

“In God we trust. All others must bring data.” – W. Edwards Deming

ANALÍTICA PREDICTIVA

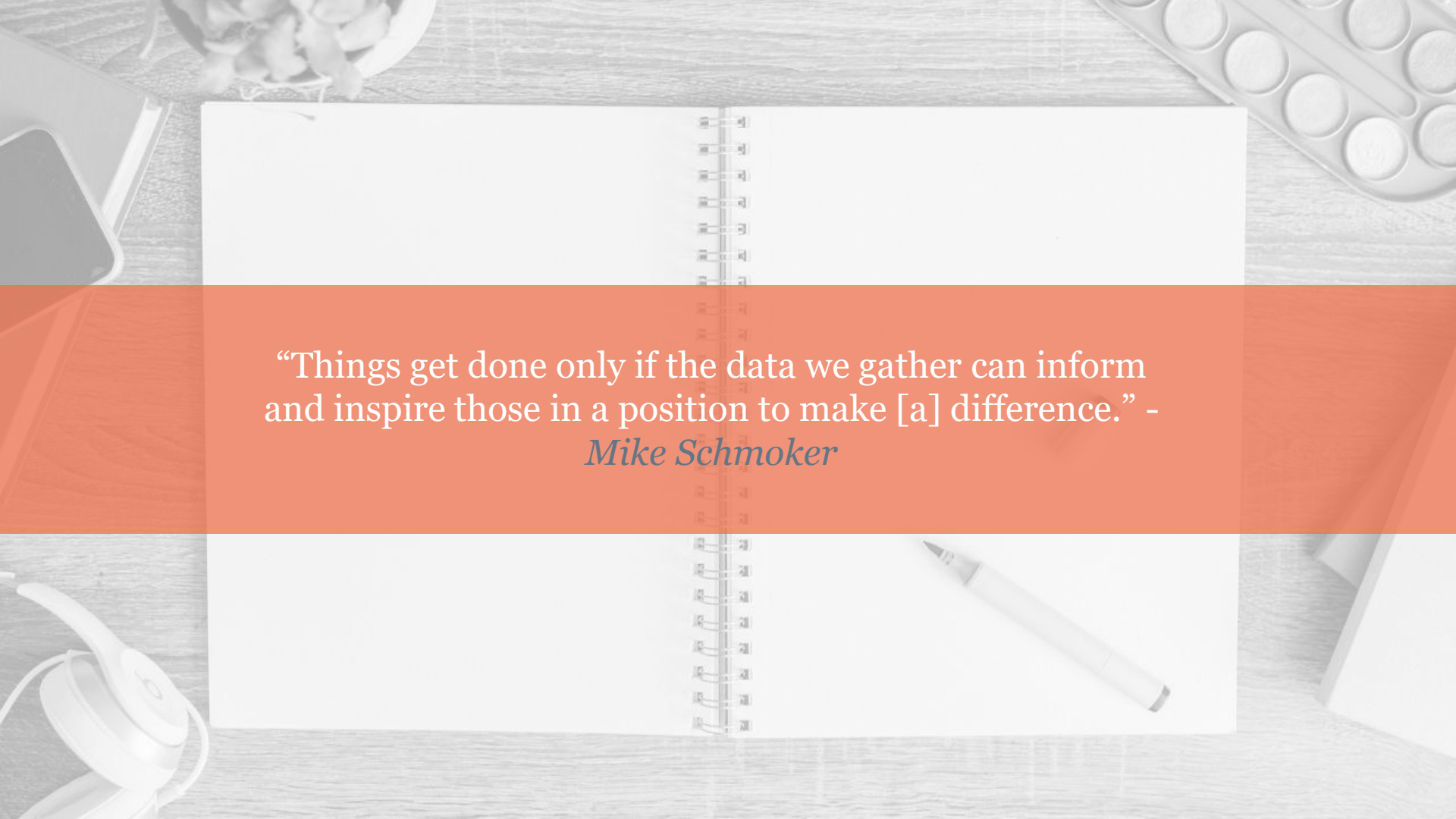
LAURA CAMILA
MOJICA LÓPEZ



The background of the slide is a blurred photograph of a desk. On the left, a pair of black over-ear headphones is visible. Next to them is a white mug. The desk surface is light-colored. In the background, there are some papers or a calendar on the wall. The right side of the slide is covered by a semi-transparent orange-red overlay.

