marzo, Javier Gil Domínguez, Nicolás Vega Muñoz

Objetivos de Negocio

- Identificación de pacientes con alto riesgo de recaída
- Optimización de los costes asociados en un ___%
- Reducción de la mortalidad en un ___%

Costes

- **Seguimiento:** *40000$*
- **Tratamiento:**
  - Costes directos + Costes indirectos * Costes directos (metástasis o local)
  - *16509,3$* + *26095$* * (*35732$* o *8271$*)
- **Cuidados paliativos:** *7832$*

Mortalidad

- **Metastasis:** *70%*
- **Local:** *14%*
- Con **cáncer** de **mama** a los **5 años:** *0,298*

O b j e t i v o s   d e   C D

- Desarrollo modelo predictivo que identifique pacientes con alto riesgo de recaída

- Mejora en la toma de decisiones clínicas
  - Proporcionando información valiosa gracias al sistema desarrollado

- Optimización en la asignación de tratamientos
  - Mediante el desarrollo de un modelo predictivo

D a t o s   S u c i o s

Patients

	ehr	birth_date	diagnosis_date	death_date
0	10011773	1959-07-05	2015-04-02	NaN
1	10020495	1953-10-02	2017-12-04	NaN
2	10030299	1966-08-16	2019-06-27	NaN
3	10030824	1953-03-03	2018-09-07	NaN
4	10041592	1959-07-06	2018-11-19	NaN
...	...	...	...	...
7495	77575810	1958-03-21	2014-02-22	NaN
7496	77578551	1948-05-18	2018-07-11	NaN
7497	77579212	1976-02-13	2022-03-22	NaN
7498	77605742	1950-03-16	2015-06-29	NaN
7499	77606485	1962-05-13	2013-11-26	NaN

	ehr	birth_date	diagnosis_date	death_date
0	72992494	13/02/1942	08/05/2018	NaN
1	73002338	22/10/1949	08/09/2018	NaN
2	73008149	04/12/1969	12/04/2018	NaN
3	73012939	18/09/1973	17/04/2014	01/01/2018
4	73019870	07/12/1955	24/05/2018	NaN
...	...	...	...	...
2995	99966287	16/12/1978	02/08/2018	NaN
2996	99981192	10/04/1949	28/10/2017	NaN
2997	99988958	21/01/1966	27/05/2013	NaN
2998	99992350	04/09/1939	17/12/2018	NaN
2999	99994417	10/02/1962	01/04/2017	NaN

Gynecological

	Unnamed: 0	ehr	pregnancy	birth	caesarean	abort	menarche_age	menopause_age
0	0	10011773	-8.0	0.0	NaN	0.0	NaN	71.0
1	1	10030299	3.0	NaN	0.0	0.0	NaN	NaN
2	2	10030824	0.0	0.0	0.0	NaN	NaN	44.0
3	3	10053435	2.0	NaN	0.0	1.0	21.0	74.0
4	4	10111454	0.0	-6.0	NaN	0.0	16.0	55.0
...	...	...	...	...	...	...	...	...
7681	7681	99948591	9.0	2.0	-4.0	0.0	12.0	47.0
7682	7682	99961100	2.0	2.0	0.0	0.0	19.0	52.0
7683	7683	99981192	3.0	3.0	0.0	-5.0	11.0	NaN
7684	7684	99992350	4.0	3.0	NaN	-6.0	11.0	55.0
7685	7685	99994417	2.0	-6.0	NaN	NaN	13.0	NaN

	ehr	er	her2	ki67	pr
0	10011773	0.0	0.0	19.0	NaN
1	10020495	1.0	0.0	9.0	0.0
2	10030299	1.0	NaN	18.0	1.0
3	10030824	1.0	0.0	NaN	1.0
4	10041592	0.0	0.0	65.0	0.0
...	...	...	...	...	...
9995	99966287	1.0	0.0	NaN	0.0
9996	99981192	1.0	0.0	14.0	0.0
9997	99988958	NaN	0.0	67.0	NaN
9998	99992350	0.0	0.0	16.0	0.0
9999	99994417	1.0	0.0	20.0	1.0

Histochemistry

	Unnamed: 0	ehr	n_tumor	t_category	n_category	m_category	t_category_after_neoadj	n_category_after_neoadj	m_category_after_neoadj	stage_diagnosis
0	0	10011773	1	IS	0	0	NaN	NaN	NaN	0
1	1	10020495	1	1	0	0	NaN	NaN	NaN	IA
2	2	10020495	2	3	1	0	2	0.0	NaN	IA
3	3	10030299	1	1	0	0	NaN	NaN	NaN	IA
4	4	10030824	1	2	1	0	2	2.0	0.0	IIIA
...	...	...	...	...	...	...	...	...	...	...
11162	11162	99966287	1	1	1	0	1	0.0	0.0	IB
11163	11163	99981192	1	1	0	0	NaN	NaN	NaN	IA
11164	11164	99988958	1	0	2	0	1	1.0	0.0	IIIA
11165	11165	99992350	1	2	0	0	3	2.0	0.0	IIA

