CC - AI

Programación para la Computación Científica - IA

Regresión Lineal

Universidad Sergio Arboleda
Prof. John Corredor

Estimación de Ganancias en una Compañia

- Una compañia de capital de riesgo quiere invertir, y esta haciendo un estudio para saber en cual compañía deberían invertir

Estimación de Ganancias en una Compañia

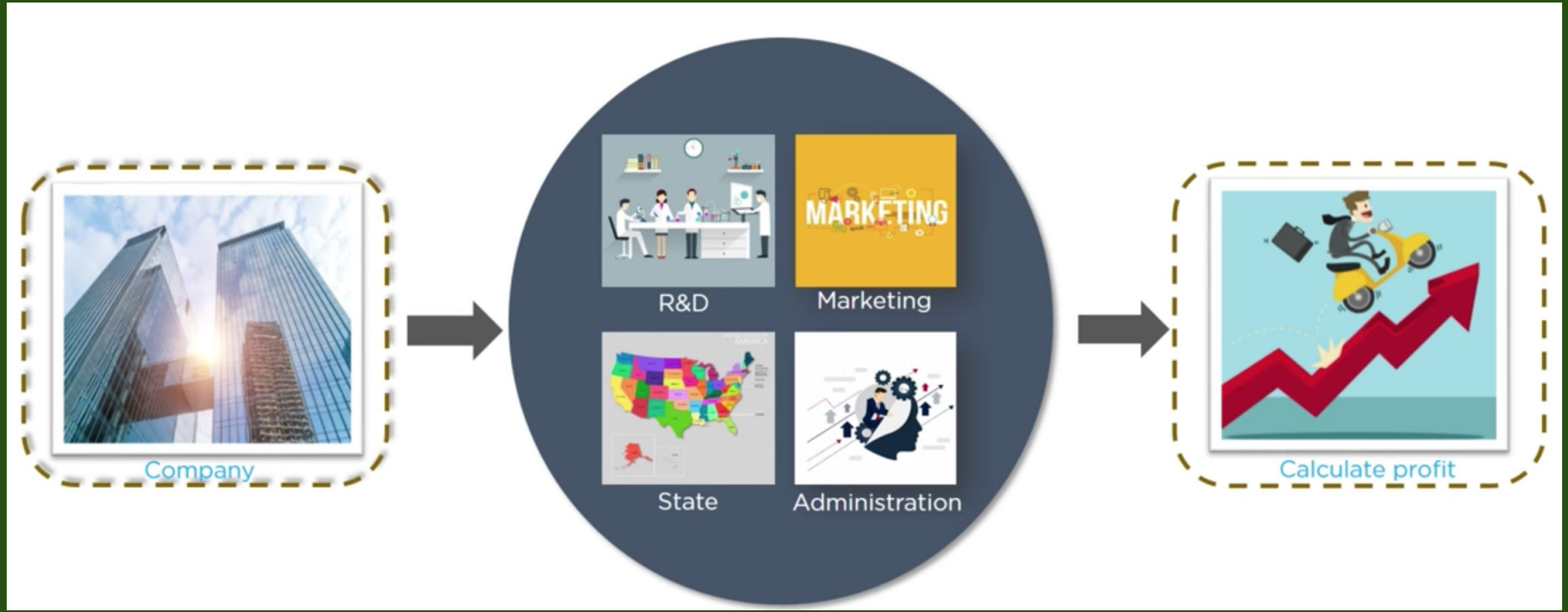
Idea

Decide companies to invest

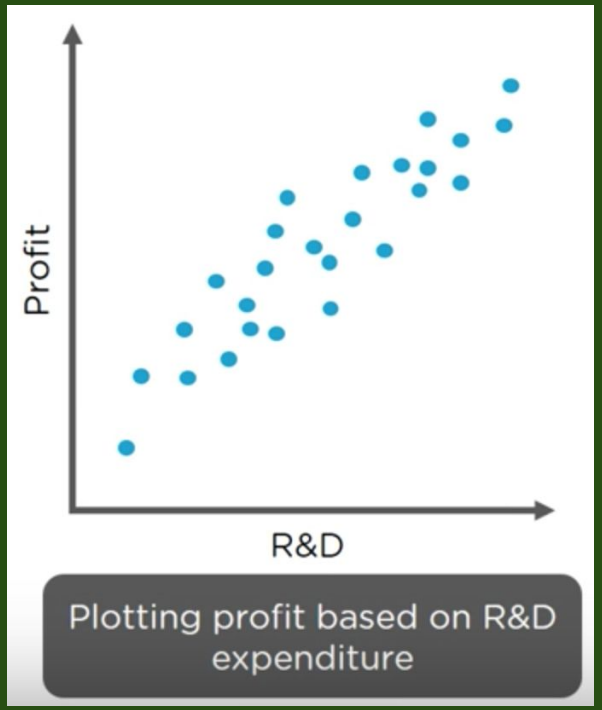
Predict the profit companies make

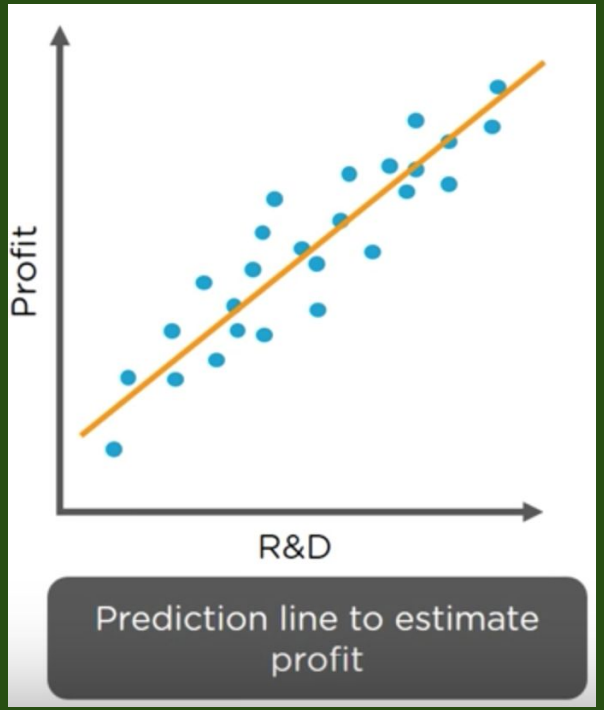
Based on companies expenses

Estimación de Ganancias en una Compañia



Estimación de Ganancias en una Compañia





Agenda

- Introducción a Machine Learning
- Algoritmos Machine Learning
- Aplicaciones de Regresión Lineal
- Comprensión de la Regresión Lineal
- Multiple Regresión Lineal
- Caso de Uso: Estimación de Ganacias de una Compañía

Introducción a Machine Learning

- basado en la cantidad de lluvia, ¿cuánto sería el rendimiento de la cosecha?

Variables Dependientes / Independientes

Independent variable

A variable whose value does not change by the effect of other variables and is used to manipulate the dependent variable. It is often denoted as **X**.

In our example:

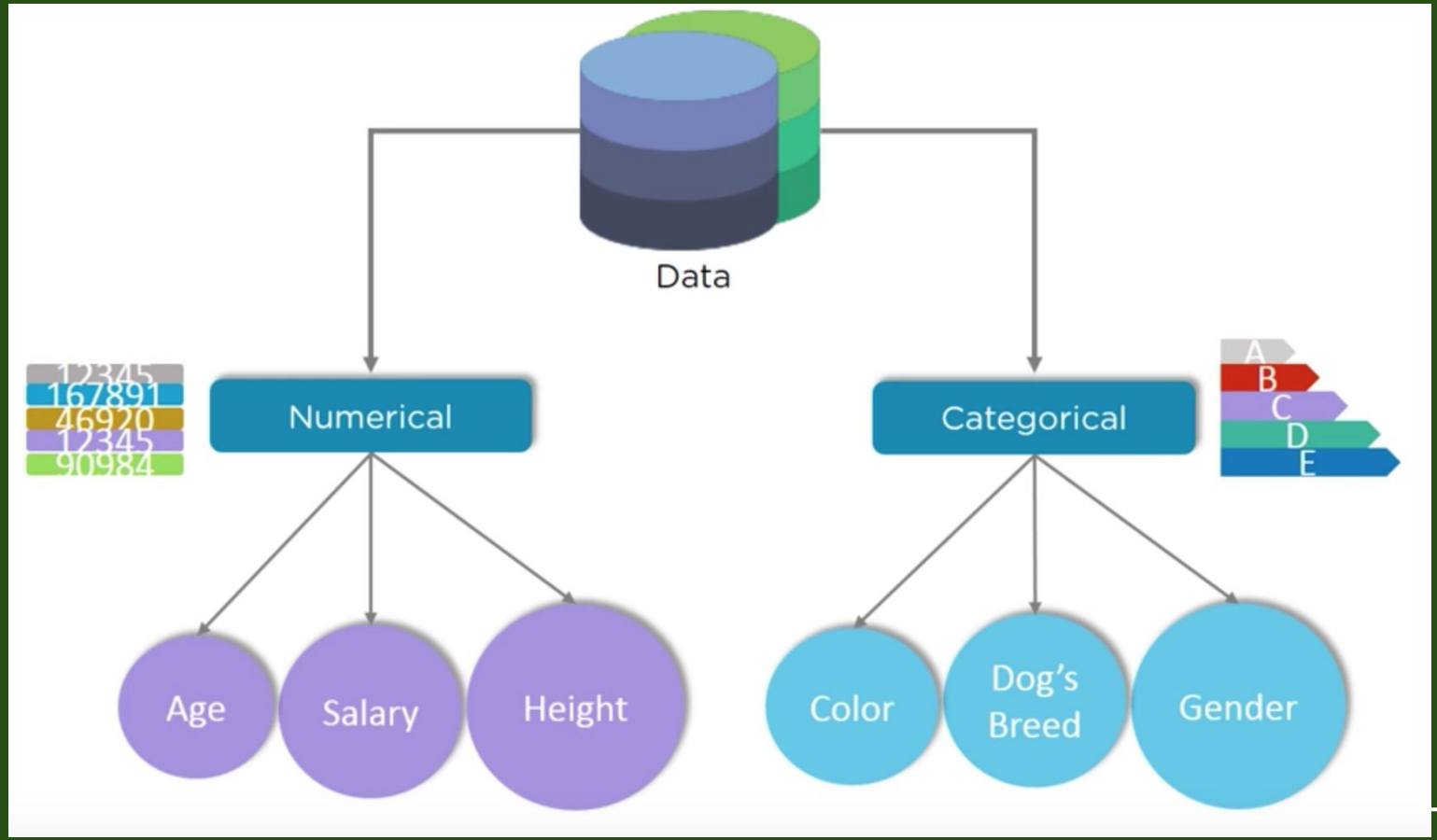
Rainfall – Independent variable

Dependent variable

A variable whose value change when there is any manipulation in the values of independent variables. It is often denoted as **Y**.

Crop yield – Dependent variable

Valores Numericos y Categoricos

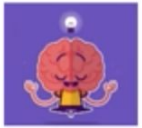


Algoritmos de Machine Learning

Machine Learning Algorithms

Supervised

Unsupervised



Reinforcement

Algoritmos de Machine Learning: Sin Supervisión

- Aquí, el algoritmo de aprendizaje automático estudia los datos para identificar patrones. No hay una clave de respuesta o un operador humano para proporcionar instrucción. En cambio, la máquina determina las correlaciones y las relaciones mediante el análisis de los datos disponibles.

- En un proceso de aprendizaje no supervisado, se deja que el algoritmo de aprendizaje automático interprete grandes conjuntos de datos y dirija esos datos en consecuencia. Así, el algoritmo intenta organizar esos datos de alguna manera para describir su estructura. Esto podría significar la necesidad de agrupar los datos en grupos u organizarlos de manera que se vean más organizados.

- A medida que evalúa más datos, su capacidad para tomar decisiones sobre los mismos mejora gradualmente y se vuelve más refinada.

Algoritmos de Machine Learning: Refuerzo

- El aprendizaje por refuerzo se centra en los procesos de aprendizajes reglamentados, en los que se proporcionan algoritmos de aprendizaje automáticos con un conjunto de acciones, parámetros y valores finales.

- Al definir las reglas, el algoritmo de aprendizaje automático intenta explorar diferentes opciones y posibilidades, monitorizando y evaluando cada resultado para determinar cuál es el óptimo.

- En consecuencia, este sistema enseña la máquina a través del proceso de ensayo y error. Aprende de experiencias pasadas y comienza a adaptar su enfoque en respuesta a la situación para lograr el mejor resultado posible.