AGENDA

- Importancia análisis de datos
- Analítica predictiva vs descriptiva
- Proceso
- Machine Learning
- Design Thinking
- Tipos de aprendizaje
- Tipos de problemas
- Familias y algoritmos
- Métricas



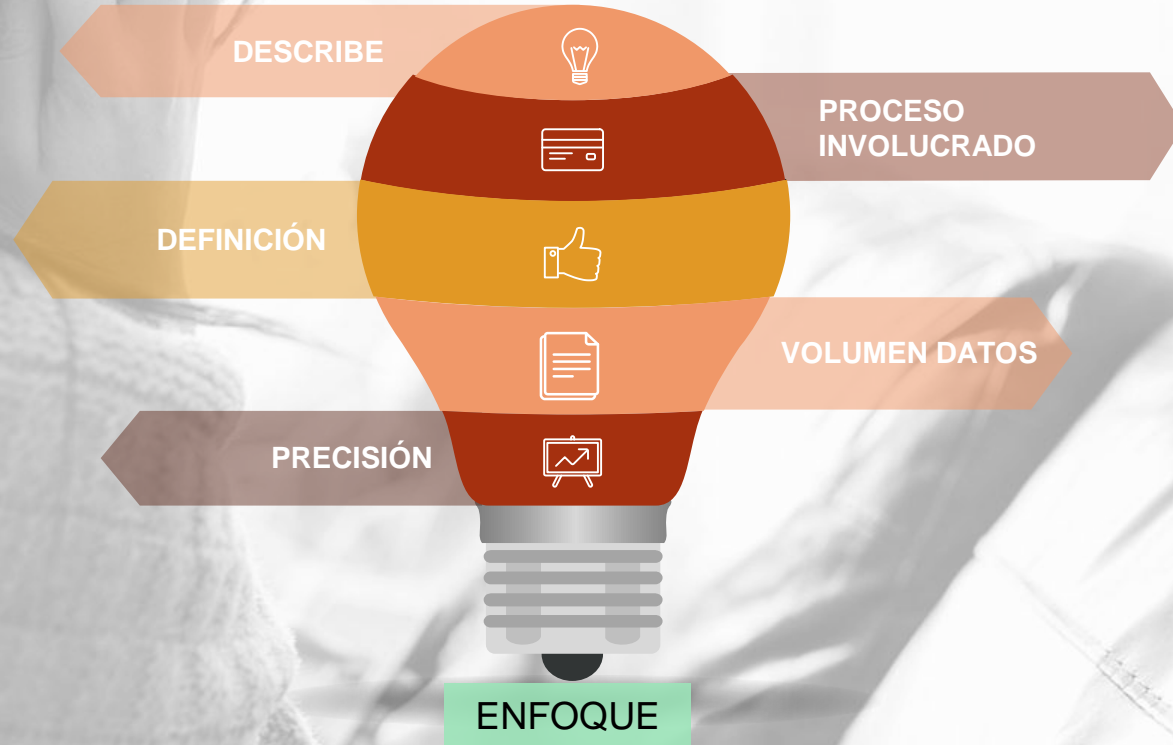
“Things get done only if the data we gather can inform
and inspire those in a position to make [a] difference.” -
Mike Schmoker

IMPORTANCIA ANÁLISIS DE DATOS

Los procesos analíticos
como Analítica Predictiva y
Analítica Descriptiva
ayudarán a una organización
a identificar el desempeño
de la empresa.



ANALÍTICA PREDICTIVA VS DESCRIPTIVA



DESCRIBE

DESCRIPTIVO



Qué sucedió en el **pasado**
usando los datos almacenados.

PREDICTIVO



Qué puede pasar en el **futuro**
utilizando los datos del pasado y
analizándolos.

PROCESO INVOLUCRADO

DESCRIPTIVO



Involucra agregación de datos y minería de datos.

PREDICTIVO



Involucra técnicas de clasificación y predicción.

DEFINICIÓN

DESCRIPTIVO



Proceso de encontrar información útil e importante al analizar una gran cantidad de datos.

PREDICTIVO



Este proceso involucra la predicción del futuro de la compañía.

