Técnicas Avanzadas en Análisis Estadístico

Este curso se enfoca en la aplicación de técnicas estadísticas avanzadas para transformar la información en conocimiento valioso.

Adriana Arango L adriana.arangol@upb.edu.co



Historia de la Ciencia de Datos

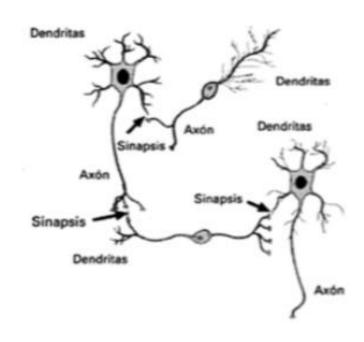
Desde las primeras aplicaciones de la estadística para analizar fenómenos sociales hasta el auge de la informática y el desarrollo de algoritmos de aprendizaje automático, la ciencia de datos ha experimentado una evolución creciente. Esta disciplina ha sido impulsada por la creciente disponibilidad de datos, el desarrollo de tecnologías de análisis avanzadas y la necesidad de tomar decisiones estratégicas basadas en información sólida.

1894

Surgimiento de la neurociencia

- Interes por entender como "computa el cerebro" y como aprende.
- Teoría de plasticidad neuronal



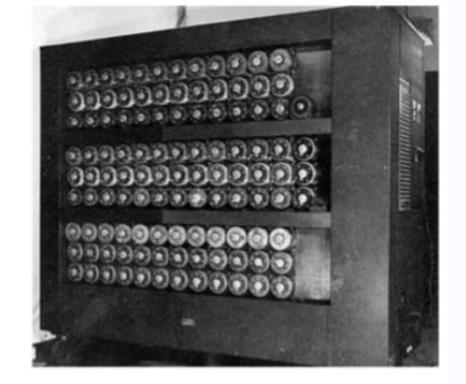


Plasticidad neuronal, neuroplasticidad o plasticidad sináptica o teoría del cerebro plástico

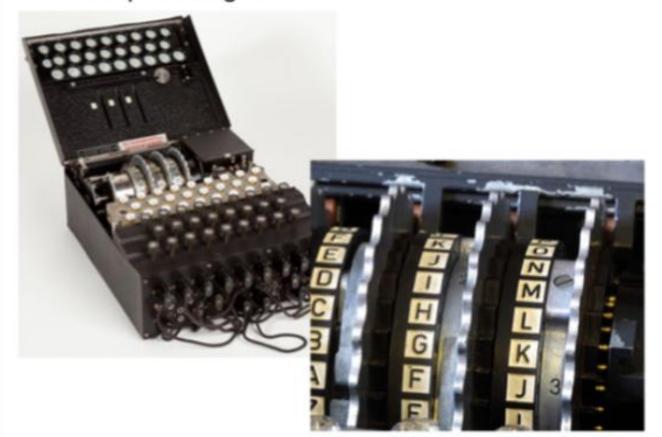
El aprendizaje se da como la modificación de la eficiencia de las conexiones inter-neuronales (y creación de nuevas conexiones) para la transferencia de impulsos electro-químicos modulando tanto la percepción como la respuesta dada ante estímulos del medio.

1940 Bombe Machine

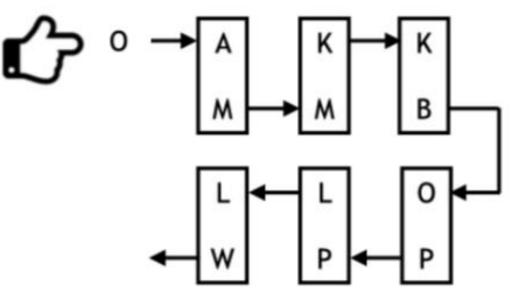
Desarrollada por Alan Turing para decodificar la máquina Enigma.



Máquina Enigma



La máquina enigma usaba tres rotores, de cinco posibles

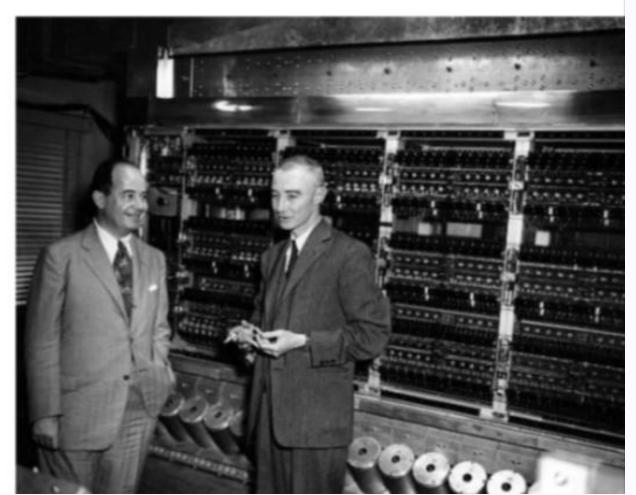


Por cada pulsación se giraban, al menos, un rotor.

