

CMP 0575

Esteban Flores

Intro

- **Conjunto de datos:** Química (126 atributos y 2 clases de salida)
- **Lenguaje utilizado:** Python
- **Librería utilizada:** scikit-learn.org
- **Proceso de Normalización:**
 - Scaling
 - Standarization
 - Normalization
- **Proceso de Reducción**
 - Modelo de búsqueda **Recursive Feature Elimination** (with **Cross Validation**)
 - Clasificador: LinearSVC

Proceso de Normalización

Algoritmos de *machine learning* pueden beneficiarse de *scaling* de atributos (ej. kNN)

Estandarización (`sklearn.preprocessing.StandardScaler`):

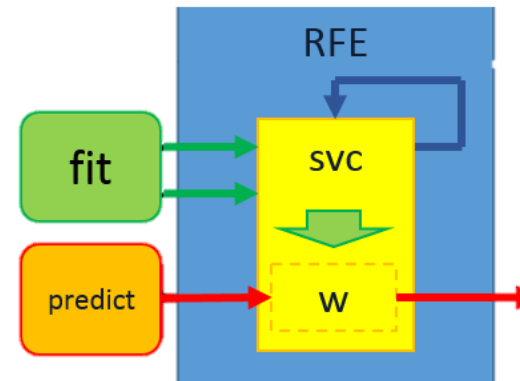
Algunos algoritmos de *scikit-learn.org* necesitan datos distribuidos normalmente: gaussianos con media cero y varianza 1.

$$z = \frac{x - \mu}{\sigma}$$

Proceso de Reducción

RFE CV:

Modelo de búsqueda **Recursive Feature Elimination** (with **Cross Validation**)



El objetivo de la eliminación recursiva de características (RFE) es seleccionar características considerando recursivamente conjuntos de características cada vez más pequeños.

Proceso de Reducción

Procedimiento en *scikit-learn.org*

La importancia de cada característica se obtiene a través de un atributo *coef_* o bien a través de un atributo *feature_importances_*. Luego, las características menos importantes se eliminan del conjunto actual de características.

RFECV realiza RFE en un bucle de validación cruzada para encontrar el número óptimo de características.