VOLUMEN DE DATOS

DESCRIPTIVO



Involucra el procesamiento de una gran cantidad de datos que están almacenados en una bodega de datos. Está limitado por datos pasados.

PREDICTIVO



Involucra el análisis de una gran cantidad de datos y luego predecir el futuro utilizando técnicas avanzadas.

PRECISIÓN

DESCRIPTIVO



Provee data precisa en reportes
utilizando datos pasados.

PREDICTIVO



Los resultados no son precisos,
no dirá exactamente qué
sucederá, pero si dirá qué
puede pasar.

ENFOQUE

REACTIVO VS PROACTIVO

Es muy importante para cualquier organización hacer uso de análisis predictivo y análisis descriptivo para que puedan tener **éxito en el mercado**.

ANÁLISIS DESCRIPTIVO

Centrado en presentación de datos y visualización a las miras de gestión.

Menor riesgo -> implica analizar los datos pasados y proporcionar un informe de lo que realmente sucedió.

ANÁLISIS PREDICTIVO

Se centra en torno al modelo estadístico que ayuda a predecir el futuro.

Mayor riesgo -> implica analizar qué sucederá exactamente en el futuro basándose en los eventos pasados, pero es posible que la condición en particular no ocurra en el futuro.

PROCESO



EXTRAER

Extraer los datos de su fuente.



PREPARAR

Limpiar, refinar y preparar.



ELEGIR

Identificar variable objetivo.



PREDECIR

Implementar algoritmo.



PLANEAR

Desarrollar plan de acción.



MACHINE LEARNING

Construir y estudiar métodos que aprendan y
hagan predicciones sobre datos.

MACHINE LEARNING

Data

Machine
Learning

Distributed
Computing

REVISEMOS ALGUNOS CONCEPTOS



OBSERVACIONES

Elementos o entidades utilizados para el aprendizaje o la evaluación, por ejemplo, correos electrónicos.



ETIQUETA

Valores / categorías asignados a las observaciones, por ejemplo, spam, no spam



CARACTERÍSTICA

Atributos (generalmente numéricos) utilizados para representar una observación, por ejemplo, longitud, fecha, presencia de palabras clave



DATOS DE ENTRENAMIENTO Y PRUEBA

Observaciones utilizadas para entrenar y evaluar un algoritmo de aprendizaje, por ejemplo, un conjunto de correos electrónicos junto con sus etiquetas

CASO DE USO DE NEGOCIO



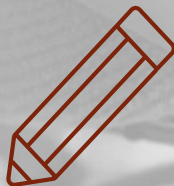
BANNER A



BANNER B

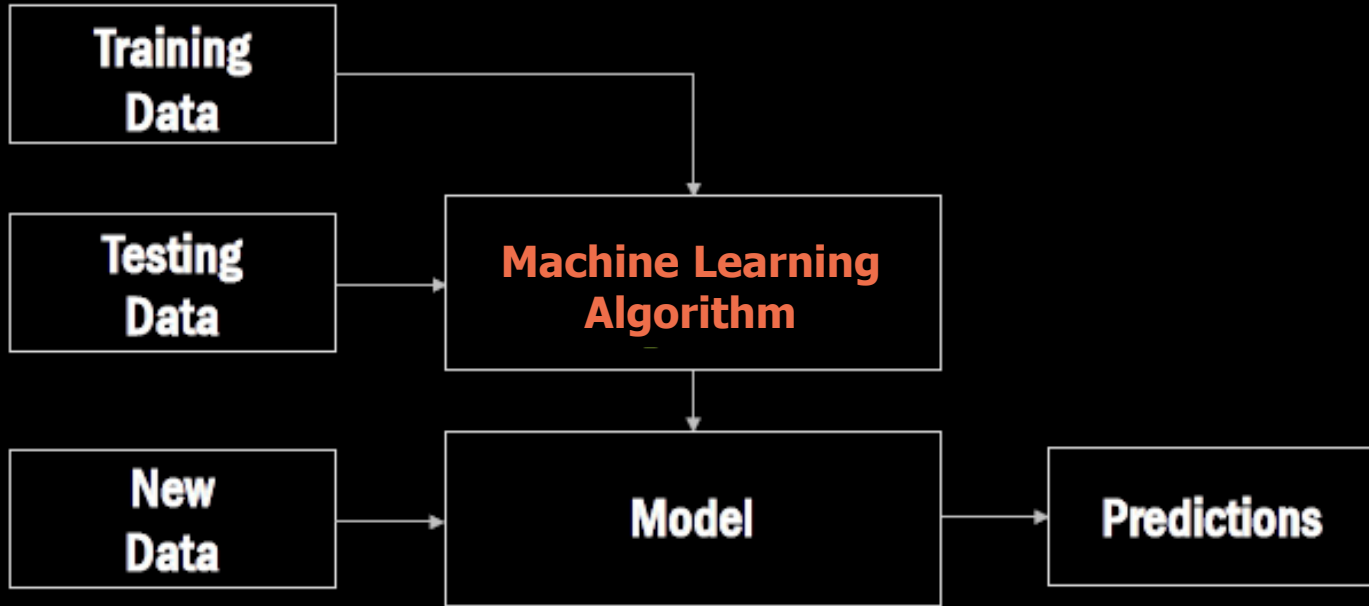
¿CUÁL ANUNCIO ELEGIRÁ
EL USUARIO?

