Sistemas de Información y Telemedicina II*

Machine Learning para Mesotelioma Maligno

Irene Estela García García Ilán Francisco Carretero Juchnowicz

Ignacio Maria Amat Hernández

April 17, 2020

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

Contents

List of Figures

Listings

1 Análisis exploratorio

1.1 Tipos de datos

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

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

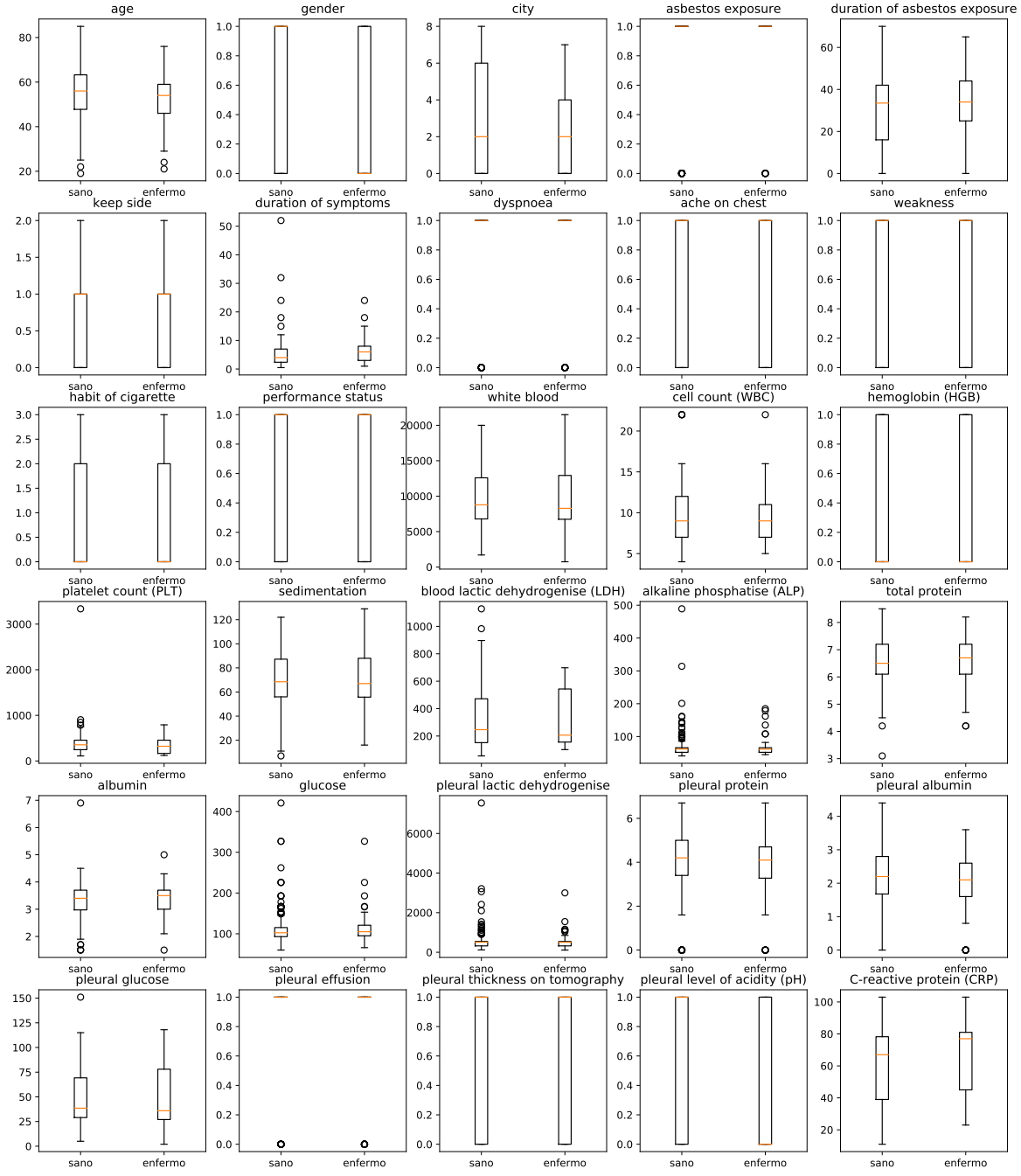

Fig. 1: Caja bigotes.

