Machine Learning

Algoritmos

Rogelio Ferreira Escutia

Profesor / Investigador
Tecnológico Nacional de México
Campus Morelia

Clasificaciones de Machine Learning

Clasificaciones (1)

- **Las clasificaciones mas generales son las siguientes:**
  - **Aprendizaje Supervisado.**
  - **Aprendizaje No Supervisado.**
  - **Aprendizaje Semi-Supervisado.**
  - **Aprendizaje por Refuerzo.**

Clasificaciones (2)

- **Otra clasificación según el tipo de algoritmo son las siguientes:**
  - **Regresión Lineal y Logística.**
  - **Árboles de Decisión.**
  - **Clasificadores Probabilísticos.**
  - **K-Means.**
  - **Support Vector Machines.**
  - **KNN.**
  - **Random Forest.**
  - **Redes Neuronales.**

Aprendizaje Supervisado

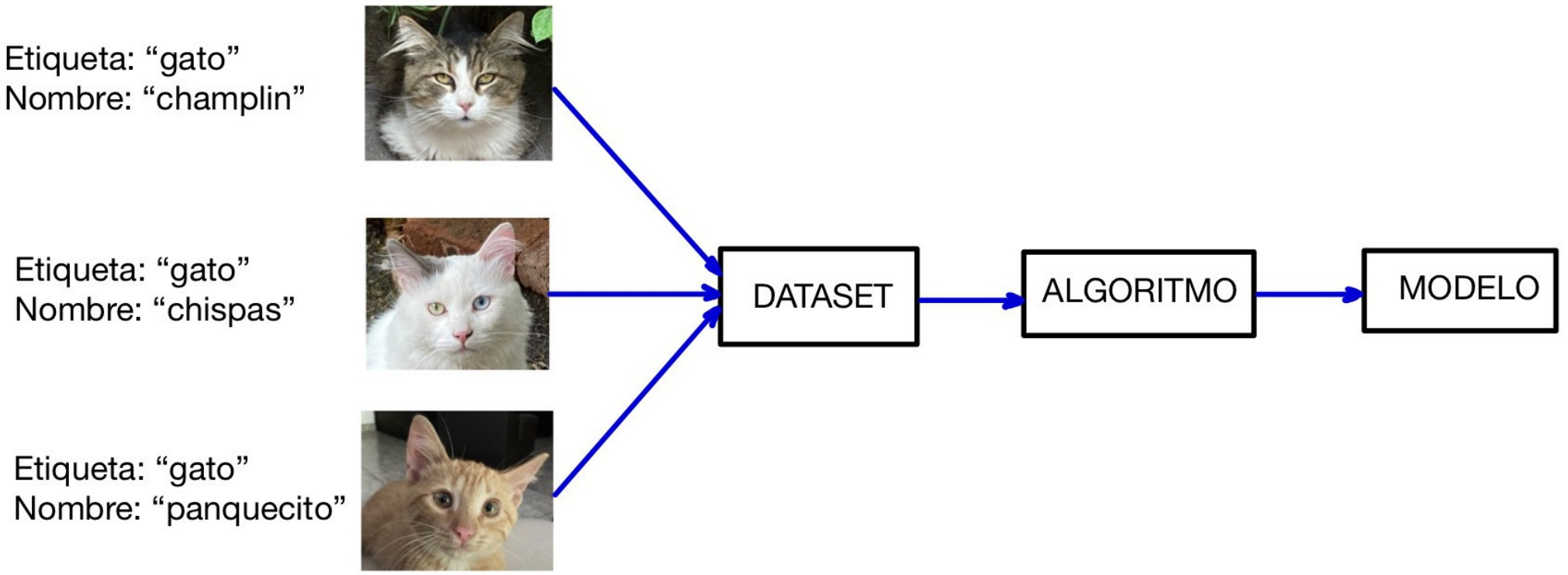
Aprendizaje Supervisado

- **Los sistemas de "Aprendizaje Supervisado" son aquellos donde los datos proporcionados viene debidamente etiquetados y clasificados, debido a esta característica se utiliza especialmente para los siguientes tipos de aplicaciones:**
  - **Clasificación**
  - **Regresión**

Entrenamiento

- **Los métodos que utilizan "*Aprendizaje Supervisado*" se dividen en 2 partes, la primera consiste en "*entrenar*" el sistema, es decir, generar un conjunto de datos debidamente etiquetados y ordenados para que el sistema pueda "*aprender correctamente*".**

