

Machine Learning

Modelos de regresión 2

M.Sc. Angelo Jonathan Diaz Soto

2025



Data&Analytics
INNOVACIÓN Y TECNOLOGÍA

Contenido



■ KNN



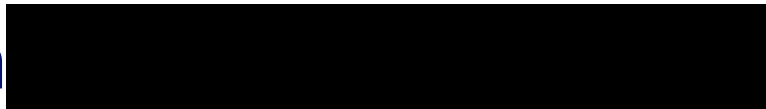
Vector de regresión de soporte ■

Árboles de decisión

■ Métodos de ensamble

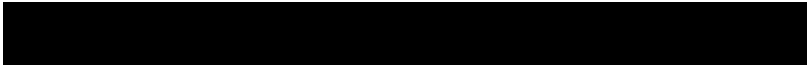
(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Métricas de Evaluación



$j=1$

\hat{y}





$$\sum_{j=1}^n$$

$$\hat{y}$$

3. Raíz del Error cuadrático medio (RMSE)

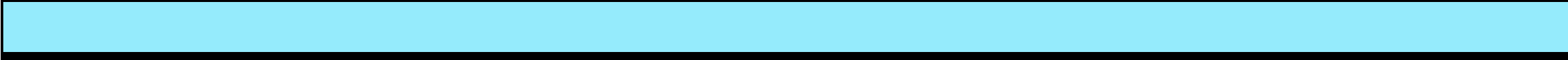
$$n. \sum$$

$$\hat{y}^2$$

$$\sum$$

$$\frac{SCT}{SCE}$$

$$\frac{SCT}{SCR}$$





KNN para regresión



- ❑ K-Nearest-Neighbor es un algoritmo basado en instancia de tipo supervisado de Machine Learning.
- ❑ Puede usarse para clasificar nuevas muestras (valores discretos) o para predecir (**regresión**, valores continuos).
- ❑ Al ser un método sencillo, es ideal para introducirse en el mundo del Aprendizaje Automático. Sirve esencialmente para clasificar valores buscando los puntos de datos “más similares” (por cercanía).



¿Qué es el algoritmo k-Nearest Neighbor?

