Ciencia de Datos Aplicada a Ciencias de la Tierra

Tema: Introducción

Martín A. Díaz-Viera[1], Farhid M. Elisea Guerrero[2]

[1)] *mdiazv@imp.mx*, [2)] *felisea@imp.mx*

2 de febrero de 2024

Contenido I

Contenido II

Motivación

- En años recientes se han popularizado los términos *Ciencia de Datos*, *Aprendizaje Automatizado* (Machine Learming) e *Inteligencia Artificial*.
- Anteriormente se manejaban términos como: Big Data, Data Mining y Soft Computing.
- Esto ha llevado a que sean confusos los límites y alcances de estos términos.

Motivación

- Es frecuente encontrar que se aplique los métodos de Aprendizaje Automatizado (ML) sin realizar previamente algún Análisis Estadístico de los Datos.
- Métodos estadísticos bien establecidos como Análisis de Regresión o de Agrupamiento se mezclan con los métodos de Aprendizaje Automatizado (ML) pero se aplican de manera mecánica, perdiéndose la parte correspondiente al análisis.

Objectivos

- Establecer qué relación existe entre estos términos.
- Revisar un Flujo de Trabajo de ML en Geociencias.
- Mostrar la importancia del Análisis Estadístico de los Datos en la aplicación de ML.

Relación entre IA, ML, DL, DS, BD

Relación entre

IA, ML, DL, DS, BD

Inteligencia Artificial (IA)

- Es la inteligencia demostrada por las máquinas, a diferencia de la inteligencia natural de los animales y los humanos.
- Anteriormente para describir máquinas que imitan habilidades cognitivas "humanas", como "aprender" y "resolver problemas".
- Ahora describen la IA en términos de racionalidad y actuación racional.
- Cualquier sistema que percibe su entorno y realiza acciones que maximizan sus posibilidades de lograr sus objetivos.

Inteligencia Artificial (IA)

- **Objetivos**: razonamiento, resolución de problemas, representación del conocimiento,planificación, aprendizaje, procesamiento del lenguaje natural, percepción, movimiento y manipulación, inteligencia social, inteligencia general.
- **Herramientas**: búsqueda y optimización, lógica, métodos probabilísticos para razonamiento incierto, clasificadores y métodos de aprendizaje estadístico, redes neuronales artificiales (aprendizaje profundo), lenguajes y hardware especializados.

Aprendizaje Automatizado (ML)

- El término Machine Learning (ML) fue acuñado por Arthur Samuel en 1959 [5].
- ML es un subconjunto de AI.
- Se utiliza en escenarios en los que se necesita que las máquinas aprendan de grandes volúmenes de datos.
- El conocimiento así adquirido se aplica a un nuevo conjunto de datos.
- ML le da a una máquina la capacidad de aprender de (o acerca de) conjuntos de datos más nuevos sin dar instrucciones explícitas.

Aprendizaje Automatizado (ML)

- Algunos de los métodos más comunes implementados para "hacer que las máquinas aprendan"son:
  - Aprendizaje supervisado
  - Aprendizaje no supervisado
  - Aprendizaje automático reforzado

Aprendizaje Profundo (DL)

- DL entra en juego cuando ML no puede ofrecer completamente los resultados deseados, porque los datos son enormes, tienen demasiadas variables, se requiere de precisión extremadamente alta.
- DL es más difícil de implementar, ya que requiere hardware especializado (por ejemplo, GPU's) para ejecutarse y más tiempo para entrenar el modelo.
- Siri, Alexa o Google Assistant son algunas aplicaciones que usan DL para comprender tus solicitudes.

Minería de Datos (MD)

- La Minería de Datos (MD) o exploración de datos (es la etapa de análisis de "knowledge discovery in databases" o KDD) es un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos.
- Utiliza los métodos de aprendizaje automátizado, estadística y sistemas de bases de datos.

Big Data (BD)

- Big Data (BD) también llamados datos masivos, inteligencia de datos, datos a gran escala es un término que hace referencia a conjuntos de datos tan grandes y complejos que precisan de aplicaciones informáticas no tradicionales de procesamiento de datos para tratarlos adecuadamente.
- El término "big data"tiende a referirse al análisis de los datos, extrayendo valor y formulando predicciones a través de los patrones observados.
- La disciplina dedicada a los datos masivos se enmarca en el sector de las tecnologías de la información y la comunicación.

