#AprendeDesdeCasa
#AprendeConLosPioneros

Online | UPC Perú

PROGRAMA DE ESPECIALIZACIÓN ANALÍTICA

ADVANCED DATA SCIENCE

Métodos No Supervisados

SESIÓN I

Docente: Jimmy Salazar

PROGRAMA DE ESPECIALIZACIÓN ANALÍTICA

ADVANCED DATA SCIENCE

Online | DiiC Perú

REGLAS

Se requiere **puntualidad** para un mejor desarrollo del curso.

Para una mayor concentración **mantener silenciado el micrófono** durante la sesión.

Las preguntas se realizarán **a través del cha**t y en caso de que lo requieran **podrán activar el micrófono**.

Realizar las actividades y/o tareas encomendadas en **los plazos determinados**.

Identificarse en la sala Zoom con el primer nombre y primer apellido.

PROGRAMA DE ESPECIALIZACIÓN ANALÍTICA

ADVANCED DATA SCIENCE

Online | DMC Perú

ITINERARIO

07:00 PM – 07:30 PM	***Soporte técnico DMC***
07:30 PM – 08:30 PM	***Módulo 1***
08:30 PM – 09:30 PM	***Módulo 2***

Horario de Atención Área Académica 09:00 am a 10:00 pm

SILABO

Agenda

1. Introducción Métodos No Supervisados
2. ´Segmentación Jerárquica
3. Segmentación K-Means
    - Metodología
    - Caso práctico real

