

Machine Learning for Finance

Caleb Terrel Orellana

Líder en Analítica Avanzada

Sección 1: Presentación del curso

Presentación del docente

Caleb Terrel Orellana

Líder de Consultoría en Advanced Analytics



Analytics Solutions
Crear nombre de usuario · Asesoría



- Experiencia:
 - +10 años consultor estadístico y capacitador
 - +6 años consultor en proyectos/pilotos de advanced analytics/data mining/machine learning/predictive modeling, y preventa
 - Capacitador en diversas empresas
 - Metodología CRISP-DM
 - Softwares: *IBM SPSS*
 - Lenguajes: *R* y *Python*



Contenido del curso

- **Sección 1:** Presentación del curso

- Presentación del docente
- Objetivos y requisitos del curso
- Configuración de entorno de trabajo (Anaconda)
- Detalles de los casos a desarrollar

- **Sección 2:** Introducción a Machine learning

- Conceptos y enfoque del Machine learning
- Tipos de aprendizaje del Machine learning
- Proceso de un proyecto de Machine learning

- **Sección 3:** Técnicas no supervisadas

- Análisis de componentes principales
- Segmentación k-means
- **Caso 1:** Segmentación de agencias financieras

- **Sección 4:** Técnicas supervisadas

- Modelos lineales: Regresión lineal
- Modelos lineales regularizados: ridge, lasso, Modelos no lineales: arbol de decisión, random forest
- **Caso 2:** Predicción de ingresos de agencias financieras

Objetivos del curso



- Comprender los conceptos, enfoque, tipos de aprendizaje y procesos de un proyecto de Machine learning
- Comprender sobre técnicas no supervisadas
- Comprender sobre técnicas supervisadas
- Aplicar lo aprendido en casos reales de Microfinancieras desarrollando: Análisis exploratorio de datos, modelamiento e interpretación de resultados y criterios.
- Comprender el lenguaje Python en el desarrollo de los casos reales.

Requisitos del curso (no indispensable)

- Hayan aprobado cursos como:



Configuración de entorno de trabajo (Anaconda)

<https://www.anaconda.com/products/individual>



Individual Edition

Your data science toolkit

With over 20 million users worldwide, the open-source Individual Edition (Distribution) is the easiest way to perform Python/R data science and machine learning on a single machine. Developed for solo practitioners, it is the toolkit that equips you to work with thousands of open-source packages and libraries.

Download

Anaconda Installers

Windows 

Python 3.8

64-Bit Graphical Installer (466 MB)


32-Bit Graphical Installer (397 MB)

MacOS 

Python 3.8

64-Bit Graphical Installer (462 MB)

64-Bit Command Line Installer (454 MB)

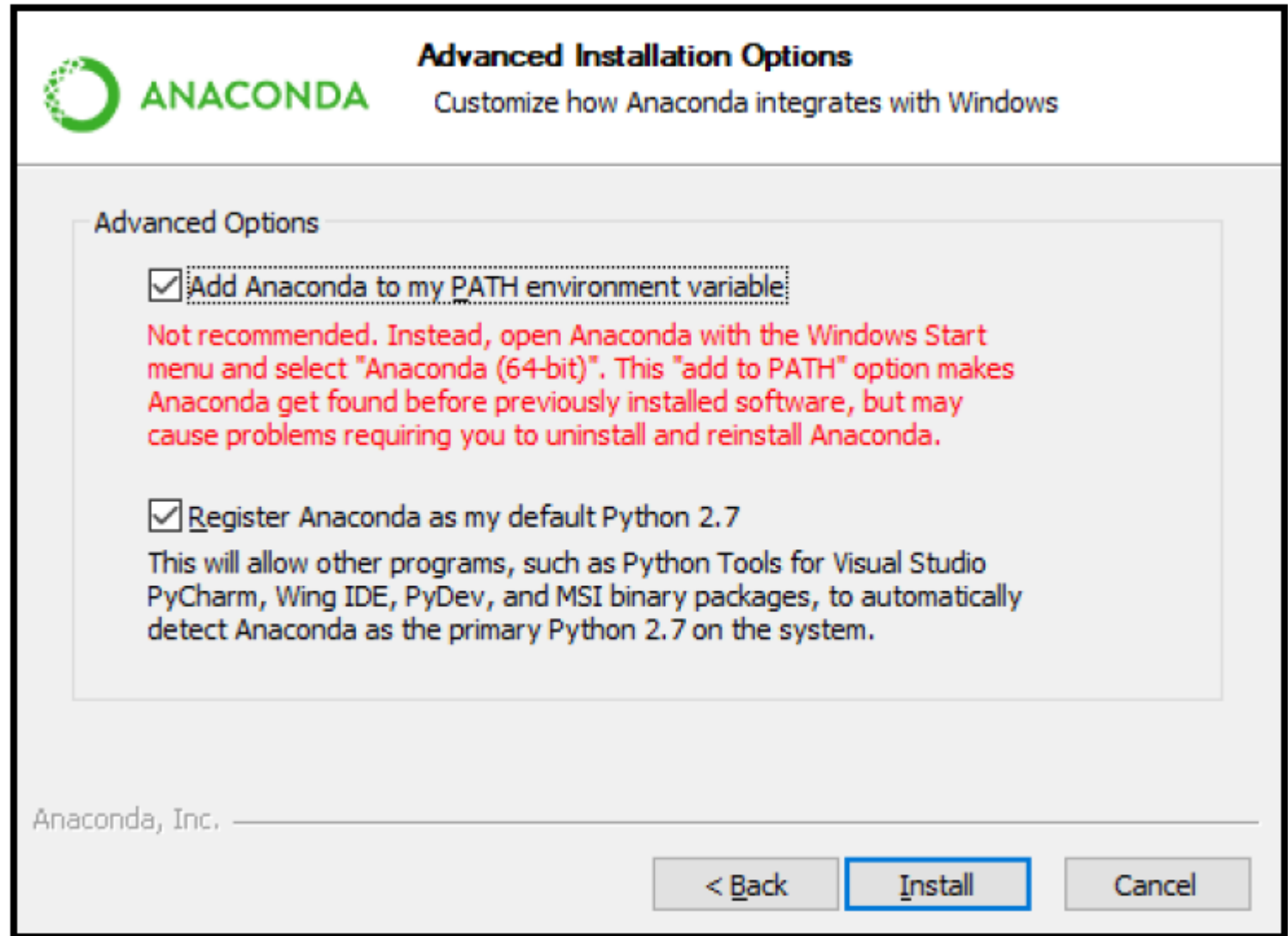
Linux 

Python 3.8

64-Bit (x86) Installer (550 MB)

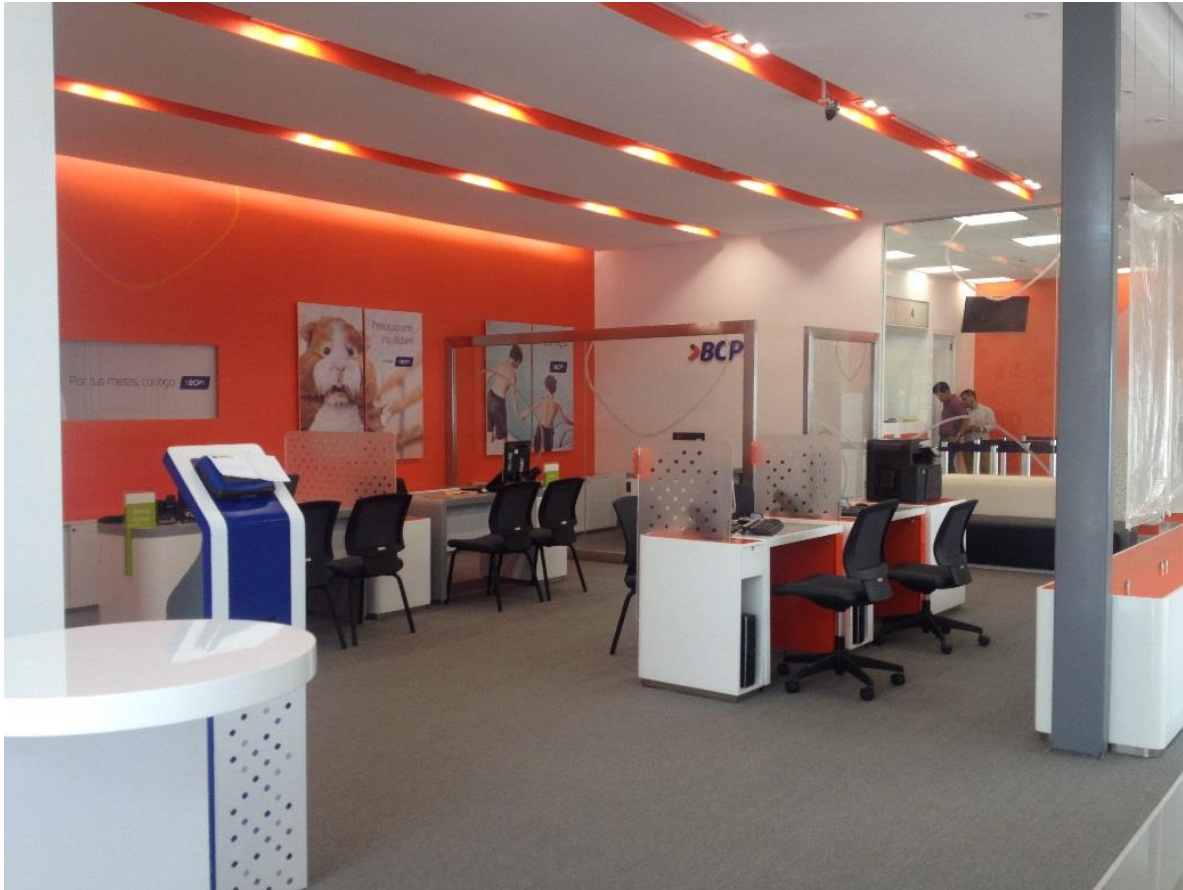
64-Bit (Power8 and Power9) Installer (290 MB)

