

Sistemas de Información y Telemedicina II*

Machine Learning para Mesotelioma Maligno

Irene Estela García García

Ilán Francisco Carretero Juchnowicz

Ignacio Maria Amat Hernández

April 20, 2020

*Grado en Ingeniería Biomédica, Escuela Técnica Superior de Ingenieros Industriales, Valencia, España.

Contents

1	Análisis exploratorio	4
1.1	Tipos de datos	4
1.2	Caja bigotes	5
1.3	Histograma	6
1.4	Kernel Density	7
1.5	Cuantil - cuantil	8
1.6	Correlaciones	9
2	Extracción de Características	10
2.1	Filter methods	10
2.2	Wrapper methods	11
2.3	PCA	12
3	Modelos de Clasificación	13
4	Interfaz	14
4.1	Interfaz sin opciones	17

List of Figures

1	Caja bigotes.	5
2	Histograma.	6
3	Kernel Density.	7
4	Cuantil - cuantil.	8
5	Correlaciones.	9
6	Diagrama de pareto.	12
7	Ventana principal.	14
8	Selección de paciente.	14
9	Importación paciente.	15
10	Cargar variables.	15
11	Selección de modelos.	16
12	Selección de modelos.	16
13	Versión limitada para médicos.	17

Listings

1	Variables de la base de datos.	4
2	Selección de variables según fscore	10

3	Mejores variables según <i>SFS</i> y <i>SBS</i>	11
4	Modelos de clasificación en Python	13

1 Análisis exploratorio

1.1 Tipos de datos

Comenzamos analizando la base de datos, consta de las siguientes variables:

```
1 RangeIndex: 324 entries, 0 to 323
2 Data columns (total 30 columns):
3 #      Column                                Non-Null Count  Dtype
4 ---  -
5 0      age                                     324 non-null    float64
6 1      gender                                324 non-null    int64
7 2      city                                  324 non-null    int64
8 3      asbestos exposure                     324 non-null    int64
9 4      duration of asbestos exposure         324 non-null    float64
10 5      keep side                             324 non-null    int64
11 6      duration of symptoms                  324 non-null    float64
12 7      dyspnoea                             324 non-null    int64
13 8      ache on chest                         324 non-null    int64
14 9      weakness                              324 non-null    int64
15 10     habit of cigarette                   324 non-null    int64
16 11     performance status                  324 non-null    int64
17 12     white blood                         324 non-null    float64
18 13     cell count (WBC)                    324 non-null    int64
19 14     hemoglobin (HGB)                    324 non-null    int64
20 15     platelet count (PLT)                 324 non-null    float64
21 16     sedimentation                       324 non-null    float64
22 17     blood lactic dehydrogenise (LDH)    324 non-null    float64
23 18     alkaline phosphatise (ALP)           324 non-null    float64
24 19     total protein                        324 non-null    float64
25 20     albumin                             324 non-null    float64
26 21     glucose                              324 non-null    float64
27 22     pleural lactic dehydrogenise         324 non-null    float64
28 23     pleural protein                      324 non-null    float64
29 24     pleural albumin                      324 non-null    float64
30 25     pleural glucose                      324 non-null    float64
31 26     pleural effusion                     324 non-null    float64
32 27     pleural thickness on tomography      324 non-null    float64
33 28     pleural level of acidity (pH)        324 non-null    float64
34 29     C-reactive protein (CRP)             324 non-null    int64
35 dtypes: float64(18), int64(12)
36 memory usage: 76.1 KB
```

Listing 1: Variables de la base de datos.

Nuestra base de datos consta de 324 entradas, cada una con 30 variables. Todos los valores son `floats` e `ints`, además no tenemos ningún `NULL`.