Entrenamiento del Sistema

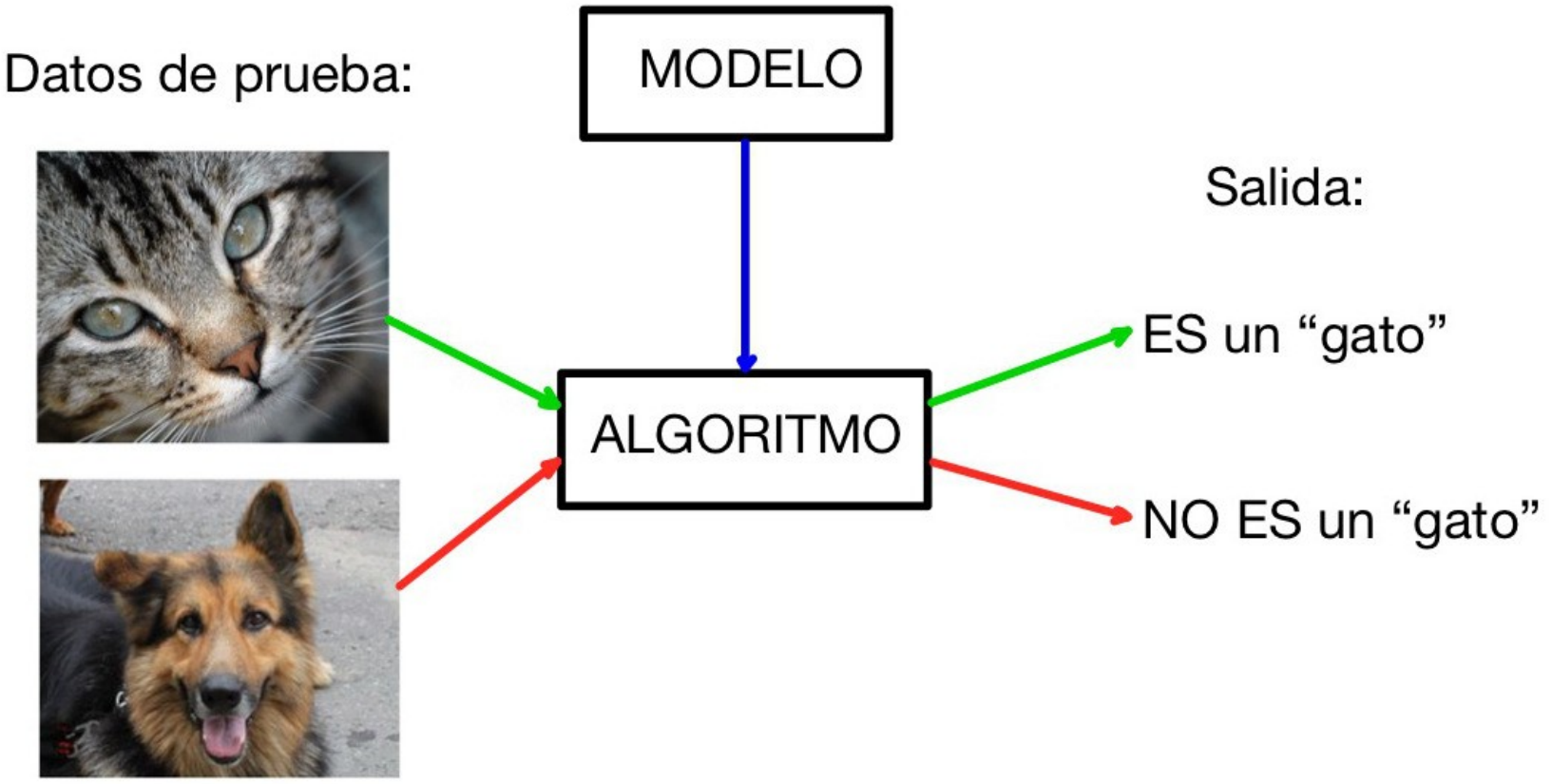


Clasificación

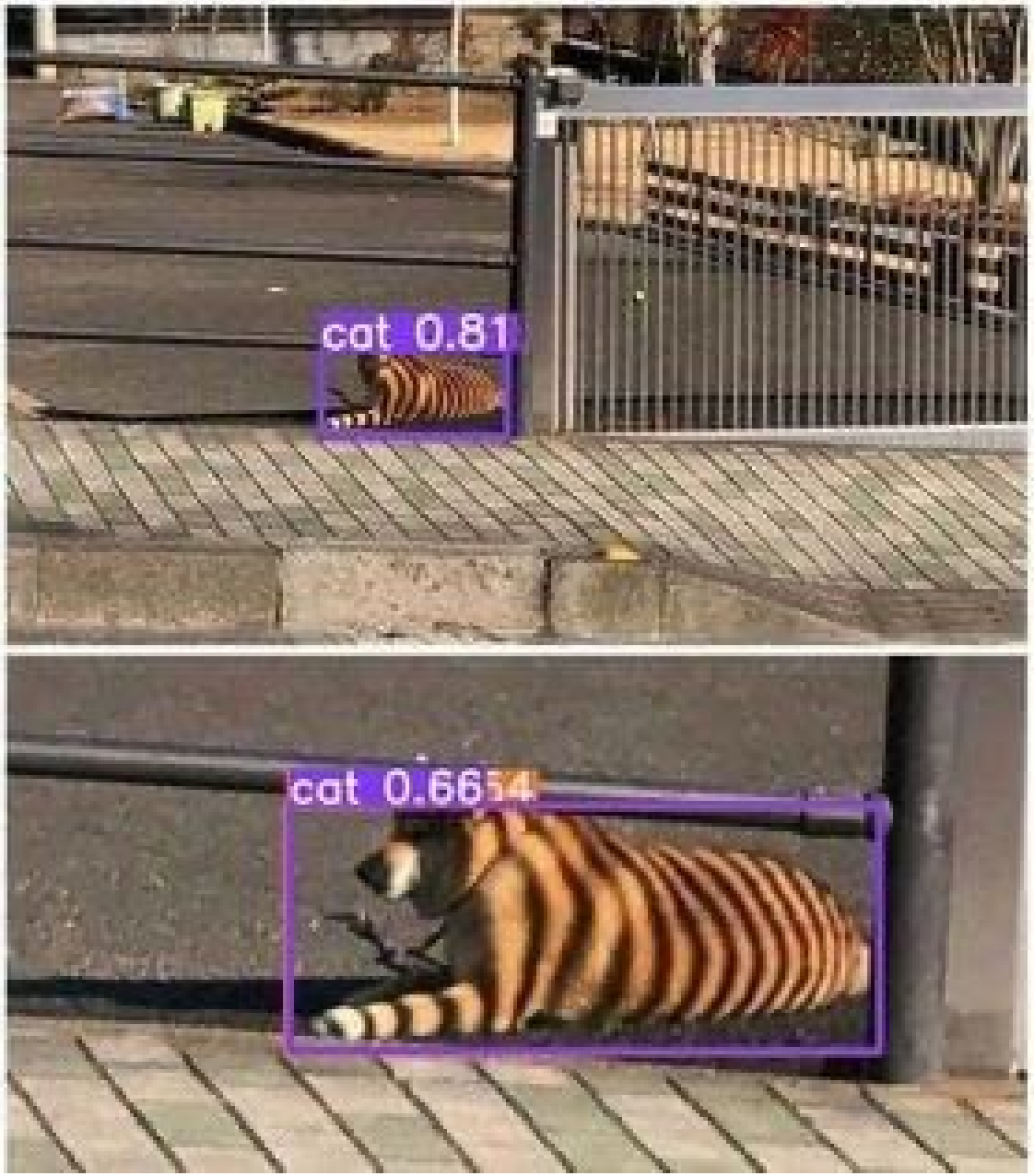
- **Una vez que tenemos entrenado nuestro "modelo", entonces ya podemos pasar a la parte de "clasificación", donde debemos de contar con otro conjunto de pruebas, el cual debe de ser diferente del que se usó para el entrenamiento.**

- **De este nuevo conjunto le damos 2 fotos nuevas para que el sistema nos diga si es ó no es un gato (recordar que el sistema sólo se entrenó para gatos).**