Configuración de entorno de trabajo (Anaconda)



Detalles de los casos a desarrollar

- Caso 1: segmentación de agencias



Cluster 1

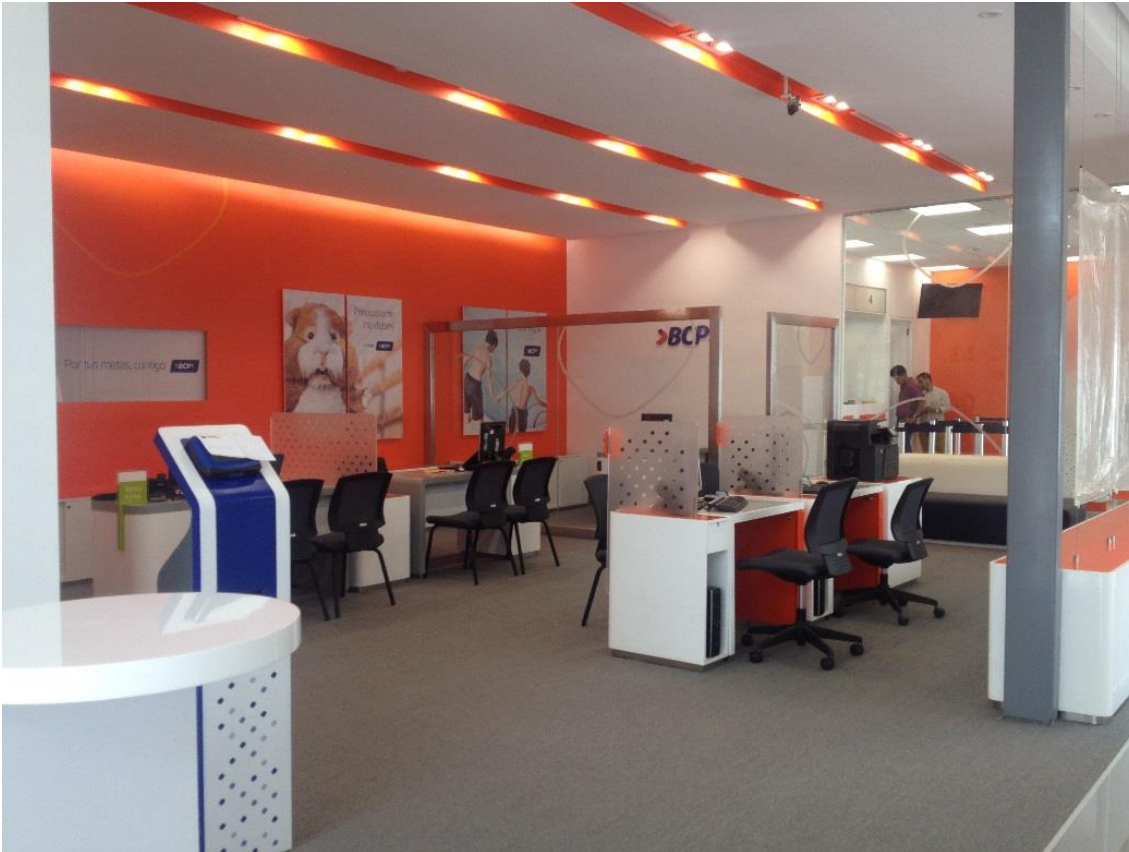
Cluster 2

...