Tumor

P a t i e n t s

- Estudio pacientes comunes
  - El número total de pacientes comunes son: 500

	ehr	birth_date	diagnosis_date	death_date
7000	72992494	1942-02-13	2018-05-08	NaN
7001	73002338	1949-10-22	2018-09-08	NaN
7002	73008149	1969-12-04	2018-04-12	NaN
7003	73012939	1973-09-18	2014-04-17	2018-01-01
7004	73019870	1955-12-07	2018-05-24	NaN
...	...	...	...	...
7495	77575810	1958-03-21	2014-02-22	NaN
7496	77578551	1948-05-18	2018-07-11	NaN
7497	77579212	1976-02-13	2022-03-22	NaN
7498	77605742	1950-03-16	2015-06-29	NaN
7499	77606485	1962-05-13	2013-11-26	NaN

Pacientes comunes batch1

	ehr	birth_date	diagnosis_date	death_date
0	72992494	13/02/1942	08/05/2018	NaN
1	73002338	22/10/1949	08/09/2018	NaN
2	73008149	04/12/1969	12/04/2018	NaN
3	73012939	18/09/1973	17/04/2014	01/01/2018
4	73019870	07/12/1955	24/05/2018	NaN
...	...	...	...	...
495	77575810	21/03/1958	22/02/2014	NaN
496	77578551	18/05/1948	11/07/2018	NaN
497	77579212	13/02/1976	22/03/2022	NaN
498	77605742	16/03/1950	29/06/2015	NaN
499	77606485	13/05/1962	26/11/2013	NaN

Pacientes comunes batch2

- Cambio formato de fecha y unión pacientes

```
pacientes_comunes_batch2['birth_date'] = pd.to_datetime(pacientes_comunes_batch2['birth_date'], format='%d/%m/%Y')
pacientes_comunes_batch2['death_date'] = pd.to_datetime(pacientes_comunes_batch2['death_date'], format='%d/%m/%Y')
pacientes_comunes_batch2['diagnosis_date'] = pd.to_datetime(pacientes_comunes_batch2['diagnosis_date'], format='%d/%m/%Y')

pacientes_comunes_batch1['birth_date'] = pd.to_datetime(pacientes_comunes_batch1['birth_date'])
pacientes_comunes_batch1['diagnosis_date'] = pd.to_datetime(pacientes_comunes_batch1['diagnosis_date'])
pacientes_comunes_batch1['death_date'] = pd.to_datetime(pacientes_comunes_batch1['death_date'])
```

	ehr	birth_date	diagnosis_date	death_date
0	10011773	1959-07-05	2015-04-02	NaN
1	10020495	1953-10-02	2017-12-04	NaN
2	10030299	1966-08-16	2019-06-27	NaN
3	10030824	1953-03-03	2018-09-07	NaN
4	10041592	1959-07-06	2018-11-19	NaN
...	...	...	...	...
9995	99966287	1978-12-16	2018-08-02	NaN
9996	99981192	1949-04-10	2017-10-28	NaN
9997	99988958	1966-01-21	2013-05-27	NaN
9998	99992350	1939-09-04	2018-12-17	NaN
9999	99994417	1962-02-10	2017-04-01	NaN

Patients

P a t i e n t s

- Cambio columnas de fechas por columnas numéricas
  - ‘birth_date’ → ‘age’
  - ‘diagnosis_date’ → ‘diagnosis_age’

```
from datetime import datetime
patients['age'] = datetime.now().year - patients['birth_date'].dt.year
patients['diagnosis_age'] = patients['diagnosis_date'].dt.year - patients['birth_date'].dt.year
```

- Cambio variable ‘death_date’ por variable binaria
  - ‘death_date’ → ‘dead’

```
patients['dead'] = patients['death_date'].notnull().astype(int)
patients.drop(columns=['death_date', 'birth_date', 'diagnosis_date'], inplace=True)
```

R e s u l t a d o   f i n a l :

	ehr	age	diagnosis_age	dead
0	10011773	65	56	0
1	10020495	71	64	0
2	10030299	58	53	0
3	10030824	71	65	0
4	10041592	65	59	0
...	...	...	...	...
9995	99966287	46	40	0
9996	99981192	75	68	0
9997	99988958	58	47	0
9998	99992350	85	79	0
9999	99994417	62	55	0

Patients

H i s t o c h e m i s t r y

- Eliminación filas con 2 o más valores nulos

```
filas_nulas = histochemistry[histochemistry.columns.drop('ehr')].isnull().sum(axis=1) >=2
len(filas_nulas[filas_nulas].index.tolist())
```

```
488
```

```
histochemistry.drop(filas_nulas[filas_nulas].index.tolist(), inplace=True)
```

- Se añade la nueva información al df de ‘patients’