Vemos que no hay diferencias notables entre los diagramas de nuestra población de pacientes sanos y enferemos. 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

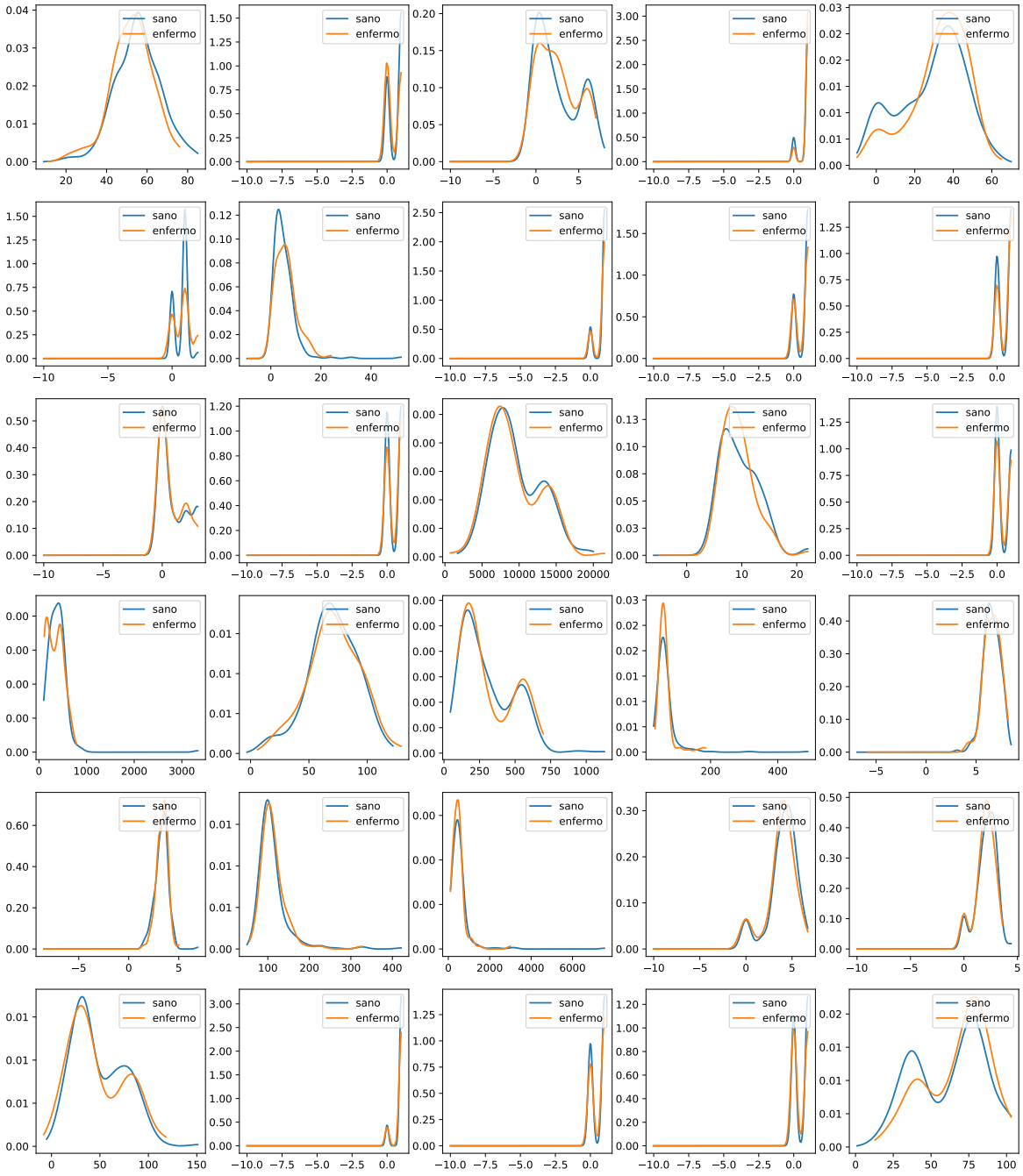

Fig. 3: Kernel Density.

