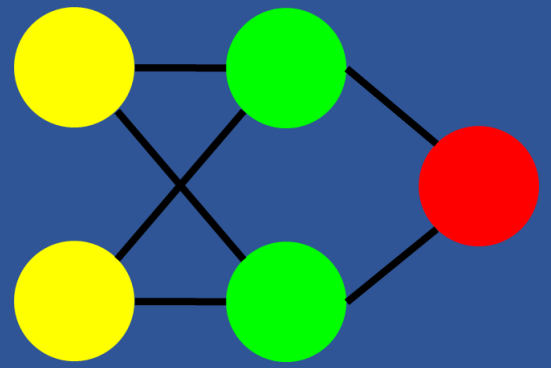
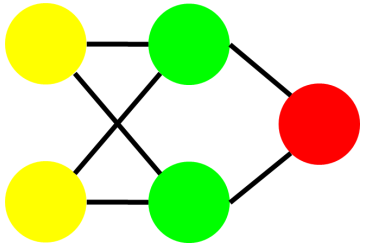
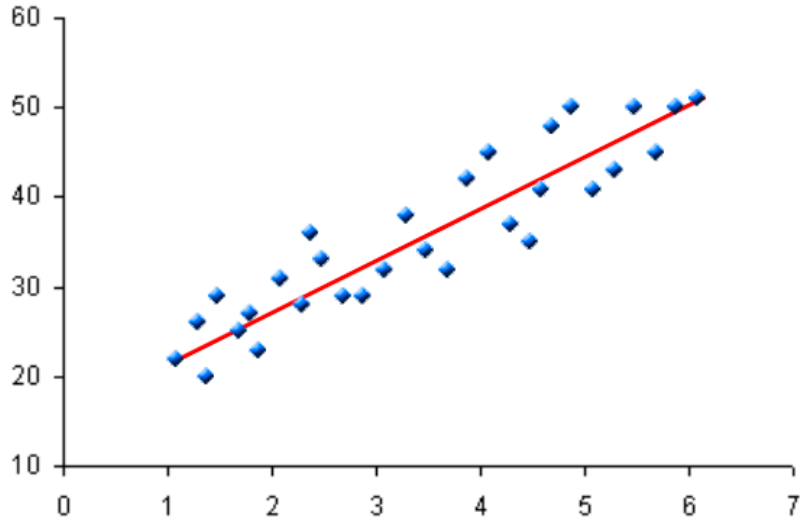
Modelos de Regresión



Algoritmos de Regresión



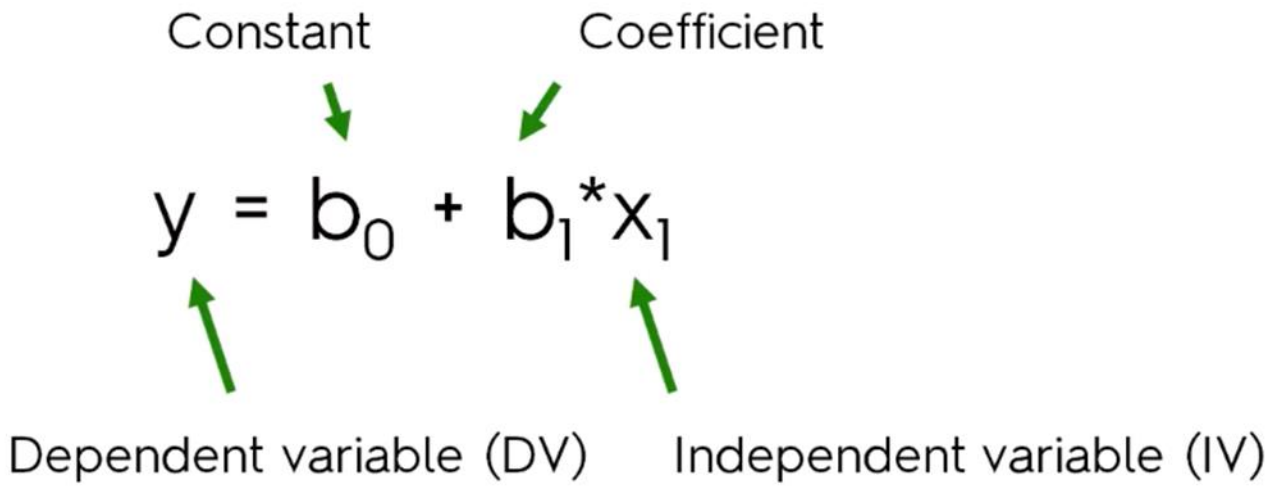
- Regresiones Lineales
- Regresiones Polinomiales
- Arboles de Decisión