Cluster n

Detalles de los casos a desarrollar

- Caso 2: predicción de ingresos de agencias

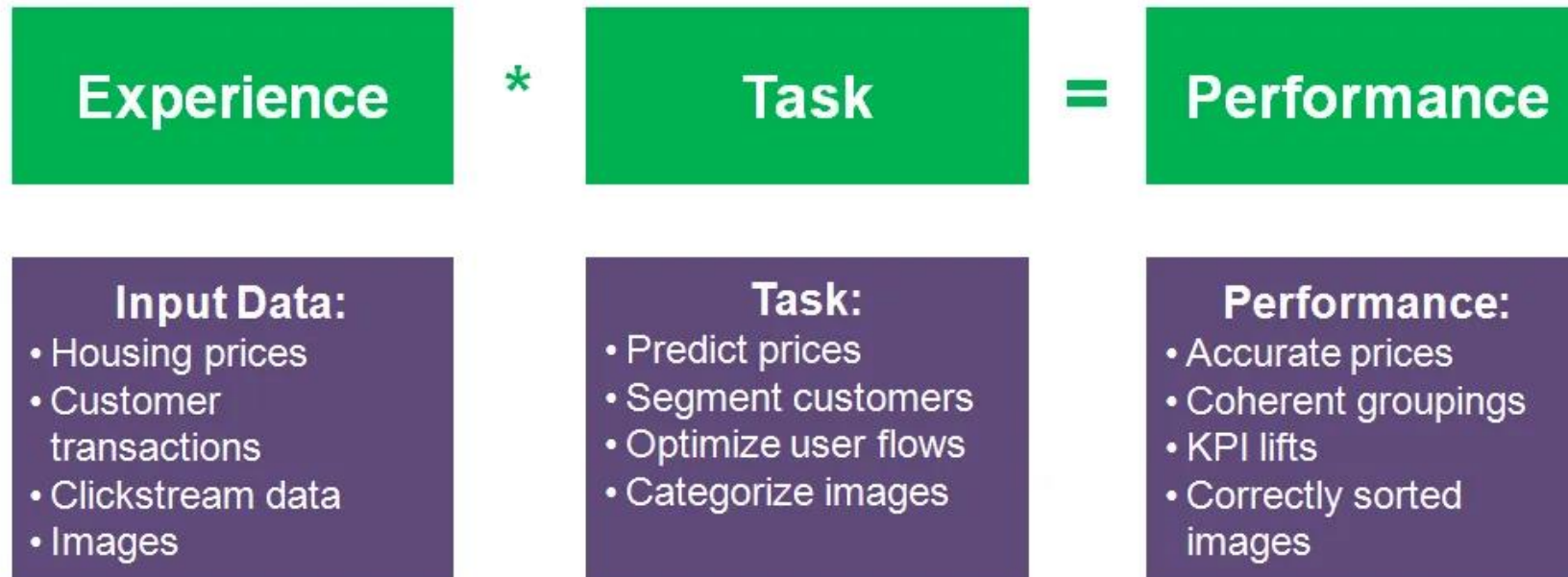


$$\text{Ingresos} = f(v1, v2, v3, \dots, vp)$$

Sección 2: Introducción a Machine learning

Conceptos y enfoque del Machine learning

$$E * T = P$$

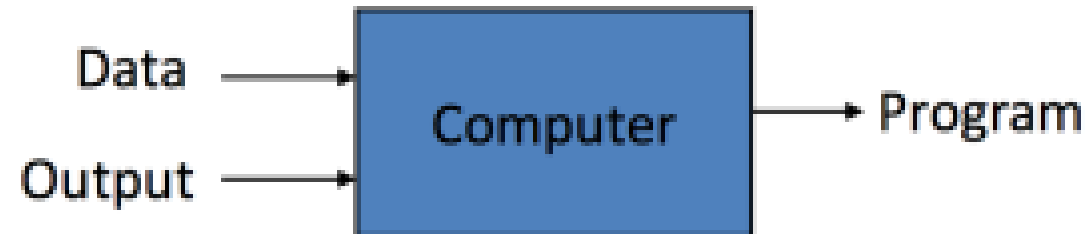


Conceptos y enfoque del Machine learning

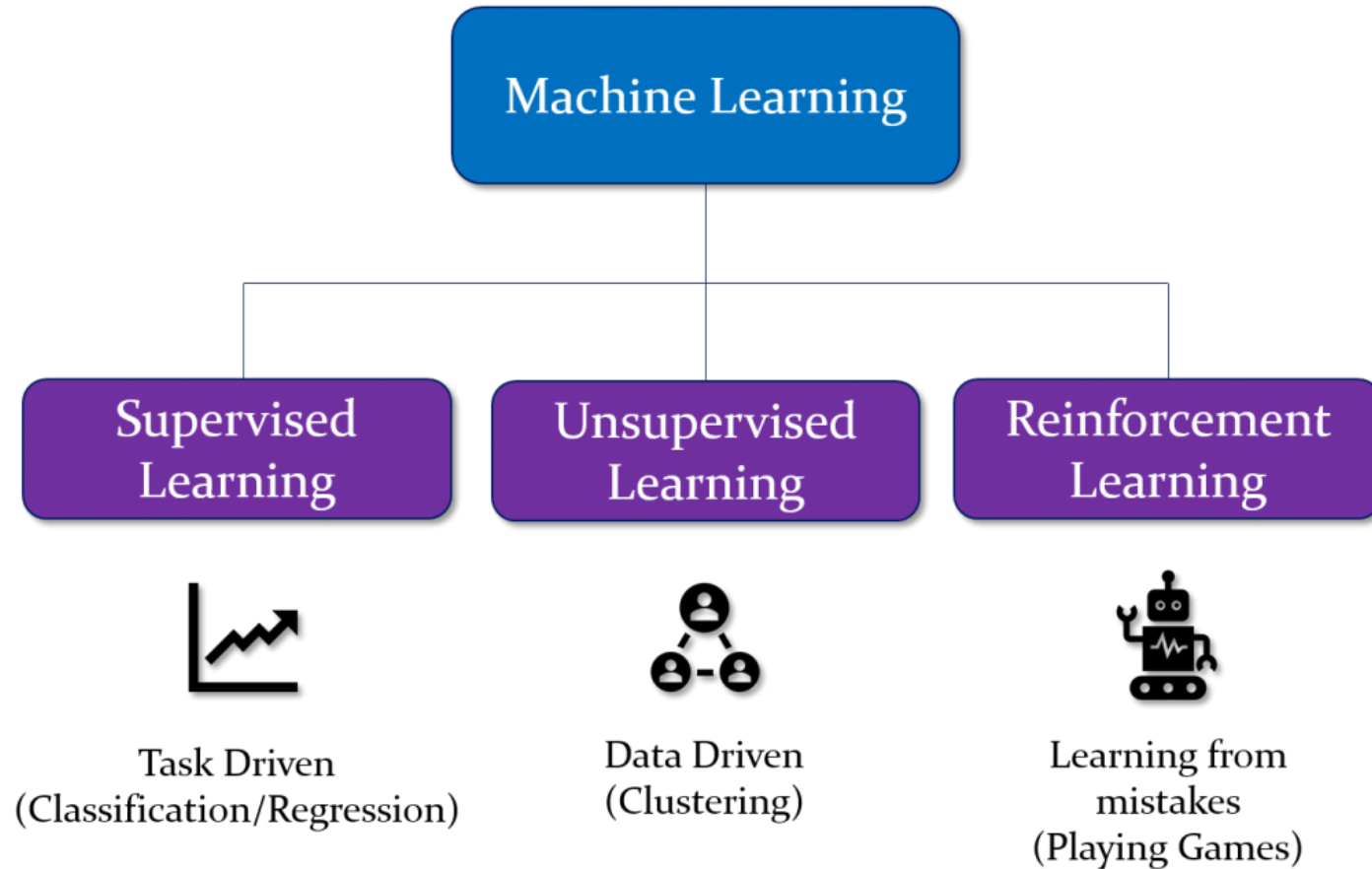
Traditional Programming



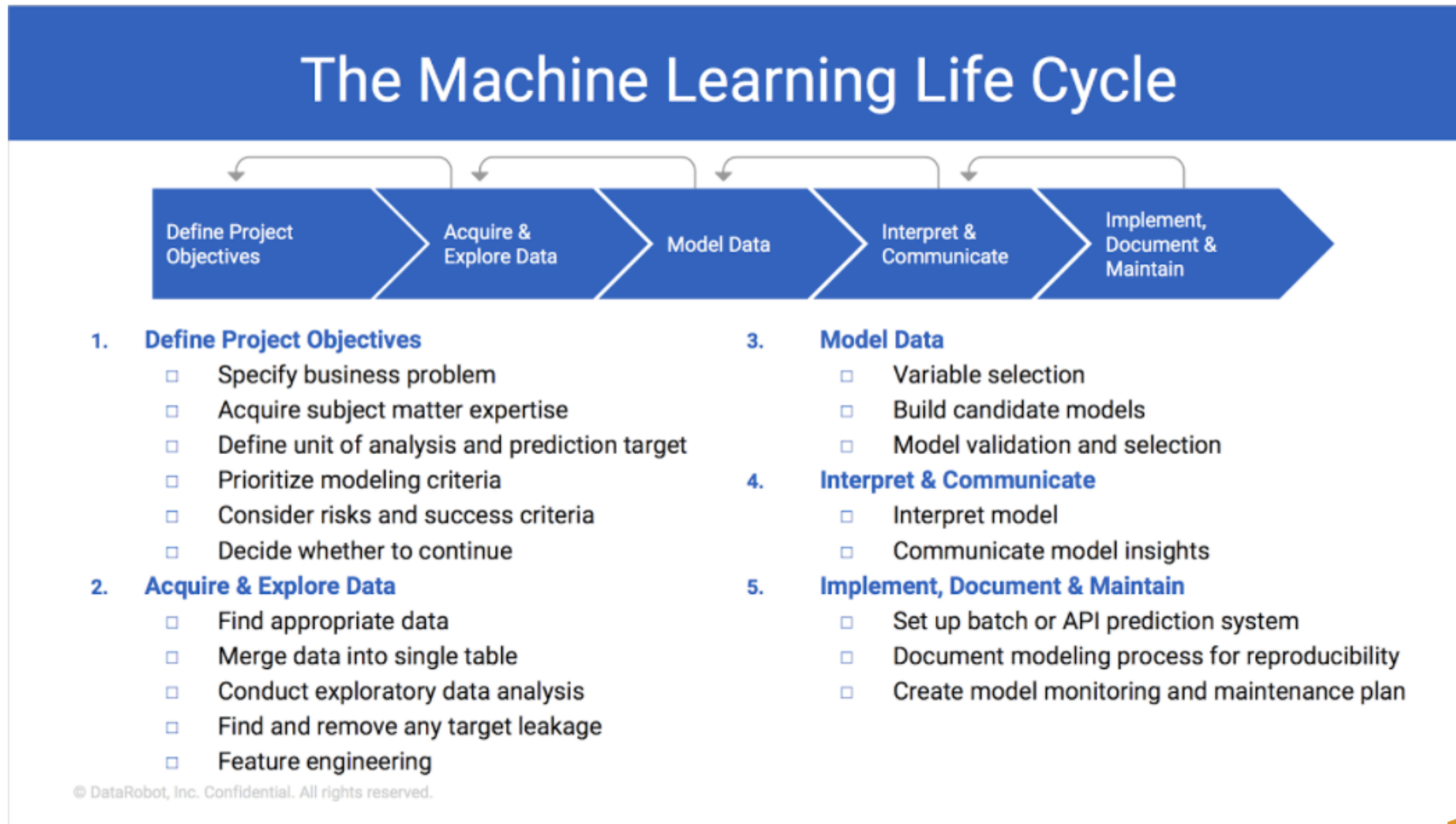
Machine Learning



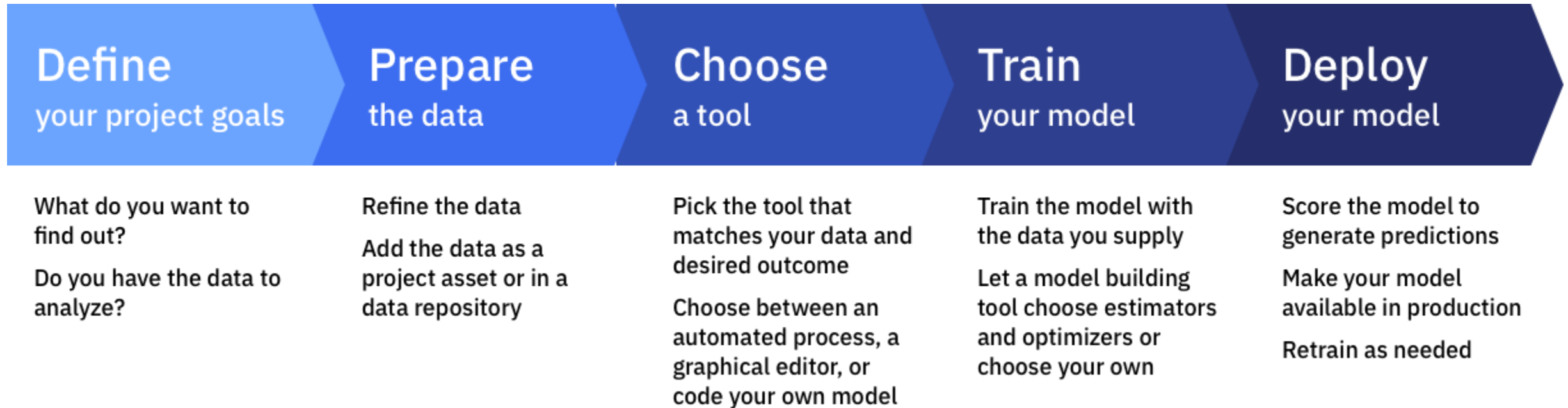
Tipos de aprendizaje del Machine learning