R e s u l t a d o f i n a l :

	ehr	age	diagnosis_age	dead	er	her2	ki67	pr
0	10011773	65	56	0	0.0	0.0	19.0	NaN
1	10020495	71	64	0	1.0	0.0	9.0	0.0
2	10030299	58	53	0	1.0	NaN	18.0	1.0
3	10030824	71	65	0	1.0	0.0	NaN	1.0
4	10041592	65	59	0	0.0	0.0	65.0	0.0
...	...	...	...	...	...	...	...	...
9507	99963879	65	62	0	1.0	0.0	18.0	1.0
9508	99966287	46	40	0	1.0	0.0	NaN	0.0
9509	99981192	75	68	0	1.0	0.0	14.0	0.0
9510	99992350	85	79	0	0.0	0.0	16.0	0.0
9511	99994417	62	55	0	1.0	0.0	20.0	1.0

Patients

G y n e c o l o g i c a l

Estudio columnas ‘pregnancy’, ‘abort’, ‘caesarean’ y ‘birth’

- Número de pacientes fuera de rango:

```
Numero de pacientes con numero de embarazos fuera de rango:  585
Numero de pacientes con numero de partos fuera de rango:  609
Numero de pacientes con numero de cesareas fuera de rango:  667
Numero de pacientes con numero de abortos fuera de rango:  617
```

- Eliminación de filas con 2 o más valores nulos:

```
gynecological = gynecological.dropna(subset=['pregnancy', 'birth', 'caesarean', 'abort'], thresh=3)
```

- Imputación nulo restante mediante la fórmula:
  - Embarazos = Partos + Cesáreas + Abortos

```
gynecological['pregnancy'] = gynecological['pregnancy'].fillna(gynecological['abort'] + gynecological['birth'] + gynecological['caesarean'])
gynecological['abort'] = gynecological['abort'].fillna(gynecological['pregnancy'] - gynecological['birth'] - gynecological['caesarean'])
gynecological['birth'] = gynecological['birth'].fillna(gynecological['pregnancy'] - gynecological['abort'] - gynecological['caesarean'])
gynecological['caesarean'] = gynecological['caesarean'].fillna(gynecological['pregnancy'] - gynecological['birth'] - gynecological['abort'])
```

- Eliminación casos atípicos que no cumplen con la fórmula

Gynecological

Estudio columnas ‘menopause_age’, ‘menarche_age’

- Creación de una variable binaria para cada una

- Valor de ‘menopause_age’ nulo y < 60 años → No menopausia
- Sí se tienen datos o > 60 años → Sí menopausia

- Para ‘menarche_age’, valores fuera de [8, 15] → Media: 12

Resultado final:

	ehr	pregnancy	birth	caesarean	abort	menarche_age	has_menopause	age	diagnosis_age	is_dead	er	her2	ki67	pr
0	10030299	3.0	3.0	0.0	0.0	12.0	0	58	53	0	1.0	NaN	18.0	1.0
1	10030824	0.0	0.0	0.0	0.0	12.0	1	71	65	0	1.0	0.0	NaN	1.0
2	10053435	2.0	1.0	0.0	1.0	21.0	1	67	60	0	1.0	0.0	NaN	1.0
3	10115313	2.0	0.0	0.0	2.0	18.0	1	67	62	0	1.0	0.0	14.0	0.0
4	10119160	1.0	1.0	0.0	0.0	15.0	1	62	55	0	0.0	1.0	52.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4492	99880060	4.0	3.0	0.0	1.0	11.0	1	66	60	0	1.0	1.0	39.0	0.0
4493	99899322	3.0	3.0	0.0	0.0	13.0	1	50	43	0	1.0	0.0	10.0	1.0
4494	99948591	9.0	2.0	7.0	0.0	12.0	1	69	67	0	1.0	0.0	12.0	1.0
4495	99961100	2.0	2.0	0.0	0.0	19.0	1	54	48	0	1.0	0.0	21.0	1.0
4496	99981192	3.0	3.0	0.0	0.0	11.0	1	75	68	0	1.0	0.0	14.0	0.0

Gynecological patients

Tumor

- Hay ids que se repiten, una persona puede tener varios tumores

- Tiene más elementos comunes con ‘patients’
    - Elementos comunes ‘patients’: 9485
    - Elementos comunes ‘gynecological_patients’: 7262

Resultado final:

	ehr	n_tumor	t_category	n_category	m_category	t_category_after_neoadj	n_category_after_neoadj	m_category_after_neoadj
0	10011773	1	IS	0	0	NaN	NaN	NaN
1	10020495	1	1	0	0	NaN	NaN	NaN
2	10020495	2	3	1	0	2	0.0	NaN
3	10030299	1	1	0	0	NaN	NaN	NaN
4	10030824	1	2	1	0	2	2.0	0.0
...	...	...	...	...	...	...	...	...
10621	99963879	1	2	3	0	3	1.0	0.0
10622	99966287	1	1	1	0	1	0.0	0.0
10623	99981192	1	1	0	0	NaN	NaN	NaN
10624	99992350	1	2	0	0	3	2.0	0.0
10625	99994417	1	0	0	0	1	0.0	0.0

