Regressions Models

Modelo/Método matemático usado para **predecir datos cuantitativos** (numéricos), buscando determinar la relación y efecto de una/s **variable/s 'Independiente'** sobre una única **variable 'Dependiente'**.

La relación es '**Lineal**' sí la razón de cambio (pendiente) es constante tanto en el eje '**X**' como en el eje '**Y**' al aumentar en 1.

Regresión Lineal Simple

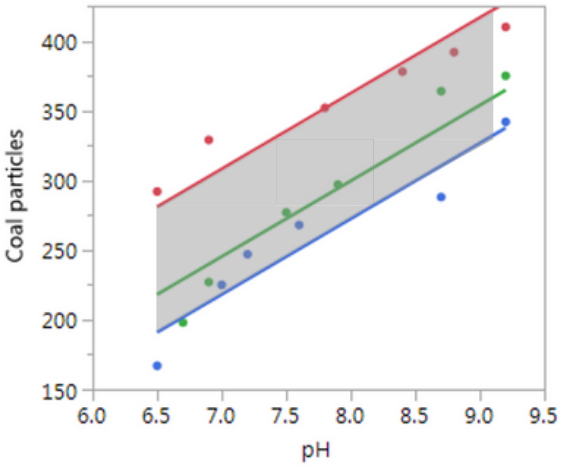
Entrenar una mejor '**media**' a través de los datos, evaluando la influencia relativa de una variable independiente (**X**) sobre una variable dependiente (**Y**) en un plano 2D.

Problemas con los valores Atípicos.
(una unidad muy distante al resto de los datos)

Regresión Lineal Múltiple

Entrenar para múltiples variables independientes (**X**) su mejor '**media**', evaluando la influencia de cada una sobre una variable dependiente (**Y**) en un plano 2D y 3D.

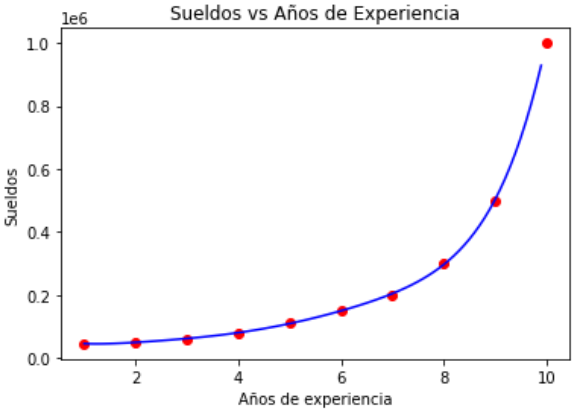
Problemas con los valores atípicos y con DataSets incompletos.



La relación es '**No Lineal**' si la razón de cambio (pendiente) es producto de operaciones matemáticas no proporcionales lo que se traduce en una función curva.

Regresión Polinomial

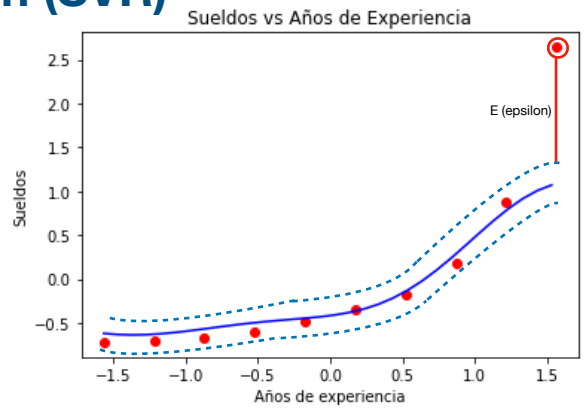
Entrena una mejor '**media**' convirtiendo los datos a '**Cuadráticos (^2)**' o '**Cúbicos (^3)**', así, determinando la exponencial de los datos obtenemos una 'media' curva. **Problemas con la relación Sesgo / Varianza.** (modelo bueno con los datos de entrenamiento pero malo en el pronostico)



Support Vector Regression (SVR)

Entrenar un mejor '**hiperplano**' (media) con dos bandas ('positiva' y 'negativa') qué cubran los datos de acuerdo a un '**rango**' (máximo margen) compuesto por ambas bandas.