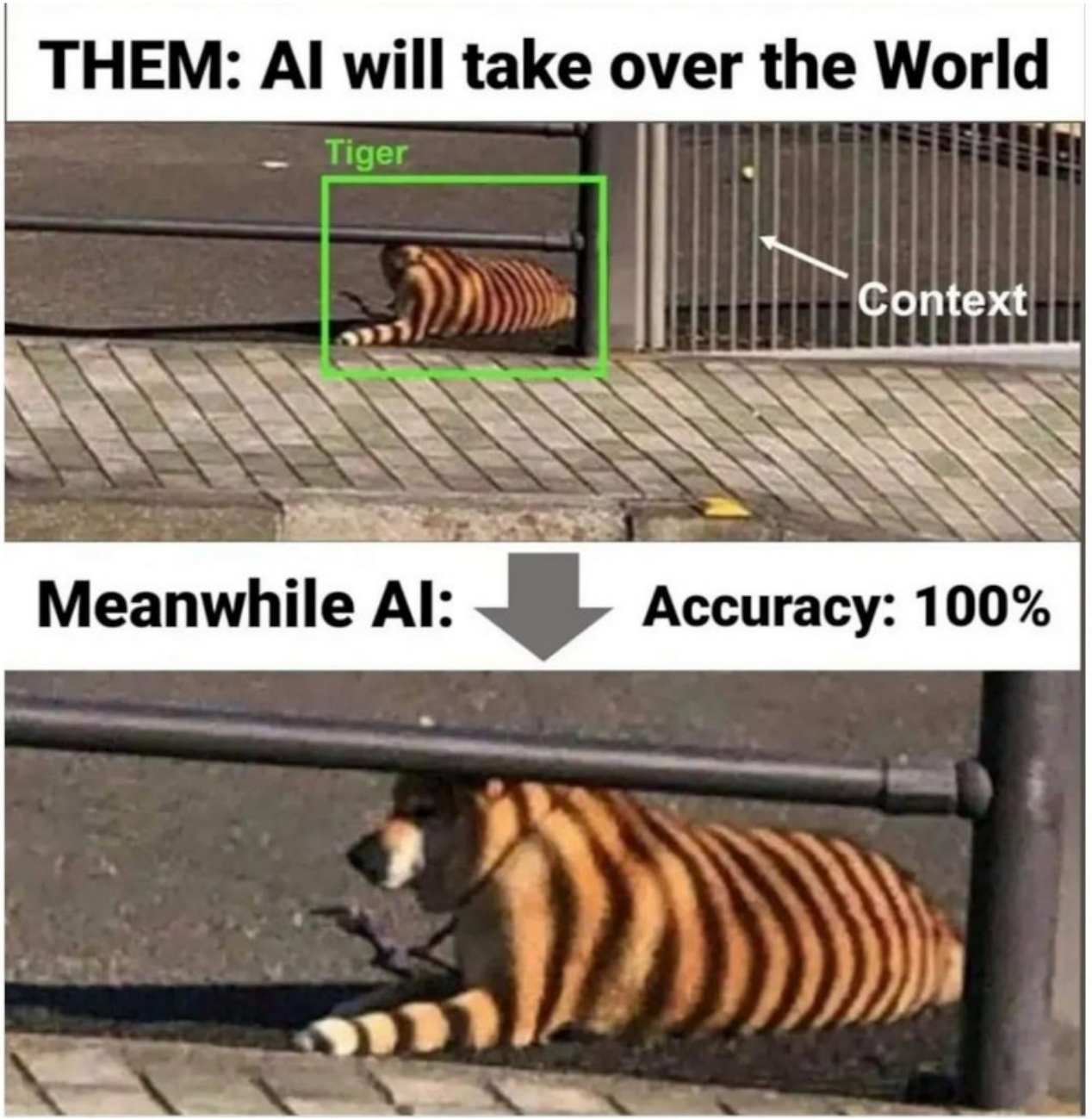
Clasificación de Imágenes



Errores



Errores


THEM: AI will take over the World
Tiger
Context
Meanwhile AI:
Accuracy: 100%

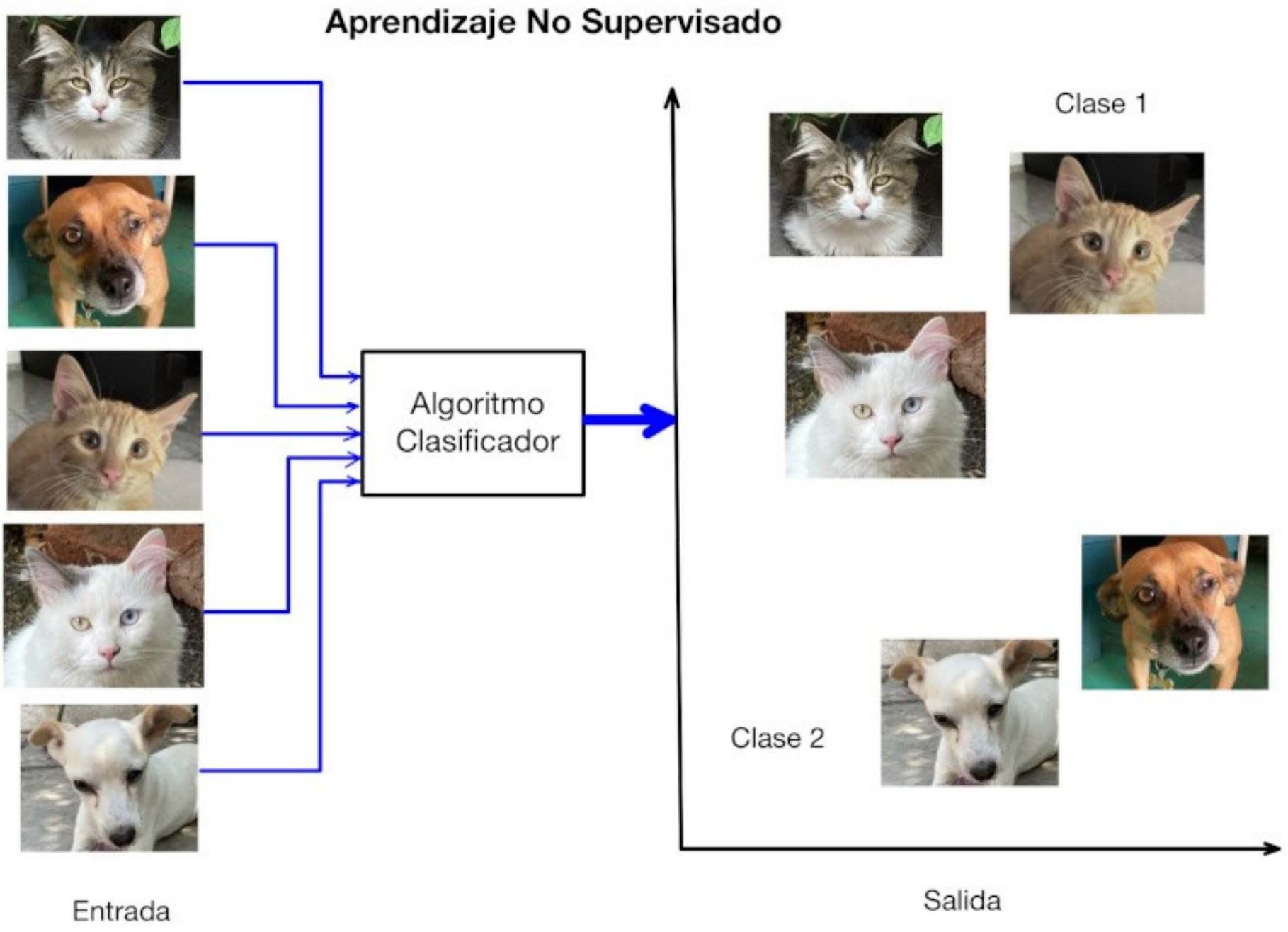
Errores

Aprendizaje
No Supervisado

Aprendizaje No Supervisado

- **Los sistemas de "Aprendizaje No Supervisado" son aquellos donde los datos proporcionados no vienen etiquetados y el sistema trata de entender el conjunto de datos que se le dieron y busca ordenarlos o clasificarlos de acuerdo a cierta similitud entre ellos.**

- **Son buenos para encontrar patrones que a veces no son fáciles de reconocer. También son buenos para detección de anomalías, es decir encontrar datos que se encuentran fuera de los patrones normales que siguen los demás datos.**


Aprendizaje No Supervisado
Algoritmo Clasificador
Clase 1
Clase 2
Entrada
Salida

Aprendizaje Semi-Supervisado

Aprendizaje Semi-Supervisado

- **Existe otro tipo de sistemas llamados "Semi-Supervisados", donde se combinan ambas técnicas anteriores, ya que una parte de los datos están etiquetados y otra no, esto permite agrupar los datos con etiquetas y después identificar y agrupar a los otros datos.**

Aprendizaje por Refuerzo

Aprendizaje por Refuerzo

- **En el "Aprendizaje por Refuerzo", a diferencia del "Aprendizaje Supervisado y No Supervisado", no se tiene un conjunto de datos (dataset) a partir del cual partir. Ahora se utiliza una técnica en donde existe un agente encargado de explorar el ambiente y determinar las acciones que se tengan que hacer para lograr un objetivo por medio de prueba y error.**

- **El sistema irá aprendiendo paulatinamente de acuerdo a las recompensas ó penalizaciones a partir de las acciones realizadas.**