Algoritmos de Machine Learning: Supervisado

Machine Learning Algorithms

Supervised

Regression

Classification

Algoritmos de Machine Learning: Supervisado

Machine Learning Algorithms

Supervised

Regression

Simple Linear Regression

Multiple Linear Regression

Polynomial Linear Regression

Apliaciones: Regresion Lineal

Economic Growth

Used to determine the Economic Growth of a country or a state in the coming quarter, can also be used to predict the GDP of a country

Product price

Can be used to predict what would be the price of a product in the future

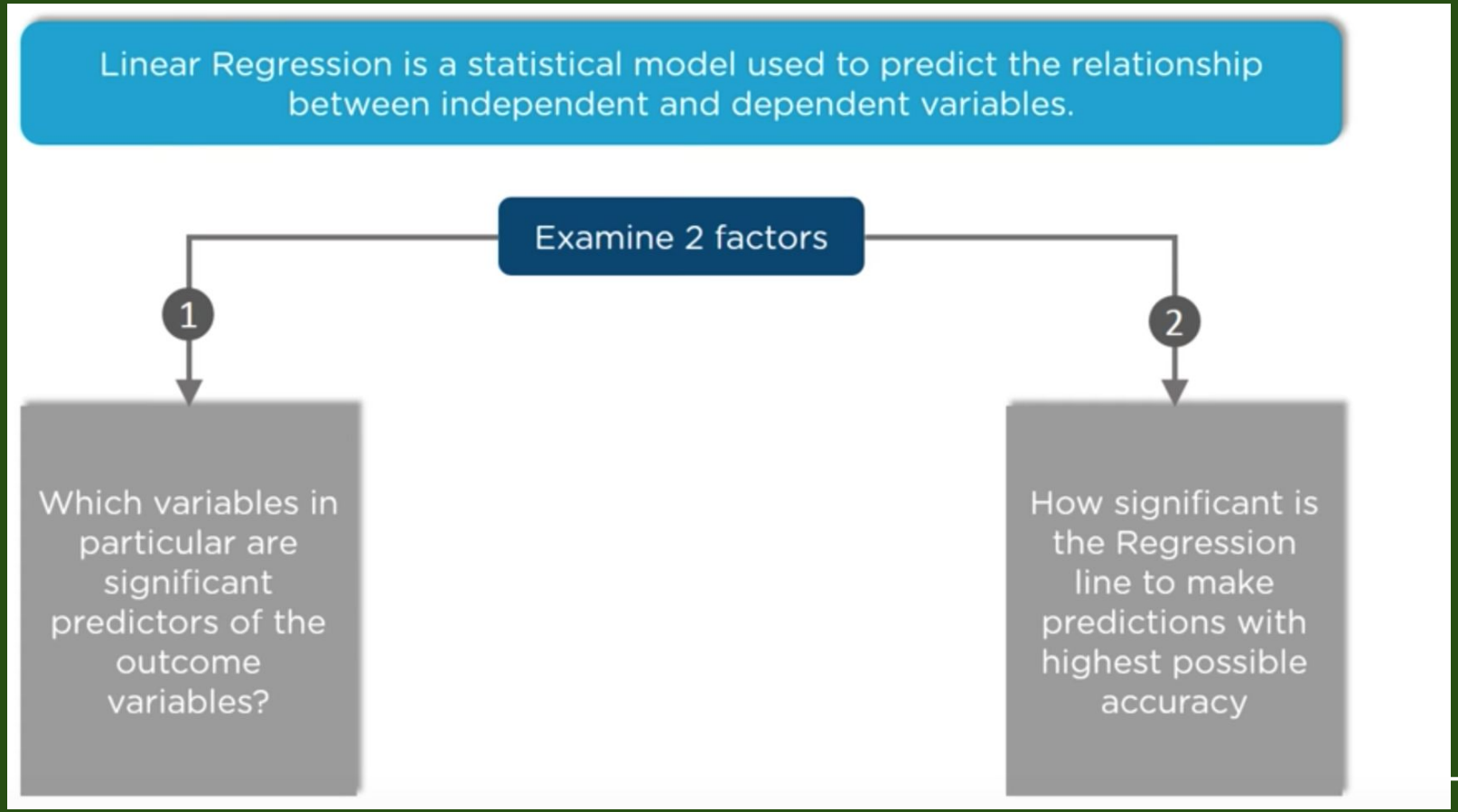
Score Prediction

To predict the number of runs a player would score in the coming matches based on previous performance

