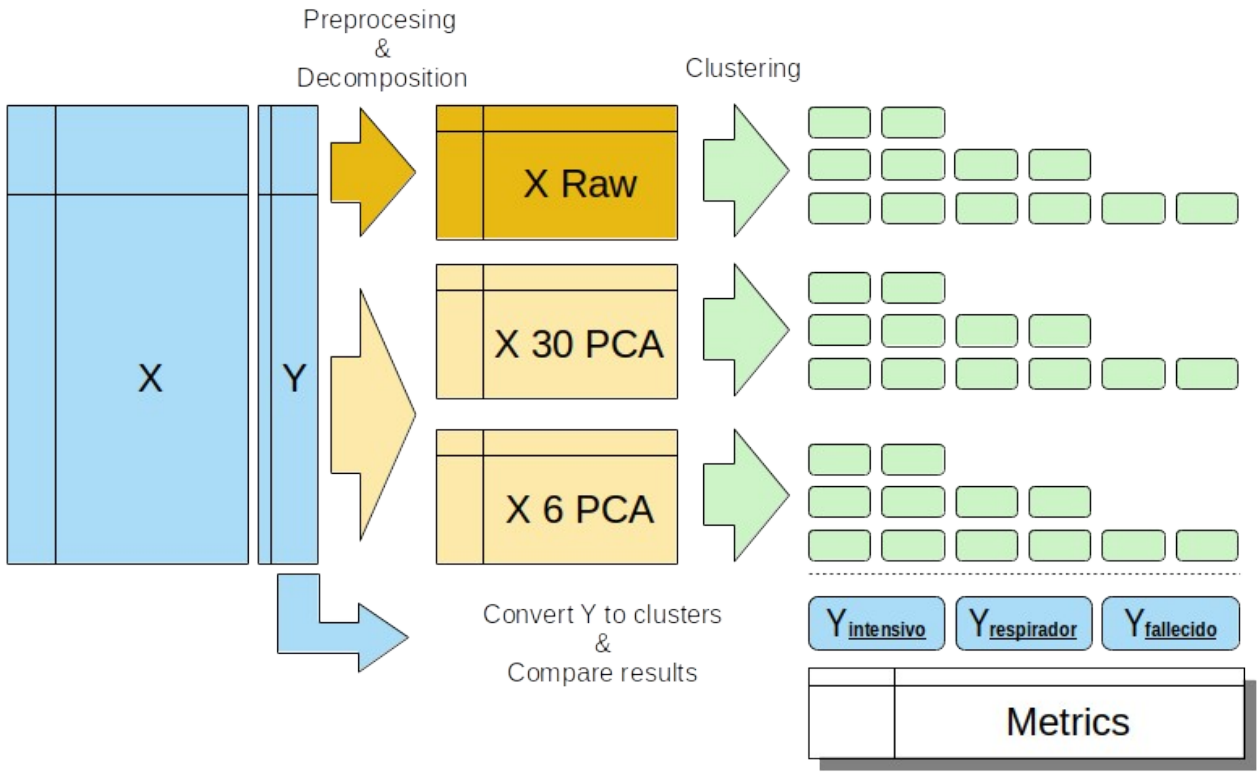
Introducción
En el presente apartado se busca conocer en mayor profundidad y detalle el comportamiento de los samples de nuestro dataset. Para esto se propone extender el Análisis Exploratorio de nuestros datos por medio de técnicas de aprendizaje no supervisado y exponer características que no se hacen evidentes simplemente aplicando técnicas más directas.

Metodología
Se pretende interpretar nuestros datos desde diferentes perspectivas para evaluar si los resultados del proceso varían significativamente. En primera instancia se separan las columnas que usaremos como control y se preprocesa el resto de nuestro dataset original. Admás se descompone en otros dos datasets de 30 y 6 componentes principales del original. Cada uno de estos nuevos datasets es sometido a los mismos tres procesos de clustering por medio del algoritmo de Kmeans. Al final de este proceso se obtendran tres sets de 2, 4 y 6 clusters por cada nuevo dataset que habíamos generado (Fig. 1).



En paralelo hemos convertido nuestras columnas de samples que requirieron “cuidados intensivos”, “asistencia respiratoria mecánica” y que han “fallecido” en clusters independientes. Estos clusters Y de control los compararemos con cada uno de los clusters generados anteriormente a través de las siguientes métricas de clasificación.

- **Accuracy**: Define la similitud de cada uno de los clusters obtenidos con el cluster de control correspondiente. Sirve únicamente para comparar entre otros clusters del mismo experimento ya que es minúscula la proporsión de positivos en cada uno de los clusters de control.
- **Sensibility**: Define la proporción de casos positivos en el cluster de control que se encuentran en el cada uno de los clusters obtenidos.
- **Overlap**: Define la proporción de casos positivos en la población total de cada uno de los clusters obtenidos. Es una métrica espontanea para este estudio, la terminología no fué tomada de ninguna bibliografía.

