Por Oscar García, Prof.
Sistemas Inteligentes

Big Data

Big Data es un término que se refiere a grandes cantidades de información (organizada o no), recopiladas a partir de cualquier dominio de actividad

Big-Data – Conceptos

- Lo importante no son los datos en sí, sino la forma en que se manipulan y analizan. Big data representa los datos que superan las capacidades de procesamiento de los sistemas de bases de datos tradicionales
- Estas cantidades de datos son demasiado grandes, se mueven demasiado rápido o no coinciden con dichas arquitecturas de base de datos. Por lo tanto, son necesarias otras herramientas que procesen datos

Big-Data – Campos de aplicación

El análisis de Big Data se puede encontrar en dominios como el sector bancario y de seguros, la industria de la salud, la educación, las redes sociales y el entretenimiento, las aplicaciones bio-informáticas, las aplicaciones geo-espaciales, la agricultura, etc.

El fenómeno de los grandes datos se puede comprender claramente al conocer las diferentes V asociadas con ellos:

- **Volumen**: denota la gran cantidad de datos producidos cada segundo, oscilando entre terabytes (10^12 bytes) y zettabytes (10^21 bytes). Estos grandes conjuntos de datos se pueden mantener utilizando sistemas distribuidos
- **Velocidad**: este término representa la velocidad a la que se producen y procesan los datos
- **Variedad**: indica la amplia gama de datos que podemos usar
- **Veracidad**: esto habla sobre la calidad de los datos. Es decir, indica los sesgos, el ruido, la anormalidad, etc. en los datos
- **Valor**: señala el valioso conocimiento revelado a partir de los datos

Aprendizaje automático

- El aprendizaje automático es utilizado para analizar los datos, automatizando la construcción de modelos analíticos
- El propósito de los algoritmos de aprendizaje automático es aprender de los datos existentes sin ser programados explícitamente
- Un aspecto importante con respecto al aprendizaje automático es que, cuando los modelos se aplican en nuevos conjuntos de datos, se adaptan independientemente, característica que proviene de la forma iterativa del aprendizaje automático. Estos modelos aprenden de cálculos anteriores para producir decisiones; y resultados ciertos y replicables

Algunos ejemplos de aplicaciones de aprendizaje automático de la vida diaria

- El automóvil autónomo de Google. Estos autos contienen sensores que detectan el objeto desde un área grande en todas las direcciones
- Entre los objetos detectados se encuentran personas a pie, ciclistas y otros vehículos, pero también bolsas de compra de plástico o aves volando. El automóvil está equipado con un software que analiza todos los datos recibidos y los procesa para una navegación segura en la calle
- Los sistemas de recomendación de Amazon y Netflix
- Detección de fraude. Este es uno de los casos de uso más importantes del aprendizaje automático en la actualidad. Tanto el negocio electrónico como el comercio on-line representan un deber desafiante porque es difícil distinguir un límite entre los sistemas para la detección de fraudes y los sistemas para la detección de intrusos en la red y sus límites se solapan

Metodologías de aprendizaje automático

- Hay dos tipos de Metodologías de aprendizaje automático: aprendizaje supervisado y aprendizaje no supervisado
- El aprendizaje supervisado a menudo se usa en problemas en los que los datos deben clasificarse, y se basa en un modelo de clasificación definido, del cual la computadora debe aprender. El aprendizaje supervisado es el método más utilizado para entrenar las redes neuronales o los árboles de decisión
- El aprendizaje no supervisado es más complicado porque la computadora debe aprender cómo realizar una tarea sin tener instrucciones
- Los pasos en el análisis de datos son:
  - Definición del problema
  - Identificación de la/s fuente/s de datos
  - Recopilación y selección de datos
  - Preparación de los datos
  - Construcción del modelo
  - Evaluación del modelo
  - Integración del modelo