1.2 Caja bigotes

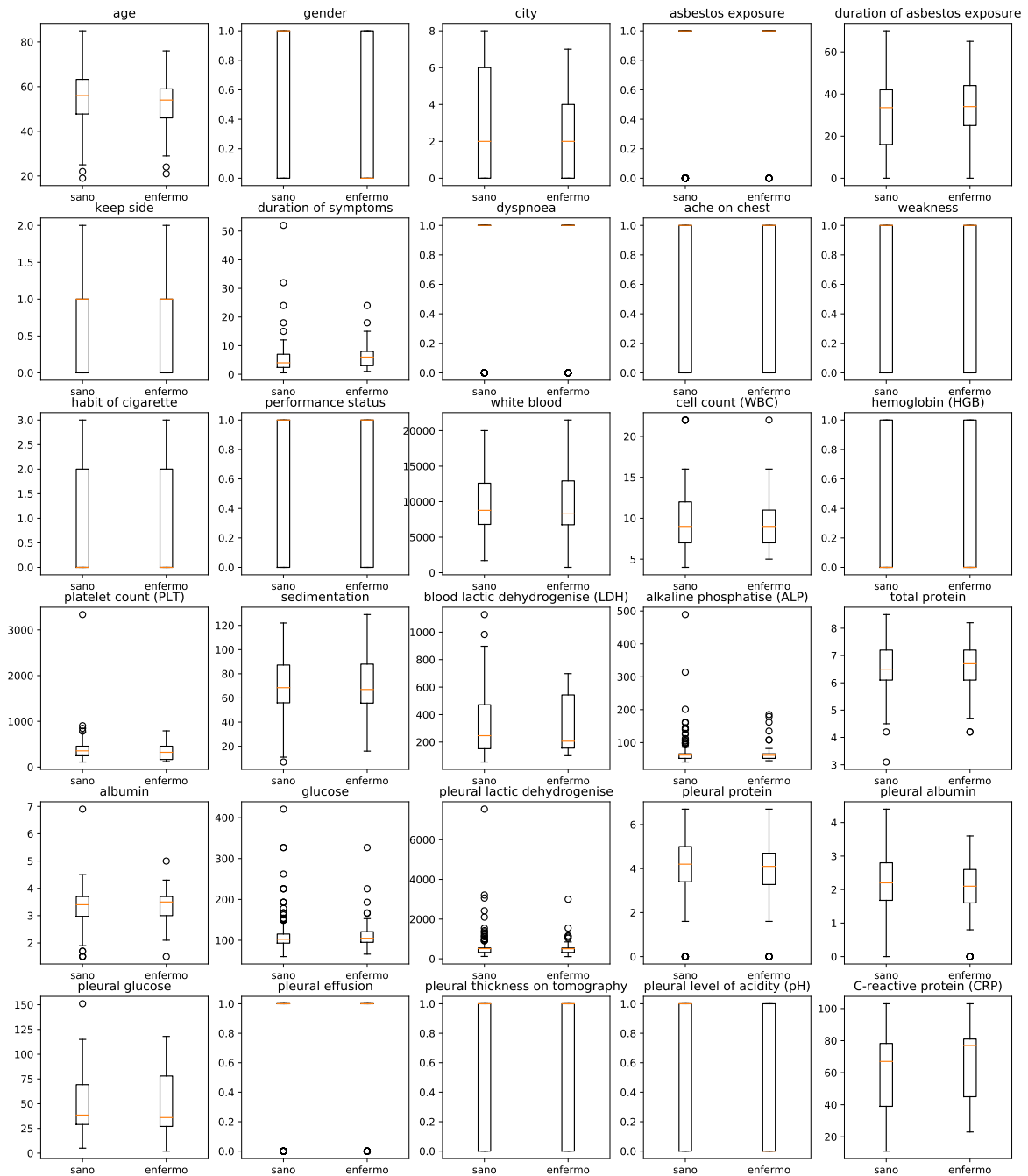


Fig. 1: Caja bigotes.

Vemos que no hay diferencias notables entre los diagramas de nuestra población de pacientes sanos y enfermos. Las variables PLT, PLT, glucosa y PLD parecen tener datos anómalos.

1.3 Histograma



Fig. 2: Histograma.

Como es de esperar tras juzgar los diagramas de caja bigotes, los histogramas también tienen una distribución casi idéntica. Lo único en lo que se diferencian es en la cuentas totales, ello indica que tenemos más observaciones de pacientes sanos que de enfermos.