Soft computing (SC)

- En contraste con **Hard computing**: algoritmos que encuentran soluciones demostrablemente correctas y óptimas a los problemas.
- **Soft computing** es una rama de la AI que abarca diversas técnicas para solucionar problemas con información incompleta, con incertidumbre y/o inexacta.
- Se enfoca a resolver problemas que resultaban inmanejables con los métodos analíticos y matemáticos convencionales.
- Soft Computing se incluyen: redes neuronales, sistemas difusos, algoritmos evolutivos, optimización de colonias de hormigas, inteligencia de enjambre, etc.

Ciencia de Datos (DS)

- El Data Science (o "ciencia de los datos" en castellano), como su nombre indica, trata sobre los datos.
- Es un campo multidisciplinario centrado en extraer INFORMACIÓN (insights) que puede ayudar a una empresa a tomar mejores decisiones.
- Data Science se superpone al campo de la inteligencia artificial en muchas áreas.
- Se utilizan herramientas como modelos estadísticos, métodos de visualización, pruebas de hipótesis y algoritmos de aprendizaje automátizado.

Ciencia de Datos (DS)

La Ciencia de Datos (DS) se ha aplicado en diferentes campos:

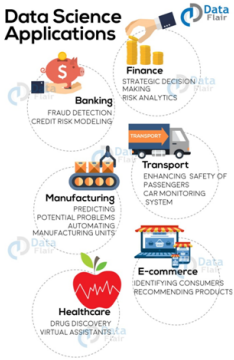

Figura 1: Campos de aplicación de la Ciencia de Datos (DS).

Relación entre IA, ML, DL, DS, BD

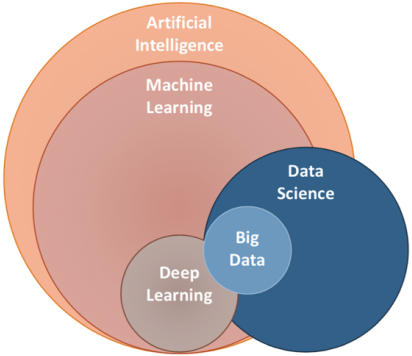

Figura 2: Relación entre IA, ML, DL, DS y BD.

Relación entre IA, ML, NN, DL

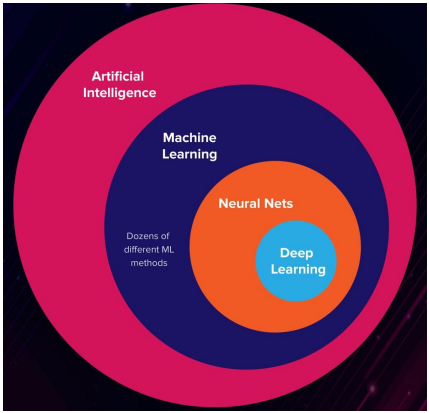

Figura 3: Relación entre IA, ML, NN, y DL.

Relación entre DS, DA, BD, DA, DM, ML

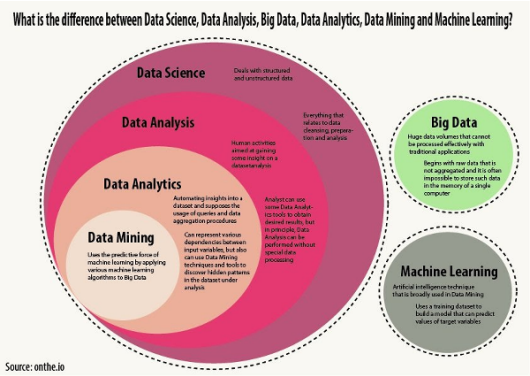

Figura 4: Relación entre DS, DA, BD, DA, DM, ML.

Aprendizaje Automatizado (ML)

Métodos de

Aprendizaje Automatizado (ML)

Línea de Tiempo de los Métodos de ML

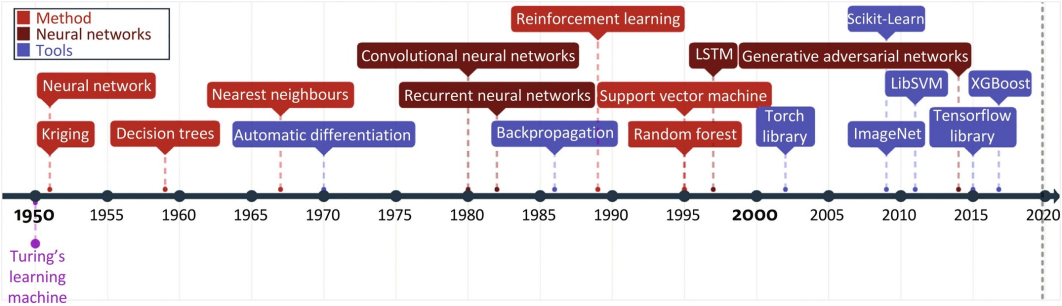

Figura 5: Línea de Tiempo de los Métodos de ML.