Proceso de un proyecto de Machine learning



Proceso de un proyecto de Machine learning

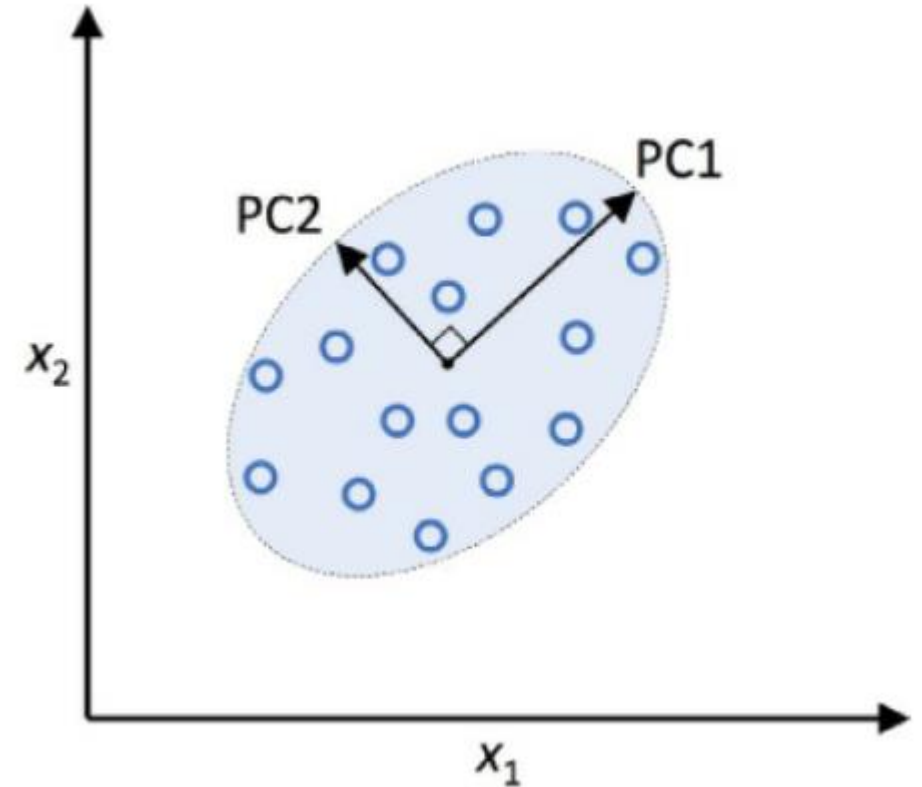


Sección 3: Técnicas no supervisadas

Reducción de dimensionalidad

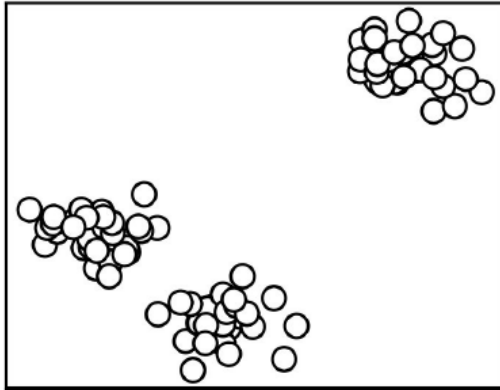
Análisis de Componentes Principales

- técnica de transformación lineal para extracción de características y reducción de dimensionalidad.
- ayuda a identificar:
 - patrones en datos basados en la correlación entre características
 - las direcciones de varianza máxima en datos de alta dimensión y las proyecta en un nuevo subespacio con dimensiones iguales o menores que el original.

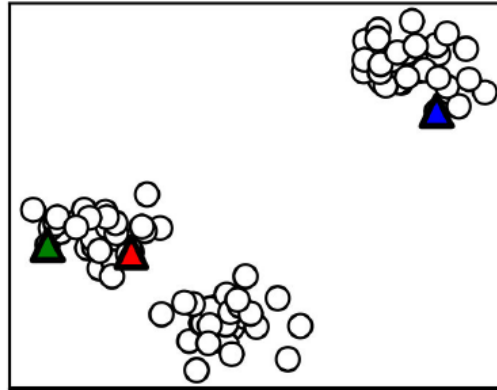


Clustering k-means

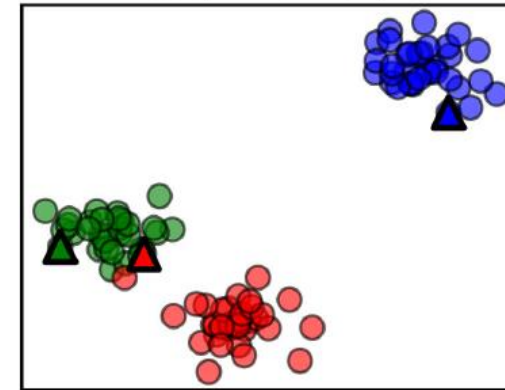
Input data



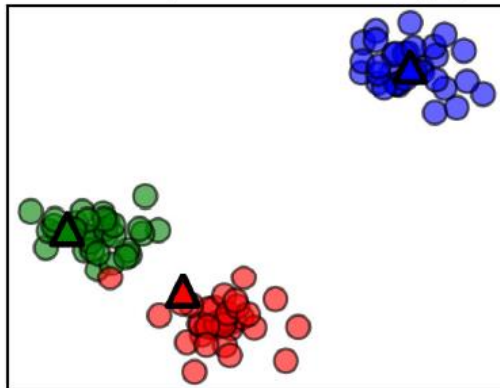
Initialization



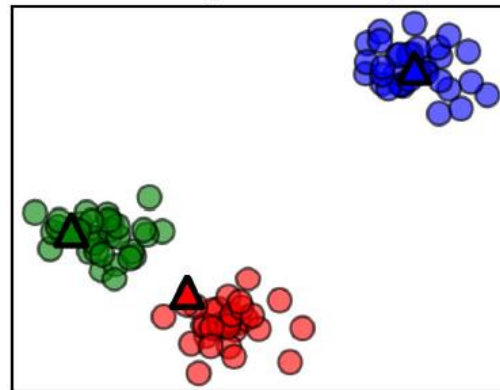
Assign Points (1)



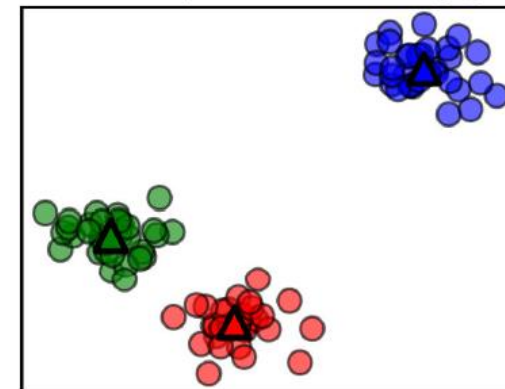
Recompute Centers (1)



Reassign Points (2)



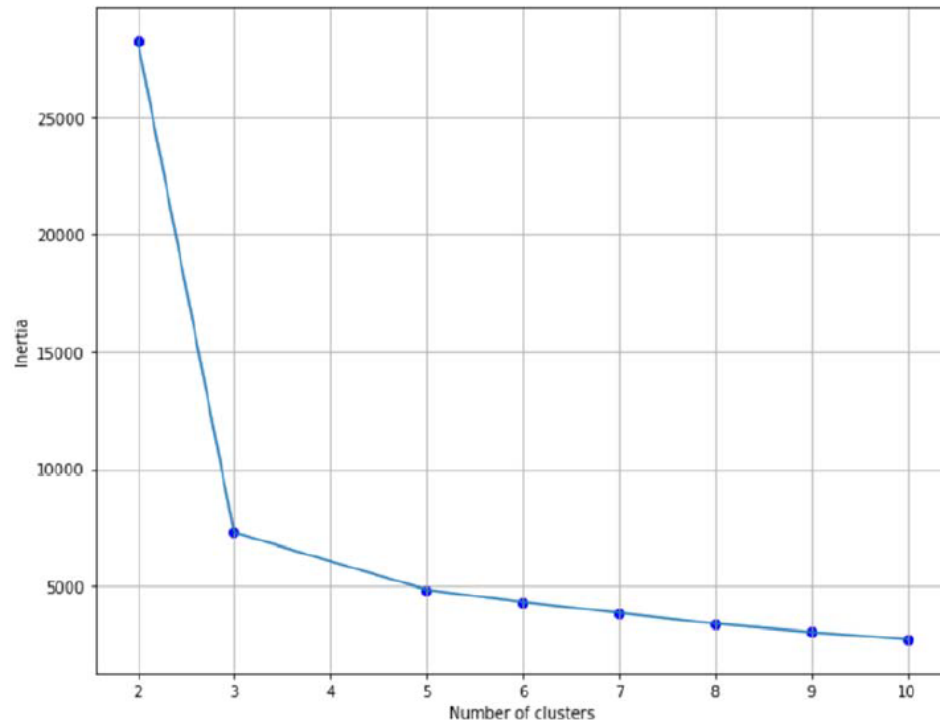
Recompute Centers (2)