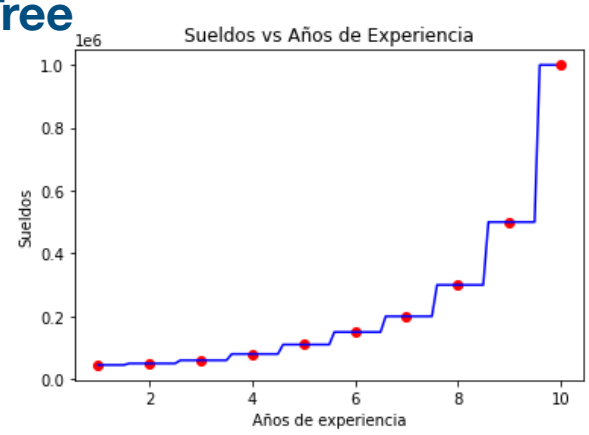
Es para DataSets chicos, requiere DataSets completos y limpios, y requiere 'Escalado' de los datos.



Regresión con Decision Tree

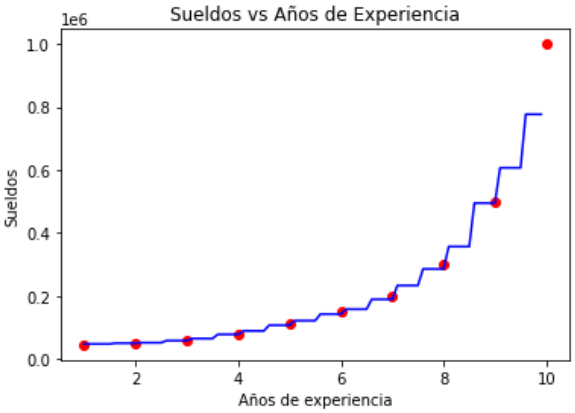
Entrenar una mejor '**vía**' (media) que represente el resultado, costos y/o consecuencias de una previa decisión en términos predictivos.

Problemas con DataSets pequeños ya que fácilmente se produce 'Overfitting' (sobre ajuste)**.**



Regresión con Random Forest

Entrenar una mejor '**vía**' (media) con un numero **N** de '**Arboles de Decisión**' independientes que al combinar los resultados se obtiene una media de predicción mas estable y precisa que la de un solo '**Arbol**'. **Overfitting, costosos y suelen ser lentos al ejecutar el modelo** (es difícil hallar el N de arboles adecuados).



Para determinar el modelo que mejor se ajuste a nuestra problemática se utiliza la técnica: '**K-Fold Cross Validation**' (se divide el conjunto de datos del DataSet en '**k contenedores iguales**', luego se determina el conjunto de 'Entrenamiento' y 'Testing', después se efectúa la técnica de validación cruzada (cada contenedor participa tanto en el conjunto de Entrenamiento como de Testing) y finalmente se promedian todas las validaciones dando como resultado una mejor '**vía**' (media)).

Evaluar Modelos de Regresión

Estas técnicas dan información sobre el modelo que estamos usando para evaluar el sentido estadístico y matemático que tienen.

R^2 (R Cuadrado)

Es un valor que mejora al tener mas variables en el modelo para predecir, a más alto, mejor.

R^2 Adj (R Cuadrado ajustado)

Es un '**rango**' entre 0 y 1 que es asociado a cada variable '**independiente**', para calificar el nivel de influencia que tiene cada una sobre la variable '**dependiente**'.

Podemos ajustar nuestro modelo haciendo eliminación hacia atrás hasta obtener el mayor sentido estadístico.

Coeficientes

Es importante saber leer he interpretar los '**Coeficientes**' estadísticos por qué pueden ser de signo, magnitud y tipo de unidades diferentes.