Proceso de Reducción

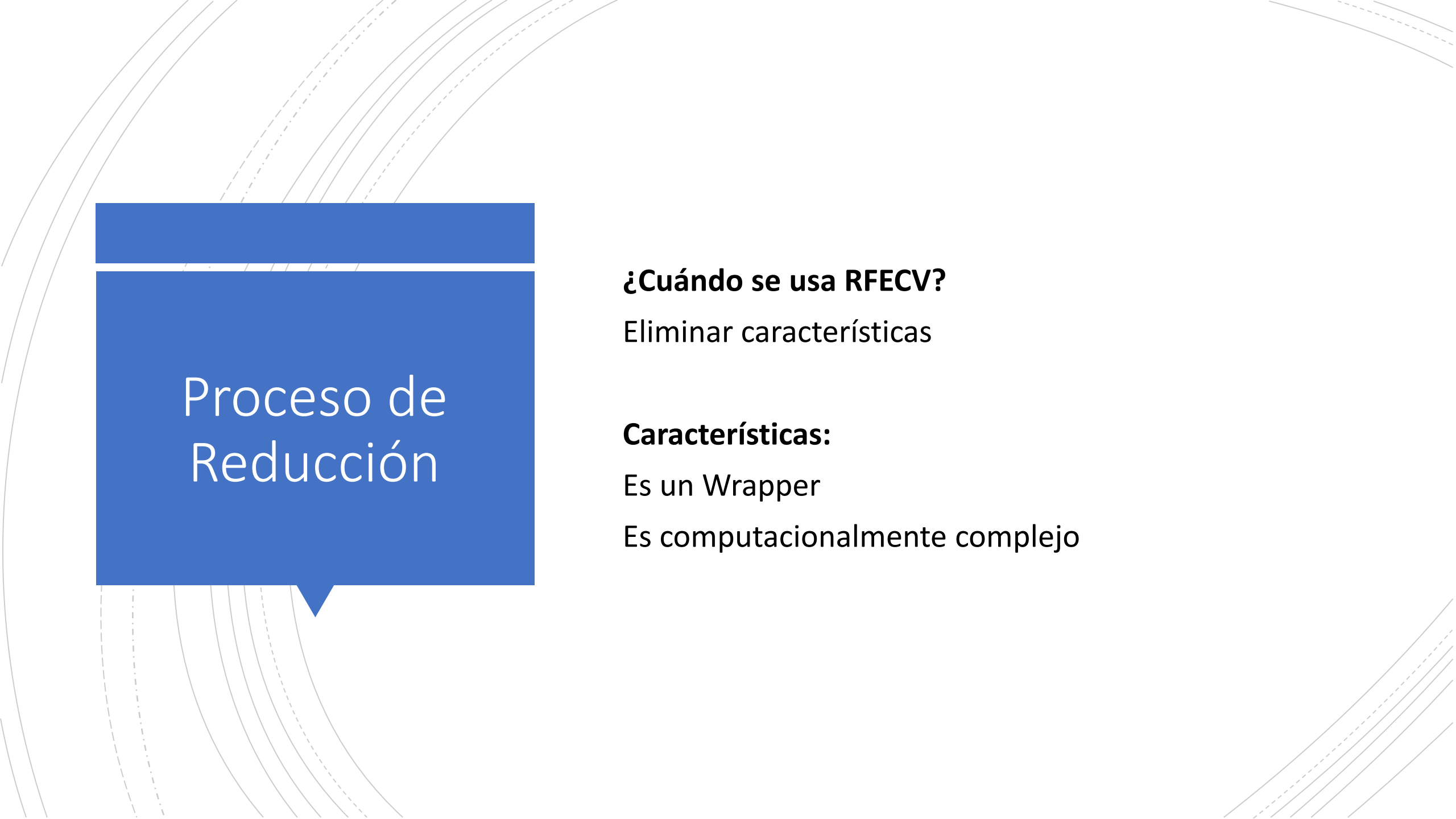
Ejemplo de RFECV

Se ejecuta un RFECV de K-Fold de 3...

La validación cruzada *K-fold* se realiza según los siguientes pasos:

1. Partición del conjunto de datos de entrenamiento original en k subconjuntos iguales. Cada subconjunto se denomina *fold* (f_1, f_2, \dots, f_k).
2. Para $i = 1$ a $i = k$
 - a) Se mantiene f_i como conjunto de validación y se mantiene todos los pliegues $k-1$ restantes en el conjunto de entrenamiento de la validación cruzada.
 - b) Se entrena el modelo de aprendizaje utilizando el conjunto de entrenamiento de validación cruzada y se calcula la precisión del modelo validando los resultados pronosticados con el conjunto de validación.
3. Se estima la precisión del modelo de aprendizaje promediando las precisiones derivadas en todos los k casos de validación cruzada.

En el método de validación cruzada *K-fold*, todas las entradas en el conjunto de datos de entrenamiento original se utilizan tanto para el entrenamiento como para la validación. Una entrada se usa para validación solo una vez.

The background of the slide features a series of thin, curved lines in a light gray color, creating a sense of motion and depth. These lines are more prominent on the left side and fade out towards the right.

Proceso de Reducción

¿Cuándo se usa RFECV?

Eliminar características

Características:

Es un Wrapper

Es computacionalmente complejo

Estimador

Estimador: LinearSVC

Support Vector Machines (SVM) conjunto de métodos de aprendizaje supervisado

Una SVM construye un hiperplano o conjunto de hiperplanos que es usado para la clasificación

Maximizar distancia a *supporting vectors*

Ventajas

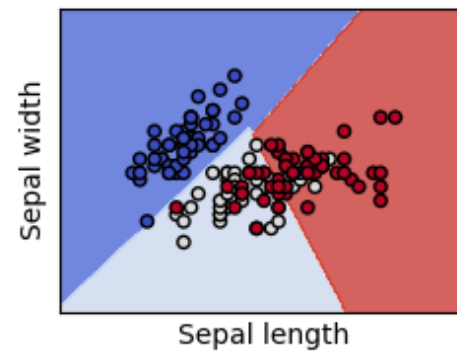
- Eficaz en espacios de alta dimensión.
- Sigue siendo efectivo en casos donde: $n_{\text{features}} > n_{\text{samples}}$
- Es eficiente en memoria.