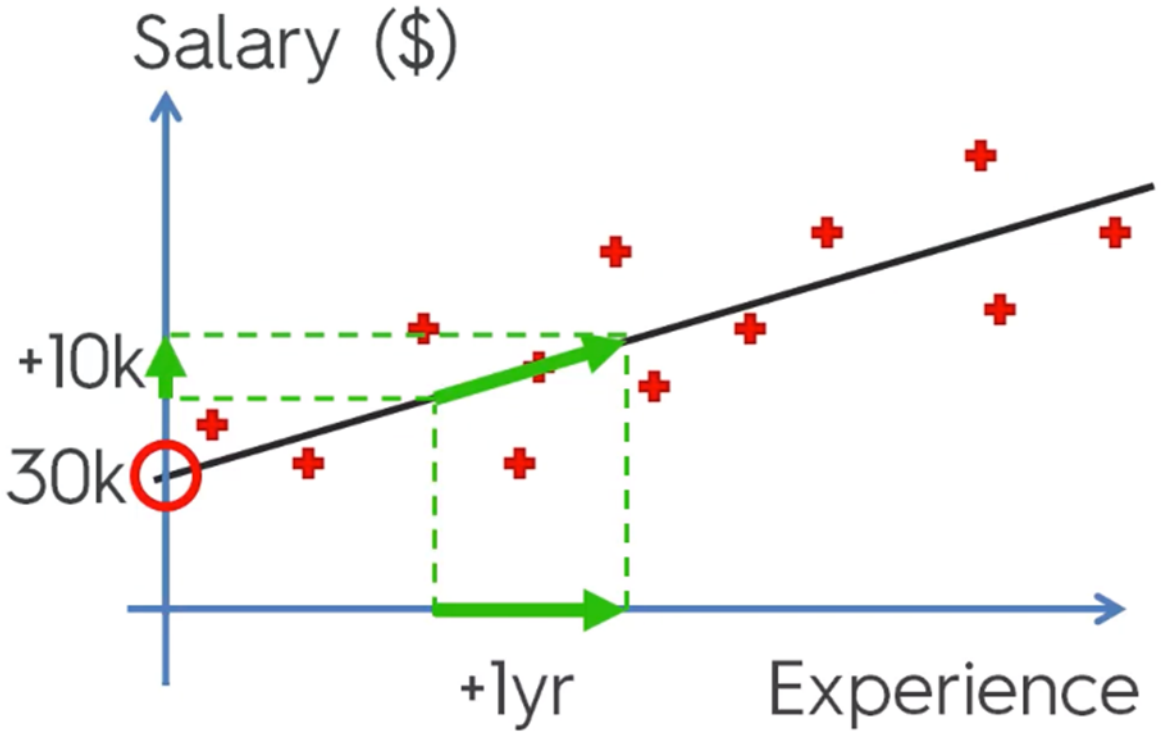
Regresiones Lineales

I. Regresiones Lineales Simples

Simple Linear Regression



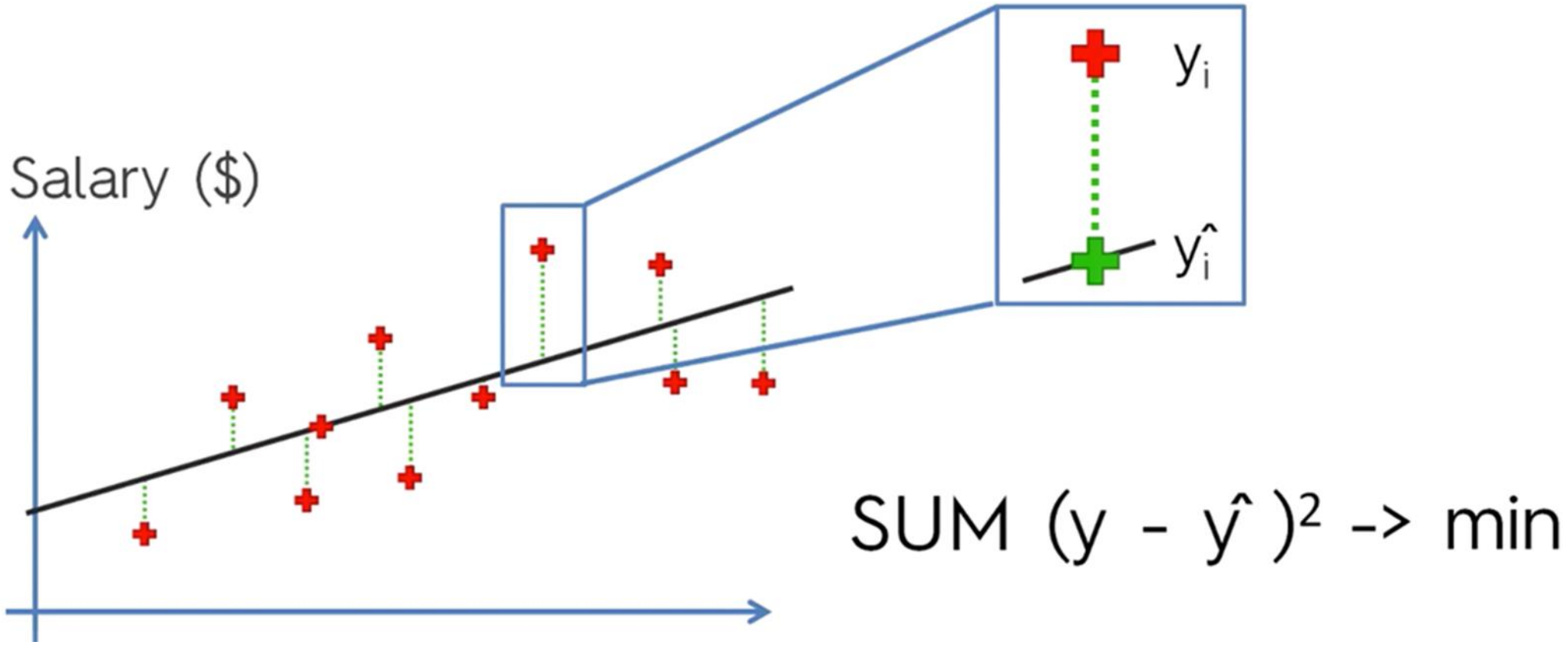
I. Regresión Lineal Simple



$$y = b_0 + b_1 * x$$