TIPOS DE APRENDIZAJE



SUPERVISADO

NO SUPERVISADO



APRENDIZAJE SUPERVISADO

Se proporcionan observaciones de entrada y salidas etiquetadas.

Las etiquetas le enseñan al algoritmo a aprender el mapeo de **observaciones -> etiquetas**.

El objetivo es aprender las **reglas generales** que asignan un nuevo ejemplo a la salida prevista.

APRENDIZAJE SUPERVISADO

ENTRADA



```
graph LR; A[ENTRADA] --> B[ ]; B --> C[SALIDA];
```

SALIDA

Ejemplo: dado un conjunto de **características de la casa** junto con los **precios de la vivienda** correspondientes, predice un precio para una casa nueva según sus características (por ejemplo, tamaño, ubicación, etc.)

APRENDIZAJE NO SUPERVISADO

Solo se proporcionan observaciones de entrada, **no hay etiquetas**.

-> No hay información explícita sobre la verdad fundamental.

El algoritmo intenta descubrir la estructura interna de los datos basándose en un conocimiento previo sobre el resultado deseado.

Puede ser **un objetivo en sí mismo** (descubrir patrones ocultos, análisis de datos exploratorios).

Puede ser un **medio para un fin** (preprocesamiento para tareas supervisadas).

APRENDIZAJE NO SUPERVISADO

ENTRADA

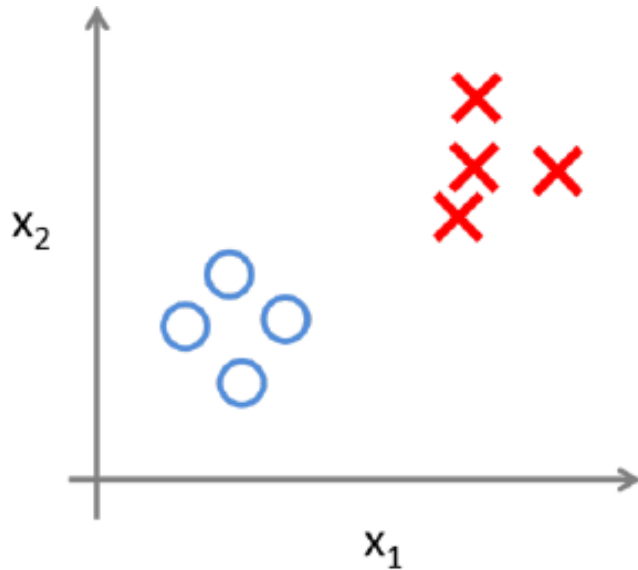


```
graph TD; A[ENTRADA] --> B[Ejemplo: dado un conjunto de transacciones de clientes, descubra cuál sería la mejor manera de agruparlos en grupos en función de la similitud de los clientes.]; B --> C[PREFERENCIA DE SALIDA]
```

