**APUNTES REGRESION LOGISTICA**

**Video explicativo:**

<https://www.youtube.com/watch?v=iMPcV8Rp-Oo>

LaRegresión Logística tiene una forma de S.

Donde los puntos o DataPoint hacen una proyección para obtener su probabilidad.

Se Utiliza como método de CLASIFICACIÓN. Teno dos clase (BINOMIAL) o mas de dos clases (MULTINOMIAL)

Para el caso Binomial la expresión matemática del modelo es:

Es decir, la operación "logit" de una variable "p" es el logaritmo de la relación de p y no p.

Los beta son los parámetros que ajusta el modelo y las x las variables independientes (Por ejemplo para la clasificación Binomial de tener cardiopatía o no tener cardiopatía x1 puede ser frecuencia cardiaca, x2: electrocardiograma debil etc..)

Como ejemplo si tenemos una variable independiente:

Proyectamos los valores de x sobre la curva logística y obtenemos p, observamos que para x>0 aumenta la probabilidad de tener cardiopatía y viceversa.

Es decir que al entrenar un modelo de regresión logística determinando los parámetro Beta obtenemos la curva de regresión logística.

Una vez que tenemos la curva realizamos las predicciones con nuevos datos.

**Algoritmo de Maxima Verisimilitud MLE (Most Likely Estimator)**

La estimación de máxima verosimilitud (MLE) es una poderosa técnica estadística que se ha utilizado ampliamente en varios campos, incluidos la economía, las finanzas, la biología y las ciencias sociales. Es un método para estimar los parámetros de un modelo estadístico, dado un conjunto de datos observados. La idea básica detrás de MLE es encontrar los valores de los parámetros que maximicen la probabilidad de observar los datos dados bajo el modelo asumido.

En nuestro caso evaluamos como la regresión logística evalúa el modelo para obtener los parámetros Beta.

Video explicativo de expresión matemática de MLE para Regresión Logística (también llamado modelo Logit):

<https://www.youtube.com/watch?v=ceDzBsXK0Iw&t=6s>

Esta expresión a la que llega el video es dificil resolverla analitivamente por ello utiliza metodos de optimizacion iterativos, en este caso el algoritmo MLE calcula un número REI que mientras mayor sea este número la curva que representa los parámetros Beta con la cuaL se hizo el cálculo de optimización son más adecuadas

En este caso REI=1.039 e-8

**Descenso del Gradiente**

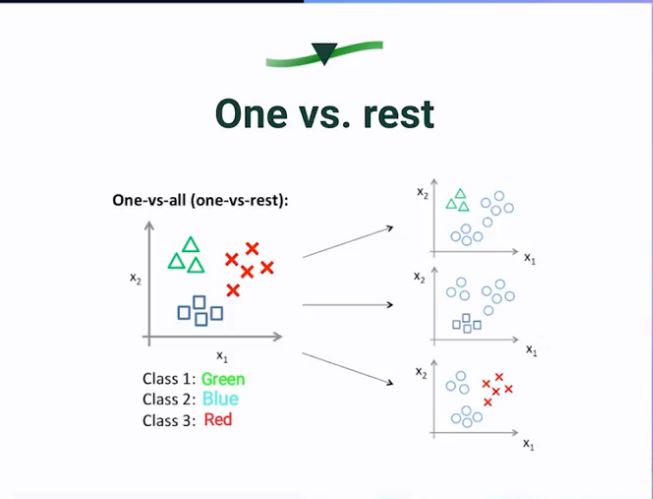
**El** descenso del gradiente es otro algoritmo de optimización usado fundamentalmente en redes neuronales.

**Regresion Logistica Multiclase**

**One vs reset**

Estrategia multiclase uno contra el resto (OvR).