1944 Proyecto Manhattan

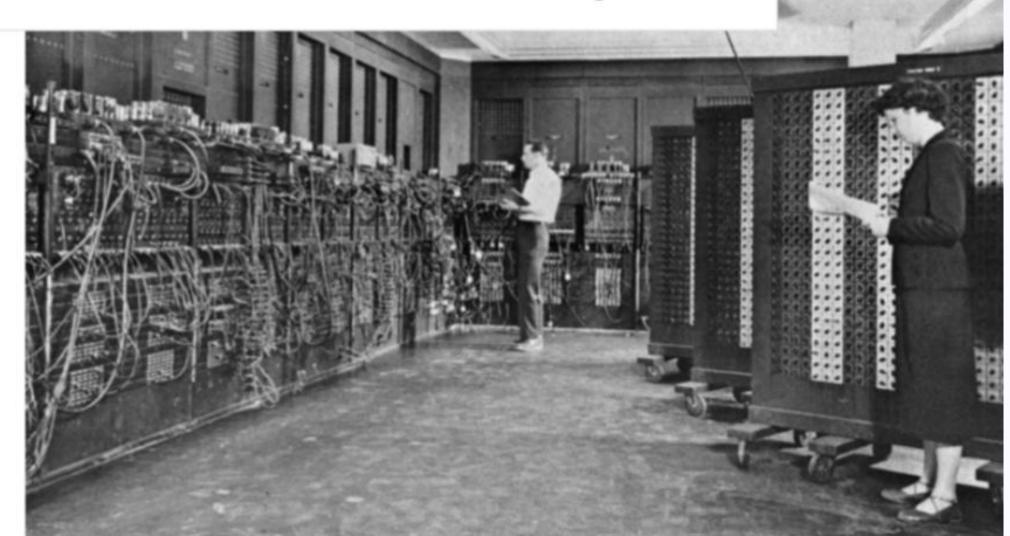
Desarrollo de la simulación numérica (método de Monte Carlo) para comprender y pronosticar el comportamiento de reacciones nucleares en cadena usando computadores análogos.

Se basa en la realización de miles de simulaciones de un proceso aleatorio para analizar posteriormente sus resultados.



ENIAC (Electronic Numerical Integrator and Computer)

- Desarrollado en la Universidad de Pennsylvania.
- · Su objetivo fue calcular tablas para fuego de artillería.
- Fue capaz de manejar 50.000 instrucciones por segundo. Comparativamente un iPhone puede manejar 5 billones. 20 horas de cálculo manual se reducían a 30 segundos



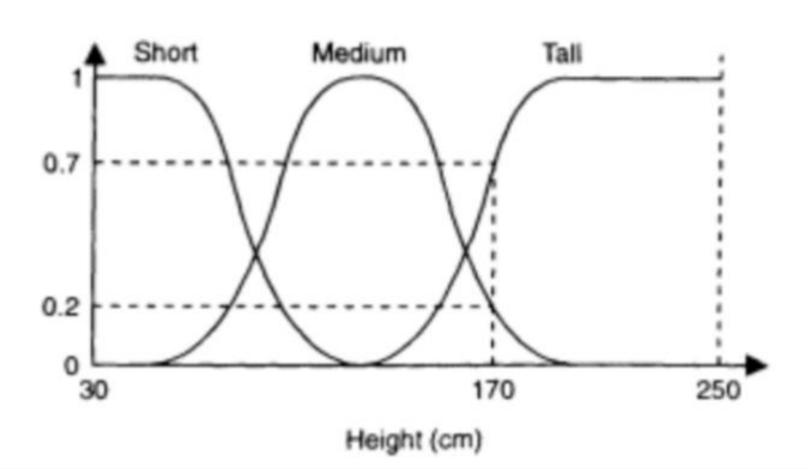
1965 Concepto de Fuzzy Set

Definición tradicional. Función característica:

- $\mu_A(u) = 1$, El elemento u pertence al conjunto A
- $\mu_A(u) = 0$, El elemento u no pertence al conjunto A

Conjunto difuso. Se admite la pertenencia parcial de un elemento a un conjunto.