Regresión Lineal

Regresión Lineal

- **Es un algoritmo de "Aprendizaje Supervisado" que se utiliza principalmente para hacer predicciones. Emplea métodos estadísticos para estimar la relación entre una variable dependiente y una o más variables independientes.**

- **La idea principal es dado un conjunto de datos, encontrar una aproximación lineal de su comportamiento y así poder estimar posibles valores nuevos.**

Regresión Lineal

Ecuación utilizada:

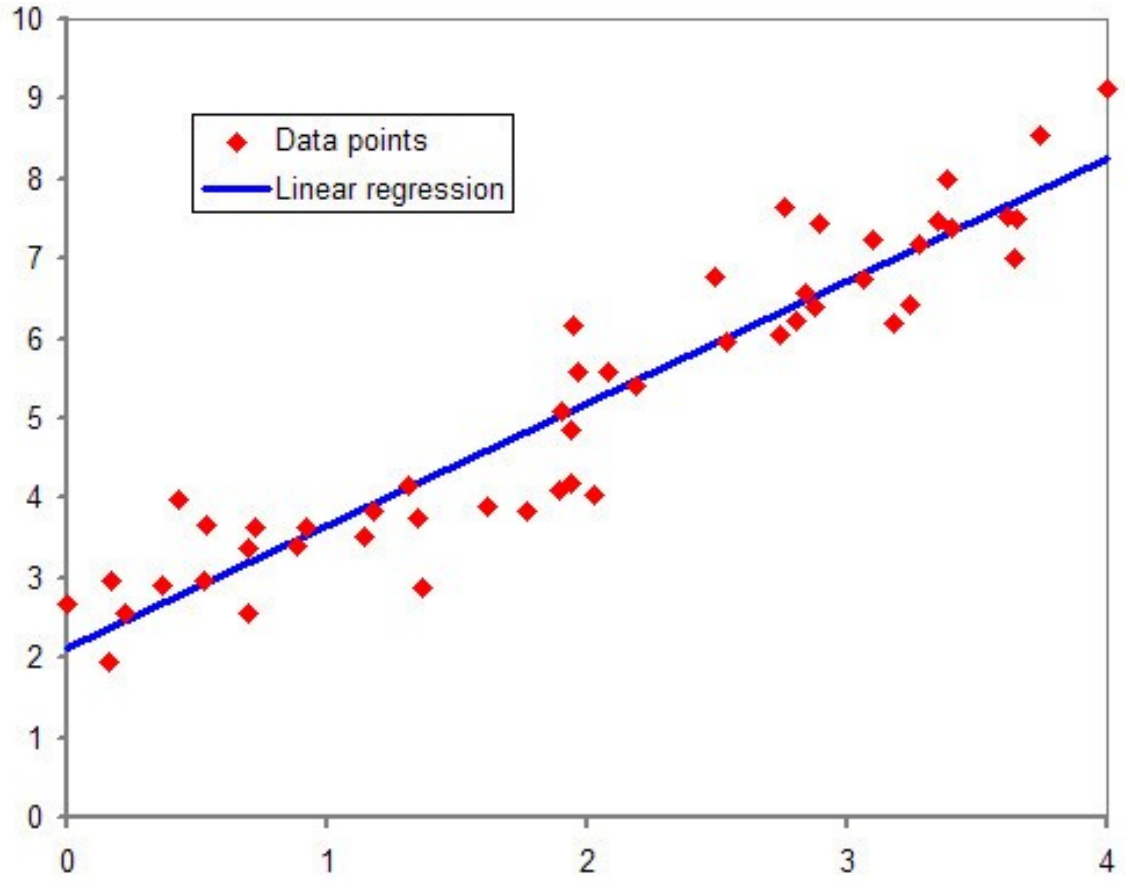
- $y = mx + b$

Donde:

- "y" es la variable dependiente.
- "m" es la pendiente de la recta.
- "x" es la variable independiente.
- "b" es un escalar.

Regresión Lineal

- **La siguiente gráfica es un una regresión lineal que utiliza 50 puntos aleatorios con una Distribución Gaussiana alrededor de la línea:**
  - **y = 1.533858x + 2.129333**



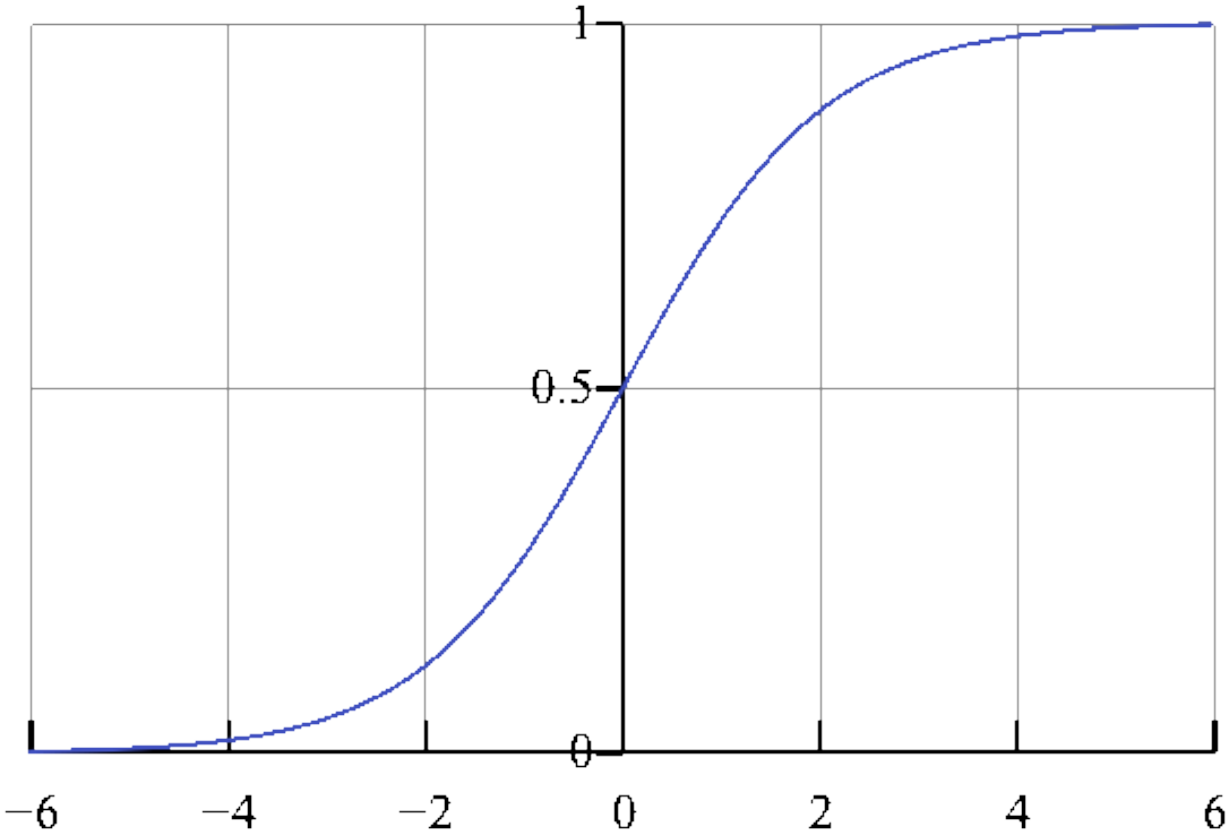
Regresión Logística

Regresión Logística

- **Es un algoritmo de "Aprendizaje Supervisado" que se utiliza principalmente para hacer clasificaciones binarias. Es ampliamente utilizado cuando se tiene un problema, en donde dados ciertos datos de entrada, se tiene que clasificar como "verdadero" ó "falso".**