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

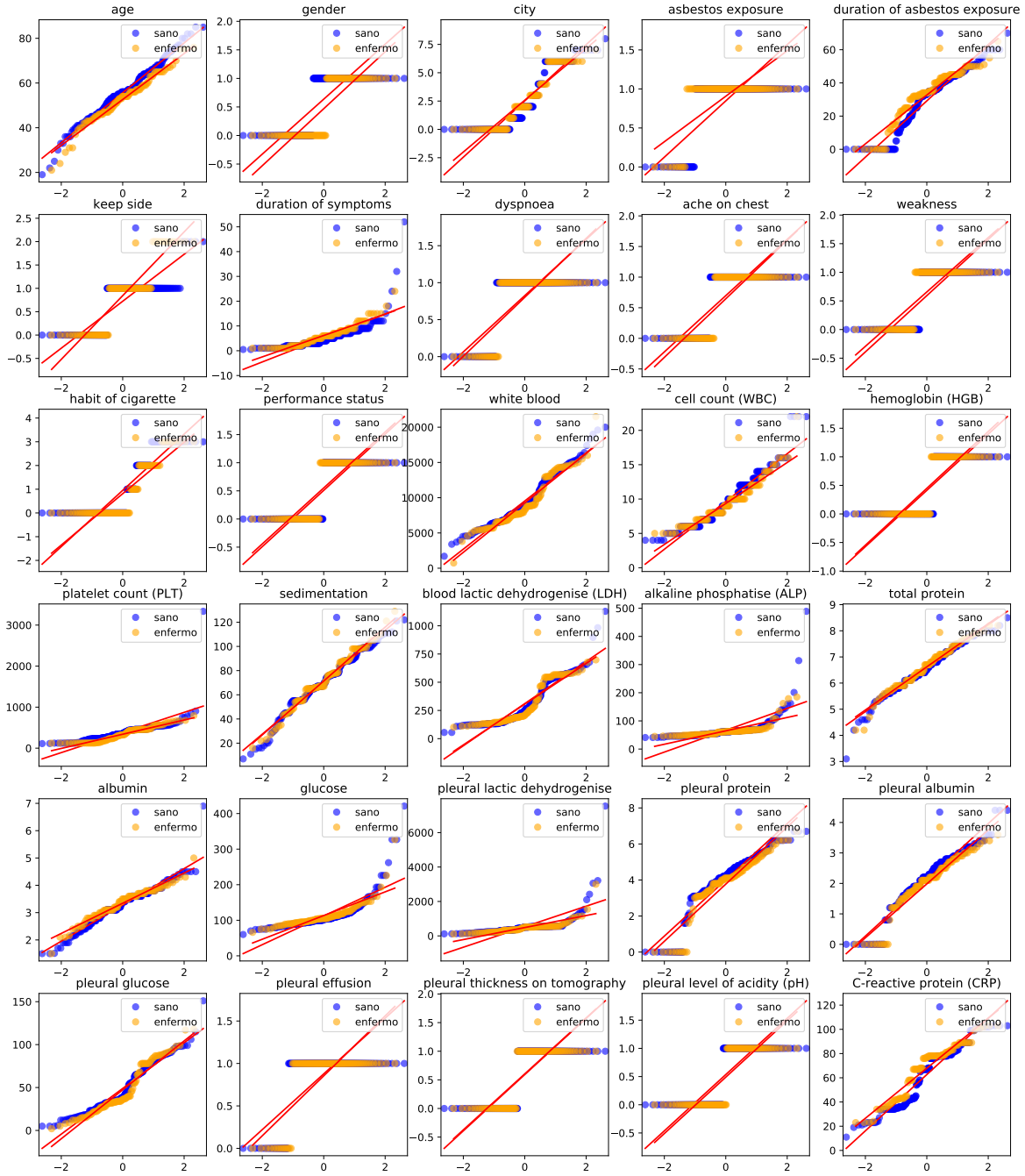

Fig. 4: Cuantil - cuantil.