1.4 Kernel Density

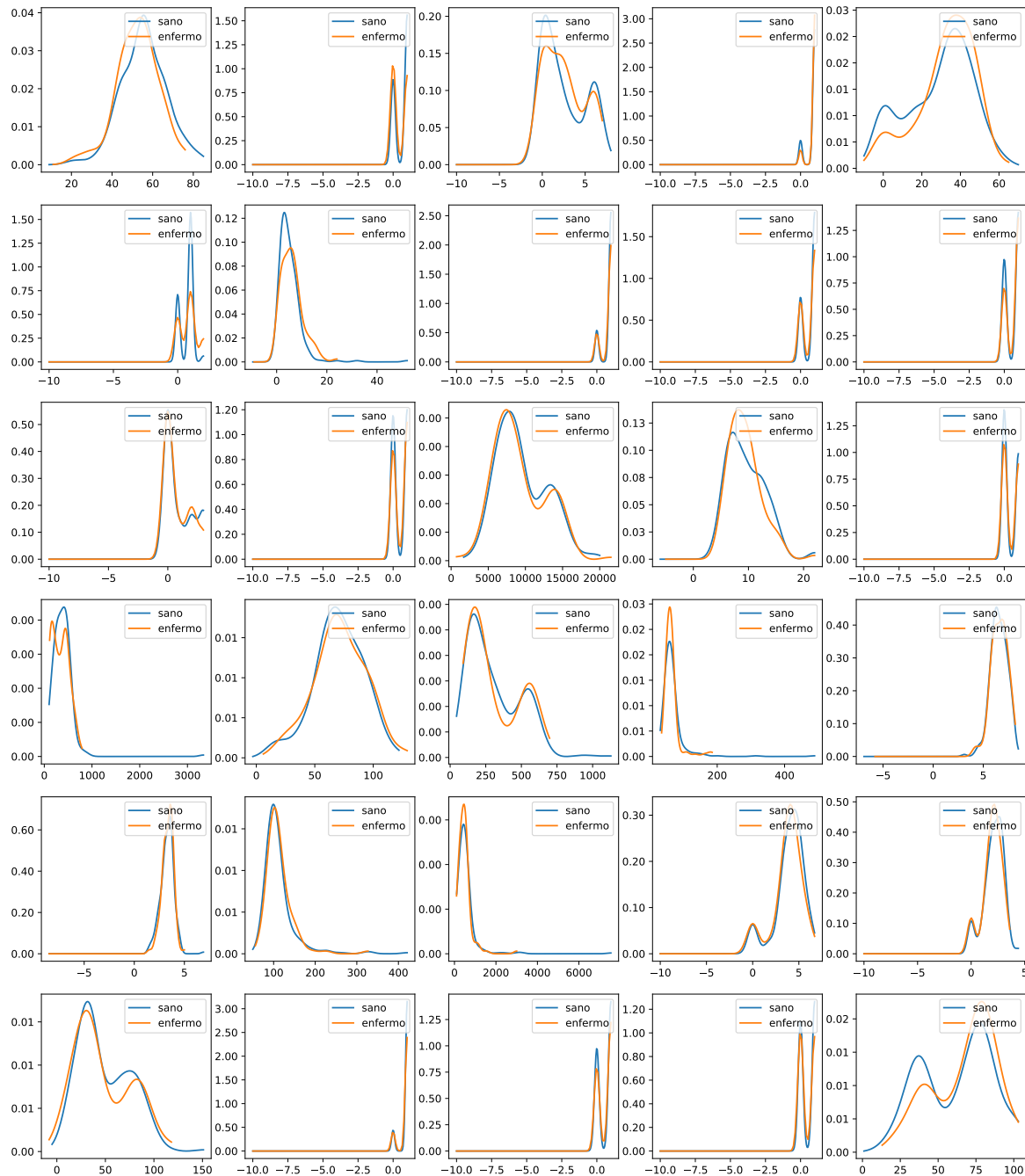


Fig. 3: Kernel Density.

A diferencia de los histogramas, al tratarse con densidades no se observan diferencias por la distinta cantidad de observaciones de sanos y enfermos. Únicamente vemos que las distribuciones son casi idénticas, como intuíamos de los histogramas.