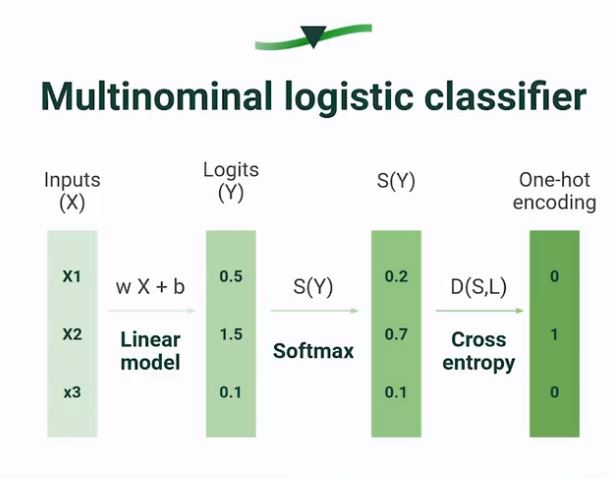
También conocida como uno contra todos, esta estrategia consiste en ajustar un clasificador por clase. Para cada clasificador, la clase se compara con todas las demás clases. Además de su eficiencia computacional (sólo n\_classes se necesitan clasificadores), una ventaja de este enfoque es su interpretabilidad. Dado que cada clase está representada por un solo clasificador, es posible obtener conocimientos sobre la clase inspeccionando su clasificador correspondiente. Esta es la estrategia más utilizada para la clasificación multiclase y es una opción predeterminada justa.



Aplicación del método en código:

<https://www.geeksforgeeks.org/one-vs-rest-strategy-for-multi-class-classification/>

**Multinomial Logic Classfier**

****

El clasificador lógico multinomial (Multinomial Logic Classifier en inglés) es un algoritmo utilizado en el campo del aprendizaje automático (machine learning) para clasificar datos en múltiples categorías o clases. Es una variante de los clasificadores lógicos que se basa en la lógica difusa y la teoría de conjuntos difusos.

En lugar de generar una única regla de clasificación para cada clase, como en los clasificadores binarios, el clasificador lógico multinomial utiliza múltiples reglas para asignar probabilidades de pertenencia a cada clase. Estas probabilidades se calculan mediante la combinación de reglas lógicas y la ponderación de las características de los datos.

El clasificador lógico multinomial puede ser especialmente útil cuando se trabaja con conjuntos de datos en los que las clases no son mutuamente excluyentes, es decir, cuando un dato puede pertenecer a más de una clase al mismo tiempo. También puede manejar datos con características continuas o discretas.

Este clasificador se utiliza en diversas aplicaciones, como clasificación de texto, análisis de sentimientos, reconocimiento de voz y diagnóstico médico, entre otros. Su capacidad para manejar múltiples clases y la incertidumbre en la clasificación lo convierte en una herramienta poderosa en el campo del aprendizaje automático.

El bloque softmax en el clasificador lógico multinomial ayuda a convertir las salidas del modelo en una distribución de probabilidad para determinar la probabilidad de pertenencia a cada clase. Esto facilita la toma de decisiones y la asignación de una clase específica a un dato en función de las probabilidades calculadas.

**Solvers**

En el método de regresión logística, los solvers son algoritmos utilizados para estimar los parámetros del modelo logístico que mejor se ajusten a los datos. Estos algoritmos buscan encontrar los valores óptimos de los coeficientes de regresión que minimicen la función de coste o maximicen la función de verosimilitud.

Existen diversos solvers disponibles para el problema de regresión logística, cada uno con sus propias características y ventajas. Algunos de los solvers más comunes son:

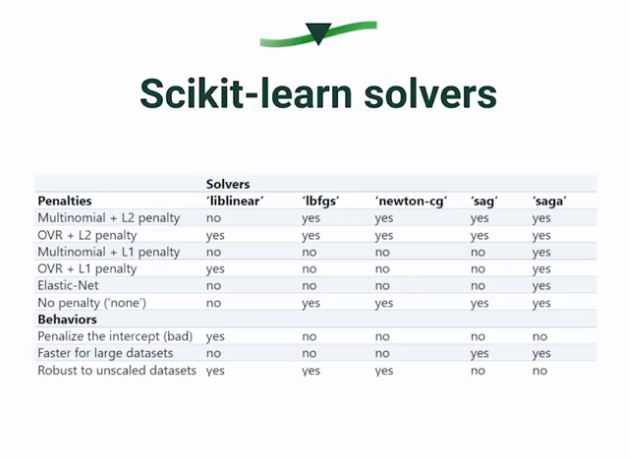
1. Solver "liblinear": Este es el solver más utilizado y es adecuado para conjuntos de datos pequeños o medianos. Utiliza un algoritmo de optimización basado en coordenadas para resolver el problema de regresión logística.
2. Solver "newton-cg": Este solver utiliza el método de Newton-Conjugate Gradient para encontrar el mínimo de la función de coste. Es adecuado para conju
3. ntos de datos grandes.
4. Solver "lbfgs": Este solver utiliza el algoritmo de Broyden-Fletcher-Goldfarb-Shanno (BFGS) para optimizar la función de coste. Es eficiente para conjuntos de datos grandes y se beneficia de la memoria limitada.
5. Solver "sag": Este solver utiliza el método de promedio de gradiente estocástico (Stochastic Average Gradient) para encontrar el mínimo de la función de coste. Es especialmente útil para grandes conjuntos de datos.

La elección del solver depende del tamaño del conjunto de datos, la complejidad del modelo y los recursos computacionales disponibles. Es importante probar diferentes solvers para encontrar el que mejor se adapte a cada problema específico.

<https://medium.com/@arnavr/scikit-learn-solvers-explained-780a17bc322d>

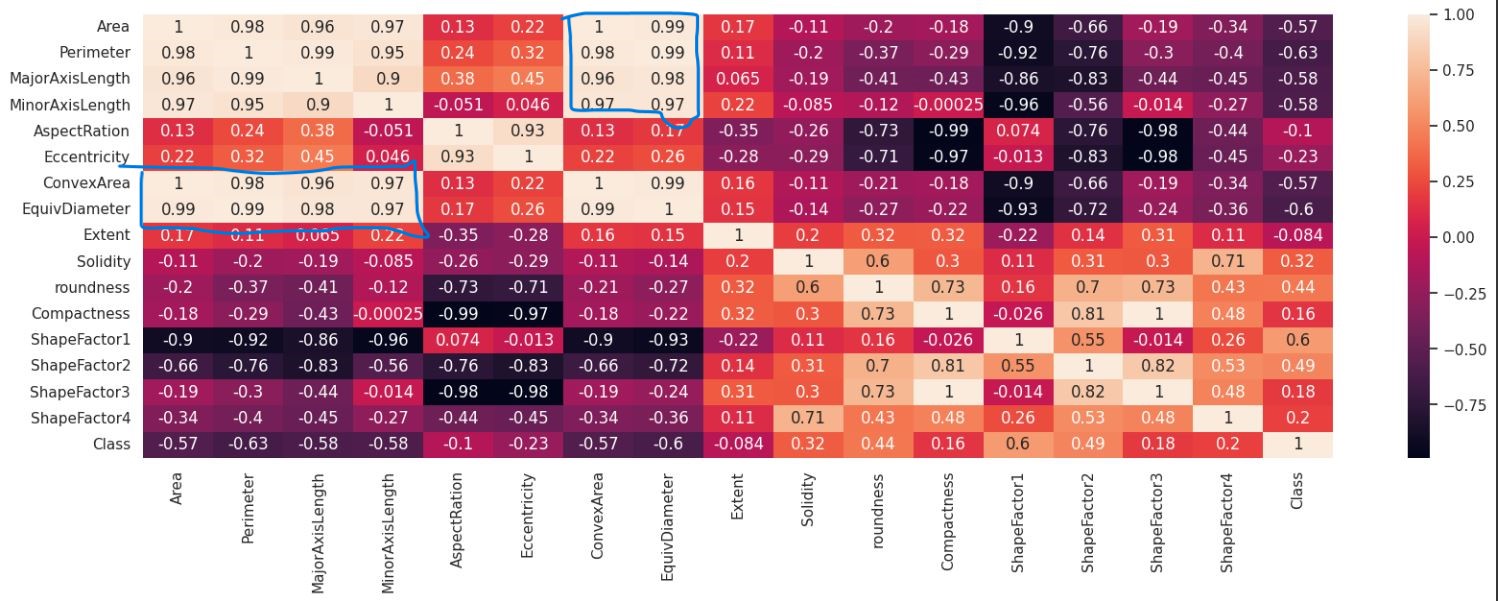
Como aplicarlos en codigo:

<https://www.geeksforgeeks.org/comparing-various-online-solvers-in-scikit-learn/>



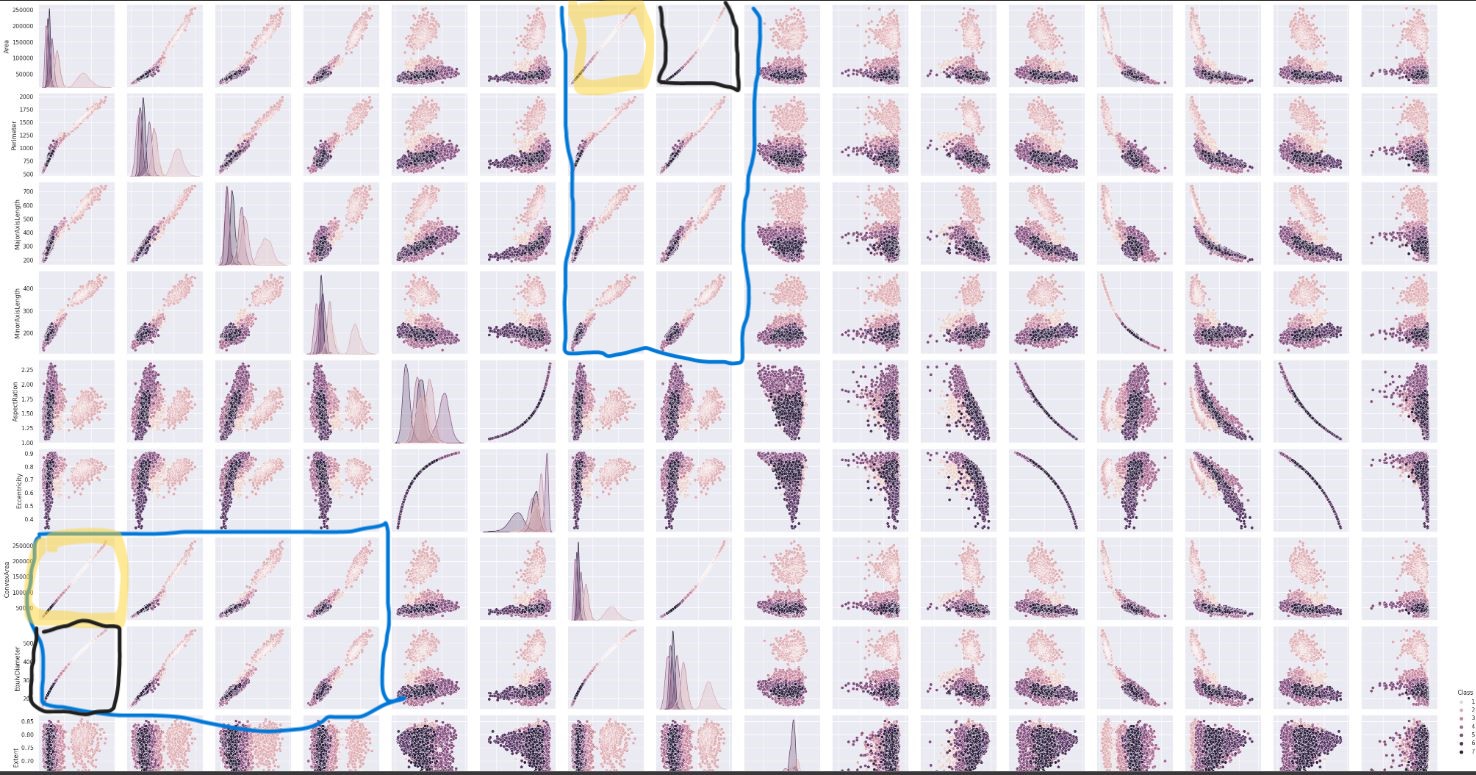
**Pasos**

* Verificar si existen datos nulos
* Verificar si existen outliers
* Verificar que las variables categóricas pasen a numéricas
* Estandarización de datos
* Variables altamente correlacionadas pueden afectar el desempeño de nuestros modelos