Ej: Medidas⟷Dólares

Classifications Models

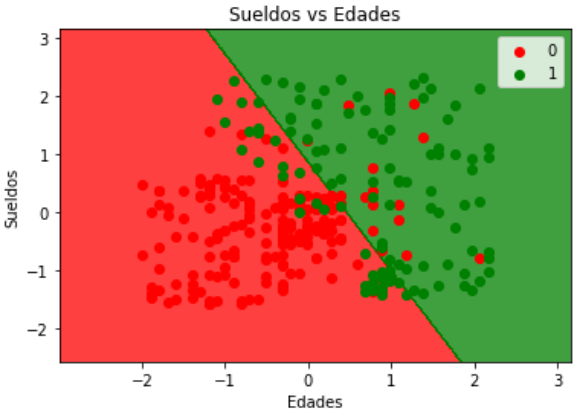
Modelo/Método matemático para asignar una '**Clase**' (tipo) a cada dato del conjunto de dato de entrenamiento, a modo de predecir la clase a la que pertenecen los nuevos datos que serán evaluados (registrados), clasificar cada dato en una clase.

La relación es '**Lineal**' sí la razón de cambio (pendiente) es constante tanto en el eje '**X**' como en el eje '**Y**' al aumentar en 1.

Regresión Logística

Entrenar una mejor '**media**' que divida homogéneamente **dos clases** que nos permita evaluar y predecir la presencia o ausencia de alguna característica en los datos.

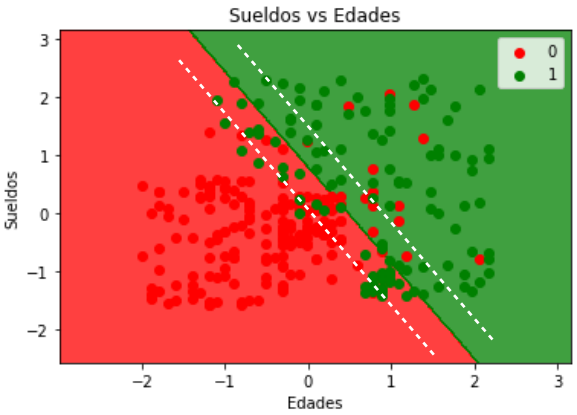
Problemas con variables Atípicas, difícil trabajar con datos No Lineal



Support Vector Machine (SVM)

Entrenar un '**Hiperplano**' qué separe de la mejor manera posible dos '**clases**' diferentes aceptando un '**Margen**' (rango) de error.

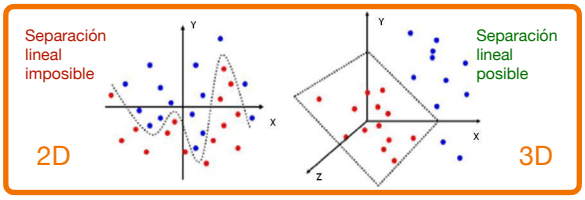
Eficaz donde el número de dimensiones es mayor al número de muestras.

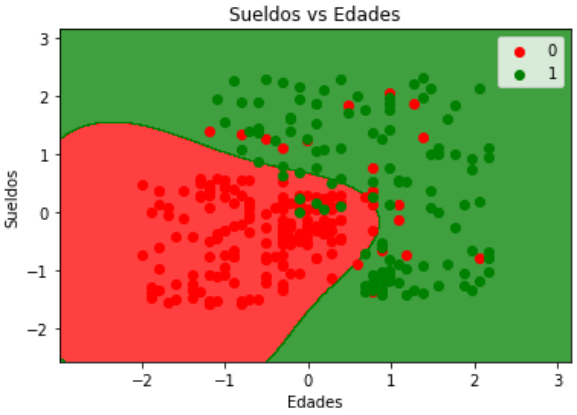


La relación es '**No Lineal**' si la razón de cambio (pendiente) es producto de operaciones matemáticas no proporcionales lo que se traduce en una función curva.

Kernel Gaussiano (RBF)

Es una '**Función de Kernel**' posible en un modelo de **SVM** que aumenta la dimensión del modelo para hacer los datos linealmente separables (efecto: Campana de Gauss) .

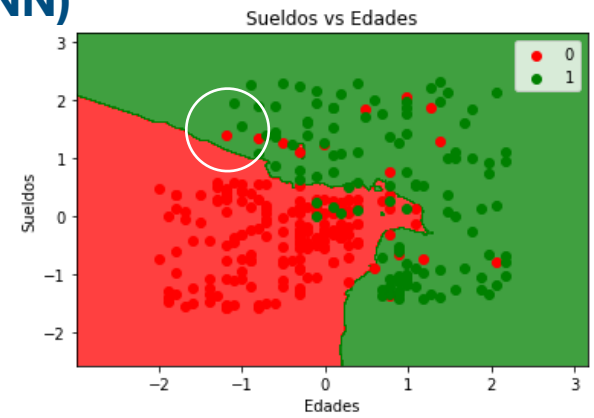




K-Nearest Neighbors (K-NN)

K Vecinos más Cercanos entrena un '**limite**' construido evaluando la '**clase**' y la '**cantidad**' de los vecinos mas cercanos (datos) de cada dato dentro de un radio proporcional.