1.5 Cuantil - cuantil

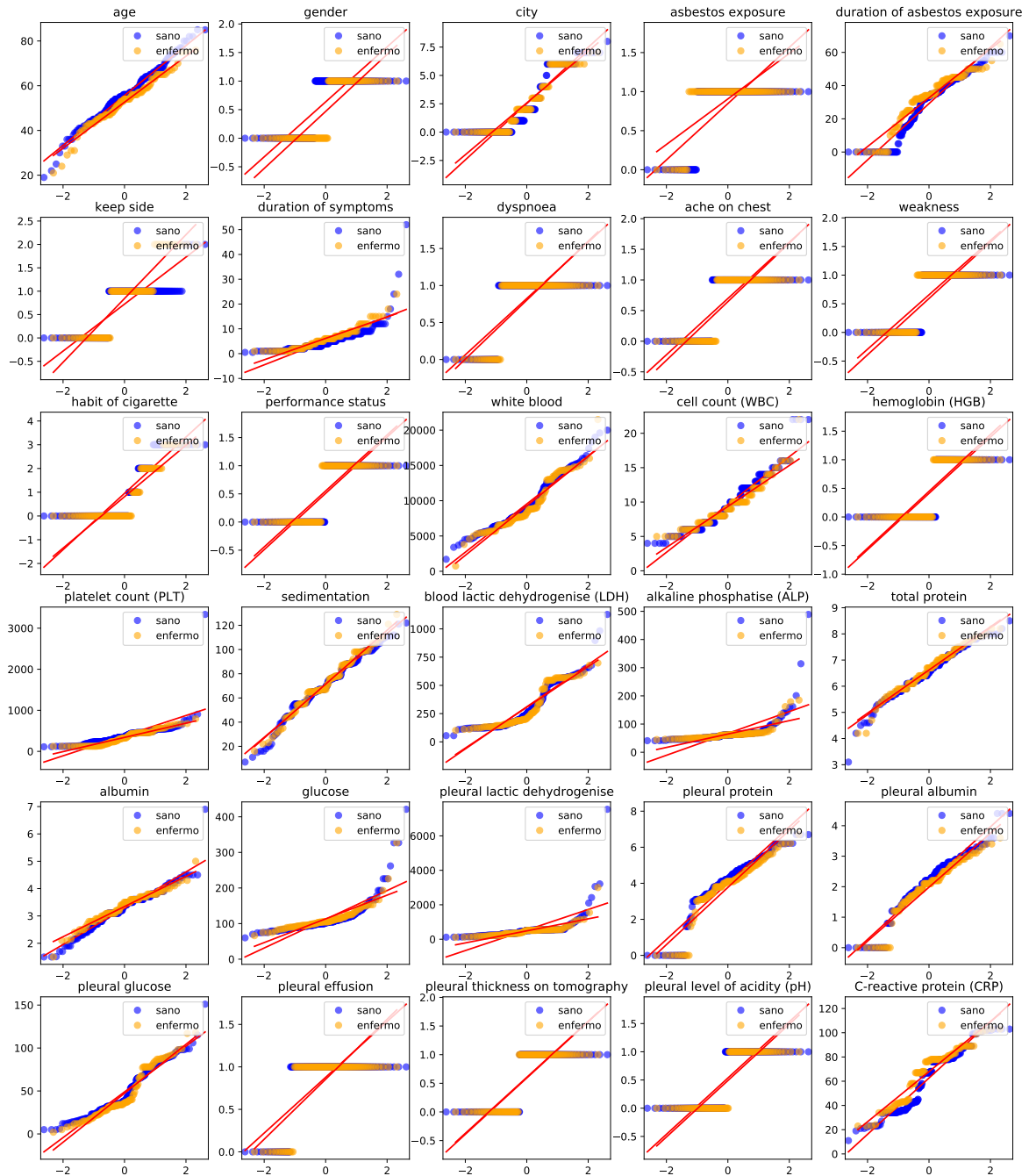


Fig. 4: Cuantil - cuantil.