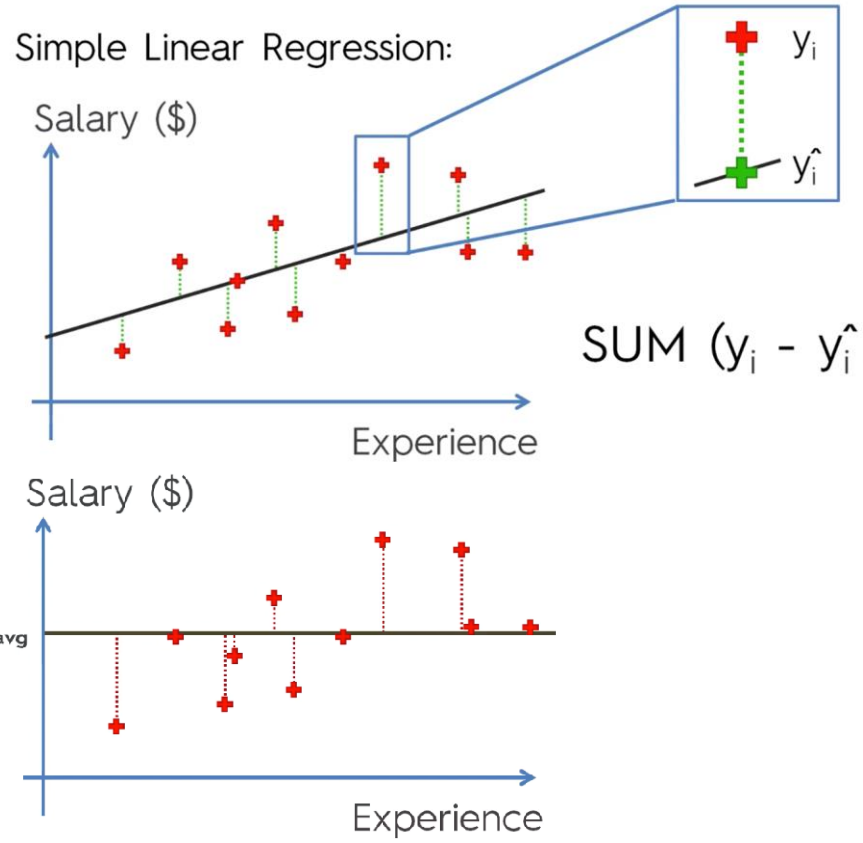
Salary = b_0 + b_1 *Experience

I. Regresión Lineal Simple



I. Regresión Lineal Simple

- Evaluando la performance de Modelos de Regresión
  - R-cuadrado