El Rand Index de cada cluster obtenido, comparándolo con cada uno de los clusters de control, es equivalente a la métrica Accuracy en este estudio.

Resultados

En la tabla siguiente se muestra la cantidad de samples agrupados en cada uno de los clusters generados a partir de cada preprocesamiento de los datos y la proporción de la población total que estos representan.

Classification of samples by cluster for each case of study

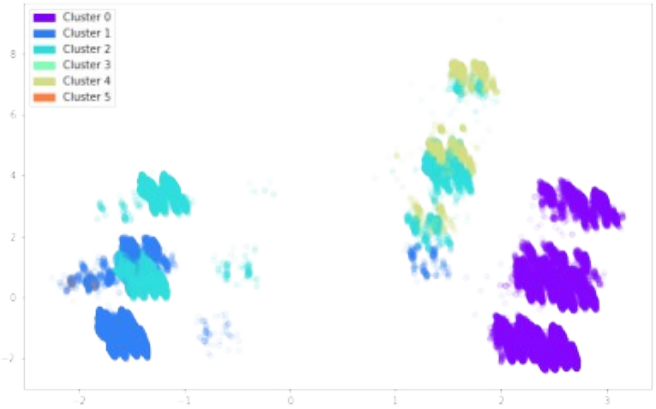
	Cluster:	0	1	2	3	4	5	Total	Silhouette Score[1]
PREPROS.	CLUSTERS	samples	samples	samples	samples	samples	samples		
raw	2	0.3822 137914	0.6178 222963					1.00 360877	0.1737
	4	0.3520 127017	0.3313 119573	0.3167 114285	0.0000 2			1.00 360877	0.1054
	6	0.3520 127017	0.3313 119550	0.3048 109990	0.0000 2	0.0119 4277	0.0001 41	1.00 360877	0.1056
30	2	0.3822 137914	0.6178 222963					1.00 360877	0.2142
	4	0.3218 116118	0.5694 205490	0.0302 10895	0.0786 28374			1.00 360877	0.2347
	6	0.2729 98487	0.0000 2	0.3520 127015	0.1630 58819	0.0302 10895	0.1819 65659	1.00 360877	0.1411
6	2	0.3822 137914	0.6178 222963					1.00 360877	0.4
	4	0.0302 10895	0.6178 222963	0.3520 127017	0.0000 2			1.00 360877	0.4319
	6	0.3217 116107	0.3177 114657	0.0302 10895	0.0786 28376	0.2517 90840	0.0000 2	1.00 360877	0.4748

[1] **Silhouette Score** is calculated by a randomly selected 50000 sample of the same size and random seed for every case.

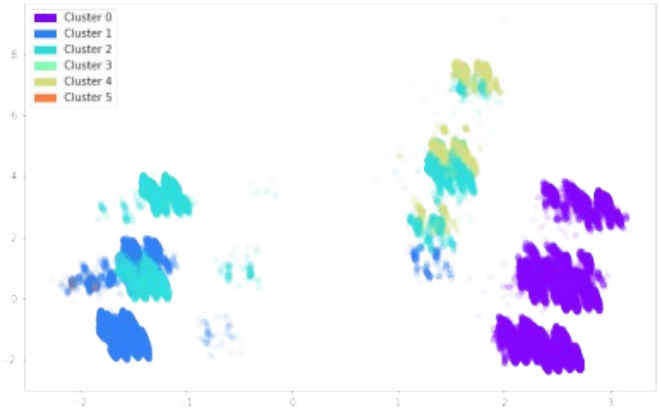
Conclusiones

A partir de la tabla de resultados, gráficos de distribución de muestras en dos dimensiones (Fig. 2) y las métricas ya mencionadas que se detallan en los archivos anexos, se proponen las siguientes o conclusiones del estudio:

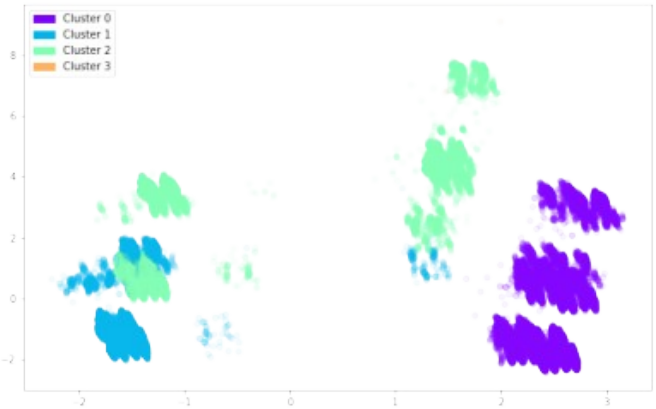
- Cuando se generan 4 o más clusters, es común que algunos queden casi vacíos. Esto demuestra la presencia de outlayers extremos y que aumentar el número de clusters provoca que estos se agrupen por su separados.
- En las tablas de métricas es más facil observar que estas clusters de outlayers poco agregan al análisis ya que no contienen samples de control positivos en ellos.
- El Silhouette Score aumenta considerablemente reducimos la dimensionalidad de nuestros datos, encontrando su máximo en el estudio cuando descomponemos en solo seis componentes principales.
- El Silhouette Score no parece corresponder especialmente con nignún número de clusters en partícular cuando variamos los features de entrenamiento del algoritmo Kmeans.
- Cuando generamos 4 o más clusters, aproximádamente el 80~90% de los casos positivos se agrupan en 2 clusters.
- Cuando generamos 2 clusters, los casos positivos se reparten de forma indiferente. Salvo en el caso de los samples fallecidos que se agrupa el 80% en uno de los clusters.



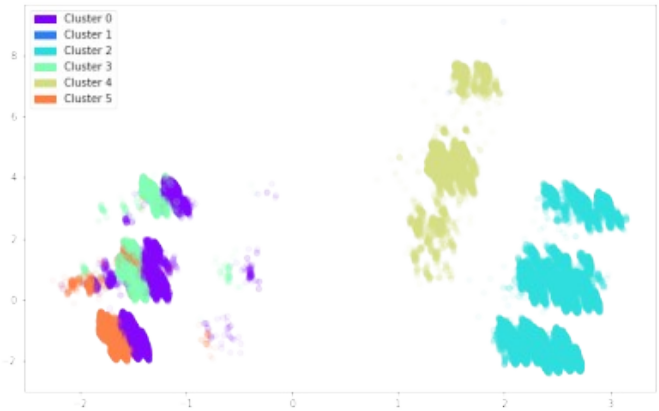
1.1. Preprocesing: Raw, Clustering 6



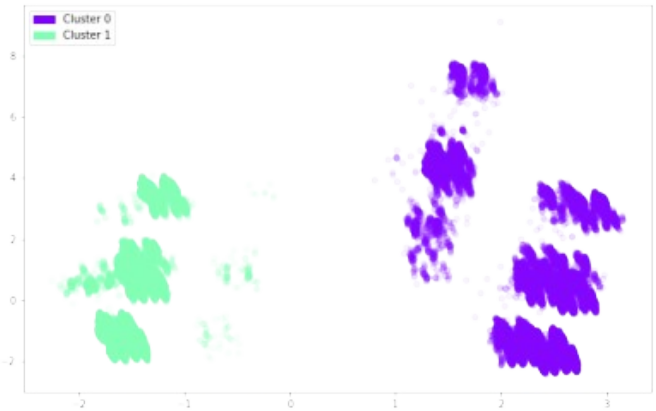
2.1. Preprocesing: Raw, Clustering 6



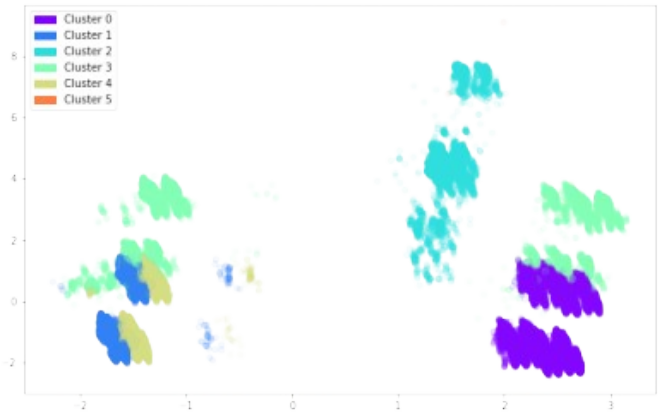
1.2. Preprocesing: Raw, Clustering 4



2.2. Preprocesing: 30PCA, Clustering 6



1.3. Preprocesing: Raw, Clustering 2



2.3. Preprocesing: 6PCA, Clustering 6

Estas observaciones no nos permiten enunciar normas absolutas para encarar el manejo de estos datos. La selección de las metodologías de preprocesamiento y agrupamiento de este dataset se verá fuertemente influida por el fin del estudio que se desea realizar con ellos. Se propone trabajar con el set de pacientes fallecidos como función objetivo al ver ser altamente consistente el agrupamiento de estos casos en un solo cluster y no encontrar mayor influencia del proceso de decomposición de los datos en sus componentes principales.