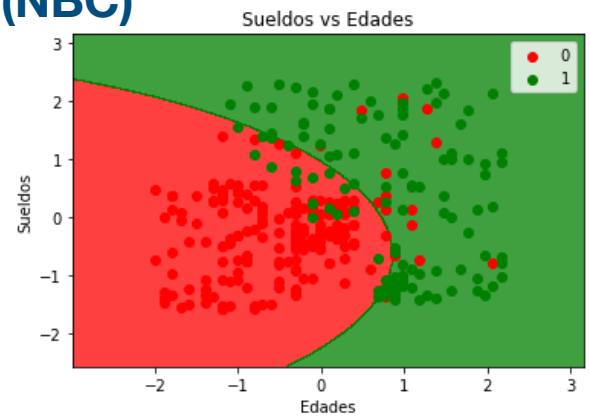
Costoso, lento para predecir, y funciona mal en DataSets grandes.



Clasificador Naïve Bayes (NBC)

Naïve : **Ingenuo |** Entrena una media entre dos '**clases**' asumiendo que no existen variables '**interdependientes**' (relacionadas) en el conjunto de datos, lo cual permite una operación rápida y superficial al efectuar una predicción.,

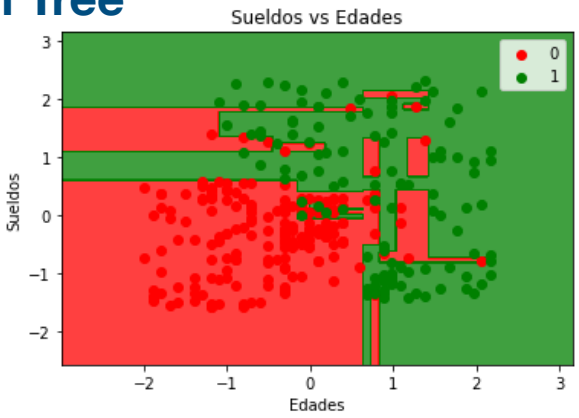
Pasa por alto patrones colectivos y relaciones complejas en los datos.



Clasificación con Decision Tree

Entrenar un único **mapa** (ruta) que secciona los datos de un DataSet para uso predictivo en clasificar nuevos datos. Además explicar desiciones relacionadas.

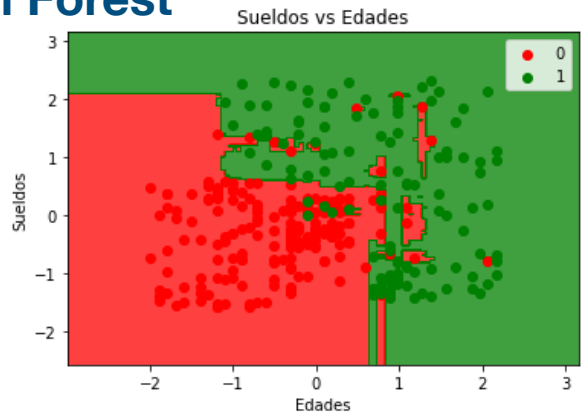
Problemas de Overfitting y exige equilibrar (equidistar extremos Δ) el conjunto de datos previamente.



Clasificación con Random Forest

Entrena un **mapa** (ruta) robusto y efectivo (promedio de cada predicción de todos los árboles) que secciona los datos del DataSet para uso predictivo en clasificar nuevos datos.

Es lento; cada árbol debe hacer una predicción diferente y una votación.
No suelen sufrir de Overfitting.