$$SUM\ (y_i - \hat{y_i})^2 \rightarrow min$$

$$SS_{res} = SUM\ (y_i - \hat{y_i})^2$$

$$SS_{tot} = SUM\ (y_i - y_{avg})^2$$

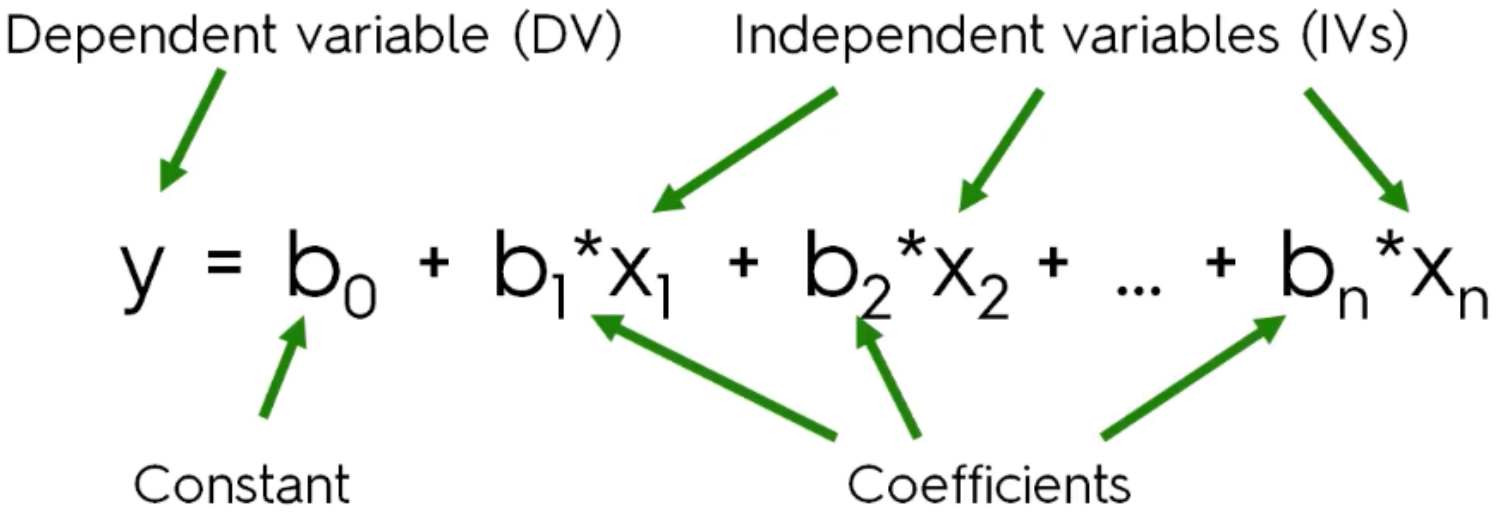
$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