$$\mu_A(u) \rightarrow [0,1]$$

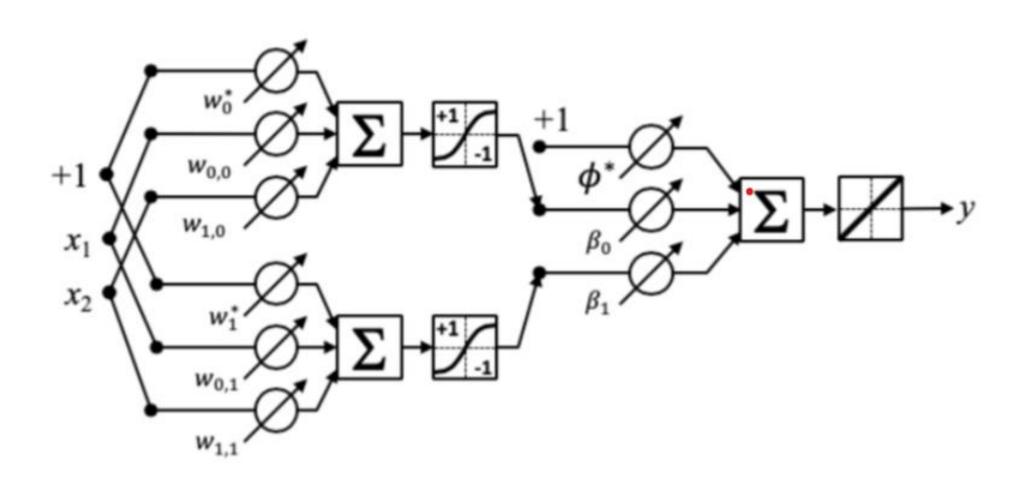


1974

Backpropagation o Regla Delta generalizada

Tesis doctoral de P. Werbos. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. Harvard, Cambridge, MA, August 1974.

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \mu \frac{\partial}{\partial \mathbf{w}(k)} \left[e^2(k) \right]$$

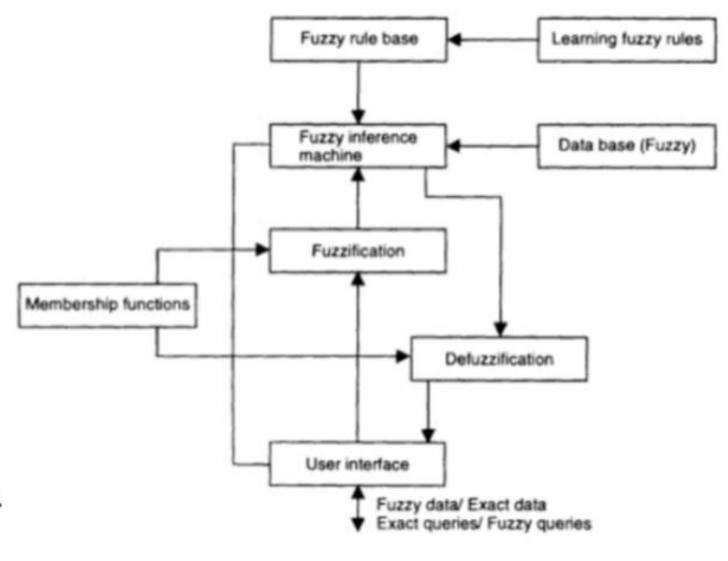


1980 Sistemas difusos de inferencia

Sistemas de representación de datos inexactos y conocimiento heurístico usando conjunto borrosos y reglas difusas en lugar de exactas.

Componentes:

- Variables difusas
- Fuzzificator: convierte entradas precisas en valores de pertenencia
- Reglas difusas: Obtenidas de especialistas o datos numéricos
- Motor de Inferencia Difusa:
- Convierte conjuntos fuzzy en conjuntos fuzzy.
- Determina como las reglas son activadas o combinadas.
- El resultado es un conjunto fuzzy
- Defuzzificator: Convierte el conjunto fuzzy de resultado en un valor preciso

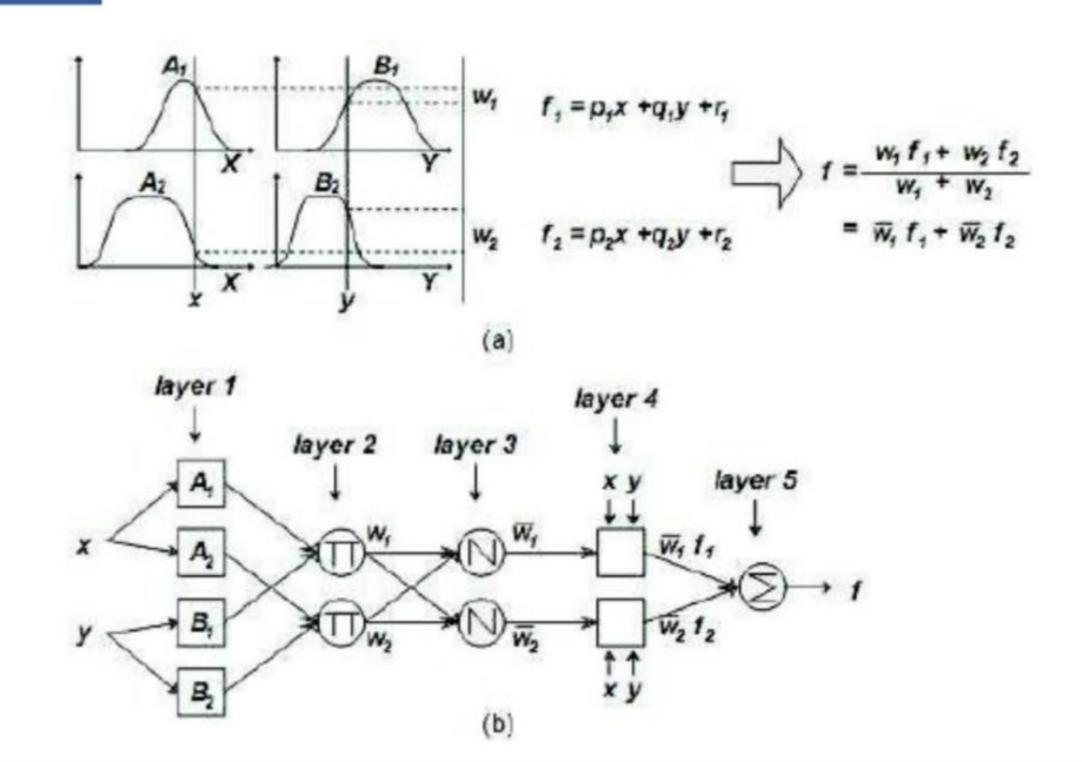


Ventajas:

- Facilidad de implementación
- Mantenimiento barato
- Robustos

1991

Adaptive Neurofuzzy Inference System



1995 Ingenieria del Conocimiento

La solución es dependiente del problema y del conocimiento disponible.

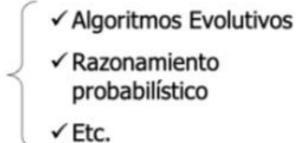
- Métodos Estadísticos
- Sistemas simbólicos basados en reglas
- Sistemas difusos
- Redes Neuronales
- Computación evolutiva
- Enjambres de partículas

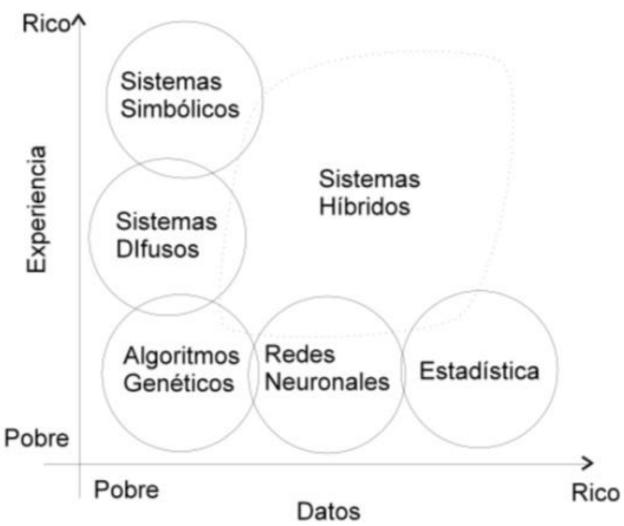
• ...

Sistemas que combinan paradigmas

- Sistemas basados en reglas
- Sistemas difusos
- Redes Neuronales

Otros





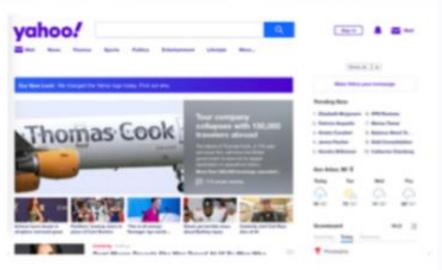
1995 Explosión de la Web

- De uso científico y gubernamental a comercial y de consumo
- Creación de Amazon en 1994, y venta de su primer libro en 1995.
- Creación de Yahoo! en 1994.
- Creación de eBay y AltaVista en 1995
- Creación de google en 1996 (algoritmo PageRank)

Precursores en la Personalización de la Experiencia online!







Comprensión del negocio

Comprensión de los datos Preparación de los datos

Modelado

Evaluación

Distribución

Resultados del proyecto

Establecer objetivos Formulación del plan preliminar Criterios de éxito

Valoración de la situación actual

Inventario de recursos Requerimientos, hipótesis y restricciones Riesgos y contingencias Terminología Costos y beneficios

Objetivos de la minería de datos

Objetivos de la DM Criterios de éxito desde el negocio Criterios de éxito desde la DM

Plan del proyecto

Desarrollo del plan de proyecto Selección de herramientas y técnicas

Recopilación de datos iniciales

Datos existentes Datos adquiridos Datos adicionales

Descripción de los datos

Cantidad de los datos Tipos de valores Esquema de codificación

Exploración de datos

Análisis exploratorio

Verificación de calidad de datos

Datos perdidos Errores de datos Errores de medición Incoherencias de codificación Metadatos erróneos

Selección de datos

Selección de elementos (filas) Selección de atributos (columnas)

Limpieza de datos

Datos perdidos Errores de datos Errores de medición Incoherencias de codificación Metadatos erróneos

Construcción de nuevos datos

Derivación de atributos (Columnas) Generación de registros (filas)

Integración de datos

Fusión (columnas) Adición (filas)

Formato de datos

Comprensión del negocio

Comprensión de los datos Preparación de los datos

Modelado

Evaluación

Distribución

Selección de la técnica (factores)

Datos disponibles Objetivos de la DM Requisitos específicos

Generación de un diseño de comprobación

Criterios de bondad Definición de datos de testing

Generación de modelos

Configuración de parámetros Listado de modelos producidos Descripción de resultados

Evaluación del modelo

Evaluación global Ajuste de parametros

Evaluación de resultados

Se cumplen los criterios?