Housing sales

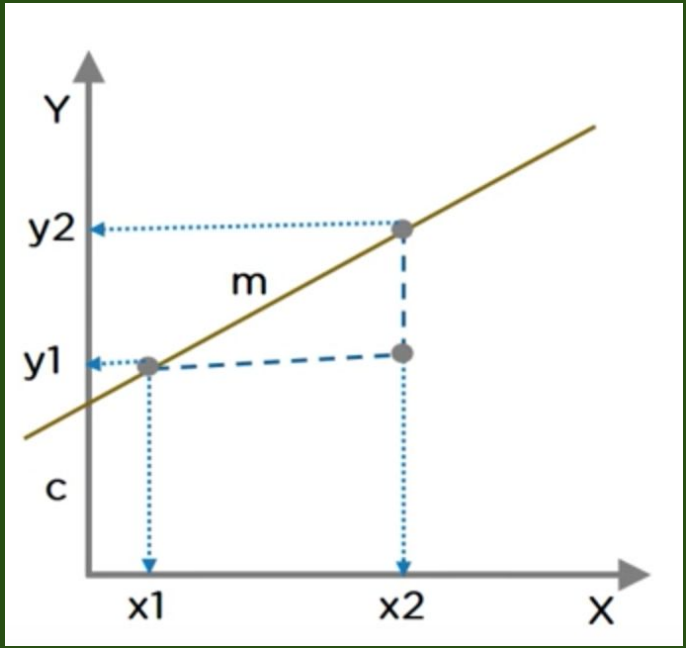
To estimate the number of houses a builder would sell and at what price in the coming months