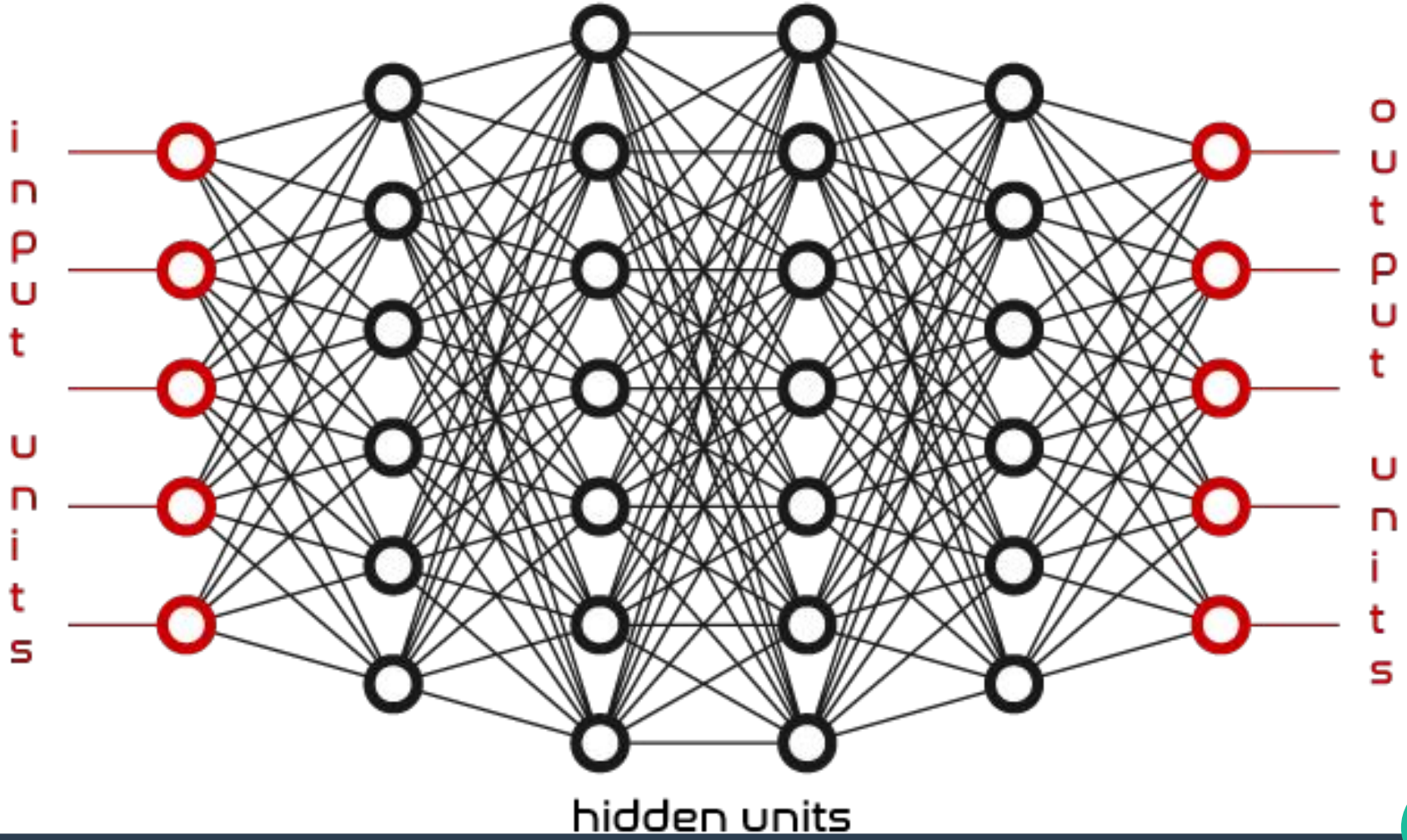
Técnicas de aprendizaje automático usadas en el tratamiento de Big-Data

- Las técnicas de aprendizaje automático más utilizadas en el tratamiento de Big Data son:
  - Redes neuronales artificiales
  - Algoritmos genéticos
  - Análisis de clusters.

    Ejemplos de algoritmos de clustering :
    - k-means - para modelos de centroide
    - Agrupación espacial basada en la densidad de aplicaciones con ruido
  - Árboles de decisión
  - Máquinas de vectores soporte (SVM)
  - Aprendizaje por refuerzo