II. Regresiones Lineales Múltiples

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression



II. Regresiones Lineales Múltiples

- La regresión lineal tiene muchos supuestos

1. Linearity
2. Homoscedasticity
3. Multivariate normality
4. Independence of errors
5. Lack of multicollinearity

II. Regresiones Lineales Múltiples

- Variables Dummy

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

II. Regresiones Lineales Múltiples

- Variables Dummy

Dummy Variables

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

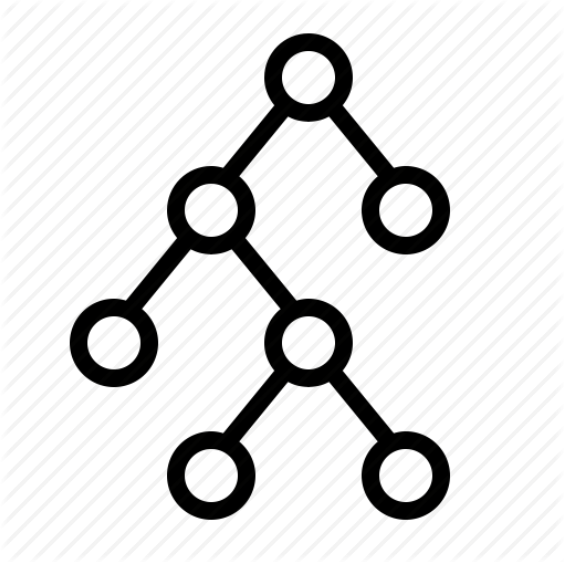
II. Regresiones Lineales Múltiples – Hands On

- DataSet

R&D Spend	Administration	Marketing Spend	State	Profit
165349.2	136897.8	471784.1	New York	192261.83
162597.7	151377.59	443898.53	California	191792.06
153441.51	101145.55	407934.54	Florida	191050.39
144372.41	118671.85	383199.62	New York	182901.99
142107.34	91391.77	366168.42	Florida	166187.94
131876.9	99814.71	362861.36	New York	156991.12
134615.46	147198.87	127716.82	California	156122.51
130298.13	145530.06	323876.68	Florida	155752.6
120542.52	148718.95	311613.29	New York	152211.77
123334.88	108679.17	304981.62	California	149759.96
101913.08	110594.11	229160.95	Florida	146121.95
100671.96	91790.61	249744.55	California	144259.4
93863.75	127320.38	249839.44	Florida	141585.52
91992.39	135495.07	252664.93	California	134307.35
119943.24	156547.42	256512.92	Florida	132602.65
114523.61	122616.84	261776.23	New York	129917.04
78013.11	121597.55	264346.06	California	126992.93
94657.16	145077.58	282574.31	New York	125370.37