PROGRAMA DE ESPECIALIZACIÓN ANALÍTICA

ADVANCED DATA SCIENCE

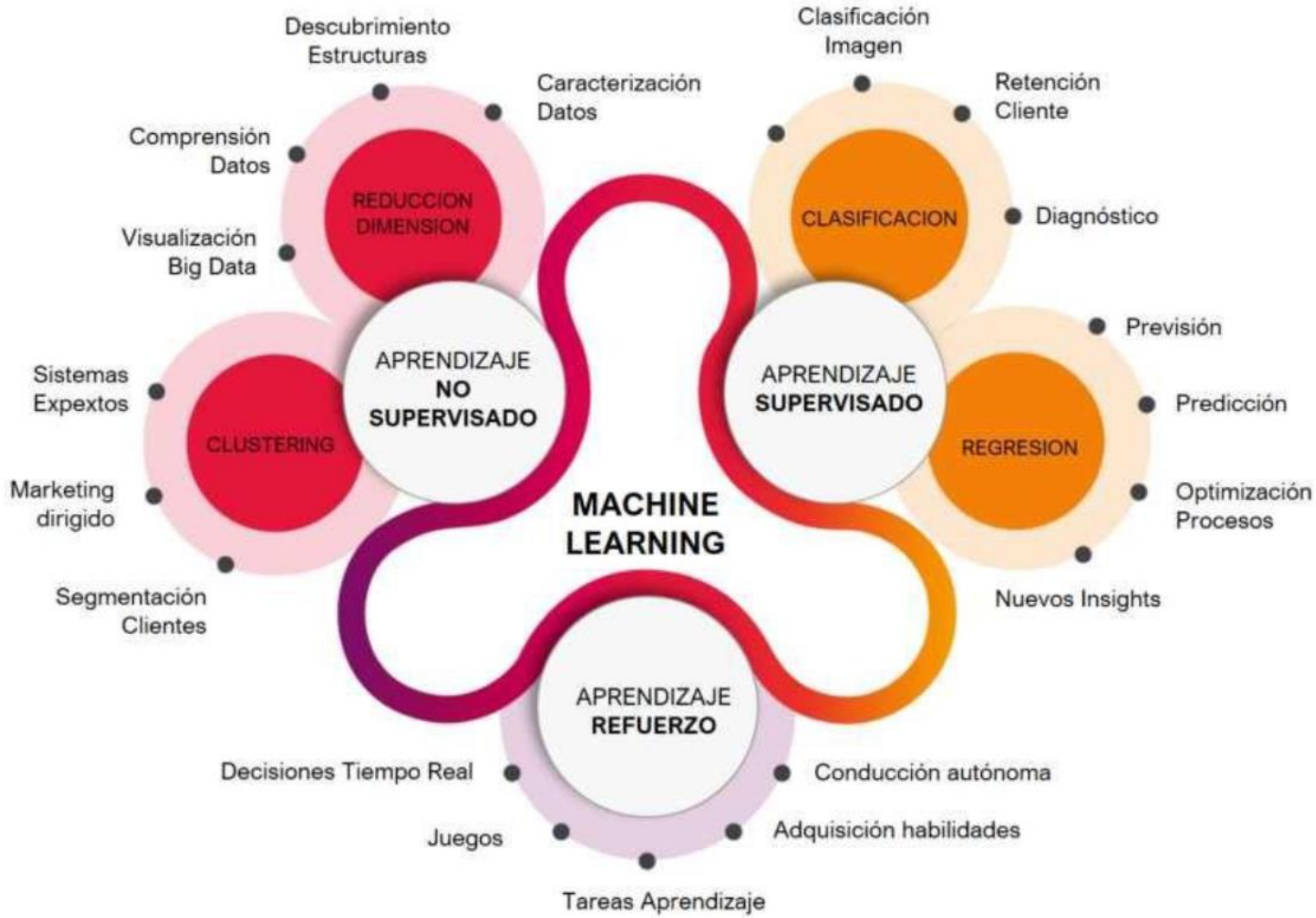
Online | DiHiC Perú

EVALUACIÓN

Asistencia (Curso): mínimo 80% sesiones para recibir la certificación

Examen Final (30%) + ***Trabajo*** (70%)

Aprendizaje No Supervisado


MACHINE LEARNING
APRENDIZAJE NO SUPERVISADO
APRENDIZAJE SUPERVISADO
APRENDIZAJE REFUERZO
REDUCCION DIMENSION
CLUSTERING
CLASIFICACION
REGRESION
Descubrimiento Estructuras
Caracterización Datos
Comprensión Datos
Visualización Big Data
Sistemas Expextos
Marketing dirigido
Segmentación Clientes
Clasificación Imagen
Retención Cliente
Diagnóstico
Previsión
Predicción
Optimización Procesos
Nuevos Insights
Decisiones Tiempo Real
Conducción autónoma
Juegos
Adquisición habilidades
Tareas Aprendizaje

Características

Pizarra 1

Aprendizaje no supervisado

Dataset de Entrenamiento

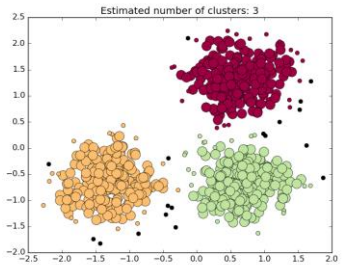
		Predictores (Features)			
		x_1	x_2	.....	x_p
Muestras	1				
	n				

n muestras de entrenamiento o instancias

p predictores

Resolución de problemas:

- Clustering
- Reducción de dimensionalidad.
- Detection de anomalias.



Datos Transformados y Agrupados en Clústeres

No tenemos etiquetas en los datos (Y)

Características

Pizarra 2

- ✓ En el **Aprendizaje NO Supervisado,** los datos de entrenamiento no están etiquetados con una **salida Y** (variable objetivo, target, etc.).

- ✓ A diferencia del aprendizaje supervisado, en el no supervisado **no hay forma determinística de verificar el performance del modelo**. Sólo se puede evaluar con conocimiento del negocio.

- ✓ Algunas aplicaciones típicas del aprendizaje no supervisado son:
  - Segmentación de clientes.
  - Detección de fraude o anomalías.

- ✓ Otra aplicación importante es la **reducción de dimensionalidad.**

PROGRAMA DE ESPECIALIZACIÓN ANALÍTICA
ADVANCED DATA SCIENCE

Online | DiiC Perú

Metodología

Proceso de Segmentación

Análisis y Exploración

- Construcción de matriz.
- Filtro de negocio.
- Estudio de variables

Transformación

- Dimensión de variables.

Se transforma en una única dimensión para las variables en estudio.

Outliers

- Análisis de outliers.
- Eliminación o agrupamiento de outliers.

Reducción de dimensión

- Análisis de reducción de dimensión de variables.
- Método ***PCA***
- Características de componentes.

Modelamiento

- Modelo de segmentación.
- Análisis de indicadores de segmentación.
- Segmento ***K-Means***.
- Segmento ***DBSCAN***

Evaluación

- Análisis de distribución de segmentos.
- Correlación de Segmentos.
- Elección del Segmento apropiado.