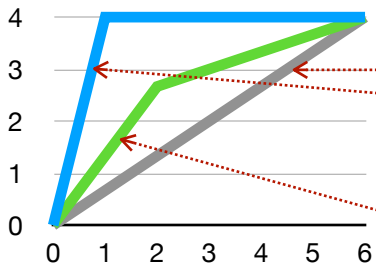
Evaluar Modelos de Clasificación

Matriz de Confusión

Nos muestra los **aciertos** y **errores** sobre la muestra de '**Testing**'

Training	N	T-N
Testing	X	T-X

Curva CAP



Encontrar el '**Modelo Aleatorio**' y '**Modelo Perfecto**' para evaluar entonces un '**Buen Modelo**'

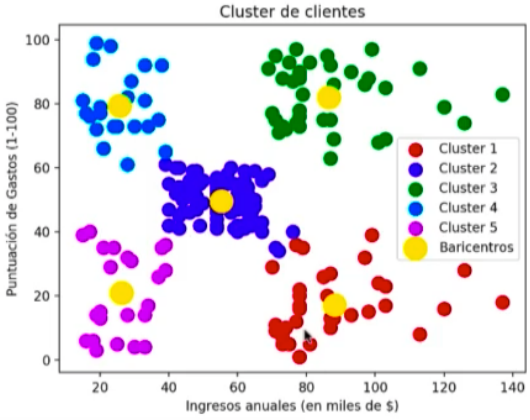
Clustering Models

Modelo matemático para identificar en un conjunto de datos patrones que permitan '**agrupar**', segmentar, aglomerar, categorizar los datos para posteriormente '**etiquetar**' los registros. La categorización puede ser llevada a cabo sin previamente conocer la características de los datos ni sus aglomeraciones.

K-Means

Agrupar/segmentar en '**K**' **grupos** (categorías) los datos. Se efectúan operaciones a los '**Baricentros**' y sus distancias respecto a los datos, hasta centrarlos bien en cada grupo.

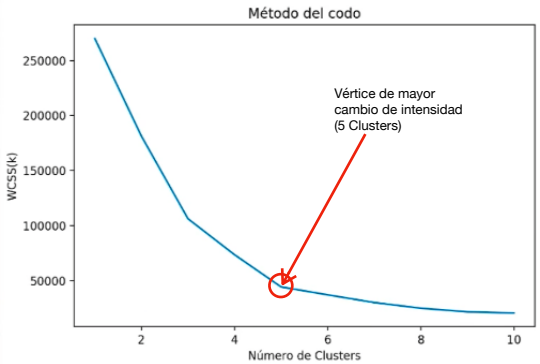
Según dónde se inicialicen los 'Baricentros' los resultados pueden ser bien diferentes.



La **Técnica del Codo** es un método para averiguar la cantidad adecuada de '**Clusters**' (grupos) que debe tener nuestro modelo.

El vértice con mayor intensidad de cambio, nos revela cuándo el sentido matemático respecto a la cantidad de '**clusters**' comienza a perder el sentido.

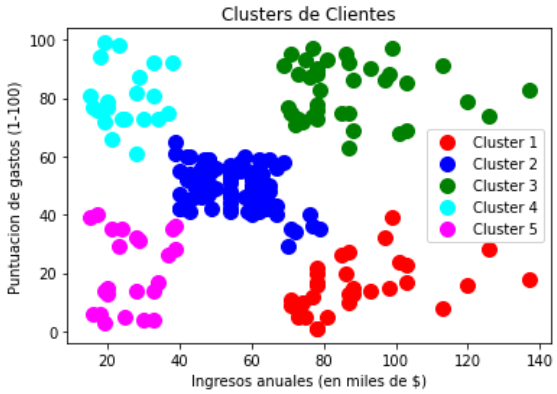
Aquel vértice nos muestra la cantidad optima de 'clusters' que debe tener nuestro modelo.



Clustering Jerárquico

Agrupar (aglomerar) y/o dividir los datos en ‘clusters’. Se averiguan todas las posibles aglomeraciones y divisiones que se les puede determinar a los datos.

