Text

Description automatically generated

Queremos saber cuales son las features mas importantes

Varias aproximaciones

1. Varianza: a mas varianza mas información

Graphical user interface, application

Description automatically generated

El umbral define el mínimo nivel de varianza con el que queremos trabajar.

Graphical user interface, application

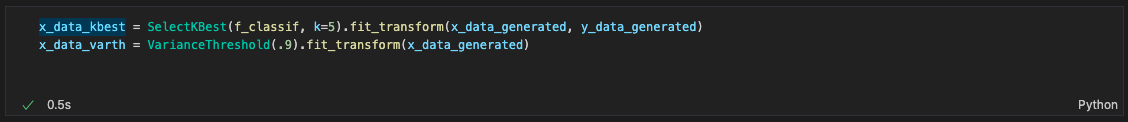
Description automatically generated

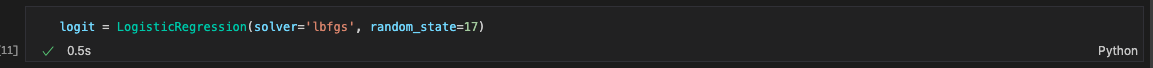
Subiendo un poco la varianza eliminamos columnas en los datos.

Graphical user interface, text, application

Description automatically generated

En supervisado necesitamos la columna target para decidir las columnas que tienen mas relación con ella.





A screenshot of a computer

Description automatically generated with medium confidence

Text

Description automatically generated

Comparamos

una regresión logística (logit)

un random forest (rf)

una combinación de la regresión logística apoyado en el random forest

Un árbol de decisión decide que features son mas importantes, y esta decisión previa es usada por la regresión logística.

El mejor scoring cuando son negativos es el mas cercano a 0 (el mas alto)

Text

Description automatically generated

Text

Description automatically generated

Vamos a decidir si sólo pudiéramos usar una columna, cual nos daría mejor resultado.

Una vez que la ha decidido, prueba combinaciones de la primera con otra, hasta llegar al número de columnas que se han indicado.

Text

Description automatically generated

Las imágenes se almacenan en una sola fila (aplanado) con un número de columnas pixels \* pixels

Dentro de ese vector, elegimos las columnas mas importantes que tendríamos que mirar para averiguar el número.

Si queremos saber las 10 columnas mas importantes y ver sus valores, sin dibujar la matriz

Cada valor es un nivel de gris.

Graphical user interface, text

Description automatically generated