Regresión Logística

- **La siguiente gráfica es una función logística con B0 + B1x + e en el eje horizontal y pi en el eje vertical:**



Árboles de Desición

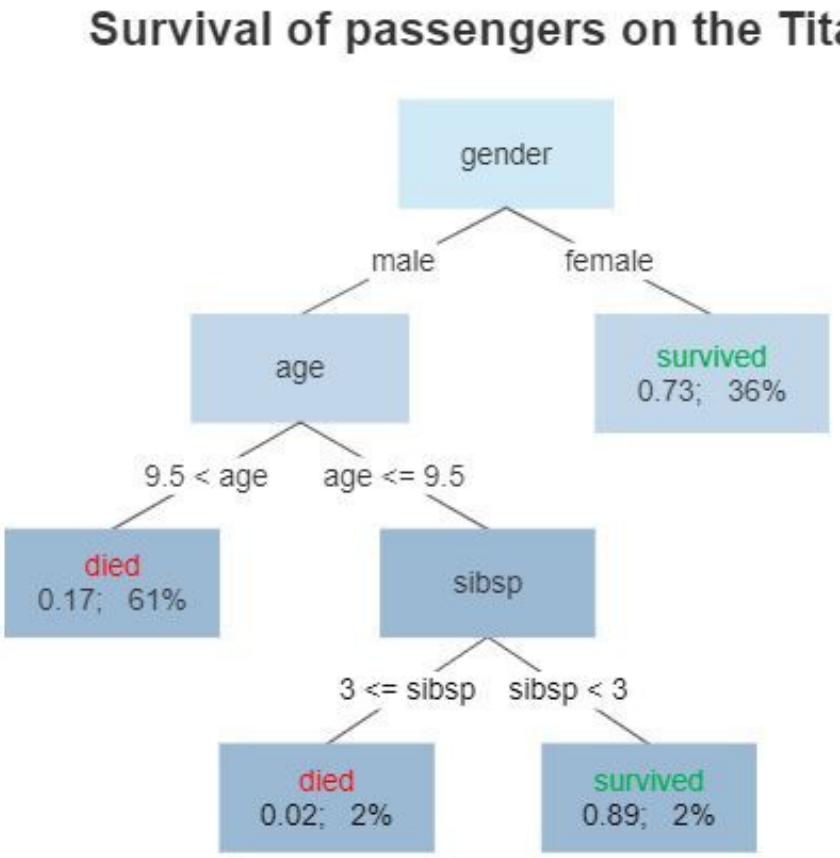
Árboles de Decisión

- **Es un algoritmo de "Aprendizaje Supervisado" que se utiliza principalmente para hacer predicciones de acuerdo a un conjunto dado de observaciones. Se utilizan "árboles de clasificación" en donde las hojas del árbol clases de etiquetas y las ramas representan conjunciones de características que llevan a clasificar a las clases de etiquetas.**

- **Se utilizan principalmente para la toma de decisiones ya que por su naturaleza gráfica resultan ser muy explícitos.**

Árboles de Decisión

- **El siguiente árbol muestra las características de los supervivientes al naufragio del "Titanic" con los datos obtenidos de los pasajeros:**



Clasificadores Probabilísticos

Clasificadores Probabilísticos

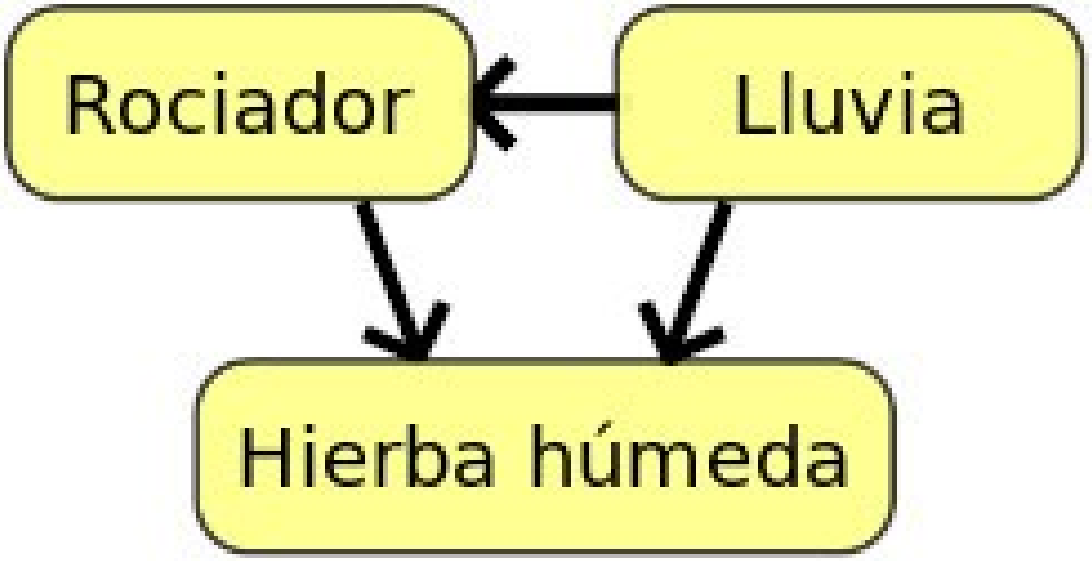
- **Es un conjunto de algoritmos que utilizan "Aprendizaje Supervisado" y que se utilizan principalmente para predicción.**

- **En este tipo de algoritmos, dado un conjunto de datos observados y de acuerdo a ciertas probabilidades, se puede predecir una posible salida.**

- **Dentro de este conjunto de algoritmos, el más utilizado es aquellos que se basan en "Redes Bayesianas", que surgen a base del "Teorema de Bayes", publicado por el matemático inglés Thomas Bayes en 1763. La implementación más simple es el "Naive Bayes" (ó "Clasificador Bayesiano Ingenuo" en español).**

Clasificadores Probabilísticos

- **Por ejemplo, se puede determinar si la hierba está mojada, de acuerdo a 2 posibles eventos, la lluvia y el agua del rociador:**



Clasificadores Probabilísticos

- **Algunas de las aplicaciones mas comunes son:**
  - **Reconocimiento de rostros.**
  - **Detección de correo SPAM.**
  - **Análisis de sentimientos.**
  - **Sistemas de recomendación.**
  - **Clasificación de artículos y textos.**