Es un
en las
que se
basado
un

-

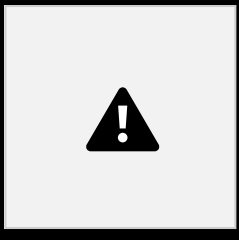
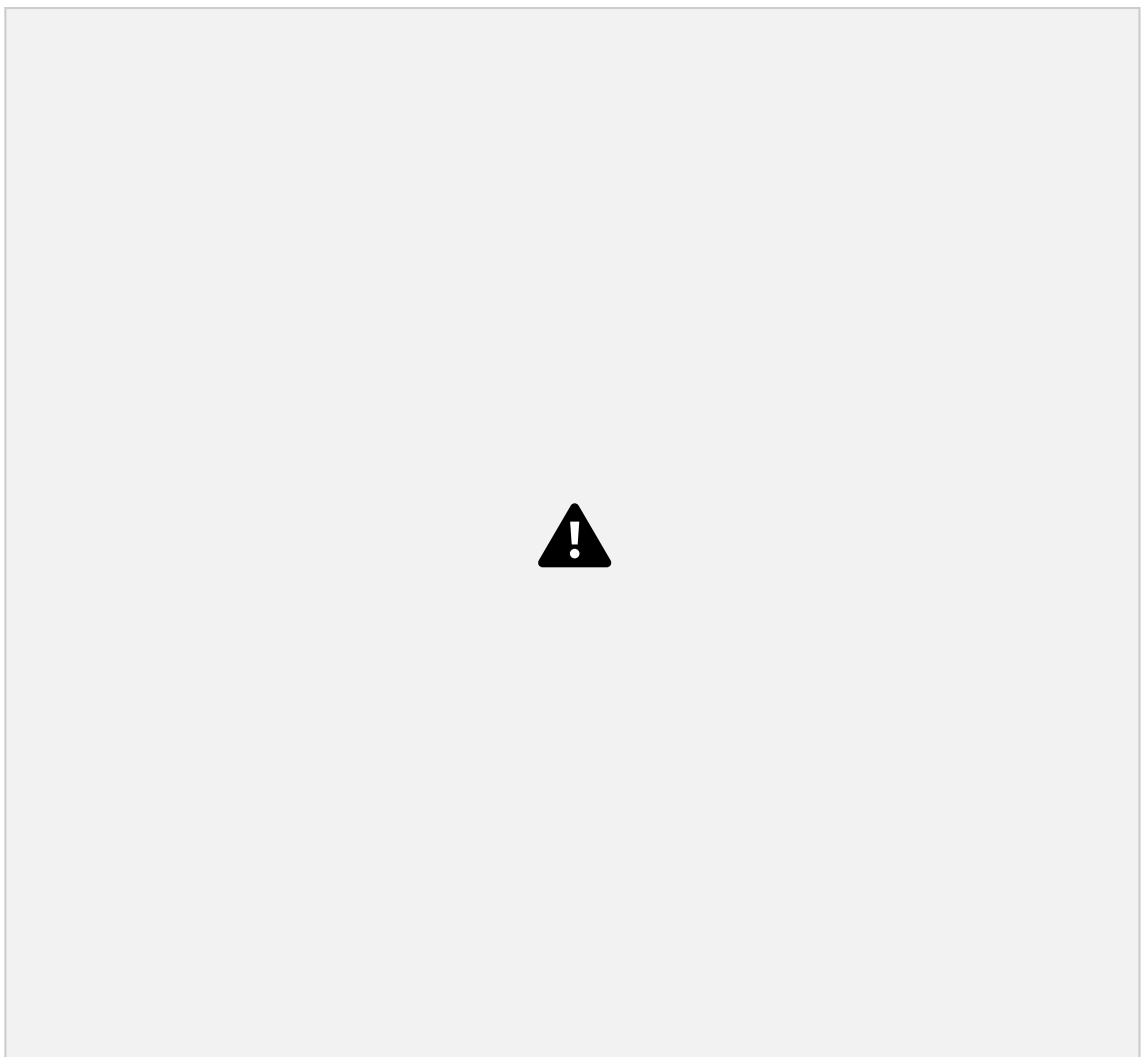
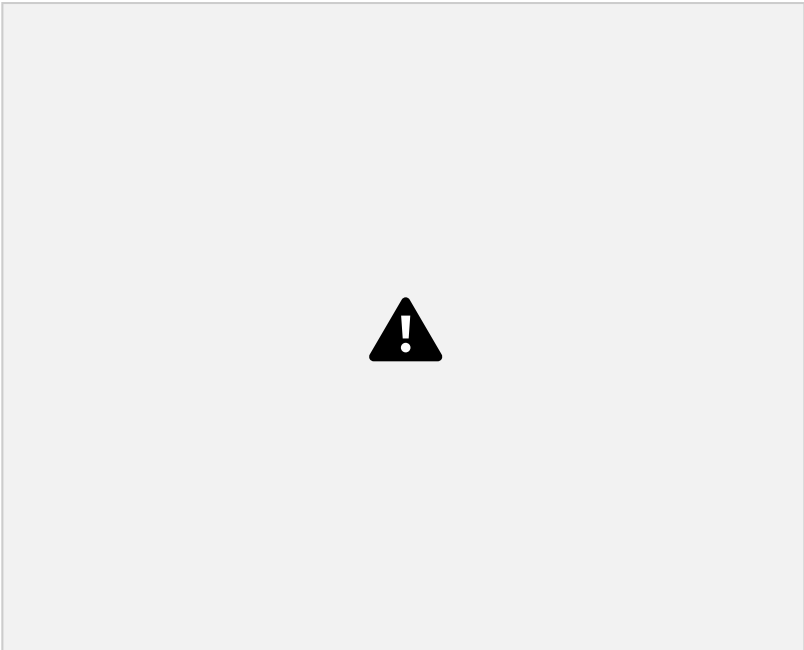


método que simplemente busca observaciones más cercanas a la que está tratando de predecir y clasifica el punto de interés en la mayoría de datos que le rodean. Como dijimos antes, es un algoritmo:

Supervisado: esto quiere decir que tenemos etiquetado nuestro conjunto de datos de entrenamiento,

con la clase o resultado esperado dada “una fila” de datos.

- **Basado en Instancia:** Esto quiere decir que nuestro algoritmo no aprende explícitamente un modelo (como por ejemplo en Regresión Logística o árboles de decisión). En cambio memoriza las instancias de entrenamiento que son usadas como “base de conocimiento” para la fase de predicción.

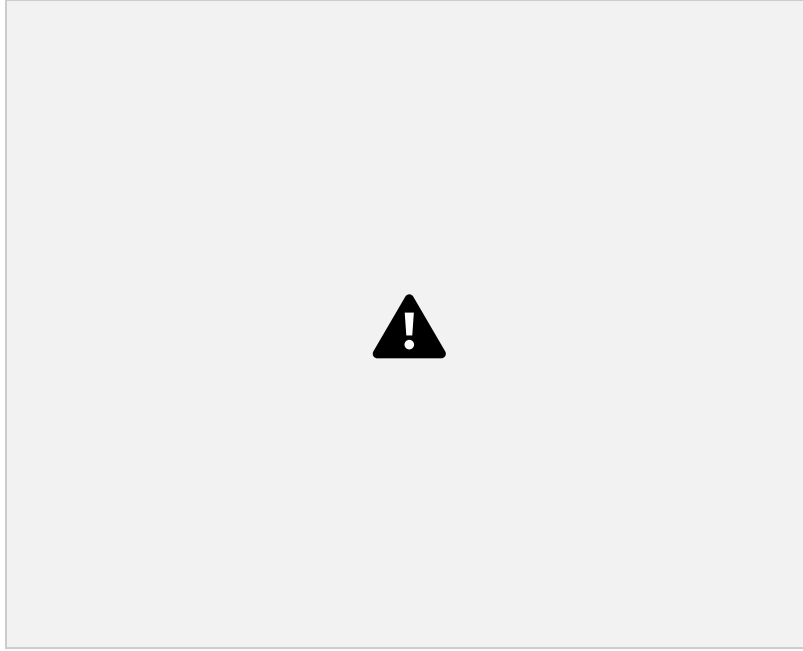


(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Pasos que se siguen para el algoritmo KNN



1. Recibimos el conjunto de datos sin procesar y sin clasificar que hay que trabajar.



2. Elegimos una matriz de distancias de la euclidiana, Manhattan o Minkowski.

3. Luego calcule la distancia entre los nuevos datos puntos y los puntos de datos de entrenamiento clasificados conocidos. 4. El número de vecinos a considerar es definida por el valor de "k". 5. Se sigue comparando con la lista de clases.

que tienen la distancia más corta y cuentan las número de veces que aparece cada clase. 6. La clase con la mayor cantidad de votos gana.

Esto significa que la clase que tiene la frecuencia más alta y ha aparecido el mayor número de veces es asignado al punto de datos desconocido.



¿Por qué utilizar el algoritmo KNN?

particularmente útil cuando:



- El algoritmo KNN no asume ninguna relación entre las características.
- Útil para un conjunto de datos donde la localización de datos es importante.
- Solo tienes que sintonizar el parámetro K, que es el número de vecino más cercano.
- No se necesita entrenamiento, ya que es un algoritmo de

El algoritmo KNN es

aprendizaje "perezoso".