Desventajas

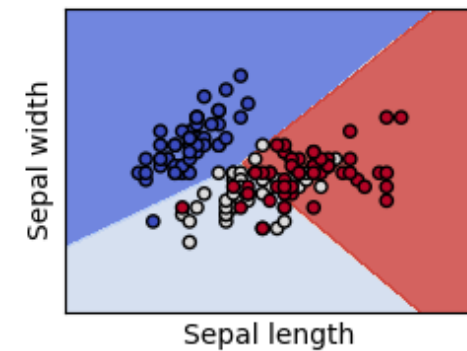
- Los SVM no proporcionan directamente estimaciones de probabilidad, se calculan utilizando una costosa validación cruzada

Estimador (LinearSVC)

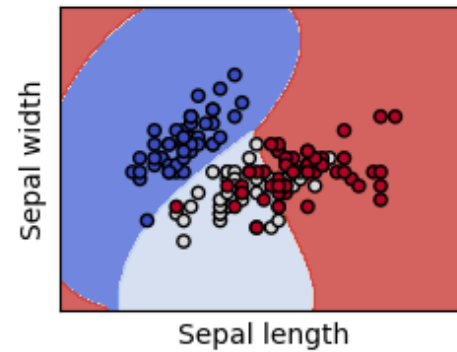
SVC with linear kernel



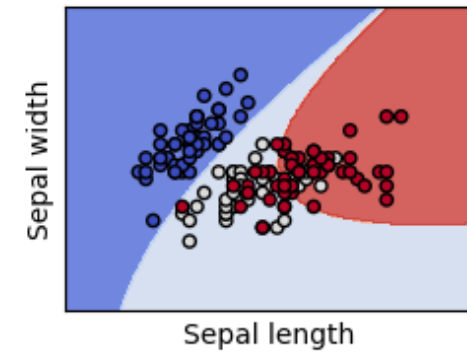
LinearSVC (linear kernel)