K-Means

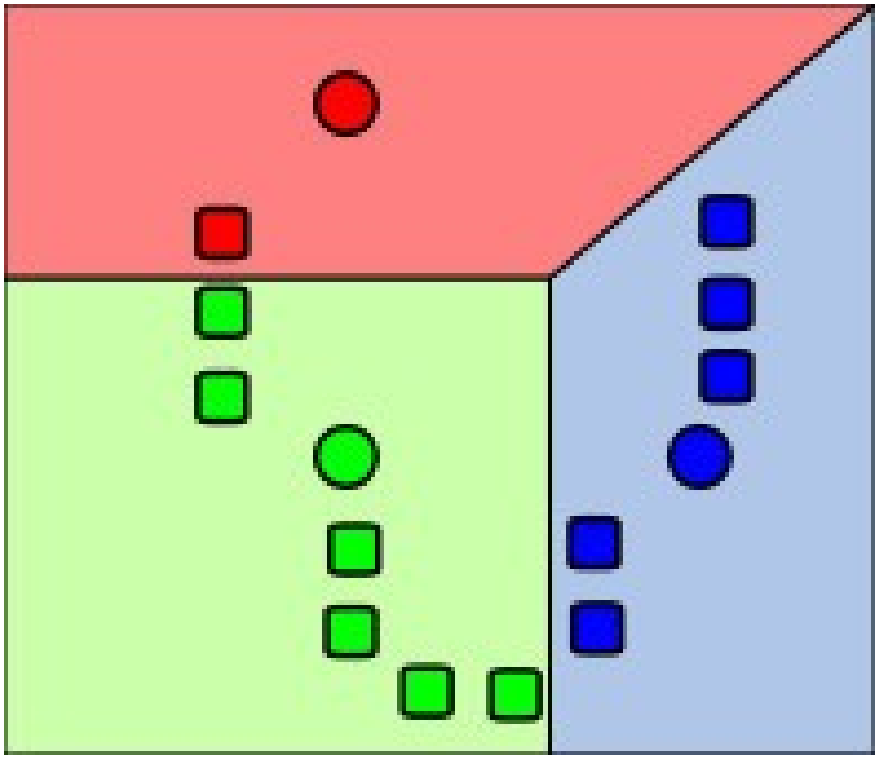
K-Means

- **Es un algoritmo de "Aprendizaje No Supervisado" que se utiliza principalmente para clasificación. De acuerdo a un conjunto dado de n datos, el algoritmo tratará de encontrar un conjunto de k agrupamientos (k-Clusters) que compartan ciertas características.**

- **Cuando se le asigne un caso de prueba, tratará de encontrar con cual de los k-Clusters tiene más afinidad, de acuerdo con la cercanía al promedio del conjunto de datos de cada uno de los k-Clusters.**

K-Means

- **En el siguiente ejemplo se observa un conjunto de datos que bajo el algoritmo "K-Means" (así se le denomina a este algoritmo) han sido encontrados 3 Clusters (agrupamientos):**



K-Means

- **Algoritmo:**

Algorithm 1 k-means algorithm

1: Specify the number k of clusters to assign.
2: Randomly initialize k centroids.
3: **repeat**
4: **expectation:** Assign each point to its closest centroid.
5: **maximization:** Compute the new centroid (mean) of each cluster.
6: **until** The centroid positions do not change.

https://realpython.com/k-means-clustering-python/

K-Means

- **Algunas de sus aplicaciones son:**
  - **Elegir mejores rutas de transporte.**
  - **Identificar noticias falsas.**
  - **Clasificar libros de acuerdo a su género.**
  - **Detección y filtrado de correo SPAM.**

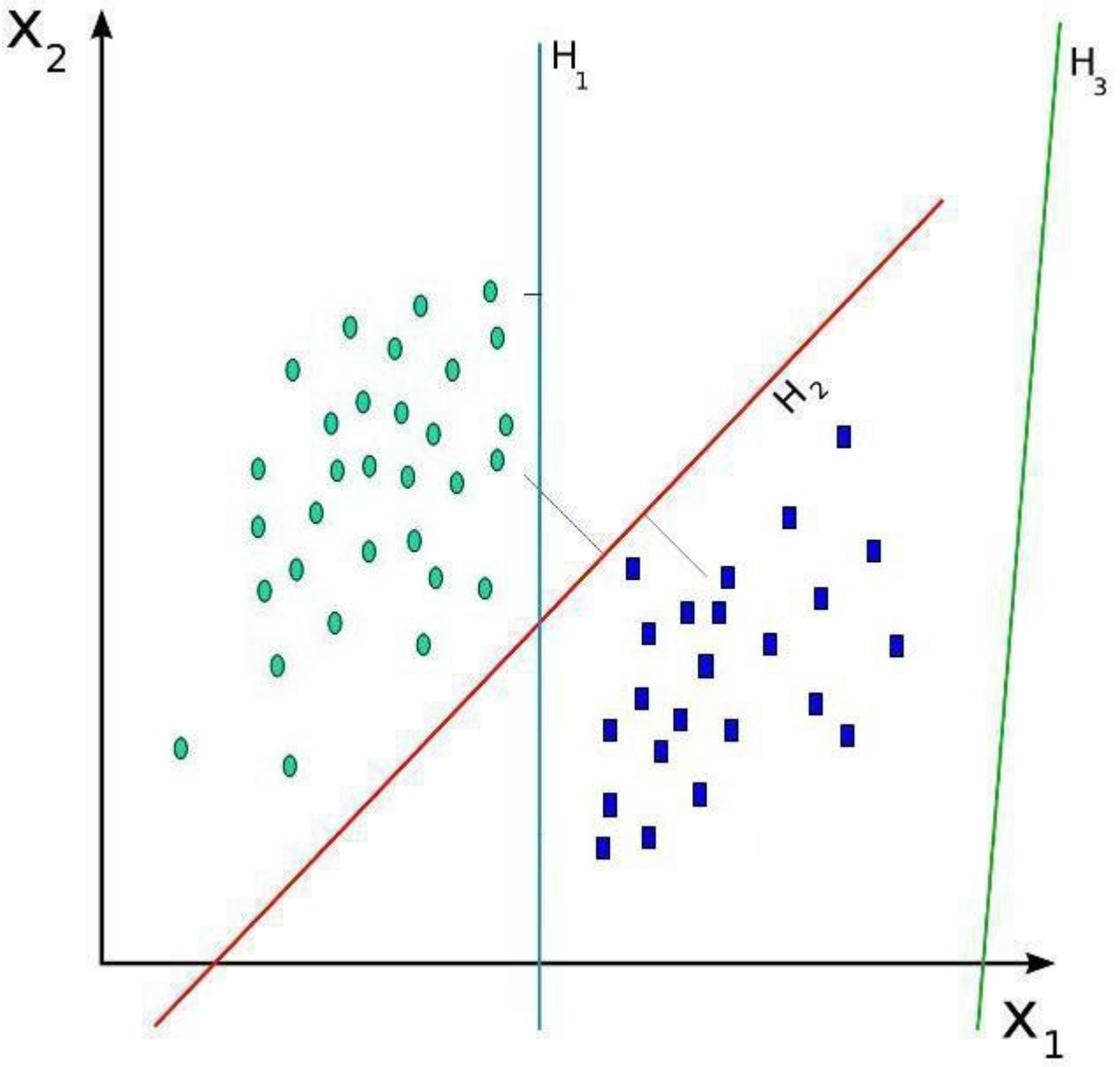
Support Vector Machines

SVM

- **Es un algoritmo de "Aprendizaje Supervisado" que se utiliza principalmente para clasificar y hacer análisis de regresión. Fue propuesto por Vladimir Vapnik de los Laboratorios Bell y que normalmente se abrevia como SVM (por sus siglas en inglés).**

- **Dado un conjunto de muestras para el entrenamiento, el SVM etiqueta las clases para que con una nueva muestra pueda predecir el tipo de clase a la que pertenece. A diferencia de otros algoritmos donde se trabaja sobre un mismo plano, el SVM genera hiperplanos con diferentes números de dimensiones, lo que permite clasificar de mejor manera algunos conjuntos de datos.**