- Sistemas de recomendación y búsqueda de similitudes semánticas entre las principales aplicaciones del algoritmo KNN.

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

¿Utilizar el algoritmo KNN?

Desventajas del algoritmo KNN:

- Tiene que encontrar el valor óptimo de K, lo cual no es fácil.

- ❑ No apto para datos de dimensiones muy elevadas.

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Support Vector Regression

- ❑ Las Máquinas de Vectores Soporte (creadas por **Vladimir Vapnik**) constituyen un método basado en

aprendizaje para la resolución de problemas de **clasificación** y **regresión**. Las SVM fueron presentadas en 1992 y adquirieron fama

- cuando dieron resultados muy superiores a las redes neuronales en el reconocimiento de letra manuscrita, usando como entrada pixeles.

El SVM es un algoritmo para encontrar clasificadores lineales

- en **espacios transformados**.

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Support Vector Regression



■ El

algoritmo de Vectores

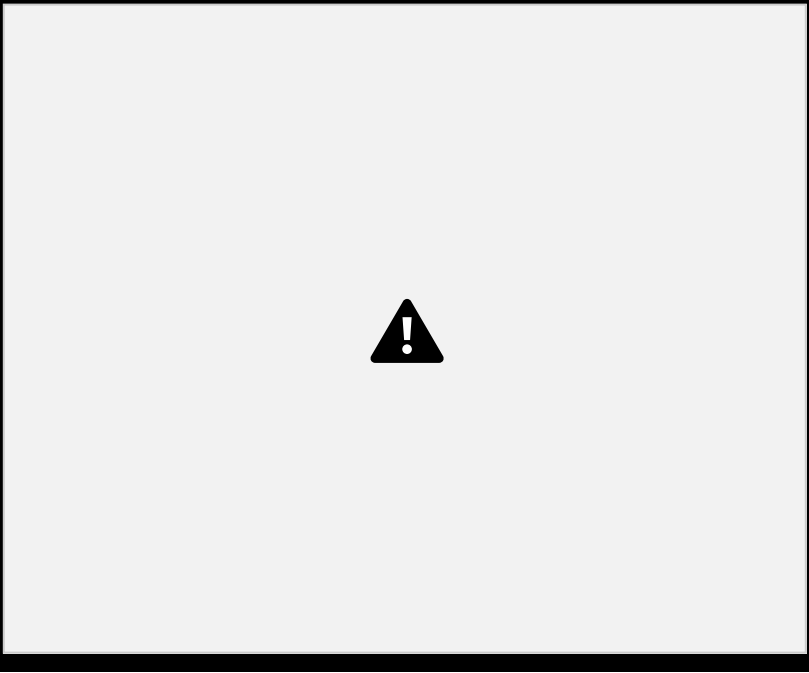
de Soporte Regresión se basa en buscar la **curva** o **hiperplano** que modele la tendencia de los datos de entrenamiento y según ella predecir cualquier dato en el futuro.

- El algoritmo de Vectores de Soporte Regresión se basa en predecir valores numéricos, dado que la salida es un número real, se vuelve muy difícil predecir la información disponible, que tiene infinitas posibilidades, sin embargo, la idea principal es siempre la misma: minimizar el error, individualizar el hiperplano que maximiza el margen, teniendo en cuenta que se tolera parte del error.

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Support Vector Regression









Support Vector Regression



- El hiperplano que se obtiene dentro de este algoritmo siempre tratará de moldear el comportamiento de los datos y esta curva siempre viene acompañada con un rango (**máximo margen**), tanto del lado positivo como en el negativo, el cual tiene el mismo comportamiento o forma de la curva.

- Todos **los datos que se encuentren fuera del rango son considerados errores** por lo que es necesario

calcular la distancia entre el mismo y los rangos. Esta distancia lleva por nombre epsilon y afecta la ecuación final del modelo.

- Este algoritmo funciona muy bien para datos lineales como no lineales.