II. Regresiones Lineales Múltiples – Hands On

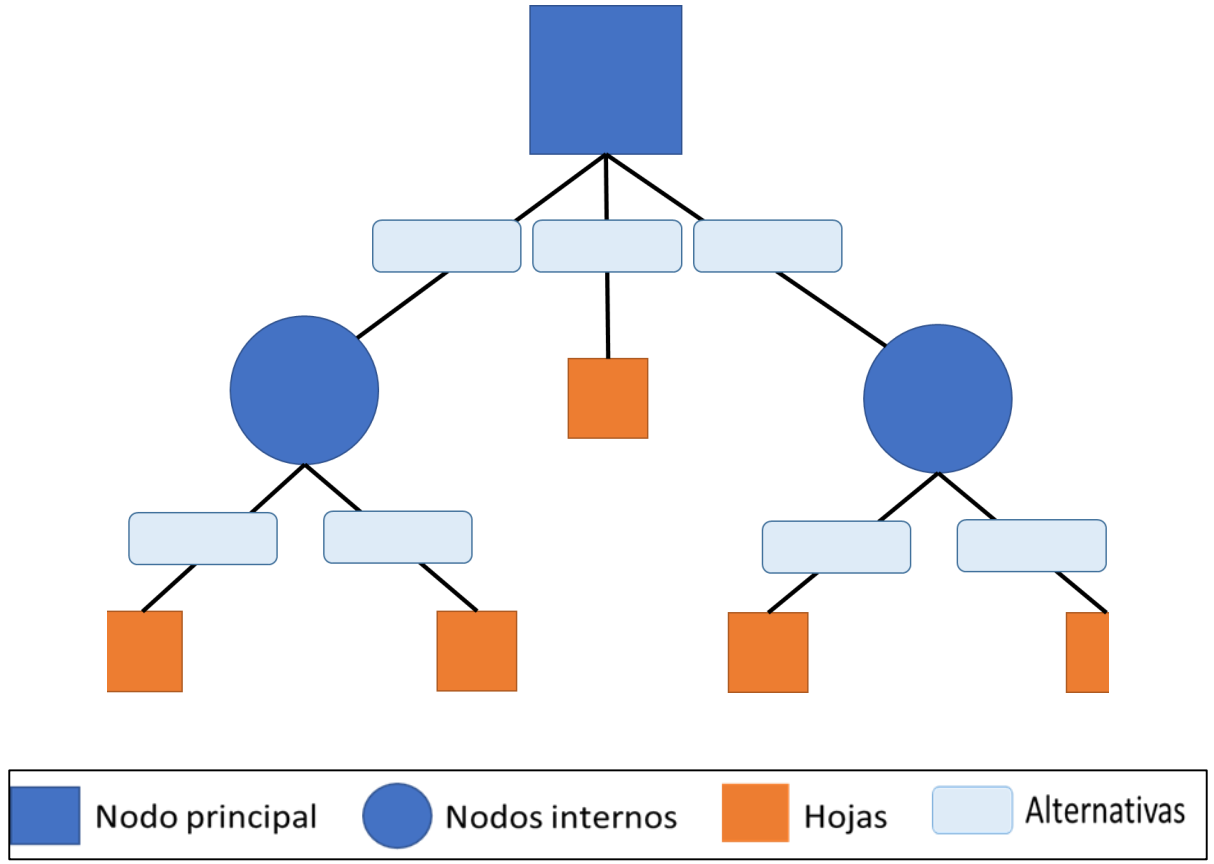
```r
# Multiple Linear Regression

# Importing the dataset
dataset = read.csv('50_Startups.csv')

# Encoding categorical data
dataset$State = factor(dataset$State,
                       levels = c('New York', 'California', 'Florida'),
                       labels = c(1, 2, 3))

# Splitting the dataset into the Training set and Test set
# install.packages('caTools')
library(caTools)
set.seed(123)
split = sample.split(dataset$Profit, SplitRatio = 0.8)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)

# Feature Scaling
# training_set = scale(training_set)
# test_set = scale(test_set)

# Fitting Multiple Linear Regression to the Training set
regressor = lm(formula = Profit ~ .,
               data = training_set)

# Predicting the Test set results
y_pred = predict(regressor, newdata = test_set)
```



Decisión Tree Regression

Arboles de Decisión

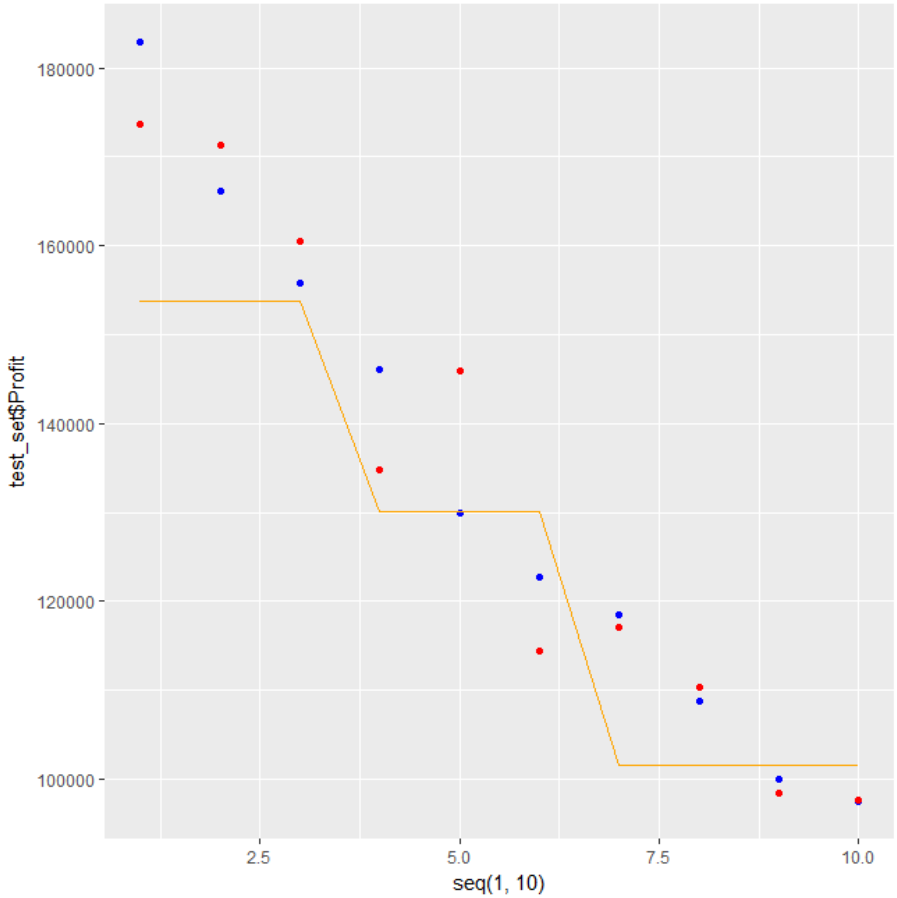
- Algoritmo de Clasificación Supervisada. Busca una variable dependiente concreta.
- Sus variables dependientes e independientes pueden ser cuantitativas o cualitativas
- Se usa los **Arboles de Regresión** para entrenar un modelo que permita predecir una variable dependiente cuantitativa.