En la imagen vemos en los recuadros ConvexArea y EquivDiameter tienen ambas una correlación Positiva Fuerte con Area, Perimeter, MajorAxisLenght, MinorAxisLenght

Viendo una grafica pairplot:



Este ultimo grafico se ve mas claro en el colab

<https://colab.research.google.com/drive/1abyyuaO0RwAQFSqAGhDmvrgEET-kWWEr#scrollTo=o7TaYCyOvhBb>

Lo que vemos es la misma correlación de variables pero con gráficos de distribución. Acá cuando hay una correlación fuerte entre las variables el gráfico de dispersión se asemeja a un recta.

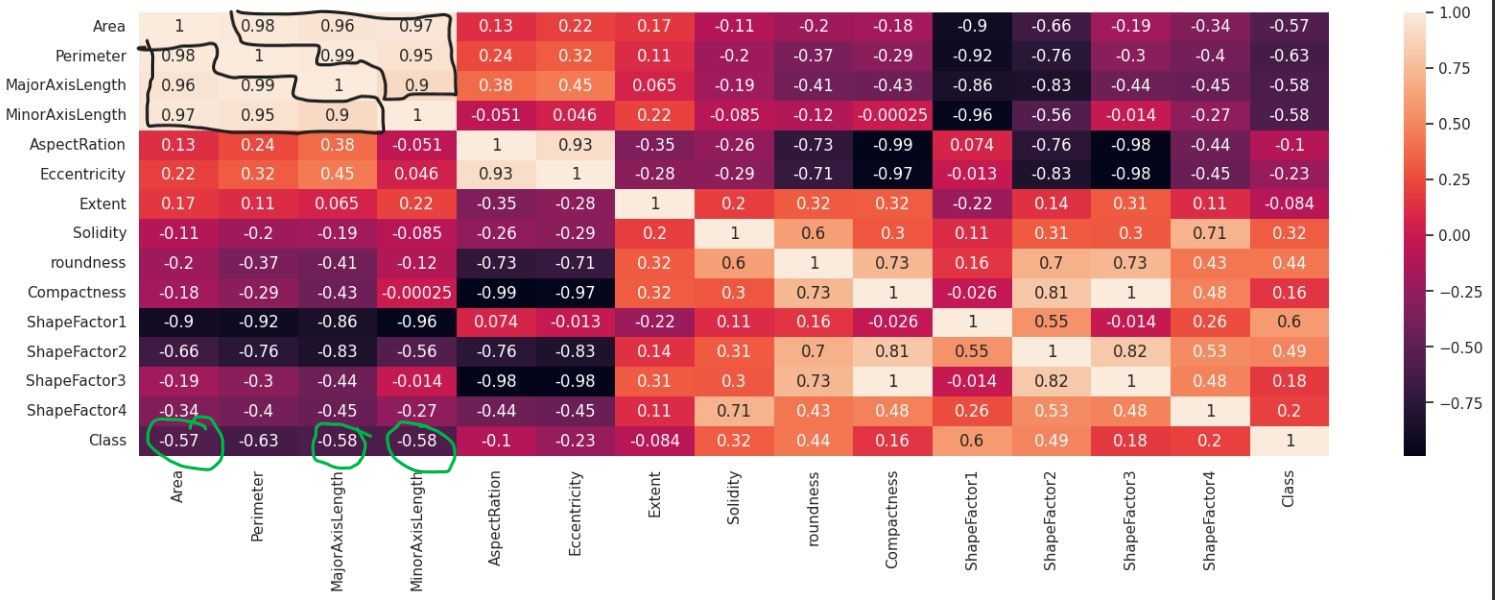
Observamos la misma correlación de variables del gráfico heatmap.