Tipos de redes neuronales I (NN)

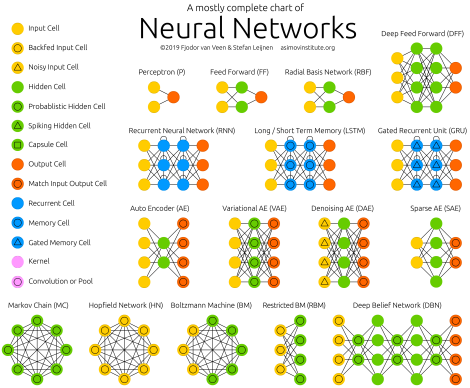

Figura 6: Tipos de redes neuronales (Leijnen and Veen, 2020 [3]).

Tipos de redes neuronales II (NN)

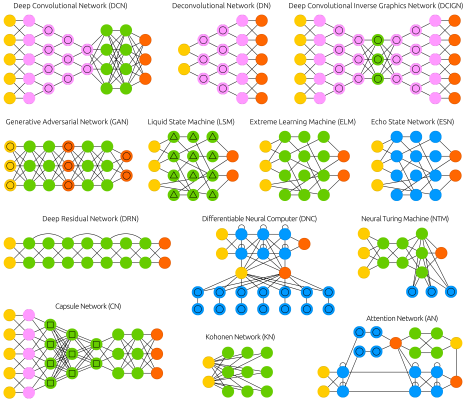

Figura 7: Tipos de redes neuronales (Leijnen and Veen, 2020 [3]).

Ciencia de Datos

Lenguajes, entornos y plataformas de la Ciencia de Datos

Ciencia de Datos

- Es un campo interdisciplinario que, con métodos científicos, extrae conocimiento de los datos.
- Es una combinación de campos de análisis de datos como estadísticas, minería de datos, aprendizaje automático y análisis predictivo.
- Incluye matemáticas, estadísticas, ciencia de datos e informática.
- Es un cuarto paradigma de la ciencia: empírico, teórico, computacional y ahora basado en datos.

Científico de Datos

El científico de datos es una mezcla de estadístico, computólogo y pensador creativo, con las siguientes habilidades:

- Recopilar, procesar y extraer valor de las diversas y extensas bases de datos.
- Comprender, visualizar y comunicar sus conclusiones a los no científicos de datos.
- Crear soluciones basadas en datos que aumentan los beneficios, reducen los costos.
- Trabajar en todas las industrias y hacer frente a los grandes proyectos de datos en todos los niveles.

Proceso básico de la Ciencia de Datos

El proceso básico se puede resumir en estos pasos:

- Extraer datos, independientemente de la fuente y de su volumen.
- Limpiar los datos, para eliminar lo que pueda sesgar los resultados.
- Procesar los datos usando métodos estadísticos como inferencia estadística, modelos de regresión, pruebas de hipótesis, etc.
- Diseñar experimentos adicionales en caso de ser necesario.
- Crear visualizaciones gráficas de los datos relevantes de la investigación.

Lenguajes de programación

- R
- Python
- SQL
- Julia
- Otros: JavaScript, Java, Matlab/Octave, C/C++, etc

Lenguaje de programación: R

- R es un lenguaje de código abierto y se ejecuta en todas las plataformas principales
- R es un lenguaje de programación para estadística comúnmente utilizado para análisis estadístico, visualización de datos y otras formas de manipulación de datos.
- R se ha vuelto cada vez más popular entre los científicos de datos debido a su facilidad de uso y flexibilidad en el manejo de análisis complejos en grandes conjuntos de datos.
- R ofrece muchos paquetes para algoritmos de aprendizaje automático, como regresión lineal, algoritmo de vecino más cercano, bosque aleatorio, redes neuronales, etc.

Lenguaje de programación: Python

- Python es un lenguaje de programación de propósito general que puede usarse para desarrollar cualquier software.
- Python es uno de los principales lenguajes de programación para la ciencia de datos.
- Python es conocido por su sintaxis simple, fácil lectura y portabilidad de código.
- También es de código abierto y se ejecuta en todas las plataformas principales

Lenguaje de programación: SQL

- SQL es un lenguaje declarativo para interactuar con bases de datos y le permite crear consultas para extraer información de sus conjuntos de datos.
- SQL se pueden ejecutar de forma interactiva desde una ventana de terminal o mediante secuencias de comandos integradas en otros programas de software.
- SQL en ciencia de datos ayuda a los usuarios a recopilar datos de las bases de datos y luego editarlos si la situación lo requiere.