1.6 Correlaciones

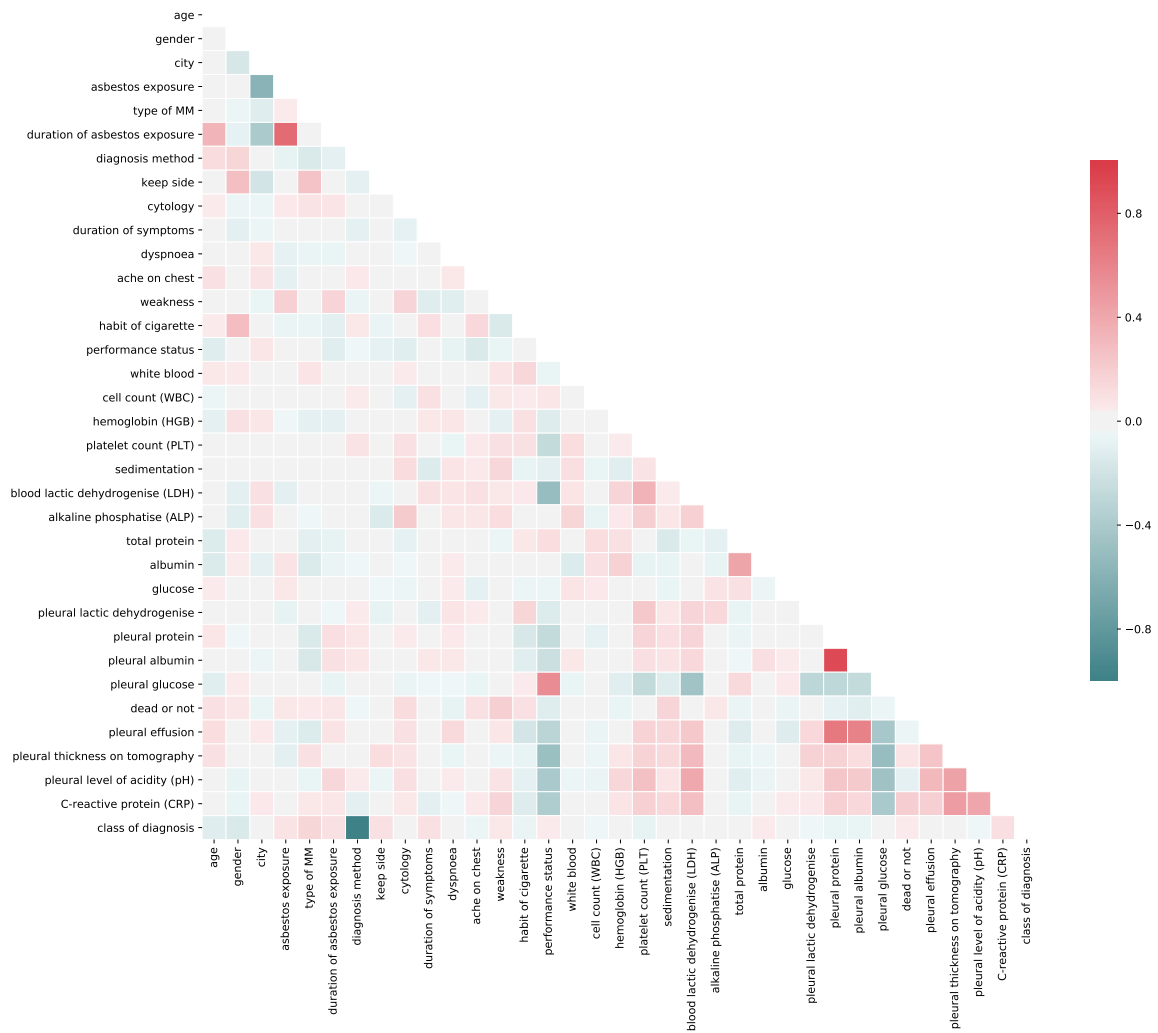


Fig. 5: Correlaciones.

2 Extracción de Características

2.1 Filter methods

Ranking	Variable
-----	-----
1	pleural level of acidity (pH)
2	C-reactive protein (CRP)
3	gender
4	pleural lactic dehydrogenise
5	pleural effusion
6	pleural glucose
7	pleural albumin
8	keep side
9	blood lactic dehydrogenise (LDH)
10	total protein
11	alkaline phosphatase (ALP)
12	white blood
13	performance status
14	cell count (WBC)
15	habit of cigarette
16	pleural protein
17	duration of asbestos exposure
18	city
19	dyspnoea
20	ache on chest
21	sedimentation
22	asbestos exposure
23	platelet count (PLT)
24	glucose
25	albumin
26	duration of symptoms
27	weakness
28	age
29	hemoglobin (HGB)
30	pleural thickness on tomography

Listing 2: Selección de variables según **fscore**

Usando la puntuación de *Fisher* clasificamos las características de mayor a menor relevancia a la hora de resolver el problema de clasificación

2.2 Wrapper methods

```
1 0.6782828282828283
2 Sequential Forward Selection ('asbestos exposure',
3                               'keep side',
4                               'weakness',
5                               'cell count (WBC)',
6                               'platelet count (PLT)',
7                               'alkaline phosphatase (ALP)',
8                               'glucose',
9                               'pleural protein',
10                              'pleural glucose',
11                              'C-reactive protein (CRP)')
12
13 0.5414285714285715
14 Sequential Backward Selection ('city',
15                                'asbestos exposure',
16                                'keep side',
17                                'duration of symptoms',
18                                'ache on chest',
19                                'performance status',
20                                'platelet count (PLT)',
21                                'alkaline phosphatase (ALP)',
22                                'pleural albumin',
23                                'C-reactive protein (CRP)')
```

Listing 3: Mejores variables según SFS y SBS

Una selección secuencial hacia adelante y hacia atrás con un modelo de `randomforest` con 100 estimadores y 10 parámetros calculamos una precisión del 67%. Los resultados son menores cuando hacemos la selección hacia atrás. También lo hemos intentado con un modelo `knn`, los resultados son peores en torno al 50%.