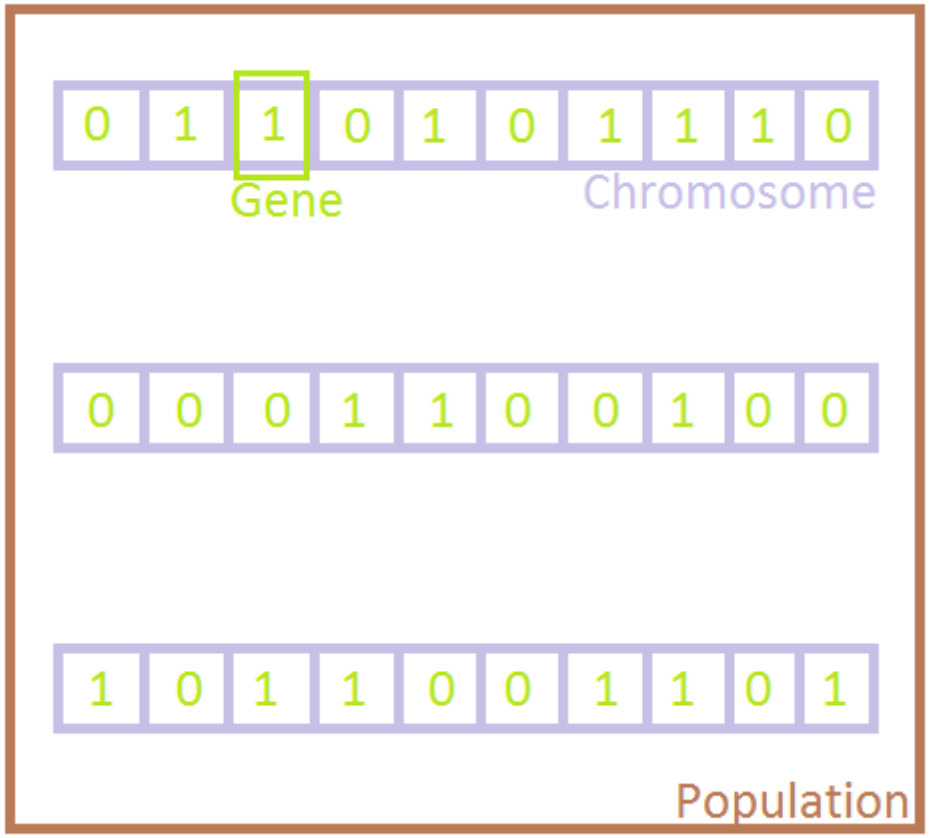
Redes Neuronales

- Red Neuronal Artificial



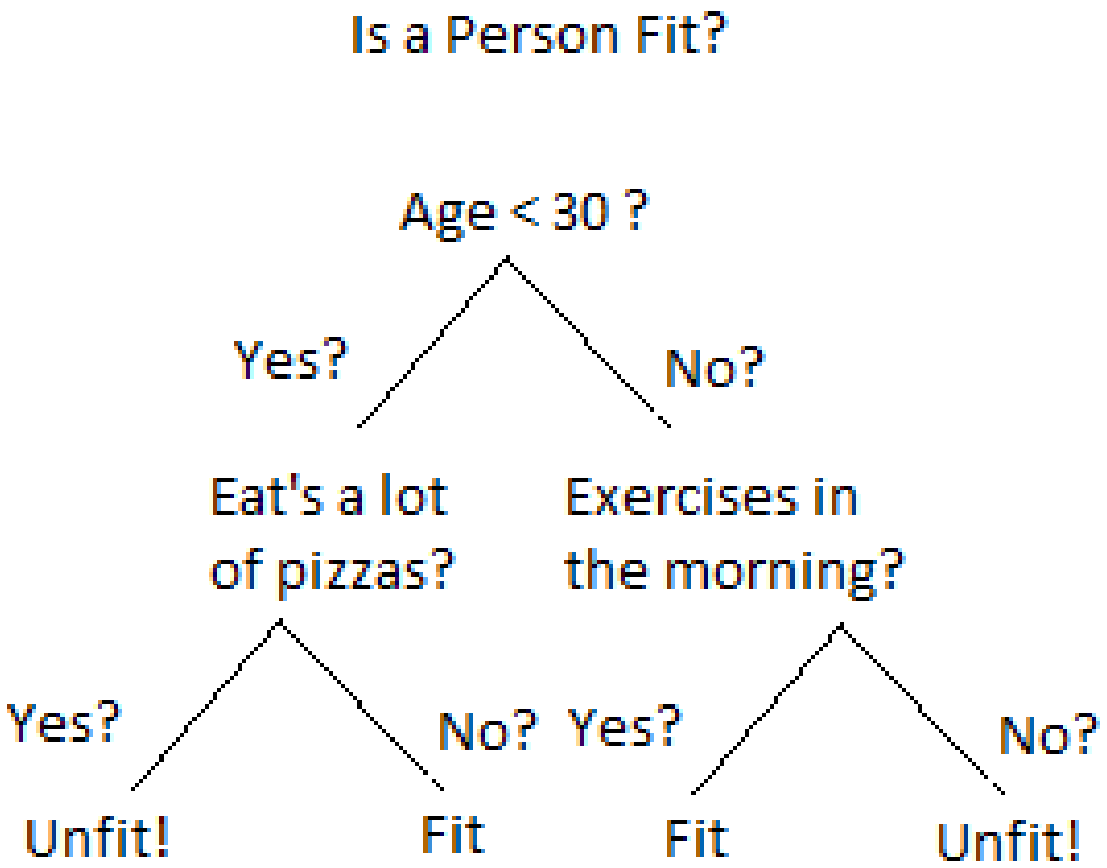
Algoritmos Genéticos

- Algoritmo Genético



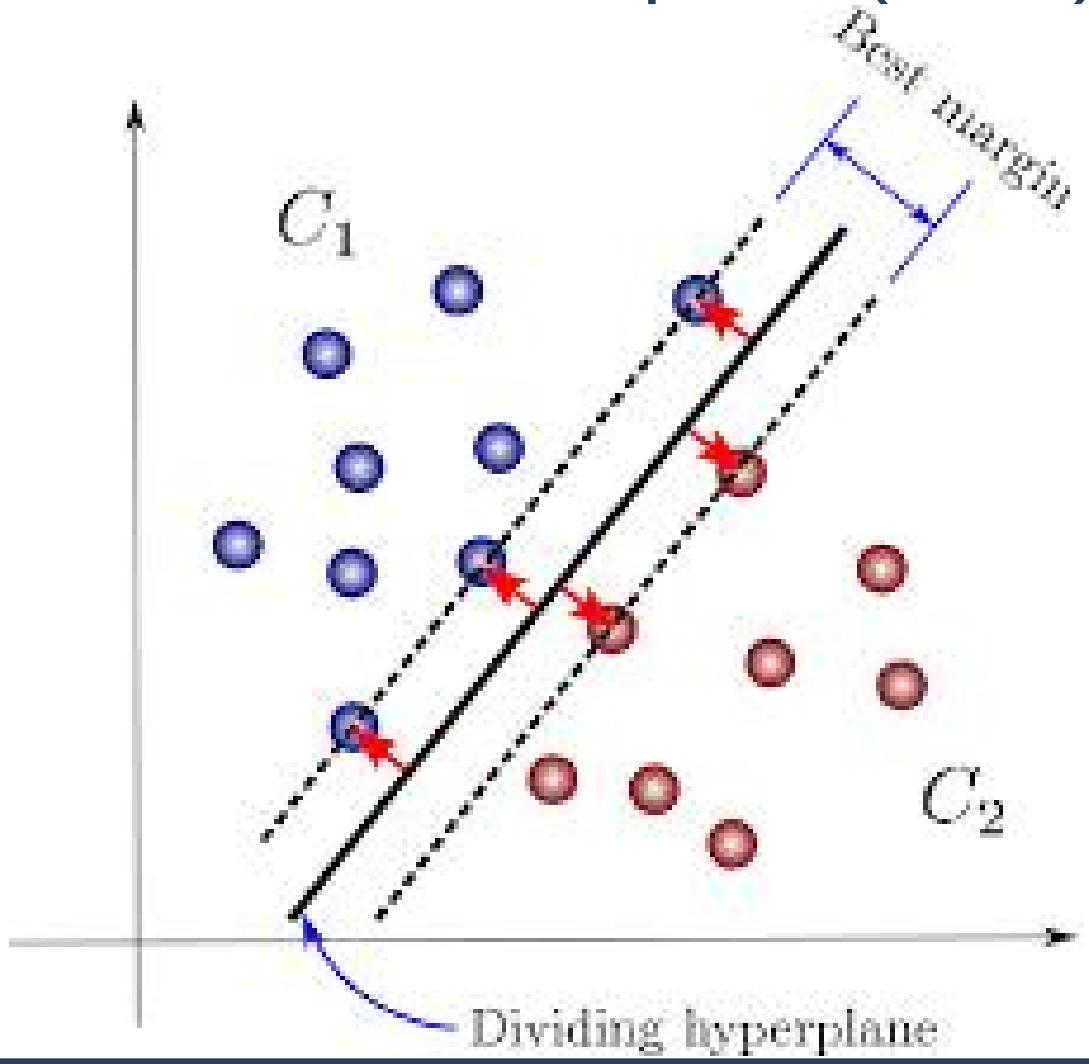
Árboles de decisión

- Árbol de decisión



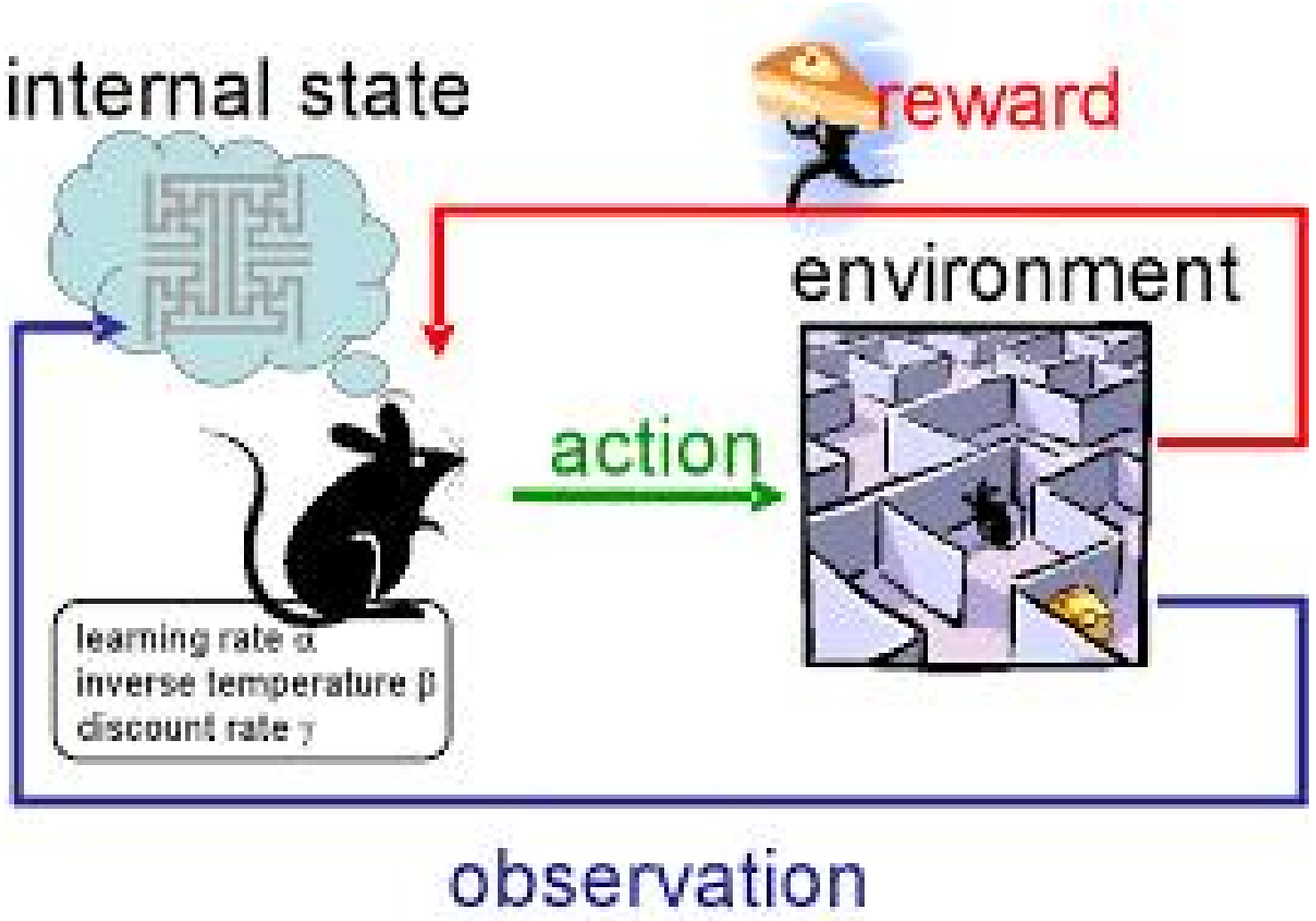
Support Vector Machines

- Máquina de vectores soporte (SVM)



Reinforced Learning

- Aprendizaje por refuerzo



Redes Neuronales y Aprendizaje Profundo

- Desde el punto de vista del procesamiento de datos, para el análisis de datos se prefieren los métodos de aprendizaje supervisados y no supervisados; y las técnicas de refuerzo para los problemas de toma de decisiones [1]
- Las técnicas de aprendizaje profundo utilizan estrategias supervisadas y no supervisadas en arquitectura profunda. Los