Tumor patients

	ehr	n_tumor	t_category	n_category	m_category	t_category_after_neoadj	n_category_after_neoadj	m_category_after_neoadj
0	10011773	1	IS	0	0	NaN	NaN	NaN
1	10030299	1	1	0	0	NaN	NaN	NaN
2	10030824	1	2	1	0	2	2.0	0.0
3	10053435	1	0	0	0	NaN	NaN	NaN
4	10111454	1	1	0	0	1	0.0	0.0
...	...	...	...	...	...	...	...	...
8154	99948591	1	2	0	0	NaN	NaN	NaN
8155	99961100	1	1	2	0	1	0.0	0.0
8156	99981192	1	1	0	0	NaN	NaN	NaN
8157	99992350	1	2	0	0	3	2.0	0.0
8158	99994417	1	0	0	0	1	0.0	0.0

Tumor patients gyn

Variable Recaída

- Creación variable binaria
  - Recaída = paciente con más de un tumor

- Porcentaje de pacientes con o sin recaída:
  - Con recaída: 12,03%
  - Sin recaída: 87,97%

Desbalanceados

Resultado final:

	ehr	n_tumor	t_category	n_category	m_category	t_category_after_neoadj	n_category_after_neoadj	m_category_after_neoadj
0	10011773	1	IS	0	0	NaN	NaN	NaN
1	10020495	1	1	0	0	NaN	NaN	NaN
2	10020495	2	3	1	0	2	0.0	NaN
3	10030299	1	1	0	0	NaN	NaN	NaN
4	10030824	1	2	1	0	2	2.0	0.0
...	...	...	...	...	...	...	...	...
10621	99963879	1	2	3	0	3	1.0	0.0
10622	99966287	1	1	1	0	1	0.0	0.0
10623	99981192	1	1	0	0	NaN	NaN	NaN
10624	99992350	1	2	0	0	3	2.0	0.0
10625	99994417	1	0	0	0	1	0.0	0.0

Tumor patients

	ehr	n_tumor	t_category	n_category	m_category	t_category_after_neoadj	n_category_after_neoadj	m_category_after_neoadj
0	10011773	1	IS	0	0	NaN	NaN	NaN
1	10030299	1	1	0	0	NaN	NaN	NaN
2	10030824	1	2	1	0	2	2.0	0.0
3	10053435	1	0	0	0	NaN	NaN	NaN
4	10111454	1	1	0	0	1	0.0	0.0
...	...	...	...	...	...	...	...	...
8154	99948591	1	2	0	0	NaN	NaN	NaN
8155	99961100	1	1	2	0	1	0.0	0.0
8156	99981192	1	1	0	0	NaN	NaN	NaN
8157	99992350	1	2	0	0	3	2.0	0.0
8158	99994417	1	0	0	0	1	0.0	0.0

Tumor patients gyn

Análisis Cols

Eliminación Cols

- Columnas ‘lobular’ y ‘ductal’

```
print(patients.lobular.unique())
patients['lobular'] = patients['lobular'].notnull().astype(int)
```

```
[nan  1.]
```

```
print(patients.ductal.unique())
patients['ductal'] = patients['ductal'].notnull().astype(int)
```

```
[nan  1.]
```

- Eliminación columnas:
  - 't_category_after_neoadj',
  - 'n_category_after_neoadj',
  - 'm_category_after_neoadj',
  - 'stage_after_neo'

- Aplicación OneHotEncoder a:
  - 't_category’
  - ‘stage_diagnosis’
  - ‘neoadjuvant’

Resultado final:

	ehr	n_category	m_category	grade	ductal	lobular	neoadjuvant	age	diagnosis_age	dead	...	t_category_IS	stage_diagnosis_0	stage_diagnosis_IA	stage_diagnosis_IB	stage_diagnosis_IIA	stage_diagnosis_IIB	stage_diagnosis_IIIA	stage_diagnosis_IIIB	stage_diagnosis_IIIC	tage_diagnosis_IV
0	10011773	0	0	1	0	0	0	65	56	0	...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	10020495	0	0	2	1	0	0	71	64	0	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	10030299	0	0	1	1	0	0	58	53	0	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	10030824	1	0	3	0	1	1	71	65	0	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
5	10041592	1	0	2	1	0	1	65	59	0	...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
10620	99963879	3	0	2	1	0	1	65	62	0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
10621	99966287	1	0	3	1	0	1	46	40	0	...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
10622	99981192	0	0	2	0	0	0	75	68	0	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10623	99992350	0	0	2	1	0	1	85	79	0	...	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
10624	99994417	0	0	3	0	0	1	62	55	0	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Patients