Regresion Lineal



Regresion Lineal

$$y = m*x + c$$



y ---> Dependent Variable

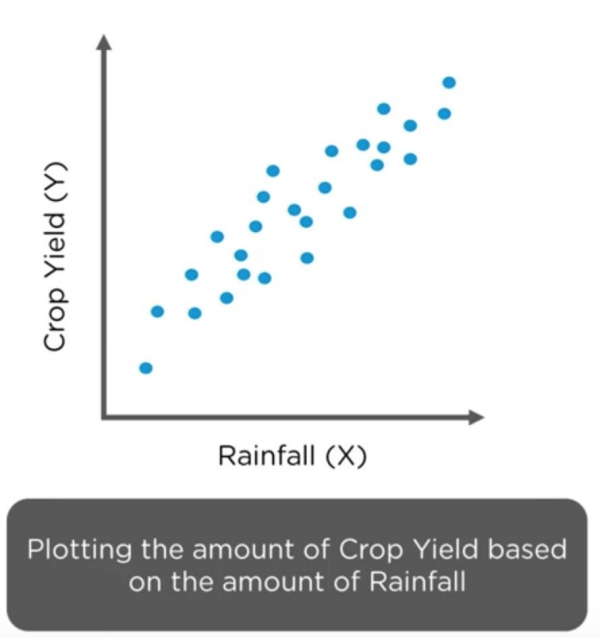
x ---> Independent Variable

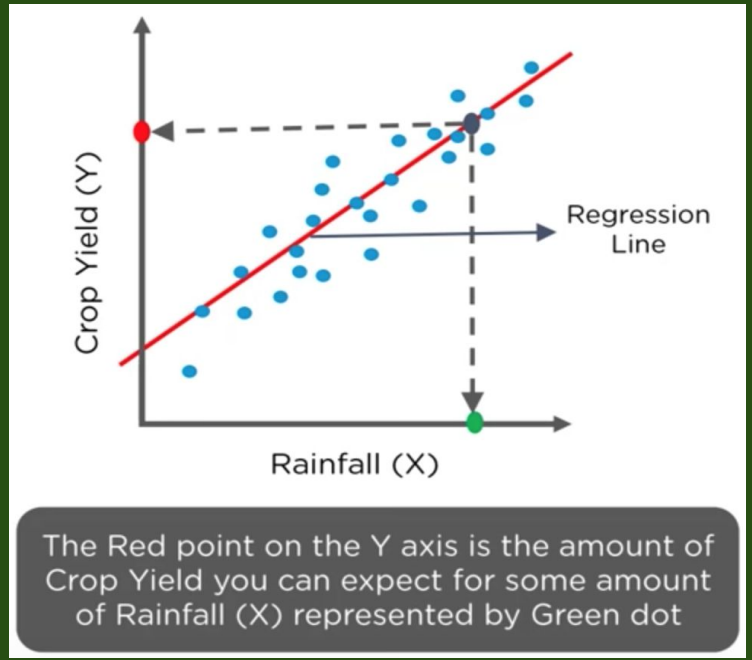
m ---> Slope of the line

$$m = \frac{y2 - y1}{x2 - x1}$$

c ---> Coefficient of the line

Predicción usando: Regresion Lineal

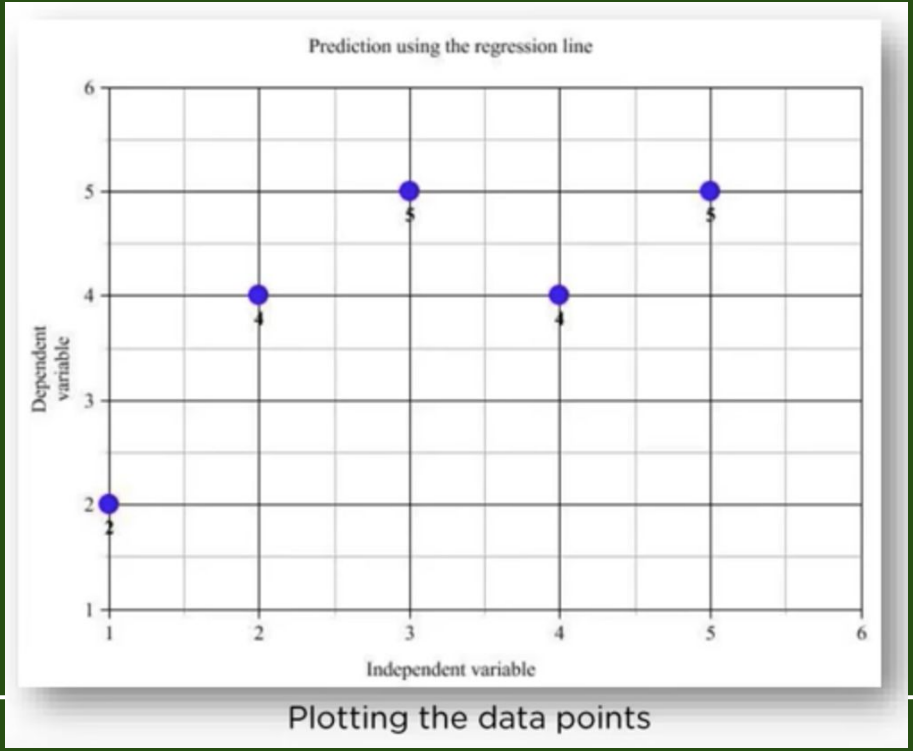




Intuición detras de Regresion Lineal

Lets consider a sample dataset with 5 rows and find out how to draw the regression line