SVM



SVM

- **Algunas de sus aplicaciones son:**
  - **Reconocimiento de escritura**
  - **Clasificación de textos.**
  - **Clasificación de imágenes.**
  - **Clasificación de proteínas.**

KNN

KNN

- **Es un algoritmo de "Aprendizaje Supervisado" que se utiliza principalmente para clasificación y regresión. El nombre completo del algoritmo es "k vecinos más cercanos", aunque es mas conocido por su nombre en inglés "k-nearest neighbors" y para abreviar se usa más KNN (por sus siglas en inglés).**

- **Este algoritmo fue desarrolado por Evelyn Fix y Joseph Hodges en 1951 y no debe de ser confundido con "K-Means", ya que son diferentes.**

KNN

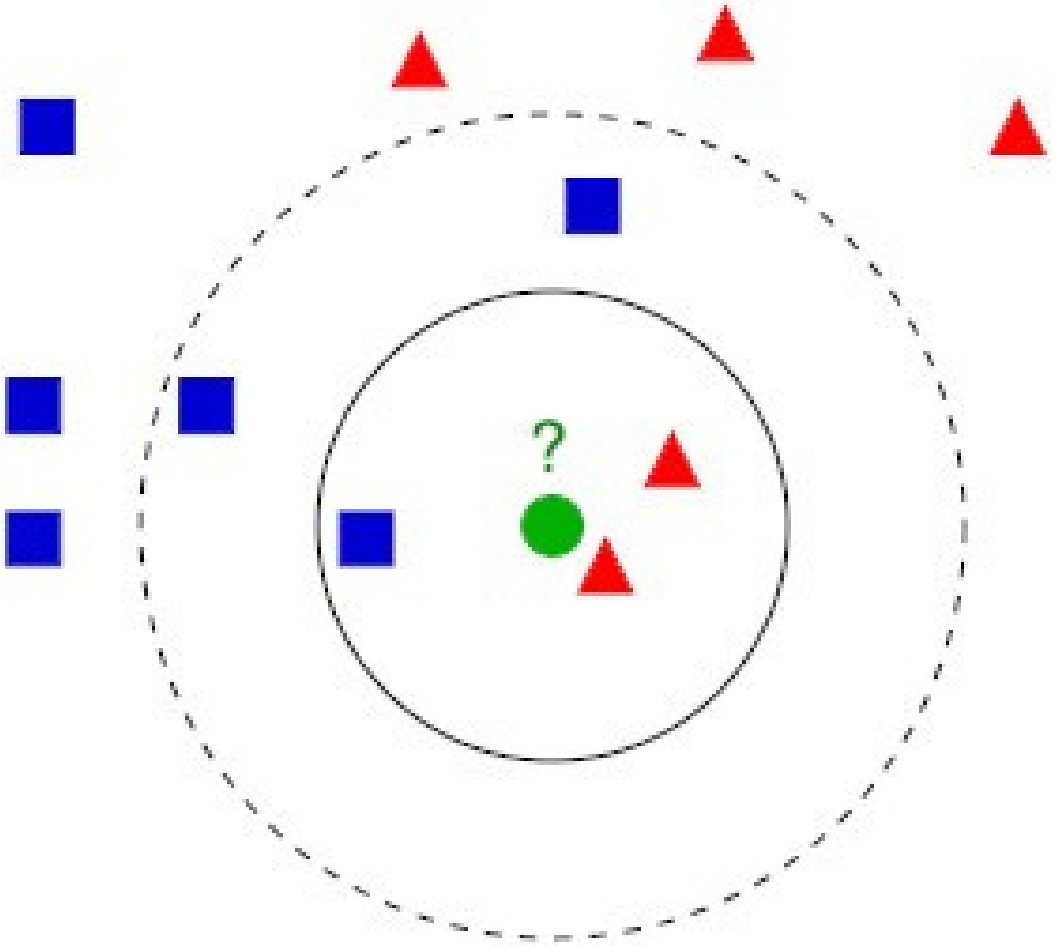
- **En este algoritmo primero se calcula la distancia de todos los datos en entrada por medio de una "función de distancia".**

- **Dado un dato de prueba y dependiendo del objetivo buscado (clasificación ó regresión) se puede hacer lo siguiente:**

  - **Si es "Clasificación" se identifica el dato de entrada con la etiqueta mas frecuente.**
  - **Si es "Regresión" se calcula la salida de acuerdo al valor promedio de los vecinos cercanos.**

KNN

- **En la siguiente figura se observa que el algoritmo KNN tratará de identificar si la figura verde (círculo) pertenece a las figuras rojas (triángulos) ó a las figura azules (cuadros).**



KNN

- **Algunas de sus aplicaciones son:**
  - **Delitos bancarios.**
  - **Análisis en el mercado de valores.**
  - **Detección de huellas digitales.**

Random Forest

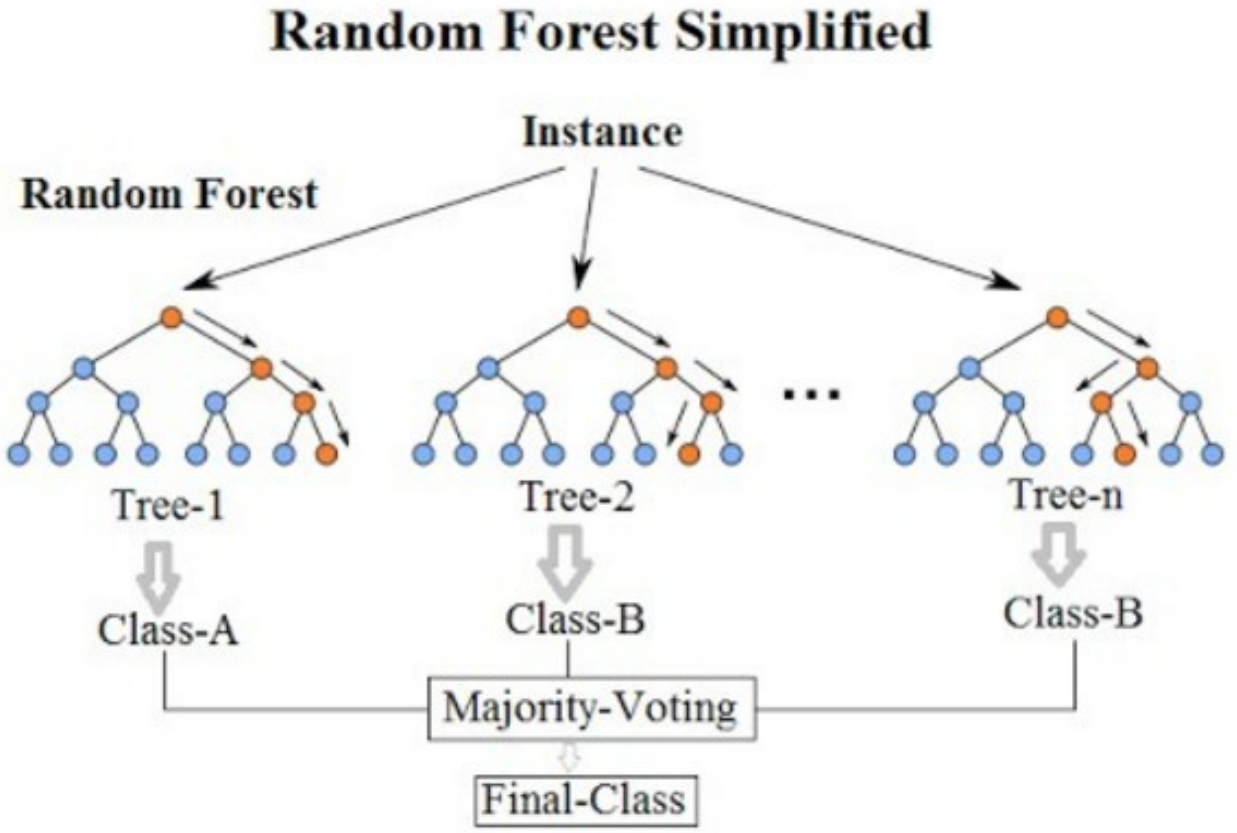
Random Forest

- **Es un algoritmo de "Aprendizaje Supervisado" que se utiliza principalmente para predecir o escoger mejores opciones.**