Imputación Nulos Normalización

- Imputación de los nulos de ‘patients’ mediante KNNImputer

	ehr	n_category	m_category	grade	ductal	lobular	neoadjuvant	age	diagnosis_age	dead	...	t_category_IS	stage_diagnosis_0	stage_diagnosis_IA	stage_diagnosis_IB	stage_diagnosis_IIA	stage_diagnosis_IIB	stage_diagnosis_IIIA	stage_diagnosis_IIIB	stage_diagnosis_IIIC	stage_diagnosis_IV
0	10011773.0	0.0	0.0	1.0	0.0	0.0	0.0	65.0	56.0	0.0	...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	10020495.0	0.0	0.0	2.0	1.0	0.0	0.0	71.0	64.0	0.0	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	10030299.0	0.0	0.0	1.0	1.0	0.0	0.0	58.0	53.0	0.0	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	10030824.0	1.0	0.0	3.0	0.0	1.0	1.0	71.0	65.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
4	10041592.0	1.0	0.0	2.0	1.0	0.0	1.0	65.0	59.0	0.0	...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9479	99963879.0	3.0	0.0	2.0	1.0	0.0	1.0	65.0	62.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
9480	99966287.0	1.0	0.0	3.0	1.0	0.0	1.0	46.0	40.0	0.0	...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
9481	99981192.0	0.0	0.0	2.0	0.0	0.0	0.0	75.0	68.0	0.0	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9482	99992350.0	0.0	0.0	2.0	1.0	0.0	1.0	85.0	79.0	0.0	...	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
9483	99994417.0	0.0	0.0	3.0	0.0	0.0	1.0	62.0	55.0	0.0	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Patients

- Normalización mediante MinMaxScaler

	n_category	m_category	grade	ductal	lobular	neoadjuvant	age	diagnosis_age	is_dead	er	...	t_category_IS	stage_diagnosis_0	stage_diagnosis_0	stage_diagnosis_IA	stage_diagnosis_IB	stage_diagnosis_IIA	stage_diagnosis_IIB	stage_diagnosis_IIIA	stage_diagnosis_IIIB	stage_diagnosis_IIIC	stage_diagnosis_IV
0	0.000000	0.0	0.0	0.0	0.0	0.0	0.527027	0.506024	0.0	0.0	...	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.000000	0.0	0.5	1.0	0.0	0.0	0.608108	0.602410	0.0	1.0	...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.000000	0.0	0.0	1.0	0.0	0.0	0.432432	0.469880	0.0	1.0	...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.333333	0.0	1.0	0.0	1.0	1.0	0.608108	0.614458	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
4	0.333333	0.0	0.5	1.0	0.0	1.0	0.527027	0.542169	0.0	0.0	...	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9479	1.000000	0.0	0.5	1.0	0.0	1.0	0.527027	0.578313	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
9480	0.333333	0.0	1.0	1.0	0.0	1.0	0.270270	0.313253	0.0	1.0	...	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
9481	0.000000	0.0	0.5	0.0	0.0	0.0	0.662162	0.650602	0.0	1.0	...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9482	0.000000	0.0	0.5	1.0	0.0	1.0	0.797297	0.783133	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
9483	0.000000	0.0	1.0	0.0	0.0	1.0	0.486486	0.493976	0.0	1.0	...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Patients norm

Creación df Patients_gyn

- Df que dispone de la información ginecológica
- Imputación de los nulos de ‘patients_gyn’ mediante KNNImputer

	ehr	n_category	m_category	grade	ductal	lobular	neoadjuvant	age	diagnosis_age	is_dead	...	stage_diagnosis_IIIA	stage_diagnosis_IIIB	stage_diagnosis_IIIC	stage_diagnosis_IV	pregnancy	birth	caesarean	abort	menarche_age	has_menopause
0	10030299.0	0.0	0.0	1.0	1.0	0.0	0.0	58.0	53.0	0.0	...	0.0	0.0	0.0	0.0	3.0	3.0	0.0	0.0	12.0	0.0
1	10030824.0	1.0	0.0	3.0	0.0	1.0	1.0	71.0	65.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	12.0	1.0
2	10053435.0	0.0	0.0	3.0	1.0	0.0	0.0	67.0	60.0	0.0	...	0.0	0.0	0.0	1.0	2.0	1.0	0.0	1.0	21.0	1.0
3	10115313.0	0.0	1.0	3.0	1.0	0.0	0.0	67.0	62.0	0.0	...	0.0	0.0	0.0	0.0	2.0	0.0	0.0	2.0	18.0	1.0
4	10119160.0	1.0	0.0	1.0	1.0	0.0	1.0	62.0	55.0	0.0	...	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	15.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4480	99880060.0	0.0	0.0	3.0	0.0	1.0	1.0	66.0	60.0	0.0	...	0.0	0.0	0.0	0.0	4.0	3.0	0.0	1.0	11.0	1.0
4481	99899322.0	2.0	0.0	2.0	1.0	0.0	1.0	50.0	43.0	0.0	...	1.0	0.0	0.0	0.0	3.0	3.0	0.0	0.0	13.0	1.0
4482	99948591.0	0.0	0.0	2.0	0.0	1.0	0.0	69.0	67.0	0.0	...	0.0	0.0	0.0	0.0	9.0	2.0	7.0	0.0	12.0	1.0
4483	99961100.0	2.0	0.0	3.0	1.0	0.0	1.0	54.0	48.0	0.0	...	0.0	0.0	0.0	0.0	2.0	2.0	0.0	0.0	19.0	1.0
4484	99981192.0	0.0	0.0	2.0	0.0	0.0	0.0	75.0	68.0	0.0	...	0.0	0.0	0.0	0.0	3.0	3.0	0.0	0.0	11.0	1.0