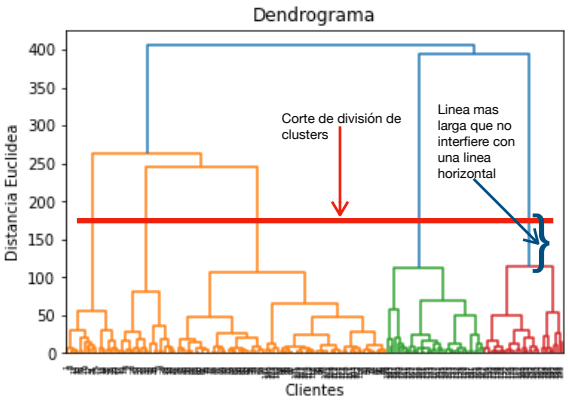
Problemas con DataSets muy grandes.



Los **Dendrograma** son una herramienta usados como método para averiguar la cantidad adecuada de ‘**Clusters**’ (grupos) que debe tener nuestro modelo.

Sé ‘**jerarquizan**’ todas las relaciones posibles y de manera visual se determine el corte que dará como resultado la cantidad de ‘**clusters**’.

El corte se recomienda efectuarlo a la linea mas larga que no interfiere con ninguna linea horizontal.



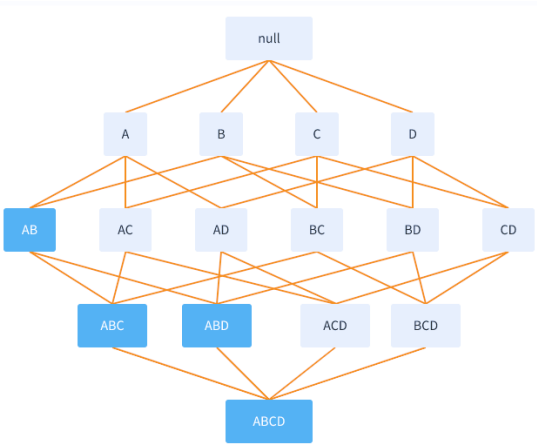
Association Rules Models

Modelo/Método matemático usado para identificar '**Reglas de Asociación**' en un conjunto de datos que ayuden a mostrar la '**probabilidad**' y/o '**relación**' entre los datos.

Apriori Algorithm

Encuentra '**asociaciones en amplitud**' entre los datos del DataSet he identifica ítems frecuentes **relacionados por reglas** con un **soporte** (% de repetirse una regla encontrada) y **confianza** (% predictivo de cumplirse una regla escogida) determinada.

Problemas con DataSets pequeños por encontrar asociaciones falsas.

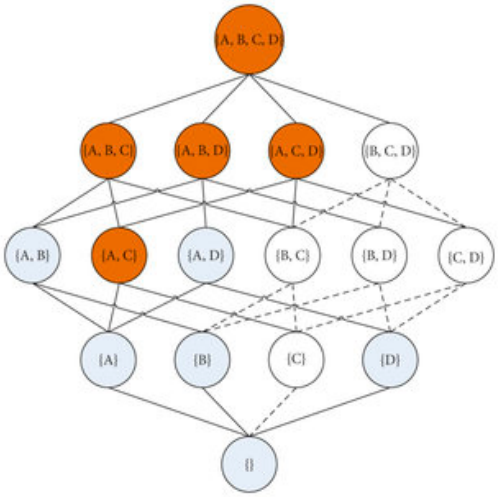


Eclat Algorithm

Encuentra '**asociaciones en profundidad**' entre los datos del DataSet con un **soporte** determinado.

Mejor ante '**Apriori Algorithm**' en velocidad y en identificar datos frecuentes.

Problemas don DataSet muy grandes.



Semiología y Conceptos

Conceptos generales sobre **AI** y **Data Science**

Artificial Intelligence
- Simular por procesos de inteligencia humana por medio de algoritmos.

Machine Learning
- Sub disciplina de la AI orientada a crear sistemas que aprendan automáticamente.

Deep Learning
- Sub disciplina del Machine Learning orientado a emular el modo de aprendizaje de los seres humanos.
- El aprendizaje profundo puede considerarse como una forma de automatizar el aprendizaje predictivo.

Data Science
- Disciplina científica que analiza y manipula grandes cantidades de datos para comprender relaciones, descubrir patrones, extraer información y apoyar la toma de decisiones.

Escalar variables
- Normalización de un conjunto de datos numéricos confinándolos en un rango [0, 1], con el objetivo de que todas las características de los datos compartan un mismo valor medio y una misma desviación media.