sistemas de aprendizaje con arquitectura de aprendizaje profundo se componen de varios niveles de etapas de procesamiento no lineal, en los que la salida de cada capa inferior se proporciona como la entrada de la capa superior inmediata. Algunos de los ejemplos son redes neuronales profundas, redes neuronales convolucionales, redes de creencias profundas, redes neuronales recurrentes, y otras. Debido al alto rendimiento de los algoritmos de aprendizaje profundo, resultan muy adecuados para aplicaciones de análisis de Big data
- Google, Amazon, Facebook, Microsoft, Ali Baba – entre otras – son las empresas que mas desarrollan y usan en la actualidad las redes de aprendizaje profundo

Escalabilidad

- Los esquemas tradicionales no pueden procesar los enormes fragmentos de datos dentro de un tiempo estipulado ya que requieren todos los datos en la misma base de datos. Para resolver este problema se ha desarrollado un nuevo campo de aprendizaje automático llamado aprendizaje distribuido. En este esquema, el aprendizaje se lleva a cabo en conjuntos de datos distribuidos entre varias estaciones de trabajo para ampliar el proceso de aprendizaje. Ejemplos de algoritmos de aprendizaje automático distribuidos son las reglas de decisión, la generalización apilada, el meta-aprendizaje y el impulso distribuido
- El aprendizaje automático en paralelo es otro esquema de aprendizaje popular en el que el proceso de aprendizaje se ejecuta en entornos de servidores de múltiples procesadores o en grupos de máquinas – ya sean virtuales o físicas – con múltiples sub-procesos o hilos (threads) [3]