Patients_gyn

- Normalización mediante MinMaxScaler

	n_category	m_category	grade	ductal	lobular	neoadjuvant	age	diagnosis_age	is_dead	er	...	stage_diagnosis_IIIA	stage_diagnosis_IIIB	stage_diagnosis_IV	pregnancy	birth	caesarean	abort	menarche_age	has_menopause
0	0.000000	0.0	0.0	1.0	0.0	0.0	0.432432	0.413333	0.0	1.0	...	0.0	0.0	0.0	0.12	0.612903	0.518519	0.571429	0.461538	0.0
1	0.333333	0.0	1.0	0.0	1.0	1.0	0.608108	0.573333	0.0	1.0	...	1.0	0.0	0.0	0.00	0.516129	0.518519	0.571429	0.461538	1.0
2	0.000000	0.0	1.0	1.0	0.0	0.0	0.554054	0.506667	0.0	1.0	...	0.0	0.0	1.0	0.08	0.548387	0.518519	0.607143	0.807692	1.0
3	0.000000	1.0	1.0	1.0	0.0	0.0	0.554054	0.533333	0.0	1.0	...	0.0	0.0	0.0	0.08	0.516129	0.518519	0.642857	0.692308	1.0
4	0.333333	0.0	0.0	1.0	0.0	1.0	0.486486	0.440000	0.0	0.0	...	0.0	0.0	0.0	0.04	0.548387	0.518519	0.571429	0.576923	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4480	0.000000	0.0	1.0	0.0	1.0	1.0	0.540541	0.506667	0.0	1.0	...	0.0	0.0	0.0	0.16	0.612903	0.518519	0.607143	0.423077	1.0
4481	0.666667	0.0	0.5	1.0	0.0	1.0	0.324324	0.280000	0.0	1.0	...	1.0	0.0	0.0	0.12	0.612903	0.518519	0.571429	0.500000	1.0
4482	0.000000	0.0	0.5	0.0	1.0	0.0	0.581081	0.600000	0.0	1.0	...	0.0	0.0	0.0	0.36	0.580645	0.777778	0.571429	0.461538	1.0
4483	0.666667	0.0	1.0	1.0	0.0	1.0	0.378378	0.346667	0.0	1.0	...	0.0	0.0	0.0	0.08	0.580645	0.518519	0.571429	0.730769	1.0
4484	0.000000	0.0	0.5	0.0	0.0	0.0	0.662162	0.613333	0.0	1.0	...	0.0	0.0	0.0	0.12	0.612903	0.518519	0.571429	0.423077	1.0

Patients_gyn_norm

C r e a c i ó n c s v

- Creación de cuatro csv:
  - ‘recaida’: formado a partir del df ‘patients’
  - ‘recaida_norm’: formado a partir del df ‘patients_norm’
  - ‘recaida_gyn’: formado a partir del df ‘patients_gyn’
  - ‘recaida_gyn_norm’: formado a partir del df ‘patients_gyn_norm’

```
patients.to_csv('recaida.csv')
patients_norm.to_csv('recaida_norm.csv')
```

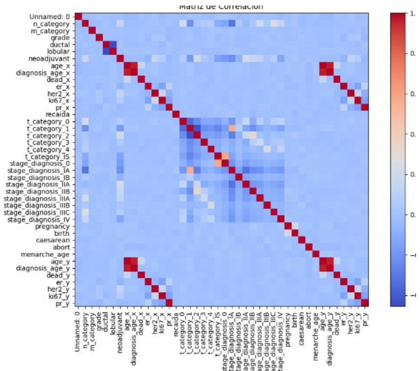
```
patients_gyn.to_csv('recaida_gyn.csv')
patients_gyn_norm.to_csv('recaida_gyn_norm.csv')
```

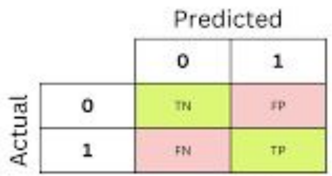
Creación modelo predictivo

- Utilizaremos estos modelo de ML para llevar a cabo estas predcciones:
  - Decision tree
  - Logistic Regression

- Métricas: Nos centraremos en conseguir un buen recall de la métrica de la clase recaída

- Análisis de correlación





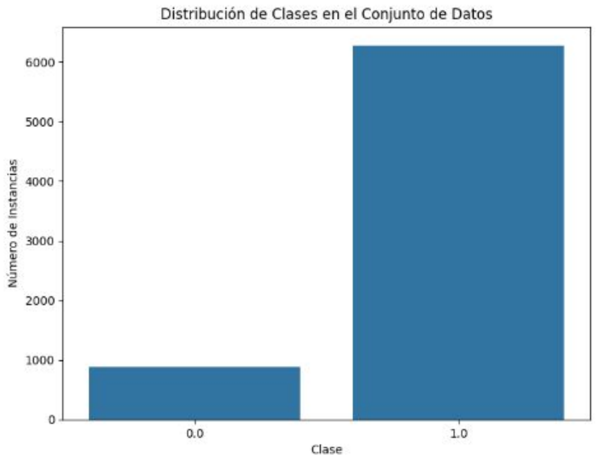
$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