Lenguaje de programación: Julia

- Julia es un lenguaje importante para la ciencia de datos que pretende ser simple pero potente, con una sintaxis similar a MATLAB o R.
- Julia también tiene un shell interactivo que permite a los usuarios probar el código rápidamente sin tener que escribir programas completos simultáneamente.
- Julia es rápido y eficiente en memoria, lo que lo hace ideal para conjuntos de datos a gran escala.
- La codificación es mucho más rápida e intuitiva, ya que le permite concentrarse en el problema sin preocuparse por las declaraciones de tipo.

Entornos y plataformas de programación

- Integrated Development Environments (IDEs): RStudio, Spyder, PyCharm, etc
- Jupyter (**Ju**-Julia, **pyt**-Python, **er**-R): Notebooks
- Colab (Colaboratory de Google)
- DataCamp (https://www.datacamp.com)

Entornos y plataformas de programación: RStudio

- **RStudio** es un entorno de desarrollo integrado para el lenguaje de programación R, dedicado a la computación estadística y gráficos.
- Incluye una consola, editor de sintaxis que apoya la ejecución de código, así como herramientas para el trazado, la depuración y la gestión del espacio de trabajo.
- RStudio está disponible para Windows, Mac y Linux
- Permite un análisis y desarrollo para que cualquiera pueda analizar los datos con R.

Entornos y plataformas de programación: Spyder

- **Spyder** es un entorno de desarrollo integrado multiplataforma de código abierto para la programación científica en el lenguaje Python.
- Spyder se integra con varios paquetes destacados en la pila científica de Python, incluidos NumPy, SciPy, Matplotlib, pandas, IPython, SymPy y Cython, así como otro software de código abierto.
- Inicialmente creado y desarrollado por Pierre Raybaut en 2009, Spyder ha sido mantenido y mejorado continuamente desde 2012 por un equipo de desarrolladores científicos de Python y la comunidad.
- Se publica bajo la licencia MIT

Entornos y plataformas de programación: Jupyter

- El Proyecto **Jupyter** es una organización sin ánimo de lucro creada para "desarrollar software de código abierto, estándares abiertos y servicios para computación interactiva en docenas de lenguajes de programación".
- El nombre del proyecto Jupyter es una referencia a los tres lenguajes de programación principales soportados por Jupyter, que son Julia, Python y R, y también un homenaje a los cuadernos de Galileo que registran el descubrimiento de los satélites de Júpiter.
- El proyecto Jupyter ha desarrollado y respaldado los productos de computación interactiva Jupyter Notebook, JupyterHub y JupyterLab, la versión de próxima generación de Jupyter Notebook.

Entornos y plataformas de programación: Colab

- **Colaboratory** (Colab para abreviar) es una herramienta de análisis de datos y aprendizaje automático que le permite combinar código Python ejecutable y texto enriquecido junto con gráficos, imágenes, HTML, LaTeX y más en un solo documento almacenado en Google Drive.
- Con Colab, puede aprovechar todo el poder de las bibliotecas populares de Python para analizar y visualizar datos.
- Con Colab, puede importar un conjunto de datos de imágenes, entrenar un clasificador de imágenes en él y evaluar el modelo, todo en solo [unas pocas líneas de código]
- Las notebooks de Colab ejecutan código en los servidores en la nube de Google, lo que significa que puede aprovechar la potencia del hardware de Google, incluidas [GPU y TPU]

Entornos y plataformas de programación: DataCamp

- **DataCamp** es una plataforma de aprendizaje en línea, establecida en el 2014.
- Este sitio se especializa casi completamente en cursos relacionados con los temas para estudiar Ciencia de Datos y programación con Python, R y SQL.
- La plataforma educativa tiene como objetivo enseñarle a las personas habilidades básicas para trabajar con datos - todo desde la comodidad de su casa (o cualquier lugar con conexión a internet).

Aplicaciones en Geociencias

- La Ciencia de Datos tiene una amplia gama de aplicaciones en las Geociencias,
- Permitiendo a los profesionales del campo analizar datos complejos, modelar fenómenos geológicos, y tomar decisiones informadas.
- Algunas de las principales aplicaciones de la Ciencia de Datos en las Geociencias son:

Aplicaciones en Geociencias

Exploración y Explotación de Recursos Naturales

- Utilización de algoritmos de aprendizaje automático para identificar posibles ubicaciones de depósitos minerales o yacimientos de petróleo y gas.
- Análisis de datos geoquímicos y geofísicos para evaluar la calidad y la cantidad de recursos naturales.