1.6 Correlaciones

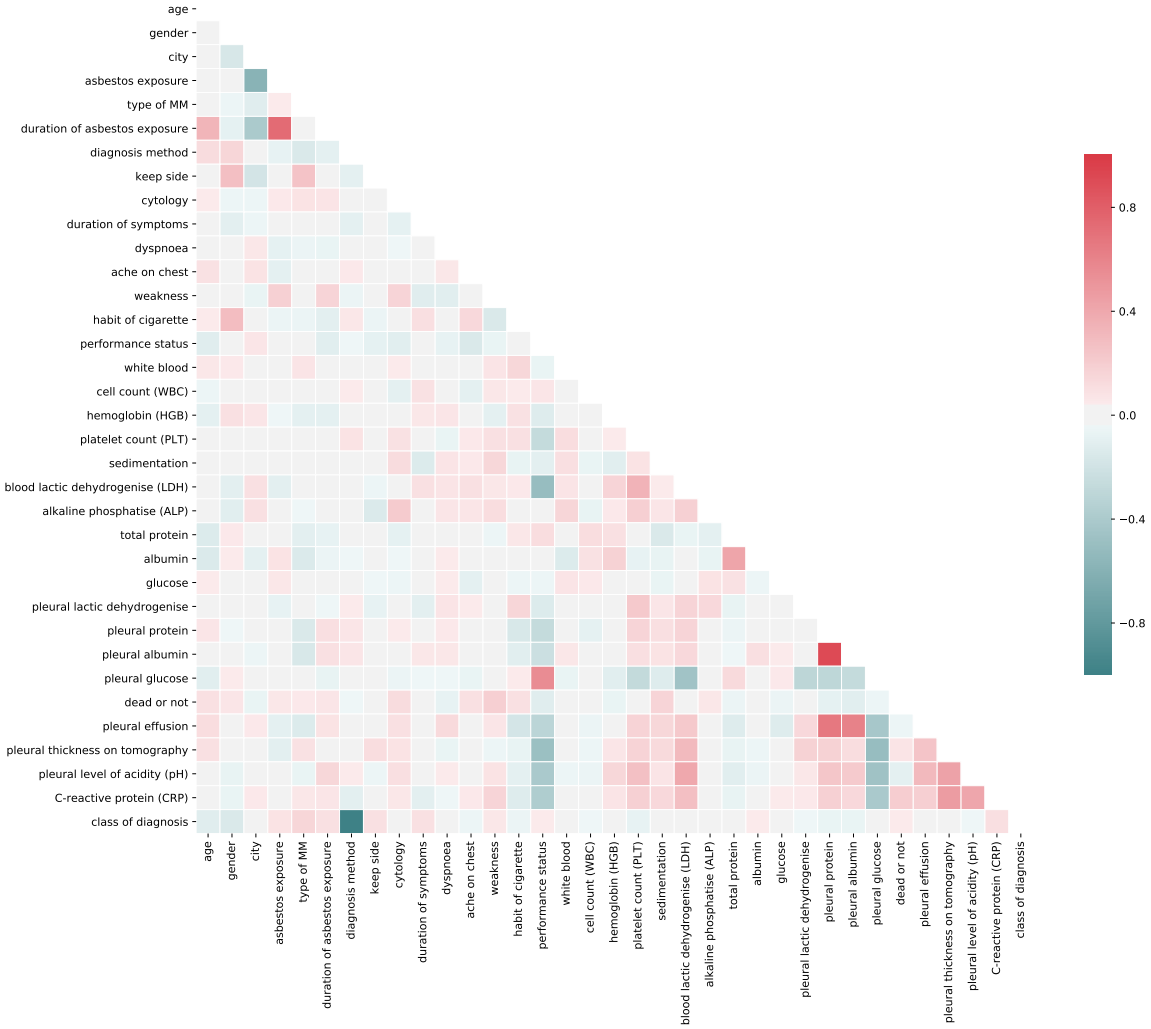

Fig. 5: Correlaciones.

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 phosphatise (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

```
0.6782828282828283
Sequential Forward  Selection ('asbestos exposure',
                               'keep side',
                               'weakness',
                               'cell count (WBC)',
                               'platelet count (PLT)',
                               'alkaline phosphatise (ALP)',
                               'glucose',
                               'pleural protein',
                               'pleural glucose',
                               'C-reactive protein (CRP)')

0.5414285714285715
Sequential Backward Selection ('city',
                               'asbestos exposure',
                               'keep side',
                               'duration of symptoms',
                               'ache on chest',
                               'performance status',
                               'platelet count (PLT)',
                               'alkaline phosphatise (ALP)',
                               'pleural albumin',
                               'C-reactive protein (CRP)')
```

Listing 3: Mejores variables según `SFS` y `SBS`

Una selección secuencial hacia adelante y hacia atras 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 atras. También lo hemos intentado con un modelo `knn`, los resultados son peores en torno al 50%.