Encontrar el número óptimo de clusters

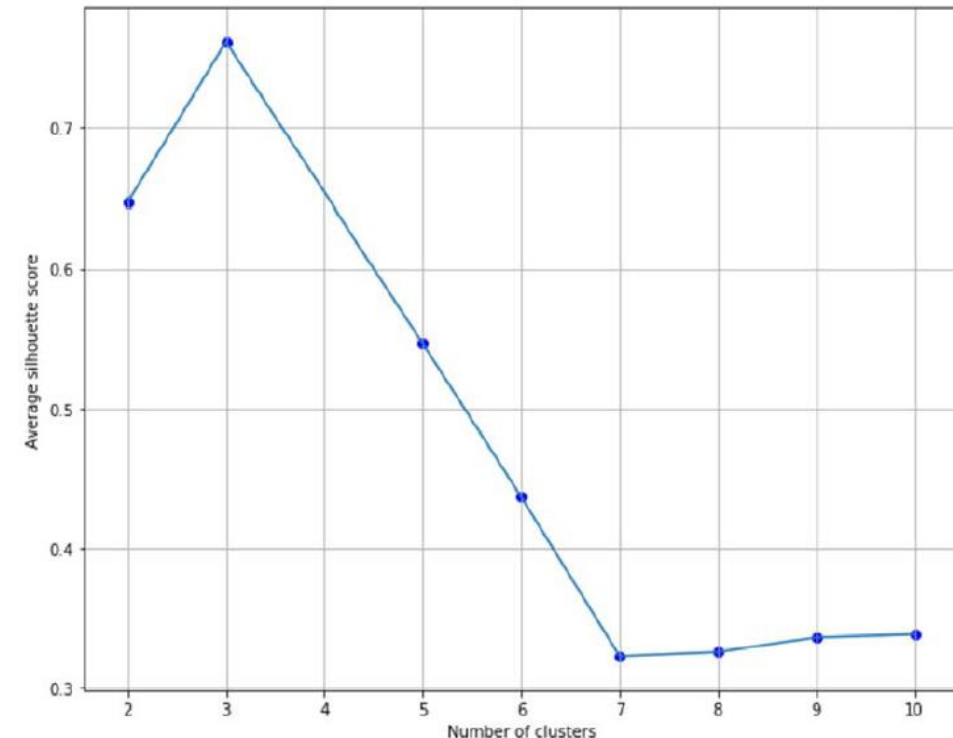
Inercia

- La inercia es la suma de todas las diferencias entre cada miembro del clúster y su centroide.
- Un número apropiado de agrupaciones debe producir una pequeña inercia.

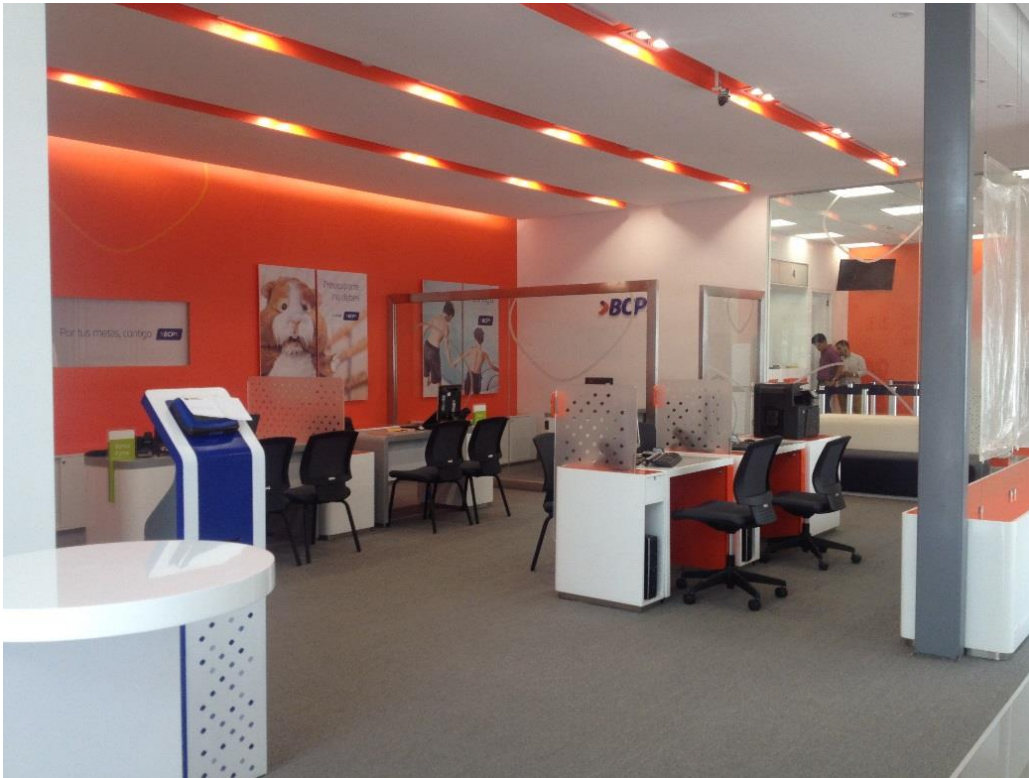


Silueta

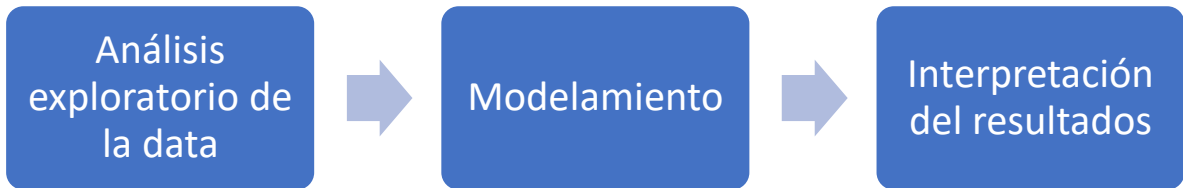
- "máxima cohesión interna y máxima separación de grupos".
- Un valor cercano a 1 es bueno
- Un valor cercano a 0 significa que la diferencia entre las medidas intra e inter conglomerados es casi nula y, por lo tanto, hay una superposición de clústeres.
- Un valor cercano a -1 significa que la muestra se ha asignado a un grupo incorrecto