2.3 PCA

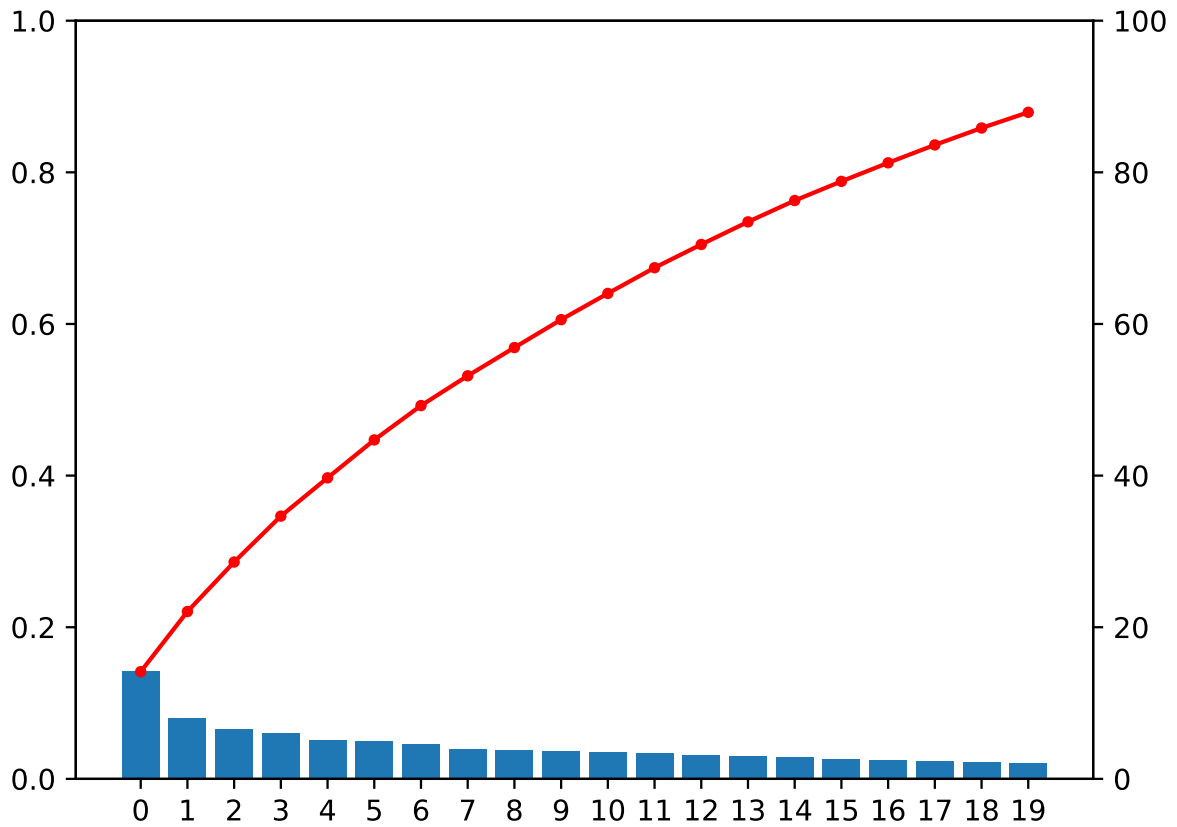


Fig. 6: Diagrama de pareto.

Calculando el diagrama de pareto vemos que necesitaríamos entorno a 17 componentes para explicar el 80% de la varianza de los datos. Con estos resultados vemos que será difícil reducir la dimensionalidad en componentes principales.

3 Modelos de Clasificación

```
1 from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
2
3 modeloLDA = LinearDiscriminantAnalysis ();
4
5 from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis
6
7 modeloQDA = QuadraticDiscriminantAnalysis ();
8
9 from sklearn.neighbors import KNeighborsClassifier
10
11 modeloKNN = KNeighborsClassifier (n_neighbors = 50)
12
13 from sklearn.ensemble import RandomForestClassifier
14
15 modelFOREST = RandomForestClassifier (
16     n_estimators = 100,
17     criterion = 'gini',
18 )
19
20 from neupy import algorithms
21
22 modeloPNN = algorithms.PNN (
23     std=5,
24     verbose=False,
25 )
26
27 from sklearn.neural_network import MLPClassifier as MLP
28
29 modeloMLP = MLP(
30     hidden_layer_sizes = (175, 100, 50, 25, ),
31     max_iter = 500,
32     random_state = 1)
33
34 from sklearn import svm
35
36 modeloSVM = svm.LinearSVC()
```

Listing 4: Modelos de clasificación en Python

Model	TP	FP	FN	TN	Accuracy	Sensitivity	Specificity	Time
LDA	58.45	10.60	21.31	7.63	0.67	0.73	0.42	5.21
QDA	55.93	12.94	22.43	6.70	0.64	0.71	0.34	4.58
KNN	68.99	0.00	29.01	0.00	0.70	0.70	nan	9.74
FOREST	66.55	2.54	23.96	4.95	0.73	0.74	0.66	181.79
SVM	45.93	22.88	19.88	9.31	0.56	0.70	0.29	20.90
PNN	61.88	7.06	25.27	3.78	0.67	0.71	0.35	9.15
MLP	48.57	20.53	20.42	8.48	0.58	0.70	0.29	287.44

Table 1: Resultados agregados tras 1000 repeticiones.