- **Al igual que el método anterior utiliza árboles de decisión, aunque ahora genera varios árboles con características diferentes para encontrar la mejor opción.**

Random Forest

- **En la siguiente figura se observan 3 árboles de decisión que se generan y de los cuales se seleccionará el que mejores resultados aporte.**



Random Forest

- **Algunas de aplicaciones mas comunes son:**
  - **Identificar enfermedades analizando historiales médicos.**
  - **Detección de fraudes en bancos.**
  - **Estimar pérdidas y ganancias en mercados de valores.**

Redes Neuronales

Redes Neuronales

- **En el proceso para entender el funcionamiento del cerebro humano, se fueron desarrollando investigaciones y teorías acerca de su funcionamiento, dando un primer paso con el desarrollo de un modelo de una neurona artificial.**

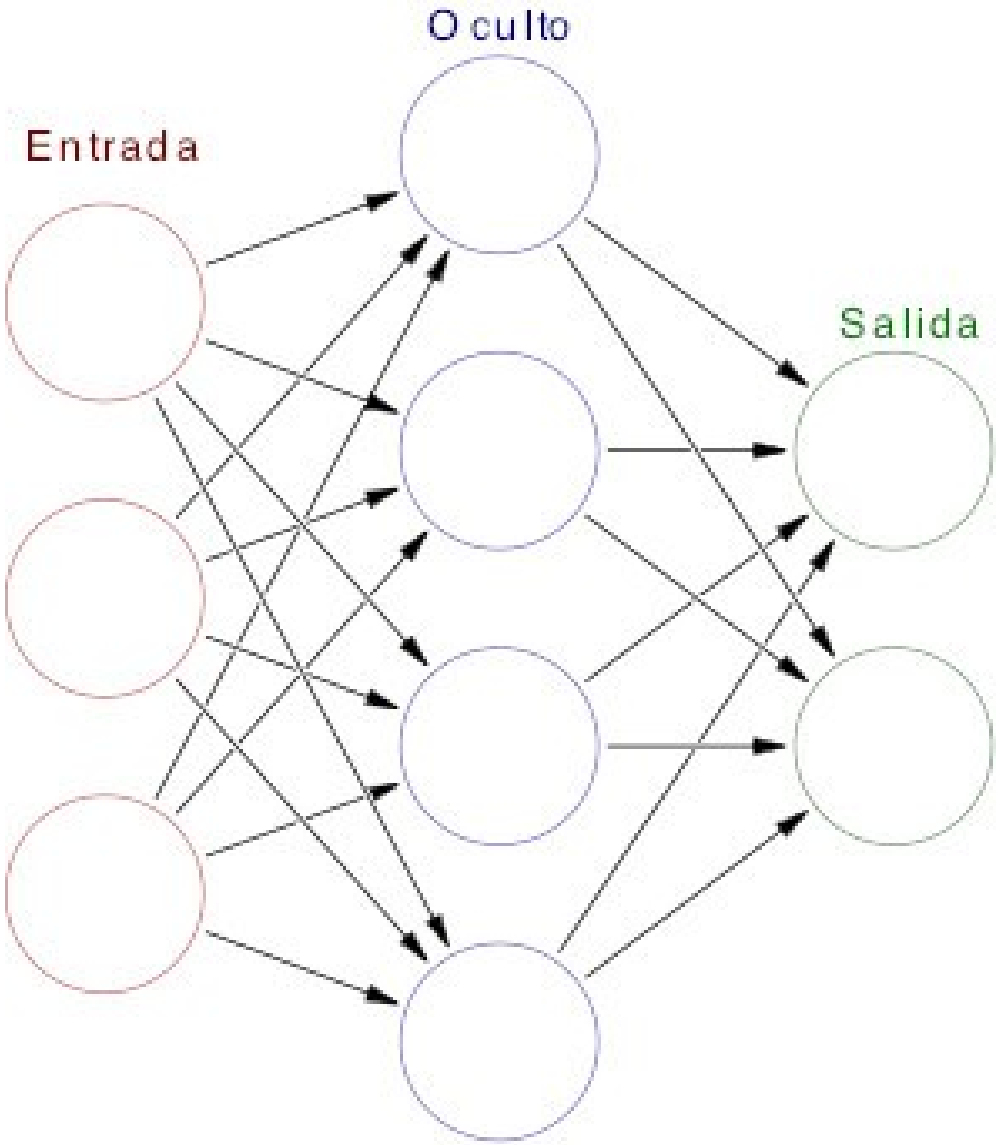
Redes Neuronales

- **Al conectar varias entre ellas se convirtió en lo que conocemos como “Red Neuronal Artificial” ó “Artificial Neural Network” ("ANN" por sus siglas en inglés).**

Redes Neuronales

- **Una Red Neuronal Artificial trata de simular los procesos de sinapsis que ocurren las neuronas del cerebro creando un modelo matemático con entradas conectadas a una primer capa de neuronas, después viene una o varias capas intermedias y finalmente una capa de salida.**

Rede Neuronal



Redes Neuronales

- **La señal de entrada se conecta a la red neuronal y entre cada conexión se va modificando su valor de acuerdo a los pesos que se les otorgue a cada una de las conexiones ó capas involucradas en el proceso. Al final se tiene una capa con valores diferentes en cual se traduce en la respuesta del sistema, en donde se puede colocar por ejemplo una imagen de entrada (por ejemplo de números escritos a mano) y a la salida nos puede indicar (por medio de valores de salida en la capa final) el número que se ha introducido en el sistema.**

Redes Neuronales

- **Para lograr lo anterior y que la respuesta sea adecuada, se debe "entrenar" el sistema con una gran cantidad de datos, es decir, se debe colocar (para el ejemplo anterior de números) una por una las imágenes de muchos números ya etiquetados con su respectiva salida para que el sistema pueda ir calculado el valor de los pesos que se tienen que asignar en cada una de las capas y de las conexiones de las neuronas. Este proceso se repite con cientos o miles de imágenes y por medio de una función de activación calcular su valores, y en caso de existir un error, tratar de minimizarlo al emplear funciones de retro-propagación (“Back Propagation”), hasta que el error desaparezca o al menos disminuya.**

Rogelio Ferreira Escutia

Profesor / Investigador
Tecnológico Nacional de México
Campus Morelia

rogelio.fe@morelia.tecnm.mx

rogeplus@gmail.com

xumarhu.net

@rogeplus

https://www.youtube.com/channel/UC0on88n3LwTKxJb8T09sGjg

rogelioferreiraescutia