Independent variable	Dependent variable
X	Y
1	2
2	4
3	5
4	4
5	5

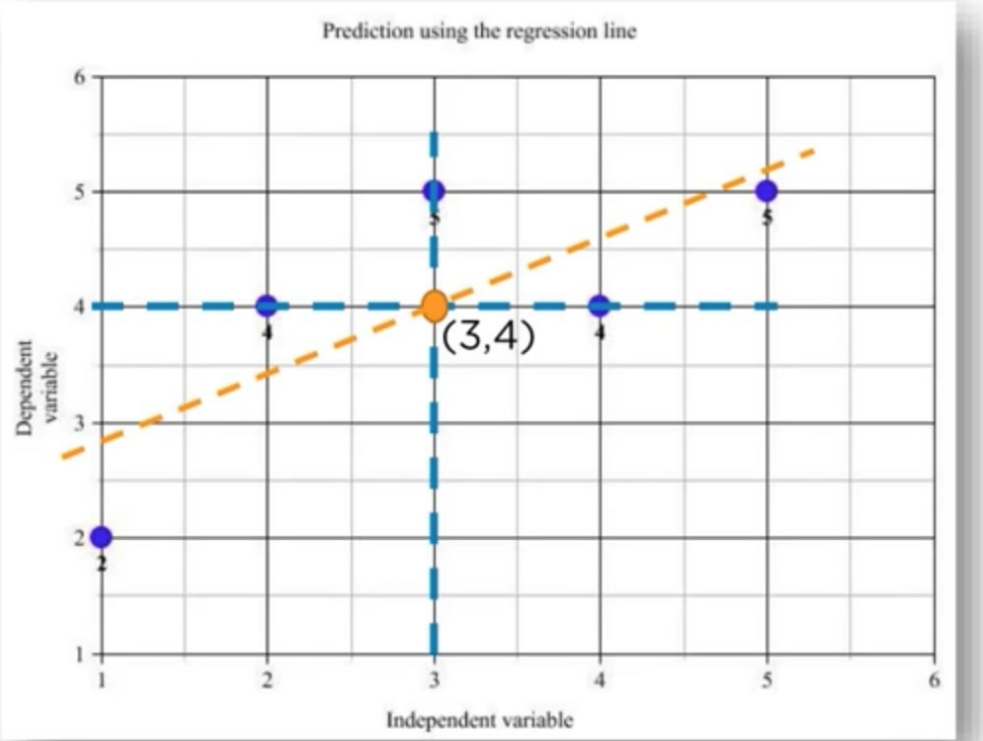


Plotting the data points

Intuición detras de Regresion Lineal

Independent variable	Dependent variable
X	Y
1	2
2	4
3	5
4	4
5	5

Mean 3 4



Regression line

Regression line should ideally pass through the mean of X and Y

Intuición detras de Regresion Lineal

Drawing the equation of the Regression line

X	Y	(X^2)	(Y^2)	(X∗Y)
1	2	1	4	2
2	4	4	16	8
3	5	9	25	15
4	4	16	16	16
5	5	25	25	25
$\sum = 15$	$\sum = 20$	$\sum = 55$	$\sum = 86$	$\sum = 66$

$$Y = m * X + c$$
$$= 0.6 * 3 + 2.2$$
$$= 4$$

Linear equation is represented as $Y = m * X + c$

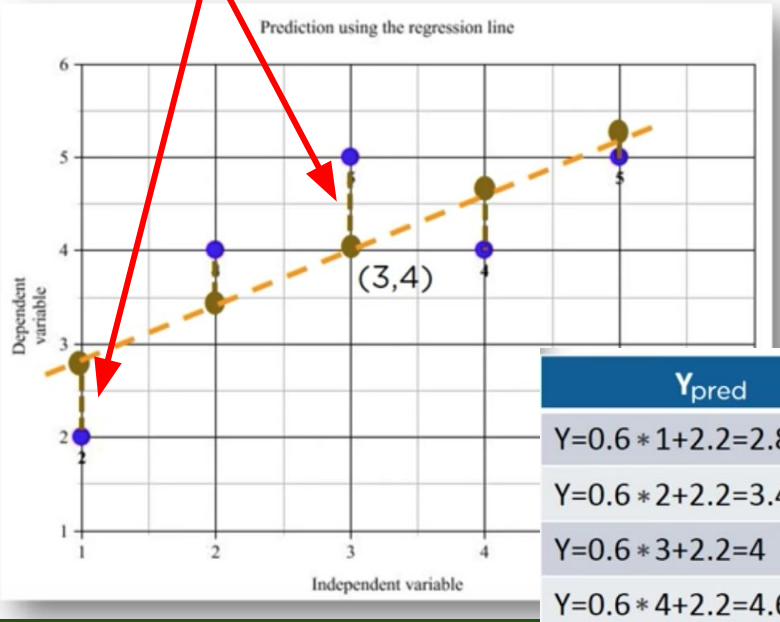
$$m = \frac{((n * \sum(X*Y)) - (\sum(X) * \sum(Y))}{((n * \sum(X^2)) - (\sum(X)^2)} = \frac{((5*66) - (15*20))}{((5*55)) - (225)} = 0.6$$

$$c = \frac{((\sum(Y) * \sum(X^2)) - (\sum(X) * \sum(X*Y)}{((n * \sum(X^2)) - (\sum(X)^2)} = 2.2$$