2.3 PCA

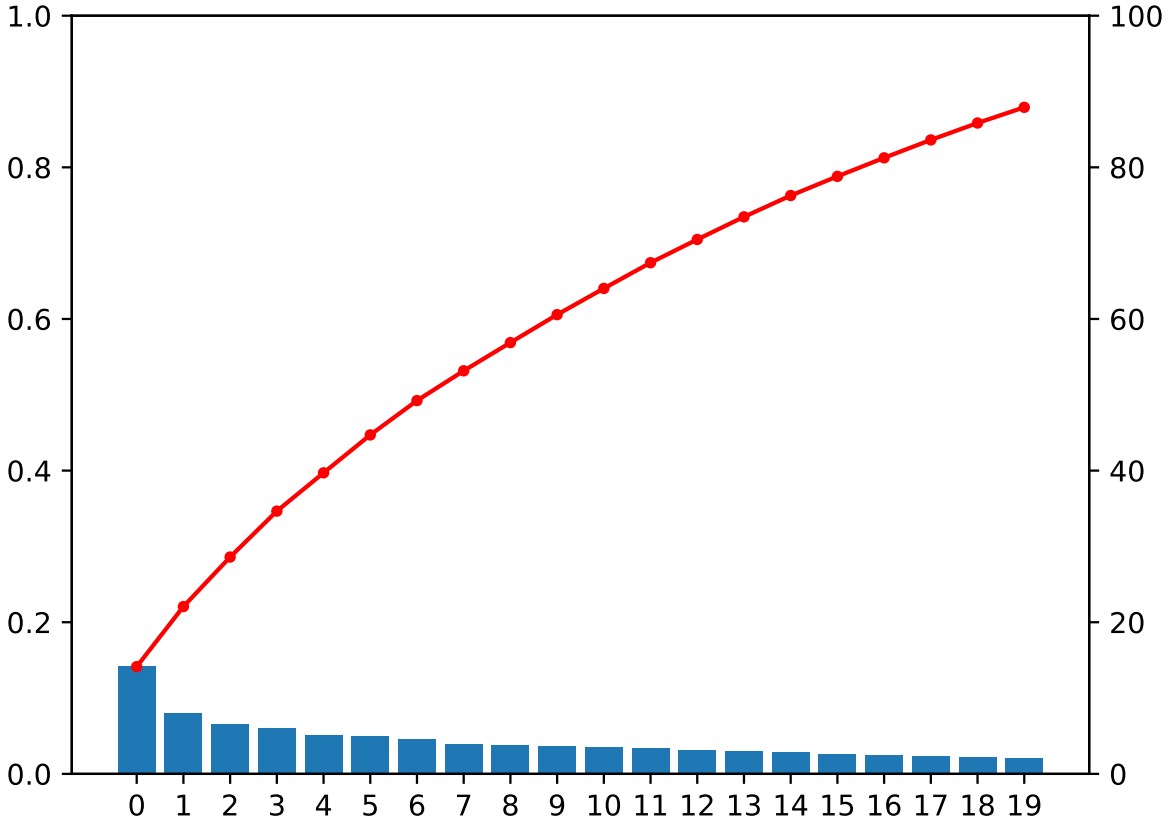

Fig. 6: Diagrama de pareto.

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 dificil reducir al dimensionalidad encomponentes principales.

3 Modelos de Clasificación

```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis

modeloLDA = LinearDiscriminantAnalysis ();

from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis

modeloQDA = QuadraticDiscriminantAnalysis ();

from sklearn.neighbors import KNeighborsClassifier

modeloKNN   = KNeighborsClassifier (n_neighbors = 50)

from sklearn.ensemble import RandomForestClassifier

modelFOREST = RandomForestClassifier (
        n_estimators = 100,
        criterion = 'gini',
        )

from neupy import algorithms

modeloPNN = algorithms.PNN (
        std=5,
        verbose=False,
        )

from sklearn.neural_network import MLPClassifier as MLP

modeloMLP = MLP(
        hidden_layer_sizes = (175, 100, 50, 25, ),
        max_iter = 500,
        random_state = 1)

from sklearn import svm

modeloSVM = svm.LinearSVC()
```

Listing 4: Modelos de clasificación en Python

Model	TP	FP	FN	TN	Accuracy	Sensitivity	Specificity	Time
LDA	58.36	10.54	21.38	7.72	0.67	0.27	0.73	5.48
QDA	56.14	12.87	22.11	6.89	0.64	0.24	0.72	5.00
KNN	68.83	0.00	29.17	0.00	0.70	0.00	0.70	10.66
FOREST	66.51	2.48	23.95	5.05	0.73	0.17	0.74	196.52
SVM	46.30	22.67	19.78	9.26	0.57	0.32	0.70	22.81
PNN	61.90	7.07	25.25	3.78	0.67	0.13	0.71	9.75
MLP	47.88	21.03	20.25	8.85	0.58	0.30	0.70	295.27

Table 1: Resultados agregados tras 1000 repeticiones.