4 Interfaz

Mesothelioma Checker

Introduce las variables:

age (0 - 100)
input var 0

gender (0 - 1)
input var 0

city (0 - 8)
input var 0

asbestos exposure (0 - 1)
input var 0

duration
input var 0

exposure (0 - 70)
input var 0

keep side (0 - 2)
input var 0

duration
input var 0

symptoms (0.5 - 52)
input var 0

Lee Vars

Modelo

Verdaderos Positivos

Falsos Positivos

Falsos Negativos

Verdaderos Negativos

Exactitud

Sensitividad

Especificidad

Diagnóstico

☐ LDA
☐ QDA
☐ KNN
☐ FOREST
☐ SVM
☐ PNN
☐ MLP

Fig. 7: Ventana principal.

Ventana principal de nuestra interfaz. A la izquierda campos de texto para introducir las variables, se muestra entre paréntesis el rango de dicha variable en nuestra base de datos.

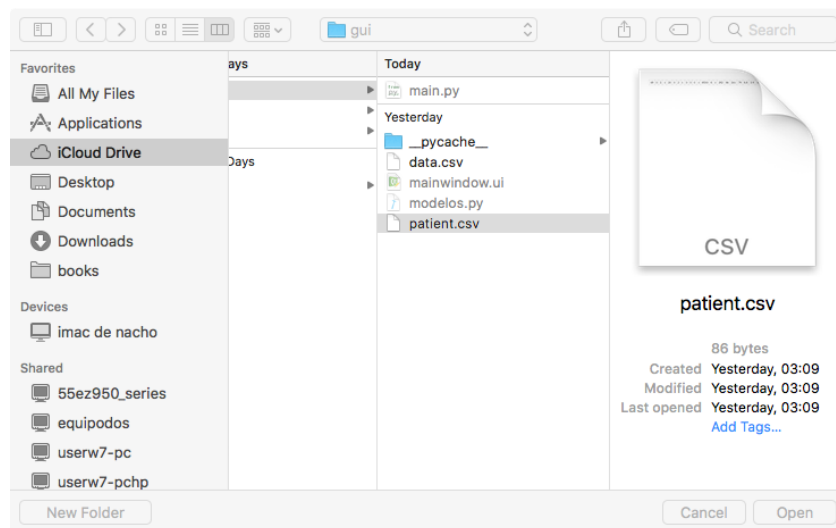


Fig. 8: Selección de paciente.

Para hacer mas cómodo su uso y evitar tener que introducir 30+ variables a mano, al presionar Ctrl + O se abre una ventana para seleccionar un archivo .csv, contiene todos los valores para un paciente separados por comas.

Mesothelioma Checker

Introduce las variables:

age (0 - 100)	<input type="text" value="49"/>	<input type="text" value="0"/>
gender (0 - 1)	<input type="text" value="0"/>	<input type="text" value="0"/>
city (0 - 8)	<input type="text" value="2"/>	<input type="text" value="0"/>
asbestos exposure (0 - 1)	<input type="text" value="1"/>	<input type="text" value="0"/>
duration		
exposure (0 - 70)	<input type="text" value="20"/>	<input type="text" value="0"/>
keep side (0 - 2)	<input type="text" value="0"/>	<input type="text" value="0"/>
duration		
symptoms (0.5 - 52)		<input type="text" value="0"/>

Modelo

☐ LDA

☐ QDA

☐ KNN

☐ FOREST

☐ SVM

☐ PNN

☐ MLP

Verdaderos Positivos

Falsos Positivos

Falsos Negativos

Verdaderos Negativos

Exactitud

Sensitividad

Especificidad

Diagnóstico

Fig. 9: Importación paciente.

Al importar el paciente se rellenan automáticamente las variables, se pueden comprobar y editar si fuera necesario.