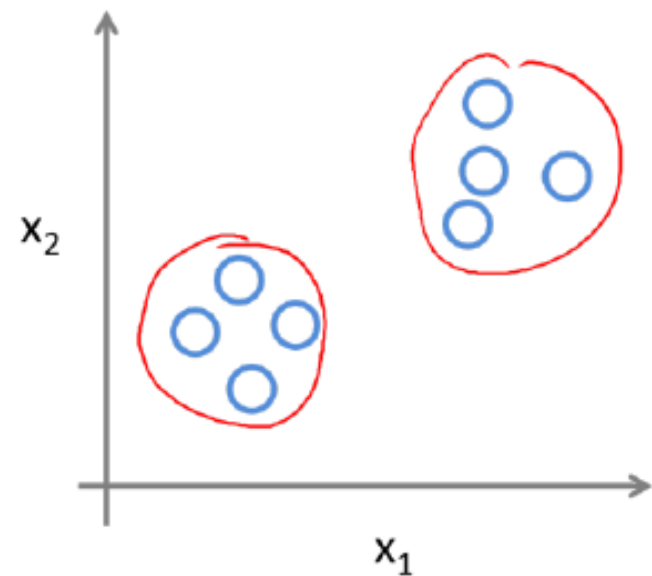
Ejemplo: dado un **conjunto de transacciones** de clientes, descubra cuál sería la mejor manera de agruparlos en grupos en función de la **similitud de los clientes**.