Caso 1: Segmentación de agencias financieras



Procesos analíticos utilizando Python

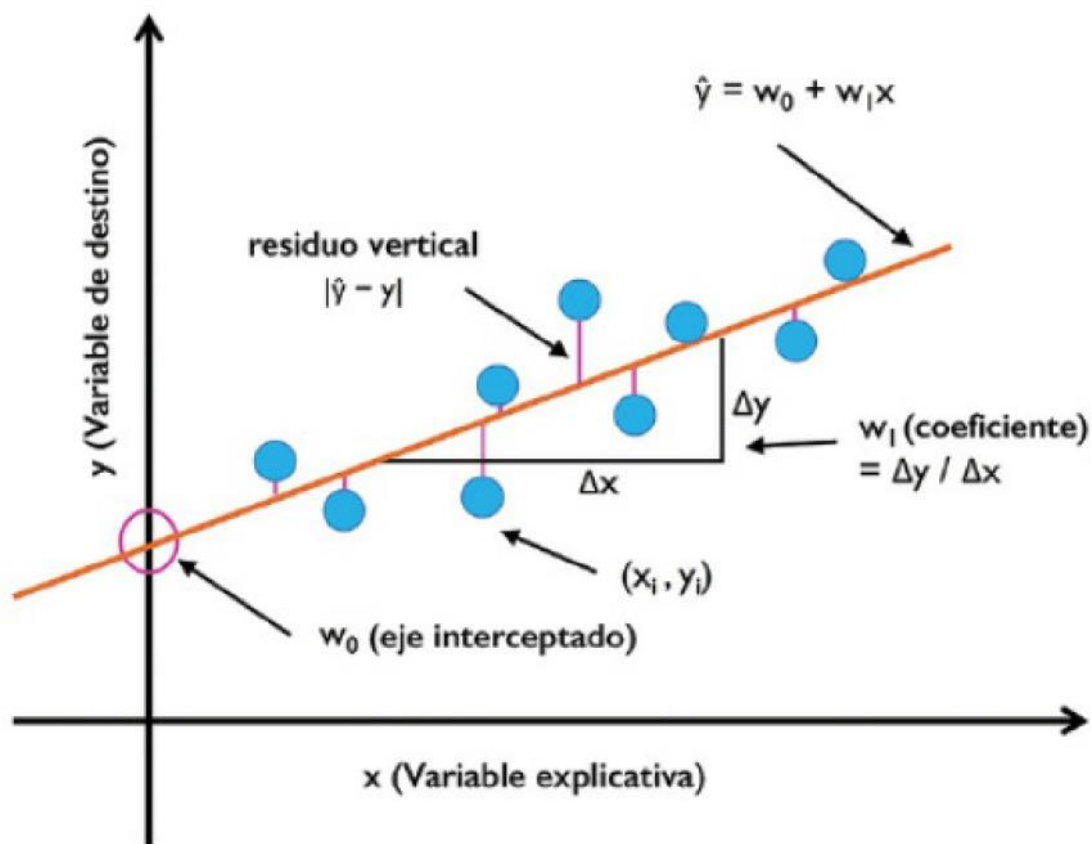


Sección 4: Técnicas supervisadas

Modelos lineales: Regresión lineal

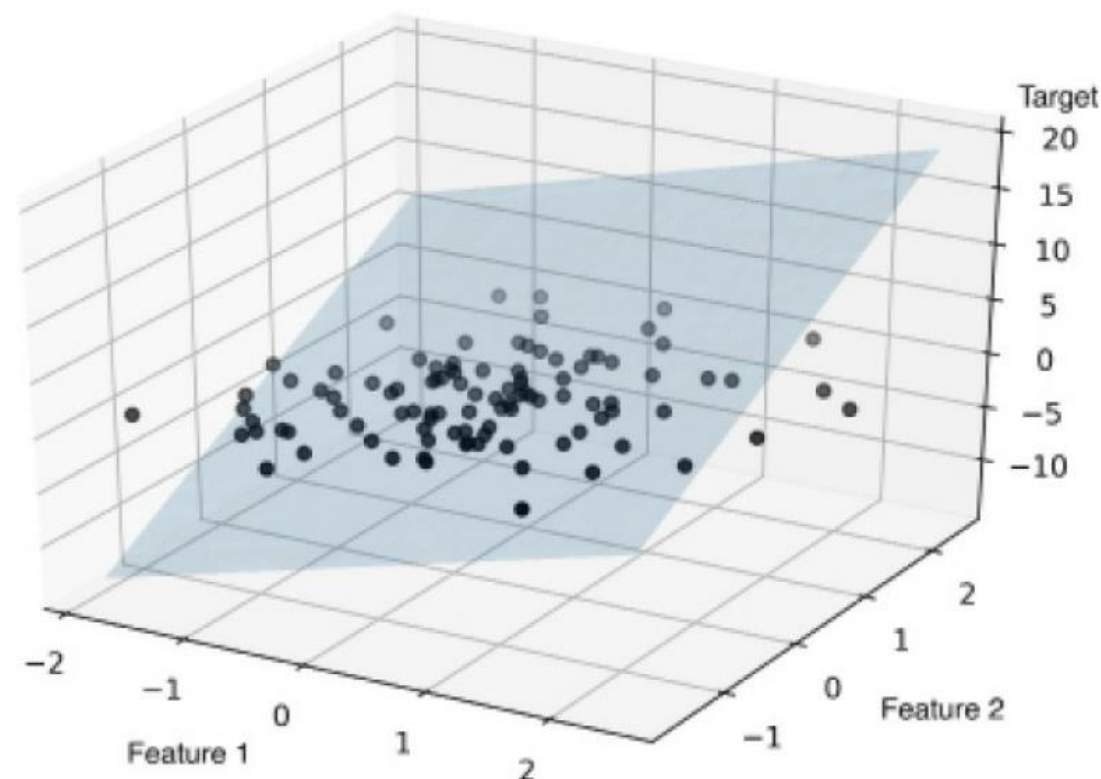
Regresión lineal simple

$$Y = \beta_0 + \beta_1 X$$



Regresión lineal múltiple

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$



Modelos lineales: Regresión lineal

- **Estimación de parámetros:** Mínimos Cuadrados Ordinarios (OLS)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad \text{and} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \frac{\left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n x_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}$$

$$S_{xy} = \sum_{i=1}^n y_i x_i - \frac{\left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n x_i\right)}{n} = \sum_{i=1}^n y_i (x_i - \bar{x})$$

$$S_{xx} = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n} = \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$e_i = y_i - \hat{y}_i$$

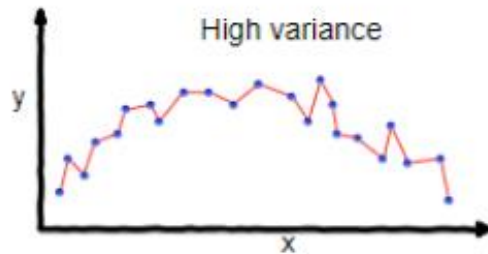
Modelos lineales: Regresión lineal

Diagnostico del modelo

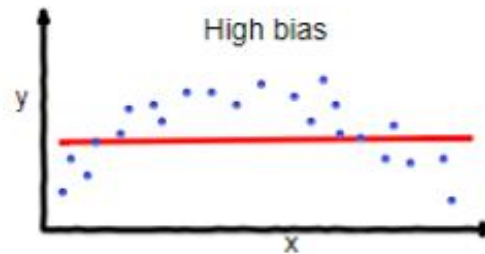
- Linealidad del modelo
- Significancia de parametros (betas) del modelo
- Los residuos del modelo se distribuyen normalmente
- Los residuos del modelo son homocedasticos
- Multicolinealidad

Modelos lineales regularizados – Intro.

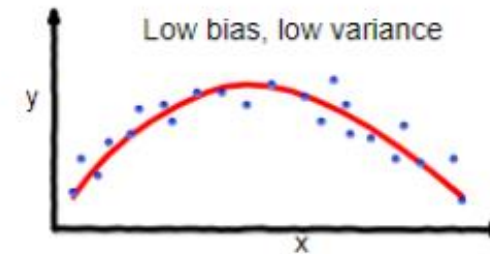
- La regularización es un método para abordar el problema del sobreajuste añadiendo información adicional y, por tanto, reduciendo los valores del parámetro del modelo para inducir una penalización contra la complejidad.



overfitting



underfitting



Good balance

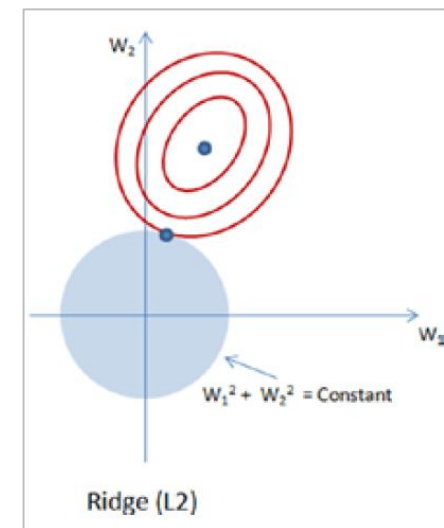
Modelos lineales regularizados – Ridge y Lasso

Ridge

- Es un modelo L2 penalizado donde simplemente añadimos la suma cuadrática de los pesos a nuestra función de coste de mínimos cuadrados.
- Aumentando el valor del hiperparámetro λ , aumentamos la fuerza de regularización y disminuimos los pesos de nuestro modelo.

$$J(w)_{\text{Ridge}} = \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2 + \lambda \|w\|_2^2$$

$$L2: \lambda \|w\|_2^2 = \lambda \sum_{j=1}^m w_j^2$$

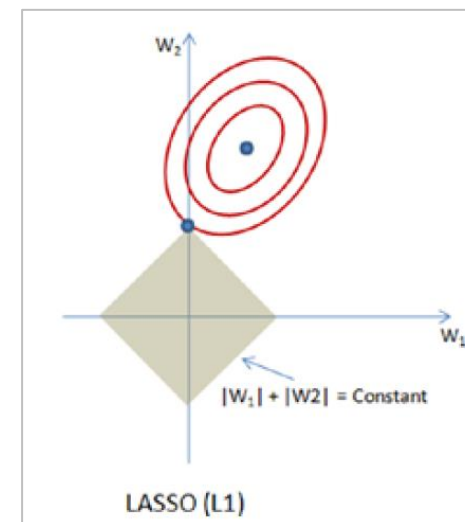


LASSO (Least Absolute Shrinkage and Selection Operator)

- Según la fuerza de regularización, determinados pesos pueden convertirse en cero, situación que hace que LASSO también sea útil como técnica de selección de características supervisada.