Support Vector Regression

datos lineales para que sea más
quemos paso a paso cómo se

1



Lo
primero
que

debemos realizar es obtener
un hiperplano que mejor




represente el comportamiento de los datos, como el ejemplo que estamos usando son datos lineales este hiperplano es simplemente una línea, pero cuando se trabaja con datos no lineales el hiperplano es mucho más complicado a este. La fórmula para este hiperplano será la misma a la de una línea:

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Support Vector Regression



 ² El siguiente paso es construir unas bandas paralelas al hiperplano que cubra la mayor cantidad de datos, a estas bandas se le

conoce como vectores de apoyo o de soporte.



Figure 14: Construir las bandas

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com



Support Vector Regression

3 Ahora bien,

como podemos observar estas bandas

no cubrieron todos los datos, todavía tenemos puntos fuera de la misma, estos datos serían los errores y los que se deben considerar para la fórmula del algoritmo. Acá lo que se calcula es la distancia entre las bandas y el punto, a esta distancia se le da el nombre de **epsilon**.

Al final la fórmula completa para el cálculo de este algoritmo, utilizando datos lineales es la siguiente:



(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Support Vector Regression





En donde:

- w es la magnitud del vector o hiperplano
- C es una constante y debe ser mayor a 0, determina el equilibrio entre la regularidad de la función y la cuantía

hasta la cual toleramos desviaciones mayores que las bandas de soporte.



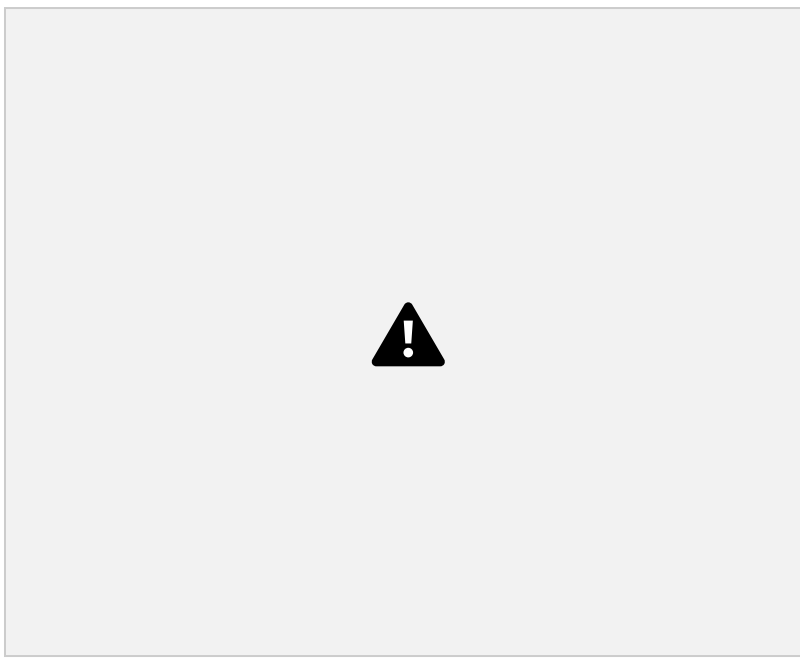
ξ y ξ^* son las variables que controlan el error cometido por la función de regresión al aproximar a las bandas.

Support Vector Regression



Por su parte **para datos no lineales** el procedimiento es exactamente igual, la diferencia es que se implementa de un





Kernel para convertir los datos lineales.





Árboles de decisión

Los árboles de decisión se destacan por su interpretabilidad y su capacidad para trabajar con datos mixtos, tanto categóricos como numéricos.





(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Aplicaciones



Los árboles de decisión tienen muchas aplicaciones valiosas en machine learning y diversas industrias:

■ Clasificación:

Identificar si un

correo es spam o no spam, o clasificar imágenes, por ejemplo, de flores o animales.