Proceso de revisión

Valor de los resultados de cada fase Mejoras posibles Errores o fallos cometidos Callejones sin salida Sorpresas Decisiones alternativas

Determinación del siguiente paso

Continuar con la fase de desarrollo Refinar o sustituir modelos

Planificación de la distribución

Resultados, modelos y descubrimientos Selección de modelos a distribuir Plan de difusión de los descubrimientos Plan de distribución de resultados

Planificación del control y mantenimiento

Factores o influencias a controlar Medición y control de la validez y precisión Desecho de modelos Reuso de modelos en problemas similares

Creación del informe final

Revisión final del proyecto

1996

Data Science



Programación



Visualización de Datos



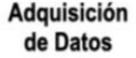
Aprendizaje de Máquinas

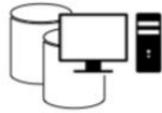


Inferencia Estadística



Limpieza de Datos





Computación Reproducible



Modelos Estadísticos



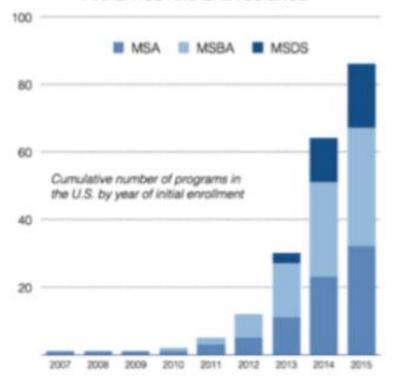
Productos de Datos



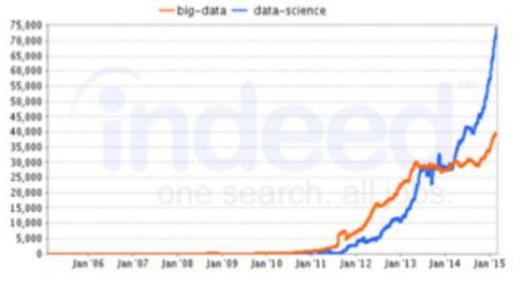
Experticia en el problema



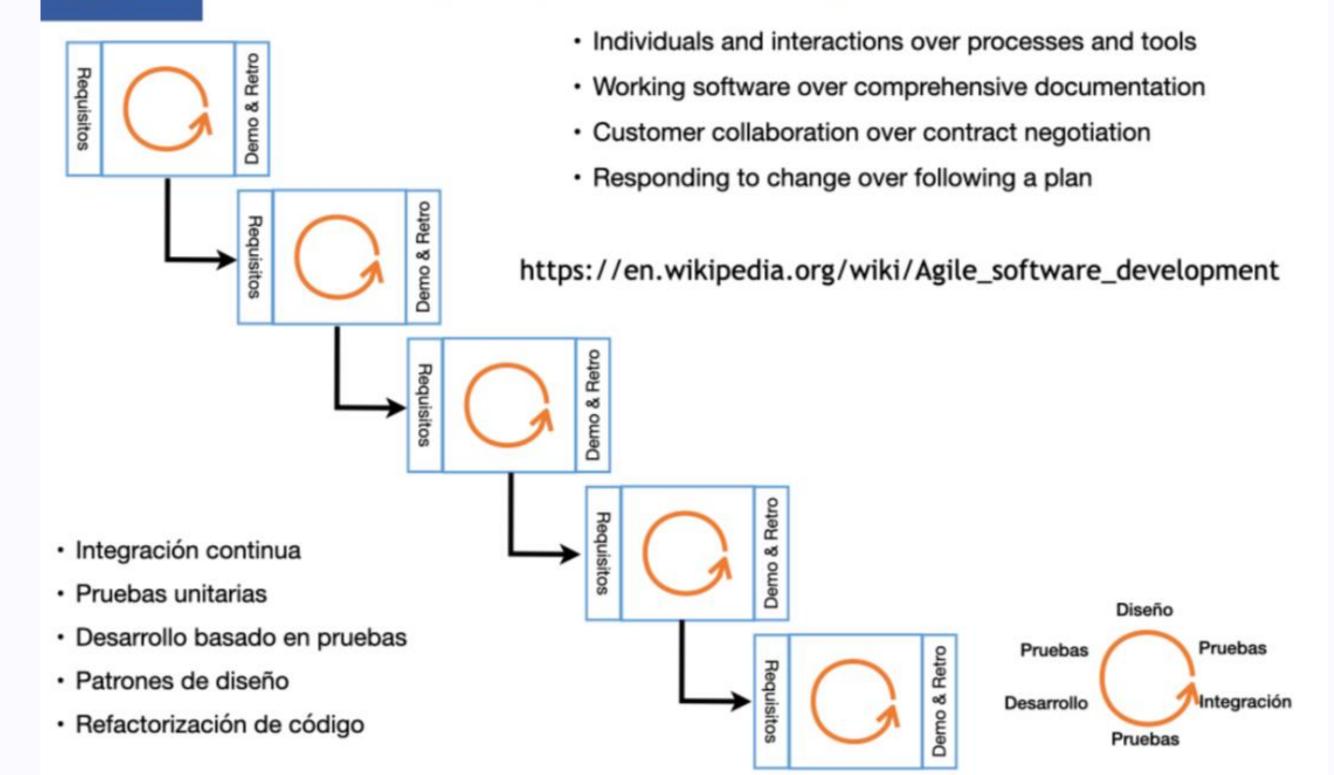
GROWTH OF MASTER'S DEGREE PROGRAMS IN ANALYTICS AND DATA SCIENCE







Métodos Agiles y manifiesto ágil



Estadística Descriptiva

Aprendizaje Estadístico

Conjunto de herramientas fundamentadas en conceptos estadísticos para la comprensión de conjuntos de datos con el fin de modelar y predecir.