Aplicaciones en Geociencias

Modelado Geoespacial y Predicción

- Desarrollo de modelos predictivos para prever fenómenos geológicos, como deslizamientos de tierra, terremotos o cambios en la cobertura del suelo.
- Integración de datos geoespaciales para comprender patrones de distribución y cambios temporales en características geológicas.

Aplicaciones en Geociencias

Caracterización de Yacimientos

- Aplicación de técnicas de análisis de datos para modelar la heterogeneidad de los yacimientos y predecir propiedades petrofísicas.
- Uso de métodos de simulación estocástica para generar modelos tridimensionales de yacimientos basados en datos geológicos y de pozos.

Aplicaciones en Geociencias

Monitoreo Ambiental y Cambio Climático

- Análisis de datos satelitales y de sensores remotos para monitorear cambios en la cobertura terrestre, la calidad del agua y la vegetación.
- Modelado climático y predicción de cambios climáticos a largo plazo basados en datos históricos y observaciones actuales.

Aplicaciones en Geociencias

Geoinformática y Sistemas de Información Geográfica (SIG)

- Utilización de SIG para integrar, visualizar y analizar datos geoespaciales, incluyendo mapas, imágenes de satélite y datos de campo.
- Desarrollo de algoritmos espaciales para realizar análisis de proximidad, interpolación y derivación de información geográfica.

Aplicaciones en Geociencias

Gestión de Recursos Hídricos y Subsuelo

- Modelado de acuíferos y simulación de flujos subterráneos para la gestión sostenible de los recursos hídricos.
- Estudio de la interacción entre aguas subterráneas y superficiales mediante técnicas de análisis de datos.

Aplicaciones en Geociencias

Aprendizaje Automátizado en Interpretación Sísmica

- Aplicación de técnicas de aprendizaje automátizado para mejorar la interpretación de datos sísmicos, identificar características geológicas y reducir el tiempo de interpretación.

Aplicaciones en Geociencias

Análisis de Riesgos Geológicos

- Evaluación de riesgos geológicos, como deslizamientos de tierra, mediante el análisis de datos geoespaciales y factores de predisposición.
- Modelado de escenarios para evaluar la probabilidad de eventos geológicos adversos.

Popularidad de la Ciencia de Datos

Sólo algunos datos:

- Se han abierto cursos, especializaciones, licenciaturas, maestrías y doctorados en Ciencia de Datos en la mayoría de las universidades del mundo.
- En la oferta y solicitud de empleos ocupa un lugar cimero en muchas áreas de la ciencia, las ingeniería y la industria.
- La industria petrolera no se queda atrás, empresas como Slumberger están contratando agresivamente a científicos de datos.
- En la SPE se abrió una seccion particular sobre el tema: "Data Science and Engineering Analytics".

Pero, ¿que sea tan popular significa que siempre sea la mejor opción?

Pero que la Ciencia de Datos sea tan popular,

¿significa que siempre sea la mejor opción?

ML: ¿Cuando es adecuado aplicar?

SI

- Conjuntos masivos de datos.
- No existe una relación clara entre las variables.
- Involucra una gran cantidad de factores (multifactorial).
- Son eficaces para la gestión de problemas a gran escala.

NO

- Si se tiene muy pocas observaciones o variables.
- Cuando existen modelos (físicos, matemáticos, numéricos, computacionales) bien establecidos.

ML: ¿Cuando es adecuado aplicar?

PERO

- Existe una alternativa intermedia en pleno desarrollo:

"Los métodos de aprendizaje automatizado informados por la física"[2]

Referencias

[1] DellÁversana, P. , "Cross-disciplinary Machine Learning - Part 1: Applications to rock classification, well log analysis and integrated geophysics", 182 pags. (2020).

[2] Karniadakis, G. E., I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Reviews Physics* **3 (6)**, 422–440 pags. (2021).

[3] Leijnen, S. and Veen, F. V., *The neural network zoo*. Proceedings, **47(1)** (2020).

[4] Rosenblatt, F., "The Perceptron: A Probabilistic Model For Information Storage And Organization In The Brain", *Psychological Review* **65 (6)** 386-408 (1958).

[5] Samuel, A. L. . "Some studies in machine learning using the game of checkers", *IBM Journal of Research and Development*, **3:3** 210–229 (1959).

Siguiente tema:

Conceptos básicos de estadística