Perfil y Visualización

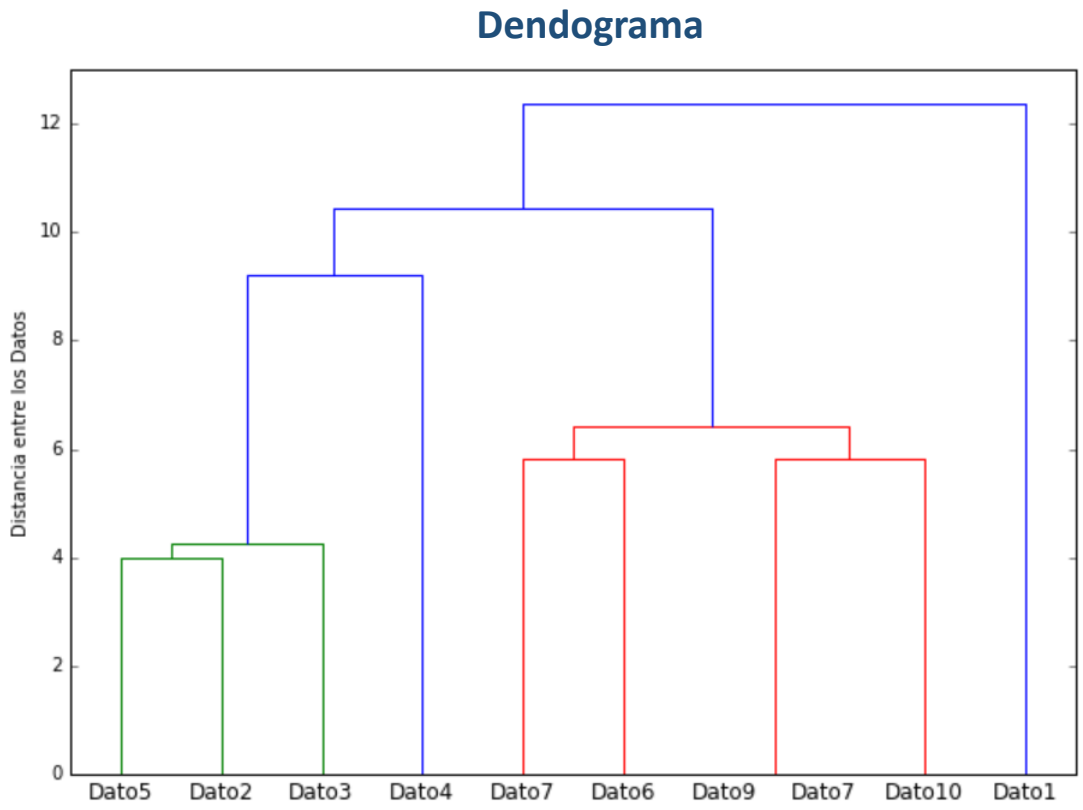
- Características de los segmentos.
- Visualización de migración de segmentos.
- Visualización **SOM** de ***Kohonen***

Retroalimentación

Segmentación Jerárquico

Agrupación Jerárquica

- ✓ Vinculación entre grupos
- ✓ La distancia es el promedio de las distancias de todos los pares de los dos conglomerados.
- ✓ La distancia es el promedio de las distancias de todos los pares de los dos conglomerados.
- ✓ Para determiner el número de cluster se analiza el histórico de agrupaciones (dendograma).



Ventaja:

- ✓ No require hacer inferencias sobre el número de clúster
- ✓ Permite visualizar las agrupaciones contínuas en forma de árbol (Dendograma)

Desventaja:

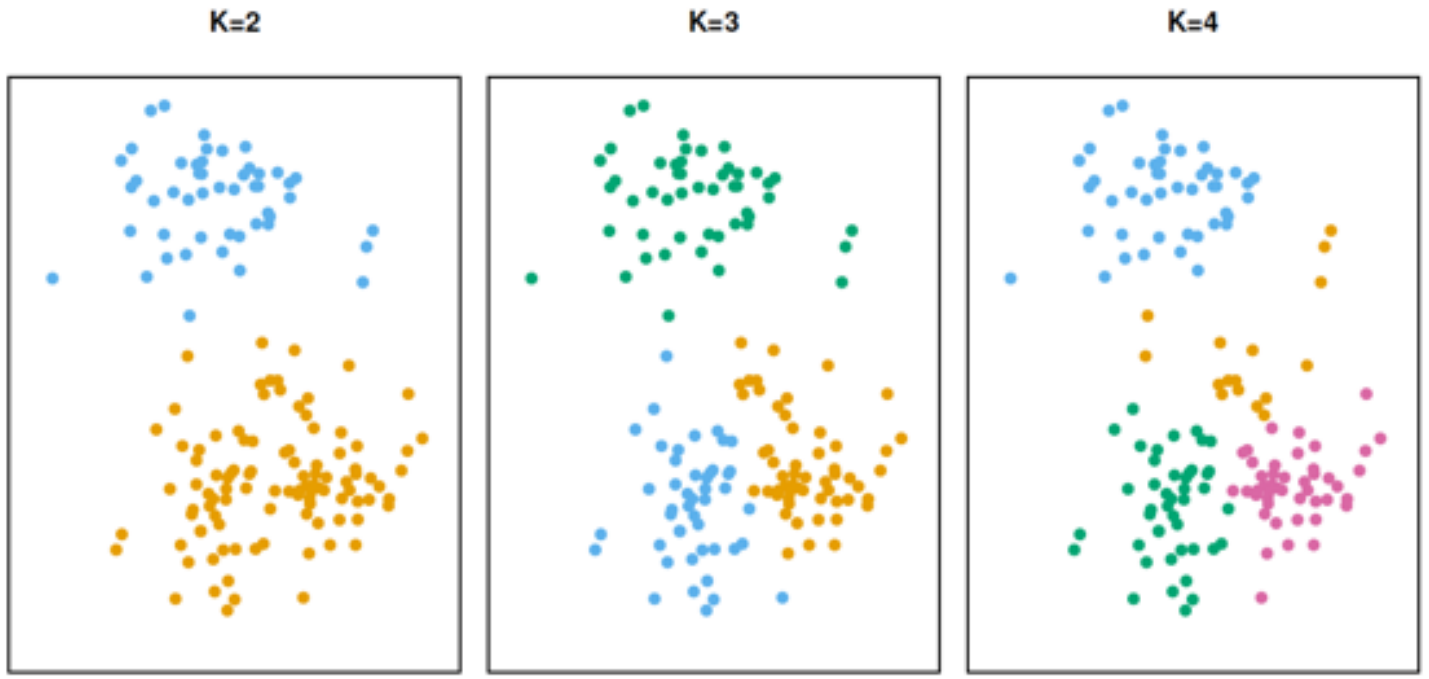
- ✓ Alto costo computacional
- ✓ Sensible respecto a las primeras agrupaciones
- ✓ Complicado de interpretar cuando el número de elementos es grande.

Segmentación
K-Means

K-means

- ✓ En K-means el objetivo es agrupar las observaciones de un dataset en un número K de clústeres.
- ✓ En número K es **hiperparámetro** que hay que darle al algorirtmo.

Ejemplo:´para el caso de *dos predictores* para distintos valores de **K**



K-means

Pizarra 3

- ✓ Al igual que el caso del PCA, las variables predictoras deben ser normalizadas antes de hacer el clustering.
- ✓ Los atributos han de ser numéricos (pues se computa la media de los mismos para reasignar los centroides).
- ✓ El **método de la silueta** suele ser un indicador para medir cómo de bien una observación se adapta a su cluster.
- ✓ Es muy sensible a los valores anómalo (outliers).