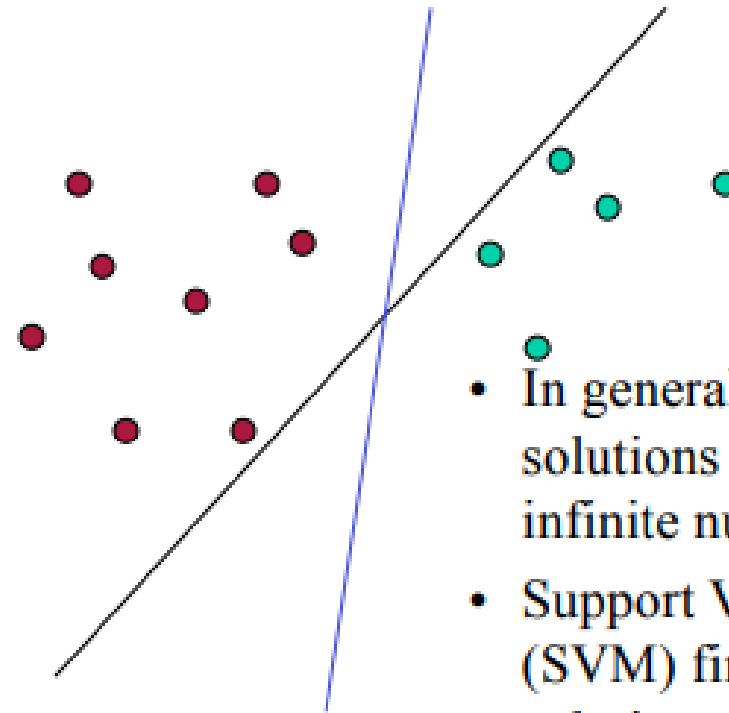
SVC with RBF kernel



SVC with polynomial (degree 3) kernel

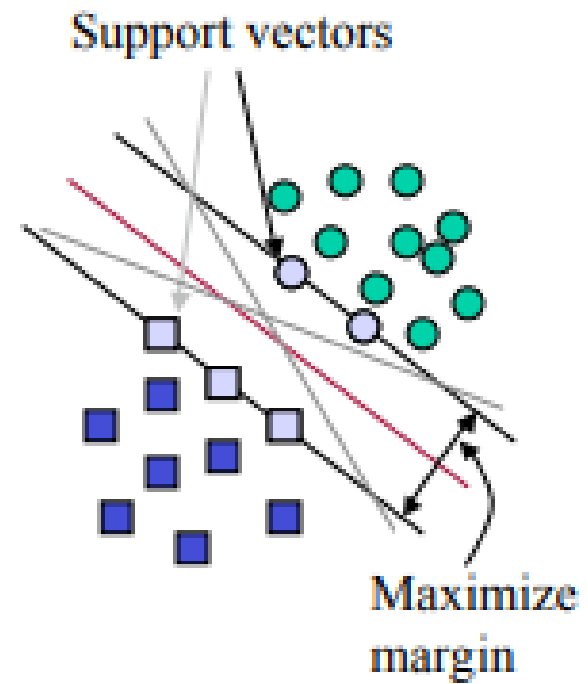


Estimador (LinearSVC)



- In general, lots of possible solutions for a, b, c (an infinite number!)
- Support Vector Machine (SVM) finds an optimal solution

Estimador
(LinearSVC)



Estimador (LinearSVC)

Score:

$$\sum_{i=1}^m \alpha_i y^{(i)} K(x^{(i)}, x) + b$$

Función K (en mi caso, lineal “LinearSVC”)