- Regresión: Predecir valores numéricos, como el precio de una casa basado en características como el tamaño y ubicación.
 - Medicina: Diagnosticar enfermedades basándose en síntomas y pruebas médicas.
 - Finanzas: Determinar riesgos de crédito al evaluar datos financieros de los solicitantes de préstamos.
 - Marketing: Segmentar a los clientes y personalizar campañas publicitarias según el comportamiento y características de los consumidores. ■
- Automoción: Sistemas de asistencia al conductor que reconocen señales de tráfico y toman decisiones en tiempo real.
- Manufactura: Detectar defectos en productos mediante imágenes y datos de sensores.

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Árboles de decisión

Un árbol de decisión se divide en varias partes clave:



■ Raíz
del árbol,

muestra

■ Nodos
Cada
prueba o

ramas
prueba.

■ Ramas

nodos y hojas, representando el resultado de una prueba en un nodo.

■ Hojas (Leaves): Los nodos terminales que no se dividen más. Representan la clase o valor final de la predicción.

■ Divisiones (Splits): Criterios que se usan para dividir los nodos internos en ramas. Cada división trata de mejorar la homogeneidad de la clase o valor resultante.



(Root): Es el nodo superior donde comienza el algoritmo. Representa toda la de datos y su división inicial.
internos (Decision Nodes): nodo interno representa una condición sobre una característica. Se dividen en basadas en el resultado de la

(Branches): Conectan

Estas partes trabajan juntas para tomar decisiones basadas en los datos de entrada de forma lógica y estructurada.



Árboles de decisión





(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Árboles de decisión



El algoritmo de un árbol de decisión usa criterios de división para crear ramas en los nodos.

Estos
tipo de

Aquí

Criterio de

■ Gini
impureza
calidad de



criterios dependen del
problema, como
clasificación o regresión.
veamos cómo se hace:

división:

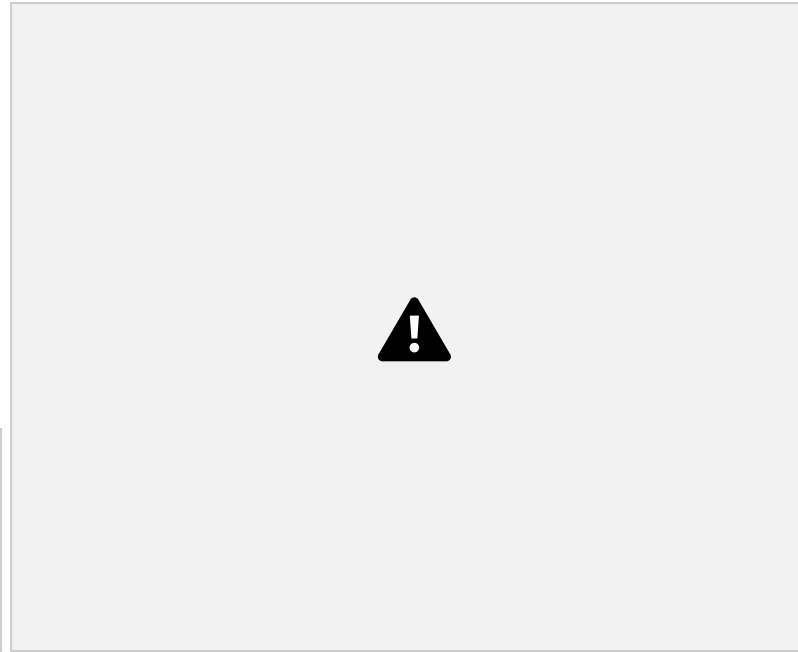
impurity: Usa la
Gini para medir la
una división. Una

división "perfecta" tendría una impureza Gini de 0.

- Entropía (Information Gain): Calcula la ganancia de información que resulta de una división. Se busca maximizar la ganancia de información.
- Mean Squared Error: Usado en regresión, mide el error cuadrático medio para evaluar divisiones.