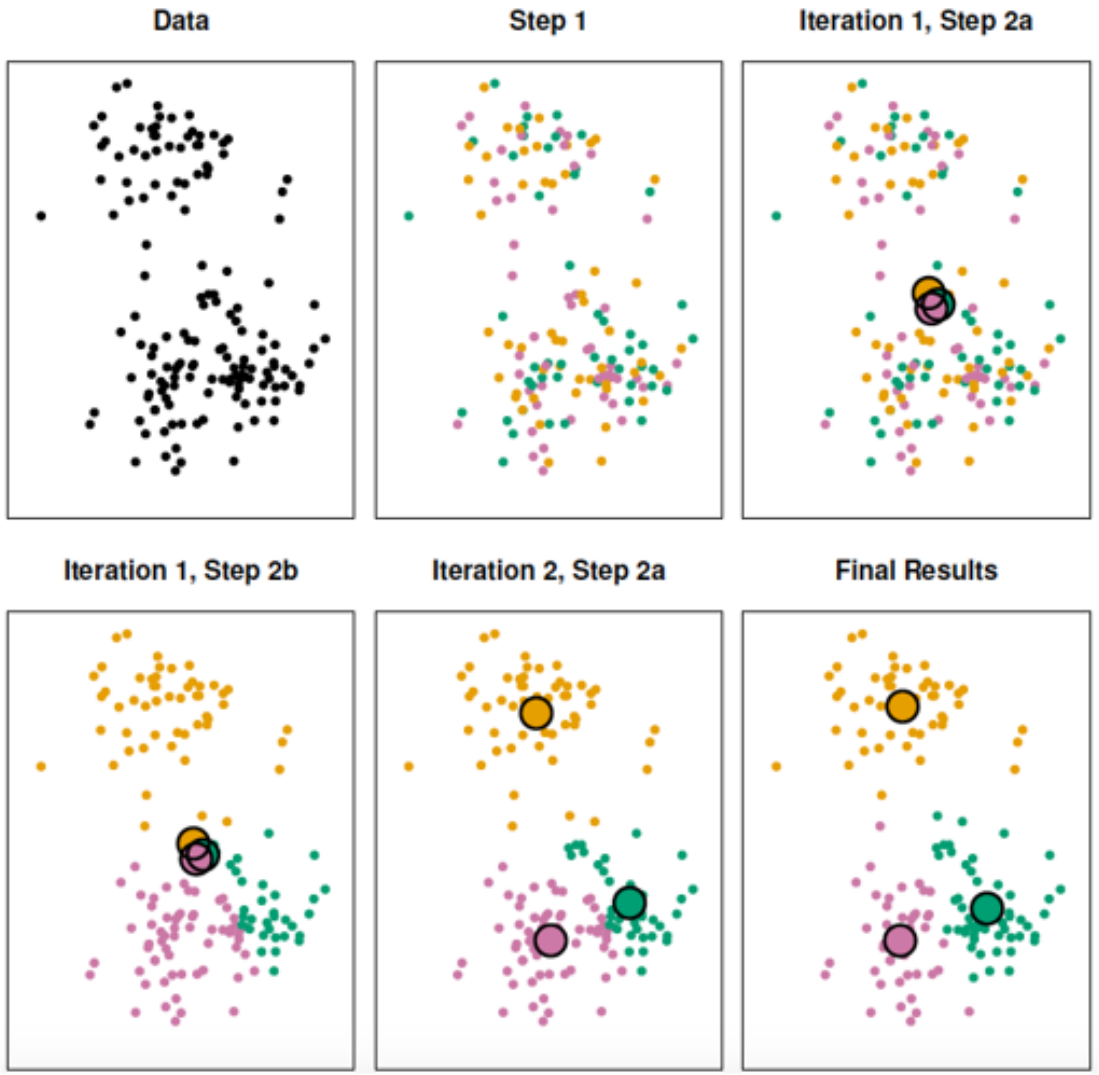
K-means

El algoritmo K-means se explica de la siguiente manera:

- ✓ De manera aleatoria asignar un número de 1 a K a cada observación. Esto será la asignación inicial a los clústeres de cada observación.

- ✓ Iterar sobre los siguientes pasos hasta que las asignaciones a los clústeres deje de cambiar:
    a. Para cada clúster, calcule el centroide será un vector compuesto por la media de los **p** predictores de las observaciones del mismo cluster.
    b. Reasigne cada observación al clúster cuyo centroide esté más cercano a la observación.

El siguiente gráfico ilustra el algoritmo K-means:



REFERENCIAS

- "Three ways to detect outliers" (Blog)
  - http://colingorrie.github.io/outlier-detection.html
- Anomaly detection with Local Outlier Factor (LOF)
  - http://scikit-learn.org/stable/auto_examples/neighbors/plot_lof.html
- How to Identify Outliers in your Data (Blog)
  - https://machinelearningmastery.com/how-to-identify-outliers-inyour-data/

PROGRAMA DE
ESPECIALIZACIÓN ANALÍTICA

ADVANCED
DATA SCIENCE

Online | DMC Perú