Aprendizaje de Máquinas•

Área de las ciencias de la computación orientada al desarrollo de sistemas inteligentes (solución de problemas como lo haría un experto humano).

Investigación de operaciones

Inteligencia de negocios .

Área enfocada al uso de técnicas estadísticas descriptivas y KPI para analizar la situación actual

Modelado Predictivo •

Área enfocada al uso de técnicas estadísticas utilizadas para pronosticar resultados (de un proceso).

Minería de Datos

Descubrimiento de patrones y conocimiento de grandes conjuntos de datos usando técnicas estadísticas, aprendizaje de máquinas y herramientas de bases de datos.

Analítica Descriptiva / Diagnostica

Análisis de la situación actual para la toma de decisiones operativas

Analítica Predictiva

Área enfocada al uso de técnicas de modelado predictivo, aprendizaje de máquinas y minería de datos para pronosticar los resultados de un proceso en un contexto organizacional.

Analítica Prescriptiva

Uso de técnicas de simulación, optimización y análisis de riesgo e incertidumbre para la toma de decisiones organizacionales

2006 Analítica Predictiva

Metodologías

Regresión lineal
Regresión logística
Agrupamiento
Árboles de decisión
Inteligencia Artificial/Aprendizaje de máquinas
Redes neuronales y aprendizaje profundo
Procesamiento del lenguaje natural

Tipos de modelos

Pronóstico Clasificación Outliers Series de tiempo Agrupamiento

Aplicaciones

Detección de fraude
Optimización de campañas de marketing
Mejora de operaciones (inventario y manejo de recursos)
Reducción de riesgo
Mantenimiento/abandono/captura de clientes
Análisis del comportamiento
Venta cruzada
Valoración

Aprendizaje de Máquinas

Aprendizaje (autónomo) a partir de los datos disponibles

Analítica predictiva

Uso de datos históricos para construir modelos matemáticos que capturan tendencias para pronosticar eventos futuros

REPORTE: ¿Qué pasó?

Herramientas de búsqueda, reportaría y consulta.

ANALÍSIS: ¿Por qué pasó?

OLAP y herramientas de visualización

MONITOREO: ¿Que está pasando ahora?

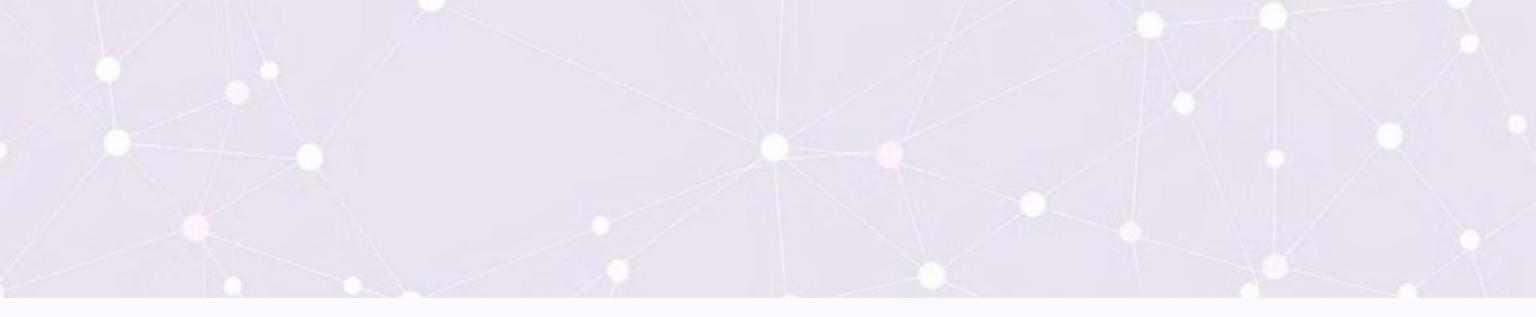
Dashboards, scoreboards

PREDICCIÓN: ¿Qué podría pasar?

Analítica predictiva

Planeamiento de la demanda Servicio al cliente Mejora de la calidad Presupuesto Gestión de personal Retail

Lenguaje natural



Técnicas Avanzadas en Análisis Estadístico

1 Análisis Multivariante

Explorar relaciones complejas entre múltiples variables.