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Árboles de decisión





(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Árboles de decisión

Selección de la mejor división:



- Evalúa cada característica

- Calcula el (Gini, entropía, posible.

- Selecciona la el mejor valor según el criterio elegido.

posible división en función de las disponibles.

criterio de división etc.) para cada división

división que produce



Recursión:

- Divide división.
- Repite hasta



los datos en ramas basadas en la mejor el proceso para cada nodo hijo que se cumpla un criterio de parada, como una profundidad máxima del árbol o una cantidad mínima de muestras por nodo.

Criterio de parada:

- Profundidad máxima del árbol.
- Número mínimo de muestras por nodo.
- Gini impurity mínimo o ganancia de información mínima.

Este proceso continúa hasta que se cumplen las condiciones de parada, resultando en un árbol que puede usar las decisiones aprendidas para hacer predicciones.

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Ejemplo

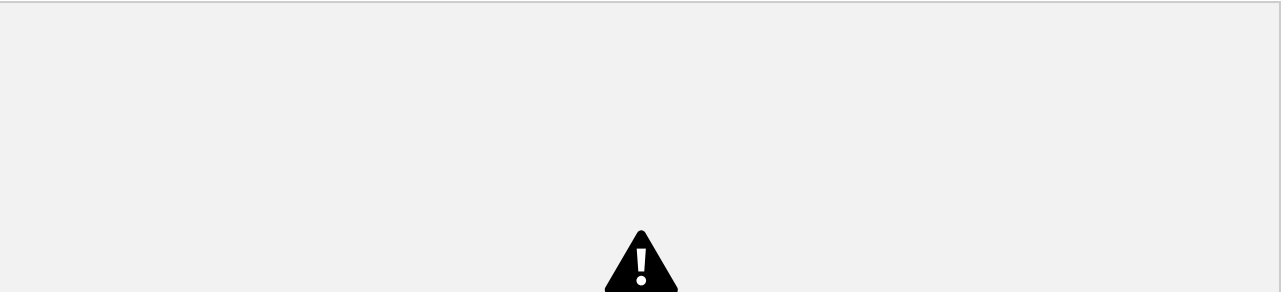
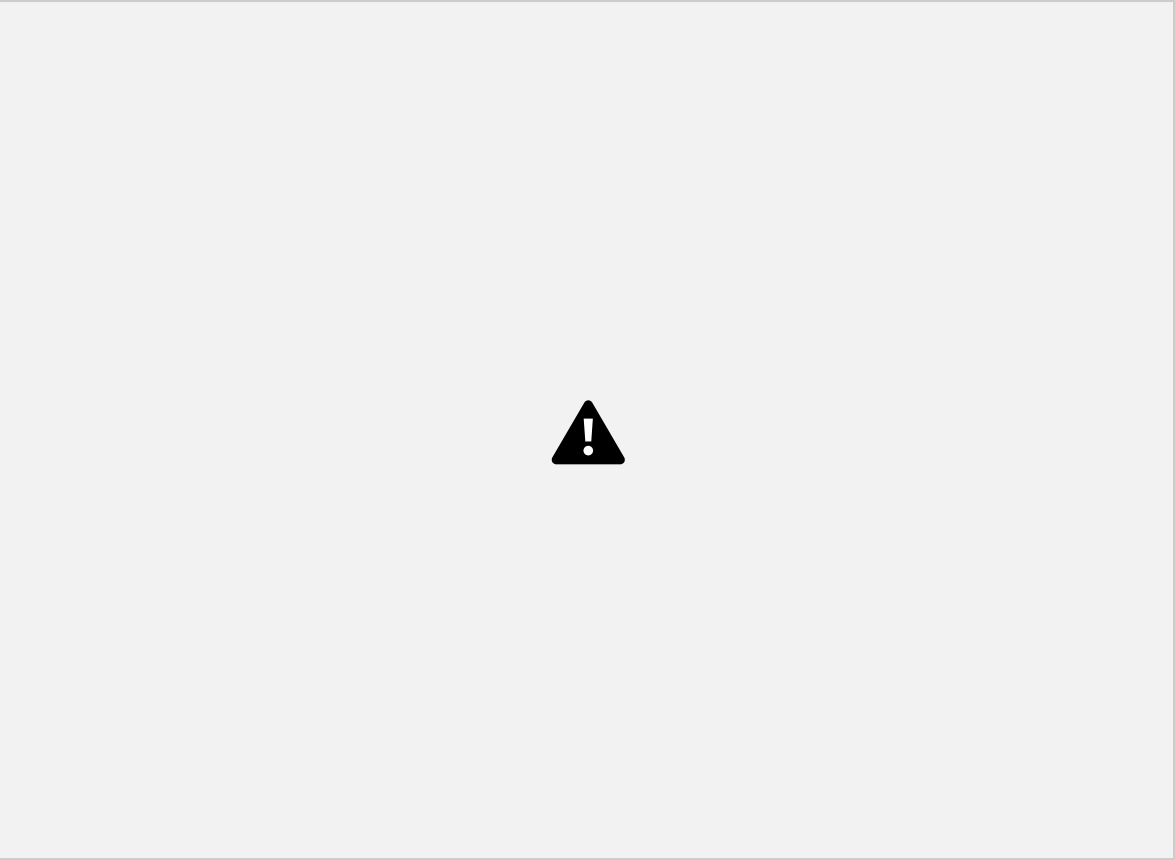
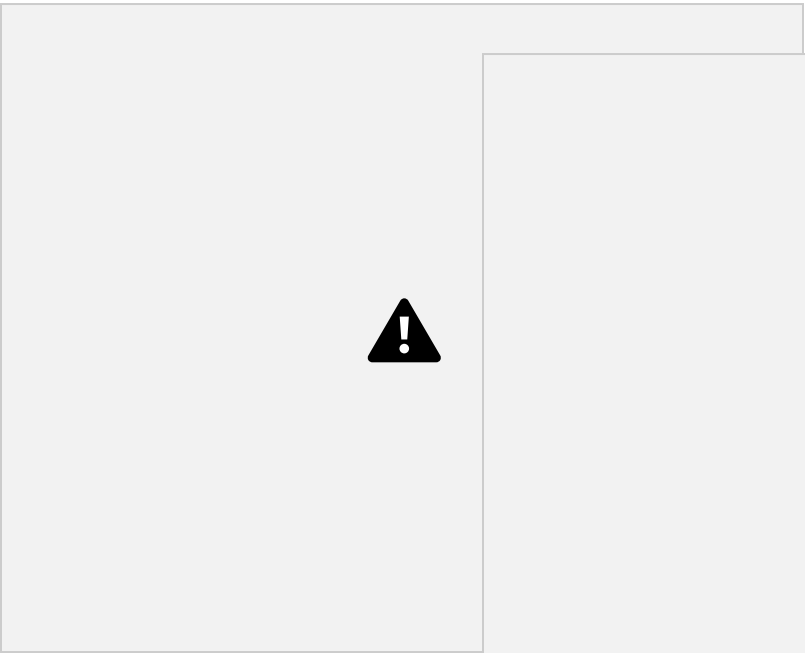




(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Ejemplo







Referencias Bibliográficas



- An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics).
- Linear Models with R by Julian J. Faraway.
- An Introduction to Statistical Learning by James, Gareth et al.
- Applied Predictive Modeling by Max Kuhn and Kjell Johnson.

□ <https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/>

<https://medium.com/datos-y-ciencia/machine-learning-supervizado-fundamentos-de-la-regresi%C3%B3n-lineal-bbcb07fe7fd>

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com





(+51) 976 760 www.datayanalytics.com info@datayanalytics.com