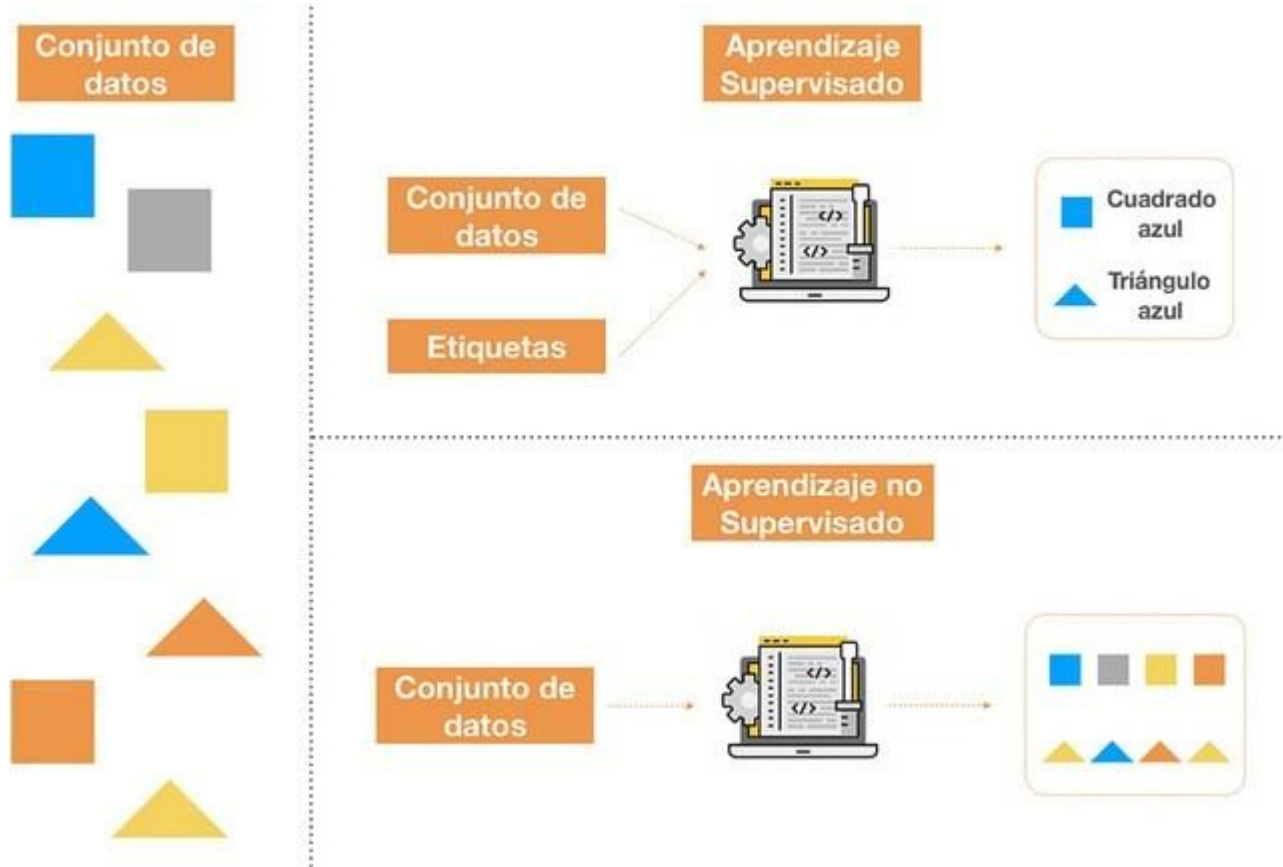
PREFERENCIA
DE SALIDA

APRENDIZAJE SUPERVISADO



APRENDIZAJE NO SUPERVISADO





TIPOS DE APRENDIZAJE

ITERACIÓN 1

DATASET DE HISTÓRICOS



BANNER A



BANNER B

Logs:

User X Features | Banner A | Click 0
User Y Features | Banner B | Click 0
User X Features | Banner B | Click 1
...

¿Aprendizaje supervisado o no supervisado?

Train
Input: Logs



ML Algorithm



Predict
Input: User Features
Output: Most preferred banner to show

TIPOS DE PROBLEMAS

CLASIFICACIÓN

REGRESIÓN

CLUSTERING

DETECCIÓN
ANOMALÍAS

REDUCCIÓN
DIMENSIONAL

CLASIFICACIÓN

Identifica a qué categoría pertenece un objeto

No hay noción de "cercanía" en entornos de clases múltiples.

Problema de **aprendizaje supervisado**

Detectar transacciones fraudulentas (**una clase**)

Clasifique los correos electrónicos por spam o no spam (**binario**)

Categorizar artículos en función de su tema (**multi-clase**)

Detectar objetos en la imagen (**multi-etiqueta**)

REGRESIÓN

Predecir un valor continuo asociado a un objeto

Problema de **aprendizaje supervisado**

Define la "cercanía" cuando se compara la predicción con la etiqueta

Predecir los precios de las acciones a partir de datos del mercado

Puntuación una solicitud de crédito basada en datos históricos

CLUSTERING

Agrupar objetos similares en grupos

Problema de aprendizaje **no supervisado**

Descubre públicos a los que apuntar en las redes sociales.

Grupo de verificación de datos basados en GEO-proximidad.

DETECCIÓN ANOMALÍAS

Identificar observaciones que no se ajusten a un **patrón esperado**.

Aborda el **aprendizaje supervisado y no supervisado**

Identificar transacciones fraudulentas o comportamiento **anormal** del cliente.

En la fabricación, detecte partes físicas que pueden fallar en un futuro cercano

REDUCCIÓN DIMENSIONAL

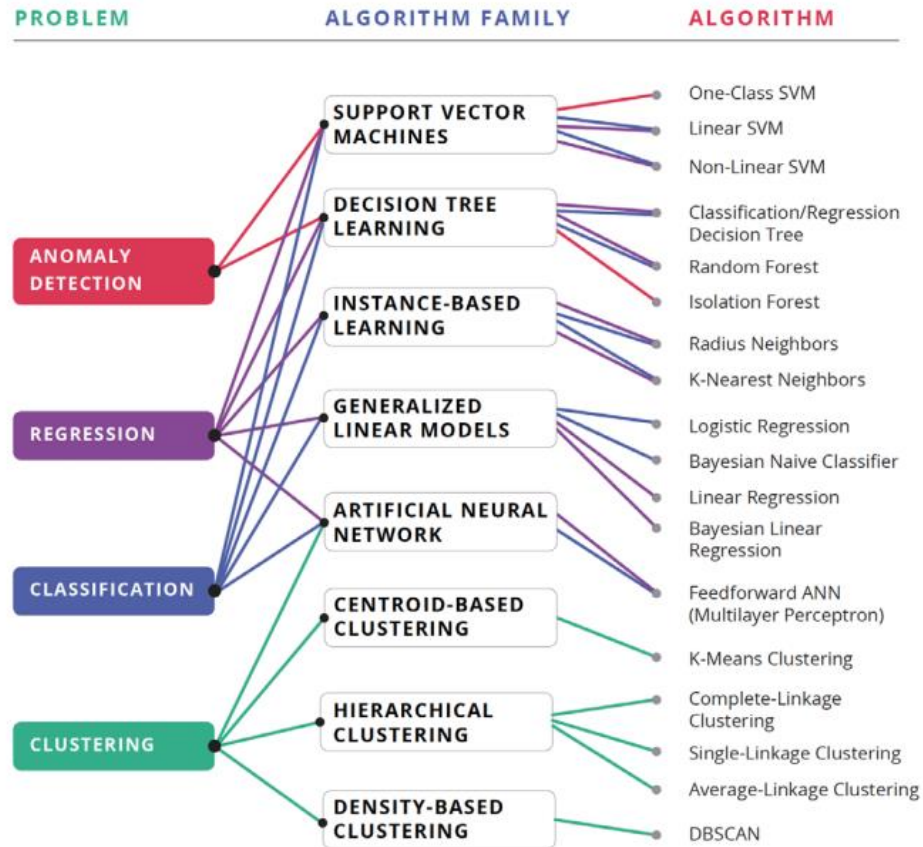
Proceso de **reducir el número de variables** aleatorias en consideración al obtener un conjunto de variables principales. Se puede dividir en selección de características y extracción de características.