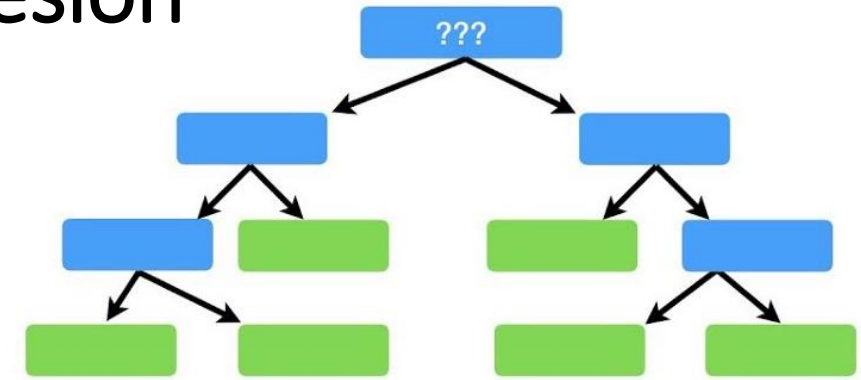
$$J(w)_{\text{LASSO}} = \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2 + \lambda \|w\|_1$$

$$L1: \lambda \|w\|_1 = \lambda \sum_{j=1}^m |w_j|$$



Modelos no lineales: Árbol de Regresión

- Dividen los datos en grupos más pequeños que son más homogéneos con respecto a la respuesta.
- Los árboles de regresión determinan:
 - El predictor para dividir y el valor de la división
 - La profundidad o complejidad del árbol.
 - La ecuación de predicción en los nodos terminales.



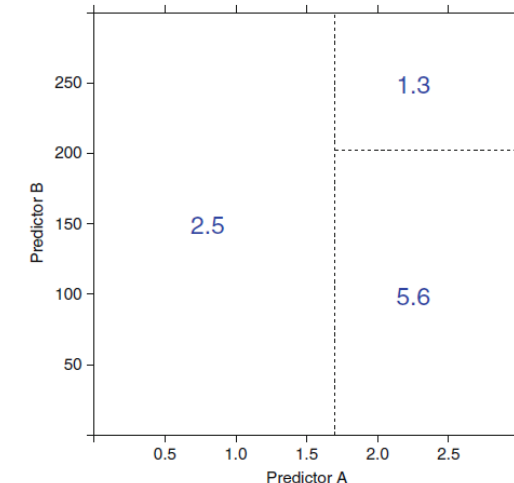
CART

- El modelo comienza con el conjunto de datos completo, S , y busca cada valor distinto de cada predictor para encontrar el predictor y el valor dividido que divide los datos en dos grupos (S_1 y S_2) de manera que las sumas generales de errores cuadrados son minimizado:

$$SSE = \sum_{i \in S_1} (y_i - \bar{y}_1)^2 + \sum_{i \in S_2} (y_i - \bar{y}_2)^2$$

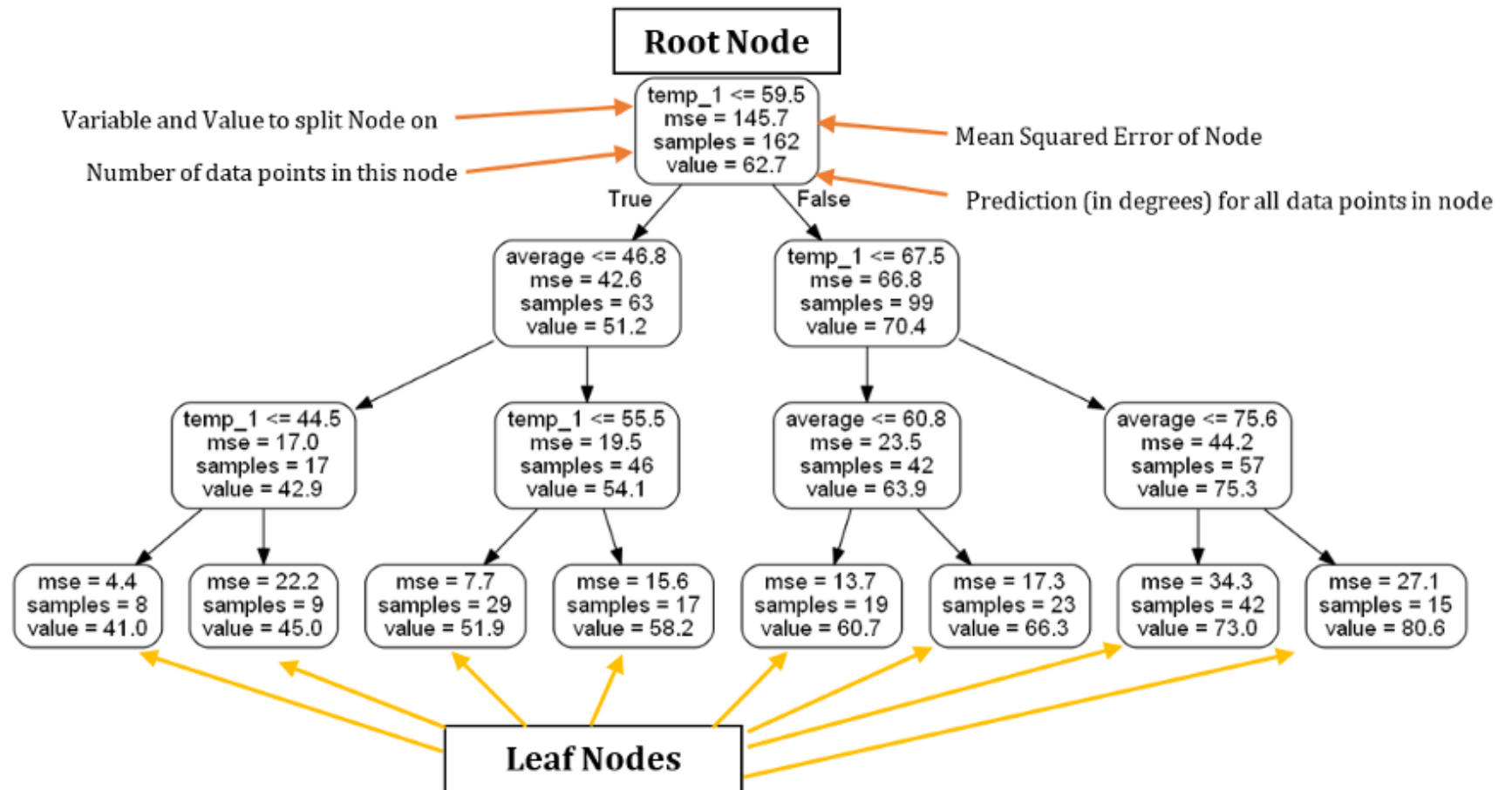
- donde \bar{y}_1 e \bar{y}_2 son los promedios de los resultados del conjunto de entrenamiento dentro de los grupos S_1 y S_2 , respectivamente. Luego, dentro de cada uno de los grupos S_1 y S_2 , este método busca el predictor y el valor de división que mejor reduce la SSE.

```
if Predictor A >= 1.7 then
|   if Predictor B >= 202.1 then Outcome = 1.3
|   else Outcome = 5.6
else Outcome = 2.5
```