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j.$$

s

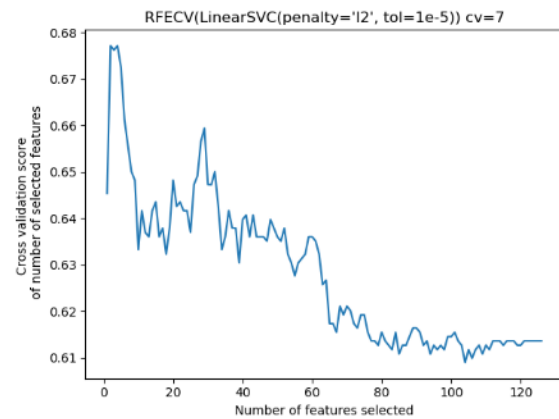
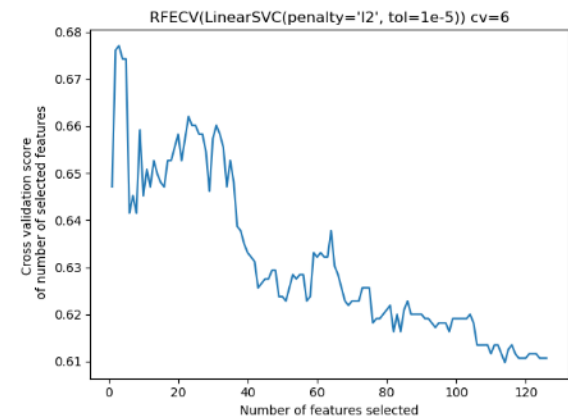
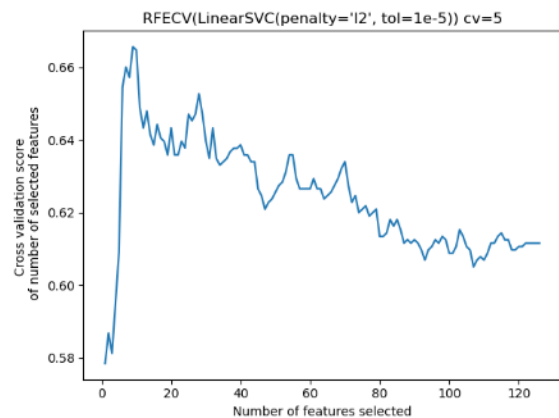
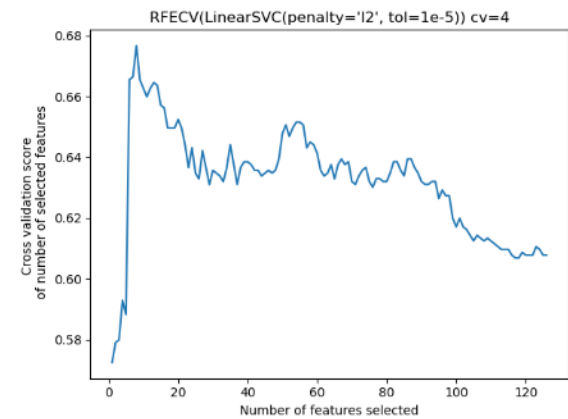
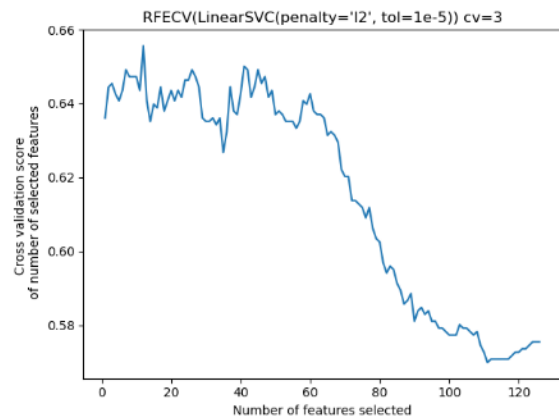
Trazabilidad (Normalización)

```
===== Scaling data(NORMALIZATION) =====  
Scaled data  
==> X:  
[[-0.35154384 -0.36546614 -1.52713662 ... -0.12904725 -0.12904725  
  0.17578687]  
 [-0.07232717 -0.26232368  0.00255608 ... -0.83180025 -0.83180025  
  0.07438824]  
 [-0.40478113  0.47247711 -1.51708682 ... -0.67132478 -0.67132478  
  0.20864378]  
 ...  
 [-1.07096793  0.45738433  0.38015985 ...  1.13353356  1.13353356  
  0.66652176]  
 [ 1.64429408  2.22542763 -1.25469749 ...  1.51109712  1.51109712  
  2.06812267]  
 [-0.05856784 -0.48908152 -0.38866151 ... -0.71861706 -0.71861706  
 -0.23204334]]
```

Trazabilidad (Subsets)

cv	# features	Selected Features	cv score
3	12	[28, 42, 43, 44, 47, 48, 49, 64, 73, 75, 119, 121]	65.56%
4	8	[42, 43, 44, 47, 48, 64, 75, 119]	67.66%
5	9	[28, 42, 43, 44, 47, 48, 64, 75, 119]	66.56%
6	3	[42, 47, 48]	67.70%
7	2	[42, 48]	67.71%
8	23	[28, 38, 42, 43, 44, 45, 46, 47, 48, 49, 55, 64, 65, 66, 73, 75, 103, 105, 106, 108, 115, 119, 121]	68.12%
9	35	[5, 6, 15, 23, 28, 38, 41, 42, 43, 44, 45, 46, 47, 48, 49, 55, 58, 59, 64, 65, 66, 68, 73, 75, 91, 99, 103, 105, 106, 108, 113, 115, 118, 119, 121]	68%
10	2	[42, 48]	66.04%
11	11	[28, 42, 43, 44, 47, 48, 49, 64, 75, 119, 121]	66.79%
12	25	[6, 28, 38, 42, 43, 44, 45, 46, 47, 48, 49, 55, 58, 64, 65, 66, 73, 75, 103, 105, 106, 108, 115, 119, 121]	67.40%

Trazabilidad (Subsets)



Trazabilidad (Subsets)