Reduce el tiempo y el espacio de almacenamiento requerido.

La eliminación de la multicolinealidad mejora la interpretación de los parámetros del modelo de aprendizaje automático.



FAMILIAS Y ALGORITMOS



CÓMO ELEGIR EL ALGORITMO DE ACUERDO AL PROBLEMA

classification



MOTIVADORES



FAMILIA

- Grandes datos
- Datos pequeños
- Datos desequilibrados
- Interpretación de resultados
- Aprendizaje en línea
- Facilidad de uso



ALGORITMO

- Precisión
- Velocidad de entrenamiento
- Velocidad de predicción
- Resistencia por sobreajuste
- Interpretación probabilística



Usted está prediciendo la categoría:

Tienes datos etiquetados:

Necesitas seguir el Enfoque de Clasificación y sus algoritmos

No tienes datos etiquetados:

Tienes que ir para el enfoque de agrupación

Si está pronosticando cantidad:

Tienes que ir para el enfoque de regresión

De otra manera

Usted puede ir para el enfoque de reducción de dimensionalidad



PARA TENER EN CUENTA

Boosting: a menudo efectivo cuando hay disponible una gran cantidad de datos de entrenamiento.

Árboles aleatorios: a menudo muy efectivos y también pueden realizar regresión.

K-NN: lo más simple que puede hacer, a menudo eficaz pero lento y requiere mucha memoria.

Redes neuronales: lentas para entrenar pero muy rápidas para correr, aún con un rendimiento óptimo para el reconocimiento de letras.

SVM: entre los mejores con datos limitados, pero perdiendo contra el crecimiento o los árboles aleatorios solo cuando hay grandes conjuntos de datos disponibles.



MÉTRICAS

Construir y estudiar métodos que aprendan y
hagan predicciones sobre datos.

COMÚNMENTE USADAS

EXACTITUD

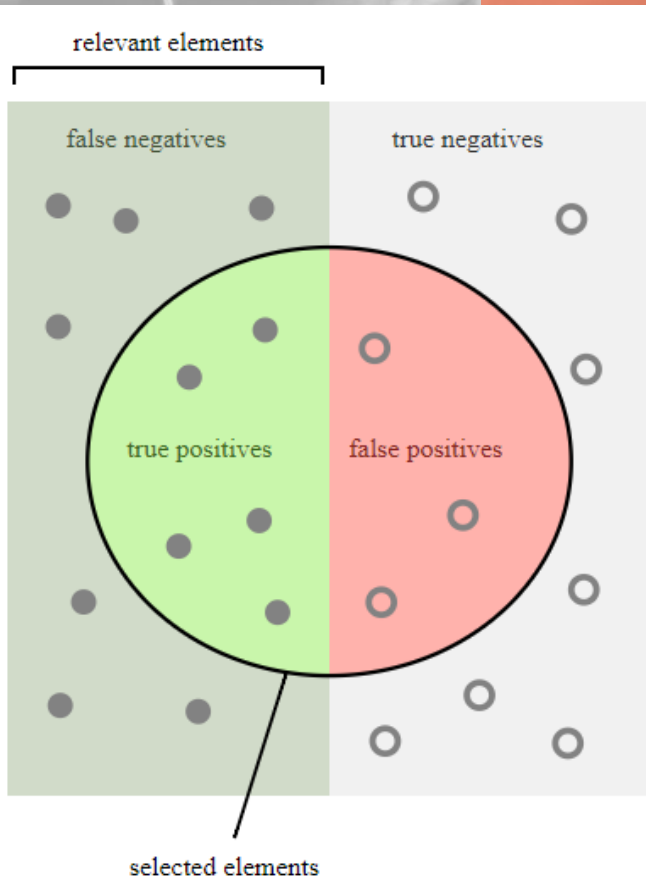
$$\frac{\# \text{ Predicciones correctas}}{\# \text{ Total de predicciones}}$$