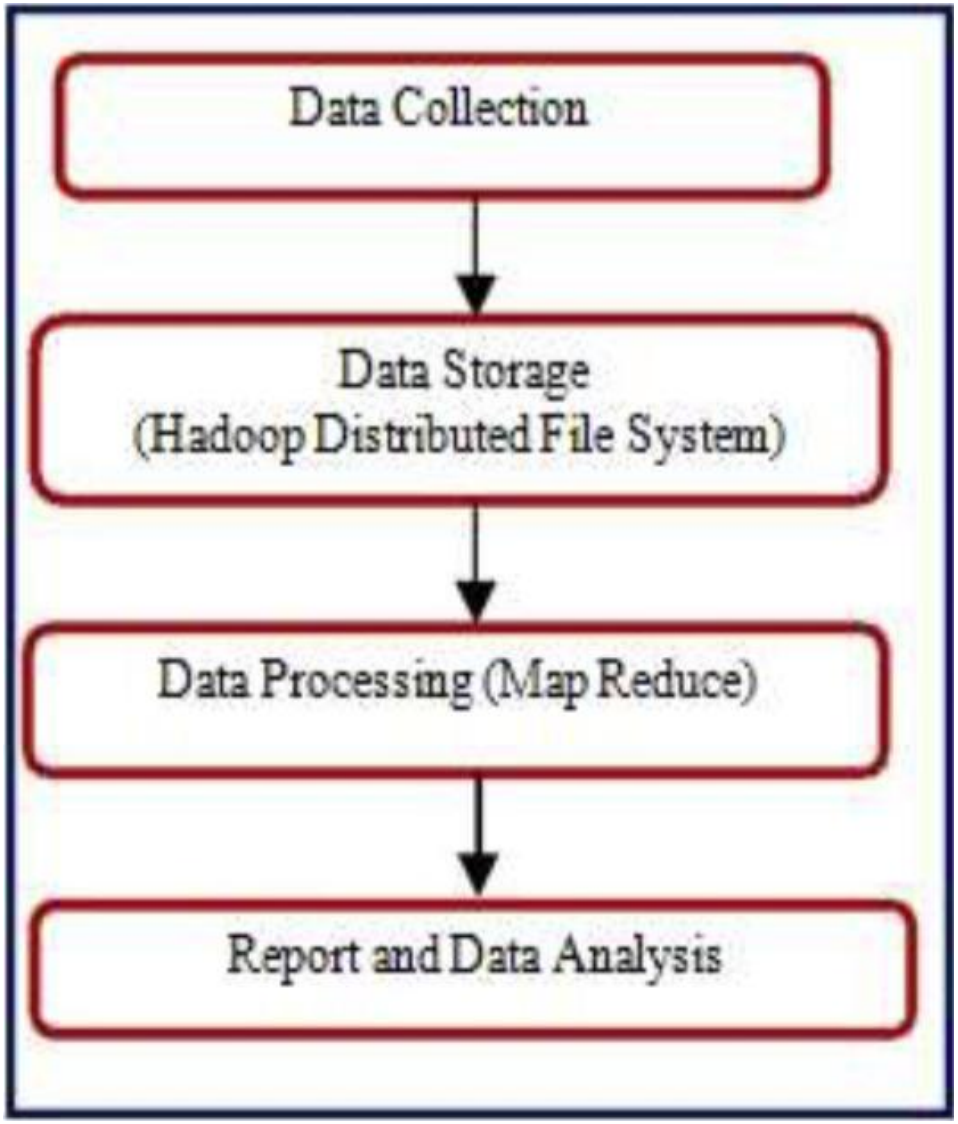
Ejemplo de Deep Learning

Ejemplo de aplicación de aprendizaje automático en el análisis de Big Data: [A Survey on Machine learning assisted Big Data Analysis for Health Care Domain] [4]

- Hay múltiples proyectos en curso dentro del sector de la salud.
- Estos proyectos pueden ayudar a los médicos y el análisis de datos ayuda a predecir enfermedades o los problemas relacionados con la salud
- En este trabajo, se ha recopilado la información relacionada con diferentes procesos de la atención médica
- Como la cantidad de datos relacionados con la atención médica aumenta cada día, y se cree que extraer conocimiento de ella mediante el proceso de análisis de datos es esencial, el análisis de Big Data no es solo una oportunidad sino una necesidad

Arquitectura de Big-Data

- Etapas de Big Data
  - Se usa Hadoop
  - Se usa MapReduce



Big-Data en el ámbito de la Salud

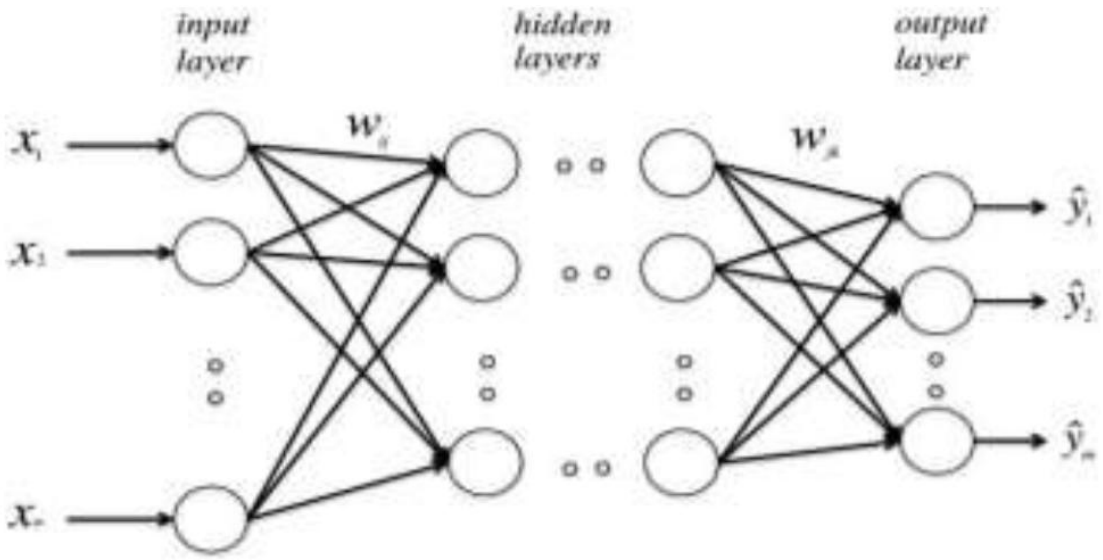
- Cantidades masivas de datos prometen respaldar una amplia gama de funciones médicas y de atención médica, incluido, entre otros, el apoyo a la decisión clínica, la vigilancia de enfermedades y la gestión de la salud de la población
- Por definición, Big-Data en salud se refiere a conjuntos de datos digitales de salud tan voluminosos y complejos que son difíciles (o imposibles) de gestionar con software y / o hardware tradicional