Problema Desbalance

- Problema debido a que hay muchas menos recaídas que no recaídas

- Soluciones que aplicaremos:
  - Undersampling
  - Oversampling
  - Undersampling y Oversampling.
  - Balanced ensembles
  - Ajuste de pesos

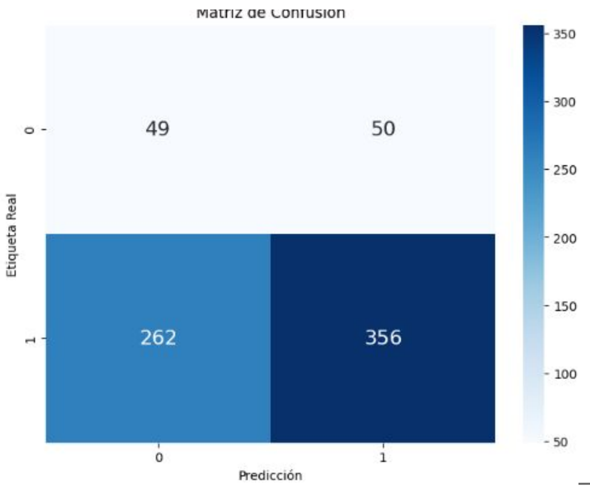


Modelo predictivo 1

- Modelo de regresión logística
- Añadimos la etiqueta de clases balanceadas
- - costes en seguimientos desaprovechado pero - recaída aceptada.

```
sampling =NearMiss()
X_res, y_res = sampling.fit_resample(X_train, y_train)

# Inicializar y entrenar el modelo de árbol de decisiones
class_weights = {0: 2, 1: 1}
modelo_arbol1 = DecisionTreeClassifier(random_state=42, class_weight=class_weights)
modelo_arbol1.fit(X_res, y_res)


# Predecir con el conjunto de validación
y_pred = modelo_arbol1.predict(X_test)
```

```
Informe de Clasificación Logistic regression:
              precision    recall  f1-score   support

         0.0       0.16      0.49      0.24        99
         1.0       0.88      0.58      0.70       618

    accuracy                           0.56       717
   macro avg       0.52      0.54      0.47       717
```



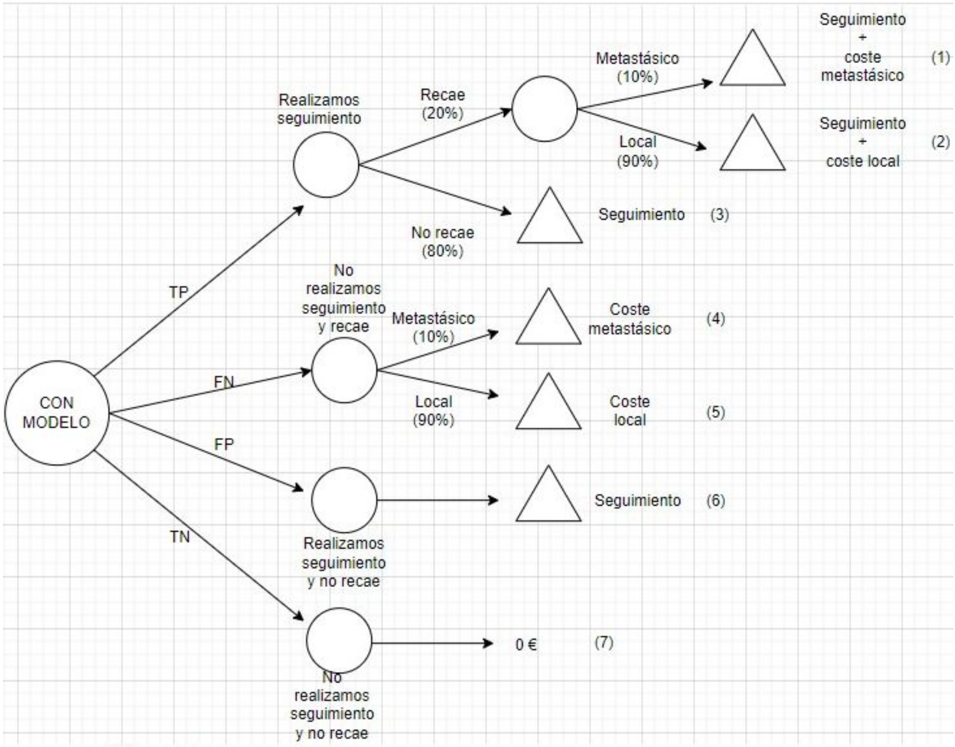
M o d e l o p r e d i c t i v o 2

- Modelo de regresión logística
- Añadimos la etiqueta de clases balanceadas
- \+ costes en seguimientos desaprovechado pero + recaída aceptada.