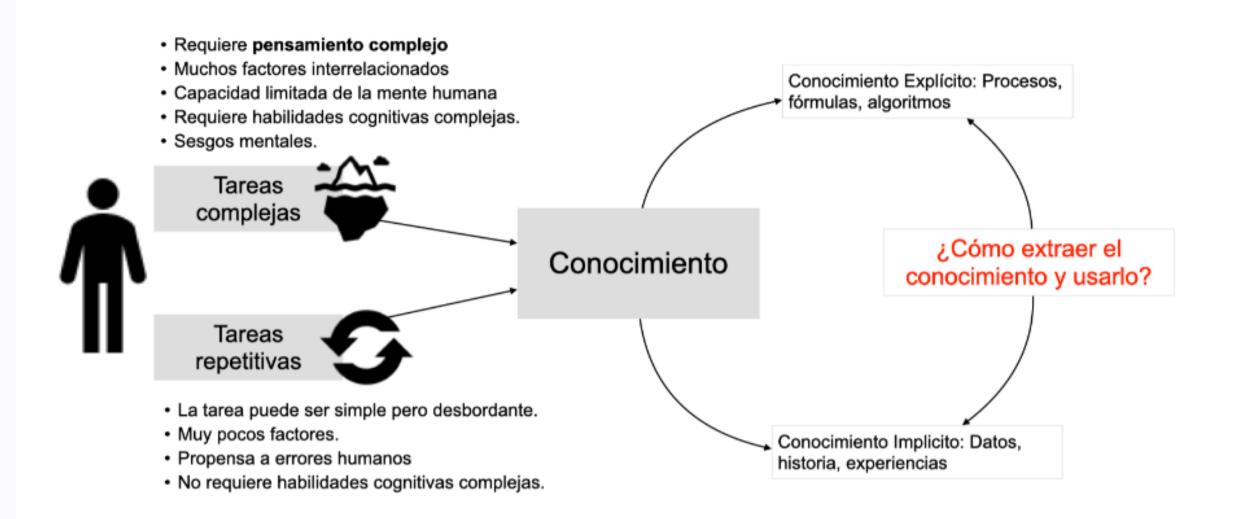
9 Procesos Estocásticos

Modelar fenómenos aleatorios que cambian con el tiempo.

3 Series de Tiempo

Analizar datos secuenciales para identificar patrones y realizar pronósticos.

¿Por qué se deben tomar decisiones basadas en datos?



La Importancia de la Ciencia de Datos

Toma de Decisiones

Los datos permiten a las empresas tomar decisiones más informadas y estratégicas.

Innovación

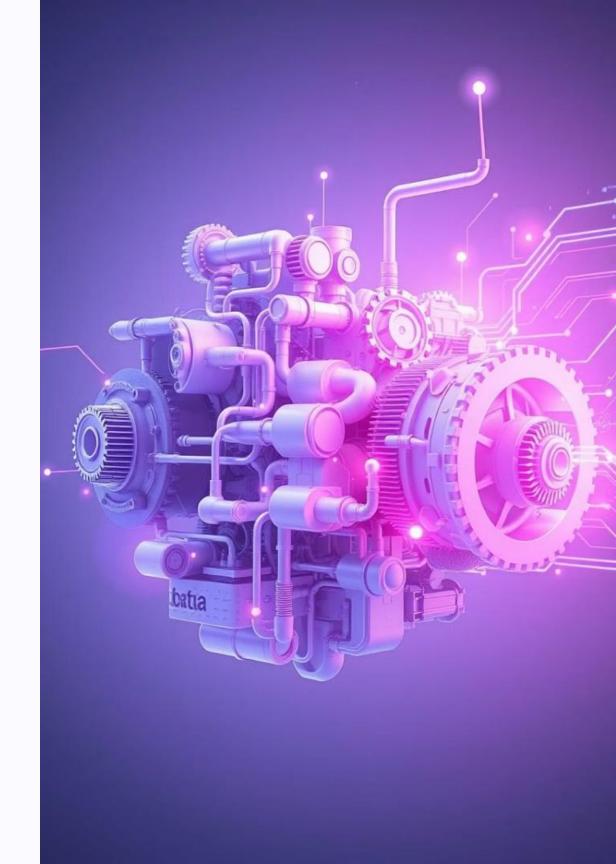
Los datos impulsan la innovación y el desarrollo de nuevos productos y servicios.

Eficiencia

Los datos permiten optimizar procesos y mejorar la eficiencia operativa.

Experiencia del Cliente

Los datos permiten
personalizar la experiencia del
cliente y ofrecer productos y
servicios más relevantes.



Análisis Multivariante: Conceptos y Aplicaciones





Agrupar datos en grupos basados en similitudes.



Análisis de Regresión

Modelar la relación entre variables para predecir resultados.