Mesothelioma Checker

Introduce las variables:

age (0 - 100)	<input type="text" value="input var"/>	<input type="text" value="49"/>
gender (0 - 1)	<input type="text" value="input var"/>	<input type="text" value="0"/>
city (0 - 8)	<input type="text" value="input var"/>	<input type="text" value="2"/>
asbestos exposure (0 - 1)	<input type="text" value="input var"/>	<input type="text" value="1"/>
duration		
exposure (0 - 70)	<input type="text" value="input var"/>	<input type="text" value="20"/>
keep side (0 - 2)	<input type="text" value="input var"/>	<input type="text" value="0"/>
duration		
symptoms (0.5 - 52)		<input type="text" value="2"/>

Modelo

☐ LDA

☐ QDA

☐ KNN

☐ FOREST

☐ SVM

☐ PNN

☐ MLP

Verdaderos Positivos

Falsos Positivos

Falsos Negativos

Verdaderos Negativos

Exactitud

Sensitividad

Especificidad

Diagnóstico

Fig. 10: Cargar variables.

Al presionar Enter con el teclado o el botón 'Leer Vars' con el ratón se guardan las variables en la memoria del programa. Para actualizar una variable no hay más que introducir un nuevo valor y pulsar enter o el ratón. Para cargar el modelo se presiona Ctrl + M, o se elige desde el menú. Aparece una venta para seleccionar el archivo .py con los modelos en python.

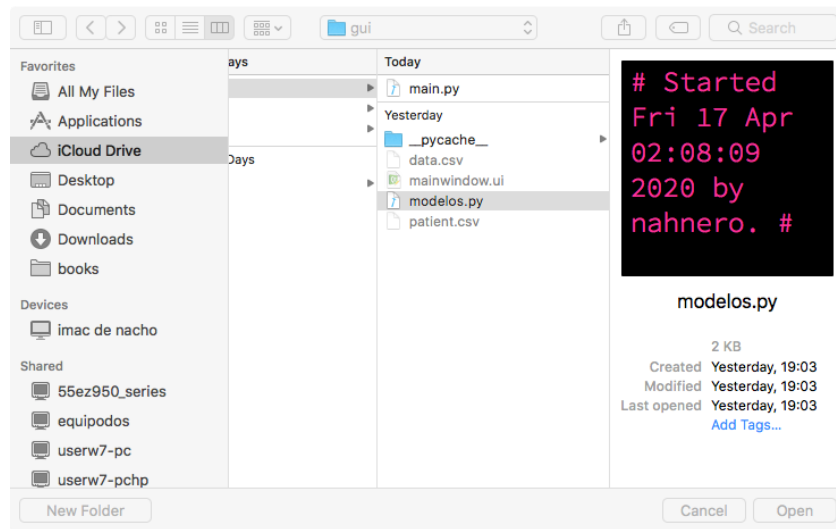


Fig. 11: Selección de modelos.

Una vez cargados los modelos se selecciona que modelo se desea ejecutar, los botones son mutuamente excluyentes. Al ejecutarse el modelo se generan las métricas de evaluación a la derecha y el diagnóstico.

Introduce las variables:

age (0 - 100)
 49

gender (0 - 1)
 0

city (0 - 8)
 2

asbestos exposure (0 - 1)
 1

duration
 20

exposure (0 - 70)
 20

keep side (0 - 2)
 0

duration
 2

symptoms (0.5 - 52)
 2

Lee Vars

Mesothelioma Checker

☒ LDA
 ☐ QDA
 ☐ KNN
 ☐ FOREST
 ☐ SVM
 ☐ PNN
 ☐ MLP

Modelo	LDA
Verdaderos Positivos	44.00
Falsos Positivos	5.00
Falsos Negativos	12.00
Verdaderos Negativos	4.00
Exactitud	0.74
Sensitividad	0.79
Especificidad	0.44
Diagnóstico	enfermo

Fig. 12: Selección de modelos.

Cuando se selecciona otro modelo se reevalúa al paciente, cada modelo tiene unas características diferentes en cuanto a exactitud, sensibilidad y especificidad. En función de lo que se busque se puede seleccionar un modelo adecuado.

4.1 Interfaz sin opciones

Esta versión restringe al usuario a usar un modelo predeterminado. Al pulsar el botón ejecutar, después de haber introducido los datos del paciente se calcula el modelo y se presentan los resultados.

Mesothelioma Checker

Introduce las variables:

age (0 - 100)

49

gender (0 - 1)

0

city (0 - 8)

2

asbestos exposure (0 - 1)

1

duration

exposure (0 - 70)

20

keep side (0 - 2)

0

duration

symptoms (0.5 - 52)

2

Lee Vars

Ejecutar

Modelo	FOREST
Verdaderos Positivos	43.00
Falsos Positivos	3.00
Falsos Negativos	17.00
Verdaderos Negativos	2.00
Exactitud	0.69
Sensitividad	0.72
Especificidad	0.40
Diagnóstico	enfermo

Fig. 13: Versión limitada para médicos.