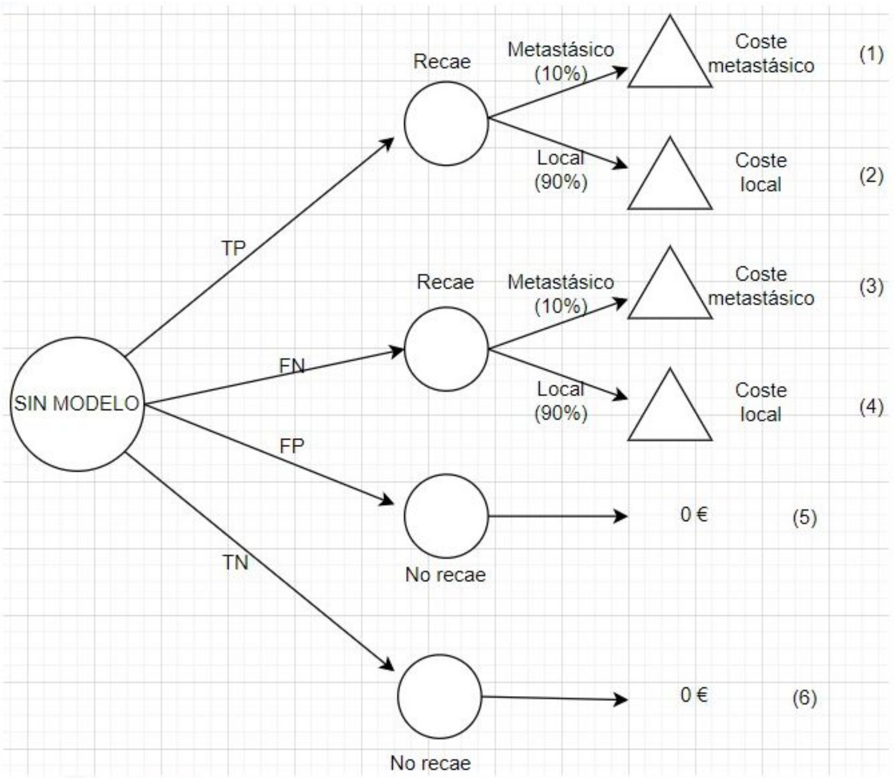
```
sampling =NearMiss()
X_res, y_res = sampling.fit_resample(X_train, y_train)

# Inicializar y entrenar el modelo de árbol de decisiones
class_weights = {0: 2, 1: 1}
modelo_arbol1 = DecisionTreeClassifier(random_state=42, class_weight=class_weights)
modelo_arbol1.fit(X_res, y_res)


# Predecir con el conjunto de validación
y_pred = modelo_arbol1.predict(X_test)
```

```
              precision    recall  f1-score   support

         0.0       0.14      0.57      0.22        99
         1.0       0.86      0.44      0.58       618

    accuracy                           0.46       717
   macro avg       0.50      0.50      0.40       717
weighted avg       0.76      0.46      0.53       717

Matriz de Confusión:
[[ 56  43]
 [345 273]]
Accuracy: 0.45885634588563456
```

Árbol con Modelo



Árbol sin Modelo



C o s t e s


```
coste_total_metastático = coste_medio_anual_metastático * duracion_metastático + coste_cuidados_paliati
coste_total_metastático
```

365152

```
coste_total_local = coste_medio_anual_local * duracion_local + coste_cuidados_paliativos
coste_total_local
```

49187

Aplicando el modelo dado a 717 clientes y con las metricas del modelo dadas, hay una reducción de 753 286.7999999998 euros.

Aplicando el modelo dado a 717 clientes y con las metricas del modelo dadas, hay una reducción de 860 899.1999999993 euros.

C o n c l u s i o n e s

- Objetivos de negocio -> se tiene que priorizar recursos resultados.
- Objetivos de CD -> hay que resolver y optimizar los datos, pero NO HAY MAGIA
- Hay que ponerse objetivos tanto económicos como de resultados-

Referencias Presentación

- Datos de costes (diapositiva 2)
  - https://www.contraelcancer.es/sites/default/files/content-file/Informe-Los-costes-cancer.pdf
- Porcentaje mortalidad y tipos de cáncer de mama (diapositiva 3)
  - https://www.cancer.net/es/tipos-de-c%C3%A1ncer/c%C3%A1ncer-de-mama/estad%C3%ADsticas
- Toxicidad financiera cáncer (diapositiva 3)
  - https://observatorio.contraelcancer.es/sites/default/files/informes/Toxicidad_financiera_cancer_mama.pdf

MUCHAS GRACIAS POR LA ATENCIÓN !!!

Objetivos

Preprocessing

Análisis
Recaída

Árbol
Decisiones

Conclusiones

Referencias

Pablo Martín Escobar, Jorge Sáenz de Miera Marzo, Javier Gil Domínguez, Nicolás Vega Muñoz