Los Desafíos de los datos masivos en el ámbito del cuidado de la salud

- Extracción de conocimiento de un conjunto de datos complejo y/o no estructurado.
- Comprensión de las notas clínicas no estructuradas en el contexto correcto
- Manejo eficiente de grandes volúmenes de datos de imágenes médicas y extracción de información potencialmente útil y bio-marcadores
  - El análisis de datos genómicos es una tarea computacionalmente intensiva y la combinación con datos clínicos estándar agrega más capas de complejidad aún
  - La plataforma de análisis de Big-Data en atención médica debe admitir las funciones clave necesarias para procesar los datos
  - El análisis de Big-Data en tiempo real es un requisito clave en la asistencia sanitaria
  - Se debe abordar el retraso (o ventana) entre la recopilación y el procesamiento de datos de manera de definir si los resultados pueden obtenerse en tiempo casi real o en diferido, y cual es la diferencia en utilidad (o valor)

Métodos Utilizados

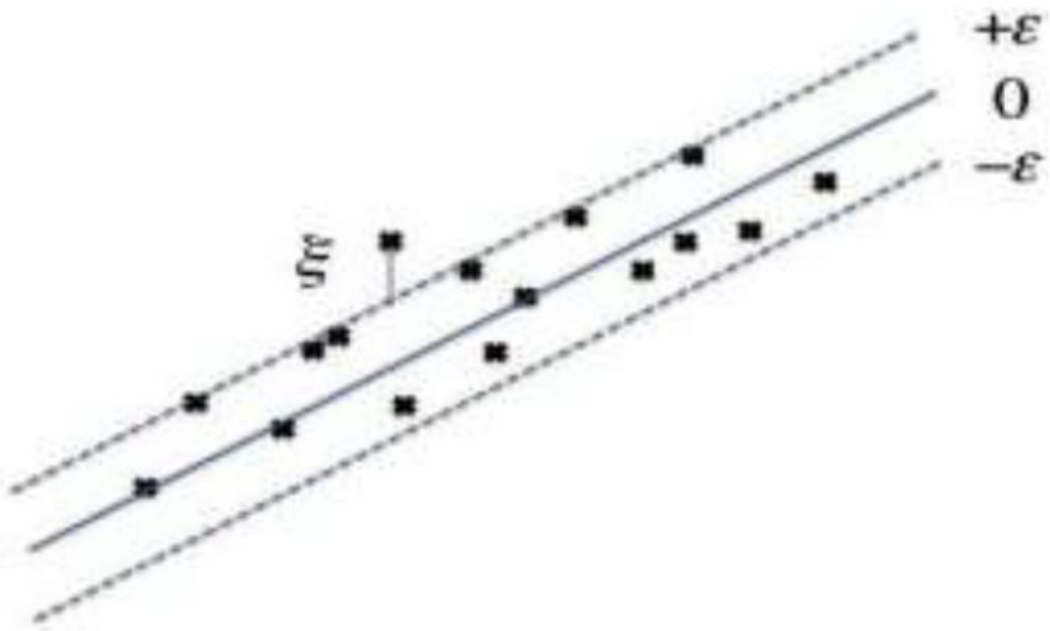
- Perceptrón multicapa (MLP)
  - MLP es un tipo de red neuronal de retroalimentación. Consiste en una capa de entrada, una o más capas ocultas y una capa de salida. Cada capa tiene varias neuronas que conectan una capa con otra
  - Las neuronas en la capa de entrada se corresponden con el número de características de entrada (es decir, rasgos de la enfermedad o síntomas) en el conjunto de datos
  - Cada instancia de datos alimenta a la capa de entrada