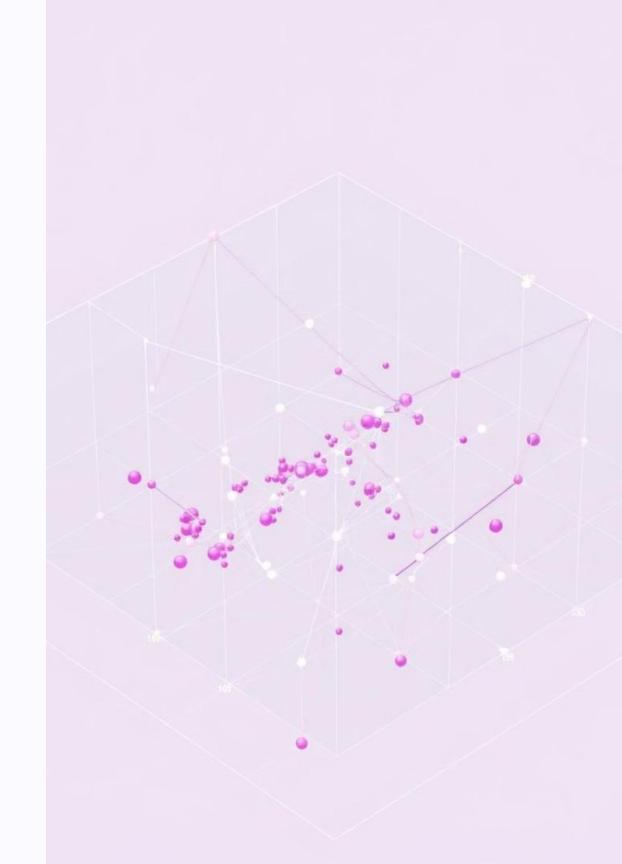
Análisis de la Varianza (ANOVA)

Comparar grupos de datos para determinar diferencias significativas.

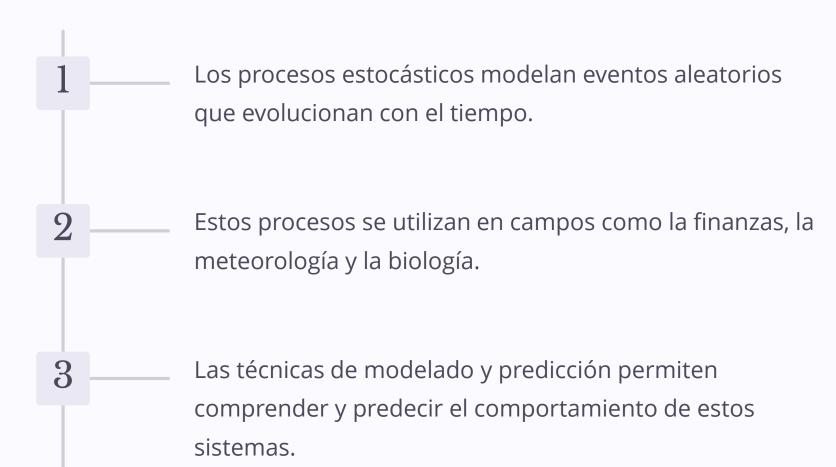


Análisis de Componentes Principales (PCA)

Reducir la dimensionalidad de los datos para simplificar el análisis.



Procesos Estocásticos: Modelación y Predicción



Series de Tiempo: Análisis y Pronósticos

1

Identificar patrones y tendencias en datos secuenciales.

2

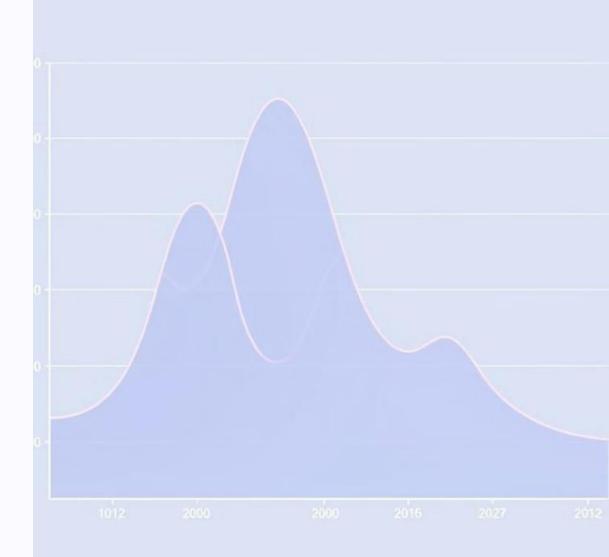
Analizar la estacionalidad, la tendencia y la aleatoriedad.

3

Realizar pronósticos de eventos futuros basados en datos históricos.

4

Optimizar procesos, planificar recursos y tomar decisiones estratégicas.



Habilidades en Ciencia de Datos: Objetivos de Aprendizaje

1	Dominio Estadístico
2	Programación
3	Machine Learning
4	Comunicación
5	Visualización

Casos de Éxito y Ejemplos Prácticos

Salud Análisis de datos para mejorar el diagnóstico y la atención médica. Finanzas 2 Detección de fraudes, gestión de riesgos y predicción de mercados. Marketing 3 Segmentación de clientes, personalización de campañas y

análisis de la experiencia del cliente.



Conclusiones y Próximos Pasos

1

2

3

Desbloquea el potencial de los datos

La ciencia de datos es una herramienta poderosa para resolver problemas y crear valor.

Amplía tus habilidades

El dominio de las técnicas estadísticas avanzadas es fundamental para el éxito en la era de los datos. Aplica el conocimiento

Utiliza las habilidades aprendidas para mejorar la toma de decisiones.