**En Conocimiento de la temática que representan esta base de datos**

**Decidimos qué variables eliminar ya que no es conveniente tener variables fuertemente correlacionadas al momento de entrenar al modelo.**

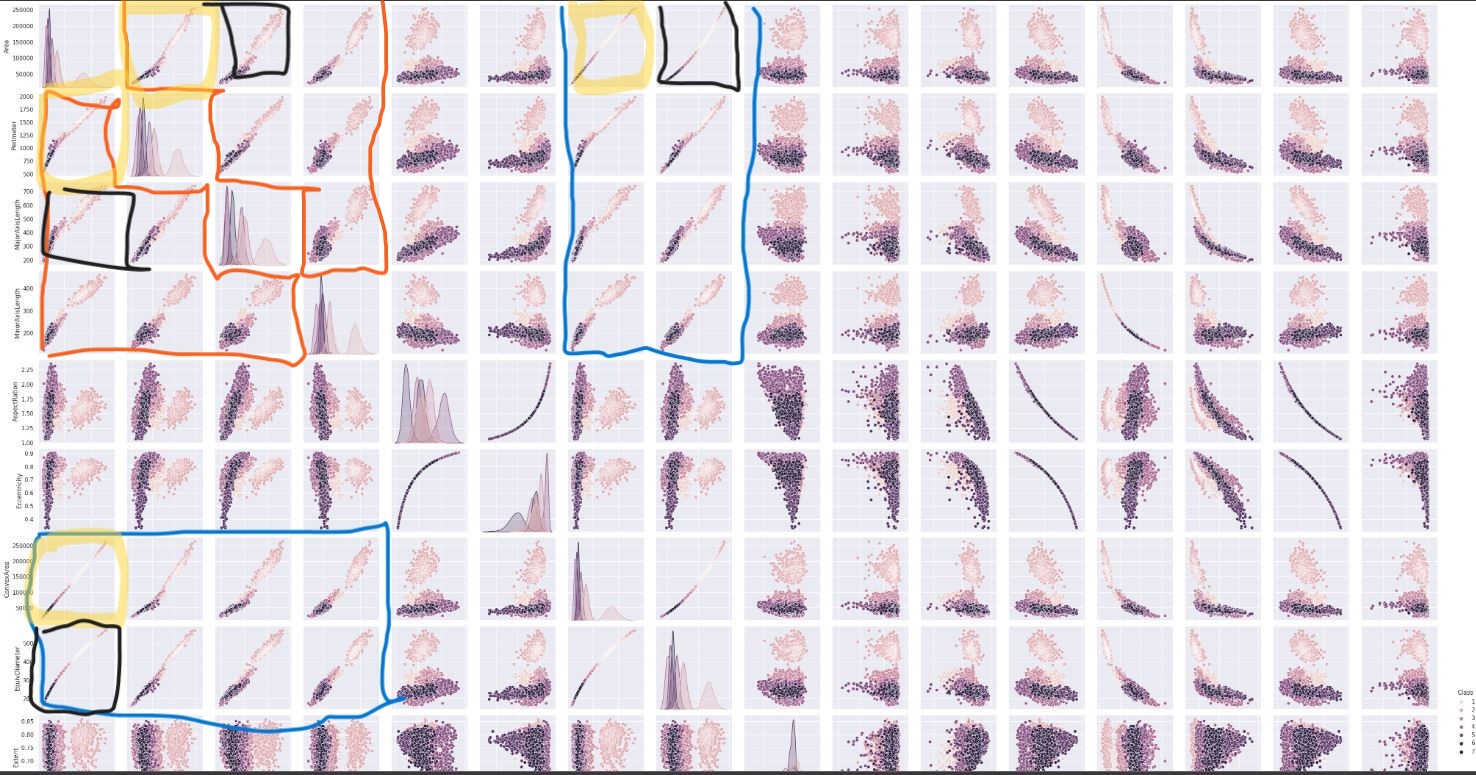
En este caso se eliminan ConvexArea y EquivDiameter

El heatmap quedaria



Que sigue mostrando una correlacion de entre Perimeter, Area, majorAxisLenght, MenorAxisLenght

Estas correlaciones las enmarcamos en naranja en pairplot:



Se ve el mismo patrón de linealidad en graficos de dispersion