Basic structure of MLP

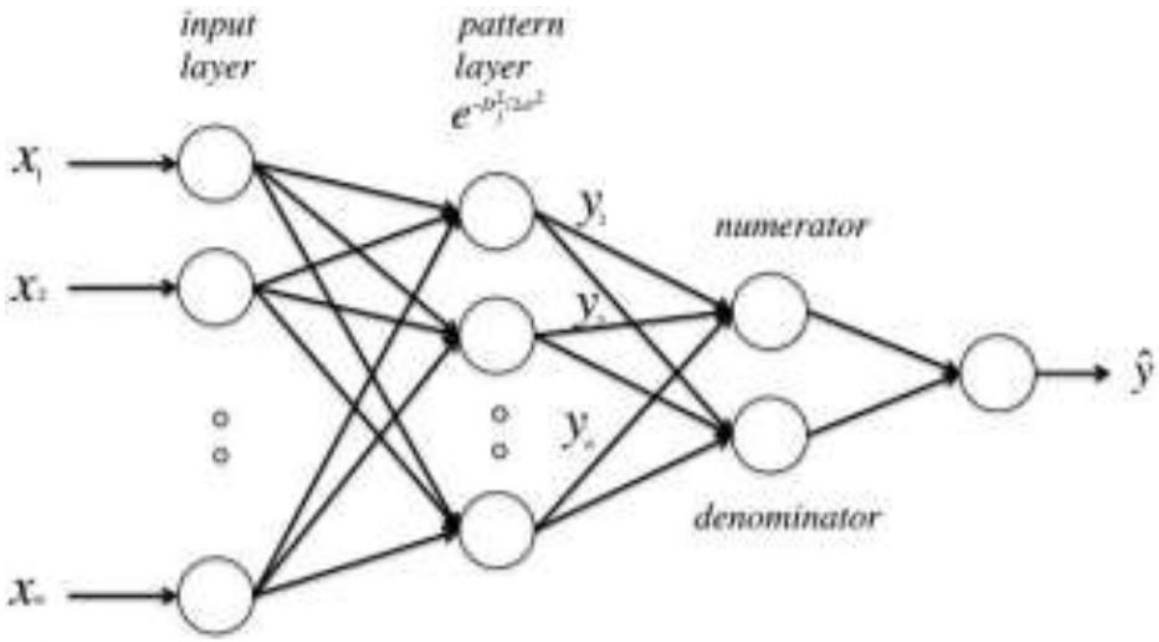
Métodos Utilizados

- Máquina de regresión de vectores soporte (SVR)
  - SVR es una extensión de SVM que se puede aplicar al problema de regresión
  - Intenta minimizar el error entre el objetivo y la predicción, encontrando un hiperplano óptimo de modo que el error de predicción para cada dato de entrenamiento no exceda un umbral



Métodos Utilizados (Continuación)

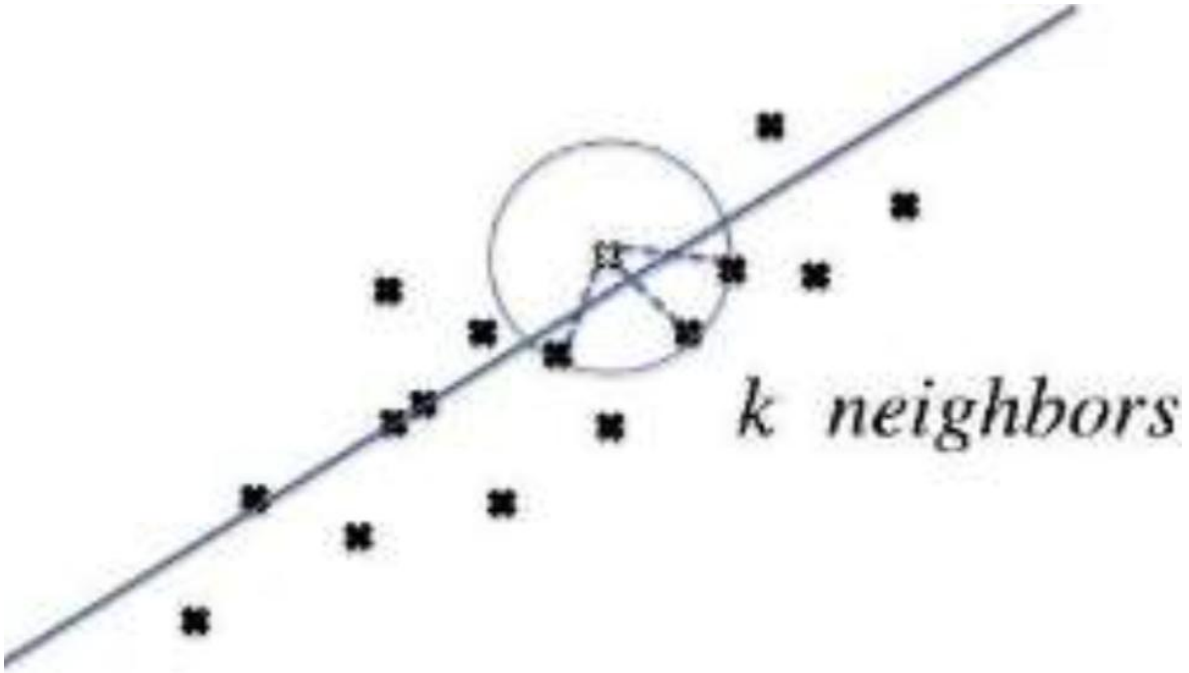
- Red neuronal de regresión generalizada (GRNN)
  - GRNN predice calculando primero la distancia entre la nueva instancia de datos de entrada y las instancias de datos de entrenamiento



Métodos Utilizados (Continuación)

- Regresión de los k vecinos más cercanos (kNN)
  - kNN predice el valor de la función f* (x) dado un nuevo dato de entrada usando los k datos vecinos más cercanos en el conjunto de datos de entrenamiento T
  - kNN permite encontrar la enfermedad relacionada más cercana que ayudará en su predicción
  - k refiere al número de datos vecinos más cercanos
  - En la figura de la derecha:
    - k = 3

Gráfica ejemplo de Regresión de los k vecinos más cercanos (kNN)



Bibliografia

[1] J. Qui, Q. Wu, G. Ding, Y. Xu and S. Feng, “A survey of machine learning for big data processing”, EURASIP Journal on Advances in Signal Processing, Springer, vol. 2016:67, pp. 1-16, 2016. DOI: 10.1186/s13634-016-0355-x

[2] “Parallel machine learning toolbox”, retrieved from http://www.research.ibm.com/haifa/projects/verification/ml_toolbox/

[3] C. A. Caligtan and P. C. Dykes, “Electronic health records and personal health records”, Semin Oncol Nurs, vol. 27, pp. 218-228, 2011

[4] “A Survey on Machine learning assisted Big Data Analysis for Health Care Domain” Publicado en el INTERNATIONAL JOURNAL OF ENGINEERING DEVELOPMENT AND RESEARCH; https://www.ijedr.org