Arboles y Regresiones

- La Regresión lineal es el método más usado en estadística para predecir valores de variables continuas debido a su fácil interpretación, pero en muchas situaciones los supuestos para aplicar el modelo no se cumplen y algunos usuarios tienden a forzarlos llevando a conclusiones erróneas.

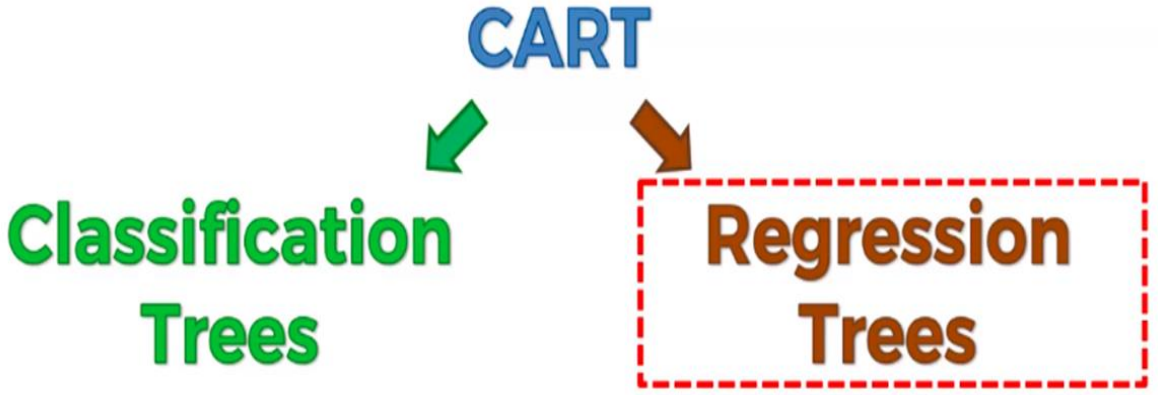
Caso	Observado	Pred_RegresionLineal	Pred_DecisionTree
1	182901.99	173687.21	153771.3
2	166187.94	171299.96	153771.3
3	155752.60	160499.08	153771.3
4	146121.95	134783.16	130087.3
5	129917.04	145873.04	130087.3
6	122776.86	114467.75	130087.3
7	118474.03	117025.30	101557.1
8	108733.99	110369.71	101557.1
9	99937.59	98447.39	101557.1
10	97483.56	97668.22	101557.1



http://www.bdigital.unal.edu.co/9474/1/71269839.2013.pdf

CART : Arboles de Clasificación y Regresión

- Los arboles de regresión CART son una alternativa de regresión que no requiere supuestos sobre los datos a analizar y es un método de fácil interpretación de los resultados.
- Los arboles de clasificación y regresión (CART) es un método que utiliza datos históricos para construir arboles de clasificación o de regresión los cuales son usados para clasificar o predecir nuevos datos. Estos arboles CART pueden manipular fácilmente variables numéricas y/o categóricas. Entre otras ventajas esta su robustez a outliers, la invarianza en la estructura de sus arboles de clasificación o de regresión a transformaciones monótonas de las variables independientes, y sobre todo, su interpretabilidad



Decisión Tree Regression - Intuición

Decisión Tree Regression - Intuición