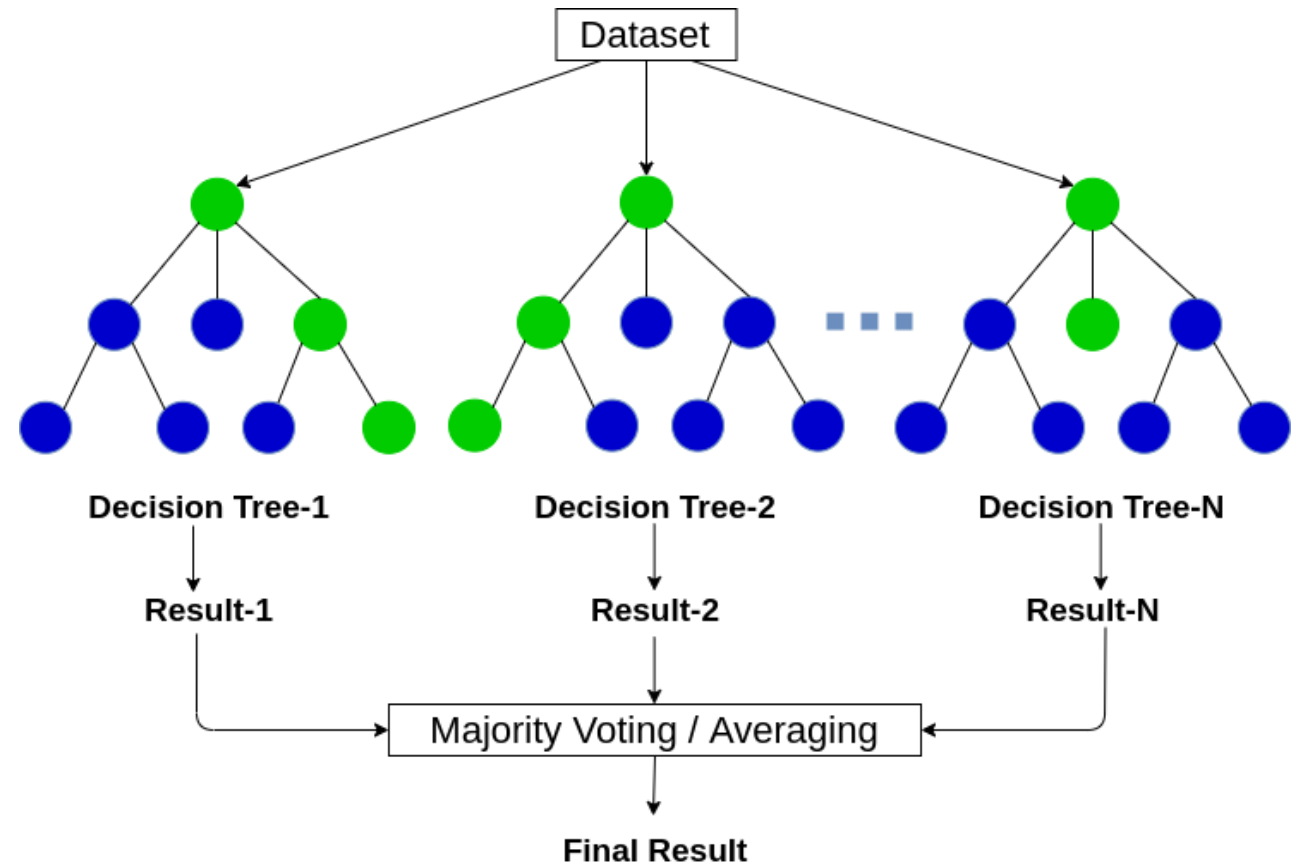
Modelos no lineales: Árbol de Regresión

- Visualización de árbol

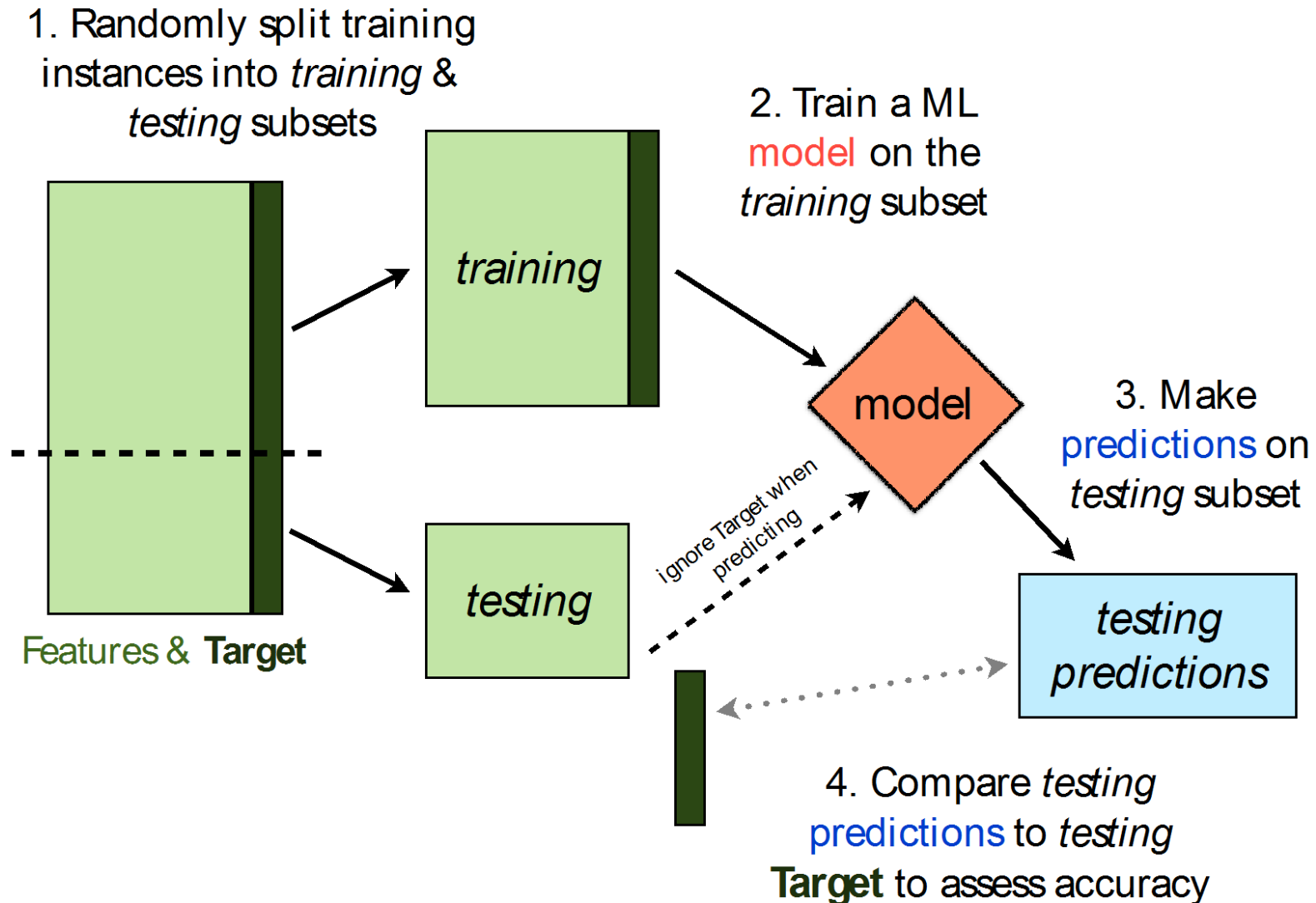


Modelos no lineales: Bosque aleatorio de Regresión

- Construye un conjunto de arboles que predican conjuntamente los datos. Cada árbol se optimiza para ajustarse sólo a algunas de las observaciones utilizando sólo algunos de los predictores.
- Es la variación en el *Bagging* de los Árboles de Decisión, reduciendo los atributos disponibles para hacer un árbol en cada punto de decisión a una sub-muestra aleatoria.
- Muy robusto al ruido



Evaluación del modelo supervisado

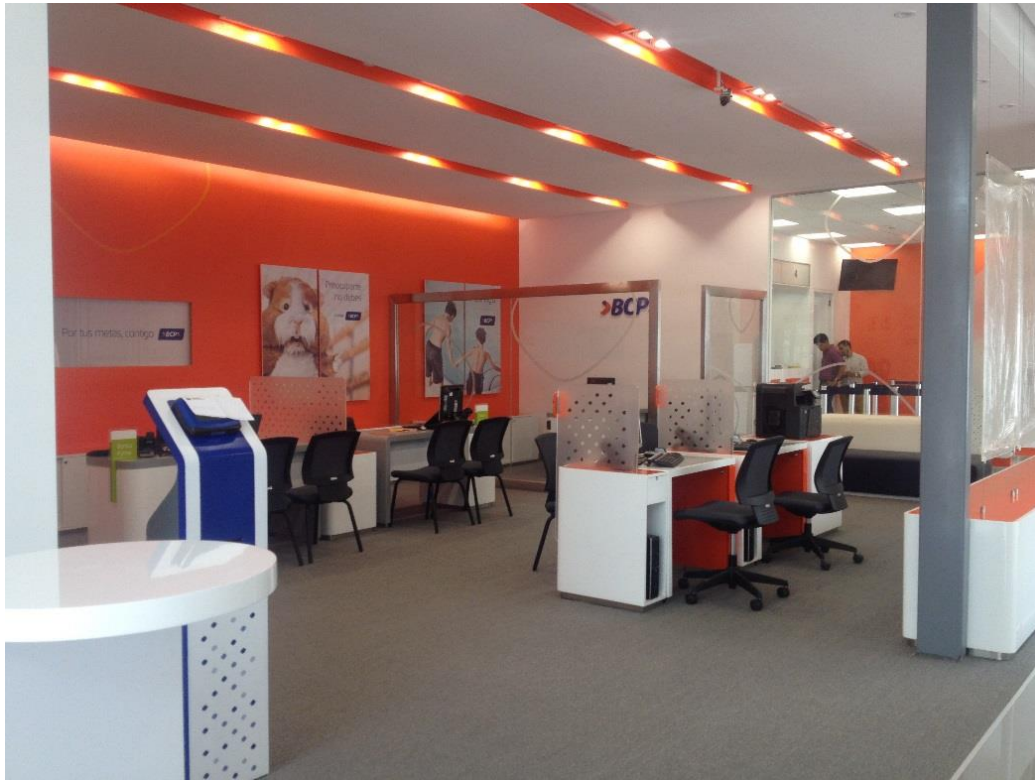


Métricas para Modelos de Regresión

- *MAE (Mean Absolute Error)*
 - Promedio de los errores de predicción del modelo.
- *RMSE (Root Mean Squared Error)*
 - Desviación promedio de los errores de predicción (unidades de la variable objetivo)
- *R^2 o Coeficiente de Determinación*
 - Porcentaje de variación explicada en los resultados, debido al modelo. Valor varia entre 0 y 1.

Detalles de los casos a desarrollar

- Caso 2: predicción de ingresos de agencias



$$\text{Ingresos} = f(v1, v2, v3, \dots, vp)$$

Procesos analíticos utilizando Python