Es la relación entre el número de predicciones correctas y el número total de muestras de entrada.

MATRIZ DE CONFUSIÓN

	Pred: No	Pred: Si
Real: No	50	10
Real: Si	5	100

Nos da una matriz como resultado y describe el rendimiento completo del modelo.



TÉRMINOS IMPORANTES

Verdaderos positivos (TP): Los casos en los que predijimos SÍ y el resultado real también fue SÍ.

Negativos verdaderos (TN): los casos en los que predijimos NO y la salida real fue NO.

Falsos positivos (FP): los casos en los que predijimos SÍ y la salida real fue NO.

Falsos negativos (FN): los casos en los que predijimos NO y la salida real fue SÍ.

$$\text{Exactitud} = \frac{TP + FN}{\# \text{Total de predicciones}} \longrightarrow \frac{100 + 50}{165} = 0,91$$

COMÚNMENTE USADAS

RECALL

$$Recall = \frac{TP}{TP + FN}$$



Es la relación entre el número de TP y el número total de muestras relevantes.

PRECISIÓN

$$Precisión = \frac{TP}{TP + FP}$$



Es la relación entre el número de TP y el número de resultados positivos predichos por el clasificador.

F1-SCORE

La media armónica entre precisión y recall. El rango para el puntaje F1 es [0, 1]. Le indica qué tan preciso es su clasificador (cuántas instancias clasifica correctamente), así como qué tan robusto es (no pierde un número significativo de instancias).

$$F1 = 2 * \frac{1}{\frac{1}{\text{precisión}} + \frac{1}{\text{recall}}}$$



Alta precisión pero menor recall, le da una precisión extrema, pero luego pierde una gran cantidad de instancias que son difíciles de clasificar.

COMÚNMENTE USADAS

MEAN ABSOLUTE ERROR

$$MAE = \frac{1}{N} \sum_{j=1}^N |y_j - \hat{y}_j|$$

Promedio de la diferencia entre los valores originales y los valores pronosticados. Nos da la medida de cuán lejos estaban las predicciones de la salida real

MEAN SQUARED ERROR

$$MSE = \frac{1}{N} \sum_{j=1}^N (y_j - \hat{y}_j)^2$$

Toma el promedio del cuadrado de la diferencia entre los valores originales y los valores predichos.

Algorithm name	Training Time	Prediction Time	Tuning Time	Initial Accuracy	Final Accuracy
Random Forest	2.61	0.47	94.44	81.61%	83.05%
KNeighbors	0.41	44.29	84.27	80.57%	83.05%
Logistic Regression	0.12	0.05	45.94	82.93%	82.93%
MLP	0.80	0.08	164.04	66.25%	82.90%
SVM	177.78	54.87	973.73	82.83%	82.83%
Linear SVM	5.93	0.04	82.91	82.69%	82.69%
Decision Trees	0.03	0.005	52.97	73.16%	82.36%
Naive Bayes	0.02	0.01	0	78.46%	78.46%

EVALUACIÓN

¿Preguntas?

MANOS A LA OBRA



BIBLIOGRAFÍA ADICIONAL

REGLAS ML

http://martin.zinkevich.org/rules_of_ml/rules_of_ml.pdf

SCI-KIT LEARN ALGORITHM CHEAT SHEET

https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

AZURE ML ALGORITHM CHEAT SHEET

<https://docs.microsoft.com/es-mx/azure/machine-learning/studio/algorithm-cheat-sheet>

MÉTRICAS

<https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>