Intuición detras de Regresion Lineal

Lets find out the predicted values of Y for corresponding values of X using the linear equation where m=0.6 and c=2.2

residual or errors



Y_{pred}
$Y=0.6*1+2.2=2.8$
$Y=0.6*2+2.2=3.4$
$Y=0.6*3+2.2=4$
$Y=0.6*4+2.2=4.6$
$Y=0.6*5+2.2=5.2$

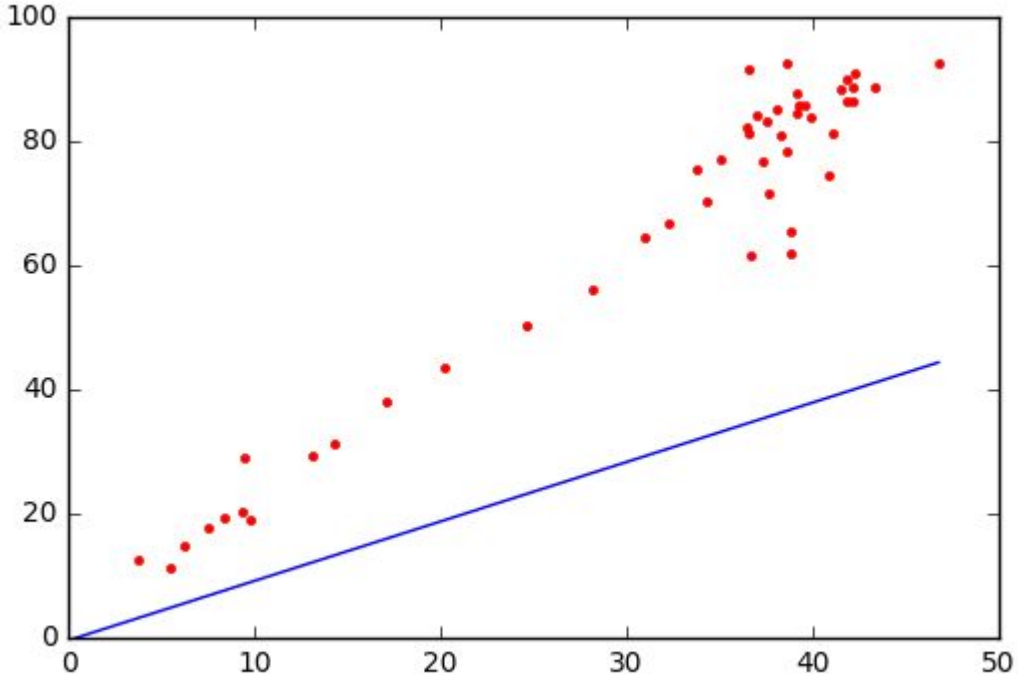
X	Y	Y_{pred}	$(Y-Y_{pred})$	$(Y-Y_{pred})^2$
1	2	2.8	-0.8	0.64
2	4	3.4	0.6	0.36
3	5	4	1	1
4	4	4.6	-0.6	0.36
5	5	5.2	-0.2	0.04

$\sum = 2.4$

The sum of squared errors for this regression line is 2.4. We check this error for each line and conclude the best fit line having the least e square value.

Intuición detras de Regresion Lineal

Minimizing the Distance: There are lots of ways to minimize the distance between the line and the data points like Sum of Squared errors, Sum of Absolute errors, Root Mean Square error etc.



ving this line through
ints to make sure the
has the least square
tween the data points
e regression line

References

- Python Programming: An Introduction to Computer Science. John Zelle
- Big Data con Python. Rafael Caballero Enrique Martín Adrián Riesco
- Aprende Python en un Fin de Semana Alfredo Moreno Muñoz Sheila Córcoles Córcoles
- Learn Python Programming Fabrizio Romano
- Python Data Analytics Fabio Nelli
- Expert Python Programming Michael Jasworski Tarek Ziadé
- Statistical analysis of questionnaires: a unified approach based on R and Stata by Francesco Bartolucci. Boca Raton: CRC Press, 2016.
- Data visualisation: a handbook for data driven design by Andy Kirk. Los Angeles: Sage, 2016.
- Learning tableau: leverage the power of tableau 9.0 to design rich data visualizations and build fully interactive dashboards by Joshua N. Milligan. Mumbai: Packt Publishing, 2015.